The Human Motion Behavior Recognition by Deep Learning Approach and the Internet of Things

Hui Li[1], Huayang Liu[2], Wei Zhao[3], Hao Liu[4]*

[1] Department of Physical Education, Luoyang Institute of Science and Technology, Luoyang, 471023 (China)
[2] College of Physical Education China West Normal University, Nanchong,637001 (China)
[3] College of Humanities and Social Sciences, Luoyang Institute of Science and Technology, Luoyang,471023 (China)
[4] College of Physical Education and Health, Guangxi Medical University.Nanning,530021 (China)

* Corresponding author: linyiliuhao@sina.com

Received 25 April 2023 | Accepted 17 May 2024 | Early Access 24 July 2024

Abstract

This paper is dedicated to exploring the practical implementation of deep learning and Internet of Things (IoT) technology within systems designed for recognizing human motion behavior. It places a particular emphasis on evaluating performance in complex environments, aiming to mitigate challenges such as poor robustness and high computational workload encountered in conventional human motion behavior recognition approaches by employing Convolutional Neural Networks (CNN). The primary focus is on enhancing the performance of human motion behavior recognition systems for real-world scenarios, optimizing them for real-time accuracy, and enhancing their suitability for practical applications. Specifically, the paper investigates human motion behavior recognition using CNN, where the parameters of the CNN model are fine-tuned to improve recognition performance. The paper commences by delineating the process and methodology employed for human motion recognition, followed by an in-depth exploration of the CNN model's application in recognizing human motion behavior. To acquire data depicting human motion behavior in authentic settings, the Internet of Things (IoT) is utilized for extracting relevant information from the living environment. The dataset chosen for human motion behavior recognition is the Royal Institute of Technology (KTH) database. The analysis demonstrates that the network training loss function reaches a minimum value of 0.0001. Leveraging the trained CNN model, the recognition accuracy for human motion behavior achieves peak performance, registering an average accuracy of 94.41%. Notably, the recognition accuracy for static motion behavior generally exceeds that for dynamic motion behavior across different models. The CNN-based human motion behavior recognition method exhibits promising results in both static and dynamic behavior recognition scenarios. Furthermore, the paper advocates for the use of IoT in collecting human motion behavior data in real-world living environments, contributing to the advancement of human motion behavior recognition technology and its application in diverse domains such as intelligent surveillance and health management. The research findings carry significant implications for furthering the development of human motion behavior recognition technology and enhancing its applications in areas such as intelligent surveillance and health management.

Keywords

Behavior Recognition, Convolutional Neural Network, Human Body Movement, Internet of Things.

DOI: 10.9781/ijimai.2024.07.004

I. Introduction

In the domain of computer vision, human motion behavior recognition stands as a prominent and widely acknowledged subject of interest. This field holds considerable application value in various domains, including intelligent monitoring, robotics, human-computer interaction (HCI), virtual reality, smart home technologies, smart security systems, and athletic training assistance [1]. A practical application of human motion analysis is content-based video retrieval, allowing for efficient retrieval of specific athlete movements, such as those observed during horizontal bars competitions in sports events. This technology not only saves users significant time and effort in querying video data [2] but also facilitates the extraction of various technical parameters, such as joint position, angle, and angular speed, contributing valuable guidance and suggestions for athletes' training and overall improvement. Moreover, this application finds relevance in sports dance movement analysis and clinical orthopedic research. Exploring human motion-tracking research presents a range of

Please cite this article as:
H. Li, H. Liu, W. Zhao, H. Liu. The Human Motion Behavior Recognition by Deep Learning Approach and the Internet of Things, International Journal of Interactive Multimedia and Artificial Intelligence, (2024), http://dx.doi.org/10.9781/ijimai.2024.07.004

theoretical and practical challenges within fields like computer vision, pattern recognition, and video image processing, especially when considering the non-rigid nature of the human body undergoing rotational motion of joints [3].

Previous research has demonstrated the significant success of convolutional neural network (CNN) models in image and video recognition tasks. However, for complex human motion behavior recognition, CNN faces challenges in terms of robustness. This paper proposes a method to enhance human motion behavior recognition by optimizing CNN model parameters. Additionally, it categorizes human motion behavior into static and dynamic actions, investigating the recognition results for different types of motion behaviors to improve accuracy. The application of the Internet of Things (IoT) introduces more diverse data scenarios for human motion behavior recognition, thereby enhancing recognition accuracy and robustness. Collecting diversified data via IoT better represents different aspects of human motion behavior, leading to a more comprehensive and accurate recognition [4]. Several challenges in human motion behavior recognition research include the immense variability in human movement patterns, resulting in identical movements manifesting in different behavioral performances. Additionally, the wide array of movements within human motion exhibits numerous distinct manifestations, posing challenges in accurately identifying human motion behaviors. Diverse viewpoints can yield various two-dimensional images of the same action, and occlusions between individuals and backgrounds present difficulties in early-action classification during feature extraction. The proposed system's insensitivity to video playback rates, exacerbated by dynamic and cluttered backgrounds, fluctuating ambient lighting conditions, and low image and video resolution, further complicates the recognition process [5]. Some researchers have addressed multi-vision and occlusion problems through the proposal of multi-camera fusion [6], a technique managed through three-dimensional (3D) reconstruction [7].

The incorporation of human behavior recognition technology in community management holds the potential to establish an efficient and secure intelligent service system. This technology enables real-time behavior recognition for individuals and groups within community monitoring. In situations where hazardous behaviors, such as illegal entry and high-altitude throwing, go unnoticed by video surveillance personnel, the automated system can promptly alert community managers, mitigating safety risks and streamlining the subsequent evidence-gathering process. The integration of CNN models, based on supervised learning, with human behavior recognition technology represents a promising research avenue deserving further exploration. Through additional research in this domain, this paper provides a robust theoretical framework and reliable technical support, laying the groundwork for future practical endeavors.

The paper aims to investigate the practical application of deep learning and IoT technology in human motion behavior recognition systems. It underscores the evaluation of performance in complex environments and addresses the challenges of poor robustness and high computational workload in traditional human motion behavior recognition methods using CNN. The focus is on evaluating the performance of human motion behavior recognition systems in real-world scenarios and optimizing them for real-time accuracy, enhancing their suitability for practical applications. This paper proposes an enhanced CNN-based human motion behavior recognition method by seamlessly integrating IoT technology and the CNN model. This method can accurately and efficiently recognize various types of human motion behaviors. The proposed approach in this paper holds significant practical relevance in real-world applications, with extensive utility in areas such as surveillance, health management, and intelligent transportation. It offers valuable technical support for real-time monitoring and recognition of human motion behavior. Moreover, the proposed method can adapt to different scales and complexities of application scenarios, laying a solid foundation for future research and applications in behavior recognition.

The paper is structured into five sections to provide a cohesive framework. Section 1 functions as an introduction, offering insights into the research background and the underlying motivation behind the paper. Section 2 presents a comprehensive review covering methodologies for target motion detection and the application of deep learning techniques in addressing challenges related to target recognition. Section 3 constitutes the paper's focal point, concentrating on the intricacies of human motion recognition. This section introduces an adaptive correlation learning module specifically based on traditional CNN, effectively calculating correlation weights between samples to enhance the recognition process. In Section 4, a series of meticulously designed experiments are conducted to empirically validate the performance of the proposed algorithm. Furthermore, the algorithm's practical significance and real-world applicability are extensively discussed. Finally, Section 5 serves as a succinct summary, encapsulating the essential findings and insights conveyed throughout the entirety of the paper.

II. Literature Review

In the field of behavior recognition technology, Guo et al. proposed a method for foreground target motion detection where the background of the target motion remains unchanged. This approach leverages the changing background as a foreground for discriminating targets. However, if the target remains stationary for a certain period, the static part is updated to the background, making it unidentifiable. Therefore, the construction of a robust background model capable of adaptive updates becomes crucial [8]. In a distinct context, Wei et al. introduced the fusion of Parametric Rectified Linear Units (ReLU) and robust initialization methods within a CNN to address applications in the ImageNet 2012 dataset. Their findings indicated that the recognition rate of the behavioral dataset surpassed that discernible by the human eye [9]. Acknowledging the complexity of the provided information, Bolanos et al. categorized behavior analysis methods into three classes: static gestures, motor behaviors, and recognition of complex processes. The static gesture recognition method primarily identifies the target's gesture within a static single-frame image, while motion behavior and complex process recognition focus on identifying video motion events [10]. Chebbout et al. performed human behavior recognition based on the spatial-temporal volume model of behavior recognition, which involves the projection of the human body onto the time axis and template matching [11]. Peng et al. engaged in feature extraction of 3D-Scale-invariant feature transform (3D-SIFT) points of interest and established a feature-based statistical histogram construction of video interest classes on the codebook. This eigenvector was then utilized for identification and training on Support Vector Machines (SVM). Lastly, leveraging the space-time trajectory method, key points in human motion were connected along the time axis to form a trajectory curve [12]. Chen et al. undertook the identification of multi-feature channels in a 3D-CNN, encompassing a grayscale image, vertical and horizontal gradients, and optical flow. Each input video sample comprised seven consecutive frames, ensuring effective utilization of time domain information. The experiments showcased the network's commendable recognition rate across real-world and Royal Institute of Technology public databases [13].

The video employs an innovative network architecture designed for human motion behavior recognition. Lin et al. formulated a 3D-CNN and conducted training on the University of Central Florida 101 (UCF101) dataset. During network training, eight convolution and four pooling operations were executed. The convolution core had dimensions of 3*3*3 with a stride of 1*1*1 [14]. In a separate study, Zhi et al. employed a 3D-CNN to extract spatial and motion features, integrating dense trajectory features into Long Short-Term Memory (LSTM) networks and embedding time series information. Subsequently, they utilized weighted averaging of the output from multiple LSTM units to obtain recognition results [15]. Jin et al. introduced a bidirectional CNN that incorporates both spatial and temporal information. This network employed two distinct paths to capture appearance information from a static frame and motion information between two frames, effectively enabling motion recognition [16]. Lu et al. developed a Trajectory-pooled CNN model, combining manual feature extraction with feature extraction derived from CNN models for motion recognition [17]. Furthermore, Guenzi et al. merged Deep Learning (DL) with Slow Feature Analysis (SFA), resulting in the construction of the Slow Feature Analysis-Deep Learning (SFA-DL) network specifically designed for behavior recognition [18]. Table I below illustrates the strategies, algorithms, and principal contributions adopted by different researchers in addressing behavior recognition problems.

Previous investigations have highlighted that targets exhibit not only spatial attributes but also temporal characteristics throughout their motion processes. The analysis of intricate behavioral processes places emphasis on human interactions and group behavior. In contrast to traditional human behavior recognition methods, CNN presents a considerable advantage by obviating the necessity for manual feature extraction. Instead, the network assimilates the characteristics that delineate the target's behavior and acts upon the target without prior experiential input. In the actual collection of human motion behavior data, challenges such as noise and incompleteness may arise, including image blurriness, occlusion, and missing data. These factors can influence the performance of the recognition model. In this study on human motion action recognition using CNN, the CNN model undergoes optimization to accommodate variations in complex environments and lighting conditions. This optimization reduces the model's parameter size and complexity, thereby mitigating computational workload, improving real-time performance and efficiency, and augmenting the CNN model's effectiveness in recognizing human motion behavior. For action behavior recognition in video sequences, this paper integrates multiple modalities of information, such as depth images and motion sensors, to capture temporal variations in action behavior, thereby enhancing recognition accuracy and robustness. Additionally, to address potential noise and incompleteness in the collected data from previous research, this paper utilizes the KTH database as the dataset for human motion behavior recognition experiments. The dataset undergoes preprocessing and enhancement, eliminating noise and supplementing missing data to fortify the model's robustness and accuracy.

III. Research Methodology

A. Research Approach

The principal aim of this paper is to proficiently utilize both labeled and unlabeled data, extracting valuable information to achieve optimal performance in semi-supervised behavior recognition. In order to fulfill this objective, this paper proposes a semi-supervised algorithm grounded in adaptive correlation learning. This algorithm capitalizes on the feature characteristics of samples to explore correlations between them and incorporates the acquired correlation information in the process of feature aggregation. Through the aggregation of features from neighboring samples for each sample, the algorithm generates more expressive and discriminative feature representations. The training process of the semi-supervised algorithm based on adaptive correlation learning is delineated in Fig. 1.

In Fig. 1, the initial step involves the preparation of both the training set and the unlabeled dataset. The training set comprises labeled samples designed for supervised learning, while the unlabeled

TABLE I. Strategies, Algorithms, and Main Contributions of Different Researchers in Behavior Recognition

Researcher	Strategies and Algorithms	Main Contributions
Guo et al. [8]	They proposed foreground target action detection method, the optimized background model	The research developed an adaptive and robust background model
Wei et al. [9]	They proposed CNN with the fusion of Parametric ReLU and robust initialization methods	The research achieved behavior recognition rates superior to human eyes on ImageNet 2012 dataset
Bolanos et al. [10]	They categorized behavior analysis into the static posture, motion behavior, and complex processes	The research classified recognition of video motion events and static postures
Chebbout et al. [11]	They used spatiotemporal volume model and template matching for human behavior recognition	The research introduced a novel approach to human behavior recognition
Peng et al. [12]	They employed 3D-SIFT and SVM for feature extraction and training of human actions	The research achieved favorable recognition results based on the spatial-temporal trajectory method
Chen et al. [13]	They utilized 3D-CNN multi-feature channels for human action recognition	The research demonstrated good recognition results in real-world and KTH public databases
Lin et al. [14]	They developed 3D-CNN for training the UCF101 dataset	The research employed specific network structure for human action recognition
Zhi et al. [15]	They used 3D-CNN to extract spatial and motion features, fused with LSTM for recognition	The research utilized LSTM's weighted average output for recognition results
Jin et al. [16]	They build bidirectional CNN for capturing spatiotemporal information	The research combined two pathways for motion and appearance recognition
Lu et al. [17]	They developed trajectory aggregation CNN model	The research integrated manual feature extraction and CNN model for motion recognition
Guenzi et al. [18]	They combined Deep Learning with Slow Feature Analysis	The research constructed SFA-DL network for behavior recognition

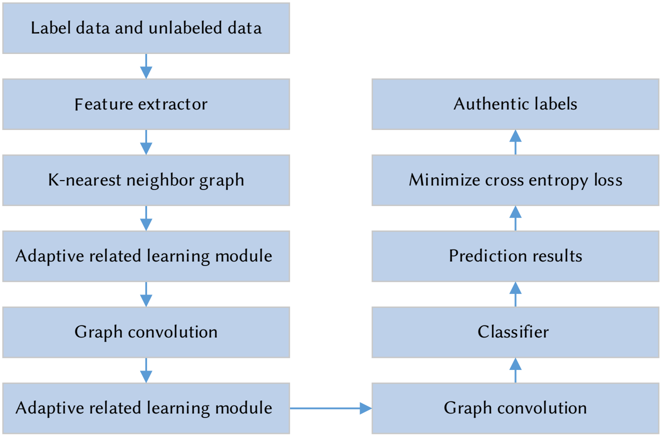

Label data and unlabeled data
Feature extractor
K-nearest neighbor graph
Adaptive related learning module
Graph convolution
Adaptive related learning module
Graph convolution
Classifier
Prediction results
Minimize cross entropy loss
Authentic labels

Fig. 1 Training process of semi-supervised algorithm based on adaptive correlation learning.

dataset includes samples devoid of labels, utilized in semi-supervised learning. Feature extraction is subsequently executed for both the training set and the unlabeled dataset, transforming raw data into more representative feature representations for subsequent learning. A pivotal stage in the semi-supervised algorithm based on adaptive correlation learning is correlation learning. The primary objective is to scrutinize the correlations between samples and incorporate this correlation information into the feature aggregation process. Unlabeled data is employed to learn these correlations. Following correlation learning, feature aggregation ensues to generate feature representations that are more expressive and discriminative. This is accomplished by aggregating the features of each sample with those of its neighboring samples, employing methods such as weighted averages or maximum pooling. Upon obtaining feature representations, a semi-supervised learning approach is employed to train the classifier. Semi-supervised learning combines both labeled and unlabeled samples for training, enhancing the classifier's generalization ability and accuracy by leveraging information from unlabeled samples. Ultimately, post-training, the model's performance on new samples is evaluated using either a validation set or cross-validation. The assessment results contribute to evaluating the model's effectiveness and generalization ability.

B. Recognition of Human Exercise Behavior

Human motion behavior recognition constitutes a pivotal research avenue within the realms of computer vision and pattern recognition, aiming to autonomously discern and comprehend diverse human motion actions through the implementation of computer algorithms and deep learning models. In the sphere of intelligent surveillance, the application of human motion behavior recognition technology in video surveillance systems facilitates the analysis and identification of pedestrian, vehicle, and other object behaviors. This integration enables functionalities like intelligent alerts and anomaly detection in behavior, significantly enhancing the efficiency and accuracy of surveillance systems. Such precise recognition contributes markedly to security personnel's ability to detect potential security risks. Within the domain of HCI, human motion behavior recognition technology finds utility in natural interaction, encompassing aspects such as posture recognition and gesture control. Identification of users' actions and postures enables computers to comprehend their intentions, thereby enhancing the convenience and intelligence of HCI. In the field of health management, the application of human motion behavior recognition technology extends to motion monitoring and rehabilitation assistance. Monitoring and analyzing human motion behavior allow for the assessment of individual movement status and the monitoring of movement performance. This, in turn, provides scientific evidence and personalized guidance for rehabilitation training, thereby augmenting rehabilitation outcomes. Despite the aforementioned applications, traditional human motion behavior recognition methods encounter challenges related to pose variations, complex backgrounds, and lighting changes, resulting in diminished recognition accuracy and weak robustness. The advent of deep learning, particularly the implementation of CNN, has substantially progressed human motion behavior recognition. CNN models demonstrate an inherent capacity to automatically learn features from data, exhibiting robust representational capabilities and adaptability. This advancement significantly elevates the accuracy and robustness of human motion behavior recognition.

The exhaustive analysis of human motion behavior encompasses several integral processes, including database construction, human motion detection, feature extraction, behavior comprehension, and recognition. The core emphasis in human motor behavior analysis lies in motion detection and feature extraction [19], as illustrated in the schematic representation of the recognition process for human motor behavior in Fig. 2.

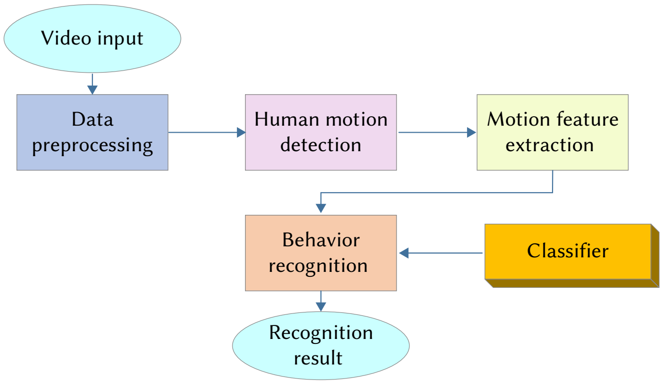


Fig. 2 Recognition process of human motion behavior.

As depicted in Fig. 2, the process of target classification detection entails extracting the region of interest from the foreground motion area identified by a moving object [20]. In intricate scenes, the foreground areas may encompass diverse targets, including pedestrians, vehicles, birds, clouds, swaying branches, among others. However, within the context of the human motion analysis system, the detection target is exclusively restricted to human movement. Hence, it becomes imperative to meticulously scrutinize, analyze, identify, and isolate human targets. The method utilized for detecting target classification is delineated in Fig. 3.

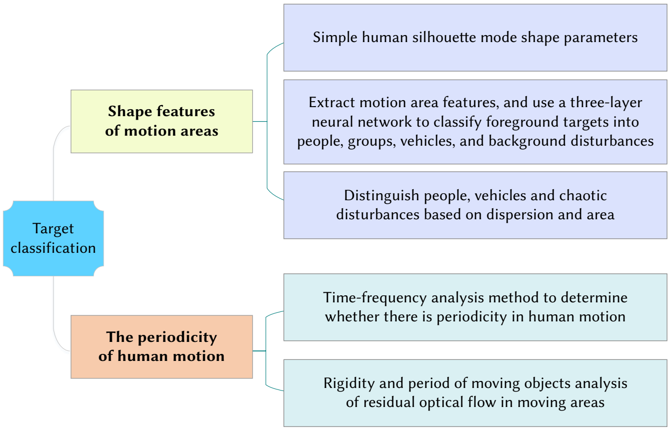

Target classification
Shape features of motion areas
Simple human silhouette mode shape parameters
Extract motion area features, and use a three-layer neural network to classify foreground targets into people, groups, vehicles, and background disturbances
Distinguish people, vehicles and chaotic disturbances based on dispersion and area
The periodicity of human motion
Time-frequency analysis method to determine whether there is periodicity in human motion
Rigidity and period of moving objects analysis of residual optical flow in moving areas

Fig. 3 Detection target classification method.

Fig. 3 entails the classification of detected targets based on the shape characteristics of the motion area and the periodicity of human movement. The assessment of residual light flow within the moving area enables the analysis of the rigidity and periodicity of the moving target. This approach proves effective due to the relatively higher residual light flow in non-rigid human movements compared to the movements of rigid vehicles, thereby facilitating the discrimination of human bodies. Currently, numerous video-based classification methods for moving objects are available, including those based on static features, dynamic features, or a combination of both. However, a single moving target feature often falls short of recognizing more than three targets or achieving satisfactory recognition accuracy. As a result, target classification studies typically select a minimum of two features. General features are indicative of characteristics applicable to all objects, while attribute characteristics represent the inherent qualities of a target, specifically reflecting its unique attributes.

Given the presence of multiple angles in the experimentally extracted foreground targets, including shadows and incomplete target area extraction, shape-based feature classification proves more suitable under such circumstances. The shape-based features of the target encompass target contour, area, aspect ratio, dispersion, centroid, and bounding rectangle [21]. The attribute characteristics of the moving target are delineated in Table 2.

TABLE II. Attribute Characteristics of Moving Targets

Sports Goals	Attribute characteristics
People	Circular Periodicity of human motion (regular changes in human gait)
Automobile	Movement speed Variation in dispersion (variation of each target)
Bicycles	The attribute is somewhere between person and automobile.

In the present study, a wide array of target features is extensively employed, encompassing aspects such as aspect ratio, area information, dispersion (regional compactness), inertial principal axis direction, invariant moment, and other regional characteristics. For experimental purposes, several attributes are defined, including the ratio of target height to the width of the target area at approximately one-third of the height, the ratio of target height to width at about two-thirds of the target area height, and the duty cycle, defined as the ratio of the background area within the target boundary rectangle to the area of the boundary rectangle. Notably, the aspect ratio feature signifies the aspect ratio of the entire target. The analysis of moving object characteristics is demonstrated using the moving target within the scene, as depicted in Fig. 4.

Fig. 4 conducts classification on the extracted moving objects, distinguishing between "bicycle" and "automobile," "automobile" and "pedestrian," as well as "crowd" by detecting the moving objects present in the scene. Remarkably, the aspect ratio of the target "person" and "automobile" demonstrates significant differences.

C. CNN Modeling

CNN represents a variant of the Multilayer Perceptron (MLP) originating from early research conducted by biologists Hubel and Wiesel on the cat visual cortex [22]. The architecture of the CNN is delineated in Fig. 5.

In Fig. 5, the architecture of the CNN involves convolution, subsampling, and fully connected layers. Each level in the CNN comprises multiple feature maps, with each feature map extracting unique features from the input through a convolution filter, housing multiple neurons. Local features are extracted as the input image and filters undergo convolution, determining their relationships with other features. The neurons in each layer receive the same input as the previous layer, establishing connected local receptive fields. The subsequent layer following each feature extraction layer is the computation layer responsible for local averaging and secondary extraction, also referred to as the feature mapping layer. This layer comprises multiple feature mapping planes with equal neuron weights. The mapping from the input layer to the hidden layer is commonly termed feature mapping. Consequently, the feature extraction layer is obtained through the convolution layer, while the feature mapping layer is achieved after downsampling [23]. The process of linking the convolution layer to the subsampling layer is illustrated in Fig. 6.

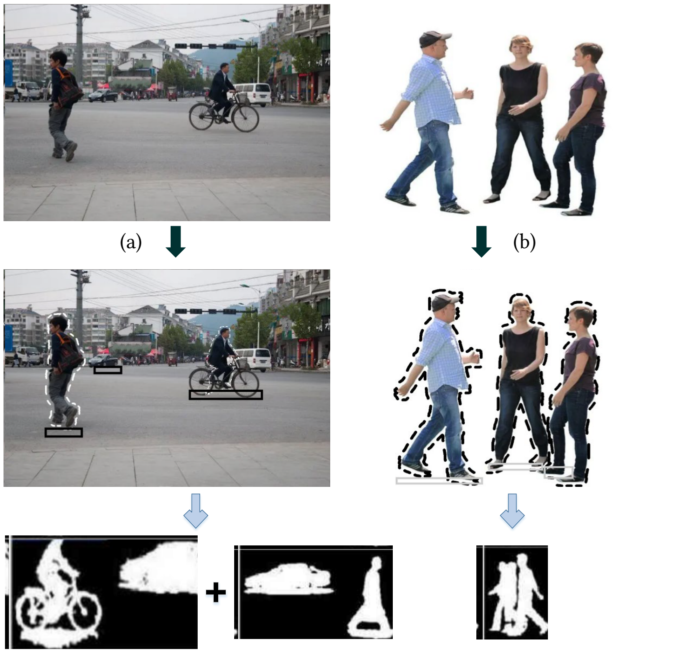


Fig. 4 Moving target in the scene (Data source: https://sucai.redocn.com/yixiang_6739023.html).

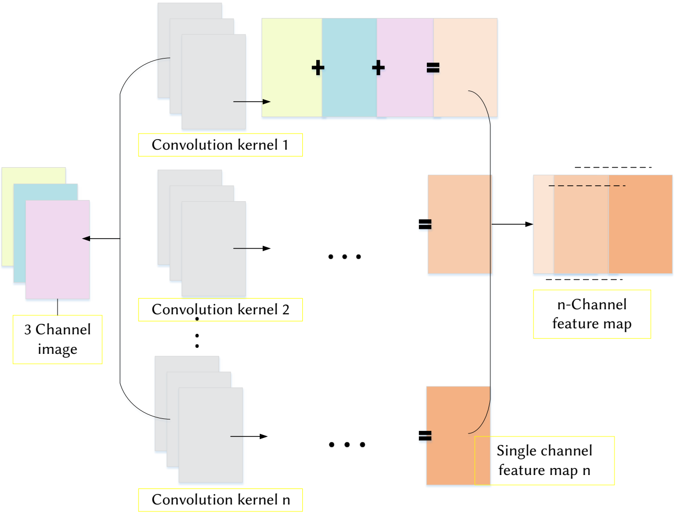


Fig. 5 Structure of CNN.

In Fig. 6, the input layer processes the normalized image, and each neuron within each layer receives input from a small local neighborhood of the previous layer. These neurons extract fundamental visual features, such as edges and corners, which are utilized by higher-level neurons. The CNN derives feature maps through the convolution operation, where cells at different locations acquire distinct features

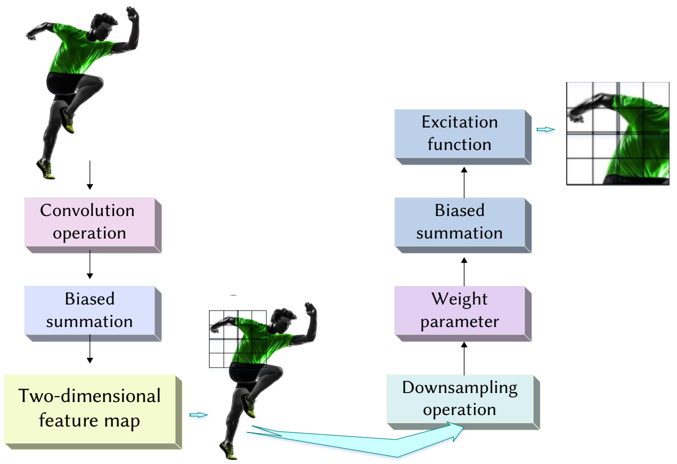


Fig. 6 Operation of connecting the convolution layer to the lower sampling layer.

from various feature maps. A convolution layer typically comprises multiple feature maps with different weight vectors, allowing for the retention of richer image features. Subsequently, the convolution layer is connected to the subsampling layer, serving dual purposes. Firstly, it reduces the image resolution and the number of parameters. Secondly, it fosters robustness to translation and deformation. The convolution and subsampling layers are interspersed throughout the network, progressively increasing the number of feature maps while decreasing the resolution [24], [25]. The calculation of the convolution layer is presented in Equation (1).

$$y_{mn} = f(\sum_{j=0}^{Q-1} \sum_{j=0}^{P-1} x_{m+i,n+j} w_{ij} + b),0 \le m < M, 0 \le n < N \tag{1}$$

In Equation (1), $x_{(m+i,n+j)}$ represents the pixel value of the input data; $(m+i,n+j)$ indicates the two-dimensional coordinates of the point; y_{mn} represents the output data after convolution; b is offset; $P \times Q$ is the size of the convolution kernel; w_{ij} represents the value of the convolution kernel in (i,j). M and N are the input image in $P \times Q$; f is the excitation function. The excitation function is shown in Equations (2) and (3):

$$f(x) = \frac{1}{1+e^{-x}} \tag{2}$$

$$ReLU(x) = \max(0,x) = \begin{cases} 0, x < 0 \\ x, x \ge 0 \end{cases} \tag{3}$$

In Equations (2) and (3), within the Sigmoid function, the output of $f(x)$ ranges between [0, 1] when a real number is input. If the output approaches zero, the neuron exhibits no response; conversely, if the output tends to 1, the neuron is activated. However, when the input is excessively large or small, the output approaches 1 or 0, respectively. In such cases, the neuron becomes saturated, impeding weight updates, and resulting in vanishing gradients. Contrastingly, in the ReLU activation function, when $x \ge 0$, the output of $f(x)$ is x, leading to rapid network convergence, and the neurons remain unsaturated. Consequently, computational efficiency is enhanced. Following convolution, the calculation of image feature size is expressed in Equation (4):

$$N = \frac{W-F+2P}{S} + 1 \tag{4}$$

In Equation (4), $W \times W$ represents the input image size; F is the size of the convolution kernel; P indicates filling. The step size is S; $N \times N$ represents the output image size. The sampling calculation expression of the down-sampling layer in $S_1 \times S_2$ is shown in Equation (5):

$$y_{mn} = f(w\frac{1}{S_1 S_2} \sum_{j=0}^{S_2-1} \sum_{j=0}^{S_1-1} x_{m\times S_1+i,n\times S_2+j} + b),0 \le m < M,0 \le n < N \tag{5}$$

In Equation (5), $x_{(m\times S_1+i,n\times S_2+j)}$ represents input data; $(m \times S_1 + i, n \times S_2 + j)$ represents the two-dimensional coordinates of the point; y_{mn} represents output data. b is offset; (S_1, S_2) is the pixel coordinate of the area; w is the weight value. Under the action of the common four excitation functions, a two-dimensional feature map is obtained with a resolution of ¼ of the original image, that is, S_{x+1} sample the layer for feature extraction again. The size of the feature map after downsampling is shown in Equation (6):

$$N = \frac{W-F}{S} + 1 \tag{6}$$

In Equation (6), $W \times W$ represents the input image size; F is the size of the downsampling template; the step size is S. $N \times N$ is the output image size. The network training process is divided into the forward and backward propagation stages. The calculation of forwarding propagation is shown in Equations (7), (8) and (9):

$$z^{(l)} = W^{(L)} \cdot a^{(l-1)} + b^{(l)} \tag{7}$$

$$a^{(l)} = f_l(z^{(l)}) \tag{8}$$

$$x = a^{(0)} \rightarrow z^{(1)} \rightarrow a^{(1)} \rightarrow z^{(2)} \rightarrow a^{(2)} \rightarrow \cdots \rightarrow z^{(l)} \rightarrow a^{(l)} \tag{9}$$

In Equations (7)-(9), l is the number of layers of CNN; $m^{(l)}$ is the number of neurons in l-layer; $W^{(L)}$ is the weight matrix; $b^{(l)}$ is offset; $a^{(l)}$ represents the output of l-layer neurons; $z^{(l)}$ is the input of l-layer neurons; $f_l(\cdot)$ is the activation function. Equation (9) represents the network prediction output $a^{(l)}$ of forward operation. The difference value of the backward propagation output layer is shown in Equation (10):

$$\delta^l = \frac{\partial J(W,b,x,y)}{\partial a^l} \odot \sigma'(z^l) \tag{10}$$

In Equation (10), δ^l represents the difference value of the output layer; $J(W, b, x, y)$ is the mean square error; z^l represents the input of layer l neurons; a^l is the output of l-layer neurons; $\sigma'(\cdot)$ indicates derivative operation. The calculation of δ^l is shown in Equation (11):

$$\delta^l = (W^{(l+1)})^T \cdot \delta^{(l-1)} \odot \sigma'(z^l) \tag{11}$$

All parameters (W, b) are updated, as shown in Equations (12) and (13).

$$\frac{\partial J(W,b,x,y)}{\partial W^l} = \delta^l (a^{l-1})^T \tag{12}$$

$$\frac{\partial J(W,b,x,y)}{\partial b^l} = \delta^l \tag{13}$$

D. IoT Technology

The IoT employs diverse connectivity technologies to meet connection requirements in various scenarios, including passive identification, short-distance wired, short-distance wireless, and long-distance wireless connections. The initial impetus for the IoT was driven by the emergence of Radio Frequency Identification (RFID) technology, although its passive reading nature limited its applicability in certain contexts. In the data collection and processing phase, the IoT integrates various sensor types into the network to capture different facets of human motion behavior. Visual sensors, such as cameras, are employed to obtain video data for recognizing human postures, actions, and motion trajectories. Motion sensors, like accelerometers, detect human movement status and acceleration changes. Concurrently, environmental sensors capture surrounding environmental information, such as light, temperature, and humidity, which may contribute to behavior recognition. Establishing real-time data transmission and communication mechanisms between sensors and the behavior recognition system is crucial. Through IoT technology, sensors can transmit collected data in real-time to the behavior recognition system, ensuring data timeliness and accuracy.

This facilitates real-time monitoring and recognition of motion behavior, thereby enhancing the system's real-time performance and efficiency [26],[27]. Additionally, the integration of edge computing into the IoT network allows local data processing and analysis, reducing the burden on centralized servers. In the behavior recognition process, tasks with high real-time requirements can be processed on edge devices where the sensors are located, mitigating the need to transmit all data to central servers for processing. This reduction in network latency improves response speed and lowers data transmission costs [28]. In order to enhance the energy efficiency of IoT devices, low-power sensors and energy-saving technologies are employed to extend sensor lifespan [29]. Energy harvesting techniques, such as solar charging or vibration energy harvesting, contribute to providing sustainable energy for sensors. By fully leveraging the potential of IoT, integrating various sensor types into the network, enabling real-time data transmission and communication, and incorporating edge computing technologies, an efficient, reliable, and real-time human motion behavior recognition system can be established [30]. Such a system is capable of addressing complex motion behavior recognition scenarios, enhancing accuracy and response speed, and providing comprehensive and reliable support for human motion behavior research and applications. The second wave of IoT was catalyzed by the maturity of short-range wireless networking technologies such as ZigBee and Wireless Fidelity (WIFI). Moreover, the ongoing evolution of cellular communication technology is anticipated to further facilitate the widespread adoption and advancement of IoT [31], [32]. In the context of ZigBee wireless networking technology, human motion behavior is collected in living environments, as illustrated in Fig. 7.

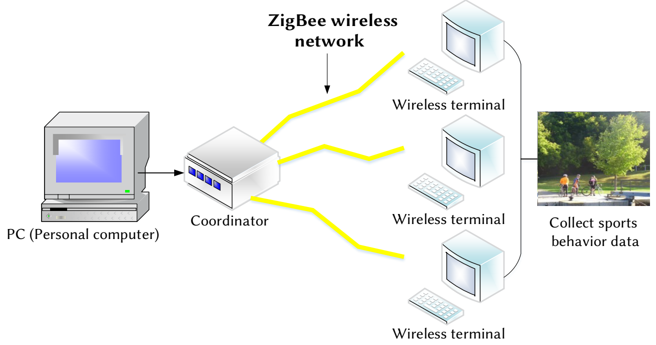


Fig. 7 Acquisition scenario under ZigBee wireless networking technology.

E. CNN Behavior Recognition Based on Adaptive Correlation Learning

In order to better represent the correlation between samples, this paper proposes an adaptive correlation learning module based on traditional CNN. This module can be used to calculate the correlation weights between samples. A shared learnable module is used to parameterize $corr(x_i, x_r)$ to obtain the specific value of w_{ir}, as shown in Equation (14):

$$w_{ir} = Z\left(Re\,L\,U\left(\frac{x_i \odot x_r}{\|x_i\|_2 \cdot \|x_r\|_2}\right)\right) \tag{14}$$

In Equation (14), Z represents a learnable weight vector and $\odot$ represents the Hadamard product. Re$LU(*)$ represents an activation function, which can increase the sparsity of features and enhance the nonlinearity of the adaptive correlation learning module.

In the original feature space, the dimensions of features are usually relatively large. The input features are mapped to a low dimensional feature space, as shown in Equation (15):

$$\hat{x}_i = W x_i \tag{15}$$

In Equation (15), W represents a learnable linear transformation matrix. An offset term can also be added to the calculation process of feature mapping.

The operation of feature mapping can not only reduce the dimensionality of input features but also enhance the expression ability of features to a certain extent, as shown in Equation (16):

$$w_{ir} = Z\left(Re\,L\,U(*)\left(\frac{\hat{x}_i \odot \hat{x}_r}{\|\hat{x}_i\|_2 \cdot \|\hat{x}_r\|_2}\right)\right) \tag{16}$$

Initially, video samples are denoted as graph nodes, and a graph structure is established using the K-nearest neighbor method, integrating both labeled and unlabeled data. Within this graph structure, adaptive correlation learning is implemented, and the original feature X of each video sample serves as the input feature for the initial layer of graph convolution. The adaptive correlation learning module calculates the adjacency matrix for each layer of graph convolution, thereby capturing correlation information. In the course of feature aggregation within the graph convolutional networks, this correlation information is harnessed to generate more expressive features for the video samples. This is achieved by aggregating the features of neighboring samples within local neighborhoods.

IV. Experimental Design and Performance Evaluation

A. Datasets Collection

The KTH database has been selected as the dataset for human motion behavior recognition [33]. This dataset comprises six distinct actions (walk, jump, run, fist, wave, and clap) performed by 25 individuals across four scenes, totaling 599 videos. It is important to note that the background remains relatively static during the recording of these human motion behaviors. Although each video may have varying durations and camera shooting angles, the background remains relatively static, facilitating a more focused recognition of human motion behavior.

In order to ensure uniformity in size and resolution, video processing tools are employed for segmenting each video into individual frames, followed by cropping and scaling. Grayscale images, which contain only brightness information, are preferred over color images for human motion behavior recognition tasks, as they enhance the visibility of the human body's form and motion features. Employing data augmentation techniques, such as random rotations, flips, translations, and other operations, enhances the diversity and generalization ability of data samples, generating additional training samples. Prior to conducting experiments, the entire dataset is partitioned into training and testing sets with a ratio of 4:1, ensuring consistency in sample distribution and features between the two sets.

Recognition outcomes based on geometric shapes or motion information from various human motion behavior databases are presented in Table 3. Notably, the KTH database demonstrates the highest recognition performance among the listed databases, achieving an impressive recognition rate of 95.77% based on geometric shape or motion information.

TABLE III. Recognition Results Based on Geometric Shape or Motion Information

Database	Accuracy of recognition
KTH [34]	95.77%
UCF [35]	86.5%
Hollywood 2 [36]	53.3%

B. The Setting of Experimental Environment and Parameters

The processor employed in this paper is the Intel (R) Core (TM) i5-7500 Central Processing Unit (CPU) @ 3.40GHz, while the operating system is Windows 10. The GTX1050, in conjunction with the Caffe framework, is utilized for GPU processing. The experimental dataset is divided into a training set and a test set with a ratio of 4:1, and each iteration encompasses 5000 generations. In order to optimize the network's recognition performance, certain network parameters undergo adjustment, including the size of the convolution kernel, the number of convolution layers, and the batch size. The optimization method involves maintaining the residual variable fixed and adjusting each individual variable until the optimal recognition rate is attained. Fig. 8 illustrates the specific parameters along with their corresponding value ranges.

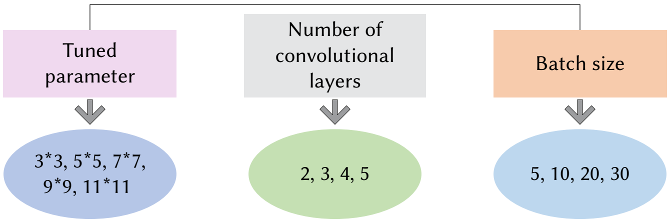


Fig. 8 Called parameters and their value ranges.

In Fig. 8, the CNN is configured to process single-channel grayscale images with a frame size of 80*80. The experiment centers on the KTH dataset and involves fine-tuning the CNN network architecture, specifically focusing on the convolution layers, kernel sizes, and other related parameters. Initially, the network configuration comprises three convolution and downsampling layers, two fully connected layers, and one Softmax layer responsible for generating identification results. The number of convolution kernels for the three convolution layers is set to 64, 128, and 128, respectively, while the two fully connected layers utilize 256 and 128 kernels. The initial learning rate is established at 0.005, and the training process is concluded after 20 Epochs.

C. Performance Evaluation

1. CNN Parameter Adjustment Results

When the batch size is defined as 10, the CNN network integrates three convolution layers and lower sampling layers. Fig. 9 illustrates the relationship between the CNN network parameters and the corresponding recognition accuracy.

In Fig. 9, the graphical representations illustrate the interdependencies between the convolution kernel size, the number of convolution layers, batch size, and their corresponding impact on recognition accuracy. Subfigure (A) elucidates the influence of the convolution kernel size on recognition accuracy, while Subfigure (B) delineates the effect of the number of convolution layers on recognition accuracy. Subfigure (C) provides insights into the relationship between batch size and recognition accuracy.

Upon meticulous examination of the findings, it is discerned that a convolution kernel size of 5*5 attains the highest recognition accuracy. Furthermore, when the number of convolution layers reaches 3, the network achieves its zenith recognition rate of 88.3%. Additionally, the network registers its peak recognition rate of 88.3% when the batch size is stipulated as 10. Consequently, the optimal configuration is ascertained to be a convolution kernel size of 5*5, three convolution layers, and a batch size of 10.

2. Analysis of Training Results Under the CNN Model

The training outcomes of the CNN model are visually represented in Fig. 10. Commencing at the initial iteration 0, the model's accuracy is documented at 0.0787. With successive iterations, there is a discernible enhancement in accuracy, coupled with a concurrent reduction in the value of the loss function. The loss function diminishes from its initial value of 1.7954 at iteration 0 to a minimal value of 0.0001 at iteration 5000. This decrease in the loss function signifies the progressive optimization of the model throughout the training process, resulting in a reduction of the disparity between predicted outcomes and actual labels. By the time the iteration count reaches 5000, the accuracy attains 92.59%. This data elucidates that the CNN model iteratively refines and assimilates knowledge during training, leading to a substantial improvement in classification accuracy on the test set.

Fig. 10 illustrates the categorization of human motion behavior into static and dynamic classifications. In order to evaluate the model's accuracy across distinct behavioral states, a comparative analysis is conducted, employing the proposed algorithm, CNN, SVM, and BPNN algorithm. The recognition accuracy of human motion behavior under these diverse models is depicted in Fig. 11.

In Fig. 11, the recognition accuracy of various models for static motion behavior generally exceeds that for dynamic motion behavior. Specifically, as depicted in Fig. 11(A), the CNN model attains the highest recognition accuracy for dynamic motion behavior, with an average accuracy of 93.61%. The proposed algorithm closely follows with an average recognition accuracy of 91.50%, while the SVM model achieves an average recognition accuracy of 83.83% for dynamic motion behavior recognition. The BPNN model records an average recognition accuracy of 90.44% for dynamic motion behavior recognition. In Fig. 11(B), concerning static motion behavior, the CNN model demonstrates the highest recognition accuracy, achieving an

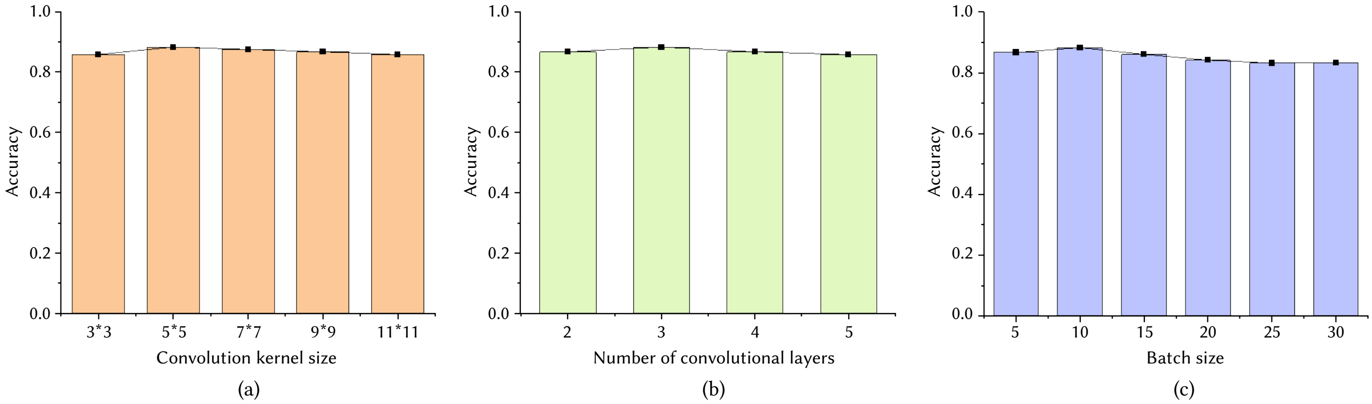


Fig. 9 Relationship between CNN network parameters and recognition accuracy.

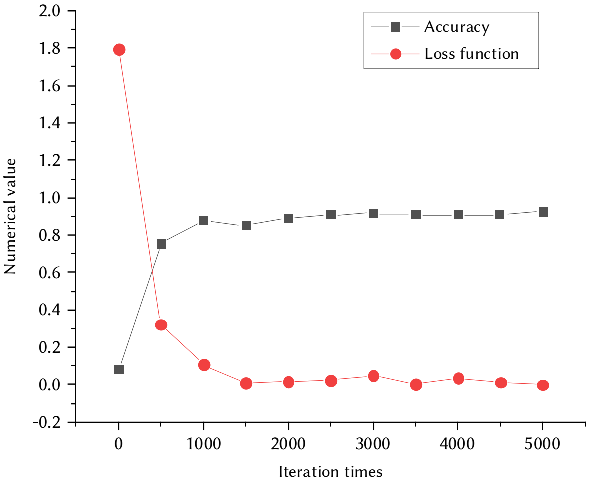


Fig. 10 Training results of the CNN model.

average accuracy of 94.41%. Following closely, the proposed algorithm achieves an average recognition accuracy of 92.76%. For dynamic motion behavior recognition, the SVM model attains an average recognition accuracy of 92.88%, and the BPNN model records an average recognition accuracy of 93.96%. Remarkably, the CNN model showcases the highest recognition accuracy overall for human motion behavior recognition.

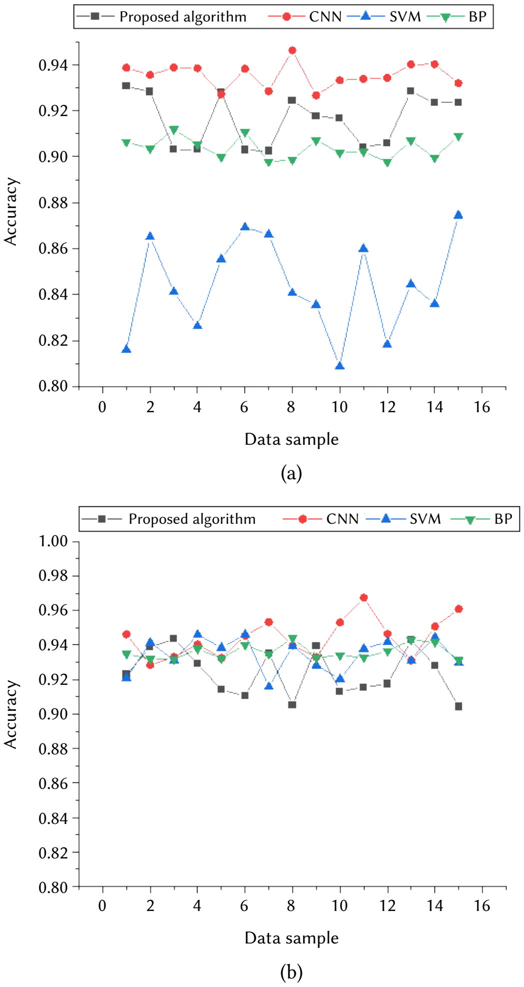


Fig. 11 Recognition accuracy of human motion behavior under different models (A: the dynamic motion behavior, B: the static motion behavior).

D. Discussion

The research findings presented in this paper underscore the substantial capability of CNN in recognizing human motion behavior. In direct comparison with alternative recognition algorithms, the CNN model emerges with the highest recognition accuracy for human motion behavior, boasting an average accuracy of 94.41%. For instance, Dong et al. delved into the application of 3D-CNN in human behavior recognition and achieved a recognition accuracy of 94.7%. In contrast, other recognition algorithms, such as the time-space domain depth CNN, recorded a comparatively lower recognition accuracy of 93.5%, underscoring the superior performance of the CNN model in this context [37]. In comparison to Dong et al.'s research outcomes, the 3D-CNN model employed in this paper showcases notably elevated recognition accuracy, reinforcing its superior performance among various recognition algorithms. The paper adopts a segmentation approach, dividing human motion behavior into static and dynamic categories and conducting separate network recognition, thereby yielding distinct recognition results for different motion behaviors. This segmentation strategy contributes to more accurate identification and differentiation of diverse types of motion behaviors, consequently enhancing recognition precision. Mahmoud conducted research on human behavior recognition utilizing LeNet-5CNN, revealing that with an increase in sample size, the recognition accuracy also improved, reaching a maximum accuracy of 78.54% [38]. In contrast to Mahmoud et al.'s research (2022), the CNN model employed in this paper demonstrates superior performance in human motion recognition. The 3D-CNN model adopted herein achieves a higher recognition accuracy in human motion behavior compared to the LeNet-5 CNN, indicating that the model utilized in this paper exhibits robust adaptability to complex motion recognition tasks.

V. Conclusion

This paper presents a methodology for extracting human motion behavior data scenes from the human living environment, leveraging the IoT framework. The primary objective is to investigate the recognition performance of human motion behavior using CNN. In order to achieve this, the KTH database is selected as the recognition dataset for human motion behavior. Rigorous parameter determination and analysis identify optimal settings for CNN recognition effectiveness, specifying a convolution core size of 5*5, three convolution layers, and a batch size of 10. The training loss function reaches a minimum value of 0.0001. Furthermore, the recognition accuracy of different models highlights CNN's superior performance in recognizing static motion behavior. While the paper introduces the concept of utilizing IoT to collect human motion behavior data characteristics, it acknowledges the challenge of processing and aggregating this data for network training due to its dispersed and complex nature. As a result, the KTH database is chosen as the training dataset instead of the collected data set. Experimental results demonstrate the recognition process and effectiveness of CNN in human motion behavior recognition using the KTH database as the training and testing dataset, yielding a commendable recognition accuracy. However, the limited number of sample data remains a consideration. In order to enhance the model's generalization ability and reliability, future research could explore the collection of more diverse and abundant human motion behavior data, validating it with other publicly available large-scale datasets. Additionally, despite utilizing an optimized CNN model to enhance human motion behavior recognition robustness, challenges may persist in complex scenarios, such as lighting variations, occlusions, and background interference, potentially affecting recognition accuracy. Future studies may delve into exploring more sophisticated network structures and data augmentation techniques to further improve the model's robustness in challenging scenarios.

References

[1] S. Wan, L. Qi, X. Xu, C. Tong, and Z. Gu, "Deep learning models for real-time human activity recognition with smartphones," Mobile Networks and Applications, vol. 25, no. 2, pp. 743-755, 2020.

[2] H. Yan, Y. Zhang, Y. Wang, and K. Xu, "WiAct: A passive WiFi-based human activity recognition system," IEEE Sensors Journal, vol. 20, no. 1, pp. 296-305,2020.

[3] A. Rudenk, L. Palmieri, M. Herman, K. M. Kitani, D. M. Gavrila, & K. O. Arras, "Human motion trajectory prediction: A survey," The International Journal of Robotics Research, vol. 8, no. 39, pp. 895-935, 2020.

[4] A. Jalal, N. Khalid, & K. Kim, "Automatic recognition of human interaction via hybrid descriptors and maximum entropy markov model using depth sensors," Entropy, vol.22, no. 8, pp. 817, 2020.

[5] B. Sun, G. Cheng, Q. Dai, T. Chen, W. Liu, & X. Xu, "Human motion intention recognition based on EMG signal and angle signal," Cognitive Computation and Systems, vol. 1, no. 3, pp. 37-47, 2021.

[6] V. Bianchi, M. Bassoli, G. Lombardo, P. Fornacciari, M. Mordonini, & I. De Munari, "IoT wearable sensor and deep learning: An integrated approach for personalized human activity recognition in a smart home environment," IEEE Internet of Things Journal, vol. 5, no. 6, pp. 8553-8562, 2019.

[7] W. Sousa Lima, E. Souto, K. El-Khatib, R. Jalali, & J. Gama, "Human activity recognition using inertial sensors in a smartphone: An overview," Sensors, vol. 14, no. 19, pp. 3213, 2019.

[8] Y. Guo, Z. Xuan, & L. Song, "Foreground target extraction method based on neighbourhood pixel intensity correction," Australian Journal of Mechanical Engineering, vol. 3, no. 19, pp. 251-260, 2021.

[9] C. Wei, G. Xue-Bao, T. Feng, S. Ying, W. Wei-Hong, S. Hong-Ri, & K. Xuan, "Seismic velocity inversion based on CNN-LSTM fusion deep neural network," Applied Geophysics, vol. 4, no. 18, pp. 499-514, 2021.

[10] L. A. Bolaños, D. Xiao, N. L. Ford, J. M. LeDue, P. K. Gupta, C. Doebeli, & T. H. Murphy, "A three-dimensional virtual mouse generates synthetic training data for behavioral analysis," Nature methods, vol. 4, no. 18, pp. 378-381, 2021.

[11] S. Chebbout, & H. F. Merouani, "A hybrid codebook model for object categorization using two-way clustering based codebook generation method," *International Journal of Computers and Applications*, vol. 2, no. 44, pp. 178-186, 2022.

[12] Y. Peng, H. Tao, W. Li, H. Yuan, & T. Li, "Dynamic gesture recognition based on feature fusion network and variant ConvLSTM," IET Image Processing, vol. 11, no. 14, pp. 2480-2486, 2020.

[13] H. Chen, C. Hu, F. Lee, C. Lin, W. Yao, L. Chen, & Q. Chen, "A supervised video hashing method based on a deep 3d convolutional neural network for large-scale video retrieval," Sensors, vol. 9, no. 21, pp. 3094, 2021.

[14] G. Lin, Y. Zhang, G. Xu, & Q. Zhang, "Smoke detection on video sequences using 3D convolutional neural networks," Fire Technology, vol. 5, no. 55, pp. 1827-1847, 2019.

[15] R. Zhi, H. Xu, M. Wan, & T. Li, "Combining 3D convolutional neural networks with transfer learning by supervised pre-training for facial micro-expression recognition," IEICE Transactions on Information and Systems, vol. 5, no. 102, pp. 1054-1064, 2019.

[16] H. Jin, J. Geng, Y. Yin, M. Hu, G. Yang, S. Xiang & H. Zhang, "Fully automated intracranial aneurysm detection and segmentation from digital subtraction angiography series using an end-to-end spatiotemporal deep neural network," Journal of NeuroInterventional Surgery, vol. 10, no. 12, pp. 1023-1027, 2020.

[17] X. Lu, H. Yao, S. Zhao, X. Sun, & S. Zhang, "Action recognition with multi-scale trajectory-pooled 3D convolutional descriptors," Multimedia Tools and Applications, vol. 1, no. 78, pp. 507-523, 2019.

[18] P. Guenzi, & E. J. Nijssen, "The impact of digital transformation on salespeople: an empirical investigation using the JD-R model," Journal of Personal selling & sales ManageMent, vol. 2, no. 41, pp. 130-149, 2021.

[19] A. Jalal, M. A. K. Quaid, & K. Kim, "A wrist worn acceleration based human motion analysis and classification for ambient smart home system," Journal of Electrical Engineering & Technology, vol. 4, no. 14, pp. 1733-1739, 2019.

[20] M. A. K. Quaid, & A. Jalal, "Wearable sensors based human behavioral pattern recognition using statistical features and reweighted genetic algorithm," Multimedia Tools and Applications, vol. 9, no. 79, pp. 6061-6083, 2020.

[21] L. Huang, X. Zhao, & K. Huang, "Got-10k: A large high-diversity benchmark for generic object tracking in the wild," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 5. no. 43, pp. 1562-1577, 2019.

[22] M. A. Khan, Y. D. Zhang, S. A. Khan, M. Attique, A. Rehman, & S. Seo, "A resource conscious human action recognition framework using 26-layered deep convolutional neural network," Multimedia Tools and Applications, vol. 28, no. 80, pp. 35827-35849, 2021.

[23] W. Qi, H. Su, C. Yang, G. Ferrigno, E. De Momi, & A. Aliverti, "A fast and robust deep convolutional neural networks for complex human activity recognition using smartphone," Sensors, vol. 17, no. 19, pp. 3731, 2019.

[24] Q. Teng, K. Wang, L. Zhang, & J. He, "The layer-wise training convolutional neural networks using local loss for sensor-based human activity recognition," IEEE Sensors Journal, vol. 13, no. 20, pp. 7265-7274, 2020.

[25] K. Wang, J. He, & L. Zhang, "Attention-based convolutional neural network for weakly labeled human activities' recognition with wearable sensors," *IEEE Sensors Journal*, vol. 17, no. 19, pp. 7598-7604, 2019.

[26] N. K. Benamara, E. Zigh, T. B. Stambouli, and M. Keche, "Towards a robust thermal-visible heterogeneous face recognition approach based on a cycle generative adversarial network," *INTERNATIONAL JOURNAL OF INTERACTIVE MULTIMEDIA AND ARTIFICIAL INTELLIGENCE,* vol. 7, no. 4, pp. 132-145, 2022.

[27] D. Dejene, B. Tiwari, and V. Tiwari, "TD2SecIoT: temporal, data-driven and dynamic network layer based security architecture for industrial IoT," *INTERNATIONAL JOURNAL OF INTERACTIVE MULTIMEDIA AND ARTIFICIAL INTELLIGENCE,* vol. 6, no. 4, pp. 146-156, 2020.

[28] A. J. F. García, J. C. P. Rodríguez, A. P. Ramos, F. S. Figueroa, and J. D. Gutiérrez, "CompareML: a novel approach to supporting preliminary data analysis decision making," *INTERNATIONAL JOURNAL OF INTERACTIVE MULTIMEDIA AND ARTIFICIAL INTELLIGENCE,* vol. 7, no. 4, pp. 225-238, 2022.

[29] S. M. Lezcano, F. López, and A. C. Bellot, "Data science techniques for COVID-19 in intensive care units," *INTERNATIONAL JOURNAL OF INTERACTIVE MULTIMEDIA AND ARTIFICIAL INTELLIGENCE,* vol. 6, no. 4, pp. 8-17, 2020.

[30] M. L. Ibáñez, A. R. Hernández, B. Manero, and M. G. Mata-García, "Computer entertainment technologies for the visually impaired: an overview," *INTERNATIONAL JOURNAL OF INTERACTIVE MULTIMEDIA AND ARTIFICIAL INTELLIGENCE,* vol. 7, no. 4, pp. 53-68, 2022.

[31] K. R. Gsangaya, S. S. H. Hajjaj, M. T. H. Sultan, & L. S. Hua, "Portable, wireless, and effective internet of things-based sensors for precision agriculture," International Journal of Environmental Science and Technology, vol. 9, no. 17, pp. 3901-3916, 2022.

[32] X. Gao, P. Pishdad-Bozorgi, D. R. Shelden, & S. Tang, "Internet of things enabled data acquisition framework for smart building applications," Journal of Construction Engineering and Management, vol. 2, no. 147, pp. 04020169, 2021.

[33] A. Abdelbaky, & S, Aly. "Human action recognition using short-time motion energy template images and PCANet features," *Neural Computing and Applications*, vol. 16, no. 32, pp. 12561-12574, 2020.

[34] J. Guo, P. V. Borges, C. Park, & A. Gawel, "Local descriptor for robust place recognition using lidar intensity," IEEE Robotics and Automation Letters, vol. 2, no. 4, pp. 1470-1477, 2019.

[35] Y. Liu, X. Yao, Z. Gu, Z. Zhou, X. Liu, X. Chen, & S. Wei."Study of the Automatic Recognition of Landslides by Using InSAR Images and the Improved Mask R-CNN Model in the Eastern Tibet Plateau," Remote Sensing, vol. 14, no. 14, pp. 3362, 2022.

[36] D. Jeong, B. G. Kim, & S. Y. Dong, "Deep joint spatiotemporal network (DJSTN) for efficient facial expression recognition," Sensors, vol. 7, no. 20, pp. 1936, 2020.

[37] M. Dong, Z. Fang, Y. Li, S. Bi, & J. Chen, "AR3D: attention residual 3D network for human action recognition," *Sensors*, vol. 5, no. 21, pp. 1656, 2021.

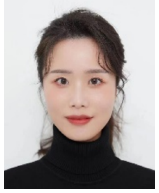
[38] H. Mahmoud, "Modern architectures convolutional neural networks in human activity recognition," Advances in Computing and Engineering, vol. 1, no. 2, pp. 1-16, 2022.

Hui Li

Hui Li was born in Luoyang, Henan, P.R. China, in 1988. He received his Ph.D. in Physical Education from Kunsan National University in Korea. Now he works in the Department of Physical Education of Luoyang Institute of Science and Technology. His research interest includes on motor learning and control, action-perception coupling, and human motor skill learning. E-mail: 200901701093@lit.edu.cn

Huayang Liu

Huayang Liu was born in Weihai, Shandong Province, People's Republic of China.He is currently a Ph.D. student at Gunsan National University of Korea, and will graduate in July 2024, and has now joined the College of Physical Education of China West Normal University in the field of sports and humanities sociology. E-mail: 690942317@qq.com



Wei Zhao

Wei Zhao was born in Xinxiang, Henan, P.R. China, in 1994. She received a "Master of Teaching Chinese to Speakers of Other Languages from Henan Normal University. She is currently working in the College of Humanities and Social Sciences of Luoyang Institute of Science and Technology. Her research interests are linguistics, psychology, and education. E-mail: 541741134@qq.com

Hao Liu

Hao Liu was born in 1995 in Linyi, Shandong Province, China. He received his doctorate from Gunsan National University . He is currently working in the School of Physical Education and Health of Guangxi Medical University, mainly studying sports biomechanics and sports statistics.Email: linyiliuhao@sina.com