Human Action Recognition From Digital Videos Based on Deep Learning

CHENWEI LIANG, Auckland University of Technology, Auckland 1010, New Zealand
JIA LU, Auckland University of Technology, Auckland 1010, New Zealand
WEI QI YAN, Auckland University of Technology, Auckland 1010, New Zealand

With the development of closed-circuit television, video-based human motion recognition has made great progress. A large number of surveillance video footages have been archived. In this paper, we implement deep learning methods to resolve human action recognition problem. We propose a new method that combines Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) together, which is able to produce a better result after expansive and extensive experiments. The experimental results of this paper show that it is feasible to implement human action recognition through deep learning algorithms, the outcome is excellent. The CNN+LSTM method proposed in this paper can better recognize human actions, which is more efficient than general deep learning methods. In addition, in this paper, we compare the differences in the recognition results using deep learning methods.

CCS Concepts: • **Computing methodologies** → *Activity recognition and understanding*.

Additional Key Words and Phrases: Human action recognition , Deep learning methods , CNN , LSTM.

ACM Reference Format:
Chenwei Liang, Jia Lu, and Wei Qi Yan. 2022. Human Action Recognition From Digital Videos Based on Deep Learning. 1, 1 (May 2022), 11 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Influenced by rapid development of monitoring and surveillance technology, a spate of digital monitoring devices have appeared in a rich assortment of scenes. The popularity and installation of these surveillance devices have generated a large number of surveillance videos, which often contain multiple classes of human actions and behaviors. The surveillance videos encapsulate tracking and detecting pedestrians or vehicle trajectories, recognizing and identifying visual objects as well as human behaviors [11] [26]. Analyzing these actions can benefit people distinguishing normal behaviors and abnormal behaviors from these videos, which is of great value in improving public safety.

In the conventional usages of surveillance videos to identify human behaviors, most of effective information can only be acquired by relying on a large number of manual operations. This brings unavoidable problems, such as labor costs and energy saving issues. With the increase of working hours, human energy will be also dropped significantly. It often needs to take a period of time for rest and recovery. Thus, it is particularly important to take use of automated methods in computer vision to replace human manual operations for video analysis [27]. In artificial intelligence, deep

Authors' addresses: Chenwei Liang, Auckland University of Technology, Auckland 1010, New Zealand; Jia Lu, Auckland University of Technology, Auckland 1010, New Zealand; Wei Qi Yan, Auckland University of Technology, Auckland 1010, New Zealand.

XXXX-XXXX/2022/5-ART $15.00
https://doi.org/XXXXXXX.XXXXXXX

learning methods have already been employed for saving human labor. The emergence of these methods has greatly uplifted the efficiency of intelligent surveillance [31].

As a field of emergence, human action recognition generally includes human actions such as falling, walking, running, and raising hands. We will implement human action recognition for the behaviors such as walking, jogging, boxing, waving, etc. The combined use of computer vision and deep learning has become increasingly popular in recent years [18]. As an important research direction in the field of computer vision, it is also necessary to apply deep learning to human motion recognition. In this paper, we will combine spatiotemporal information together and make use of our algorithm to recognize human behaviors. The focus is on recognizing human behaviors by using spatiotemporal information. It is worth noting that our algorithm is implemented by using MATLAB, which provides a new implementation method. This algorithm takes use of a combination of LSTM and CNN to produce the experimental result that matches our expectations. In addition, in order to have a much intuitive understanding of the results, we conduct a side-by-side comparison between this method and other methods.

CNN is a deep learning algorithm that is unique from general neural networks. Its characteristics are very obvious, generally consisting of input layer, convolutional layer, pooling layer, and fully connected layer [17]. LSTM is a type of recurrent neural networks (RNN), which not only solves the gradient vanishing problems, but also is suitable for classification and prediction in time series analysis [21]. By combining these two neural networks, we get a very new method for human motion detection.

In the next section of this paper, we will briefly review the existing methods related to human action recognition. In the third part, we depict our method of this paper in detail. In the fourth section, the experimental results are demonstrated. Our conclusions are drawn in Section V.

2 LITERATURE REVIEW

Human motion recognition, as a research subject that attracted much attention, has emerged in the past few decades, a large amount of related research work on human motion recognition has been published. A great deal of new algorithms or frameworks have been proposed, explored and exploited. The emergence of the literature has played a pivotal role in the development of human motion recognition. Among the research methods, there are not only conventional machine learning methods for extracting visual features and predicting motion, but also popular deep learning methods proposed in recent years. In the rest of this section, we will give a brief overview of the related work.

The related work like using Transformers to recognition [1] , 3D human action recognition with LSTM[30] , skeleton Fourier for motion trajectory recognition[13], Human gait recognition using frame-by-frame gait energy images [23] and human gait recognition using multi-channel convolutional neural networks [24] [10], etc. These works have enriched the methods of human action recognition and have played a central role in future research.

Trajectory analysis based on neural network is one of the main methods for human action recognition. A method was proposed for human action recognition based on trajectory analysis using neural networks. The method was tested by using CAVIAR database, the activity description vector (ADV) was employed to describe basic description of the scene understanding whilst describing the scene. The final results of the experiments show that the use of ADV enables the detection system to clearly distinguish and identify human behaviors and scenes in complex environments [5].

A method was proffered for predicting human motion trajectory to recognize actions and behaviors for human action recognition in video sequences [6]. In the proposed method, the prediction and recognition of human behaviors can finally be implemented from the input video. The reason is that they took use of full human motion trajectory while training the model. This

assists the model to evaluate human actions with specific scenarios. In order to improve the accuracy of the experiments, a slew of experiments were conducted with multiple classifiers by using three datasets. The experimental results show that the proposed method is able to accurately improve the accuracy of early identification of human behaviors.

As the most popular method in early human motion recognition, prediction has been widely utilized. On the basis of previous research outcomes, Almeida and Azkune added deep learning framework of LSTM as a new model for predicting human actions [3]. The neural network models the behaviors of human interacting with the environment. Finally, a probabilistic model is put forward, which enables the algorithm not only to predict human actions, but also further to identify abnormal human actions. This algorithm was finally applied to a project called City4Age H2020 which aims at improving the quality of lives of urban residents.

Deep learning has been developed rapidly in the field of computer vision in recent years. It is exceptional in the field of human action recognition. In order to automatically identify human behaviors from our screens, Ji et al al. [12] created a 3D-CNN model by using deep learning methods. Unlike prevailing methods at that time, which required constructing a classifier, CNN is a convenient and easy-to-use model. But CNNs could only handle 2D inputs. In order to make this model more broadly applied, a 3D-CNN network was developed for human action recognition. The basic principle of this model is to extract the features of spatial and temporal relationships from the input video through a special 3D convolution, and obtain the motion information in multiple adjacent frames, and finally attain the result. After numerous experiments, 3D-CNN model is finally employed for human behavior recognition in the environment compared with other methods, the outcome was excellent.

In view of increasing popularity of deep learning for human action recognition in recent years, Wu et al. made a brief review of popular deep learning methods for human action recognition as well as labelled datasets. Single-view datasets and GRU-RNN methods have been employed to deal with continuous convolutions. The datasets have an obvious attribute that the background is often static, and the video is usually from a single viewpoint of a digital camera. This means that this type of datasets have occlusion problems. Other methods take use of multiview datasets or RGB datasets. The best way is to take use of fuzzy CNN so as to achieve best result based on motion information of videos [25].

With the further improvement of computing speed and processing power, the prospect of deep neural networks (DNN) becomes more and more light. In order to distinguish it from the traditional human action recognition algorithm that extracts global features of the input image so as to recognize human actions. Lu, et al [16] completed the research work based on human action recognition by using YOLOv3 model in 2018. The final accuracy of the work was up to 80.20%. In order to achieve this goal, a few of public datasets are adopted, special datasets were created for experiments. In addition, the network learning rate with the highest accuracy was attained by continuously adjusting the net structure. In order to make the experimental results more convincing, YOLOv2 was implemented for the purpose of comparisons. In the model, the accuracy is nearly as good as that of YOLOv3. This makes the method much efficient and accurate.

Dataset training is equally important in deep learning. In terms of training method, Wan et al [22] proposed a real-time human activity classification method based on convolutional neural network (CNN). This method includes CNN module for feature extraction. To demonstrate the superiority of the method, they also adopted various classification models such as CNN, LSTM, etc. on the UCI and Pamap2 datasets. The final result proves that the CNN classification method outperforms other methods.In 2017, Lu et al [15] also investigated the classification of different devilishness. They took use of three feature extraction methods: Histogram of Oriented Gradients (HOG), Local Binary Patterns (LBP), and Scale-Invariant Local Ternary Patterns (SILTP). The INRIA

and Weizmann datasets are taken into acount to compare the experimental results. It turns out that both LBP and HOG perform well on the INRIA dataset. However, based on the Weizmann dataset, only HOG feature performs better for classification.

In 2020, Lu, et al. further optimized deep nets for human action recognition. They took advantage of a more advanced approach than YOLOv3 to solve the human action recognition problem [14]. This new method has adopted the network architecture YOLOv4 + LSTM to achieve the results. The network is very successful, the accuracy has reached to 97.87% after extensive experiments. The reason is that they added event information and spatial information to the identification method. Besides, they also validated a new Selective Kernel Network (SKNet) model with attention mechanism in the experiments. The use of this model yielded even better results, achieved 98.70% accuracy rate across multiple datasets.

Yu, et al. proposed specific models and methods for human action recognition in 2020. Human action recognition is thought as the process of making the machine to understand the actions of people in digital videos and label the human actions accordingly. In addition, because human behavior is affected by external factors such as lighting conditions, the background is greatly affected. To this end, three different action recognition algorithms were designed and implemented. These methods are all based on convolutional neural networks (CNN), namely 3D-CNN, Two-Stream CNN, and CNN+LSTM. These models are merged together after extensive testing based on HMDB-51 dataset. After compared the final experimental results, it is found that the CNN+LSTM method is able to effectively avoid interference factors [28].

In this paper, our method is different from the existing work. Our contribution is to implement a fast deep learning method for human action recognition. In addition, we also compare the accuracy between CNN+LSTM and other models. Therefore, the model has a positive impact on human action recognition.

3 OUR METHODS

In this paper, we elaborate a new deep learning model. Compared with those existing deep learning methods, this proposed method is much efficient. In order to make this method excellent, we chose LSTM to extract temporal features. The reason is that in conventional machine learning methods, the efficiency of most deep learning methods often depends on the feature extraction. In addition, due to the complicated information existing in videos, including a large amount of temporal and spatial data, the use of LSTM facilitates our model to extract temporal information from each sequence of video frames. CNN algorithm and LSTM algorithm are employed in the whole framework.

A complete CNN model was created with a variable number of convolutional layers, pooling layers, and fully connected layers. Among them, the convolutional layer is the core of CNN [2]. The performance of the entire network can be improved by tuning the parameters in the convolutional layers [2]. The basic principle of CNN is to reduce the resources required for operations by reducing the dimensionality of the image, while retaining the basic features of the image. After convolution operations, the image does not change substantially. Among them, the convolutional layer is mainly responsible for extracting visual features of the image, and the pooling layer is adopted after the convolutional layer to reduce the dimension of the entire network and prevent the network against overfitting. It is precise because of these advantages of the CNN networks [17].

LSTM, as a recurrent neural network (RNN), is applied in this paper to select temporal features, which is different from general RNNs. Compared with ordinary neural networks, RNNs are able to process data after a sequence of changes. However, if the length of the sequential data for network training is too long, it will cause problems such as vanishing and exploding gradient problems [21]. LSTM consists of memory units and various internal gates, such as input gates and forget gates.

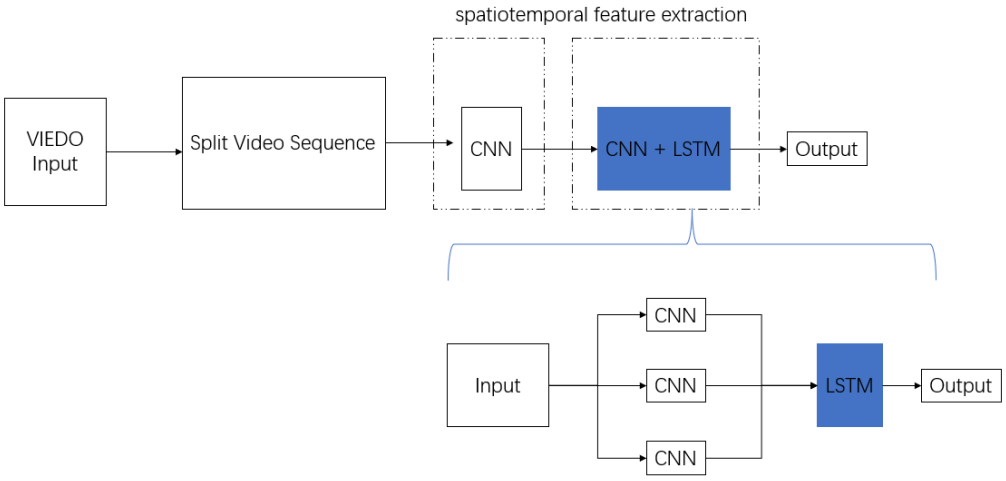

Fig. 1. Action recognition model using deep learning including CNN+LSTM model

Adjusting LSTM through these gates can solve the vanishing gradient problem well and store the content that needs to be memorised for a long time [20]. In this paper, because the duration of video is far beyond the processing length of ordinary RNN, LSTM with better performance is selected as the recurrent network to store the content that needs to be deposited for a long time.

Figure 1 shows the structure of this model. After the model segments the input video clips, it separates the video frame by frame, then harnesses the CNN network to generate the feature data set of each action and utilizes CNN+LSTM to collect spatiotemporal information in the labelled dataset. We take use of these data to train the deep net, and finally export it as a recognizable action for the final human action recognition.

4 EXPERIMENTAL RESULTS

In this paper, we choose the open public dataset KHT to our tests. The KHT dataset contains six classes of human actions, each class includes 100 videos which are grouped into 25 topics and 4 scenarios [19]. The videos were taken with a homogeneous background using a still camera with a frame rate of 25 fps, the resolution of these videos was 160×120, the average length was 4 seconds. There are 25 classes in this dataset. Specific human actions include walking, jogging, running, boxing, waving, and clapping. In our experiments, we selected all six action categories for this paper.

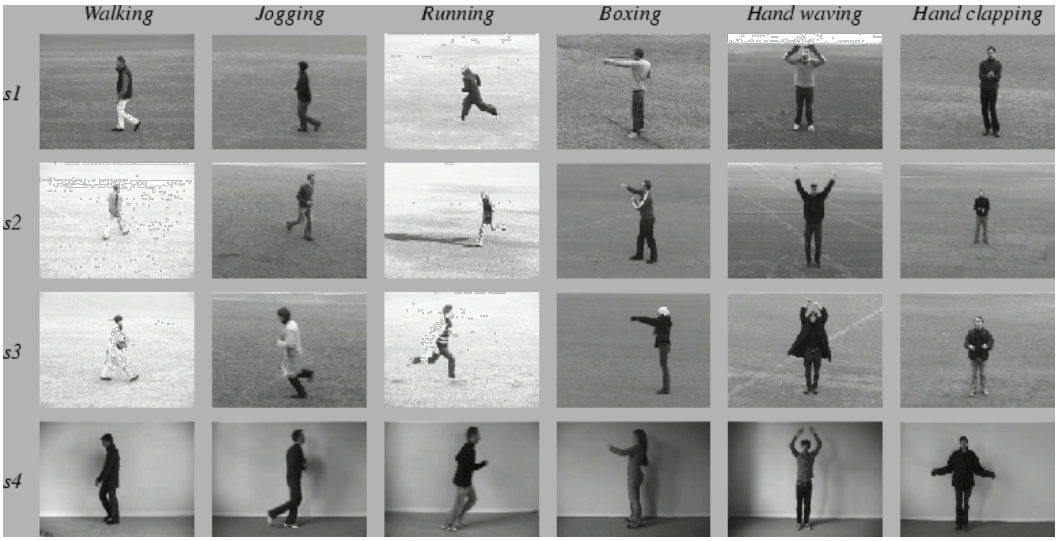

Fig. 2. Samples of KHT dataset

The focus of our work is on human action recognition using deep learning methods. We compare the results of human action recognition by using LSTM+CNN model, CNN, KNN, and STIP+KNN. Finally, four different accuracies were obtained. In addition, in this paper, we also verify whether the efficiency of recognition will be affected by the different implementation methods. In Figure 3, we show the results for two videos in the selected dataset.

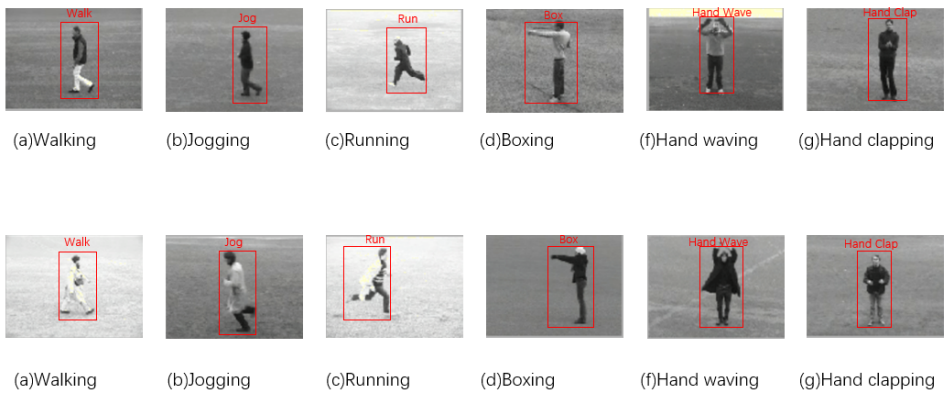

Fig. 3. Two human action recognition results show that the recognition of six actions in two sets of images

In this paper, we compare four different methods, namely LSTM+CNN, CNN, KNN, STIP+KNN, which have various accuracy rates. In the experiments of this paper, as the number of iterations increases, the experimental results gradually converge, the models finally stop at a stable level. All three models gradually started converging after the 200-th iteration.

Table 1. Multiple methods for human action recognition

S.No	Method	Features	Accuracy
1	Ismail et al.	CNN on block of frames	70.37%
		Optical Flow + Bag-of-Words + SVMAccuracy	78.24%
		CNN on block of frames+ optical flow	90.27%
2	Bilen et al. [8]	2D-CNN	86.80%
3	Basha et al. [7]	3D-CNN	95.27%
4	Our Method	CNN+LSTM	88.30%

In the final results, the accuracy of LSTM+CNN model in this paper is 88.00% and 89.00%, respectively. Among other models, KNN model is up to 83.00%, and the CNN model reaches 86.00% and 87.00%, STIP+KNN model is at 84.00% and 85.00%, respectively. Compared to the method implemented using the same method, we see that the LSTM+CNN model, CNN model, and STIP + KNN model are very less accurate. The exact change of the KNN model is not very large. This shows that though the accuracy of action recognition using deep learning models has declined due to different environments, the decrease is not significant. It does not have a big impact on the recognition results. In addition, we also see that the accuracy of the LSTM+CNN method is different in the two environments, but still very good among the four models.

Moreover, we see that if a neural network is employed for recognition, such as CNN and KNN, the recognition accuracy is often not high. If we add a network that captures spatial information, such as LSTM and STIP, to the network, the accuracy is greatly improved. This further illustrates that it is very correct and important for us to choose an appropriate structure or network to capture temporal information when building a model. If there is no part of the model that extracts temporal information, our recognition accuracy will drop significantly.

In Table 1, we show the accuracy of human action recognition based on the KTH dataset. From Table 1, we see that compared with the result of Ismail et al. based on the KHT dataset in 2017, the accuracy of the proposed model in this paper is higher. The methods based on CNN and Optical Flow + Bag-of-Words + SVM achieve 78.24% and 70.37% accuracy, respectively, which are lower than our model. But the accuracy of CNN on block of frames+ optical flow method is higher than our model.

Compared to the work of Bilen et al., the accuracy of our model is 2% higher than that of 2D-CNN. The model proposed by Basha et al. is slightly more accurate than ours, which has achieved 95.27% accuracy using 3D-CNN. The results show that the performance of the model in this paper is good, though there is still a gap with some methods, it is still at top stream level in general. This further demonstrates the performance of our model.

Figure 4 shows the accuracy rates of the six actions in the dataset with multiple models. We see from the graph that the recognition rate for each action using the LSTM+CNN model is above 0.85. The performance of other three models is relatively poor. The action accuracy of the CNN model is between 0.84 and 0.89. The accuracy of the KNN model is between 0.80 and 0.86. The accuracy of STIP+KNN model is between 0.82 and 0.87.

Table 2. Single action recognition results

	Walking	Jogging	Running	Boxing	Waving	Clapping
LSTM+CNN	0.91	0.85	0.92	0.90	0.88	0.87
CNN	0.89	0.84	0.89	0.86	0.87	0.85
KNN	0.86	0.80	0.85	0.84	0.84	0.82
STIP+KNN	0.87	0.82	0.85	0.85	0.84	0.84

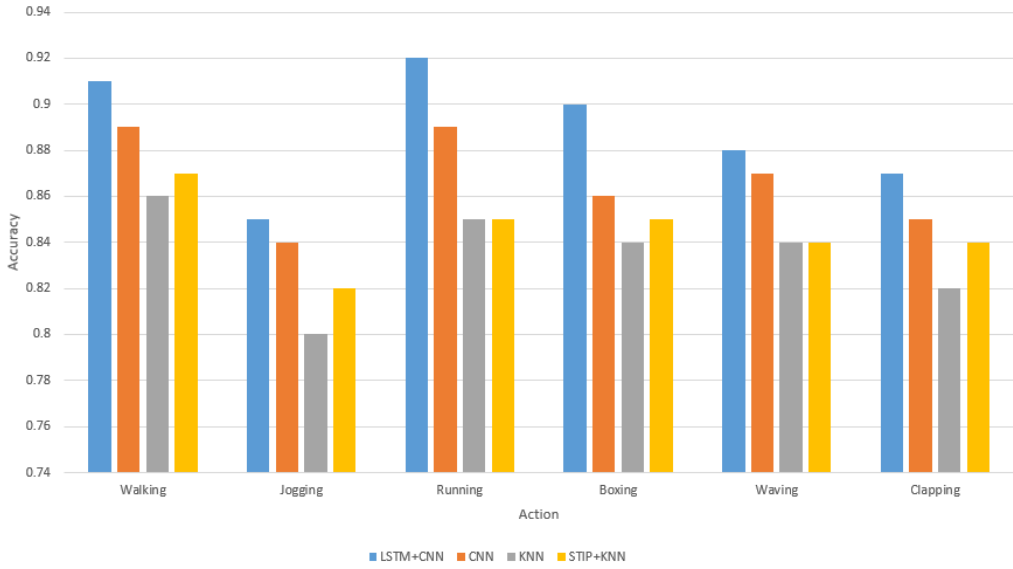

Fig. 4. Single action recognition results

Among them, the performance of KNN model is much lower than our proposed method. STIP+CNN is relatively better than CNN. The accuracy of LSTM+CNN model is still the highest one among the four models. Among all the recognized actions, walking and running actions have higher accuracy. The accuracy of other movements is slightly lower, such as the action jogging, which has the lowest accuracy of all movements. The highest one is only 0.85. The accuracy of each action in different models is shown in Table 2.

Figures 5 shows the wrong recognition results. We see that the wrongly classified action is Jogging. After our analysis, we believe that the reason for the low recognition accuracy is that the action is similar to walking and running in motions, which leads to the problem that it is easy to generate errors in human action recognition. In response to this problem, we made our adjustments to the algorithm and increased the number of layers of the neural network during feature extraction.

Finally, in order to see if the recognition accuracy varies with the length of the video sequences in the dataset, we are use of MSR action dataset. The dataset includes 16 video sequences, and the actions in the videos are generally classified into three classes, namely clapping, waving, and boxing. Each video sequence contains multiple actions. The resolution of the video is not high. Video length is between 32 and 76 seconds [29]. After our experiments, the final test results are shown in Table 3. We see that while using LSTM+CNN to identify these three actions, when the

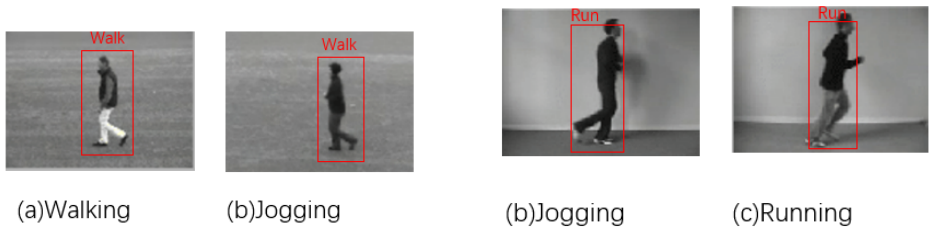

Fig. 5. Wrong recognition results

Table 3. MSR Action Dataset Result.

Single Action	Waving	Clapping	Boxing
Accuracy	0.8947	0.8774	0.8983
Multiple Actions	Waving+Clapping+Boxing		
Accuracy	0.7312		

actions are identified individually, the accuracy rate does not change much, and the overall stability is around 85.00%. But if these actions occur consecutively, the recognition accuracy drops drastically. This proves that our chosen method is less suitable for some more complex environments.

Overall, the four models are able to accurately identify most human actions, with the LSTM+CNN model performing the best. CNN and STIP+KNN are the next, and KNN is the worst. However, by properly adjusting the parameters in the network, such as the network depth and the number of iterations, higher accuracy can be obtained and the occurrence of action recognition errors can be reduced. Furthermore, human action recognition by using deep learning methods was successful. However, the algorithm still has the problems that it cannot be adaptive to the complicated environment, the accuracy of a certain action is not high.

5 CONCLUSION AND FUTURE WORK

Deep learning methods are most suitable for the field of computer vision. In this paper, we conduct experiments for human action recognition by using deep learning methods. In our experiments, we clearly see that deep learning models are well implemented. Among them, the method of using CNN+LSTM network to obtain spatiotemporal information is helpful for human action recognition which shows excellent results. From the experimental results, we see that based on different deep learning methods, the accuracy will also vary. However, this loss will not have a significant impact on the results.

In the near future, we will implement increasingly complex deep learning models, such as spatial temporal interest point (STIP). STIP can effectively extract interest points in videos for human action recognition. It is a mature method for human action recognition. We will take use of deconvolution or dilated convolution in the recognition network [9]. We will combine spatial and temporal interest points with LSTM as part of our recognition network to extract temporal features. In the recognition of a single action, jogging has the lowest recognition accuracy, and the recognition of this behavior needs to be improved in future. In the continuous recognition of multiple actions in complex environments, the accuracy of recognition is not high, which needs to be improved in future work. Based on these identified behaviors, we further predict what the next

behavior will be when one behavior occurs. In addition, we will improve the accuracy of the model for multiple and continuous action recognition in complex environments [4, 26, 27].

REFERENCES

[1] Tasweer Ahmad, Junaid Rafique, Hassam Muazzam, and Tahir Rizvi. 2015. Using discrete cosine transform based features for human action recognition. *Journal of Image and Graphics* 3, 2 (2015), 96–101.

[2] Saad Albawi, Tareq Abed Mohammed, and Saad Al-Zawi. 2017. Understanding of a convolutional neural network. In *IEEE International Conference on Engineering and Technology (ICET)*. 1–6.

[3] Aitor Almeida and Gorka Azkune. 2018. Predicting human behaviour with recurrent neural networks. *Applied Sciences* 8, 2 (2018), 305.

[4] Na An and Wei Qi Yan. 2021. Multitarget tracking using Siamese neural networks. *ACM Transactions on Multimedia Computing, Communications and Applications* 17, 2s (2021), 1–16.

[5] Jorge Azorín-López, Marcelo Saval-Calvo, Andrés Fuster-Guilló, and José García-Rodríguez. 2013. Human behaviour recognition based on trajectory analysis using neural networks. In *IEEE International Joint Conference on Neural Networks (IJCNN)*. 1–7.

[6] Jorge Azorin-Lopez, Marcelo Saval-Calvo, Andres Fuster-Guillo, and Jose Garcia-Rodriguez. 2016. A novel prediction method for early recognition of global human behaviour in image sequences. *Neural Processing Letters* 43, 2 (2016), 363–387.

[7] SH Basha, Viswanath Pulabaigari, and Snehasis Mukherjee. 2020. An information-rich sampling technique over spatio-temporal CNN for classification of human actions in videos. *arXiv preprint arXiv:2002.02100* (2020).

[8] Hakan Bilen, Basura Fernando, Efstratios Gavves, Andrea Vedaldi, and Stephen Gould. 2016. Dynamic image networks for action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*. 3034–3042.

[9] Debapratim Das Dawn and Soharab Hossain Shaikh. 2016. A comprehensive survey of human action recognition with spatio-temporal interest point (STIP) detector. *The Visual Computer* 32, 3 (2016), 289–306.

[10] Muhammad Hassan, Tasweer Ahmad, Nudrat Liaqat, Ali Farooq, Syed Asghar Ali, and Syed Rizwan Hassan. 2014. A review on human actions recognition using vision based techniques. *Journal of Image and Graphics* 2, 1 (2014), 28–32.

[11] Sutrisno Warsono Ibrahim. 2016. A comprehensive review on intelligent surveillance systems. *Communications in Science and Technology* 1, 1 (2016).

[12] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 2012. 3D convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 1 (2012), 221–231.

[13] Naresh Kumar and Nagarajan Sukavanam. 2018. Motion trajectory for human action recognition using fourier temporal features of skeleton joints. *Journal of Image and Graphics* 6, 2 (2018), 174–180.

[14] Jia Lu, Minh Nguyen, and Wei Qi Yan. 2020. Deep learning methods for human behavior recognition. In *IEEE International Conference on Image and Vision Computing New Zealand (IVCNZ)*. 1–6.

[15] Jia Lu, Jun Shen, Wei Qi Yan, and Boris Bačić. 2017. An empirical study for human behavior analysis. *International Journal of Digital Crime and Forensics (IJDCF)* 9, 3 (2017), 11–27.

[16] Jia Lu, Wei Qi Yan, and Minh Nguyen. 2018. Human behaviour recognition using deep learning. In *IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. 1–6.

[17] Keiron O'Shea and Ryan Nash. 2015. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458* (2015).

[18] Niall O'Mahony, Sean Campbell, Anderson Carvalho, Suman Harapanahalli, Gustavo Velasco Hernandez, Lenka Krpalkova, Daniel Riordan, and Joseph Walsh. 2019. Deep learning vs. traditional computer vision. In *Science and Information Conference*. Springer, 128–144.

[19] Christian Schuldt, Ivan Laptev, and Barbara Caputo. 2004. Recognizing human actions: A local SVM approach. In *IEEE International Conference on Pattern Recognition*, Vol. 3. 32–36.

[20] Sreelekshmy Selvin, R Vinayakumar, EA Gopalakrishnan, Vijay Krishna Menon, and KP Soman. 2017. Stock price prediction using LSTM, RNN and CNN-sliding window model. In *International Conference on Advances in Computing, Communications and Informatics*. 1643–1647.

[21] Alex Sherstinsky. 2020. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Physica D: Nonlinear Phenomena* 404 (2020), 132306.

[22] Shaohua Wan, Lianyong Qi, Xiaolong Xu, Chao Tong, and Zonghua Gu. 2020. Deep learning models for real-time human activity recognition with smartphones. *Mobile Networks and Applications* 25, 2 (2020), 743–755.

[23] Xiuhui Wang and Wei Qi Yan. 2020. Human gait recognition based on frame-by-frame gait energy images and convolutional long short-term memory. *International Journal of Neural Systems* 30, 01 (2020), 1950027.

[24] Xiuhui Wang, Jiajia Zhang, and Wei Qi Yan. 2020. Gait recognition using multichannel convolution neural networks. *Neural Computing and Applications* 32, 18 (2020), 14275–14285.

[25] Di Wu, Nabin Sharma, and Michael Blumenstein. 2017. Recent advances in video-based human action recognition using deep learning: A review. In *International Joint Conference on Neural Networks (IJCNN)*. 2865–2872.
[26] Wei Qi Yan. 2019. *Introduction to Intelligent Surveillance: Surveillance Data Capture, Transmission, and Analytics*. Springer.
[27] Wei Qi Yan. 2021. *Computational Methods for Deep Learning*. Springer.
[28] Zeqi Yu and Wei Qi Yan. 2020. Human action recognition using deep learning methods. In *IEEE International Conference on Image and Vision Computing New Zealand (IVCNZ)*. 1–6.
[29] Junsong Yuan, Zicheng Liu, and Ying Wu. 2011. Discriminative video pattern search for efficient action detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33, 9 (2011), 1728–1743.
[30] Seyma Yucer and Yusuf Sinan Akgul. 2018. 3D human action recognition with siamese-LSTM based deep metric learning. *arXiv preprint arXiv:1807.02131* (2018).
[31] Qingchang Zhu, Zhenghua Chen, and Yeng Chai Soh. 2018. A novel semisupervised deep learning method for human activity recognition. *IEEE Transactions on Industrial Informatics* 15, 7 (2018), 3821–3830.