Human Action Recognition Based on YOLOv7

Chenwei Liang and Wei Qi Yan
Auckland University of Technology, 1010 New Zealand

ABSTRACT

Human action recognition is a fundamental research problem in computer vision. The accuracy of human action recognition has important applications in robotics. In this book chapter, we are use of a YOLOv7-based model for human action recognition. To evaluate the performance of the model, the action recognition results of YOLOv7 were compared with those using CNN+LSTM, YOLOv5, and YOLOv4. Furthermore, a small human action dataset suitable for YOLO model training is designed. This data set is composed of images extracted from KTH, Weizmann, MSR data sets. In this book chapter, we make use of this data set to verify the experimental results. The final experimental results show that using the YOLOv7 model for human action recognition is very convenient and effective, compared with the previous YOLO model.

Keywords: Human action recognition, Attention mechanism, Deep learning, YOLO, YOLOv7

INTRODUCTION

Typically, surveillance recordings often have a sequence of actions (Yan, 2019). The act of recognizing the movements shown in these videos may provide significant advantages, such as promptly recognizing individuals who have had a fall down and providing them with assistance to mitigate the subsequent issues arising from the accident (An, 2020; An & Yan, 2021; Lu, et al., 2021; Zhu & Yan, 2022; Liu & Yan, 2022). Hence, the performance of human action recognition on videos has significant importance. The term "human action recognition" often refers to the process of evaluating or analysing the categories of human activities seen in videos (Soomro, et al., 2014). In a succinct manner, the objective is to accurately group human behaviours into established action categories (Herrera et al., 2008).

While recognizing these actions, it also entails a substantial amount of labor. Hence, the significance of a rapid and efficient action recognition approach cannot be overstated. The pertinent techniques in the field of deep learning have the capability to fulfil the specified criteria and effectively address this issue. Deep learning has gained significant popularity and widespread use since its inception (Yan, 2021). The objective of this methodology is to facilitate the training of computers, enabling them to analyse and discern certain data (Gao et al., 2021). The use of computing machines for human action recognition leads to an improvement in recognition efficiency as time progresses.

Human action recognition has garnered significant attention within academic discourse. Historically, a substantial body of research project on the recognition of human actions has included typical machine-learning techniques, which involve the extraction of visual characteristics or motion trajectories. Additionally, more contemporary approaches have emerged in the form of deep learning methods. The introduction of deep learning has considerable significance as a pivotal component within the field of machine learning. The use of this method is prevalent not just within the domain of computer vision but also across several disciplines, including natural language processing and robotics (Wiriyathammabhum et al., 2016). Since the advent of deep learning, there has been a notable increase in the use of human action recognition. Currently, several recognition algorithms have been suggested consecutively, including CNN (Khan et al., 2020), Two-Stream (Simonyan et al., 2014), C3D (Convolution 3 Dimension) (Tran et al., 2015), and RNN (Du et al., 2017), etc.

Similar to the Convolutional Neural Network (CNN) architecture, You Only Look Once (YOLO) model has an input layer, convolutional layer, pooling layer, and fully connected layer. The aforementioned study conducted by Redmon et al. (2016) establishes the fundamental framework for a comprehensive Convolutional Neural Network (CNN) architecture. However, YOLO exhibits a clear differentiation from the conventional CNN model. The achievement of end-to-end object detection necessitates the use of a distinct CNN model. This enhancement results in improved computational efficiency of the YOLO model. This is one of the reasons why this book chapter selects the YOLOv7 model for human action recognition.

This book chapter employs YOLOv7 framework to construct a comprehensive network for human action recognition. The YOLO algorithm, which stands for “You Only Look Once”, is a visual object identification method that utilizes a convolutional neural network. This approach was firstly introduced in 2016. One of the key benefits of this particular approach is in its inherent simplicity and efficiency, which allows for swift execution. According to Cao et al. (2023), the YOLOv7 model exhibits notable advancements in terms of both running speed and structure. This research effort primarily focuses on the investigation of fundamental human actions. This work presents an evaluation of the efficiency of the YOLOv7 model in human action recognition, therefore contributing to the existing body of knowledge in this field.

In the subsequent sections of this work, we will provide an overview of the current methodologies used in human action recognition, namely in section two. In section three, the methodology is expounded upon. In section four, an analysis and discussion of the experimental outcomes will be conducted. In the concluding part, we provide a comprehensive summary of our research and provide a perspective on potential future endeavors.

LITERATURE REVIEW

Deep learning is a unique approach in the field of human action recognition. CNN is the most representative one. However, the CNNs tend to only be able to handle 2D inputs when dealing with incoming data. In order to make this model more widely applicable, and also to automatically recognize human actions on the screen, Ji et al. (Ji et al., 2012) created a 3D-CNN network for human action recognition by using deep learning methods. The basic principle of this method is to perform a special 3D convolution on the input video, extract features of the spatial and temporal relationships of the input, find the motion information between different adjacent frames according to the obtained features, and acquire the final recognition result based on this information. Unlike the widely popular classifier approach at the time, this CNN model was much simpler. It is a very convenient and easy-to-use model. The final experimental results show that the 3D-CNN model is very effective for human action recognition in the environment, compared with other methods at the time, the effect of this method is the best.

The Two-Stream method is another mainstream direction of deep learning in the direction of human action recognition. It was firstly proposed in 2014 (Simonyan & Zisserman, 2014). The structure of Two-Stream CNN network is roughly divided into two parts. One part is applied to process RGB images, the other is employed to process optical flow images. Finally, the processed images are jointly trained and classified. Furthermore, they demonstrate that training multi-frame density optical flow on ConvNet networks can achieve promising results with limited training data. A multi-task training method was offered to combine two different action classification datasets so as to increase the training data, finally the achieved results are based on both datasets. The experimental results significantly outperformed past recognition results using deep learning networks.

C3D (3-Dimensional Convolution) was proposed in 2015 as one of the mainstream methods at the same time as Two-Stream (Tran et al., 2015). This approach demonstrates rapid and effective acquisition of spatiotemporal features. This is due to the use of deep 3D Convolutional Networks (3D ConvNets) for the processing of spatiotemporal information. Furthermore, the network acquires characteristics that are accompanied by a straightforward linear classifier. The conclusive examination results indicate that the model exhibits a notable degree of computational efficiency. Specifically, the model demonstrates a streamlined end-to-end training process and has a succinct network topology that facilitates ease of training and utilization.

Subsequently, Feichtenhofer et al. made enhancements to the two-stream network with the aim of attaining superior performance (Feichtenhofer et al., 2016). By integrating the two networks into a convolutional layer rather than a SoftMax layer, it achieves parameter efficiency without compromising performance. Optimal network fusion was shown to be most effective at the last convolutional layer. Furthermore, it has been shown that using additional fusion at the class prediction layer might enhance the accuracy of the model. A novel architectural design was proposed based on the aforementioned enhancements. Ultimately, the suggested design has been validated as the capable of attaining state-of-the-art results.

Despite the considerable success obtained by the two-stream strategy, it is important to acknowledge that this approach still has a significant limitation. In other words, the model is incapable of accurately representing videos of extended duration. Hence, the TSN network, also known as Temporal Segments Networks, was introduced by Wang et al. (2016) as a modification of the two-Stream CNN approach. To address this issue, a very innovative approach is employed inside this network. The proposed approach involves partitioning a lengthy video into segments, followed by the random selection of a short segment from each partition. Subsequently, the aforementioned two-stream method is applied to analyse and process the selected segments. Finally, a fusion method is employed to integrate the outcomes. The existing approaches for two-stream-based methodologies mostly rely on the use of Temporal Segment Networks (TSN) as the underlying framework.

Among various methods for human action recognition by using deep learning, prediction is the most widely employed one. A method to predict human motion trajectories to recognize actions and behaviours was proposed. This method will be employed for human action recognition in videos (Azorin-Lopez et al., 2016). On the basis of previous research results, the deep learning framework of LSTM was treated as a new model for predicting human behaviour (Almeida & Azkune, 2018). In this model, the neural network includes a probabilistic model by learning and simulating the different behaviours of human interaction with the environment, which enables algorithms not only to predict human behaviour but also to further identify abnormal human behaviour.

On the basis of previous research results, the deep learning framework of LSTM was treated as a new model for predicting human behaviour (Almeida & Azkune, 2018). In this model, the neural network proposes a probabilistic model by learning and simulating the different behaviours of human interaction with the environment, which enables algorithms not only to predict human behaviour but to further identify abnormal human behaviour.

In order to deal with the dimensional information in the videos, especially the temporal information. The processing of time-required classes has always been a point in human action recognition. As we all know, the RNN network is very suitable for dealing with time-series problems. Therefore, RNN networks were expected to resolve the problem of human action recognition. A recurrent pose-attention network (RPAN) was proposed (Du et al., 2017, Cao, 2022, Cao & Yan, 2022). In this approach, a mechanism called pose-attention was proposed. This allows the RNN models to learn more complex motion structures over time. In addition to this, the method can also perform coarse pose labeling of actions in videos. The proposal of this method enriches the kinds of networks that make use of RNN for human action recognition. Finally, the experimental results prove that the network is very excellent.

In 2018, a new multi-level recurrent residual network (MRRN) was proposed (Zheng et al., 2017). Three recognition streams were combined in this network. Each distinct stream consists of a residual network (ResNet) and a recurrent model. The model captures spatiotemporal information by learning spatial representations from static frames using two optional meshes and modelling temporal dynamics by using Stacked Simple Recurrent Units (SRUs). The streams of three independently learned low-, mid- and high-level representations are fused by computing a weighted average of SoftMax scores to obtain a complementary representation of the video. In addition, the complexity of this model is greatly reduced. This reason is that the model reduces complexity by using shortcut connections and trains end-to-end with greater efficiency. Unlike previous models that improved performance at the expense of time and space complexity, our experimental results demonstrate that the final performance of MRRN is significantly improved. It achieves an accuracy of 81.9% based on the UCF-101 dataset.

A recurrent structure called a coupled recurrent network (CRN) (Sun et al., 2018) was proffered. The network can handle action recognition from multiple input sources. In CRN, parallel streams of RNNs are coupled together. Among them, the Circular Interpretation Block (RIB) plays a key role. This is the reason why this module enables the network to learn reciprocal feature representations from multiple signals in a recurrent fashion. Meanwhile, an efficient CRN training strategy was proposed, which differs from RNNs that stack training losses at each time step or the last time step. The final experiments not only demonstrated the effectiveness of the method but also achieved progress in human action recognition and multi-person pose estimation.

Human action recognition is in fact the process of enabling machines to understand human actions in digital videos and label them accordingly. Human behaviour is greatly affected by external factors such as lighting conditions and background. Hence, three specific models and methods for human action recognition were employed. These methods are all based on Convolutional Neural Networks (CNN), namely, 3D-CNN, Two-

Stream CNN, and CNN+LSTM, respectively. These models are merged together after extensive testing on the HMDB-51 dataset. Compared to the final experimental results, it is found that only the CNN+LSTM method can effectively avoid interference factors (Yu & Yan, 2020).

With the further improvement of computing speed and processing power, the application prospect of deep neural networks (DNN) is getting brighter. YOLO was proposed. Lu et al. (Lu et al., 2018) completed the research work based on human action recognition using YOLOv3 model in 2018, the final accuracy rate of the work was as high as 80.20%. In this experiment, the public datasets were applied to conduct experiments. In addition, by continuously adjusting the network structure, the network learning rate with the highest accuracy is obtained. To make the experimental results more convincing, YOLOv2 was compared with YOLOv3. The experimental results show that the method is very effective and accurate.

Further optimized deep networks were designed for human action recognition (Gowdra, et al. 2021). The advanced methods were developed after YOLOv3 to solve the human action recognition problem (Lu et al., 2020). This new method employs the network architecture YOLOv4 + LSTM to achieve the results. After extensive experiments, the accuracy of the network reached 97.87%. This is the reason why they recognize temporal and spatial information in the recognition method. Furthermore, they also experimentally validated a Selective Kernel Network (SKNet) model with an attention mechanism. The use of this model yielded better results.

There are other methods for human action recognition. For example, Wang et al. took use of frame-by-frame recognition of human gait energy images (Wang & Yan, 2020; Liu &Yan, 2020) and later human gait recognition based on multichannel convolutional neural networks (Wang, Zhang, & Yan, 2020). Both methods have achieved good experimental results in the end. However, there are still few methods to realize human action recognition based on YOLO. Most are using other deep learning techniques like CNN.

Nguyen & Bui (2023) introduced a novel approach that integrates deep learning and machine learning techniques for the purpose of categorizing the activities and behaviours of numerous human subjects. The suggested methodology encompasses a three-step recognition procedure, which includes the use of the YOLOv5 model for object detection, the implementation of a media pipeline for skeletal visualization, and the utilization of an LSTM network for action identification. The experimental findings pertaining to target detection using YOLOv5 indicate that the loss accuracy consistently maintains below 5% throughout the training and validation processes. The mean precision (mAP) of the YOLOv5 model created for all the examined case studies consistently exceeds 99%.

Cao et al. (2023) conducted a study that focuses on the development of a pedestrian detection and identification system using YOLOv7. The objective of the algorithm is to effectively differentiate between regular walkers and power grid maintenance inspectors who are approaching to a high level of accuracy. The work provides a description of the algorithm design process, which involves categorizing pedestrians and workers into two distinct groups. This categorization is based on discernible characteristics, such as the presence of safety helmets and the State Grid insignia.

Upon conducting a comprehensive analysis of the existing methodologies for human action recognition, it has been observed that the study pertaining to the utilization of YOLOv7 for this purpose lacks of a comparative evaluation of the performance across different iterations of the YOLO series. We will conduct a series of comparisons for this work.

METHODOLOGY

In this book chapter, we present a novel direction that distinguishes itself from previous research work. This method employs YOLOv7 framework for the purpose of human action recognition. In this study, we evaluate the efficacy of YOLOv7 across a range of criteria. Furthermore, we conducted a comparative analysis between YOLOv7 and three other models, namely, YOLOv4, YOLOv5, and CNN+LSTM.

Dataset

There are a large number of datasets currently available for human action recognition, such as KTH, The HMDB-51 Dataset, UCF-101, etc. In this book chapter, we choose the open public dataset KTH for testing for convenience. The KTH dataset contains six classes of human behaviours, which are further divided into four classes in the specific classification, each class contains 25 topics. There are 100 videos in total (Schuldt et al., 2004). The videos were all shot with a static camera with a frame rate of 25 fps. The shooting backgrounds are all uniform backgrounds, the resolution of the video is 160 × 120, the average length is 4 seconds. Specific human actions include "walking", "jogging", "running", "boxing", "waving", and "clapping". In our experiments, we selected all six action classes for this book chapter. But our dataset does not directly use video data as our training data.

Regarding the selection of the data set, a total of 1,200 images were chosen as representatives. This selection process was based on the integration of the KTH dataset with the Weizmann and MSR datasets. Each human action is represented by a collection of 200 photographs, resulting in a total of six distinct groups. Each collection of images represents a distinct and consistent action shown within the dataset. The training set for our model consists of 80% of the dataset. To conduct our analysis, we will designate a test set consisting of 20% of the data set, along with an additional 100 sample movies.

Network structure

By using this approach, we successfully accomplish our objective of recognizing human actions. The YOLOv7 model (Wang, Bochkovskiy, & Liao, 2022) is the most recent one to the YOLO series of models. In the YOLOv7 model, the network underwent a complete resizing process, where the input picture was scaled to a resolution of 640×640. The process begins by feeding an image into the backbone network, which then generates three layers of feature maps with various sizes via the head layer network. Ultimately, the prognostication outcomes are exported via the Rep and Conv layers. The architectural design of the whole network remains on Convolutional Neural Networks (CNN). One of the components of the system is the backbone network, which is linked to an ELAN (Extended Local Area Network) by a series of four Convolutional layers followed by Batch Normalisation and SiLU (Sigmoid Linear Unit) activation functions (referred to as CBS). This is then further augmented by three Maxpooling layers and a combination of CBS and ELAN structures. Each structure is associated with the output of a primary layer. Figure 1 illustrates the simplified structural diagram of the whole YOLOv7 model.

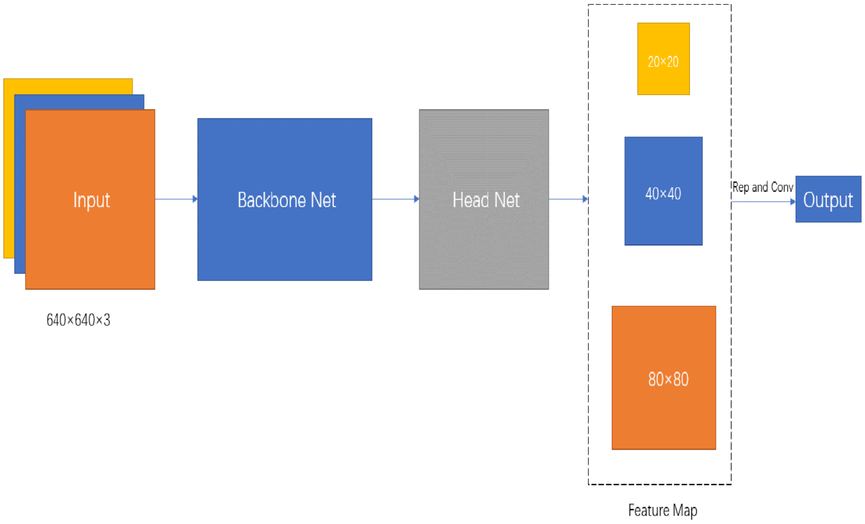

Fig 1. The simplified structure of YOLOv7 model

RESULT ANALYSIS

In this book chapter, we designated the batch size as 16 and the number of threads as 8. Consequently, we fed 16 samples into the network, distributing these data among the 8 threads for network training. Additionally, the epoch number is 250. In our experiments, we are use of GPU acceleration as a means to enhance the efficiency of our training process, hence mitigating time consumption. Followed the completion of model training, the data was next employed to evaluate the performance in the domain of human action recognition. The analysis yielded the following outcomes. In order to do the tests, the batch size remains fixed at 16 and the number of threads is set at 8.

Figure 2 shows the results of human action recognition using YOLOv7, which are all generated based on test video frames, including (a) boxing, (b) clapping, (c) waving, (d) jogging, (e) walking and (f) running.

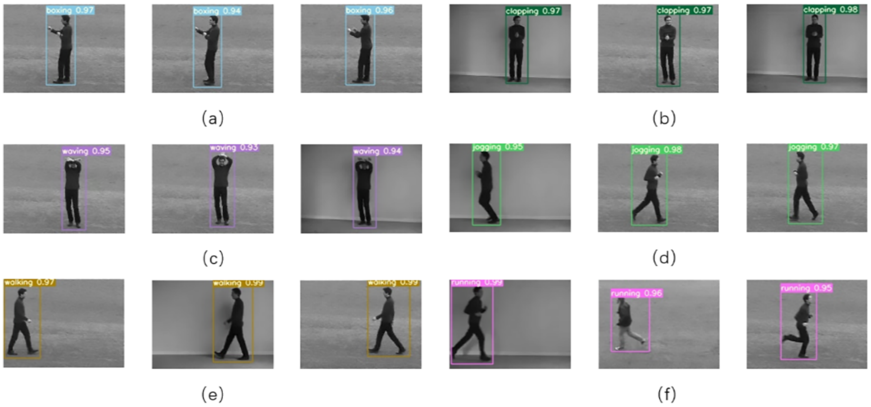

Fig 2. The result of human action recognition by using YOLOv7

In this experiment, it is important to assess the impact of the model. Recall and accuracy are employed as the evaluation metrics for assessing the performance of our trained model. The precise trajectories of the two indicators are shown in Fig 3. It is evident that with an increase in the number of epochs, there is a tendency for both metrics to converge and stabilize at about 80. Both exhibit a large magnitude. This demonstrates that our trained model has both high accuracy and a broad scope. Furthermore, in Fig 3, it is seen that the loss value shown on the graph decreases progressively with each iteration. The graphs representing Objectness, Classification, Validation Box, Validation Objectness, and Validation Classification exhibit similar trends. The mean average precision at a threshold of 0.5 (mAP@0.5) demonstrates an inverse relationship with the values shown in the figures. A lower value in these figures indicates a higher performance of the model. Additionally, as the number of iterations increases, both mAP@0.5 values steadily improve until they reach a stable state.

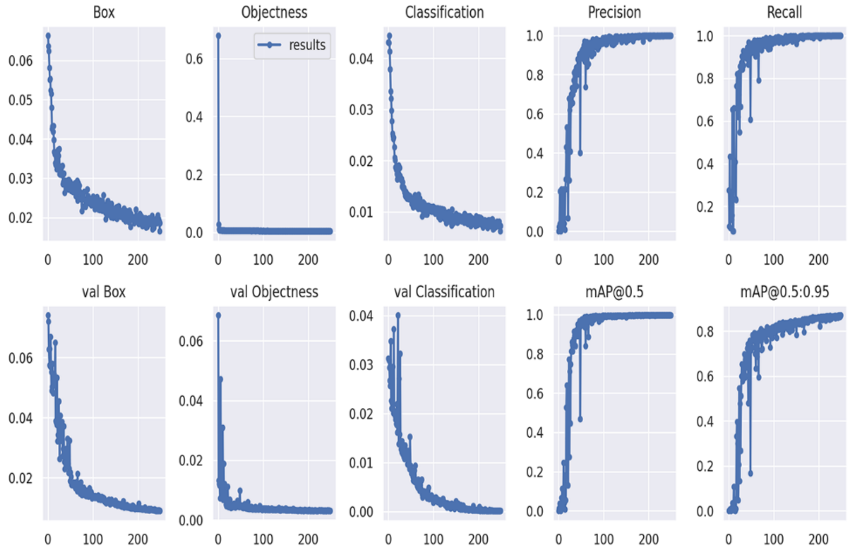

Fig 3. The result

YOLOv4 and YOLOv5 models were adopted in this work. The trials were conducted by using the two models individually. In all phases of model training, we established a total of 250 epochs, while maintaining a consistent batch size of 16. A portion of the experimental findings is shown in Fig 4 and Fig 5. The experimental findings of YOLOv4 are shown in Fig 4. The experimental findings of YOLOv5 are shown in Fig 5.

Based on the partial findings shown in the figure, it is evident that the accuracy of the two YOLO models is marginally inferior to that of the YOLOv7 model. In order to assure precision, we opted identical data sets for testing purposes across three distinct models. The outcomes of these tests are visually shown in Fig 6.

Fig 6 illustrates the sequential presentation of the outcomes obtained from the analysis of the same dataset using YOLOv4, YOLOv5, and YOLO7, respectively. Both YOLOv4 and YOLOv5 exhibit a commendable accuracy of 0.95 when applied to the task of action boxing. The YOLOv7 model achieves an accuracy of 0.96. The precision of the remaining two models for the task of clapping action is 0.96. The YOLOv7 model achieves an accuracy of 0.97. It is evident that the identification accuracy of YOLOv7 surpasses that of the other two models by a small margin. Furthermore, we have identified other issues that were not detected inside a singular model. The temporal efficiency of the three models likewise varies. In a given dataset, the execution time for YOLOv7 epochs is around 90 seconds on average. The relative running times for the YOLOv5 and YOLOv4 models are 120 seconds per epoch and 135 seconds per epoch. Compared with the three models, it is seen that YOLOv7 exhibits a little higher speed in execution, while maintaining the same level of GPU acceleration.

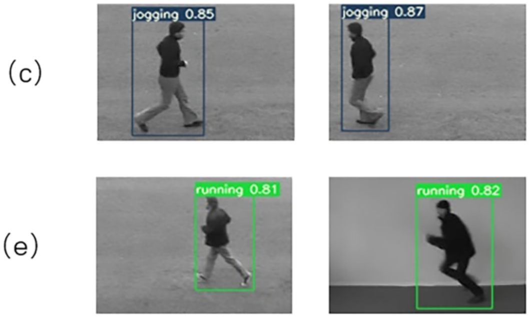

Fig 4. YOLOv4 part results

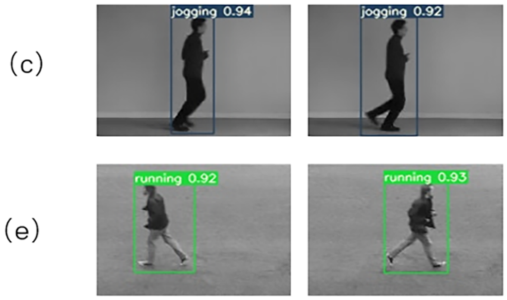

Fig 5. YOLO5 part results

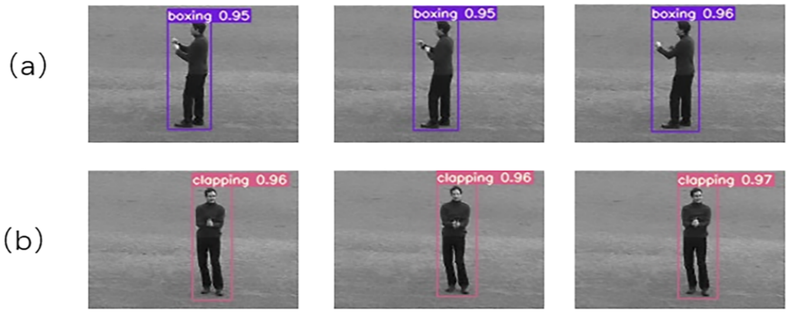

Fig 6. The results of three different YOLO

The ultimate accuracy is obtained by developing the YOLOv7 model. Furthermore, we also compared the results with the recognition accuracy of the CNN+LSTM model using the same dataset (Liang, Lu, & Yan, 2022). The data shown in Table 1 was acquired. In contrast, it is evident that the use of YOLOv7 yields superior accuracy. The overall accuracy of the system stays about 0.96. Of the six motions, boxing moves exhibit the greatest level of precision.

Table 1. Classification precision for action recognition

Models	Walking	Jogging	Running	Boxing	Waving	Clapping	Total
YOLOv7	0.96	0.96	0.96	0.98	0.96	0.97	0.96
YOLOv5	0.95	0.94	0.95	0.96	0.95	0.95	0.95
YOLOv4	0.91	0.92	0.93	0.94	0.93	0.93	0.93
CNN+LSTM	0.91	0.85	0.92	0.90	0.88	0.87	0.88

In addition, we also added the attention mechanism CBAM (Woo, Park, Lee, & Kweon, 2018) and the attention mechanism SimAM (Yang, Zhang, Li, & Xie, 2021) to the network structure of YOLOv7 respectively. By adding this structure, we can further evaluate whether YOLOv7 is suitable for human action recognition. The results are shown in Table 2.

Table 2. Classification Accuracy of Action Recognition Based on YOLOv7.

Models	Walking	Jogging	Running	Boxing	Waving	Clapping	Total
YOLOv7	0.96	0.96	0.96	0.98	0.96	0.97	0.96
CBAMYOLOv7	0.963	0.965	0.97	0.976	0.966	0.975	0.96
SimAMYOLOv7	0.959	0.961	0.968	0.982	0.976	0.957	0.961

From Table 2, we observe that the action can still be recognized effectively after using the attention mechanism. Using the attention mechanism also improves recognition accuracy to a certain extent. This shows that the applicability of YOLOv7 is excellent.

To determine whether the effect of recognition will be affected by the dataset. We also chose “Running” and “Walking” classes from the Weizmann dataset (Gorelick, Blank, Shechtman, Irani, & Basri, 2007) to test our model. The Weizmann dataset is as same as the KTH dataset, the actions in the data are performed by only one person. The experimental results are shown in Fig 7. The results show that changing such datasets does not have much impact on the recognition performance. The model using YOLOv7 effectively recognizes the corresponding action. The recognition accuracy is also not much different from the original dataset. But there are also drawbacks, such as the detection box will incorrectly identify the background. This is where improvement is needed.

In order to examine the potential variation in recognition accuracy in response to complicated backdrop changes within the dataset, the MSR dataset was adopted. The dataset only comprises three distinct motions, namely “Clapping”, “Waving”, and “Boxing”. One video within this dataset encompasses all activities, resulting in an average video duration of over 30 seconds. Moreover, the dataset comprises individuals engaging in various activities (Yuan, Liu, & Wu, 2011). The experimental findings are shown in Figure 8. It is evident that the YOLOv7-based model is capable of accurately identifying and recognizing the three specified activities. However, the accuracy of the system is influenced by intricate environmental factors and other contextual elements, resulting in a significant decrease in overall precision by a factor of twelve. This phenomenon may be attributed to the insufficient amount of relevant data available in the dataset used for training purposes. Additional data is necessary to adequately address the intricacies of some complicated ecosystems.

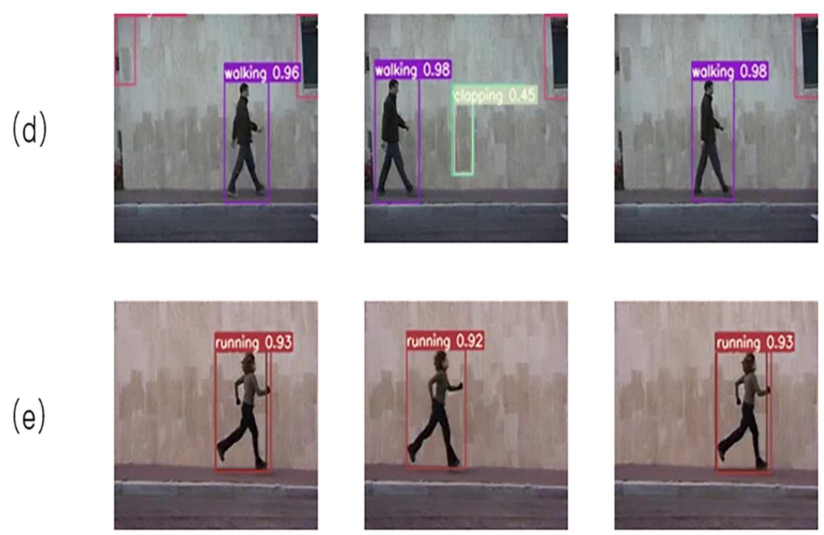

Fig 7. The results of Weizmann dataset.

Fig 8. The results MSR dataset

The aforementioned test findings indicate that the use of YOLOv7 for the purpose of human action recognition consistently demonstrates efficacy and feasibility, irrespective of variations in the dataset. Nevertheless, there are some constraints. The selection of a data set has a significant impact on the ultimate outcome of recognition. There exists ample potential for enhancing model advancements.

CONCLUSION

In this book chapter, we present an investigation into three distinct deep learning-based YOLO model detection methodologies, specifically focusing on their efficacy in pedestrian and human action identification. The preliminary evaluation of human action recognition using YOLOv7 yields good findings. This study presents the first evidence supporting the applicability of YOLOv7 in the domain of human action recognition. A novel action recognition dataset, specifically designed for compatibility with the YOLOv7 model, is developed by using the existing datasets. At present, this study demonstrates the capability to accurately identify and classify six pre-determined behaviours with consistent performance.

Nevertheless, there are also some issues that need to be addressed. The absence of data pertaining to further acts remains a notable gap, therefore indicating a potential avenue for future research and investigation.

In further research endeavours, our efforts will be made to further enhance the current YOLOv7 model. There is an expectation that the enhancement of the model will lead to a significant improvement in the recognition impact. Furthermore, we will add pre-judgment conditions to the model. This reason is from our experimental results. When there are multiple tags in the recognized image. For example, in the action of a video, multiple tags alternate or exist simultaneously. While adding this condition, the action will only be recognized if the conditions are met. We will continue to find suitable datasets for deeper evaluation of human action recognition models using YOLOv7 (Liu, et al. 2023).

REFERENCES

Almeida, A., & Azkune, G. (2018). Predicting human behaviour with recurrent neural networks. Applied Sciences, 8 (2), 305.

An, N., Yan, W. (2021) Multitarget tracking using Siamese neural networks. *ACM Transactions on Multimedia* Computing*, Communications and Applications.*

An, N. (2020) *Anomalies Detection and* Tracking *Using Siamese Neural Networks.* Master's Thesis. Auckland University of Technology, New Zealand.

Azorin-Lopez, J., Saval-Calvo, M., Fuster-Guillo, A., & Garcia-Rodriguez, J. (2016). A novel prediction method for early recognition of global human behaviour in image sequences. Neural Processing Letters, 43 (2), 363–387.

Bochkovskiy, A., Wang, C.-Y., & Liao, H.-Y. M. (2020). YOLOv4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934.

Cao, W., Li, L., Gong, S., & Dong, X. (2023, May). Research on human behaviour feature recognition and intelligent early warning methods in safety supervision scene video based on YOLOv7. Journal of Physics: Conference Series (Vol. 2496, No. 1, p. 012019). IOP Publishing.

Cao, X. (2022) *Pose Estimation of Swimmers from Digital Images Using Deep Learning*. Master's Thesis, Auckland University of Technology.

Cao, X. and Yan, W. (2022) Pose estimation for swimmers in video surveillance. Multimedia Tools and Applications, Springer.

Du, W., Wang, Y., & Qiao, Y. (2017). Rpan: An end-to-end recurrent poseattention network for action recognition in videos. In IEEE International Conference on Computer Vision (pp. 3725–3734).

Feichtenhofer, C., Pinz, A., & Zisserman, A. (2016). Convolutional two-stream network fusion for video action recognition. In IEEE Conference on Computer Vision and Pattern Recognition (pp. 1933–1941).

Gao, X., Nguyen, M., & Yan, W. Q. (2021). Face image inpainting based on generative adversarial network. In 36th International Conference on Image and Vision Computing New Zealand (IVCNZ) (pp. 1–6).

Gorelick, L., Blank, M., Shechtman, E., Irani, M., & Basri, R. (2007, December). Actions as space-time shapes. IEEE Transactions on Pattern Analysis and Machine Intelligence, 29 (12), 2247–2253.

Gowdra, N. (2021) Entropy-*Based Optimization Strategies for Convolutional Neural Networks.* PhD Thesis, Auckland University of Technology, New Zealand.

Herrera, A., Beck, A., Bell, D., Miller, P., Wu, Q., Yan, W. (2008) Behaviour analysis and prediction in image sequences using rough sets. *International Machine Vision and Image Processing Conference* (pp.71-76)

Ji, S., Xu, W., Yang, M., & Yu, K. (2012). 3d convolutional neural networks for human action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35 (1), 221–231.

Khan, M. A., Javed, K., Khan, S. A., Saba, T., Habib, U., Khan, J. A., & Abbasi, A. A. (2020). Human action recognition using fusion of Multiview and deep features: An application to video surveillance. Multimedia Tools and Applications, 1–27.

Liang, C., Lu, J., & Yan, W. (2022). Human action recognition from digital videos based on deep learning. ACM ICCCV 2022.

Liu, C., Yan, W. (2020) Gait recognition using deep learning. *Handbook of Research on Multimedia Cyber Security* (pp.214-226) IGI Global.

Liu, J., Yan, W. (2022) Crime prediction from surveillance videos using deep learning. *Aiding Forensic* Investigation *Through Deep Learning and Machine Learning Frameworks*. IGI Global.

Liu, Y., Nand, P., Hossain, A., Nguyen, M., Yan, W. (2023) Sign language recognition from digital videos using feature pyramid network with detection Transformer. *Multimedia Tools and* Applications.

Lu, J., Yan, W. Q., & Nguyen, M. (2018). Human behaviour recognition using deep learning. In IEEE AVSS (pp. 1–6).

Lu, J., Nguyen, M., & Yan, W. Q. (2020). Deep learning methods for human behaviour recognition. In IEEE IVCNZ (pp. 1–6).

Lu, J., Nguyen, M., Yan, W. (2020) Comparative evaluations of human behaviour recognition using deep learning. Handbook *of Research on Multimedia Cyber Security*, 176-189.

Lu, J., Nguyen, M., Yan, W. (2021) Sign language recognition from digital videos using deep learning methods. *International Symposium on Geometry and Vision.*

Lu, J. (2021) *Deep Learning* Methods *for Human Behaviour Recognition*. PhD Thesis. Auckland University of Technology, New Zealand.

Nguyen, A. T., & Bui, H. A. (2023, March). Multiple target activity recognition by combining YOLOv5 with LSTM network. In The International Conference on Intelligent Systems & Networks (pp. 400-408). Singapore: Springer Nature Singapore

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In IEEE Conference on Computer Vision and Pattern Recognition (pp. 779–788).

Sanjaya, S. A., Adi Rakhmawan, S. (2020). Face mask detection using MobileNetv2 in the era of COVID-19 pandemic. International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI).

Schuldt, C., Laptev, I., & Caputo, B. (2004). Recognizing human actions: A local SVM approach. In IEEE International Conference on Pattern Recognition (Vol. 3, pp. 32–36).

Shen, D., Chen, X., Nguyen, M., Yan, W. Q. (2018). Flame detection using deep learning. International Conference on Control, Automation and Robotics (ICCAR). https://doi.org/10.1109/iccar.2018.8384711

Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. Advances in Neural Information Processing Systems, 27.

Singh, S., Ahuja, U., Kumar, M., Kumar, K., &amp; Sachdeva, M. (2021). Face mask detection using YOLOv3 and Faster R-CNN models: COVID-19 environment. Multimedia Tools and Applications, 80(13), 19753–19768.

Soomro, K., & Zamir, A. R. (2014). Action recognition in realistic sports videos. In Computer Vision in Sports (pp. 181–208). Springer.

Sun, L., Jia, K., Shen, Y., Savarese, S., Yeung, D. Y., & Shi, B. E. (2018) Coupled recurrent network (CRN). arXiv preprint arXiv:1812.10071.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3D convolutional networks. In IEEE International Conference on Computer Vision (pp. 4489–4497).

Venkateswarlu, I. B., Kakarla, J., &amp; Prakash, S. (2020). Face mask detection using mobilenet and global pooling block. IEEE Conference on Information &amp;amp; Communication Technology (CICT).

Wang, H., Yan, W. Q. (2022). Face detection and recognition from distance based on Deep Learning. Advances in Digital Crime, Forensics, and Cyber Terrorism, 144–160.

Wang, C.-Y., Bochkovskiy, A., & Liao, H.-Y. M. (2022). YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. arXiv:2207.02696.

Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., & Gool, L. V. (2016). Temporal segment networks: Towards good practices for deep action recognition. In European Conference on Computer Vision (pp. 20–36).

Wang, X., Yan, W. (2019) Human gait recognition based on self-adaptive hidden Markov model. *IEEE/ACM* Transactions *on Biology and Bioinformatics*.

Wang, X., Yan, W. (2019) Cross-view gait recognition through ensemble learning. *Neural Computing* and *Applications.*

Wang, X., & Yan, W. Q. (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), 1950027.

Wang, X., Zhang, J., & Yan, W. Q. (2020). Gait recognition using multichannel convolution neural networks. Neural Computing and Applications, 32 (18), 14275–14285.

Wang, X., Yan, W. (2020) Non-local gait feature extraction and human identification. *Springer Multimedia Tools and Applications.*

Wang, X., Yan, W. (2020) Cross-view gait recognition through ensemble learning. *Neural* Computing *and Applications* 32 (11), 7275-7287.

Wang, X., Yan, W. (2022) Human identification based on gait manifold. *Applied Intelligence.*

Wiriyathammabhum, P., Summers-Stay, D., Ferm¨uller, C., & Aloimonos, Y. (2016). Computer vision and natural language processing: recent approaches in multimedia and robotics. ACM Computing Surveys (CSUR), 49 (4), 1–44.

Woo, S., Park, J., Lee, J.-Y., Kweon, I. S. (2018). CBAM: Convolutional block attention module. ECCV, 3–19. https://doi.org/10.1007/978-3-030-01234-2_1

Wu, P., Li, H., Zeng, N., Li, F. (2022). FMD-YOLO: An efficient face mask detection method for COVID-19 prevention and control in public. Image and Vision Computing, 117, 104341.

Yan, W. (2019) *Introduction to Intelligent Surveillance: Surveillance Data Capture,* Transmission*, and Analytics*. Springer London.

Yan, W. (2023) *Computational Methods for Deep Learning: Theory, Algorithms, and Implementations*. Springer.

Yang, G., Feng, W., Jin, J., Lei, Q., Li, X., Gui, G., & Wang, W. (2020). Face mask recognition system with YOLOv5 based on image recognition. IEEE 6th International Conference on Computer and Communications (ICCC).

Yang, L., Zhang, R.-Y., Li, L., & Xie, X. (2021). Simam: A simple, parameterfree attention module for convolutional neural networks. In International Conference on Machine Learning (pp. 11863–11874).

Yu, J., Zhang, W. (2021). Face mask wearing detection algorithm based on improved Yolo-V4. Sensors, 21(9), 3263.

Yu, Z., & Yan, W. Q. (2020). Human action recognition using deep learning methods. In IEEE IVCNZ (pp. 1–6).

Yuan, J., Liu, Z., & Wu, Y. (2011). Discriminative video pattern search for efficient action detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 33 (9), 1728–1743.

Zheng, Z., An, G., & Ruan, Q. (2017). Multi-level recurrent residual networks for action recognition. arXiv:1711.08238.

Zhou, D., Fang, J., Song, X., Guan, C., Yin, J., Dai, Y., &amp; Yang, R. (2019). IOU loss for 2D/3D object detection. 2019 International Conference on 3D Vision (3DV). https://doi.org/10.1109/3dv.2019.00019

Zou, Z., Chen, K., Shi, Z., Guo, Y., &amp; Ye, J. (2023). Object detection in 20 years: A survey. Proceedings of the IEEE, 111(3), 257–276.

Zhu, Y., Yan, W. (2022) Ski fall detection from digital images using deep learning. *ACM ICCCV.*