International Journal of Computer Applications Technology and Research
Volume 13–Issue 05, 23 - 25, 2024, ISSN:-2319–8656
DOI:10.7753/IJCATR1305.1006

Human Action Recognition for Outdoor Monitoring System

Panca Mudjirahardjo
Dept. of Electrical Engineering
Faculty of Engineering, Brawijaya University
Malang, Indonesia

Rahmadwati
Dept. of Electrical Engineering
Faculty of Engineering, Brawijaya University
Malang, Indonesia

Abstract: Action recognition is one of computer vision task that involves recognizing human actions in videos or images. The goal is to classify and categorize the actions being performed in the video or image into a predefined set of action classes. Human Action Recognition (HAR) is a challenging task that has various applications, such as intelligent video surveillance and human-computer interaction. In this paper we describe a method of HAR for outdoor monitoring system. For detection a person we utilize YOLO v.8. then from the sequential frames we extract a motion feature. We use pre-trained VGG-16 model for spatial feature extraction. Then, this feature vector is fed to action classifier.

Keywords: human action recognition; outdoor environment; YOLO; VGG-16 model; LSTM

1. INTRODUCTION

Human Activity Recognition (HAR) is an exciting research area in computer vision and human-computer interaction. HAR objective is to identify human actions that are depicted in videos [1, 2]. In general, HAR research involves enhancing monitoring processes from environmental, spatial and temporal data.

Automatic detection of human physical activity has become crucial in pervasive computing, interpersonal communication, and human behavior analysis.

The broad usage of HAR benefits human safety and general well-being. Health monitoring can be done through wearable devices tracking physical activity, heart rate, and sleep quality. In smart homes, HAR-based solutions allow for energy saving and personal comfort by detecting when a person enters or leaves a room and adjusting the lighting or temperature. Personal safety devices can automatically alert emergency services or a designated contact. And that's just the tip of the iceberg.

There are many HAR framework based on input, feature extraction and action classifier. Agahian, et.al [4] using skeleton joints as input, skeleton information as feature and SVM and ELM as action classifier. Kilis, et. Al [5] using skeleton joints as input, Directed Graph Neural Network (DGNN) to compute feature vector, and Graph Convolutional Network (GCN) for action recognition.

2. THE PROPOSED METHOD

Our proposed method is shown in Figure 1.

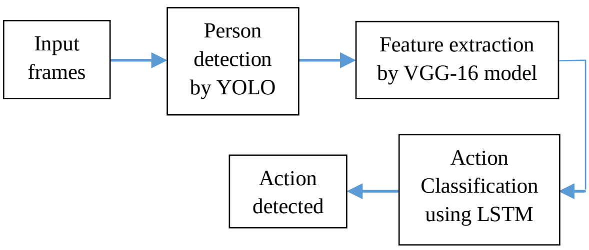

Figure 1. Our proposed method

Our system input are video frames captured by surveillance camera. To localize person as interest input, we utilize YOLO v.8. Then in each frame, we have person to be recognized his/her action. We utilize the pre-trained VGG-16 model to extract motion feature. Finally, the feature vector is fed to classifier. We utilize LSTM as action recognition.

2.1 Human Detection

In outdoor environment, there are many moving object captured by a camera surveillance, such as human, vehicles, pets, goods carried by a person, and so on. To process a moving person, first we localize the person.

We utilize YOLO v.8 to detect and localize a person. As we know YOLO (You Only Look Once) is a robust object detection framework in various poses.

YOLO is a groundbreaking real-time object detection algorithm. Its unique approach treats object detection as a regression problem, utilizing a single convolutional neural network to spatially separate bounding boxes and associate probabilities with detected objects. This innovation enables YOLO to perform real-time object detection with unprecedented speed and accuracy, making it a versatile solution across various domains [6][7].

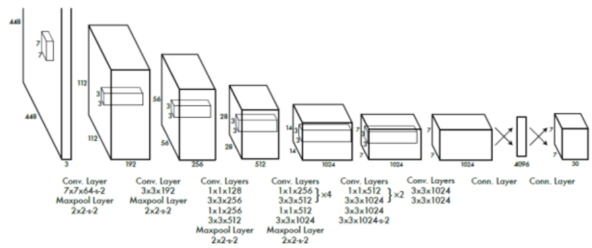

Figure 2. YOLO architecture

2.2 Feature Extraction

Currently, CNNs are widely used as effective spatial feature extractors which better than traditional handcrafted feature extraction methods. The handcrafted feature extraction method has a drawback, i.e. it is difficult to identify the correct combination of features for the classification task [3]. CNN architecture is depicted in Figure 3.

CNN has solution for this drawback, due to it has capability for automatic feature extraction. For feature extraction, our method uses a pre-trained VGG-16 model. The VGG-16 network consists of sixteen convolution layers, five max-pooling layers, and three dense layers and uses relatively small filters of size 3×3, which is helpful for local feature extraction. The last pooling layer of the VGG-16 network is used for extracting local features from the frames of a video clip. The extracted feature vector size is 7×7×512 for each frame, which is given as input to the LSTM network for classification.

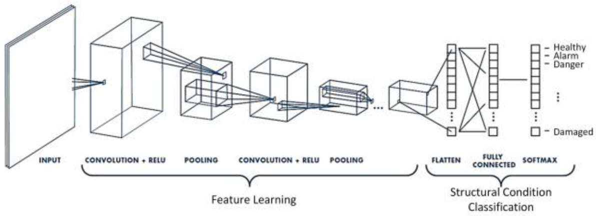

Figure 3. CNN architecture

2.3 Classification

Long Short-Term Memory (LSTM) is a form of Recurrent Neural Network (RNN) which has been effectively used for a variety of sequential data-related tasks, including Human Activity Recognition (HAR). LSTM architecture is shown in Figure 4.

LSTM models, like other RNNs, are designed to analyze data sequences and save internal memories of prior inputs, enabling them to retain the temporal connections between different sections of the sequence.

The main benefit of LSTMs over all other RNNs is their capacity to forget or retain information from previous time steps consciously. This aids in solving the issue of vanishing gradients, which frequently occur in regular RNNs. LSTMs can effectively simulate long-term dependencies inside the input sequence. They're well-suited for complicated HAR tasks such as identifying anomalies and recognizing complex human actions.

LSTM-based models demonstrated significant gains in HAR tasks in various benchmark datasets, attaining state-of-the-art performance. They have also shown resilience in detecting complicated activities and dealing with variable-length input sequences. However, just like other models based on deep learning, LSTMs have several drawbacks for HAR: the requirement for vast volumes of labeled data, computational cost, and model interpretability.

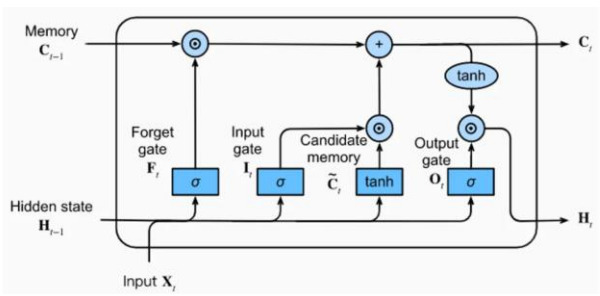

Figure 4. LSTM architecture

2.4 Evaluation

The proposed method is evaluated by confusion matrices. We use our dataset as outdoor video. The activity to be evaluated are riding a bicycle (RB), walking (W), hand waving (HW), kick a ball (KB) and running (R). Frames of video that consist of the activities is shown in Figure 5. There are 50 activities for each activity to be evaluated with our method.

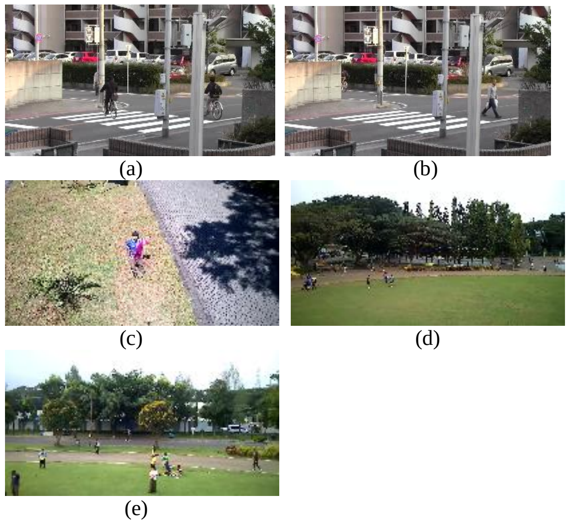

Figure 5. Activity to be recognized, (a) riding a bicycle (b) walking (c) hand waving (d) kick a ball (e) running.

3. THE EXPERIMENTAL RESULT

In this section, we explain our experimental result. The confusion matrices is shown in Figure 6. We evaluate 50 activities in every activity. The highest accuracy is for hand waving of 94%. This is due to no shifting, this activity is only hand moving. The lowest accuracy is for running activity of 72%. This is due to the activity similar to riding a bicycle and walking.

	RB	W	HW	KB	R
RB	40	5			5
W	1	40			9
HW			47	3	
KB		5		45	
R	4	10			36

Figure 6. Confusion matrices for each activity

4. CONCLUSIONS

We have demonstrated our method for HAR in outdoor environment. The accuracy of recognition is satisfy enough, i.e. between 72% - 94%.

The future work are improving the foreground extraction and motion feature extraction.

5. ACKNOWLEDGMENTS

We would like to express our thanks to the Faculty of Engineering, Brawijaya University, which has supported this research funding.

6. REFERENCES

[1] Donahue, J., Hendricks, L.A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T. 2015. *Long-term recurrent convolutional networks for visual recognition and description*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[2] Wang, Y., Xiao, Y., Xiong, F., Jiang, W., Cao, Z., Zhou, J. T., Yuan, J. 2020. *3dv: 3d dynamic voxel for action recognition in depth video*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[3] Parui, S.K., Biswas, S.K., Das S. 2023. *An Efficient Human Action Recognition System Using Deep-Learning Based Method*.

[4] Agahian, S., Negin, F., Kose, C. 2019. *An efficient human action recognition framework with pose-based spatiotemporal features*. Engineering Science and Technology, an International Journal.

[5] Kilis, N., Papaioannidis, C., Mademlis, I., Pitas, I. 2022. *An efficient framework for human action recognition based on graph convolutional networks*. International Conference in Image Processing (ICIP).

[6] Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. *You Only Look Once: Unified, Real-Time Object Detection*. IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[7] Pebrianto, W., Mudjirahardjo, P., Pramono, S.H., 2022 .*YOLO Method Analysis and Comparison for Real-Time Human Face Detection*. 11th Electrical Power, Electronics, Communications, Controls and Informatics Seminar (EECCIS).