2020 2nd International Conference on Sustainable Technologies for
Industry 4.0 (STI), 19-20 December, Dhaka

Human Activity Recognition System from Different Poses with CNN

Md. Atikuzzaman[*],Tarafder Razibur Rahman[†], Eashita Wazed [‡],
Md. Parvez Hossain[§], and Md. Zahidul Islam[¶]
Department of Computer Science and Engineering, Green University of Bangladesh, Dhaka, Bangladesh[*†‡§¶]
Email: atikuzzaman524[*]@gmail.com, rajiburrahmantrafder[†]@gmail.com, eashitawazed[‡]@gmail.com,
parvezhossain92[§]@gmail.com, zahid[¶]@cse.green.edu.bd

2020 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI) |978-1-6654-0489-1/20/$31.00 ©2020 IEEE | DOI: 10.1109/STI50764.2020.9350508

***Abstract*—In the principle of Human Activity Recognition, a variety of real-life implementations are available using different types of sensors such as fitness monitoring, day-life monitoring, health monitoring, etc. Especially for the elders, sensor-based applications are not feasible due to many reasons such as carrying a mobile phone or gadgets. In this paper, we focused on CCTV videos and camera images to detect human poses using HAAR Feature-based Classifier and recognize the activities of the human using the Convolutional Neural Network (CNN) Classifier. Our Human Activity Recognition System was trained using our own collected dataset which is composed of 5648 images. The approach accomplished an efficacious detection accuracy of 99.86% and recognition accuracy of 99.82% with approximately 22 frames/second after 20 epochs.**

***Index Terms*—Human Activity Recognition, CNN, HAAR Classifier, Human Poses Recognition**

I. Introduction

Recognition of human behavior from different positions The use of CNN is an inevitable part of the process of recognition. The aim of the system is to detect human activity and to recognize different classes of body movement from videos. Detection of human behavior The use of CNN is a mixture of modules such as object detection , segmentation and recognition. Alongside image collection and processing,we have been applied to a breakthrough method called HAAR Feature-based Classifier [1]. It is a machine learning based approach where a cascade function is trained from a lot of positive and negative images. Recognition of the activity, which used to identify the category of actions, that helps in different type of activity, walking, running, standing, sitting and laying.

The dataset includes human photographs from various events that are clicked in various poses. And the system will train and test and send the output to the dataset. In our study, we proposed a separate approach that differs from traditional work, which can define and understand the human activity class.

The suggested approach is outlined in this section and is divided into two categories: dataset and system design. System design is divided into three categories: human detection / localization, segmentation, and video frame recognition. Complex features and temporal information could be used simultaneously to detect and classify activity classes in order to achieve high precision for variations in environments. However, the computational load has to increase dramatically as more characteristics are measured. When detecting and recognizing the activity class from videos or images, we must build a system that can minimize the computation time and achieve the most accuracy.

II. Related Work

Piergiovanni *et al.* present the idea of filters of temporal attention and explain how they can be used from videos to identify human behavior. [2]. They used HMDB5 and DogCentric datasets for image training testing and used four CNN architecture feature model underlays. The author confirms that the suggested definition of filters for temporal attention benefits the identification of behavior.

Bevilacqua *et al.* propose to use Convolutional Neural Network (CNN) to classify human activities [3]. They discuss many combinations of activities and sensors, illustrating how motion signals can be adapted using various network architectures to be fed into CNN. They use CNN for behavior detection and recognition here. They show that the output is better in this paper and this means that collinearities occur among the signals sampled on distinct placements with sensors.

Ignatov *et al.* introduce a first approach to HAR based on deep learning models [4]. In order to feed real images to a convolutionary neural network, they create a spectrograph image from an inertial signal. This approach overcomes the need to reshape the signals for a CNN in an acceptable format, but the step of spectrogram generation simply replaces the function extraction process, adding initial overhead to the training of the network. CNN with real-time behavior classification for local feature extraction.

Bevilacqua *et al.* propose to use convolutional nural network to classify human activities [5]. For data collection, they use 5 kinds of IMU sensors. This paper demonstrates how, using various network architectures, motion signals can be modified to be fed into CNN. They also evaluate the performance of different sensor classes, investigating the ability of single, double and triple sensor systems to be graded. Their experimental findings were collected on a dataset of 16 very promising lower-limb behaviors, gathered from a group of participants using five separate sensors.

978-1-6654-0489-1/20/$31.00©20XX IEEE

TABLE I
IMAGE CLASS AND QUANTITY OF THE DATASET

Image Class	Quantity
Laying	566
Running	1016
Sitting	1152
Standing	1591
Walking	1323

Shinde *et al.* provide an approach to identify, locate and recognize near-real - time behavior of interest from frames collected from a continuous stream of video data that can be recorded from a surveillance camera. [6]. They show that in the Liris Human Activities dataset, YOLO is an efficient tool and comparatively quick for recognition and localization. After a given time, the model takes input frames and is able to offer action labels based on a single frame. They predicted the action mark for the video stream by averaging findings over a particular time.

Singh *et al.* present a deep learning model that learns without using any previous experience to identify human activities. [7]. For this function, on three real world smart home datasets, a Long Short Term Memory Recurrent Neural Network was implemented. The findings of these experiments show that in terms of accuracy and efficiency, the proposed method outperforms the current ones.

Zhang *et al.* develop a gait recognition system based on the Siamese neural network to automatically extract stable and discriminatory gait characteristics for human identification. [8]. Instead of a raw sequence of gaits, they composite the gait energy images. By applying Siamese network-based gait recognition to classify individuals, they suggested a structure.

He *et al.* proposed a new technique that can detect objects in an image effectively while generating a high-quality segmentation mask for each instance at the same time, called Mask-RCNN. [9]. Mask R-CNN is easy to train and adds only a modest overhead, operating at 5 fps, to Faster R-CNN. With the COCO16 dataset, we use this process. And they explain how accurate it is for detection to perform well.

III. DATASET AND DATA COLLECTION

The experiment was conducted by our own collected dataset and the collection has been done with the laptop camera and CCTV camera and image frames from different videos. We have collected 5648 different activity images in several kinds of weather and light conditions. Our dataset contains 60 humans data. The data is collected from indoor and outdoor, rooftop, room, the road in low light and bright light conditions. As a result, the image resolution is not good enough which introduces a challenge to correctly detect and recognize human activities. The dataset is composed of only five different classes of human activity and the quantity of the images is not the same for all classes. The number of images of different classes is shown in table I. To train and test the system, we used 80% and 20% images respectively.The dataset consists of five different classes and the classes are : Walking, Running, Standing, Sitting, and Laying.

An example of different classes of the dataset is shown in Fig. 1. are Running, Walking, Standing, Sitting, Laying.

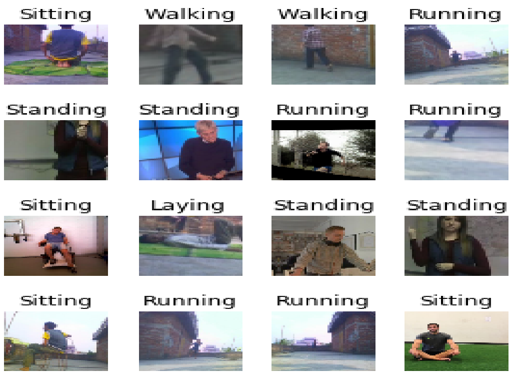

Fig. 1. Example Images of Various Classes from Our Dataset

IV. METHODOLOGY

The proposed human activity recognition system is separated into three subsections human detection, segmentation, and recognition of human activity from the videos or images with high precision and lower time complexity. The system architecture of human activity recognition is shown in Fig. 2.

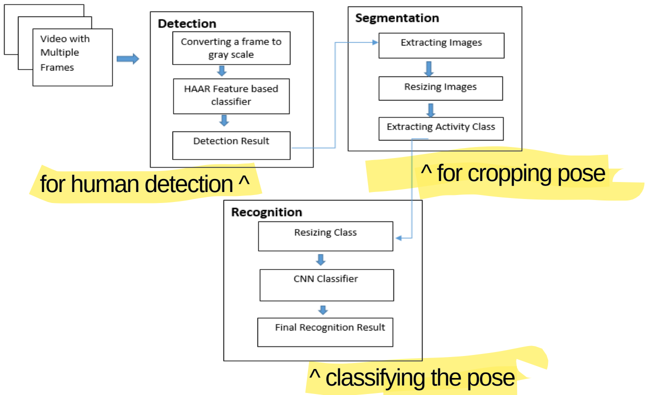

Fig. 2. Proposed Human Pose Detection And Recognition System Architecture

A. Human Pose Detection

The detection intends to prepare all frames from an input video in real-time because continuous frames may hold human or not that is obliged to develop a proper human activity recognition system. Pose activities may not be recognized in delivered frames due to the various reasons, for example, the human may be hidden or partially hidden by other objects or mostly shaded when they are out of focus. We trained our HAAR Feature-based classifier using only human images to detect the human poses from the input videos or images.HAAR Feature-based classifier [10] is a machine learning-based approach and

data occlusion

TABLE II
CONFIGURATION OF HAAR FEATURE-BASED CLASSIFIER

boostType	GAB
featureType	HAAR
BG threshold	80
Invert	FALSE
Max intensity deviation	40
Max x angle	1.1
Max y angle	1.1
Max z angle	0.5
Show samples	FALSE
Width	64
Height	64
Max Scale	-1

the cascade the classifier is trained by plenty of the true and false images then it applied to detect objects in other images. The adaptation of the HAAR feature-based classifier of human detection is described in the table. IV-A.

$$Y = (0:299 * R) + (0:587 * G) + (0:114 * B) \qquad (1)$$

To detect the human from a video frame or images, the input image needs to be converted into gray-scale and this is done by the Function. 1. An example of RGB to gray-scale is shown in Fig. 3. The final result of the HAAR Feature-based classifier is the (X, Y,H, W) where (X, Y) is the origin coordinate of the human and H Height of the human and W is the Width human of the input image.

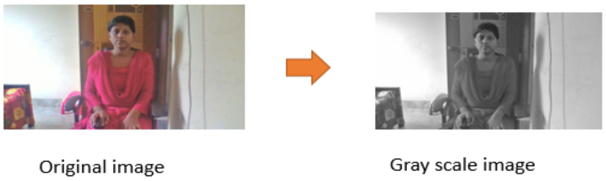

Fig. 3. Converting RGB to Gray-scale Image

B. Human Pose Segmentation

Human segmentation is essential for so many reasons such as extracting the two or many people in one single frame, removing the background, etc. The human segmentation is done by the result of the HAAR feature-based classifier where we get (X, Y, H, W), (X, Y) as starting coordinate and H height of the human and W is the width of the human. After deriving the human we resized the image into 64 X 64 resolution. An example of the input and segmented image is shown in Fig. 4.

C. Activity Recognition

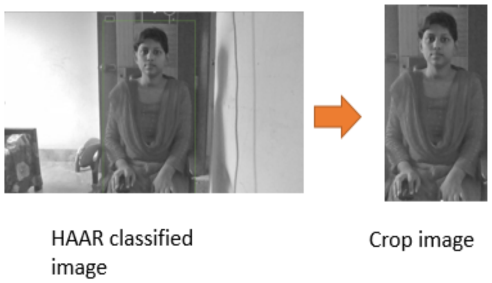

Fig. 4. Segmentation of Input Image

Convolution Neural Network has been adequately applied in the recognition system. And it's frequently applied to solve the problems of deep learning and pattern recognition. CNN's are inspired by biological procedure, the layers of Convolution Neural Network have neurons that are grouped in height, width, and depth. The neurons in the layer must be linked to a small part of the layer before it. Hidden layers form with the Convolutional layer, ReLU layer (Activation function), Pooling layer, Fully connected layer, and Normalization layer. The convolutional layer calculates the output neurons that are linked to the native area in the input. ReLU layer use activation function like maximum (0, x) threshold zero. The pooling layer accumulates the outputs of neurons bunch at one layer and makes the single layer in the following layer. A fully connected layer associated with each neuron in one layer to another layer and determine the class score which is shown in Fig. 5.

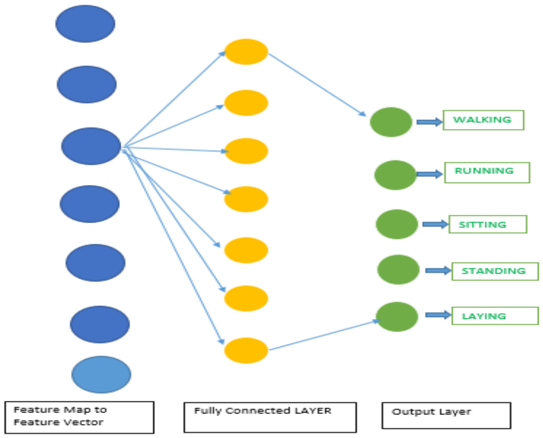

Fig. 5. CNN Classifier with Different Layers

To preprocess the input dataset of human activity recognition of the CNN network, we resized the images with 64 x 64 pixels of extracted human images. To train the CNN classifier 80% of the dataset of five different is used and 20% is used to validate the classifier. The summary of our CNN network is given in the table. IV-C. It is a deep CNN classifier [11]. with input, convolution, normalization, activation, dense, flatten, dropout layers.

TABLE III
CNN LAYERS SUMMARY OF ACTIVITY RECOGNITION

Layer (type)	Output Shape	Parameters
conv2d$_1$($Conv2D$)	(None, 64, 64, 64)	1792
activation$_1$($Activation$)	(None, 64, 64, 64)	0
batch normalization$_1$($Batch$)	(None, 64, 64, 64)	256
conv2d$_2$($Conv2D$)	(None, 64, 64, 64)	36928
activation$_2$($Activation$)	(None, 64, 64, 64)	0
batch normalization$_2$($Batch$)	(None, 64, 64, 64)	256
max pooling2d$_1$($MaxPooling2$)	(None, 32, 32, 64)	0
dropout$_1$($Dropout$)	(None, 32, 32, 64)	0
conv2d$_3$($Conv2D$)	(None, 32, 32, 128)	73856
activation$_3$($Activation$)	(None, 32, 32, 128)	0
batch normalization$_3$($Batch$)	(None, 32, 32, 128)	512
conv2d$_4$($Conv2D$)	(None, 32, 32, 128)	147584
activation$_4$($Activation$)	(None, 32, 32, 128)	0
batch normalization$_4$($Batch$)	(None, 32, 32, 128)	512
max pooling2d$_2$($MaxPooling2$)	(None, 16, 16, 128)	0
dropout$_2$($Dropout$)	(None, 16, 16, 128)	0
conv2d$_5$($Conv2D$)	(None, 16, 16, 256)	295168
activation$_5$($Activation$)	(None, 16, 16, 256)	0
batch normalization$_5$($Batch$)	(None, 16, 16, 256)	1024
conv2d$_6$($Conv2D$)	(None, 16, 16, 256)	590080
activation$_6$($Activation$)	(None, 16, 16, 256)	0
batch normalization$_6$($Batch$)	(None, 16, 16, 256)	1024
max pooling2d$_3$($MaxPooling2$)	(None, 8, 8, 256)	0
dropout$_3$($Dropout$)	(None, 8, 8, 256)	0
flatten$_1$($Flatten$)	(None, 16384)	0
dense$_1$($Dense$)	(None, 256)	4194560
activation$_7$($Activation$)	(None, 256)	0
batch normalization$_7$($Batch$)	(None, 256)	1024
dropout$_4$($Dropout$)	(None, 256)	0
dense$_2$($Dense$)	(None, 256)	65792
activation$_8$($Activation$)	(None, 256)	0
batch normalization$_8$($Batch$)	(None, 256)	1024
dropout$_5$($Dropout$)	(None, 256)	0
dense$_3$($Dense$)	(None, 5)	1542
activation$_9$($Activation$)	(None, 5)	0

V. RESULT ANALYSIS

The performance analysis of our system has been completed in two separate levels, firstly we have analyzed the detection of the human poses, and then recognition of human activities which divide into five different classes.

A. Performance of human poses Detection

To perceive the human pose detection accuracy using the HAAR classifier, the following equation is used.

$$Detection\ Accuracy = \frac{Accurate\ Pose}{Total\ Pose} \times 100\%$$

In total 1130 human pose images are used to measure the accuracy where the system accurately detected 1114 human poses. Thus we got an exquisite accuracy rate of 99.86% for the human pose detection.

B. Performance of Human Activity Recognition

To investigate human activity recognition precision, We tested the CNN Classifier with 1039 human pose images and got an excellent accuracy rate of 99.82% with more moderate computational time. The train and validate loss for human activity recognition is also satisfactory which is shown in Fig. 6.

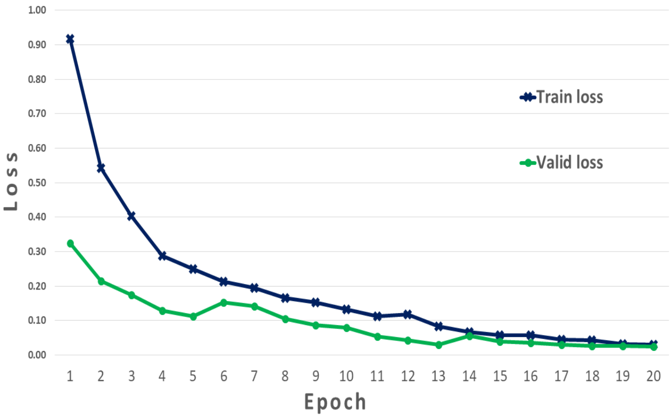

Fig. 6. Loss Rate with Several Epochs

After 20 epochs the train and validate loss of human activity recognition is 3.56% and 1.07% respectively. An accuracy rate concerning epochs is shown in Fig. 7.

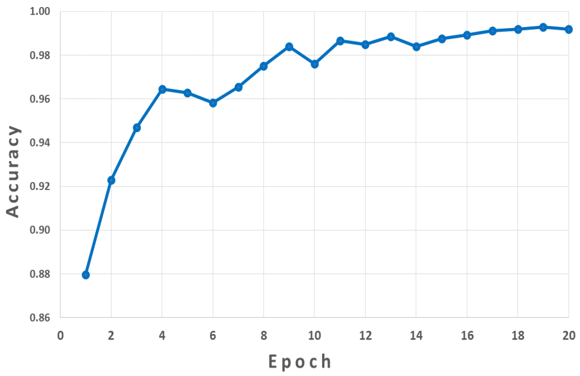

Fig. 7. Accuracy Rate with several Epochs

confusion?

A confession matrix of human activity recognition along with five distinct classes is shown in Fig. 8. Where two images of human pose activity are missed classified.

In table. IV Human pose activity recognition along with different classes recall and precision is represented in detail.

C. Comparison

Our system is a combination of HAAR Feature-based Classifier and Convolution Neural Network Classifier which is so

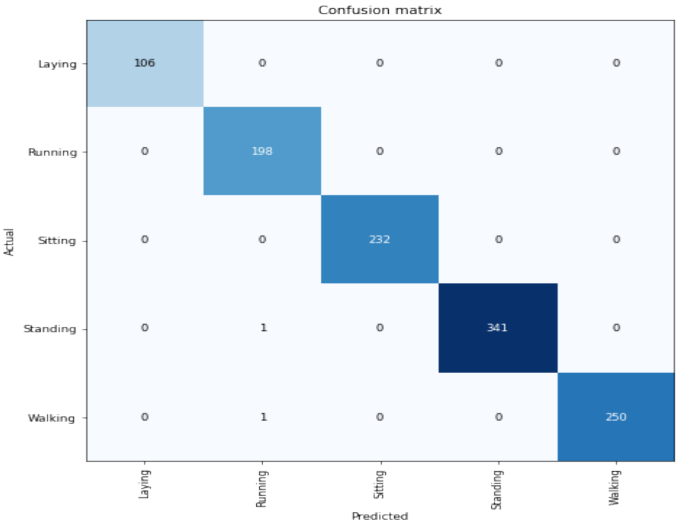

Fig. 8. Confusion Matrix of Human Activity Recognition

TABLE IV
RECALL AND PRECISION FOR HUMAN ACTIVITY RECOGNITION

Class	Recall	Precision
Laying	100%	100%
Running	99%	100%
Sitting	100%	100%
Standing	100%	99.708%
Walking	100%	99.602%

much distinct approach from other methods. The comparison of our system has been done with four different well established papers and each paper is distinct with others approaches. Given papers are complete their work in the circumstances of different poses. Piergiovanni *et al.* uses temporal filters and LSTMs Networks over HMDB5 Dataset and get accuracy 81.4% [2]. But they were unable to replicate the TDD's reported recognition performance of 63.2 % with the code provided by the authors. This probably is due to the difference in detailed parameter settings and engineering tricks.

From our paper we got average accuracy 99.82 % accuracy for using VGG19 CNN architecture.And we got less noise from dataset for using HAAR classifier,for this we have got a good accuracy.We teste it using our own dataset and we get better performance .

VI. CONCLUSION

We have developed a Human activity recognition system with the accuracy of 99.82% where the total no of images is counted. Which will help to track human activity to monitoring.

Since our machine had no GPU and low computing power, we were unable to train our system using RGB images. We only used uniform gray images to train and evaluate the CNN classifier on a 4 GB RAM general purpose computer and a 2.4 GHz Core i5 processor. More memory is needed for using some other proponent or using a deeper CNN model. A broad CNN model on the GPU system would certainly boost the overall performance of the recognition system for human activity.

- Due to the unavailability of various tasks, it can only identify five types of activity groups.
- Due to a lack of a broad data collection, we could not test our proposed human activity recognition system with distinct weather and light conditions.

We are gathering more data from various activities for future study, as we believe it will increase the diversity of recognition of other class letters. In addition, under certain strategies, our device can be used to detect and identify other events. We have also intended to establish a method of recognition of human behavior.

ACKNOWLEDGMENT

This research work is partially supported by the Green University of Bangladesh.

REFERENCES

[1] M. Atikuzzaman, M. Asaduzzaman, and M. Z. Islam, "Vehicle number plate detection and categorization using cnns," in *2019 International Conference on Sustainable Technologies for Industry 4.0 (STI)*. IEEE, 2019, pp. 1–5.

[2] A. Piergiovanni, C. Fan, and M. S. Ryoo, "Learning latent subevents in activity videos using temporal attention filters," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[3] A. Bevilacqua, K. MacDonald, A. Rangarej, V. Widjaya, B. Caulfield, and T. Kechadi, "Human activity recognition with convolutional neural networks," 09 2018.

[4] A. Ignatov, "Real-time human activity recognition from accelerometer data using convolutional neural networks," *Applied Soft Computing*, vol. 62, pp. 915–922, 2018.

[5] A. Bevilacqua, K. MacDonald, A. Rangarej, V. Widjaya, B. Caulfield, and T. Kechadi, "Human activity recognition with convolutional neural networks," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2018, pp. 541–552.

[6] S. Shinde, A. Kothari, and V. Gupta, "Yolo based human action recognition and localization," *Procedia computer science*, vol. 133, pp. 831–838, 2018.

[7] D. Singh, E. Merdivan, I. Psychoula, J. Kropf, S. Hanke, M. Geist, and A. Holzinger, "Human activity recognition using recurrent neural networks," in *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*. Springer, 2017, pp. 267–274.

[8] C. Zhang, W. Liu, H. Ma, and H. Fu, "Siamese neural network based gait recognition for human identification," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 2832–2836.

[9] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[10] K. Visakha and S. S. Prakash, "Detection and tracking of human beings in a video using haar classifier," in *2018 International Conference on Inventive Research in Computing Applications (ICIRCA)*. IEEE, 2018, pp. 1–4.

[11] Y. Chen and Y. Xue, "A deep learning approach to human activity recognition based on single accelerometer," in *2015 IEEE International Conference on Systems, Man, and Cybernetics*. IEEE, 2015, pp. 1488–1492.