Engineering and Technology Journal e-ISSN: 2456-3358

Volume 09 Issue 06 June-2024, Page No.- 4186-4189

DOI: 10.47191/etj/v9i06.02, I.F. – 8.227

© 2024, ETJ

Application of Human Activity Recognition in Supermarkets with Human Pose Estimation

Evander Christian Dumalang[1], Roberto Davin[2], Lina[3]

[1,2,3] Faculty of Information Technology, Tarumanagara University, Jl. Letjen S. Parman 1, Jakarta, 11440, Indonesia

ABSTRACT: Human Activity Recognition is the ability to interpret human body movements or movements through sensors and to determine human activities or actions. Most everyday human tasks can be simplified or automated if they can be identified through an activity management system. The introduction of human activities in supermarkets has many benefits in its development, such as Just Walk Out technologies such as Amazon Go and the prevention of theft of goods. Proposed research with the theme of human activity recognition that has a focus on supermarkets. The methods that will be used in this research are Human Pose Estimation and Random Forest Classifier with data collected from the internet. The designed application will first detect skeletons with Blazepose from human objects, which will then be classified by the Random Forest Classifier. Activities that can be classified by the designed application are standing, walking, and fetching. The output of the application is a classification of activities carried out in real-time with the camera. The test results on the training data get an accuracy value of 100%, precision of 100%, recall of 100%, and F1-Score of 100%. Using the test data produces a confusion matrix which shows that the model being trained has an accuracy value of 100%, precision of 100%, recall of 100%, and an F1-Score of 100%.

KEYWORDS: Human Activity Recognition, Human Pose Estimation, Random Forest Classifier, Activity Detection, Supermarket

I. INTRODUCTION

Human Activity Recognition is one of the fields of computer vision that is currently being explored. The widespread use of human activity recognition in everyday life has prompted significant research in this area. Human Activity Recognition applications for surveillance cameras, military systems, and health care systems have a major impact because they can overcome system weaknesses that people cannot [1].

Visual-based and sensor-based techniques are the two main techniques in human activity recognition [2]. Sensor-based solutions are more difficult to implement as they require the deployment of additional devices to track activity. This technique not only requires the use of more devices but also reduces user comfort by requiring the attachment of additional devices to the body.

As an alternative to sensor-based human activity recognition, visual-based human activity recognition has gained popularity. Without using a device connected to the body, visual-based Human Activity Recognition will detect objects directly from the image obtained by the camera.

Visual-based Human Activity Recognition experiments include using Deep Convolutional Neural Networks to group images directly from the camera, producing photos by segmenting human movement against the background image. The weakness of this method is that it fails when the human body and the background have the same composition, causing the segmentation process to fail [3]. The coordinates of the observed body parts will be used to train machine learning models such as the Random Forest Classifier and Decision Tree while using Human Pose Estimation with Blazepose to identify body parts [4].

The use of Deep Learning can be used in various cases. In supermarkets, there are many actions or actions taken by humans, such as pushing shopping carts, picking up goods, walking (looking for goods), and others. The development of Deep Learning can be applied by detecting human activity in supermarkets. This development has many benefits, such as Just Walk Out technology from Amazon Go, which uses human activity detection to find out what items are taken by customers without requiring a cashier, so it can make the store not require a long queue for each visitor or buyer at the store.

II. METHODS

The system is designed to receive input in the form of images directly from the camera or video. The image obtained is processed by BlazePose to obtain the skeleton, which will then be converted into a one-dimensional matrix to be predicted by the Random Forest Classifier model that has been trained. In this system, activity prediction will be carried out by making predictions on each issued frame. The system workflow can be seen in Figure 1 below.

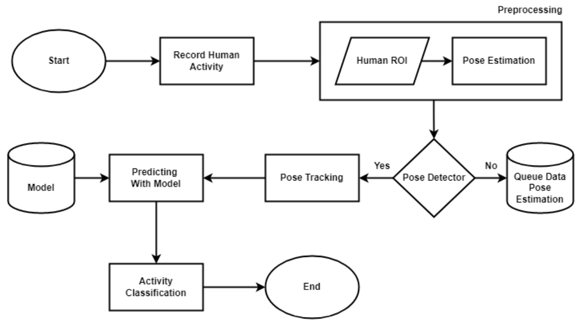


Figure 1. The Flowchart of the Human Activity Recognition System

The activity dataset is collected from the internet and augmented by flip, rotation, skew left/right, skew top/bottom, and shear operations. The activities collected were standing, walking, and picking up things. The open leg position will be considered as walking even though the object is not actually walking. The closed leg position will be considered as standing even though the object is not actually standing. The hand raised up position will be considered as picking up the item even though the object is not actually picking up the item. Details of the amount of data per class and its distribution can be seen in Table 1 below.

Table 1. Total Data By Activity Class

Activity Class	Data Amount
Standing	19.910
Walking	19.920
Picking	18.044
Total	**57.874**

Each activity image data has varying positions and lighting. From the dataset that has been obtained, it will be further processed with Human Pose Estimation (HPE) BlazePose to extract the skeleton data. The skeleton data thus formed will be converted into a one-dimensional matrix. Each of the 33 keypoints has 4 values, namely, x, y, z, and visibility. An illustration of this conversion process can be seen in Figure 2 below.

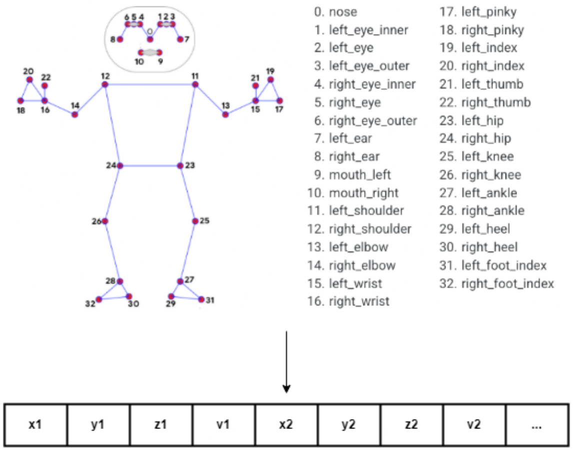


Figure 2. The process of converting skeleton data into matrix form

The HPE method that will be used in this system is BlazePose [5]. The architecture used by BlazePose to predict the skeletons of human objects is a modified stacked hourglass-based architecture. An encoder-decoder system is used to predict the heatmap of all joints, which is then followed by an encoder to calculate the coordinates of a particular joint.

BlazePose uses a detector-tracker system where every first frame, BlazePose will detect the region-of-interest (ROI) of human objects, and then on the next frame, BlazePose will track human objects on the previous ROI. If a human object is not detected, the detector will be run again.

The BlazePose detector detects human objects by detecting a fixed body part, namely the head. Thus, BlazePose's basic assumption in detecting humans is that the head must be visible in the frame. The number of dots that will be detected by BlazePose is 33 keypoints. The locations of 33 points can be seen in Figure 3.

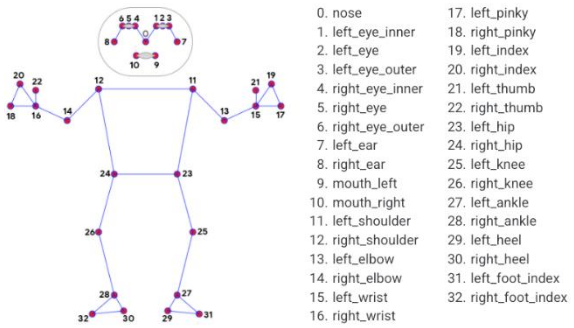


Figure 3. Location of 33 coordinate points of BlazePose

Each point that BlazePose detects will have 4 values, namely the x, y, z coordinates, and the visibility value. The visibility value is obtained from the implementation of the visibility classification system per point, which will indicate whether a point is blocked or has a low level of accuracy. This implementation allows BlazePose to estimate points that are either out of frame or completely blocked.

The Random Forest Classifier is a supervised machine learning approach that is based on ensemble learning. The Random Forest Classifier is the most adaptable and simple algorithm to learn and run. Ensemble learning is a type of learning in which different algorithms or the same method are combined multiple times to produce a robust predictive model. In other words, this method combines many algorithms of the same type, which implies that the Random Forest Classifier has multiple decision trees. One of its characteristics is that the Random Forest Classifier can control over-fitting [6].

Then, several hyperparameter values were tested on the Random Forest Classifier method with the initial configuration, namely test size 50%, decision tree 100, and max depth 0. The purpose of testing with several types of hyperparameters is to find the best hyperparameters that can be used for training the Random Forest Classifier model.

III. RESULTS AND DISCUSSIONS

There are several testing methods carried out to get the best model with several well-known metrics, namely accuracy, precision, recall, and f1-score. Accuracy is an indicator of the amount of data that is classified correctly from the total amount of data. Accuracy results are calculated by the following formula:

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (1)$$

Precision is the ratio of correctly classified data to the total amount of data that is predicted to be correct. Precision results are calculated by the following formula:

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

Recall is the ratio of correctly classified data to the total number of correct data. The recall results are calculated by the following formula:

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

An F1-score is a comparison of the median precision and recall, which is weighted. The F1-Score results are calculated by the following formula:

$$F1\ Score = 2 * \frac{(Recall * Precision)}{Recall + Precision} \quad (4)$$

The hyperparameter model testing is done iteratively, starting from the learning rate, batch size, and epoch. The best value of the predecessor hyperparameter will be taken and used in testing the next hyperparameter value. The best test results with hyperparameters can be seen in Table 2.

Table 2. Hyperparameter Test Results

Hyperparameter	Accuracy	Precision	Recall	F1-Score
Test Size 15%	100%	100%	100%	100%
Decision Tree 100	100%	100%	100%	100%
Max Depth 0	100%	100%	100%	100%

Based on the test results above, the most optimal hyperparameter is a test size of 15, a decision tree of 100, and a max depth of 0.

After getting a suitable hyperparameter configuration to use, the research team conducted an experiment to test the success of activity detection. This experiment was conducted using six pieces of video data containing activity scenarios. The scenarios for each video can be seen in Table 3.

Table 3. Tested Video Scenarios

Video	Time	Ground-Truth
1	28 seconds	Standing
2	8 seconds	Standing
3	11 seconds	Walking
4	15 seconds	Walking
5	9 seconds	Picking
6	7 seconds	Picking

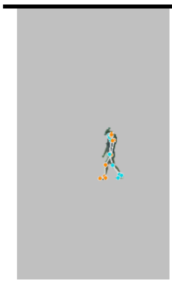
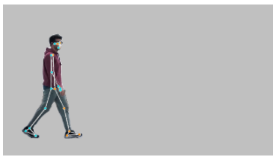
The test that will be carried out is a test with predictions for every frame and every second. The accuracy of testing various scenarios for the model that has been trained can be seen in Table 4. Examples of test video frame samples can be seen in Table 5.

Table 4. Scenario Video Predictions

Video	Prediction Type	Prediction Class	Prediction
1	Frame	Standing	95.172%
	Second	Standing	95.285%
2	Frame	Standing	92.35%
	Second	Standing	92.745%
3	Frame	Walking	84.853%
	Second	Walking	84.9%
4	Frame	Walking	95.199%
	Second	Walking	95.046%
5	Frame	Picking	96.537%
	Second	Picking	96.792%
6	Frame	Picking	87.42%
	Second	Picking	87.095%

Table 5. Sample Frame Prediction Results of Video Scenario Testings

Image	Activity	
	Ground-Truth	Prediction
	Standing	Standing
	Standing	Standing

	Walking	Walking
	Walking	Walking
	Picking	Picking
	Picking	Picking

CONCLUSIONS

Based on the results of the tests carried out, several conclusions can be drawn as follows: The designed system successfully detects children's activities with image datasets collected from the internet. The results of the evaluation of the training model obtained by the system have a value of 100% accuracy, 100% precision, 100% recall, and 100% F1-score. The Use of Human Pose Estimation in Human Activity Recognition means activities that have the same skeleton form can be misclassified. Like the activity of walking and standing, this is because there is no environmental context that can assist detection. The position of the camera during image capture is very influential. Therefore, we need a dataset that has varying camera angles. In the condition where the head of the human object is obscured by an object, BlazePose fails to detect the skeleton, thus making the system fail to predict. In general, detection every second produces better results. This is because the HPE results are unstable (changing). In future research, the use of BlazePose can be replaced with pre-trained human pose estimation models, which are more accurate in detecting poses. In addition, a dataset is also needed for more varied activities so that the detection results can be more accurate.

REFERENCES

1. Prastika, K., 2020, Aplikasi Pendeteksi Aktivitas Individu dalam Ruangan Menggunakan Metode CNN AlexNet, Computatio: Journal of Computer Science and Information Systems.
2. Hussain, Z., Sheng, M., dan Zhang, W. E., 2019, Different Approaches for Human Activity Recognition – A Survey, Journal of Network and Computer Applications, No. 102738, Vol. 167.
3. Gruosso, M., Capece, N., dan Erra, U., 2021, Human segmentation in surveillance video with deep learning, Multimedia Tools and Applications, Vol. 80, pp. 1175-1199.
4. Gupta, A., Gupta, K., Gupta, K., Gupta, K. Human Activity Recognition Using Pose Estimation and Machine Learning Algorithm, CEUR Workshop Proceedings, Vol. 2786.
5. Bazarevsky, V., Grishchenko, I., Raveendran, K., Zhu, T., Zhang, F., dan Grundmann, M., 2020, BlazePose: On-device Real-time Body Pose tracking, https://arxiv.org/abs/2006.10204, Last Accessed 1 September 2021.
6. Dogru, N., dan Subasi, A. Traffic accident detection using random forest classifier. In 15th learning and technology conference (L&T), pp. 40-45. IEEE.