JURNAL INFOTEL
Informatics - Telecommunication - Electronics
Website: http://ejournal.st3telkom.ac.id/index.php/infotel
ISSN: 2085-3688; e-ISSN: 2460-0997

Static and dynamic human activity recognition with VGG-16 pre-trained CNN model

Mawaddah Harahap[1,*], Valentino Damar[2], Sallyana Yek[3], Michael Michael[4], M. Ridha Putra[5]
[1,2,3,4,5]Universitas Prima Indonesia
[1,2,3,4,5]Jl. Sampul, Sei Putih Barat, Medan 20118, Indonesia
[*]Corresponding email: mawaddah@unprimdn.ac.id

Received 8 February 2023, Revised 12 March 2023, Accepted 28 March 2023

Abstract — Despite human activity recognition being an important research area for modern devices such as virtual reality and smart home technology with cameras, there is a problem with developing an accurate method for classifying human activities based on data recorded by these devices. This study aims to evaluate the performance of the pre-trained visual geometry group-16 model in classifying two types of human activity: static and dynamic. The dataset used for testing includes both public and local images, with the objective of achieving high accuracy. The dataset consists of 1,680 public datasets, divided into 80 % for training, 10 % for validation, and 10 % for testing (Test Data I). Additionally, 100 local images are used for further testing (Test Data II). The training and testing process was conducted to avoid overfitting. The study achieves impressive results, with a testing accuracy of 98.8 % using the public dataset and 97 % using the local dataset. This demonstrates the effectiveness of the visual geometry group-16 pre-trained model in accurately classifying static and dynamic human activities based on data recorded by camera devices.

Keywords – deep convolutional neural network, human activity recognition, image classification, visual geometry group-16

I. Introduction

Human activity recognition (HAR) aims to identify a person's specific movements or actions by utilizing data recordsmovement such as image o,r video data, and processed by an algorithm like deep learning, as shown in Fig. 1 [1]. The benefits generated with this technology are quite useful for modern devices, for instance, wearable health tracking technology [2], virtual reality devices [3], and smart home technology with surveillance cameras [4]. Moreover, HAR can be used to monitor the activities of patients with conditions such as Parkinson's disease [5], Alzheimer's disease [6], and depression [7]. This can help healthcare providers to understand better and manage these conditions.

Previous research demonstrate that HAR is necessary for various reasons, including improving human-computer interaction, developing smart home technology, and enhancing security systems [8]. However, building a deep learning algorithm from scratch to detect human activity is a very difficult and requires a significant amount of labeled data and computational resources. Therefore, the solution to this challenge is to use the transfer learning method to reuse a pre-trained deep learning model that has been previously trained with many objects. However, more studies must be conducted to evaluate the merits of pre-trained deep learning models for HAR.

In the context of binary classification, two previous studies employed the convolutional neural network algorithm with the University of California Irvine HAR (UCI-HAR) dataset [9], [10], which consisted of 10,299 data points and yielded accuracies of 90 % and 91.63 %. Numerous pre-trained models have been extensively explored for human activity detection, achieving accuracy rates surpassing 90 %. These models include InceptionV2 [11], YOLOv3 [11], ResNet-152 [12], and visual geometric group with 16 layers (VGG-16) [11]–[13]. Among these models, VGG-16 has attracted significant attention as one of the most extensively researched pre-trained models. Developed by Zisserman and Simonyan from the University of Oxford [14], VGG-16 encompasses 16 layers and achieved victory in the ImageNet competition. Consequently, this study aims to assess the capabilities of

the VGG-16 algorithm.

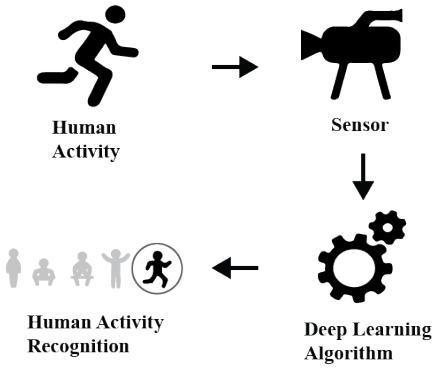

Fig. 1. Human activity recognition with deep learning.

The contribution to this research is divided into three, namely 1) Reporting the performance of the transfer learning method, which is believed to produce higher accuracy with a limited amount of data 2) Testing the VGG-16 pre-trained model, which has not been widely used to detect human activity, 3) Testing a different dataset that has a smaller amount of data compared to previous studies. The amount of human activity data examined in this study was only 1,680 public datasets in 160 $\times$ 160 pixel JPG format images. The dataset has been investigated using the semisupervised recurrent convolutional attention model [15] and the temporal information convolutional neural network [16]. In addition, 100 local data obtained from the internet will be used to test the model further.

II. Research Method

The research methodology employed in this study involved the following steps: data input, preprocessing, data spil, training, testing, and evaluation. First, a collection of human activities images was assembled and imported into the Jupyter Notebook program. The dataset used in this research was derived from two sources: a public dataset and a local dataset. Then, the data underwent preprocessing to ensure compatibility with the convolutional neural network (CNN) algorithm. The preprocessing steps included resizing the images to the standardized size of 224 $\times$ 224 pixels, conforming to the VGG-16 model requirements. Additionally, the dependent variable's data type was transformed to binary, with 0 representing "Static" and 1 representing "Dynamic." No further data augmentation or modifications were applied in this research.

After that, the human activities dataset underwent a data split process, dividing the data into three subsets: 10 % Test Data I, 10 % validation data, and 80 % training data. Furthermore, an additional set of 100 locally gathered samples was allocated as Test Data II. Then, the VGG-16 model was trained using the publicly available dataset, which was further divided into train data and validation data subsets. The trained VGG-16 model was subjected to testing using the designated test data. The model's performance was assessed by computing the accuracy metric. Moreover, a second round of testing was conducted using the local dataset, providing additional evaluation of the VGG-16 model's capabilities. Finally, the model's performance was evaluated through the application of a confusion matrix, which facilitated a comprehensive analysis of its classification performance.

A. Dataset

The dataset for this experiment comes from the Data Sprint 76 - Human Activity Recognition dataset shared by Karthick *et al.* on aiplanet.com, as seen in Fig. 2. The dataset consists of human activity images that are typically found on the internet and labeled with the type of activity in the image. Originally there were 18,002 data with 15 types of activity in this dataset. Still, this study will only use two types of activity, namely sitting and running, that can be called static and dynamic, respectively, to be more aligned with previous research. Moreover, the dataset size is 1,680 data.

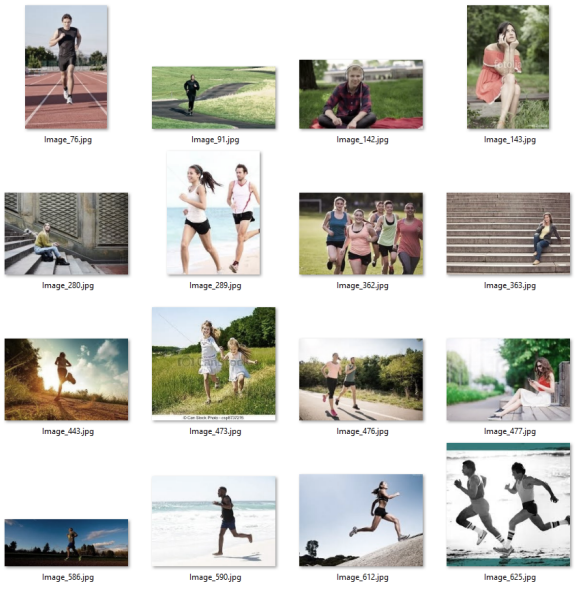

Fig. 2. Public HAR still images dataset.

In addition, there are also 100 local image datasets, as illustrated in Fig. 2. These local datasets are utilized in the testing process of this study and are further divided into 50 sitting data and 50 running data. The local dataset was acquired from the website https://images.google.com by conducting searches using the Indonesian keywords 'sitting' and 'running'. Once the data is successfully saved, it undergoes a preprocessing process where the image size is reduced to 250 pixels for either the length or width of the image (maintaining the original proportions) in order to optimize storage memory usage. It is important to note that the ideal image size for VGG-16 is 224 $\times$ 224 pixels. Hence, the dataset for this research retains a size

of 250 × 250 pixels without compromising the image quality. Furthermore, the local dataset has also been combined with the public data, which is accessible through the Google Drive site.

As shown in Fig. 3, the VGG-16 is the CNN architecture used to win the Imagenet competition 2014. This model is considered one of the best vision architectures [17]. Furthermore, it has been studied for HAR [18], [19] and shows reliable results.

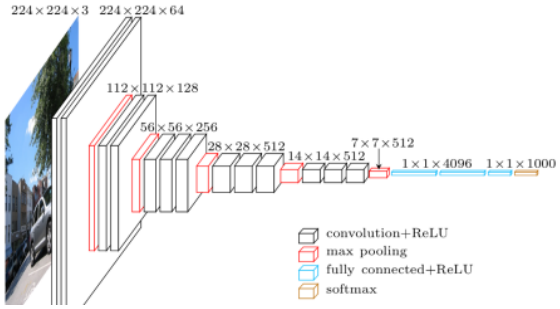

Fig. 3. VGG-16 architecture.

The process of VGG-16 for HAR can be seen in Fig. 5 from image input with a size of 224 × 224 pixels to a typical CNN process. The difference is VGG-16 has 16 layers and was pre-trained with many images before but not specifically HAR images. The VGG-16 model is chosen as the pre-trained model in this study because the VGG-16 model may shorten the training time of the CNN algorithm for classifying human activities and maintaining a good performance even with a limited dataset [8], [21]. The hyperparameters settings used for the VGG-16 model in this experiment can be seen at Table 1.

Table 1. Hyperparameter Setting.

Hyperparameter	Value
Image size	224
Epochs	20
Train Batch Size	77
Pooling	'avg'
Weights	'imagenet'
Optimizer	'adam'
Hidden layer neurons	512
Hidden layer activation	'relu'
Output activation	sigmoid

The schematic representation in Fig. 4 provides an overview of the sequential steps involved in the processing of input data by the VGG-16 model. Initially, the input image with dimensions of 224 × 224 pixels undergoes convolution and pooling operations. This process is repeated multiple times, gradually reducing the size of the matrix. Subsequently, the matrix is passed through dense layers, where the predicted output is obtained by converging through multiple softmax or sigmoid nodes.

B. Evaluation Metrics

The experiment conducted in this research uses the two metrics commonly used to observe the training and validation process, namely accuracy and loss [22]. The

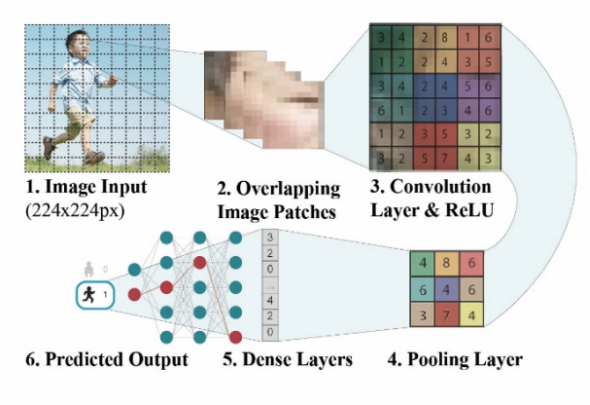

Fig. 4. VGG-16 model classification flow.

accuracy metric is also used for the model evaluation of the predicted testing data. Accuracy is the percentage of the correct prediction made. Meanwhile, loss is the distance between the true values of the problem and the values predicted by the model. A higher loss value can indicate more errors we made on the data.

This study uses binary classification, therefore there can be four possible outcomes between the actual values and predicted values which are the True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN), also commonly known as the Confusion Matrix [23] as shown in Fig. 5.

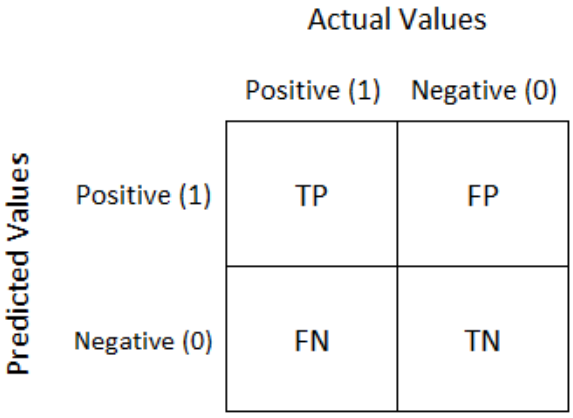

Fig. 5. Confusion matrix.

Therefore, the accuracy percentage can be calculated using (1).

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

III. Result

This study employed a data categorization scheme comprising train data, validation data, and test data, with respective proportions of 80 %, 10 %, and 10 %. The static and dynamic labeled images were balanced, accounting for approximately 49.4 % and 50.6 % of the dataset, respectively. Additionally, a local dataset was collected specifically for the purpose of further evaluating the proposed model, as detailed in Table 2.

Table 2. Data Split

Description	Count
Public Dataset (Train)	1,344
Public Dataset (Validation)	168
Public Dataset (Test I)	100
Local Dataset (Test II)	100

The transfer learning method was adopted in this study, utilizing a pre-trained model for human activity recognition. This approach facilitated a shorter training process and minimized the dataset requirements while achieving high accuracy.

1) Training Accuracy: The accuracy result from the training process is 98.9 %, as shown in Fig. 6.

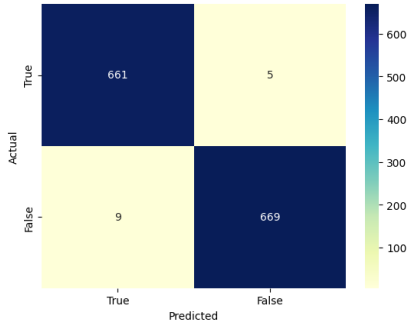

Fig. 6. Training performance.

2) Validation Accuracy: For the validation process, the accuracy is 99.4 %, as shown in Fig. 7.

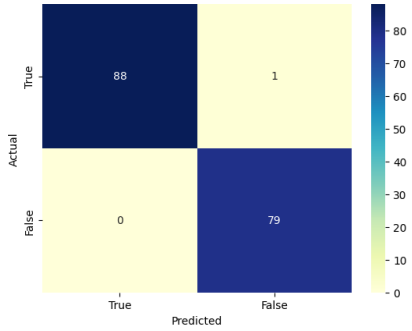

Fig. 7. Validation performance.

3) First Testing Result: The accuracy for the first test result with the same dataset source as training and validation yielded 98.8 %, as shown in Fig. 8.

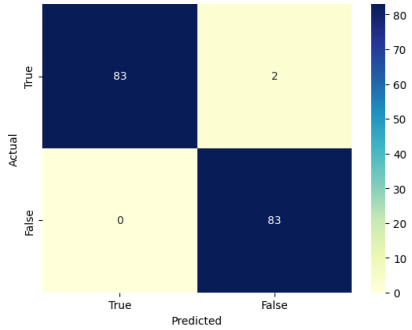

Fig. 8. Testing I result.

4) Second Testing Result: The accuracy for the second test result with a different dataset source, training and validation, yielded 97 %, as shown in Fig. 9.

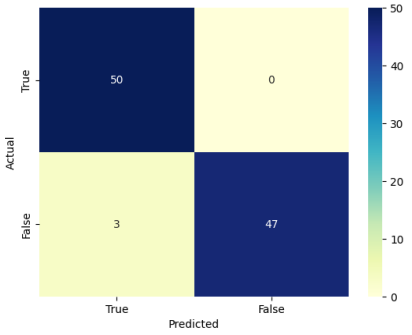

Fig. 9. Testing II result.

IV. Discussion

The existing literature on transfer learning and HAR has been extensively reviewed. However, in this study, we conducted empirical experiments using the VGG-16 model for HAR, focusing on a dataset comprising both dynamic and static human images. The results obtained from our research contribute to the ongoing evaluation of the effectiveness of VGG-16 for HAR.

One notable aspect observed in this research is the absence of overfitting issues when using the VGG-16 model for HAR, both with public and local datasets. The VGG-16 pre-trained model was developed to detect human activity, specifically classifying static and dynamic activities. Through the training process involving 1,344 training data, 168 validation data, 168 Test Data I, and 100 Test Data II obtained locally, our model exhibited a minimal difference in accuracy between the training and validation stages, with only a 0.005 % discrepancy. Furthermore, when comparing the results of test I with the public dataset and test II with the local dataset, the difference was merely 1.8 %. Consequently, our model demonstrated success in avoiding overfitting issues.

Moreover, our research achieved higher accuracy compared to previous studies. Past research efforts employed the CNN algorithm to predict static and dynamic activities in still images, achieving a maximum accuracy of 96.5 %. In contrast, our study involved different datasets, encompassing a total of 1,680 data points. By utilizing the VGG-16-based CNN model with transfer learning, we achieved superior performance on two test data sets. Specifically, the accuracy reached 98.8 % for the public test data and 97 % for the local test data. The improved accuracy in our study is attributed to the utilization of the VGG-16 transfer learning method rather than relying on the gait history image descriptor used in prior research [14].

V. Conclusion

This research has demonstrated the process of developing a model created from transfer learning techniques using the visual geometric group with 16 layer (VGG-16) model to detect human activity. Two types of human activity are examined in this study: static and dynamic. The training process with the training and validation shows a low difference in accuracy which is a good sign that indicates no overfitting. To further examine the overfitting of the Testing process, the model developed in this study was also tested with two kinds of datasets, one is a public dataset, and the other one is a local dataset gathered in this research. Both Test Data produced high accuracy of 97 % and 98.8 %, indicating no overfitting issue. This research has demonstrated a Transfer Learning technique with the VGG-16 model to detect static and dynamic human activities that can produce a good accuracy and low overfitting issue.

References

[1] K. Hirooka, M. A. M. Hasan, J. Shin, and A. Y. Srizon, "Ensembled transfer learning based multichannel attention networks for human activity recognition in still images," *IEEE Access*, vol. 10, pp. 47051–47062, 2022.

[2] Y. Wang, S. Cang, and H. Yu, "A survey on wearable sensor modality centred human activity recognition in health care," *Expert Syst Appl*, vol. 137, pp. 167–190, 2019.

[3] S. Zhang et al., "Deep learning in human activity recognition with wearable sensors: A review on advances," *Sensors*, vol. 22, no. 4, p. 1476, 2022.

[4] K. Kim, A. Jalal, and M. Mahmood, "Vision-based human activity recognition system using depth silhouettes: A smart home system for monitoring the residents," *Journal of Electrical Engineering & Technology*, vol. 14, pp. 2567–2573, 2019.

[5] G. Sarapata, G. Morinan, Y. Dushin, B. Kainz, J. Ong, and J. O'Keeffe, "Video-based activity recognition for automated motor assessment of Parkinson's disease," *TechRxiv*, preprint, 2022. [Online]. Available: https://doi.org/10.36227/techrxiv.21610251.v1.

[6] Y. Asim, M. A. Azam, M. Ehatisham-ul-Haq, U. Naeem, and A. Khalid, "Context-aware human activity recognition (CAHAR) in-the-Wild using smartphone accelerometer," *IEEE Sens J*, vol. 20, no. 8, pp. 4361–4371, 2020.

[7] A. K. M. Masum, E. H. Bahadur, and F. A. Ruhi, "Scrutiny of mental depression through smartphone sensors using machine learning approaches," *International Journal of Innovative Computing*, vol. 10, no. 1, 2020.

[8] L. M. Dang, K. Min, H. Wang, M. J. Piran, C. H. Lee, and H. Moon, "Sensor-based and vision-based human activity recognition: A comprehensive survey," *Pattern Recognit*, vol. 108, p. 107561, 2020.

[9] S. Zebhi, S. M. T. AlModarresi, and V. Abootalebi, "Transfer learning based method for human activity recognition," in *2021 29th Iranian Conference on Electrical Engineering (ICEE)*, pp. 761–765, 2021.

[10] A. Ferrari, D. Micucci, M. Mobilio, and P. Napoletano, "On the personalization of classification models for human activity recognition," *IEEE Access*, vol. 8, pp. 32066–32079, 2020.

[11] T. Mustafa, S. Dhavale, and M. M. Kuber, "Performance analysis of inception-v2 and Yolov3-based human activity recognition in videos," *SN Comput Sci*, vol. 1, pp. 1–7, 2020.

[12] S. Khan, M. A. Khan, M. Alhaisoni, U. Tariq, H. S. Yong, A. Armghan, and F. Alenezi, "Human action recognition: a paradigm of best deep learning features selection and serial based extended fusion," *Sensors*, vol. 21, no. 23, pp. 7941, 2021.

[13] L. Wang, Y. Xu, J. Cheng, H. Xia, J. Yin, and J. Wu, "Human action recognition by learning spatio-temporal features with deep neural networks," *IEEE access*, vol. 6, pp. 17913–17922, 2018.

[14] A. Ullah, K. Muhammad, I. U. Haq, and S. W. Baik, "Action recognition using optimized deep autoencoder and CNN for surveillance data streams of non-stationary environments," *Future Generation Computer Systems*, vol. 96, pp. 386–397, 2019.

[15] A. Banerjee, S. Roy, R. Kundu, P. K. Singh, V. Bhateja, and R. Sarkar, "An ensemble approach for still image-based human action recognition," *Neural Comput Appl*, vol. 34, no. 21, pp. 19269–19282, 2022.

[16] M. Safaei and H. Foroosh, "Still image action recognition by predicting spatial-temporal pixel evolution," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 111–120, 2019.

[17] M. E. Agus, S. Y. Bagas, M. Yuda, N. A. Hanung, and Z. Ibrahim, "Convolutional neural network featuring VGG-16 model for glioma classification," *JOIV: International Journal on Informatics Visualization*, vol. 6, no. 3, pp. 660–666, 2022.

[18] S. Deep and X. Zheng, "Leveraging CNN and transfer learning for vision-based human activity recognition," in *2019 29th International Telecommunication Networks and Applications Conference (ITNAC)*, pp. 1–4, 2019.

[19] L. Wang, Y. Xu, J. Cheng, H. Xia, J. Yin, and J. Wu, "Human action recognition by learning spatio-temporal features with deep neural networks," *IEEE access*, vol. 6, pp. 17913–17922, 2018.

[20] S. Zebhi, S. M. T. AlModarresi, and V. Abootalebi, "Transfer learning based method for human activity recognition," in *2021 29th Iranian Conference on Electrical Engineering (ICEE)*, pp. 761–765, 2021.

[21] S. Liu and W. Deng, "Very deep convolutional neural network based image classification using small training sample size," in *2015 3rd IAPR Asian conference on pattern recognition (ACPR)*, pp. 730–734, 2015.

[22] J. Rawat, D. Logofătu, and S. Chiramel, "Factors affecting accuracy of convolutional neural network using VGG-16," in *International Conference on Engineering Applications of Neural Networks*, pp. 251–260, 2020.

[23] A. Bagaskara and M. Suryanegara, "Evaluation of VGG-16 and VGG-19 deep learning architecture for classifying dementia people," in *2021 4th International Conference of Computer and Informatics Engineering (IC2IE)*, pp. 1–4, 2021.