Recognition of Real-Time Angklung Kodály Hand Gesture using Mediapipe and Machine Learning Method

2023 International Conference on Computer Science, Information Technology and Engineering (ICCoSITE) | 979-8-3503-2095-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICCoSITE57641.2023.10127808

Muhammad Zulfikar Fauzi
Department of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
mzulfikarfauzi09@gmail.com

Riyanarto Sarno
Department of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
riyanarto@if.its.ac.id

***Abstract*— The application of machine learning in the arts and culture can help preserve these arts and culture and, in some cases even facilitate access to the arts to the public. In this study, a hand gesture recognition system was developed for the Kodaly method hand gesture to play angklung directly from a personal computer. Many methods of hand gesture recognition still need an additional hardware such as gloves and expensive sensors that may not be accessible to some people. This study proposes a computer vision-based approach utilizing the personal computer internal webcam without additional hardware and employs multi-layer perceptron classifier for real-time recognition and Mediapipe for feature extraction. The developed system achieves an average accuracy of 99% and runs with low latency between recognition and audio output. The effective range of hand gesture recognition is up to 1.5 meters in environment with a light level as low as 25 LUX.**

Keywords—Multi-layer Perceptron, Hand gesture recognition, Computer vision, Angklung

I. Introduction

Hand gesture is one form of communication that uses the arms and hands to create meaning. Recently, human gesture recognition has been one of the trending topics due to the development of machine learning technology [1]. One of the uses of hand gestures is a musical hand gesture, where a conductor signals the performer to play their instruments based on the gesture signed by them.

Saung Angklung Udjo is a tourist attraction in Bandung, Indonesia. Their most well-known program is a live session of practicing and playing angklung together, led by a conductor using a hand gesture system; this session is called interactive angklung [2]. Angklung is a musical instrument from the Sundanese people in Indonesia made of various bamboo tubes attached to a bamboo frame.

Kodály's method, a musical hand signs system that evolved organically in music schools in Hungary under Kodály's instruction and guidance, is used because of its interactivity and engagement [3]. There are a few modifications to the original, which can be seen from the 'fa' and 'si' sign. The 'fa' sign was changed from thumb down while the 'si' sign faced backward. The signs used in Saung Angklung Udjo are shown in Fig. 1.

Chin-shyurng, et al. [4] used a Kinect optical sensor to perform vision-based sensing in musical gesture recognition in real-time application with an average accuracy of 89.17% in 30 frames per second video format. A single-depth camera is used to capture the input and for the gesture recognition system itself. A Kinect sensor is still needed to perform the method proposed.

This study developed a computer vision-based hand gesture recognition system to recognize the hand gestures used in Saung Angklung Udjo. The method used in this system utilizes a webcam that is commonly used by the public; this allows for this system to be widely used by people who have computers or laptops. Moreover, in this study, a program was also developed to play the sound of angklung utilizing the computer's audio device, according to hand gestures given in real time by the user. By using the Mediapipe framework for feature extraction, programs can run using relatively light computing power [5].

Fig. 1. Hand signs used in saung angklung udjo

II. Literature Reviews

Yun, et al. [11] proposed hand gesture recognition using a device-based approach. Multi-layer Perceptron is applied in the study producing a high recognition result at 99.8% with 700 epochs of training. As the purpose of the study is to develop a device for teaching purposes, using a device-based approach is more inconvenient because of the necessity to make the device and maintain them. This study will address these setbacks.

E. M. Budi, et al. [2] developed a hand gesture recognition for Kodály's method using a vision-based approach. The study uses a Kinect sensor to capture human movement using two infrared and digital cameras. The front-capturing method shows higher accuracy than side-capturing, but the problem of differentiating between palm and hands arises since both overlaps. Kinect sensors address this problem with their innate depth information from the sensor embedded within. In this study, the issue with information extraction will be addressed using the proposed method while only utilizing a daily-use webcam.

III. Proposed Method

In this study, our proposed method will be divided into three main parts. Dataset creation for the Kodály method hand gesture will be described in the data acquisition and pre-processing section. The model developed for the Kodály method gesture recognition will be described in the model training and evaluation section. The system development to simulate a real angklung will be described in the system development and system testing section. The diagram for the proposed method can be seen in Fig. 2.

A. Data Acquisition

Data acquisition describes how the raw data used in this research is produced. An internal laptop webcam with a resolution of 1280 x 720 pixels is used for image capturing. OpenCV and python are utilized to automate the data acquisition process.

A total of 1800 images will be captured using this method which consists of 200 images for each class, which is the notes starting from do until high do, and an additional idle state class. These raw image data are saved in a directory and will be pre-processed in the data pre-processing section.

B. Pre-processing

Pre-processing describes how the raw image data acquired will be processed before classification is carried out. The labeled numerical dataset in Comma Separated Values (CSV) format is the output of this pre-processing. Pre-processing steps will also be done in the hand gesture recognition without the data labeling step.

Image color format is converted from Blue Green Red (BGR) to Red Green Blue (RGB); this process is done because Mediapipe works with RGB color format; meanwhile, the raw images acquired are in BGR. The RGB images are processed using Mediapipe holistic solution for feature extraction to get human landmark key points. This process will return a series of float numbers representing the coordinates in the x, y, and z-axis and the visibility of each key point from the camera perspective. Mediapipe is a machine-learning solution used to extract key points of humans from a given image or video [6].

The results from Mediapipe will be converted into a one-dimension array. Feature selection is carried out manually. Kodály method musical gesture only utilizes the right hand; hence, only the right hand's key points and the human posture from half to top are kept for the dataset features.

Since the image data is already separated by directory, labeling is done according to each image's respective directory. In total, the dataset has nine classes.

C. Model Training

Model training describes the method used to train the model using the machine learning technique. Multi-layer Perceptron (MLP), a fully connected class of feedforward artificial neural network, will be used as the algorithm for classification. MLP is chosen because of its reliability in its learning process [8]. The parameters are tuned to find the best possible combination of values.

To optimizes the parameters used in the model training, the grid searching technique is used to find the best combination of values for each parameter. This step increases the model's predictive accuracy and reduces the chance of producing suboptimal results [7]. Max iteration is set to 200 in this tuning process to save the time and computing power needed. Cross-validation is also implemented to get a more accurate result of accuracy [12]. The parameters tuned are as follows:

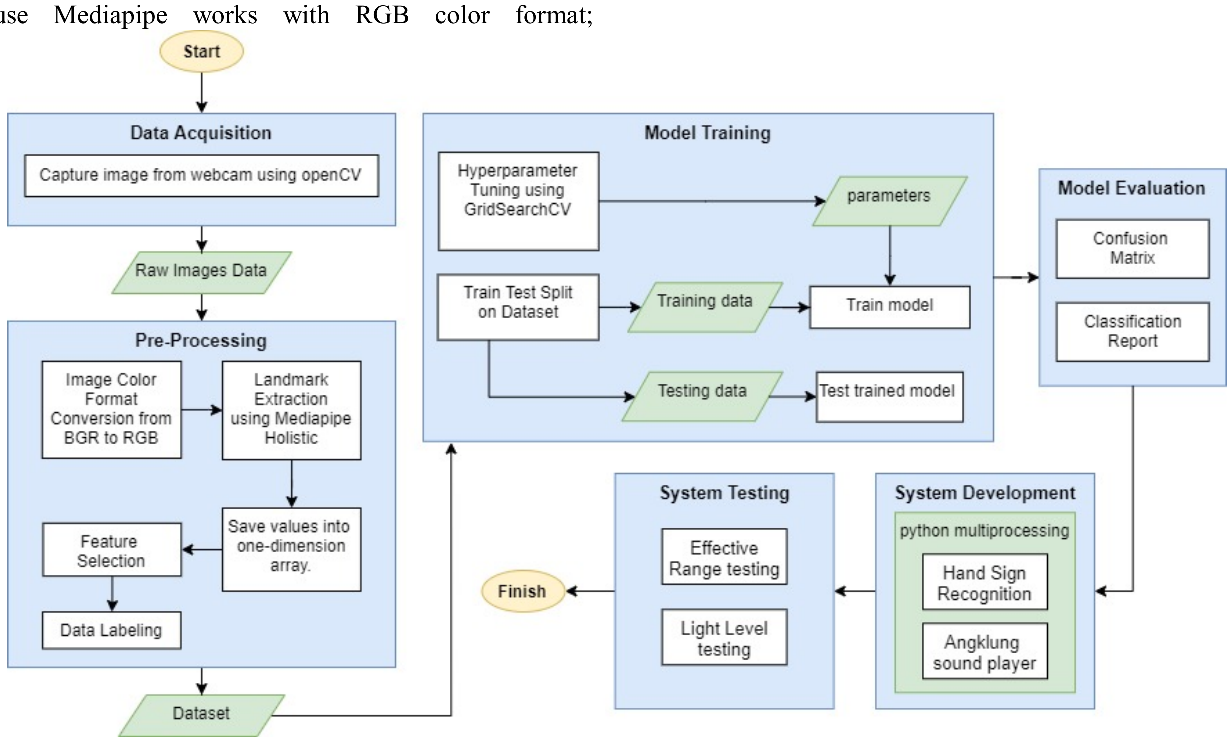

Fig. 2. Proposed method

Fig. 3. Image annotated with keypoints

- hidden_layer_sizes

This parameter determines the number of neurons in the hidden layer and the number of hidden layers themselves.

- learning_rate

This parameter controls the algorithm's learning rate method for weight updates.

- learning_rate_init

This parameter sets the value of the initial learning rate used.

The optimized parameter will be used in the model training. The dataset will be split into training and testing data with a ratio of 6:4, resulting in 1080 data for training and 720 data for testing.

D. Model Evaluation

The confusion matrix and classification report will be used in this study to evaluate the model performance. By using both methods, the model's predictive performance in each class will be visible and can be analyzed to improve further the model's performance [9-10].

E. System Development

System development describes the environment, hardware, and software used in the e-angklung system development and the underlying methods and algorithm used in the system. The system uses python as the programming language with the help of anaconda and visual studio code for the environment and Integrated Development Environment (IDE). The hardware specifications used in this research are the following:

- Processor: AMD Ryzen 5 Mobile 3500U 2.10 GHz
- RAM: 8 GB DDR4 2400MHz
- GPU: Integrated Radeon Vega 8 Mobile Gfx 2GB Integrated Memory
- Webcam: 720p HD Internal Laptop Webcam
- Operating System: Windows 10 Home 64bit
- Audio Device: SonicMaster

The system mainly comprises two components: hand sign recognition and the angklung sound player. The hand sign recognition component will handle all the processes necessary for the system to detect and recognize the hand sign that the user is doing; this includes the camera accessing, pre-processing, etc. The angklung sound player component handles the system's process to produce audio smoothly; in this case, it is the sound of each note of angklung.

Python multiprocessing will be utilized in the system development to reduce delay and make the system perform better. Two processes will be used, one for each of the system components.

F. System Testing

System testing describes how the system performance will be measured. Since the system is computer vision-based, its effective range of usage and ability to perform with dim lighting will be tested. The effective range of usage testing will be carried out using the following scenario:

- The user will perform Kodály's method "mi" hand gesture.
- The user will be placed at 1 meter, 1.5 meters, and 1.8 meters in front of the webcam.

System's ability to recognize the hand gesture testing will be carried out using the following scenario:

- Two light levels will be used for each scenario in the effective range of usage testing: bright and dim. Dim lighting has a light level of 25 Lux on average, and bright lighting has a light level of 90 Lux on average.

IV. Results and Discussion

A. Dataset

The raw image data acquired from the data acquisition part of this are a collection of images captures the upper half of the user doing the Kodály hand gesture. The format of each image is then converted to RGB to be processed by Mediapipe for human key points annotation. Fig. 3 shows the original image annotated with key points. In this study, Mediapipe used the following parameters:

- static_image_mode = True
- model_complexity = 2
- min_detection_confidence = 0.6

static_image_mode was set to True because the raw data are images, and model_complexity was set to 2 to increase the complexity of the landmark key points model to get better accuracy in value.

The red dots, which indicate the key points, are extracted in the form of numerical values describing the positions of each dot relative to the camera. The results of Mediapipe processing include all the visible and non-visible vital points. Feature selection was done to remove the unnecessary key points leaving only the key points representing the right hand and the human posture visible in the frame.

The final dataset has the following specifications:

- One hundred thirty-one features of each key point's x, y, and z coordinates and an additional visibility value for the human pose key points (excluding the key points in the human hand).
- One feature for the label.

B. Model Evaluation

Hyperparameter tuning was done using a grid searching technique and cross-validation of 5-folds. In this study, hidden layers of the MLP were set to 3 layers, with each layer consisting of neurons ranging from eight to twelve. The results of hyperparameter tuning can be seen in Table 1.

The results show that 12 neurons in each layer with adaptive learning rate and 0.001 initial value produce the best average test score accuracy of 0.985.

TABLE I. RESULTS OF HYPERPARAMETER TUNING

No.	hidden_layer_sizes	learning_rate	learning_rate_init	mean_test_score
1	12,12,12	adaptive	0.001	0.985556
2	12,12,12	constant	0.001	0.985556
3	10,8,10	constant	0.001	0.983333
4	10,8,10	adaptive	0.001	0.983333
5	12,12,12	adaptive	0.010	0.980556

Both model training and testing used the tuned parameters and 300 epochs of training. The model training results can be seen in Fig. 4 for the confusion matrix and Table 2 for the classification report.

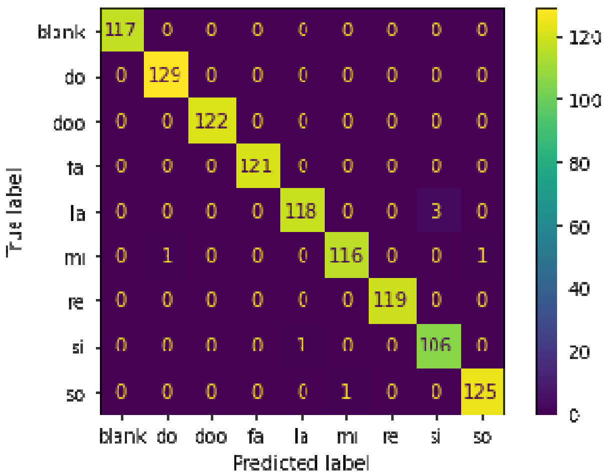

Fig. 4. Model training confusion matrix

Table 2 shows that the training accuracy ranges from 0.97 to 1.00, with the lowest when predicting the "si" class. The average accuracy across the board is 99%.

TABLE II. MODEL TRAINING CLASSIFICATION REPORT

	precision	recall	F1-score	support
blank	1.00	1.00	1.00	117
do	0.99	1.00	1.00	129
doo	1.00	1.00	1.00	122
fa	1.00	1.00	1.00	121
la	0.99	0.98	0.98	121
mi	0.99	0.98	0.99	118
re	1.00	1.00	1.00	119
si	0.97	0.99	0.98	107
so	0.99	0.99	0.99	126
accuracy	-	-	0.99	1080
macro avg	0.99	0.99	0.99	1080
weighted avg	0.99	0.99	0.99	1080

Fig. 5 and Table 3 show the result of model testing. The accuracy decreased from the training, albeit not much, and the average accuracy is still 99%. The class "la" has the lowest precision score.

C. System Evaluation

Fig. 6, Fig. 7, and Fig. 8 show the testing result of the class "mi" variations in range with bright lighting. The model successfully predicted the hand gesture in all ranges in this testing.

TABLE III. MODEL TESTING CLASSIFICATION REPORT

	precision	recall	F1-score	support
blank	1.00	1.00	1.00	83
do	0.99	1.00	0.99	71
doo	0.99	1.00	0.99	78
fa	0.99	0.99	0.99	79
la	0.96	1.00	0.98	79
mi	0.99	0.96	0.98	82
re	1.00	1.00	1.00	81
si	1.00	0.96	0.98	93
so	0.97	0.99	0.98	74
accuracy	-	-	0.99	720
macro avg	0.99	0.99	0.99	720
weighted avg	0.99	0.99	0.99	720

Fig. 9, Fig. 10, and Fig. 11 show the testing result of the class "mi" variations in range with dim lighting. The model failed to predict the hand gesture in the 1.8-meter test.

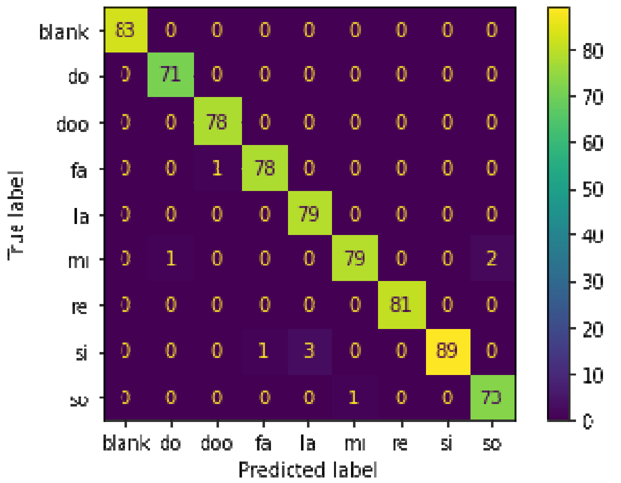

Fig. 5. Model testing confusion matrix

D. Discussion

A few problems arise when system testing is done. Playing music requires continuous gesture to be signed, and this causes the system to pick up unnecessary detection in between signs, for example when the user changes from "do" to "si", the system sometimes return a false prediction if the user takes too long to transition from a class to class.

The system runs well up to 1.5 meters range with sufficient lighting. In a 1.8-meter range and dim lighting, the model fails to predict the class "mi" and mistook it for "do". The model's ability to predict accurately goes down as the user goes further away from the webcam.

Fig. 6. "mi" testing at 1-meter range and bright lighting

Fig. 7. "mi" testing at 1.5-meter range and bright lighting

Fig. 8. "mi" testing at 1.8-meter range and bright lighting

Fig. 9. "mi" testing at a 1-meter range and dim lighting

Fig. 10. "mi" testing at 1.5-meter range and dim lighting

Fig. 11. "mi" testing at 1.8-meter range and dim lighting

While testing, the system runs at 10 frames per second and barely has any noticeable delay between recognition and producing the audio output; this is because both recognition and audio systems run in parallel using multiprocessing technology.

The note produced by the system will change in quick succession according to the sign that the user was giving, so the duration of the note is the same as how long the user holds their sign for.

V. Conclusion

This study aims to create a system for Kodály hand gesture recognition with audio as the output. From the experiments conducted, the following results were obtained:

- Using the Multi-Layer Perceptron algorithm, the model achieved an average accuracy of 99% in hand gesture recognition. This was proven in the model evaluation section.
- The system can run in 10 frames per second without a noticeable delay between the hand recognition and audio output.
- The system's effective range of use is from one meter up to about one and a half meters with sufficient lighting (90 Lux). This was proven by the testing conducted in this study.
- Utilizing a camera with better specifications may improve the system's ability to recognize hand gestures at further distances; this can be implemented in future works.

Acknowledgment

This research was funded by Lembaga Pengelola Dana Pendidikan (LPDP) under Riset Inovatif-Produktif (RISPRO) Program; the Indonesian Ministry of Education and Culture under Penelitian Terapan Unggulan Perguruan Tinggi (PTUPT) Program; and Institut Teknologi Sepuluh Nopember (ITS) under project scheme of the Publication Writing and IPR Incentive Program (PPHKI).

References

[1] M adni, M., and C. Vijaya, “Hand gesture recognition using semi vectorial multilevel segmentation method with improved ReliefF algorithm”, Int. J. Intell. Eng. Syst 14: 447-45, 2021.

[2] E. M. Budi, A. A. Rochim, H. K. Dipojono, A. Handojo, and J. Sarwono, “Musical gesture recognition for interactive angklung robot”, In: 3rd International Conference on Instrumentation Control and Automation (ICA), pp. 149-154, 2013.

[3] Howard, P. M., "Kodály Strategies for Instrumental Teachers: When the Kodály method is extended to string and other instrumental instruction, it offers opportunities for enhancing student musicianship", Music Educators Journal, 82(5), 27–33, 1996.

[4] Chin-Shyurng, Fahn, Shih-En Lee, and Meng-Luen Wu, "Real-time musical conducting gesture recognition based on a dynamic time warping classifier using a single-depth camera", Applied Sciences 9, no. 3: 528, 2019.

[5] Fan Zhang, Valentin Bazarevsky, Andret Vakunov, Andrei Tkachenka, George Sung, Chang, Chuo-Ling, Matthias Grundmann. "MediaPipe Hands: On-device Real-time Hand Tracking", 2020.

[6] Ghanbari, Sara, Zahra Parvin Ashtyani, and Mehdi Tale Masouleh, "User Identification Based on Hand Geometrical Biometrics Using Media-Pipe", In: 30th International Conference on Electrical Engineering (ICEE). IEEE, 2022.

[7] Anggoro, Dimas Aryo, and Salsa Sasmita Mukti, "Performance Comparison of Grid Search and Random Search Methods for Hyperparameter Tuning in Extreme Gradient Boosting Algorithm to Predict Chronic Kidney Failure", Int. J. Intell. Eng. Syst 14, no. 6: 198-207, 2021.

[8] Mhd Hasan, Riyanarto Sarno, Shoffi Sabila, "Optimizing Machine Learning Parameters for Classifying the Sweetness of Pineapple Aroma Using Electronic Nose", International Journal of Intelligent Engineering and Systems, 13. 122-132, 2020.

[9] Agus Tri Haryono, Riyanarto Sarno, Rachmad Abdullah, "Aspect-Based Sentiment Analysis of Financial Headlines and Microblogs Using Semantic Similarity and Bidirectional Long Short-Term Memory", International Journal of Intelligent Engineering and Systems, Vol.15, No.3, pp.233-241, 2022.

[10] Dedy Rahman Wijaya, Riyanarto Sarno, Enny Zulaika, "Noise filtering framework for electronic nose signals: An application for beef quality monitoring", Computers and Electronics in Agriculture, Volume 157, Pages 305-321, ISSN 0168-1699, 2019.

[11] Yun, Young Bae, Young Man Kwon, Myung Jae Lim, Jeong Jun Park, and Dong Kun Chung, "Development of Kodály Hand Sign Educational Tools using Multi-Layer Perceptron", Turkish Journal of Computer and Mathematics Education (TURCOMAT) 12, no. 13: 6169-6174, 2021.

[12] Riyanarto Sarno, Dedy Wijaya, Muhammad Mahardika, "Music fingerprinting based on bhattacharyya distance for song and cover song recognition", International Journal of Electrical and Computer Engineering. 9. 1036-1044, 2019.