Development of a Hand Gesture Detection-Based Robot System with MediaPipe

[1*]Selamat Muslimin, [2]Ekawati Prihatini, [3]Tri Martin
[1,2]Departement of Electrical Engineering, Politeknik Negeri Sriwijaya, Indonesia
Email: [1]selamet_muslimin@polsri.ac.id, [2]ekawati_p@polsri.ac.id, [3]trimartin488@gmail.com

Article Info

Article history:
Received Jul 02nd, 2025
Revised Oct 06th, 2025
Accepted Nov 27th, 2025

Keyword:
Gesture Detection
Hand Waving
Intelligent Robot
MediaPipe
Raspberry Pi

ABSTRACT

This research presents the development of an intelligent robot that can be summoned simply by waving a hand, without the need for physical buttons or voice commands. The system utilizes *MediaPipe* technology to detect and recognize hand gestures in real time through a camera. When a user waves their hand toward the camera, the system processes the motion and identifies it as a signal to call the robot. Image processing is handled by a Raspberry Pi, while movement control is managed by an Arduino, which regulates the direction and speed of the motors. The robot automatically moves toward the user and stops at a certain point to wait for further confirmation. Test results show that the robot can accurately detect gestures under various lighting conditions and distances. This approach enables more natural and efficient human–robot interaction, making it well-suited for modern contactless service systems.

Corresponding Author:
Selamat Muslimin
Departement of Electrical Engineering, Politeknik Negeri Sriwijaya
Srijaya Negara St.Bukit Lama, Ilir Barat 1 District, Palembang City 30128, South Sumatra, Indonesia
Email: selamet_muslimin@polsri.ac.id

DOI: http://dx.doi.org/10.24014/ijaidm.v8i3.37678

1. INTRODUCTION

With the advancement of technology, the field of robotics has also experienced significant developments, particularly in computer science and control systems. These advancements have made a major contribution to the design and development of robots, both manually operated and automated [1]. Innovations in robotics technology have led to various applications of self-service robots, which are now being increasingly used across multiple sectors, including restaurants, healthcare facilities, shopping centers, and educational institutions. However, the development of self-service robots still faces challenges, particularly in creating more natural and intuitive interactions between humans and robots. One of the main obstacles lies in the process of summoning and controlling robots by users. To address this challenge, a smarter and more responsive interaction approach is needed, such as through the use of hand gesture recognition technology using MediaPipe [2].

Jochen Wirtz, a professor at the National University of Singapore who is known as an expert in technology-based services, stated that service robots equipped with artificial intelligence and motion recognition systems have the potential to revolutionize interactions between humans and machines, making them more natural and efficient [3]. On the other hand, MediaPipe technology also contributes to improving the autonomous navigation capabilities of robots. Dewangga et al. (2024) demonstrate that this technology can be utilized to control various devices, including smart home devices and robots, in a more efficient and responsive manner. By utilizing motion recognition systems, robots can autonomously navigate toward users who require assistance, facilitating timely and error-free service delivery. Additionally, the integration of navigation sensors enables robots to move stably and accurately, even in complex and obstacle-filled environments [4].

In July 2019, Google released MediaPipe as an open-source platform designed to support real-time visual data processing. MediaPipe uses the Single Shot Detection (SSD) method to detect palms as part of the object identification process. Once the palm is detected, a hand landmark model is used to extract important information to recognize the position and movement of the fingers. In general, MediaPipe is a machine learning-based algorithm designed to detect and track human hands in live video streams. This tracking capability can be utilized by anyone with a camera device, such as a webcam. This technology enables the system to identify up to 21 three-dimensional (3D) coordinate points from a single hand image. Simultaneous and real-time tracking of multiple hands is possible through a three-step approach: palm detection, hand landmark point extraction, and hand movement recognition [5].

The hand detection process aims to identify the presence of hands in images or videos, while hand tracking focuses on monitoring the position and movement of hands from one frame to the next. This technology plays a crucial role in various applications, such as gesture recognition and human-computer interaction. MediaPipe provides a hand detection model trained using a large dataset of hand images, enabling high-precision detection and tracking processes. Based on previous studies, hand detection has great potential for application in various fields, such as gesture-based control systems, virtual reality (VR), and augmented reality (AR) [6]. In the development of the hand detection system, MediaPipe and OpenCV were integrated to achieve efficient and accurate tracking performance. MediaPipe is responsible for identifying and tracking key points (landmarks) on the hand, while OpenCV is used to capture and process video data directly from the webcam. The final result is visualized in the form of a display of hand landmark points on the screen. This system demonstrates optimal performance in real-time testing, characterized by a high frame rate (FPS) and good detection accuracy [7].

2. RESEARCH METHOD

This research began with the use of a digital camera as the main visual sensor to capture images in real time. The visual data obtained was then processed using two main software programs, OpenCV and MediaPipe. OpenCV is used for image acquisition and processing, while MediaPipe is responsible for extracting body poses and performing real-time image segmentation. During the body pose tracking stage, MediaPipe detects and tracks keypoints on the human body. These points represent important body parts, such as the shoulders, elbows, wrists, hips, knees, and ankles. MediaPipe then connects these points to form a body skeleton that depicts the user's posture or pose in real-time [8].

Body pose tracking in this system utilizes the concept of Human Pose Estimation, which is an important field in computer vision that focuses on the process of identifying and analyzing human posture and movement [9]. In this study, the process of detecting and tracking body poses is carried out using MediaPipe Pose, a machine learning-based framework developed by Google [10]. MediaPipe Pose offers a ready-to-use solution that enables real-time detection of human body poses through images or videos.

Body pose can be defined as the arrangement of human joints in a specific configuration. Therefore, the problem in Human Pose Estimation refers to the process of localizing or identifying the positions of human body joints, which are referred to as landmarks. In the context of images and videos, pose estimation encompasses various types, such as body, face, and hand poses, which are important components in the field of computer vision. Pose estimation from video data plays a significant role in various applications, including physical activity monitoring, sign language recognition, and overall body movement-based control [8]. Landmark models are used as visual markers to indicate the main locations of human body joints. In this system, input images are processed using the MediaPipe library to detect the user's key body points. The output of this process consists of three-dimensional coordinates (X, Y, and Z) for 33 key points on the human body. These coordinates represent the position of each detected body part in the input image. The output from MediaPipe contains information in the form of a list of coordinates for the user's key body points in the image [11].

MediaPipe is a framework designed to build machine learning pipelines for processing time-series data, such as video, audio, and other types of sequential data. The platform is cross-operating system and can be run on desktop/server devices, Android, iOS, and embedded devices such as Raspberry Pi and similar devices. In July 2019, Google released MediaPipe as open source software. One of MediaPipe's key capabilities is hand detection using the SSD approach, which is an object detection technique. After the detection process is completed, a hand landmark model is used to extract the necessary information to recognize finger positions. MediaPipe was developed using machine learning technology and is designed to detect the presence of individual hands in real-time video streams. This system enables real-time hand tracking for anyone using a webcam. Using machine learning-based techniques, the system can identify up to 21 three-dimensional (3D) coordinate points from a single hand image. Real-time hand tracking, whether single or dual, is enabled through a three-step approach that includes hand palm detection, hand landmark identification, and hand movement recognition [12].

MediaPipe Pose is an open-source cross-platform framework developed by Google that estimates two-dimensional (2D) human joint coordinates in each image frame. MediaPipe Pose builds a video-based cognitive data processing pipeline using machine learning technology [13]. This system can extract up to 33 2D landmark points on the human body, as shown in Figure 1. One of the architectures used in this framework is BlazePose, a lightweight machine learning model designed to deliver real-time performance on both mobile devices and personal computers by performing inference via the CPU. The pose estimation process is performed using normalized coordinates. Among all the landmarks generated, the landmark with index number 16 is specifically used in this study to detect poses and movements such as hand gestures. This movement is recognized when the right wrist is above the eyebrow line, indicating the occurrence of a hand gesture signal.

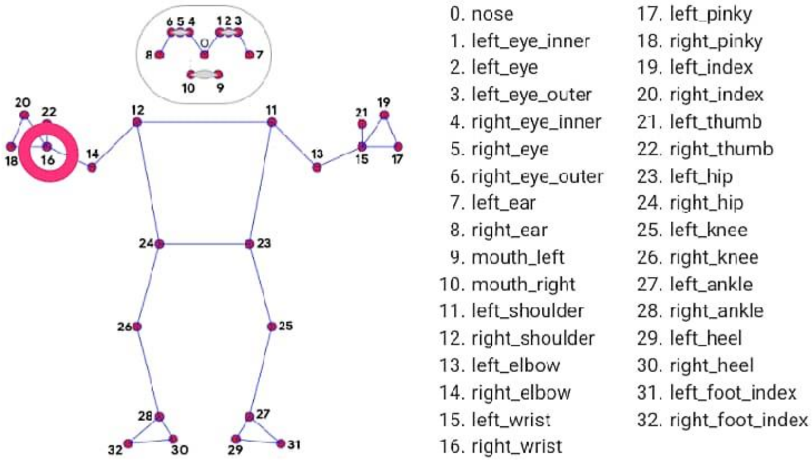
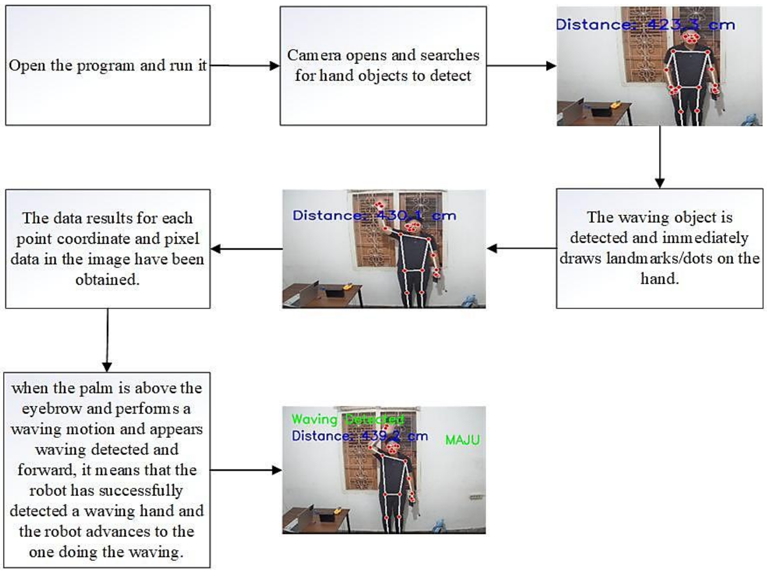


Figure 1. Definition of landmarks in MediaPipe Pose

Human pose estimation technology is currently a rapidly growing research topic around the world, with widespread applications in various fields such as sports, surveillance systems, work activity monitoring, home care for the elderly, independent physical training, entertainment, motion-based control, and implementation in robotic systems. In general, human pose estimation methods can be classified into several categories, namely two-dimensional (2D) and three-dimensional (3D) coordinate estimation, single-subject and multi-subject approaches depending on the number of subjects being observed, monocular and multi-view image-based approaches based on the number of cameras used, and single-image and video data-based approaches depending on the type of input used [14], [15], [16], [17], [18].

To perform real-time hand gesture recognition and determine the range of hand gestures that can be detected by MediaPipe. This is intended to determine the best performance of the MediaPipe system. The performance referred to is MediaPipe's ability to estimate the accuracy of the distance between the webcam and the user. The hand detection process from MediaPipe can be seen in Figure 2.



Open the program and run it

Camera opens and searches for hand objects to detect

The waving object is detected and immediately draws landmarks/dots on the hand.

The data results for each point coordinate and pixel data in the image have been obtained.

when the palm is above the eyebrow and performs a waving motion and appears waving detected and forward, it means that the robot has successfully detected a waving hand and the robot advances to the one doing the waving.

Figure 2. Flowchart of the hand wave gesture detection system

The process of recognizing hand gestures begins with activating the camera device, in this case using a webcam. Once the camera is active, the system begins receiving input in the form of real-time hand images. Hand object detection is performed using the VideoCapture command, which captures images and recognizes hand features with the help of color processing. When the color parameters match the characteristics of the hand, the system processes the data to generate hand landmark visualizations using MediaPipe. This visualization is displayed as a series of green lines connecting the red points on the hand within a single image frame. The formation of these hand landmarks indicates that the system has successfully detected the hand gesture object. Subsequently, the coordinates of each detected point on the hand are stored by the program for analysis or further processing.

MediaPipe is closely related to OpenCV because both support features relevant to computer vision, such as face detection and object detection. The collaboration between MediaPipe and OpenCV enables efficient video data processing [19]. In this study, the programming language used is Python, which has the capability to run scripts that integrate MediaPipe and OpenCV to perform various computer vision tasks, including object detection and hand gesture recognition. OpenCV itself is a widely used computer vision library for image and video processing. Using OpenCV, developers can easily acquire and process video streams from camera devices, such as webcams. Additionally, OpenCV provides various functions for image manipulation, object identification, and pattern recognition, making it a highly supportive tool in the development of hand gesture recognition systems [20].

2.1 System Design

This robot uses a Raspberry Pi as the image processing center for the camera to detect hand waves. In this system, the camera sends images to the Raspberry Pi for processing. The Raspberry Pi processes the visual data, communicates with the Arduino Mega for sensor data synchronization, and sends the information to the HMI as the user interface. The Arduino Mega controls the motor driver to move the DC motor. The integration of all these components results in a responsive and interactive automation system. To integrate all components into a fully functional robot, a program consisting of three main blocks is required: image processing on the Raspberry Pi, communication and motor control on the Arduino Mega, and the implementation of a PID algorithm to ensure precise and stable robot movement. For high-precision navigation, the robot is equipped with an encoder that measures distance traveled by counting wheel rotations and a compass sensor that provides directional orientation, enabling the robot to move accurately along the specified route. Figure 3 is a block diagram and 3D design of the robot.

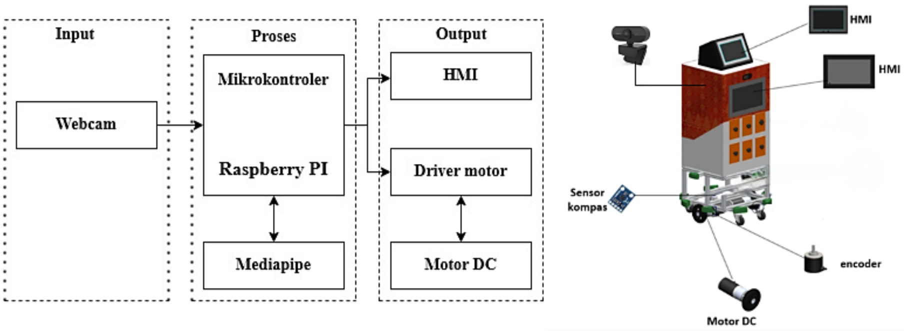


Figure 3. Block diagram and 3D design of the robot

2.2 System Flow Chart

In this study, an intelligent robot system was designed that is capable of responding to commands through the detection of hand gestures. The process begins with the initialization of the system, during which all sensors and hardware are activated. Once the system is active, the robot moves automatically along a predetermined path or trajectory. During its movement, the system continuously detects objects using a camera to locate the user's hands. If a waving hand gesture is detected, the system will proceed to the image processing stage using MediaPipe, an artificial intelligence-based library capable of detecting and tracking hand landmarks in real-time. The output of this process is a visualization of the hand's coordinate points (landmarks) displayed on the screen, which also serves as confirmation that the gesture has been correctly recognized. After that, the robot automatically moves toward the user who performed the hand wave gesture. Upon reaching the user, the robot will wait for confirmation in the form of a button press within a certain timeframe (±30 seconds). If confirmation is successfully completed within that time, the process is deemed complete. If not, the robot will automatically return to its track. Figure 4 shows the system flowchart in this study.

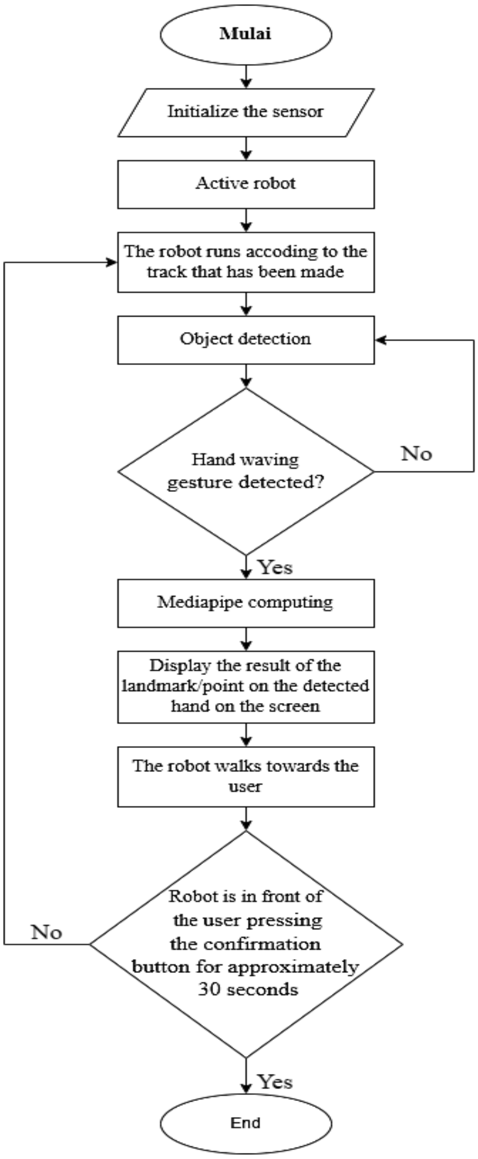

Mulai
Initialize the sensor
Active robot
The robot runs accoding to the track that has been made
Object detection
Hand waving gesture detected?
No
Yes
Mediapipe computing
Display the result of the landmark/point on the detected hand on the screen
The robot walks towards the user
Robot is in front of the user pressing the confirmation button for approximately 30 seconds
No
Yes
End

Figure 4. System Flow Chart

3. OBJECT DETECTION RESULTS

Analysis of the data from detecting hand wave objects in this study shows that testing was conducted under two lighting conditions (Flux), namely, bright and dark conditions. Testing was conducted at three different distances, as shown in Table 1.

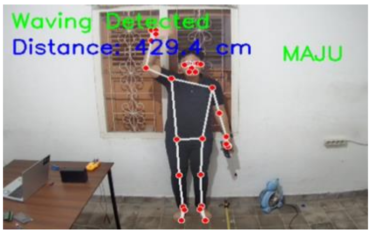
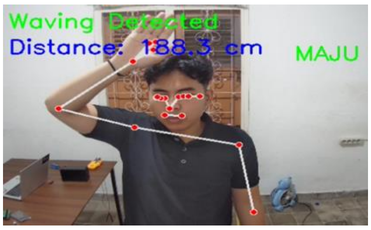
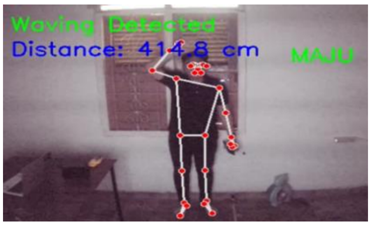
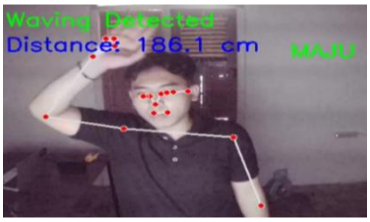
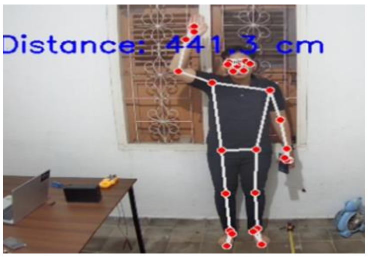
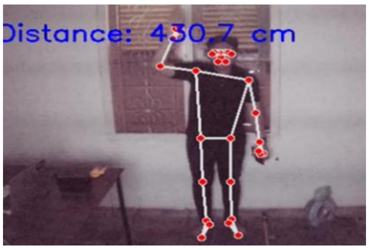
Table 1. Hand Gesture Detection Results

No	Distance	Intensity Position (Flux)		Fps Average
		Bright	Dark	
1	429,4 cm	22	20	5
2	188,3 cm	24	20	5
3	414,8 cm	63	24	5
4	186,1 cm	56	10	5
5	441,3 cm	63	20	5
6	430,7 cm	24	20	5

From the data above, it can be concluded that the further the distance, the intensity detected in bright conditions tends to increase quite sharply, while in dark conditions the increase is lower. Meanwhile, the average fps remains stable at 5 at all distances, indicating that the detection speed is not affected by changes in distance or lighting conditions.

3.1 Robot Results Detecting Objects (See Table 2)

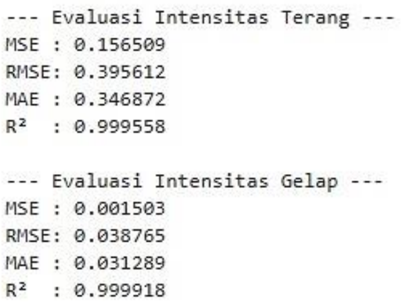
Table 2. Robot Results Detecting Objects

No.	Camera Detection Results	Description
1.		In this image, the robot successfully detected a waving motion at a distance of approximately 429.4 cm. The screen displays the message “Waving Detected” and the command “FORWARD,” indicating that the robot recognized the waving motion as a signal to move forward. The lines and dots on the body are clearly visible, indicating that the system detected the body pose accurately.
2.		At a closer distance, namely 188.3 cm, detection becomes sharper. The waving hand position is also clearly visible, with red dots marking the body joints perfectly readable. The “FORWARD” command remains visible.
3.		In dark conditions and from a distance of approximately 414.8 cm, the robot can still recognize body poses and detect hand waves even though the distance is more than 4 meters. The body lines are still visible even though the image is somewhat dark.
4.		Up to the closest distance, approximately 186.1 cm. Even in dark lighting conditions, the robot can still detect hand gestures and issue the command “FORWARD.” This proves that the system works quite stably in low light conditions.
5.		At a distance of 441.3 cm in bright conditions, the system was still able to detect body position well. However, the text “Waving Detected” did not appear. This indicates that the hand movement was not recognized as waving by the system because the right wrist was below the eyebrows, but was only considered to be a raised hand without a waving motion.
6.		At a distance of 430.7 cm in dark conditions, the system was still able to recognize body position well. However, the message “Waving Detected” did not appear. This means that the system did not recognize the hand movement as waving.

3.2 Light intensity prediction results

Figure 5 shows the results of evaluating the system's performance in predicting light intensity under two different conditions, namely bright and dark. The evaluation was conducted using four primary statistical metrics: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R²). MSE is the average of the squared differences between the actual values and the predicted values. A low MSE value indicates that the prediction error relative to the actual data is relatively small. RMSE is the square root of MSE, which also describes the magnitude of the error but in the same units as the original data. MAE measures the average absolute value of the difference between the prediction and the actual data, and the smaller the value, the better the model performance. Meanwhile, R² or the coefficient of determination shows how much of the variation in the actual data can be explained by the prediction model. An R² value close to 1 indicates that the model has a very high level of accuracy.

In bright conditions, the MSE value is 0.156509, RMSE is 0.395612, MAE is 0.346872, and R² is 0.999558. While in dark conditions, the system produces an MSE value of 0.001503, RMSE of 0.038765, MAE of 0.031289, and R² of 0.999918. R² values that are very close to 1 in both conditions indicate that the model is able to predict the data very well. In addition, the low MSE, RMSE, and MAE values, especially in the dark condition, indicate that the developed prediction system is able to produce highly accurate outputs. With these results, it can be concluded that the system used is quite reliable in predicting light intensity under both high and low lighting conditions. Figure 6 show Final optimized bright and dark intensity graph.



```
--- Evaluasi Intensitas Terang ---
MSE : 0.156509
RMSE: 0.395612
MAE : 0.346872
R²  : 0.999558

--- Evaluasi Intensitas Gelap ---
MSE : 0.001503
RMSE: 0.038765
MAE : 0.031289
R²  : 0.999918
```

Figure 5. Results of Evaluating

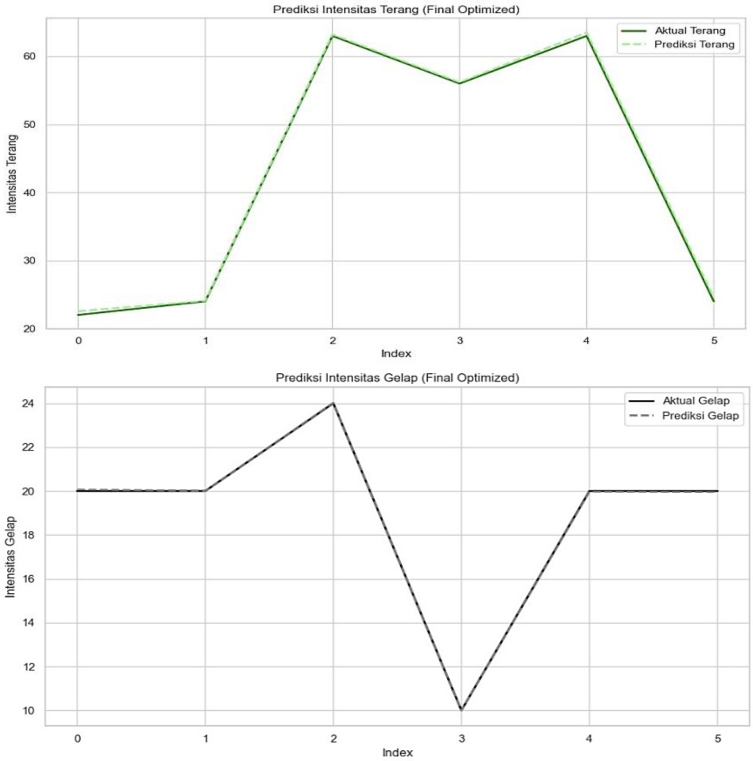


Figure 6. Final Optimized Bright and Dark Intensity Graph

4. CONCLUSION

This research successfully developed an intelligent robot system that can respond to user commands naturally through waving gesture detection using MediaPipe technology. This system integrates a Raspberry Pi as an image processor and an Arduino Mega as an actuator controller, allowing the robot to move automatically towards a user who gives a waving gesture in front of the camera. Based on the test results, the system demonstrates good performance in detecting gestures in real-time with high accuracy across various lighting conditions, including both bright and dark environments, as well as at different distances. Detection in bright conditions provides more optimal results, while in dark conditions, the system remains stable with a consistent average detection speed (FPS). Evaluation of the system using statistical parameters such as MSE, RMSE, MAE, and R² showed very satisfactory results. R² values close to 1, as well as low error values, prove that the prediction system works accurately in identifying light intensities that affect detection quality. This innovation allows for a more practical and intuitive interaction between humans and robots without physical contact, so it has great potential to be applied in automated services in various public sectors such as hospitals, restaurants, campus environments, or other public service areas. Overall, the use of MediaPipe proved effective in improving the quality of human-robot interaction in a modern, responsive and efficient manner.

REFERENCES

[1] P. Purnama, “Superkritis Karbon Dioksida–Cosolvent sebagai Alternatif Media Pemrosesan Biopolimer Polilaktida dalam Menanggulangi Permasalahan Lingkungan,” *Arus Lintas Indonesia Proxy untuk Prediksi Perubahan Iklim Regional dan Global (Indonesian Throughflow, ITF: A Proxy for Regional and Global Climate Prediction)*, pp. 70–71, 2009.

[2] T. Engineering, “Fascicle of Management and Technological Engineering Accuracy Measurement Of The National Instrument Starter Kit,” no. May, 2015.

[3] J. Wirtz, W. H. Kunz, and S. Paluch, “What are the Implications of Service Robots in Hospitality for Consumers ? By Jochen Wirtz , Professor of Marketing , National University of Singapore ; Werner Kunz , Professor of Marketing , University of Massachusetts ; and Stefanie Paluch , Professor o,” no. October, 2022.

[4] S. A. Dewangga, M. Subianto, and W. Swastika, “Implementation of Hand Gesture Recognition as Smart Home Devices Controller,” vol. 06, no. 02, pp. 63–68, 2024, doi: 10.52985/insyst.v6i2.372.

[5] M. Wameed, A. M. ALKAMACHI, and E. Erçelebi, “Tracked Robot Control with Hand Gesture Based on MediaPipe,” *Al-Khwarizmi Engineering Journal*, vol. 19, no. 3, pp. 56–71, 2023, doi: 10.22153/kej.2023.04.004.

[6] M. Arif *et al.*, “Teknik dan Multimedia Sistem Pendeteksi Tangan Berbasis Mediapipe dan OpenCV untuk Pengenalan Gerakan,” *Biner : Jurnal Ilmu Komputer*, vol. 2, no. 2, pp. 173–177, 2024, [Online]. Available: https://journal.mediapublikasi.id/index.php/Biner

[7] S. N. Budiman, S. Lestanti, S. M. Evvandri, and R. K. Putri, ‘Pengenalan Gestur Gerakan Jari Untuk Mengontrol Volume Di Komputer Menggunakan Library Opencv Dan Mediapipe’, Antivirus : Jurnal Ilmiah Teknik Informatika, vol. 16, no. 2, Art. no. 2, Nov. 2022, doi: 10.35457/antivirus.v16i2.2508

[8] F. Zhang, V. Bazarevsky, A. Vakunov, A. Tkachenka, G. Sung, C.-L. Chang, dan M. Grundmann, "MediaPipe Hands: On-device Real-time Hand Tracking," *arXiv preprint arXiv:2006.10214*, Jun. 2020.

[9]. I. Irfan, M. A. Muthalib, K. Kartika, dan S. Meliala, "Pengiraan pose model manusia pada repetisi kebugaran AI pemograman Python berbasis komputerisasi," *INFOTECH Journal*, vol. 9, no. 1, pp. 11–19, Jan. 2023.

[10] .Pranav, D. (2022, 1 Maret). MediaPipe: The ultimate guide to video processing.LearnOpenCV.https://learnopencv.com/introduction-to-mediapipe/

[11] R. Josyula and S. Ostadabbas, “A Review on Human Pose Estimation,” Oct. 2021, [Online]. Available:

[12] C. Lugaresi *et al.*, “MediaPipe: A Framework for Building Perception Pipelines,” 2019, [Online]. Available: http://arxiv.org/abs/1906.08172

[13] Bazarevsky, V.; Grishchenko, I. Pelacakan Pose Tubuh Real-Time di Perangkat dengan MediaPipe BlazePose, Google Research. Tersedia daring: https://ai.googleblog.com/2020/08/on-device-real-time-body-pose-tracking.html

[14] Sarafianos, S.; Boteanu, B.; Ionescu, B.; Kakadiaris, IA Estimasi pose manusia 3D: Tinjauan pustaka dan analisis kovariat. *Comput. Vis. Image Underst.* 2016 , *152* , 1–20.

[15] Chen, Y.; Tian, Y.; He, M. Estimasi pose manusia monokuler: Survei metode berbasis pembelajaran mendalam. *Comput. Vis. Image Underst.* 2020 , *192* , 102897.

[16] Wang, J.; Tan, S.; Zhen, X.; Xu, S.; Zheng, F.; He, Z.; Shao, L. Estimasi pose manusia 3D yang mendalam: Tinjauan. *Comput. Vis. Image Underst.* 2021 , *210* , 103225.

[17] Yurtsever, MME; Eken, S. BabyPose: Dekode komunikasi non-verbal bayi secara real-time menggunakan estimasi pose berbasis video 2D. *IEEE Sens.* 2022 , *22* , 13776–13784.

[18] Alam, E.; Sufian, A.; Dutta, P.; Leo, M. Sistem deteksi jatuh pada manusia berbasis penglihatan menggunakan pembelajaran mendalam: Tinjauan. *Comput. Biol. Med.* 2022 , *146* , 105626.

[19] T. C. A.-S. Zulkhaidi, E. Maria, and Yulianto3, "Pengenalan Pola Bentuk Wajah dengan OpenCV," JURTI (Jurnal Rekayasa Teknologi Informasi), vol. 3, 2019.

[20]]M. W. Julius Sembiring, Vara Susilowati, Vinsensius Reinard, “Evaluasi Jarak Deteksi Antara Gestur Tangan Dan Kamera Webcam Dengan Metode Mediapipe,” *Informatika dan Teknik Elektro*, vol. 2, no. 2, pp. 86–92, 2023.

BIBLIOGRAPHY OF AUTHORS

Dr. Selamat Muslimin has been a lecturer at Sriwijaya State Polytechnic since 2007. He has professional experience in private companies and banking, and is actively involved in the Mechatronics Laboratory. He holds a Bachelor’s degree in Electrical Engineering from Sriwijaya University and a Master’s degree in IT Infrastructure from Bina Darma University. He is currently pursuing a Doctoral degree in Electrical Engineering and actively conducts research, teaching, and community service.

Ekawati Prihatini is an Assistant Professor in Electrical Engineering at Sriwijaya State Polytechnic with over 20 years of academic and research experience. Her research interests include robotics, artificial intelligence, renewable energy, and assistive technology. She has led numerous projects, serves as an editor for national conferences, and is an active member of IAENG and PII. Her work has received national and international awards for innovation and social impact.

Tri Martin is an undergraduate student in Electrical Engineering at Sriwijaya State Polytechnic. He actively participated in the MSIB–MBKM program, gaining practical experience to strengthen his professional skills. He also has industry experience at PT Semen Baturaja, a state-owned cement company, where he gained insights into industrial operations, business processes, and professional workplace practices that align with his academic background.