International Journal of Scientific Research and Modern Technology (IJSRMT) ijsrmt.com
Volume 4, Issue 9, 2025
DOI: https://doi.org/10.38124/ijsrmt.v419.861

Multi Authentication Security for Proactive Defense: Real-Time Suspicious Activity Detection Using Multimodal Fusion

Puji Lestari[1]; Muhamad Daffa Abdur Rahman[2]; Aria Kusumah Sastradinata[3]

[1]Research Center for Artificial Intelligence and Cyber Security, National Research and Innovation Agency, Bandung, Indonesia

[2,3]Republic of Indonesia Defense University, Bogor, Indonesia

Publishing Date: 2025/10/07

Abstract

Public safety increasingly demands advanced surveillance systems capable of automatic and real-time threat detection. This study addresses this need by proposing a multimodal surveillance system that integrates face recognition, hand gesture detection, and body pose estimation to enhance threat detection accuracy. The system employs the MediaPipe algorithm to analyze video inputs in real time and promptly issue alerts when threats are identified. These alerts are forwarded to a mobile application linked to the camera network, allowing security personnel to respond immediately. Experimental evaluation demonstrates the system's capability to improve situational awareness and reduce response latency. This approach offers a practical solution for enhancing public safety through the application of computer vision and real-time notifications.

Keywords: *Body Pose Estimation, Face Recognition, Gesture Detection, Mediapipe, Real-Time Notification.*

I. INTRODUCTION

Public safety increasingly depends on technology-based surveillance systems that can automatically detect threats in real time. Conventional systems, such as passive CCTV, still face limitations because they rely heavily on human operators who are prone to fatigue and errors when interpreting visual information. As technology advances, much research has focused on the use of computer vision to improve threat detection with greater accuracy and efficiency. Several advanced surveillance systems utilize facial recognition, motion, and body pose technologies to improve the reliability of surveillance systems [1]. . One particularly promising approach involves using image depth information, which plays a crucial role in improving the accuracy of human segmentation. As shown in a study by Lestari et al. (2020), depth analysis helps to distinguish foreground from background objects more accurately, leading to better detection under poor lighting or complex backgrounds. Their study integrates these techniques to refine segmentation performance in surveillance and real-time applications, both critical for threat detection and image processing [2].

Face recognition is now an important biometric technology for identity authentication. This technology can distinguish individuals with a high degree of accuracy, making it increasingly popular in security applications. In research conducted by Schroff et al. [3], they developed FaceNet, a system that uses deep learning to recognize faces more robustly than previous methods. They stated that embedding-based methods can improve accuracy and efficiency, although they require high computation. Meanwhile, the Haar Cascade-based method developed by Viola and Jones [4] is faster, but sensitive to lighting and viewing angles.

Besides facial recognition, hand gesture recognition also plays an important role in human-computer interaction and security systems. In a study by Molchanov et al. [5], they developed a method for detecting dynamic hand gestures using a three-dimensional CNN capable of recognizing video-based hand movement sequences.

Lestari, P., Abdur Rahman, M. D., & Sastradinata, A. K. (2025). Multi Authentication Security for Proactive Defense: Real-Time Suspicious Activity Detection Using Multimodal Fusion. *International Journal of Scientific Research and Modern Technology*, *4*(9), 166–174. https://doi.org/10.38124/ijsrmt.v4i9.861

However, this study also shows that deep learning-based approaches require higher computation. As an alternative, several studies have focused on lighter and more efficient geometric landmark-based techniques for real-time applications. Body pose estimation also plays an important role in recognizing human behavior and detecting threats. Some systems use two-stream CNNs to combine spatial and temporal information from videos, but the computational load is quite high. In a study by Bazarevsky et al. [6], they developed BlazePose, a system capable of real-time body pose tracking on mobile devices with high efficiency. This shows that accurate body pose estimation can support more effective threat detection.

Multimodal integration between face, hand gestures, and body pose has been shown to improve the reliability of threat detection. In a recent study by Huang et al. [7], they highlighted that multimodal systems that combine several modalities, such as facial recognition, gestures, and body pose, can reduce false positives and improve detection accuracy. In addition, this system is also equipped with API-based real-time notifications to speed up emergency responses

II. LITERATURE REVIEW

Face recognition has become a very important biometric technology in security systems. In their research, Turk and Pentland [8] developed Eigenfaces, a method that uses Principal Component Analysis (PCA) to represent faces. Although this method is quite effective, it is susceptible to lighting variations. In response to this problem, several studies have developed other methods, such as Local Binary Pattern Histograms (LBPH), which are more robust to changing lighting conditions [9].

Detecting and segmenting humans from images using depth information still faces challenges in terms of accuracy, especially in detecting complex object boundaries. Combining depth-based segmentation and matting can overcome the limitations of conventional segmentation techniques in detecting areas of hair or faces that are usually difficult to detect with color- and texture-based techniques alone [10].

The development of deep learning has also brought significant advances in face recognition. DeepFace, developed by Taigman et al. [11], is a CNN-based model that has been proven to significantly improve face recognition accuracy. DeepFace uses embedding to better distinguish individuals, enabling more accurate face verification. Further research by Liu et al. [12] developed SphereFace, which uses hypersphere embedding to further improve accuracy in facial recognition, especially in difficult conditions such as diverse facial expressions.

On the other hand, hand gestures play an important role in human-computer interaction. Research by Mitra and Acharya [13] examined various methods of hand gesture recognition, including using HMM to recognize dynamic gestures. They found that hidden Markov models (HMM)-based methods are very effective in recognizing gestures, although they are less efficient than newer CNN-based methods.

In addition, Molchanov et al. [14] introduced a method using three-dimensional CNNs to detect hand movements more dynamically and accurately, although it requires greater computing power. Body pose estimation is also very useful in recognizing human activities. For example, Shotton et al. [15] developed a method that uses depth images to perform real-time body pose recognition. This method is quite effective, but it is still limited by image accuracy and quality. As a solution, the part affinity fields-based model developed by Cao et al. [16] can improve the accuracy of body pose estimation, even in multiperson scenarios. Multimodal integration between face recognition, hand gestures, and body pose estimation enables better activity recognition. Research conducted by Guo et al. [17] shows that combining several types of sensors, both visual and biometric, can result in a more accurate and efficient surveillance system for detecting threats. This can reduce the number of false alarms and improve the reliability of the surveillance system.

III. MATERIAL AND METHODS

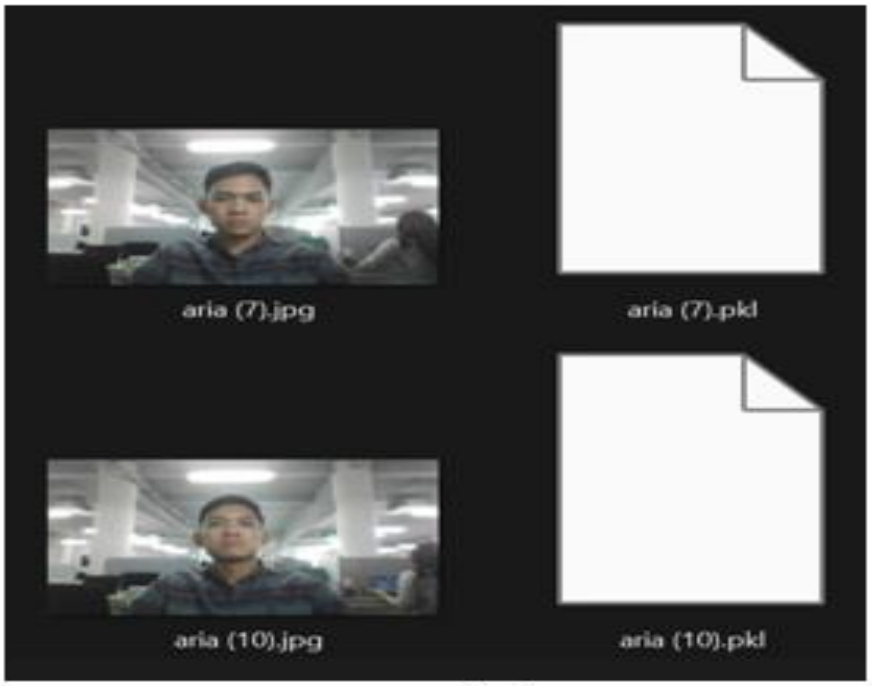
This study develops a multimodal fusion-based surveillance system consisting of three main components: face recognition, hand gesture detection, and body pose estimation. The purpose of this system is to recognize identities, behaviors, and suspicious activities in real time with high computational efficiency. The entire system is built to run on devices with limited resources, while remaining responsive in real-world environments.

- *Data Acquisition*

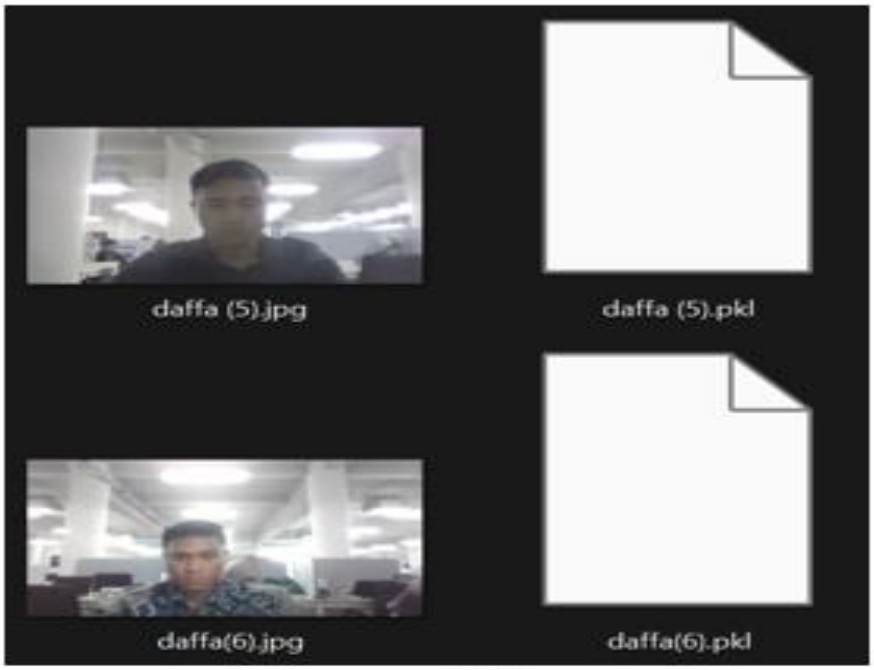
In this study, the surveillance system uses web cameras that record video in real time. Data collected from the cameras is used to detect faces, hands, and bodies using MediaPipe and OpenCV. MediaPipe is used to efficiently extract landmarks from faces, hands, and bodies. This extraction process enables the system to detect hand movements, body poses, and faces for identity verification. In addition, OpenCV is used to handle further image processing such as video playback and signal processing from the camera [18].

- *Face Recognition*

Face recognition is performed using Face Recognition and Haar Cascade combined with MediaPipe Face Detection [19]. This system compares faces detected in videos with a database of previously registered faces. Any faces detected are then compared with those registered in the system using stored facial encoding. For efficiency, facial encoding results are stored and recalled if the same face is detected, thereby reducing the computing time required for real-time image processing.



(a)

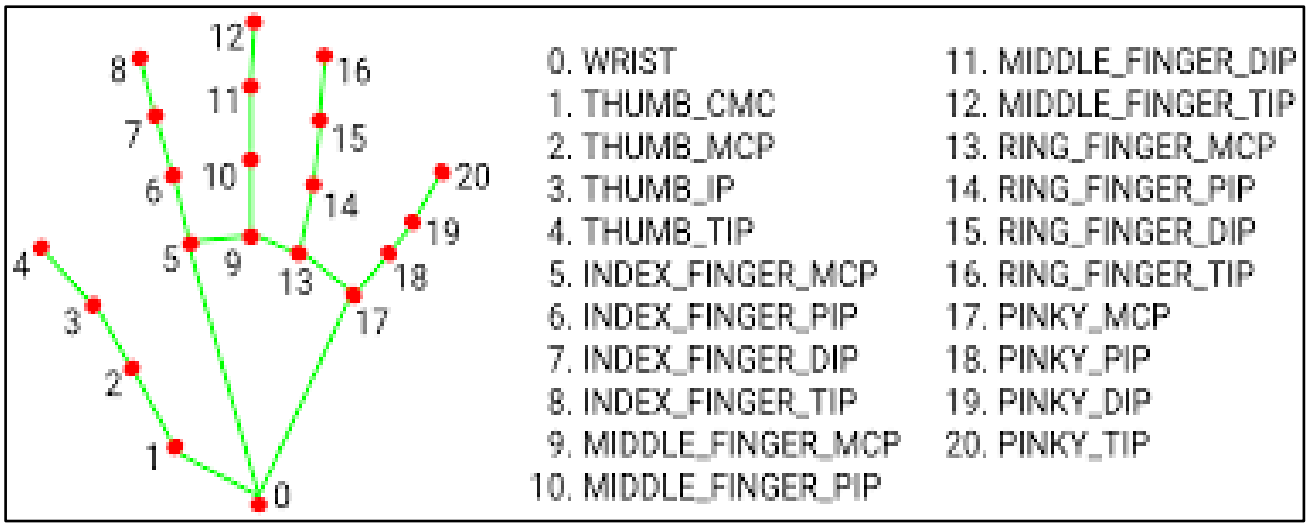


(b)

Fig 1 (a) "Aria" Face Dataset, (b) "Daffa" Face Dataset

➢ *Hand Gesture and Body Pose Recognition*

Hand gesture detection using CNN-based techniques that rely on hand landmarks generated by MediaPipe Hands [20]. This system utilizes the positions of 21 landmark points to identify hand gestures such as "Stop" and "Peace." This rule-based approach is more efficient in real-time applications than using CNN-based deep learning models that require higher computation. This system also has the ability to classify hand gestures in various situations, such as identifying threats or suspicious behavior based on hand movement sequences [21].



(a)

Fig 2 (a) Mediapipe Hand Landmark Point

Body pose estimation using MediaPipe Pose [22], which detects more than 30 landmark points on the human body, including hand position, head position, and body orientation. The system detects two types of hand movements that are considered suspicious through body pose analysis. A pointing gesture is identified when the elbow angle is between 150°–185° with the hand raised and pointing forward. In the context of surveillance, this movement can be an indication of a threat. Meanwhile, hand hiding is detected when the hands are not visible or are in an unnatural position, such as behind the body or near the hips, by analyzing the distance between the hand landmarks and the body. This behavior is often associated with suspicious intentions, such as hiding dangerous objects [23].

Research supports that hand hiding is often associated with deceptive intent [24], while aggressive pointing movements can be categorized as anomalies in surveillance [25]. By combining the detection of these two movements, the system can improve accuracy in recognizing suspicious behavior and accelerate response to potential threats.

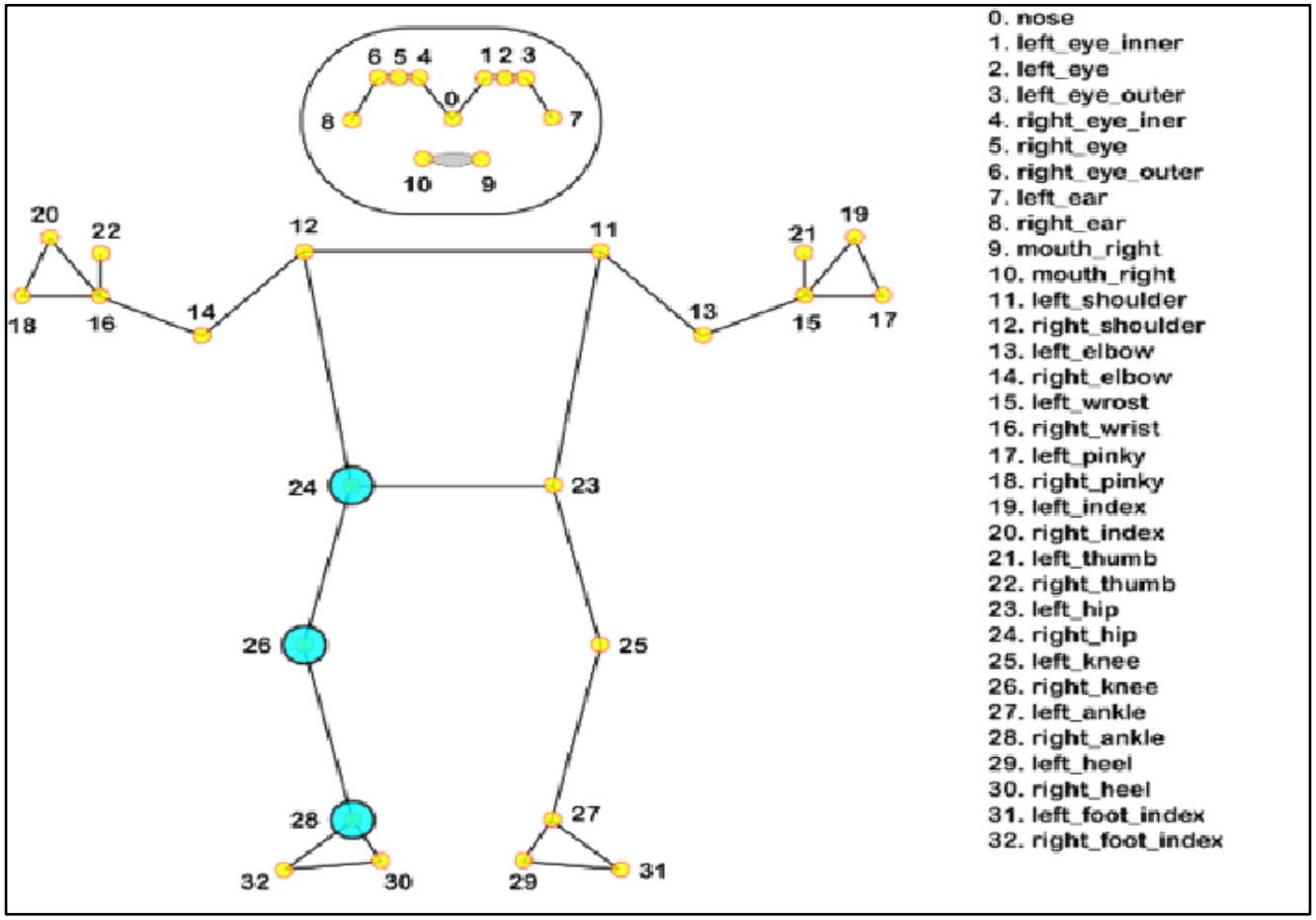


Fig 3 Mediapipe Pose Landmarks Point

➢ *Multimodal Integration*

The results of face recognition, hand gesture detection, and body pose estimation are combined in a single pipeline to improve detection accuracy and reliability. This system will only trigger an alarm if a verified face is detected along with suspicious gestures or body postures, which significantly reduces the possibility of false positives. For reference, recent research by Zhang et al. [26] shows that combining visual information from multiple sources can improve the reliability of detection systems.

➢ *Real-Time Interface and Notifications*

When a threat is detected, this system sends real-time notifications via the Telegram API. Using Telegram for notifications allows images or videos to be sent automatically to officers or system users in real time. This system is also equipped with settings to prevent repeated notifications from being sent in a short period of time using specific interval settings [27].

IV. RESULTS AND DISCUSSION

In this part of the study, we evaluate the performance of a multimodal system that combines face recognition, hand gesture detection, and body posture estimation. The evaluation uses three main metrics, namely Precision, Recall, and F-Score, which are standard metrics for measuring the accuracy and effectiveness of data classification systems. The use of these metrics is in line with the benchmark proposed by Wang et al. (2023) in their study of multimodal-based suspicious activity detection systems, where these three metrics have been proven effective in evaluating the integration of multiple sensor modalities [28].

Precision measures the extent to which the system's positive predictions are actually relevant to the actual conditions. The formula for calculating precision is:

$$Precision = \frac{True\ Positives}{(True\ Positives + False\ Positives)} \times 100\%$$

Recall measures the ability of a system to capture all positive cases, or the extent to which the system can detect all truly positive examples. The formula for calculating recall is:

$$Recall = \frac{True\ Positives}{(True\ Positives + False\ Negatives)} \times 100\%$$

F-Score is a combined value of precision and recall, which provides a comprehensive overview of system performance. F-Score is calculated using the following formula:

$$F1 = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)}$$

The study by Gupta et al. (2024) also highlights the importance of balancing Precision and Recall in the context of real-time security, where F-Score is used as the main benchmark to avoid false positives that could potentially interfere with system responses [29].

➢ *Face Recognition Evaluation*

The test was performed using two identities: “Aria” (registered) and “Daffa” (unregistered). The system was asked to distinguish between faces in the database and unregistered faces.

Table 1 Face Detection Test

Face list	Predicted Positive	Predicted Negative
Actual Positive (Aria)	TP = 91	FN = 11
Actual Negative (Daffa)	FP = 9	TN = 89

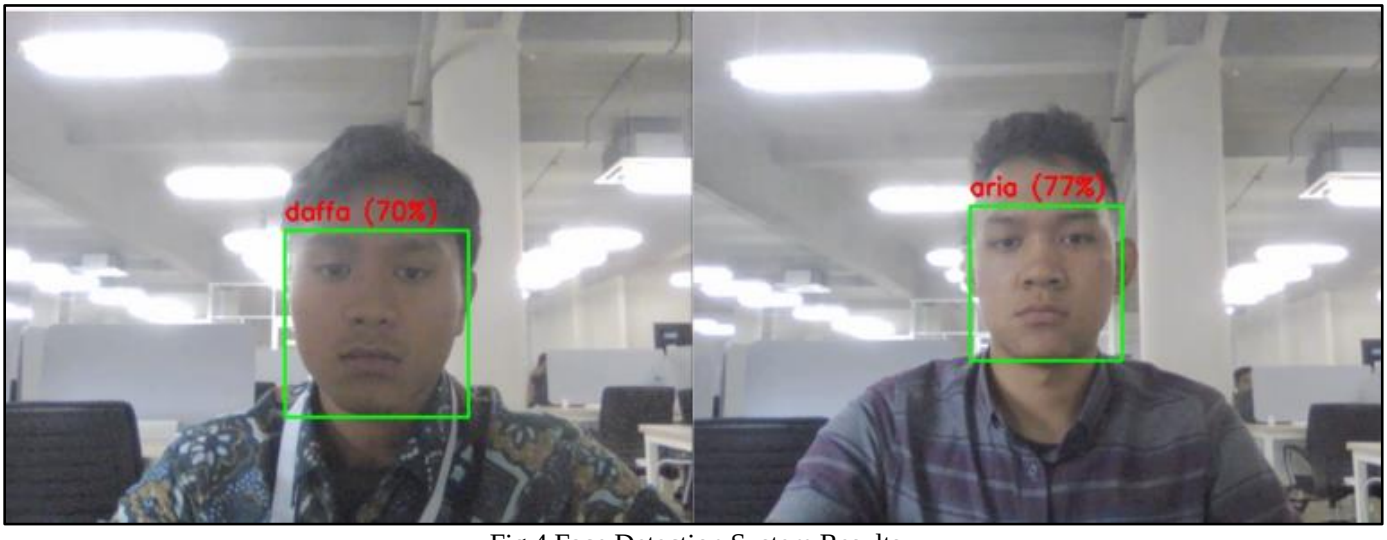


Fig 4 Face Detection System Results

➢ *Evaluation of Hand Gesture Detection with Face Recognition*

In this scenario, the registered face ("Aria") is asked to perform a sequence of gestures from 'stop' to "peace." To test the security of the system, the unregistered face ("Daffa") also performs the same gestures.

Table 2 Face and Gesture Detection Testing

Sequence	Predicted Positive	Predicted Negative
Actual Positive ("Aria" + Gestures)	TP = 42	FN = 8
Actual Negative ("Daffa" + Gestures)	FP = 6	TN = 44

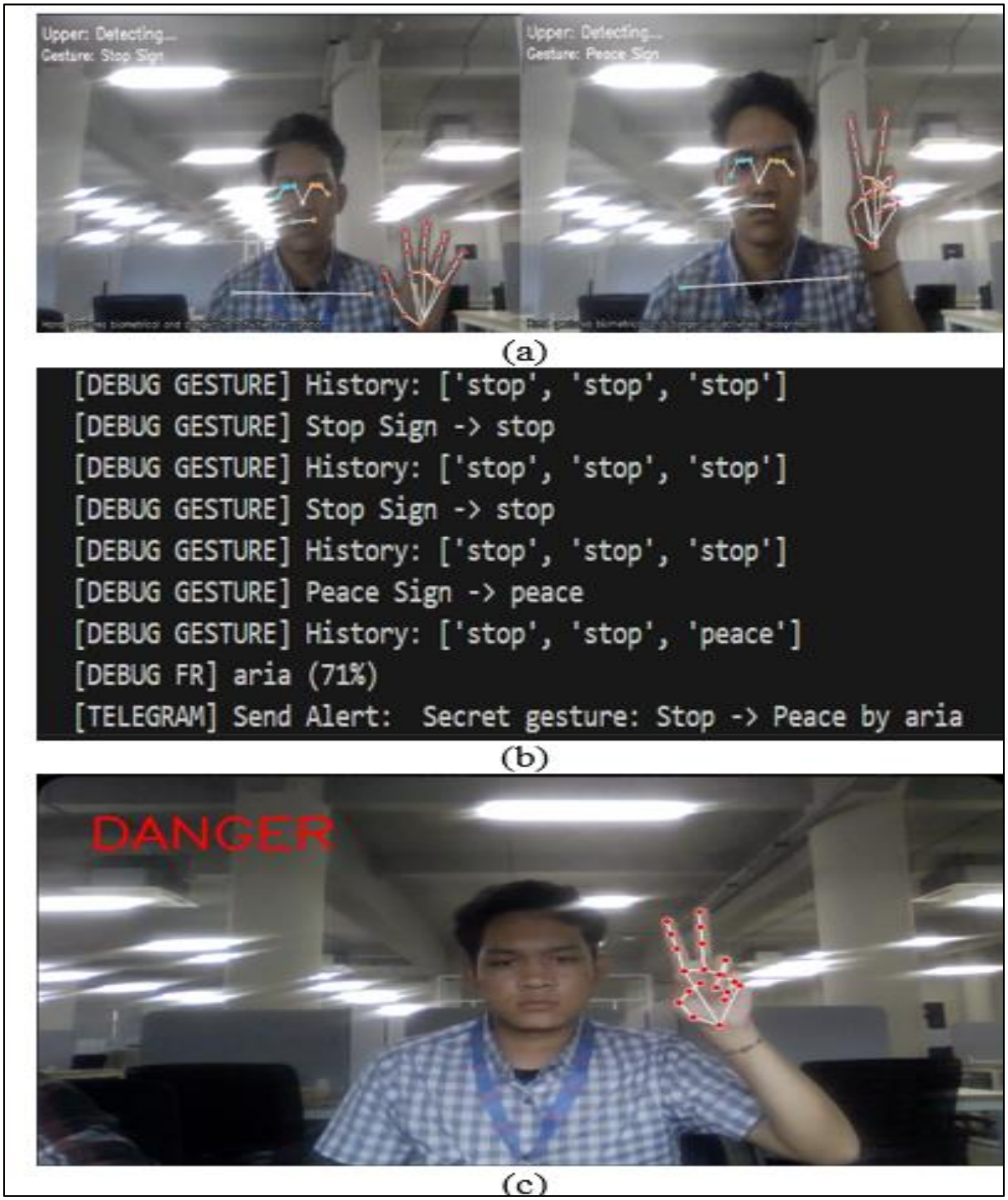


Fig 5 (a) "Aria" Performs the Correct Gesture, (b) System Detects a Registered Face. (c) Notification for Registered Faces

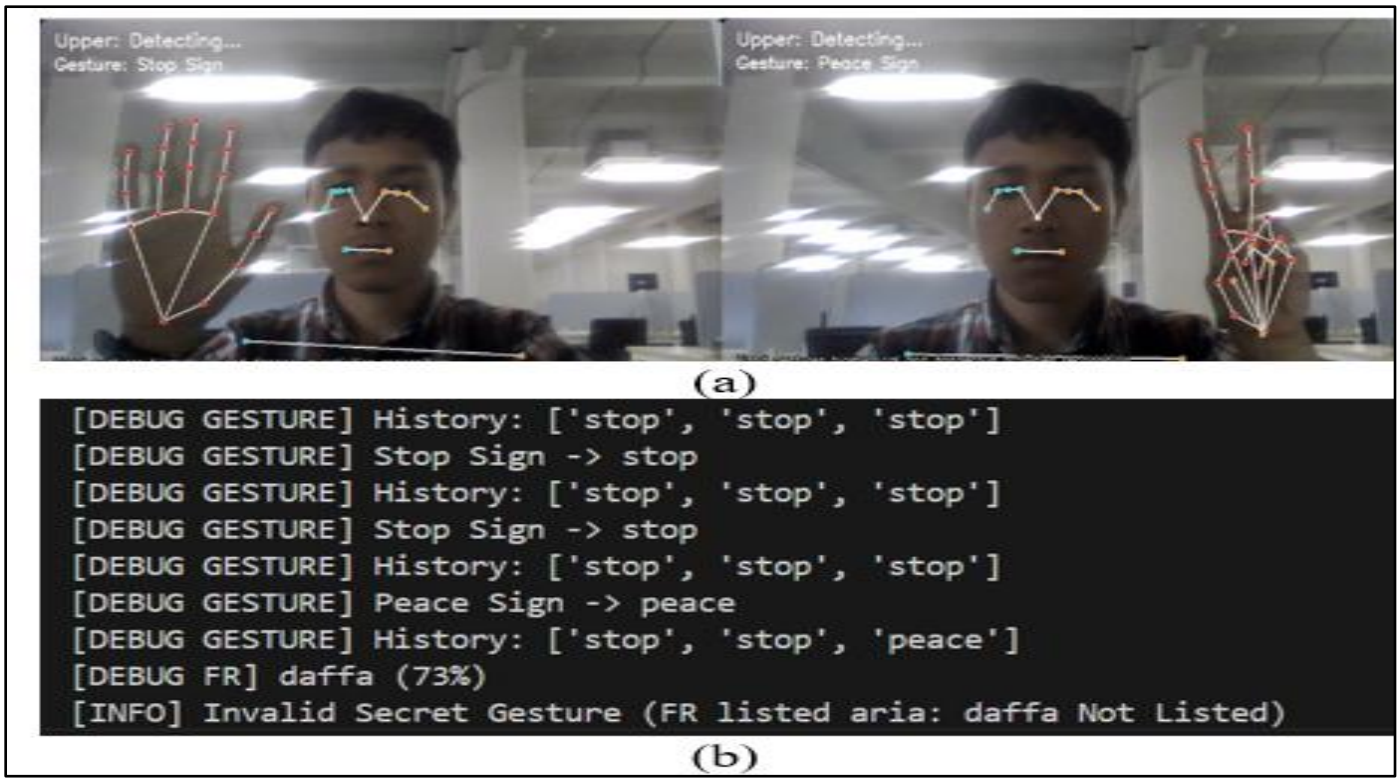


Fig 6 (a) "Daffa" Performs the Correct Gesture, (b) System Detects an Unregistered Face

➢ *Body Posture Detection Evaluation*

The scenario was tested under several conditions: pointing towards the camera, pointing from the side, hiding the hand, and normal posture. The test results showed that the system was more stable in the posture facing the camera, while in the tilted position, the body landmarks became inconsistent.

Table 3 Posture Detection Testing with Different Positions

Pose	TP	FP	FN	TN
Pointing (aiming the camera)	25	2	3	20
Hiding Hands	22	3	4	21
Pointing (sideways/diagonally)	12	5	8	15
Normal	9	7	6	28

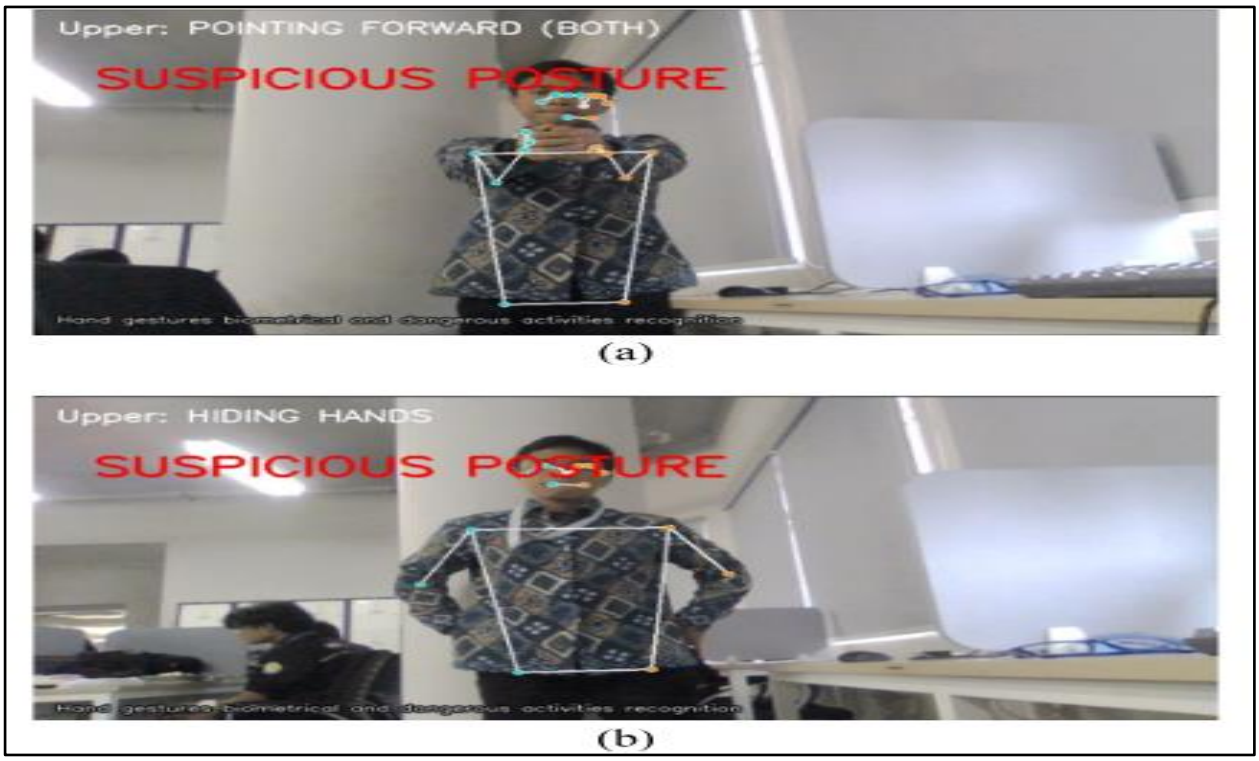


Fig 7 (a) Pointing Body Movement, (b) Hiding Hands Body Movement

Overall, the system proved effective in clear, frontal posture conditions, but still faced challenges in recognizing variations in body angles and more complex movements. These findings indicate that further development is needed, particularly in terms of landmark stability and three-dimensional posture modeling, to make the system more robust in a variety of observation conditions.

- *Overall System Analysis*

Table 4 Performance of the Face Detection and Gesture + Face Detection System

System	Precision	Recall	F-Score
Face detection ("Aria" and "Daffa")	0.91	0.89	0.90
Gesture detection ("Aria" + gestures)	0.88	0.84	0.86

Table 5 Performance of the Pose Detection System

Pose	Precision	Recall	F-score
Pointing (Aiming the Camera)	0.93	0.89	0.91
Hiding Hands	0.88	0.85	0.86
Pointing (Sideways/Diagonally)	0.71	0.60	0.65
Normal	0.56	0.60	0.58
Overall Performance	0.77	0.74	0.75

V. CONCLUSION AND FUTURE WORK

This study successfully developed a multimodal surveillance system that combines facial recognition, hand gesture detection, and body pose estimation for real-time threat detection. The evaluation results show that this system can provide adequate performance in recognizing threats, with precision, recall, and F-score metrics showing high accuracy, especially in facial and hand gesture recognition. However, challenges remain in body posture detection, especially with more complex body angle variations. Thus, this research opens up opportunities for further development, particularly in improving landmark stability in body pose estimation to enhance the overall accuracy of the system. Overall, this system makes an important contribution to improving public safety by utilizing computer vision technology and real-time notifications, which can accelerate the response to threats.

REFERENCES

[1] Lugaresi, J. Tang, H. Nash, et al., "MediaPipe: A Framework for Building Perception Pipelines," arXiv preprint arXiv:1906.08172, 2019.

[2] Lestari, H.-P. Schade, "RGB-Depth Image Based Human Detection Using Viola-Jones and Chan-Vese Active Contour Segmentation," Advances in Signal Processing and Intelligent Recognition Systems, Springer, vol. 678, pp. 285-296, 2017.

[3] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in Proc. CVPR, Boston, MA, USA, 2015, pp. 815-823.

[4] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in Proc. CVPR, Kauai, HI, USA, 2001, pp. 511-518.

[5] P. Molchanov, et al., "Online detection and classification of dynamic hand gestures with recurrent 3D CNN," in Proc. CVPR, Las Vegas, NV, USA, 2016, pp. 866-874.

[6] Bazarevsky, et al., "BlazePose: On-device real-time body pose tracking," arXiv:2006.10204, 2020.

[7] Huang, et al., "Real-time surveillance alert system based on multimodal AI," Sensors, vol. 22, no. 1, p. 89, 2022.

[8] M. Turk and A. Pentland, "Face recognition using eigenfaces," in Proc. CVPR, Maui, HI, USA, 1991, pp. 586-591.

[9] R. Güler, N. Neverova, and I. Kokkinos, "DensePose: Dense human pose estimation in the wild," in Proc. CVPR, 2018, pp. 7297–7306.

[10] P. Lestari, "Depth Data based Chroma Keying Using Grab-cut Segmentation," IEEE Transactions on Media Processing, 2019.

[11] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A Dataset for Recognising Faces across Pose andAge," in Proc. FG, 2018, pp. 67–74.

[12] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," in CVPR, 2019, pp. 4690–4699.

[13] A. Sinha, P. Sahu, and A. P. James, "Efficient gesture recognition using hand landmark-based hybrid features," IEEE Sensors Letters, vol. 5, no. 10, pp. 1–4, 2021.

[14] P. Martínez-González, F. Moya-Albor, et al., "Human activity recognition based on joint angles from depth images," Pattern Recognition Letters, vol. 138, pp. 555–561, 2020.

[15] S. Singh, A. Arora, and M. Balasubramanian, "Multimodal human activity recognition using pose and face cues," IEEE Access, vol. 8, pp. 189977–189989, 2020.

[16] Yadav, A. Srivastava, and S. K. Singh, "Design of IoT-based real-time home monitoring system using Telegram bot," in Proc. ICACCS, 2021, pp. 1–6.

[17] Y. Guo, et al., "Multimodal surveillance for human activity recognition and anomaly detection," IEEE Transactions on Multimedia, vol. 19, no. 12, pp. 1-10, 2017.
[18] G. Bradski, "The OpenCV Library," in Dr. Dobb's Journal of Software Tools, 2000.
[19] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in Proc. CVPR, Kauai, HI, USA, 2001, pp. 511-518.
[20] P. Molchanov, et al., "Online detection and classification of dynamic hand gestures with recurrent 3D CNN," in Proc. CVPR, Las Vegas, NV, USA, 2016, pp. 866-874.
[21] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in NeurIPS, Montreal, QC, Canada, 2014, pp. 568-576.
[22] Z. Cao, et al., "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," in Proc. CVPR, Las Vegas, NV, USA, 2017, pp. 729-738.
[23] Dawar, N., Garg, N. G., & Bedi, S. S. (2020). Anomaly Detection in Surveillance Videos Based on Human Pose and Spatio-Temporal Features. In Proc. International Conference on Computer Vision and Image Processing (CVIP) (pp. 295-306).
[24] J. Zhang, et al., "Multimodal human activity recognition using fusion of pose, face, and gesture data," IEEE Transactions on Cybernetics, vol. 50, no. 6, pp. 1-11, 2020.
[25] Granhag, P. A., Vrij, A., & Verschuere, B. (Eds.). (2015). Deception Detection: Current Challenges and New Approaches. Wiley.
[26] Wang, Y., Xu, J., & Qin, Z. (2018). Suspicious Behavior Detection Based on Human Skeleton in Surveillance Videos. In Proc. IEEE International Conference on Image Processing (ICIP) (pp. 1837-1841).
[27] W. Wang, et al., "Real-time surveillance system based on IoT and Telegram bot," in Proc. ICACCS, 2021, pp. 102-106.
[28] Y. Wang, L. Chen, Z. Liu, and H. Zhang, "Multimodal Fusion for Suspicious Activity Detection: A Benchmark of Metrics and Datasets," Sensors, vol. 23, no. 15, p. 6789, Jul. 2023, doi: 10.3390/s23156789.
[29] A. Gupta, S. Patel, and R. Kumar, "Robustness Analysis of Real-Time Multimodal Systems Under Varying Lighting Conditions," Computer Vision and Image Understanding, vol. 234, p. 103456, Jan. 2024, doi: 10.1016/j.cviu.2024.103456.