Human Interaction and Emerging Technologies (IHIET-FS 2025), Vol. 196, 2025, 206–218
https://doi.org/10.54941/ahfe1005969

Transforming Elderly Care With Vision-Based Hand Gesture Recognition: A Deep Learning Framework

Riazul Islam and Seyed Ali Ghorashi

Department of Engineering and Computing School of Architecture, Computing and Engineering, University of East London, United Kingdom

ABSTRACT

In today's rapidly evolving healthcare landscape, there is a growing trend on integrating technology with human experience to enhance patient care and outcomes. This paper presents a vision-based hand gesture recognition (HGR) system as a human-computer interface (HCI) that translates natural movements into intuitive device and application controls. Designed to empower the elderly and improve patient monitoring, this non-intrusive approach bridges compassion with cutting-edge technology, fostering a more connected and responsive healthcare environment. The study introduces a deep learning-based detec tion framework that enables the practical application of HGR, supporting gesture-based virtual assistants for digital interactions and facilitating communication between elderly residents in care homes and remote caregivers. The proposed framework consists of two deep learning models: one for region-of-interest (ROI) detection and another for gesture classification. The recognition process begins with capturing RGB images, selected for their versatility, affordability, and compatibility. These images are processed using YOLOv8 (You Only Look Once, version 8) for ROI detection. Following detection, several image processing techniques are applied to overcome common HGR constraints, including ROI extraction and grayscale conversion for faster computation, resizing for model optimization and distance adapt ability, HSV conversion for lighting invariance, and histogram equalization for enhanced feature extraction. The processed image is then fed into a convolutional neural network (CNN) for gesture classification. The proposed HGR system achieves an impressive accuracy of 97% in training and validation, significantly improving upon conventional HGR systems, which are often developed and tested under controlled conditions. By demonstrating its effectiveness in real-world applications, this framework addresses key limitations in existing HGR research and highlights its potential for practical implementation in healthcare and beyond.

Keywords: Hand gesture recognition, Human-computer interaction, Deep learning, YOLOv8, Convolutional neural network, Elderly care

INTRODUCTION

Human-Computer Interaction (HCI) refers to the process of data exchange between an end user and a computer system and hand gestures serve as a powerful tool for bridging the communication between humans and machines. (Aggarwal and Arora, 2022). Hand gestures, considered a core aspect of nonverbal communication, can be conveyed through the central part of the palm, the positioning of the fingers, and the overall appearance of the hand (Suarez and Murphy, 2012), and have found applications ranging from assisting deaf–mute communities and controlling robots to automating homes and enabling medical interventions (Rahim et al., 2020). In the context of HCI, HGR is regarded as an emerging input medium that promises a more intuitive and convenient mode of interaction (Panwar and Mehra, 2011). According to Premaratne and Premaratne (2014), the origin of hand gesture recognition used for computer control is related to the introduction of glove-based controlled applications when researchers found that employing gestures inspired from sign languages could be used to replicate simple command for computer interface, and the accuracy of those command has seen an uprise when integrated with accelerometers, infrared cameras and fibreoptic bend-sensors. Although sensor-based approaches often yield strong performance, they do possess inherent limitations (Oudah et al., 2020). First and foremost, an adequate hardware configuration is necessary, which could turn out to be quite costly. In addition, it obstructs the free movement of the hand, making it difficult to perform gestures. As an alternative to sensor-based recognition, vision-based approaches were introduced, which eliminates specialized hardware requirements and hand-movement limitations (Ahuja and Singh, 2015). The efficacy and adaptability of vision-based HGR is heavily dependent on the image acquisition process. RGB-D images, captured using depth sensors like Microsoft Kinect, offer enhanced hand gesture recognition by combining colour and depth data. How ever, these sensors often produce low-density depth images and require significant computational power, making them costly and less accessible (Suarez and Murphy, 2012; Gu et al., 2017). Infrared cameras, on the other hand, perform well in low-light conditions by detecting heat emissions but struggle with boundary details and require precise calibration due to temperature variations (Geng and Yin, 2020). Despite its limitations, RGB imaging remains the most accessible and cost-effective option for HGR, as most consumer devices are equipped with RGB cameras. However, challenges such as background noise and lighting variations complicate recognition, leading to the necessity of advanced image processing techniques for improved accuracy (Hu and Zhu, 2019).

HGR systems also heavily rely on classification and identification algorithms, which are crucial for interpreting extracted information and precisely identifying gestures. Traditional machine learning techniques, such as Support Vector Machines (SVM) and K-Nearest Neighbour (K-NN), have been widely used for data categorization and clustering (Liu et al., 2008; Taunk et al., 2019). While effective, these methods often struggle with complex variations in gestures and environmental conditions. To address these limitations, Convolutional Neural Networks (CNNs) have emerged as a powerful deep learning approach capable of automatically extracting essential features from images through convolution operations, learning hierarchical representations, and generalizing well across different lighting conditions and backgrounds, making them highly effective for hand gesture recognition (Chung et al., 2019; Chauhan et al., 2018).

This study focuses on developing a HGR system that facilitates effective communication between elderly individuals and their caregivers. The proposed framework employs a two-stage deep learning architecture, integrating YOLOv8 for precise ROI detection and a CNN for gesture classification, thereby enhancing accuracy by eliminating background noise and improving feature extraction. To ensure stable recognition, a semi-continuous recognition strategy is introduced, reducing miss-classification caused by rapid frame fluctuations. Additionally, the framework incorporates an adaptive pre-processing pipeline, including grayscale conversion, HSV transformation, and histogram equalization, to optimize feature visibility across diverse conditions. The system's efficacy is demonstrated through real-world testing, enabling seamless gesture-based control of applications, highlighting its practical applicability beyond theoretical models. A user evaluation further validates its effectiveness, achieving a 9.03/10 real-time accuracy score, underscoring its reliability and applicability in real-world scenarios.

The remainder of this paper is organized as follows: Section II presents a comparative review of previous studies in the HGR domain, highlighting their real-time applications, methodologies, image processing techniques, and accuracy. Section III provides a brief overview of the proposed framework, while Section IV details the experimental results. Finally, Section V concludes the study by summarizing the key findings and outlining future research directions.

RELATED WORK

The evolution of hand gesture recognition in human computer interaction has progressed from traditional machine learning models to advanced deep learning approaches, improving accuracy and robustness. Early studies, such as Gajjar, Mavani, and Gurnani (2017), focused on feature-based methods, using the Haar-Like model to identify hand movements and enable interactive applications like a Real-Time Paint Toolbox. Islam et al. (2019) advanced the field by incorporating CNN based techniques, leveraging data augmentation strategies such as rescaling, zooming, and shifting, which improved accuracy from 92.87% to 97.12%. Hu and Zhu (2019) introduced an RGB only gesture recognition system that refined background detection and landmark identification before applying an adapted CNN model for classification. Further enhancements, such as those by Chung, Chung, and Tsai (2019), integrated deep learning with tracking mechanisms, using modified versions of AlexNet and VGGNet alongside kernelized correlation filters to achieve high accuracy in home automation applications. More recent approaches, like those of Rahim et al. (2020), have focused on refining segmentation techniques, evaluating multiple algorithms (YCbCr, SkinMask, and HSV) to enhance recognition under varying environmental conditions. The following Table 1 provide an overview of the current state of HGR technologies.

Table 1: Findings from the literature.

Paper	Processing	Model	Real-Time	Limitations	Accuracy
Babu et al., 2015	Morph. ops, HSV/gray conv.	Contour template matching	App control (MS Word)	Needs black background	95%
Gajjar et al., 2017	Binary conv., colour filtering	Haar-like features & ML	Paint Toolbox	Needs controlled background	96%
Islam et al., 2019	K-Gaussian, grayscale	CNN	Not mentioned	One hand only	97.12%
Hu and Zhu, 2019	HSV YCbCr conv., Markov Fields	Two-step CNN recog.	Not mentioned	Only 249 images tested	81.2%
Chung et al., 2019	YCbCr conv., morph. ops	ROI detect., TCF, CNN TL	Not mentioned	No light adaptation	99.1% (Train) 95.61% (Test)
Rahim et al., 2020	HSV, YCbCr, skin-mask	CNN / SoftMax	Not mentioned	Small dataset (2000 images), no real-time testing	97.43%

This review of existing literature highlights that most studies have not focused on real-time application integration. While some studies have achieved high recognition accuracy, they have often relied on small datasets or controlled environments, raising the question: would the model have been as accurate during real-world testing as claimed, where challenges such as lighting variations were apparent? Although many approaches have employed ROI detection or background subtraction to enhance accuracy and reduce computational complexity, the potential of neural networks for this purpose has remained largely unexplored due to extensive data requirements. Furthermore, no study has systematically evaluated recognition speed in real-world applications or explored semi-continuous recognition, where processing has occurred at short intervals rather than on every captured frame. This study has endeavoured to mitigate issues related to lighting accuracy and distance to some extent while focusing on achieving a high recognition rate in a semi-continuous process.

FRAME-WORK

This research introduces a novel hand gesture recognition (HGR) method that leverages the advanced capabilities of YOLOv8 for precise Region of Interest (ROI) detection. The proposed framework first detects and isolates hand regions from an image, reducing background interference and improving classification accuracy. Once the ROI is identified, it is cropped and pre-processed using a series of image enhancement techniques before being fed into a CNN for classification. This systematic approach ensures robust and efficient gesture recognition, addressing key limitations in conventional HGR systems. As shown in the following Figure 1, the overall process of the system is explained in the following sections.

Dataset Collection & ROI Detector Training

This study utilizes the HaGRID dataset, a large-scale col lection of high-resolution RGB images covering 18 distinct gesture classes with over 550,000 images and a total size of approximately 770 GB. Each class contains more than 30,000 images, capturing diverse lighting conditions and background variations, ensuring model robustness in real-world applications (Kapitanov et al., 2024). To optimize computational efficiency, a subset comprising 5.40% of the dataset across five gesture classes is selected for analysis. For precise ROI detection, the YOLOv8 model is trained on 27,000 RGB images extracted from the HaGRID dataset, with 5,000 labelled as 'no gesture' and 27,000 labelled as 'gesture.' To improve generalization, several data augmentation techniques are applied, including random rotation (±10°), scaling (0.5), and brightness adjustment (hsvv = 0.4). Additionally, loss function parameters are optimized, assigning a weight of 8.5 to the box loss and 0.8 to the classification loss, prioritizing precise bounding box localization and gesture differentiation. The trained YOLOv8 detector effectively marks the hand region with a bounding box and enters into preprocessing.

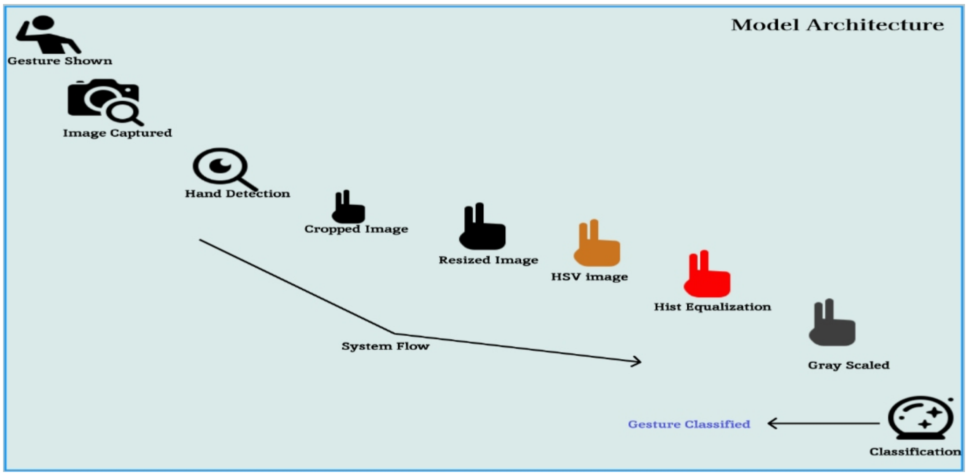

Figure 1: Flowchart of the HGR framework.

Image Pre-Processing

The pre-processing phase begins by extracting the ROI from the detected bounding boxes provided by YOLOv8, as shown in Figure 2. The cropped region is obtained using:

$$\text{cropped region} = \text{image}\,[y_1 : y_2, x_1 : x_2]$$

where (x1, y1) and (x2, y2) denote the top-left and bottom right coordinates of the bounding box, respectively. These coordinates are directly derived from YOLOv8's detection output, ensuring precise localization of the hand gesture before proceeding with image enhancement and feature extraction.

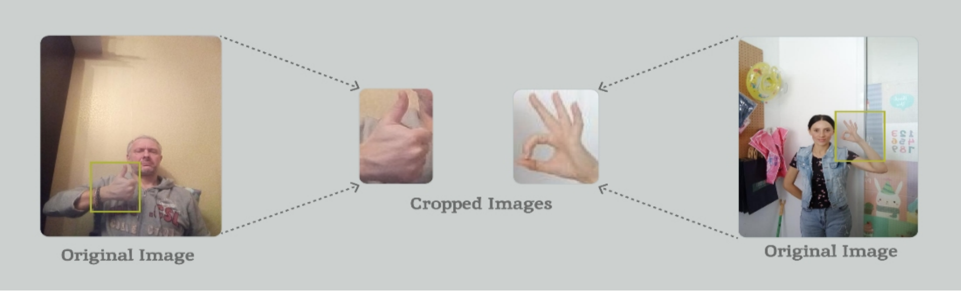

Figure 2: Image cropping based on bounding box.

Image Resizing: Once the extraction of the ROI component from the images is completed, the cropped regions were resized to a fixed dimension (h,w) to maintain a uniform input size for the CNN. The resizing process is mathematically defined using inter-cubic interpolation as:

$$I'\left(x',y'\right) = \sum_{i=0}^{h-1}\sum_{j=0}^{w-1} I\left(i,j\right)\cdot K\left(x'-i,y'-j\right)$$

Where ($I^{'}\left(x^{'},y^{'}\right)$)represents the interpolated pixel value at coordinates (x′,y′), I(i,j) denotes the original pixel intensity at (i, j), and K (x′ −i,y′ −j) is the inter-cubic interpolation kernel used to estimate new pixel values. This approach ensures smooth and accurate resizing while preserving essential spatial features, which is critical for stabilizing CNN training and improving model performance (Keys, 1981).

HSVConversion: One of the preprocessing tactics used to tackle issues arising from variations in lighting conditions among photos was the conversion of images from the RGB colour space to the HSV (Hue, Saturation, Value) colour space. This transformation facilitates the segregation of colour data from brightness data, enabling the model to concentrate more efficiently on the intrinsic characteristics of the images, rather than being influenced by fluctuations in lighting.

Histogram Equalization: Another preprocessing step involved applying histogram equalization to the value (V) channel of the HSV images. Histogram equalization enhances image contrast by redistributing pixel intensity values, ensuring a more uniform intensity distribution across the full dynamic range. The transformation function is defined as:

$$T(v) = \frac{(L-1)}{N}\sum_{i=0}^{v} h(i)$$

where T(v) represents the new intensity value after equalization, L is the maximum intensity level, N is the total number of pixels, and h(i) denotes the histogram count of intensity level i (Gonzalez and Woods, 2002).

Grayscale Conversion & Data Augmentation: To reduce computational complexity while preserving essential structural features, the pre-processed images were converted to grayscale, removing redundant colour information. Addition ally, data augmentation techniques were employed to enhance dataset diversity. Random horizontal flipping was applied to generalize gestures for both hands, while random rotation (±30°) improved robustness against variations in gesture orientation. Figure 4 illustrates the effects of these augmentation strategies.

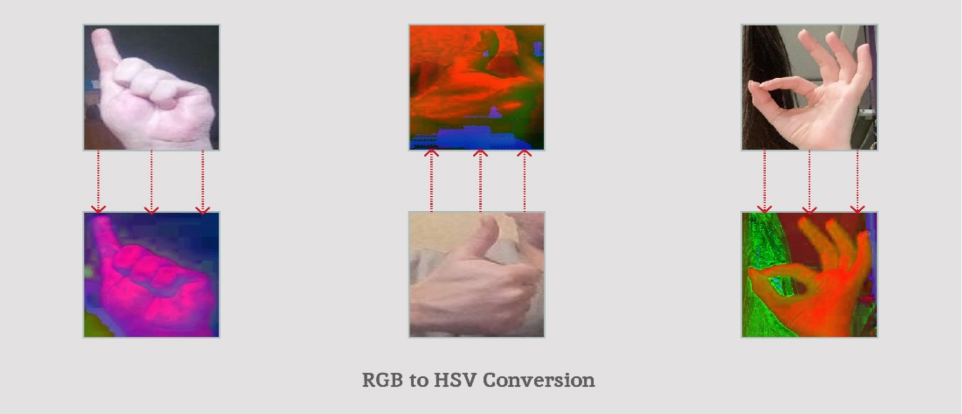

Figure 3: RGB to HSV conversion.

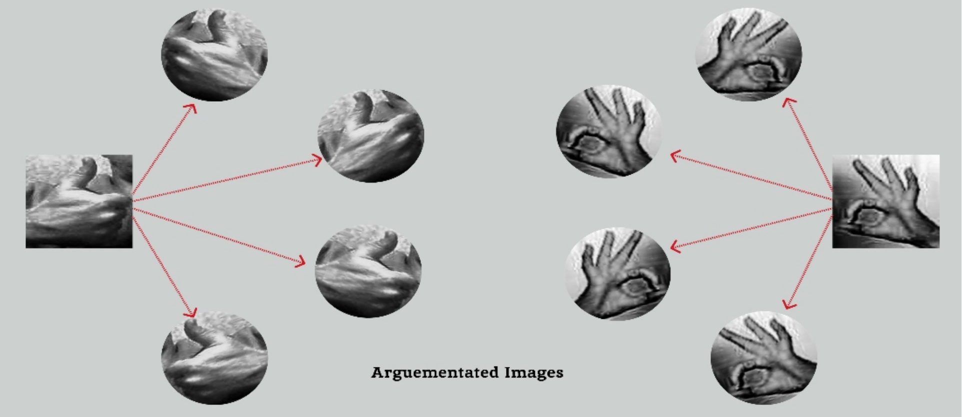

Figure 4: Data augmentation.

Gesture Classification

Following pre-processing, a CNN-based classifier was developed using the Keras Sequential framework to effectively recognize hand gestures. The model consists of multiple layers optimized for feature extraction, normalization, and classification. The architecture begins with a convolutional layer employing 256 kernels of size k × k, where spatial filtering is per formed to detect gesture features. The convolution operation is defined as:

$$F_l^{(j)} = \sigma\left(\sum_{i=1}^{M} W_{i,j}^{(l)} F_{l-1}^{(i)} + b_j^{(l)}\right)$$

where ($F_l^{(j)}$) is the feature map at layer l, ($W_{i,j}^{(l)}$) represents filter weights, b(l) j is the bias term, and σ $(\cdot)$ is the activation function. To introduce non-linearity and enhance gradient propagation, the ReLU activation is applied:

$$f(x) = \max(0, x)$$

Feature normalization is achieved through batch normalization, stabilizing activations by transforming them into a standardized distribution:

$$\widehat{x_i} = \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}}$$

where xi is the activation, μB and σ2 B denote the batch mean and variance, and ϵ prevents division by zero. To retain dominant features while reducing spatial dimensions, max pooling is performed:

$$P_l^{(j)} = F_l^{(i)}$$

The extracted features are then processed through dense layers, where classification is performed using the SoftMax activation function. Dropout layers are incorporated to prevent overfitting by randomly deactivating neurons during training, thereby enhancing model generalization.

RESULTS

The effectiveness of the proposed HGRS is evaluated based on the detector and classifier performance, followed by a comprehensive real-time assessment.

Detector Performance

The detection model was validated on an independent dataset comprising over 2,500 images, including both gesture and non-gesture instances. The validation results exhibited a consistent decrease in both box loss and classification loss, as shown in Figure 5, mirroring the trend observed during training and confirming effective model convergence. The precision metrics further substantiate the model's robustness, with an mAP50 exceeding 99% and an mAP50-95 reaching approximately 85%, indicating high detection accuracy across varying intersection over union (IoU) thresholds. These results demonstrate the model's strong ability to localize hand regions accurately while maintaining minimal false positives.

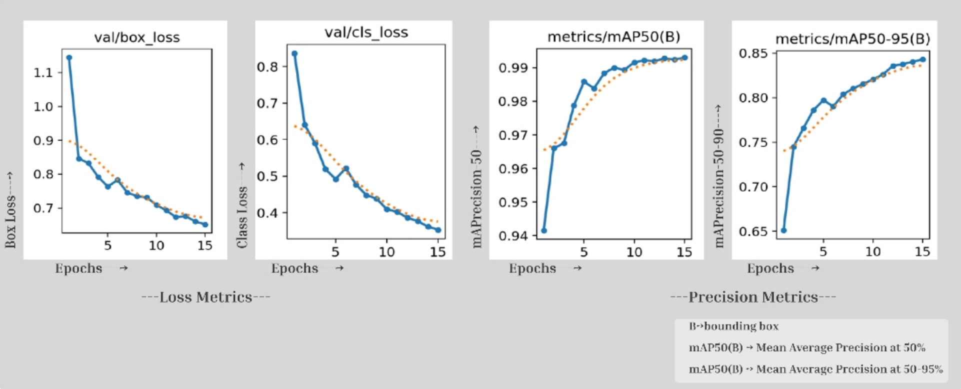

Figure 5: Performance on validation.

Classifier Performance

The classification model underwent rigorous validation on a dedicated test set while ensuring a balanced representation of gesture classes. The training process was optimized using the Adam optimizer, which adaptively refined the learning rate to minimize categorical loss. The model demonstrated progressive improvements in classification accuracy, ultimately achieving 97.71% on the training set and 97.77% on the validation set, confirming its ability to generalize effectively across unseen samples (Figure 6).

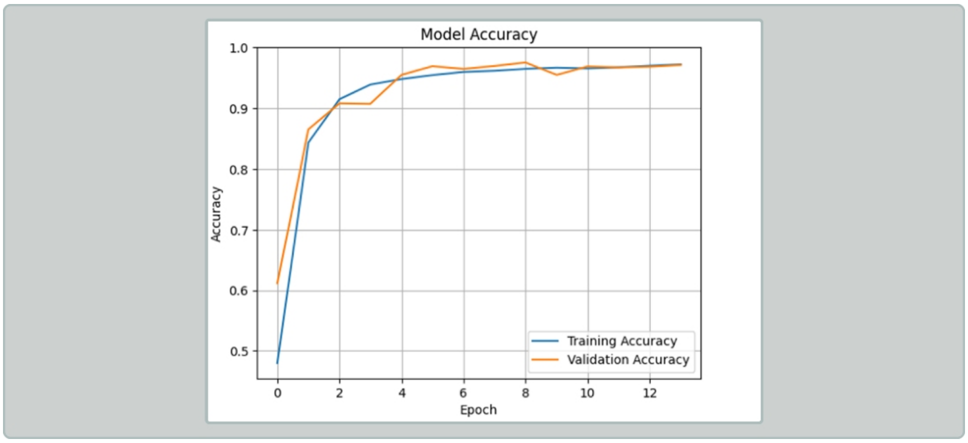

Figure 6: Model accuracy.

Real-Time Evaluation

To assess the practical applicability of the system, the model was deployed in a real-world organizational environment as seen in Figure 7, where classification reliability was evaluated under natural conditions, including variations in lighting, hand positioning, and dynamic user interactions. The assessment combined empirical analysis and subjective user feedback, wherein participants executed predefined gestures to interact with the system. A user-based evaluation measured system responsiveness and usability, with survey results yielding an accuracy rating of 9.03 out of 10, further supporting the model's effectiveness in real-time deployments (Figure 8).

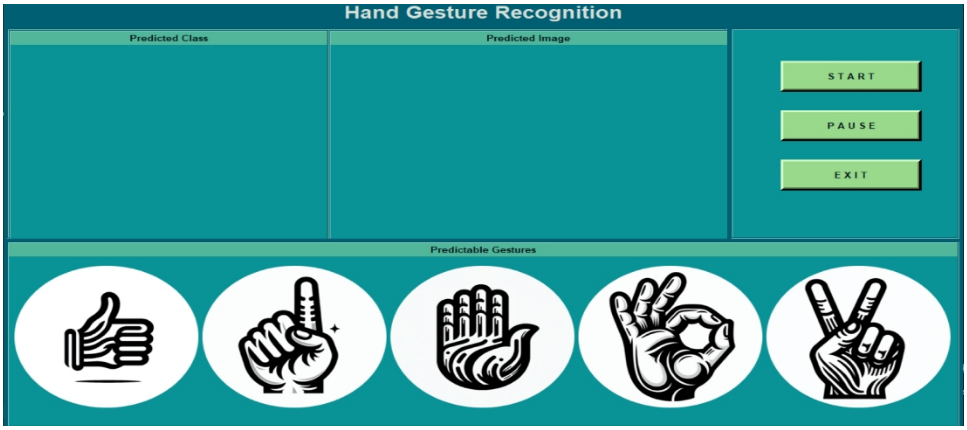

Figure 7: Deployed GUI.

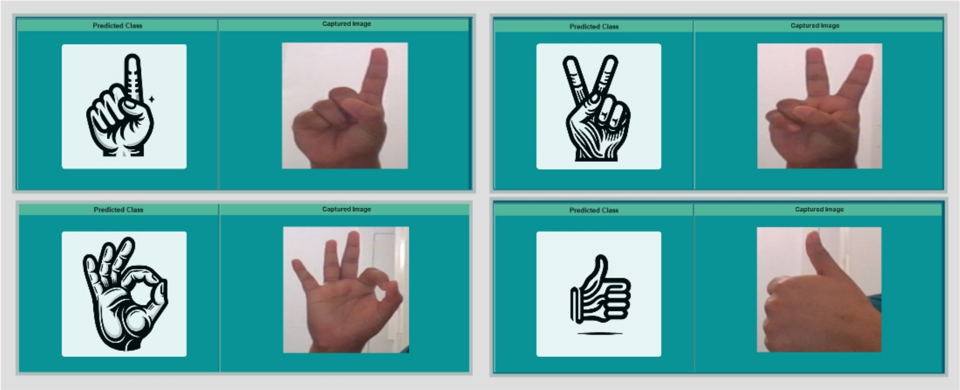

Figure 8: Real time testing.

During testing, two primary issues were identified affecting classification accuracy. The first was the lack of a mechanism to handle false positives, as the classifier continuously pro cessed inputs, attempting to classify even non-gesture frames, leading to misclassifications. Additionally, variations in subject distance impacted the cropped image quality; distant subjects resulted in lower-resolution ROIs, causing a loss of pixel intensity after resizing and reducing classification accuracy.

The developed system was evaluated in an experimental setting where participants performed a predefined set of hand gestures, each mapped to a specific message or request. Upon recognizing a gesture, the system accurately identifies it and transmits the corresponding instruction to caregivers through a suitable medium (e.g., a pager, speaker, etc.). These instructions may include requests for assistance, environmental adjustments (such as turning on lights or adjusting room temperature), or emergency alerts. This approach enables intuitive and contactless communication, particularly beneficial for elderly individuals or those with speech impairments, eliminating the need for verbal interaction (Figure 9).

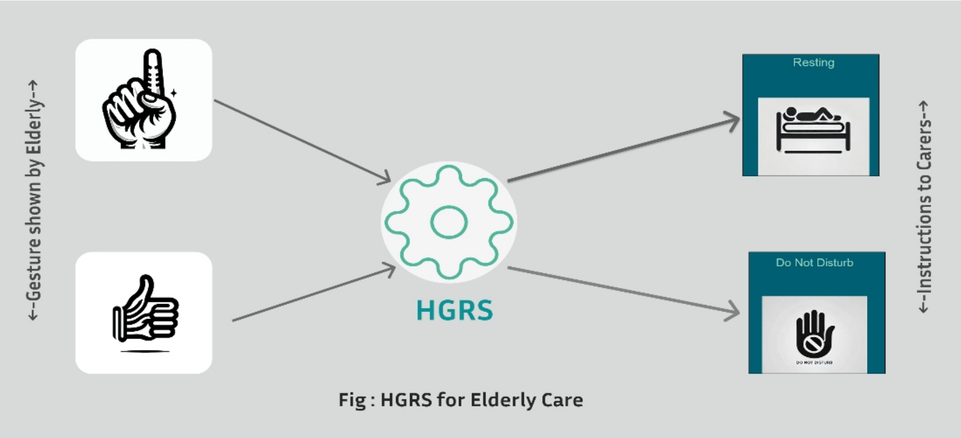
←-Gesture shown by Elderly-→
HGRS
Resting
Do Not Disturb
←-Instructions to Carers-→
Fig : HGRS for Elderly Care

Figure 9: Real time testing.

CONCLUSION

This study proposed a vision-based Hand Gesture Recognition System (HGRS) that integrates a two-stage deep learning architecture, utilizing YOLOv8 for precise ROI detection and a CNN for classification. The system effectively addresses challenges associated with real-time gesture recognition, such as variations in lighting, background complexity, and subject distance. The experimental results demonstrate the system's high accuracy of 97.77%, reinforcing its robustness and adaptability for practical applications. Additionally, a real time evaluation within an organizational setting validated its reliability, with a user satisfaction rating of 9.03 out of 10. The findings confirm that deep learning-based HGR systems can bridge the gap between theoretical models and real-world applications, making contactless interaction more intuitive and efficient. Future work will explore the hybridization of deep learning frameworks with advanced optimization techniques to further enhance system performance. This approach can significantly contribute to transforming elderly care by enabling gesture-based control for assistive technologies, allowing seamless communication for sending instructions without physical contact. To improve the results, it is recommended to apply a threshold to eliminate low-confidence recognitions, ensuring more reliable predictions. Additionally, utilizing the entire dataset is advised to enhance diversity and provide the model with a better understanding of unseen real-world data. Introducing an extra class is also recommended to minimize the occurrence of false positives, improving overall system accuracy. Furthermore, to make the system independent of distance—within a certain range—it is suggested to avoid using image resizing functions and instead focus on developing a CNN-based classifier capable of making predictions regardless of image size, ensuring robustness in varying real-world conditions.

REFERENCES

A. A. Babu, S. Varma, and R. Nikhare, "Hand gesture recognition system for human-computer interaction using contour analysis," *IJRET*, vol. 4, no. 3, 2015.

A. Kapitanov, K. Kvanchiani, A. Nagaev, R. Kraynov, and A. Makhliarchuk, "Hagrid–Hand Gesture Recognition Image Dataset," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 4572–4581.

H.-Y. Chung, Y.-L. Chung, and W.-F. Tsai, "An efficient hand gesture recognition system based on deep CNN," in *2019 IEEE International Conference on Industrial Technology (ICIT)*. IEEE, 2019, pp. 853–858.

J. Suarez and R. R. Murphy, "Hand gesture recognition with depth images: A review," in *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 2012, pp. 411–417.

K. Geng and G. Yin, "Using deep learning in infrared images to enable human gesture recognition for autonomous vehicles," *IEEE Access*, vol. 8, pp. 88227–88240, 2020.

K. Aggarwal and A. Arora, "An approach to control the PC with hand gesture recognition using computer vision technique," in *2022 9th International Conference on Computing for Sustainable Global Development (INDIACom)*. IEEE, 2022, pp. 760–764.

K. Taunk, S. De, S. Verma, and A. Swetapadma, "A brief review of nearest neighbor algorithm for learning and classification," in *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*. IEEE, 2019, pp. 1255–1260.

M. A. Rahim, A. S. M. Miah, A. Sayeed, and J. Shin, "Hand gesture recognition based on optimal segmentation in human-computer interaction," in *2020 3rd IEEE International Conference on Knowledge Innovation and Invention (ICKII)*. IEEE, 2020, pp. 163–166.

M. K. Ahuja and A. Singh, "Static vision-based hand gesture recognition using principal component analysis," in *2015 IEEE 3rd International Conference on MOOCs, Innovation and Technology in Education (MITE)*. IEEE, 2015, pp. 402–406.

M. Panwar and P. S. Mehra, "Hand gesture recognition for human-computer interaction," in *2011 International Conference on Image Information Processing*. IEEE, 2011, pp. 1–7.

M. Oudah, A. Al-Naji, and J. Chahl, "Hand gesture recognition based on computer vision: A review of techniques," *Journal of Imaging*, vol. 6, no. 8, p. 73, 2020.

M. Z. Islam, M. S. Hossain, R. ul Islam, and K. Andersson, "Static hand gesture recognition using convolutional neural network with data augmentation," in *2019 Joint 8th International Conference on Informatics, Electronics & Vision (ICIEV) and 2019 3rd International Conference on Imaging, Vision & Pattern Recognition (ICIVPR)*. IEEE, 2019, pp. 324–329.

P. Premaratne and P. Premaratne, "Historical development of hand gesture recognition," *Human Computer Interaction Using Hand Gestures*, pp. 5–29, 2014.

R. Chauhan, K. K. Ghanshala, and R. Joshi, “Convolutional neural network (CNN) for image detection and recognition,” in *2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC)*. IEEE, 2018, pp. 278–282.

R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 2nd ed. Prentice Hall, 2002.

R. Keys, “Cubic convolution interpolation for digital image processing,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 6, pp. 1153–1160, 1981.

S. Gu, W. Zuo, S. Guo, Y. Chen, C. Chen, and L. Zhang, “Learning dynamic guidance for depth image enhancement,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3769–3778.

V. Gajjar, V. Mavani, and A. Gurnani, “Hand gesture real-time paint toolbox: Machine learning approach,” in *2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)*. IEEE, 2017, pp. 856–860.

Y. Liu, Z. Gan, and Y. Sun, “Static hand gesture recognition and its application based on support vector machines,” in *2008 Ninth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing*. IEEE, 2008, pp. 517–521.

Z. Hu and X. Zhu, “Gesture detection from RGB hand image using modified convolutional neural network,” in *2019 2nd International Conference on Information Systems and Computer Aided Education (ICISCAE)*. IEEE, 2019, pp. 143–146.