TIPTEKNO'24
MEDICAL TECHNOLOGIES CONGRESS
10-12 October 2024
Yasmin Bodrum Resort, Muğla, Türkiye

Increasing the Accessibility of Disabled Individuals with AI Integrated Smart Home Systems

Ufuk BESENK, Tugce TOPRAK M. Alper SELVER, Ahmet OZKURT
Dept. of Electrical Electronics Eng. & BioIzmir
Dokuz Eylul University
Izmir, Turkey
contact.ufuk.besenk@gmail.com

Omur OCAL
Core Bina Otomasyon Teknolojileri
San ve Tic Ltd S¸ti
Izmir, Turkey

***Abstract*—This paper introduces a smart home control system leveraging gesture-based interactions to enhance the accessibility and usability of home automation technologies, particularly for users with physical disabilities. Traditional control mechanisms, such as touchscreens and remotes, often fail to meet the needs of users with disabilities. This research addresses this gap by developing an intuitive, non-verbal communication system that interprets human hand gestures as commands using deep learning techniques. A ceiling-mounted camera detects and tracks humans, with fine-tuning of the YOLO-V8 model allowing for precise hand detection and classification. Experimental results on HaGRID (HAnd Gesture Recognition Image Dataset) using 554,800 images over 37,583 subjects in various scenes and lighting conditions demonstrate the system's effectiveness and accuracy, making it suitable for extended smart home automation and human-computer interaction by transforming gestures into com- mands that control various devices and increase the accessibility of individuals with disabilities.**

***Keywords*—Convolutional Neural Networks, Gesture recognition, Human-machine systems, Internet of Things, Image Classification, Object Detection, Real-time systems.**

I. Introduction

The rapid evolution of smart home technology enables more intuitive and inclusive user interfaces. Traditional control mechanisms, such as touchscreens and remote controls, often fail to accommodate the needs of individuals with physical disabilities. This limitation not only impacts the usability of smart home systems but also hinders their adoption among a broader range of users. To address this gap, we propose a gesture-based control system for smart homes, enabling users to interact with their environment through hand gestures—providing an intuitive, non-verbal communication method. Smart home systems have gained immense popularity in recent years due to their ability to enhance convenience, security, and energy efficiency in residential environments. However, the reliance on conventional input devices like touchscreens, remote controls, and voice commands poses significant challenges. For instance, individuals with mobility impairments or speech disabilities may find it difficult to use these interfaces effectively [1]. Moreover, even able- bodied users may prefer hands-free interaction methods for convenience and hygiene reasons, particularly in situations where their hands are occupied or dirty [2]. Previous research in gesture recognition for human-computer interaction has shown promising results, but there remains a significant gap in creating systems that are both highly accurate and computationally efficient for real-time applications. Early systems relying on template matching and rule-based approaches achieved limited success due to their inability to generalize across diverse gestures and varied lighting conditions [3]. The advent of machine learning techniques, particularly convolutional neural networks (CNNs), has revolutionized gesture recognition by enabling the automatic extraction of intricate features from visual data [4]. Our system leverages state-of-the-art deep learning tech- niques, such as utilizing YOLO-V8 [5] and MobileNetV3 [6]. The primary objective of this study is to develop a smart home control system that interprets static hand gestures as commands for home device control, thereby enhancing ac- cessibility and presenting an intuitive user interface that does not rely on physical touch or fine motor skills. This system is designed to robustly mitigate environmental variables such as lighting variations and complex backgrounds, ensuring consistent performance. In Figure 2 and 2, the system showcases two scenarios where users perform a gesture directed towards the ceiling- mounted camera, which then triggers the corresponding action. In Figure 1, this action corresponds to the shutdown of all active technological devices in the room. In Figure 2 a pregnant woman opens the blinds and windows. These simple visual demonstrations highlight two practical applications of the gesture recognition system in a real-world smart home environment.



Fig. 1: Illustration of the proposed system which shuts down all devices in the room according to hand gesture.

In smart homes, the developed gesture recognition system allows users to control their homes effortlessly without physi-

979-8-3315-2981-9/24/$31.00 ©2024 IEEE

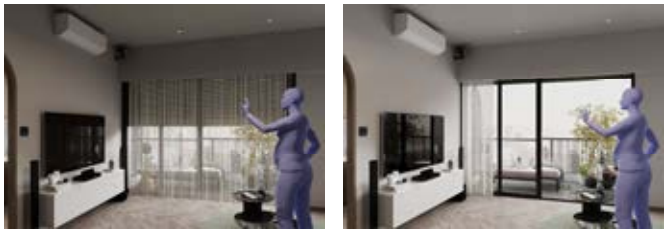


Fig. 2: Demonstration of a system use-case where a pregnant woman opens the blinds and windows without moving.

cal touch or voice commands. This touch-free control is particularly beneficial for individuals with limited mobility, enabling easier management of living spaces. To optimize computational efficiency, this research incorporates mobile algorithms, particularly utilizing models like MobileNetV3 [6] designed for accuracy and training efficiency. These models support high accuracy while being computa- tionally lightweight, mak- ing them exceptionally suitable for gesture recognition tasks within a smart home context. Their streamlined architecture ensures rapid processing and respon- siveness, critical for real- time applications in accessibility- focused smart environments [4]. In addition to these models, the You Only Look Once (YOLO) series of detectors have established themselves as efficient and practical tools. They designed for high accuracy and high inference speed and easy to deploy for downstream applications [7].

II. Related Works

The journey of hand gesture recognition began with early systems relying heavily on conventional techniques such as template matching, rule-based methods, and statistical approaches. These systems often struggled with low accuracy, particularly in complex environments with varying lighting conditions. Traditional machine learning methods like Support Vector Machines (SVMs) and K-Nearest Neighbors (KNN) improved accuracy rates to around 80% to 90%, but these approaches were still limited by their inability to effectively generalize across diverse gestures and complex hand movements [8], [9]. The introduction of deep learning, particularly Convolu- tional Neural Networks (CNNs), marked a significant turning point in HGR. CNNs have proven exceptionally effective in automatically learning hierarchical features from visual data, enabling the recognition of intricate patterns that are crucial for accurately identifying complex gestures under varied conditions. Krizhevsky et al. introduced the AlexNet architecture, which set a new standard in image classification and laid the groundwork for applying deep CNNs to gesture recognition [10]. Subsequent models, such as VGG16 by Simonyan and Zisserman [11], and MobileNetV2 by Sandler et al ., have been widely adopted, achieving accuracy rates above 95%, setting new benchmarks in the field.

Several studies have explored the application of hand gesture recognition for smart home automation. Kurian et al. (2023) developed a visual gesture-based home automation system using MediaPipe and CNNs, highlighting the potential of gesture recognition to improve accessibility for individuals with disabilities [12]. Their system utilized a camera to capture hand gestures, which were then processed and classified to control home devices. Similarly, Kumar et al. (2023) employed MediaPipe and CNNs for real-time American Sign Language (ASL) gesture recognition, achieving an accuracy of 99.95% [13]. These studies demonstrate the feasibility and effectiveness of using deep learning techniques for gesture-based control in smart home environments.

The development of real-time gesture recognition systems has been a focal point in recent research, especially with the increasing demand for mobile and embedded platforms. Real-time systems are critical for applications in smart environments, interactive robotics, and mobile computing. Lightweight models optimized for mobile devices, such as those based on MobileNetV3 [14] and 3D CNNs, have been developed to achieve real-time performance without compromising accuracy. Yin et al. demonstrated a significant reduction in model size—by 81%—while maintaining high accuracy, making it feasible to deploy HGR on mobile devices [15]. Advanced techniques like the combination of learning-based pose initialization with physical-based model fitting have also been explored for real-time 3D hand pose estima- tion. Che et al. introduced a novel human-mobile interaction interface using a depth sensor that provided a seamless user experience in real-time environments. These approaches are particularly valuable in scenarios requiring responsive and interactive user interfaces, such as virtual and augmented reality applications [16]. The ability to process gestures in real-time and provide immediate feedback is essential for enhancing user experience in these environments.

III. Methodology

A. Dataset

In this study, the HaGRID (HAnd Gesture Recognition Image Dataset) dataset [17] is employed. HaGRID is a comprehensive dataset designed to facilitate the development of hand gesture recognition systems. It contains 554,800 images, with 18 gesture classes and one additional class for "no gesture." HaGRID's diversity in terms of subjects, scenes, lighting conditions, and subject-to-camera distances makes it ideal for training robust gesture recognition models. For evaluation purposes, we partitioned 90,000 images from the dataset, ensuring each class contains 5,000 images, for the test set.

B. Model Parameters for Classification

To enhance the model's robustness, various data augmentation techniques are applied. These techniques include: Color Jitter, Brightness and Contrast Adjustments, Grid Distortion, and Elastic Transformation. Each image is resized to 224x224 pixels and normalized using the dataset's calculated mean and standard deviation values. This pre-processing ensures that the model is trained on a diverse set of images, making it more resilient to variations in real-world conditions.

TABLE I: Comparison of Training Parameters Across Models

Parameter	MobileNetV3	EfficientNetV2	ResNet18
Activation function	Leaky Relu	Leaky Relu	Leaky Relu
Optimizer	AdamW	AdamW	AdamW
Batch size	32	32	32
Initialization	He	He	He
Learning rate	0.005	0.032	0.026
Dropout rate	0.283	0.166	0.225
L1 lambda	1e-5	1e-5	1e-5
L2 lambda	1e-4	1e-4	1e-4
Weight decay	1.09e-06	8.72e-06	6.70e-05
Criterion	Cross Entropy	Cross Entropy	Cross Entropy
Epoch	25	25	25

C. Model Parameters for Object Detection

Each image is resized to 640x640 pixels and and no other data processing was done.

TABLE II: Comparison of Training Parameters Across Object Detection Models

Parameter	YOLO-World	YOLO-V8s
Optimizer	SGD	SGD
Batch size	32	32
Learning rate	0.01	0.01
Momentum	0.9	0.9
Weight decay	0.0005	0.0005
Epoch	25	25

IV. Results

A. Results of Classification

The results presented in Table I provide a comparative analysis of various deep learning models. The models evaluated include MobileNetV3Large [6], EfficientNetV2 [18], and ResNet18 [19], with their performance assessed based on accuracy and model size. Among the models tested, MobileNetV3Large achieved the highest accuracy of 0.9905, outperforming all other models. This indicates that MobileNetV3Large effectively captures the necessary features for the task, potentially due to its architec- tural design optimized for resource-constrained devices while maintaining high performance. In terms of model size, there is a notable difference be- tween MobileNetV3Large and EfficientNetV2, with the former being significantly smaller at 16.4 MB compared to the latter's 89 MB. Despite the smaller size, MobileNetV3Large models achieve higher accuracy. This suggests that MobileNetV3Large is not only more efficient in terms of memory usage, but also more effective in performance, making it a preferable choice for applications where both accuracy and model size are critical factors. The findings indicate that for tasks requiring high accuracy with limited model size, MobileNetV3Large is a preferable choice.

B. Results of Object Detection

Table IV provides a comparative analysis of two object detection models, YOLO-World [7] and YOLO-V8 [5], across four key performance metrics: mean Average Precision at 50% IoU (mAP50), mean Average Precision averaged across IoU thresholds from 50% to 95% (mAP50-95), inference time, and model size. Both models exhibit exceptional performance in terms of detection accuracy, with YOLO-World and YOLO-V8 achiev- ing identical mAP50 scores of 0.9946. This high accuracy indicates that both models are highly effective at detecting objects with at least 50% Intersection over Union (IoU) accuracy. In terms of mAP50-95, which provides a more rigorous evaluation by averaging precision across various IoU thresh- olds, YOLO-World scores marginally higher at 0.9036, com- pared to YOLO-V8's 0.9033. Although this difference is mini- mal, it suggests that YOLO-World may offer a slight advantage in more challenging detection scenarios where higher precision is required across varying IoU thresholds. YOLO-V8, however, demonstrates superior efficiency in terms of inference speed, achieving an inference time of 1.35 ms, significantly faster than YOLO-World's 2 ms. This reduced latency is particularly advantageous for real-time applications where quick decision-making is critical. Additionally, YOLO- V8 has a smaller model size of 21.34 MB compared to YOLO-World's 24.47 MB. The reduced size of YOLO-V8 makes it more suitable for deployment in resource-constrained environments, such as on mobile devices or edge computing platforms, where memory and storage are limited. In summary, while both models deliver near-identical detec- tion accuracy, YOLO-V8's faster inference time and smaller model size make it a more suitable choice for applications requiring low-latency and efficient resource usage.

TABLE III: Results of models

Model	Accuracy	Size
MobileNetV3Small	0.9855	6.6 MB
MobileNetV3Large	0.9905	16.4 MB
EfficientNetV2	0.9867	89 MB
Resnet18	0.9800	43.8 MB

TABLE IV: Results of Object Detection Models

Model	mAP50	mAP50-95	Inference	Size
YOLO-World	0.9946	0.9036	2 ms	24.47 MB
YOLO-V8s	0.9946	0.9033	1.35 ms	21.34 MB

V. Conclusion

This research explores the application of hand gesture recognition to the field of smart home automation particularly for people with disabilities, It contributes to creating more inclusive and accessible smart home environments. By addressing key challenges in real-time processing, model efficiency, and user accessibility, the study establishes a robust foundation for future innovations in human-computer interaction and smart home technology. Our work demonstrates the potential of integrating deep learning technologies, particularly through the use of person detection and hand gesture classification using the latest versions of YOLO (i.e. YOLO-V8 and YOLO-world), within IoT ecosystems to create smarter, more accessible living environments. The proposed system, characterized by its scalability and ease of integration, offers a versatile

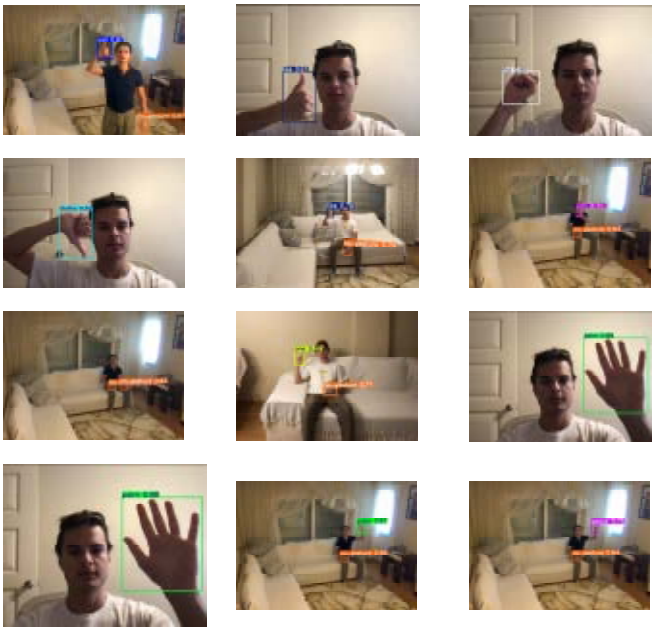


Fig. 3: Demonstration of the system for different hand gestures. The images correspond to different video frames and are processed in real-time.

solution that can be adapted across various applications, from residential to commercial and healthcare settings. The system delivers high accuracy and efficiency by reducing computational load through mobile-friendly algorithms and optimizing for edge computing, making it particularly suitable for resource-constrained environments.

In future works, we plan to explore several key areas. Firstly, we will investigate the integration of Edge Computing to enhance data privacy and system performance. This approach ensures faster processing by handling data locally, particularly focusing on its underexplored application in hand gesture recognition (HGR) systems, providing both privacy and reduced latency. Additionally, we aim to optimize real-time performance by employing lightweight deep learning models such as other versions of YOLOv8, along with compression and quantization techniques, to create a high-accuracy HGR system suitable for low-power IoT devices. Another direction is reducing false positives and negatives through a two- step verification system, incorporating movement consistency checks and auditory feedback for reliable user interaction. We will also focus on improving accuracy at varying distances and angles, overcoming current limitations by making our system robust to such changes. Lastly, we plan to develop a multi- user command prioritization system, enabling multiple users to interact simultaneously, ensuring secure and efficient operation in multi-user environments like smart homes or care facilities.

REFERENCES

[1] S. E. Ovur, X. Zhou, W. Qi, L. Zhang, Y. Hu, H. Su, G. Ferrigno, and E. De Momi, "A novel autonomous learning framework to enhance semg-based hand gesture recognition using depth information," *Biomedical Signal Processing and Control*, vol. 66, p. 102444, 2021.

[2] Y. Liu, X. Jiang, X. Yu, H. Ye, C. Ma, W. Wang, and Y. Hu, "A wearable system for sign language recognition enabled by a convolutional neural network," *Nano Energy*, vol. 116, p. 108767, 2023.

[3] J. Ding and N.-W. Zheng, "Rgb-d depth-sensor-based hand gesture recognition using deep learning of depth images with shadow effect removal for smart gesture communication." *Sensors & Materials*, vol. 34, 2022.

[4] Q. Gao, Y. Chen, Z. Ju, and Y. Liang, "Dynamic hand gesture recognition based on 3d hand pose estimation for human–robot interaction," *IEEE Sensors Journal*, vol. 22, no. 18, pp. 17 421–17 430, 2021.

[5] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLO," Jan. 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[6] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan *et al.*, "Searching for mobilenetv3," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1314–1324.

[7] T. Cheng, L. Song, Y. Ge, W. Liu, X. Wang, and Y. Shan, "Yolo- world: Real-time open-vocabulary object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 16 901–16 911.

[8] V. I. Pavlovic, R. Sharma, and T. S. Huang, "Visual interpretation of hand gestures for human-computer interaction: A review," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 19, no. 7, pp. 677–695, 1997.

[9] M. Van den Bergh and L. Van Gool, "Combining rgb and tof cameras for real-time 3d hand gesture interaction," in *2011 IEEE workshop on applications of computer vision (WACV)*. IEEE, 2011, pp. 66–72.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[12] B. Kurian, J. Regi, D. John, P. Hari, and T. Y. Mahesh, "Visual gesture-based home automation," in *2023 3rd International Conference on Advances in Computing, Communication, Embedded and Secure Systems (ACCESS)*. IEEE, 2023, pp. 286–290.

[13] R. Kumar, A. Bajpai, and A. Sinha, "Mediapipe and cnns for real-time asl gesture recognition," *arXiv preprint arXiv:2305.05296*, 2023.

[14] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[15] Y. Yin, C. McCarthy, and D. Rezazadegan, "Real-time 3d hand-object pose estimation for mobile devices," in *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021, pp. 3288–3292.

[16] Y. Che, Y. Qi, and Y. Song, "Real-time 3d hand gesture based mobile interaction interface," in *2019 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. IEEE, 2019, pp. 228–232.

[17] 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.

[18] M. Tan and Q. Le, "Efficientnetv2: Smaller models and faster training," in *International conference on machine learning*. PMLR, 2021, pp. 10 096–10 106.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.