

Research Article

Implementation and Analysis of AI-Based Gesticulation Control for Impaired People

S. Nivash,¹ E. N. Ganesh,² T. Manikandan,³ Arvind Dhaka ,⁴ Amita Nandal ,⁴
Vinh Truong Hoang ,⁵ Adarsh Kumar ,⁶ and Assaye Belay ⁷

¹Department of ECE, VISTAS, Chennai, India

²Dept. of ECE, VISTAS, Chennai, India

³Department of ECE, Rajalakshmi Engineering College, Chennai, India

⁴Department of Computer and Communication Engineering, Manipal University Jaipur, India

⁵Faculty of Computer Science, Ho Chi Minh City Open University, 97 Vo Van Tan, Ward Vo Thi Sau, District 3, Ho Chi Minh City, Vietnam 70000

⁶University of Petroleum & Energy Studies, Dehradun, Uttarakhand, India

⁷Department of Statistics, Mizan-Tepi University, Ethiopia

Correspondence should be addressed to Amita Nandal; amita_nandal@yahoo.com,
Adarsh Kumar; adarsh.kumar@ddn.upes.ac.in, and Assaye Belay; abstat23@gmail.com

Received 20 February 2022; Accepted 4 May 2022; Published 26 May 2022

Academic Editor: Kuruva Lakshmana

Copyright © 2022 S. Nivash et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This paper presents an intelligent human PC intuitive framework. In this proposed work, artificial intelligence is utilized for home mechanization, which perceives human motions with the assistance of a camera and performs tasks appropriately. The idea of perceiving the motions depends on three layers: detection, tracking, and recognition. We use a camera-PC interface that can catch the developments, and later use PC vision innovation and AI calculations to comprehend the fundamental example and match the information with a preprepared dataset. When it comes to safes, an extra layer of security is provided by using face recognition, and the safe is opened if the individual is recognized from the dataset.

1. Introduction

A number of recent studies have used sensors and gloves to recognize hand motions. The portable glove-based sensor strategy and the image vision-based sensor methodology are the two main methodologies for hand motion analysis. Wearable sensors directly placed towards the hands with the glove were previously used to identify hand gestures. The information gathered was then analyzed, and the information collected was then stored on a processor that was linked to the glove. Some of the projects addressed by [1] proposed a wheelchair where the physically unable or elderly could move around freely.

The hand sign interpreter-based communication system proposed in [2] includes flex sensors. With all due respect, despite their impressive results, the methods outlined above

may not be suitable for the elderly, who may find themselves in a state of confusion and disarray due to wire association concerns. Similarly, senior persons with chronic conditions that induce muscular atrophy may find it difficult to put on and take off gloves, which can be uncomfortable and inconvenient when worn for lengthy stretches of time. In people with sensitive skin or those who are enduring consumers, these sensors may induce skin injury, pollution, or antagonistic reactions. Also, a few sensors are very costly. Lamberti and Camastra [3] addressed some of these concerns in their experiment, which was based on a computer vision (CV) system based with colored checked gloves and may be utilized online. It was not necessary to employ sensors for this test, but the usage of colored gloves was.

It was discovered by referring to several articles that frameworks incorporating both hand signal and face motion

were not discovered. The following publications collect information on PC vision and face recognition. Murthy et al. [4] investigated the role of human-computer interaction in terms of behavior to recognition, layout, and their uses, demonstrating the control of PC vision in diverse scenarios. Khan et al. [5] published another study that introduced a recognition system that addressed the issues of highlight attribute extraction, gesture sorting, and test implementation. After a thorough examination of face detection, Bonsor [2] discusses facial skin and color models.

Accurate recognition and interpretation of hand and face motions are the primary objectives of this research and also to make it easier for those with disabilities to use gestures to control their household equipment.

Screen readers and a regular keyboard are two common methods of communicating with computers. Listening to what is being shown on the screen and writing on a keyboard with feedback whether the input is accurate are possible in this manner for blind people. With features like larger buttons that generate sound, buttons that can be felt individually and voice dialling as well as audio feedback from the phone, screen readers for mobile phones and finger gestures, it is easier for visually impaired and blind people to interact with mobile phones and smart devices. Devices that allow the user to engage with a wearable computing system are a vital part of the subject matter. Such gadgets may be integrated into clothing, worn on the wrist, or carried in the hand. A wide variety of input devices, such as the Twiddler, are now available. While these input devices are convenient for those who can see, they have limits for those who are blind or visually challenged. Finger-Braille interfaces [6] and keyboards with sound feedback [7] are among the options for interaction. In a congested space, sound-based interfaces have a disadvantage in terms of usability. Keyboard solutions need undivided attention, which interferes with daily routines.

Mobility is one of the most difficult problems to overcome for persons with impairments.

Artificial intelligence (AI) has the potential to improve accessibility in any setting, including the home. When it comes to those with impairments, a virtual assistant may make a huge difference in their quality of life. Because of the advancements in artificial intelligence (AI), persons with disabilities have access to a more inclusive environment. The world is becoming a more inclusive place as technology adapts and aids in the development of artificial intelligence accessibility. Because AI equalises people with and without impairments, there is a feeling of fairness.

The following is the analysis of the paper's formation: AI-based gesticulation control for persons with disabilities is summarised in Section 2. Section 3 entails the proposed system and methodology of the collection module and training. The paper in Section 4 has the results and discussion in terms of various performance metrics. Sections 5 and 6 are concluded with the realistic constraints and future scope of the proposed work.

2. Related Work and Background

Over the last few years, egocentric vision (also known as a first-person vision (FPV)) applications have grown in popu-

larity. The wearable camera's location (typically on the helmet) allows it to capture exactly what the wearer sees in front of them, including hands and handled items. Andrea Bandini [6] analyzes the literature on egocentric vision and hands, classifying the techniques into three categories: constrain (what happened to the hands or parts of them?), translation (what exactly are the hands up too?), and requisition (e.g., systems that solved a given issue using egocentric hand cues).

Human activity recognition based on skeletons in recent times has become a popular subject. To deal with noisy skeleton data and variations in viewpoints, using a 3D bioconstrained skeleton model, Nie et al. [7] suggest a view-invariant technique for human action identification that recovers damaged skeletons and visualizes the body-level motion data collected from the recovery process using photographs. Joint Euler angles and the Euclidean distance matrix between joints (JEDM) are two new motion characteristics used to characterize human activity. JEDM comprises the body's global structural information. Many multimedia applications take into account the modeling and identification of human activity in 3D. Despite the widespread usage of latent state techniques to represent activities, past research has assumed single-attribute latent states [8]. For expressing sophisticated action structures, this assumption is wrong. Wei et al. [8] propose a new composite latent structure (CLS) model to express and recognize 3D skeletal sequences in human activities, based on the idea that latent states have composite features. A method (action-fusion) for human action recognition utilizing depth maps and posture data was developed using convolutional neural networks (CNN). There are two input descriptors used to represent activities [9].

Three CNN channels are trained to utilize a range of inputs to maximize feature extraction for successful action classification. In order to come up with the ultimate action categorization, CNN's three networks aggregate their action forecasts. Kamel et al. [9] propose a number of fusion scoring processes to maximize the score of the right action. According to the study, integrating the outputs of three channels produces better results than using one channel or fusing two channels separately. CNNs, despite their impressive performance in image recognition, have yet to reach the same remarkable results in video motion detection [10]. In part, CNN's incapacity to simulate long-term temporal trends is to blame for this, notably if there are separate action steps involved, which are necessary for the identification of human conduct. Spatiotemporal vector of locally aggregated descriptor (ActionS-ST-VLAD) approach is proposed by Tu et al. [10] to aggregate relevant deep features throughout the full video based on adaptive video feature segmentation and adaptive segment feature sampling (AVFS-ASFS). Using an RGBF modality, it gathers action-related motion-sensitive patches in RGB images. Human action recognition relies on a domain-invariant (view-invariant, modality-invariant) feature representation [11].

The suggested MDMTL system simultaneously addresses the challenges of domain-invariant feature extraction and multitask modeling. In the field of Ambient

Assisted Living (AAL), human action recognition (HAR) is commonly used to facilitate human-computer interaction. In certain situations, people cannot be asked to act in an unnatural manner [12]. A new descriptor termed body directional velocity and real-time categorization [12] are used to accomplish this goal. Robustness to perspective shifts and rapid object recognition are two of the most important issues for robotic applications.

Rahmani et al. [13] introduce a robust nonlinear knowledge transfer model for recognizing human behavior from new angles (R-NKTM). Nonlinear transformations that link the perspectives and transmit human activity information from each unknown viewpoint are identified using the R-NKTM [13] neural network, which is a deep fully connected neural network.

In terms of cross-view action recognition, there is no better system than the R-NKTM. It is less user-friendly than R-NKTM, which is taught by using fake labels and does not need previous camera expertise. A supervised temporal t-stochastic neighbor embedding (st-tsne) and incremental learning approach for human action recognition are for human contour sequences [14]. It is suggested to comprehend the underlying relationship related to scope in a duct, wherein the target class knowledge and temporal information are offered to accurately represent those frames from the same action class [14], which is inspired by some and its adaptations. The effectiveness of explicit linear representations based on the local neighbor connection in maintaining the intrinsic action structure is investigated. A low- and memory-efficient hand segmentation framework for the detection of realistic hand motions is developed in the paper [15].

A stream-handed form tracing unit [15] and a quick contour filling unit [16] are also employed to achieve both high memory efficiency and low latency. On-chip memory is reduced by 1.68 times and latency is reduced by 1.65 times when just 34.8 KB of on-chip memory is used, resulting in a 7.14 ms latency [15]. When dealing with nonuniform blurring induced by camera tilts and rotations, existing methods for face detection depend on the convolution model [16]. In the presence of space-varying motion blur, Punnappurath et al. [16] propose a method for face detection utilizing arbitrarily shaped kernels. The hazy face is defined as a valid blend of geometrically changed versions of the focused gallery face [16], indicating that all photos obtained by nonuniformly blurring a given image form a convex set. It is an effective way to account for deviation instances as well.

The positioning of features is an important aspect of image processing. The picture information is preprocessed before any feature extraction approach is done, and several preprocessing strategies are applied to images, such as binarization, thresholding, and standardization, among others. Highlights are then removed and used for grouping reasons. Zhang et al. [17] present a method for detecting extra and stable picture highlights that makes use of symmetric qualities from visual input. The territorial highlights are framed using a subjective symmetry administrator that uses quantitative balance range data.

Hasan [18] used nongeometric features to recognize hand motions using a multivariate Gaussian distribution. The input hand image is segmented using skin color-based segmentation using the HSV (hue, saturation, and value) color model and clustering-based thresholding methods [19]. With the new direction analysis algorithm, the direction of the hand motion is employed to determine the slope and trend of the object (hand) in the data [18].

Color feature hierarchies, which assign distinct shades of color to the user's hand and the backdrop in order to detect and eliminate the background, or algorithms that regard each finger as a cluster and remove the empty spaces in between them may be used to do this. For example, finger and thumb status, skin color, finger alignments, and palm position [20] are all factors that might be taken into consideration depending on the application.

Gevers et al. [21] examine the extraction and classification of local picture structures. The majority of image processing and computer vision tasks, such as object recognition, stereo vision, and 3D reconstruction, need the algorithm to extract spatial image structure. Based on geometric and photometric data, they proposed a method for categorizing the physical nature of local image structure. Damasio and Musse [22] designed a system for recognizing hand postures that included a data glove and an artificial neural network system.

This device uses specially developed gloves with flexible sensors to collect and transmit data to a computer, allowing real and virtual individuals to interact. This problem statement system focuses to deliver a Touchless User Interface system and make the system more reliable by eliminating the dependency on sensors and thereby ease the lives of differently abled people [23, 24]. Also, enhance the security of certain appliances using face recognition. The current systems for user applications have only hand gesture recognition systems which are complex and unwieldy. Thus, the proposed system yields higher accuracy and less operating time using the customized CNN model (GestureNet) and Mobile FaceNet model, tackling the problem of huge model files that are generated while processing [25, 26].

3. Proposed System and Methodology

Smart homes use top developments, for example, progressed light components and security frameworks to carry solace and comfort to your living experience. As savvy houses become more known and the mechanical advances are all the more broadly utilized, possessing a shrewd home has its benefits.

This proposed framework goes about as an AI helper which makes a difference in the automation of home appliances and helps meet various ends together. Through this framework, the utilization of facial acknowledgment and hand motion to help the differently abled also makes it a superior spot to live for the visually impaired. This framework is utilizing computer vision; this tech has been utilized for human PC collaboration. HCI utilizes a physical medium where hand motion and facial acknowledgment play a significant job. Hand motions have been done since the

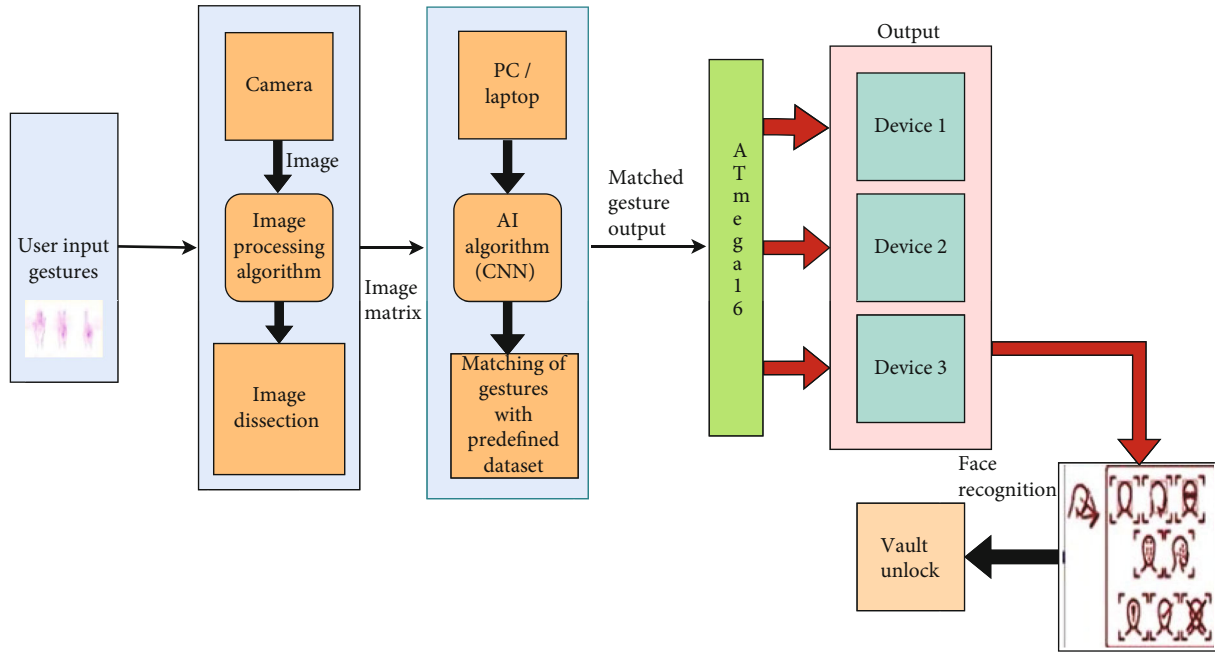


FIGURE 1: Functioning of the pattern.

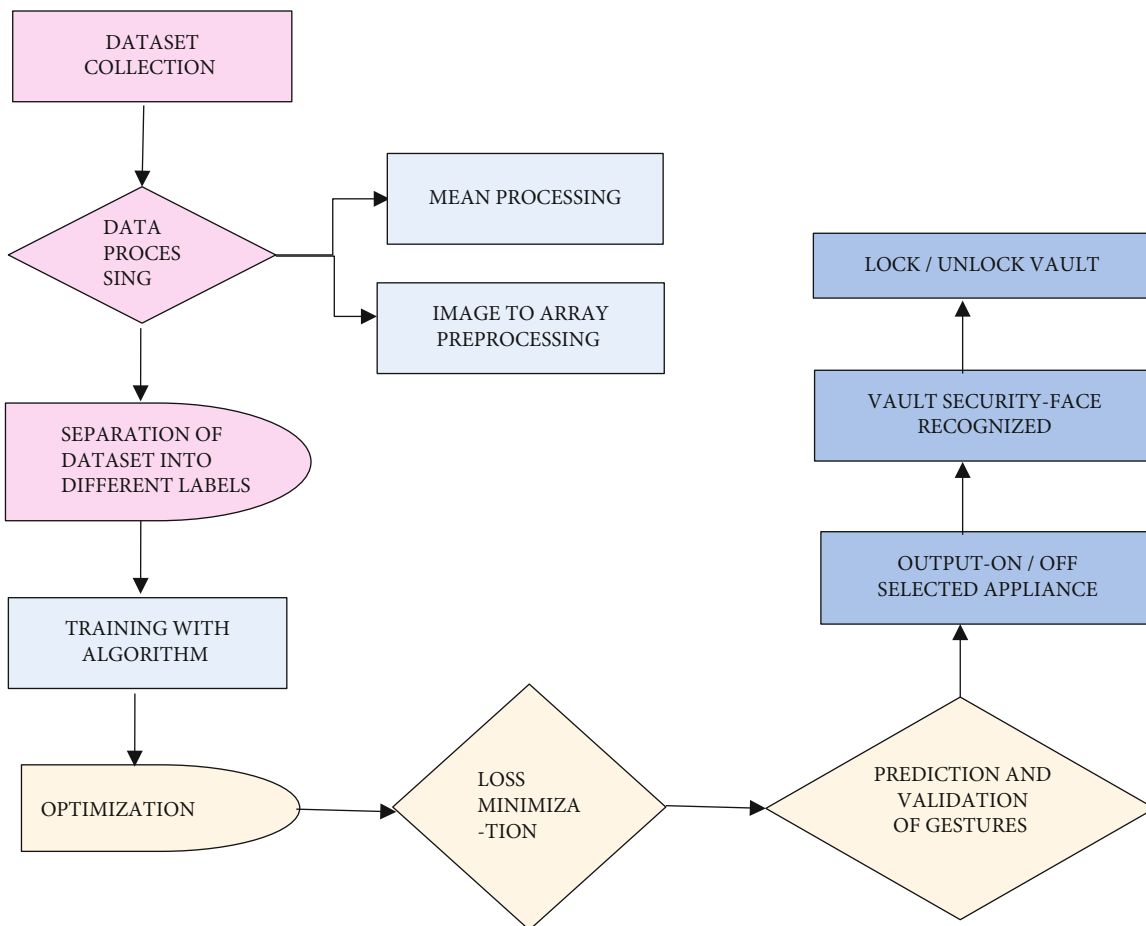


FIGURE 2: Flow of the proposed system.

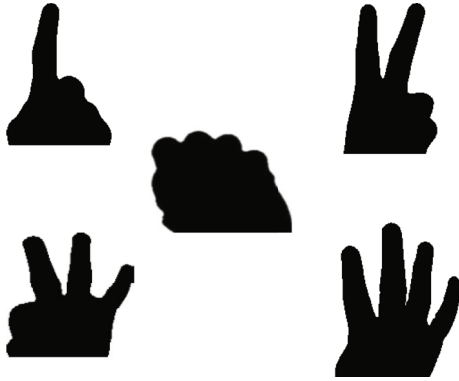


FIGURE 3: Primary gestures of the dataset.

TABLE 1: A number of datasets and training data.

Gesture	No. of data	Trained data
One	100	75
Two	100	75
Three	102	76
Four	101	76

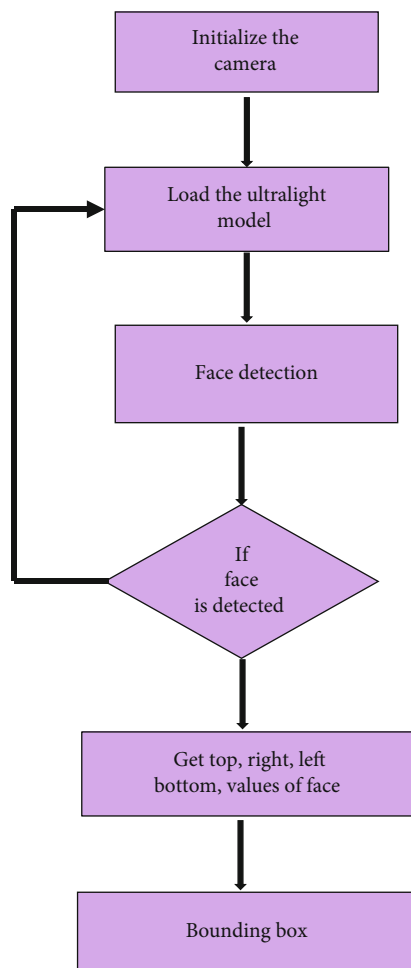


FIGURE 4: Flow diagram of ultralight face detector.

beginning of human advancement and have different significance relying upon the topographical area. PC vision has additionally been created like the wearable gloves, yet they are excessively expensive as they need sensors and other equipment gadgets that require various algorithms for hand signal acknowledgment like KNN (K Nearest Neighbor). The majority of the algorithms require an enormous sum of databases for training the model.

Despite the fact that the tactics mentioned above have yielded excellent outcomes, they have different constraints that make them unacceptable for many due to various reasons. Due to these shortcomings, promising and practical technologies that did not need the use of heavy gloves were developed. Camera vision-based sensor enhancements are the name for these methods. Thanks to the open-source software's rise frameworks for coding, identifying hand signals that may be used for a variety of applications is now easier than ever.

From Figure 1, it is illustrated that the proposed framework is to help the differently abled to work the home apparatuses without any human assistance. In this proposed work, artificial intelligence is utilized for home computerization, which perceives human gestures with the assistance of a camera and performs tasks likewise. The idea of perceiving the gestures depends on three layers: detection, tracking, and recognition. We utilize a camera-PC interface that can catch the developments and later use computer vision innovation and AI calculations to comprehend the fundamental gesture and match the information with a predefined dataset. Face detection is used in safes to add an extra layer of security, and the safe is unlocked if the person is identified from the database. This proposed system centers around conveying an auto home apparatus control utilizing hand motions by utilizing Touchless User Interface (TUI) to facilitate hands-free control of gadgets and furthermore to upgrade the security of specific apparatuses utilizing face recognition.

3.1. Data Collection Module. Working with deep learning projects necessitates a massive amount of data, as AI models cannot be prepared without it. Figure 2 illustrates that gathering and setting up the dataset are perhaps the most urgent part while making an AI project. The innovation applied behind any deep learning project cannot work as expected if the dataset is not decidedly ready and pre-handled. During the advancement of the work, the engineers totally depend on the datasets. The proposed system mainly consists of a dataset collected in the following methods along with real-time data obtained while training the GestureNet module.

- (i) Scraping from websites: this includes manually checking and retrieving images from the Internet, which takes a lot of time
- (ii) Third party: because data is such an important resource in the deep learning period, many startups have decided to sell their own picture sets. Third-party datasets are what they are called

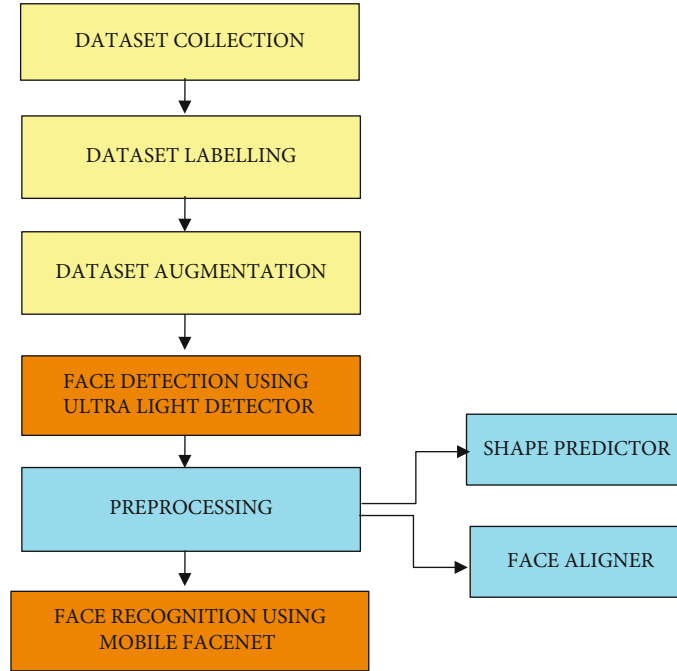


FIGURE 5: Face recognition system.

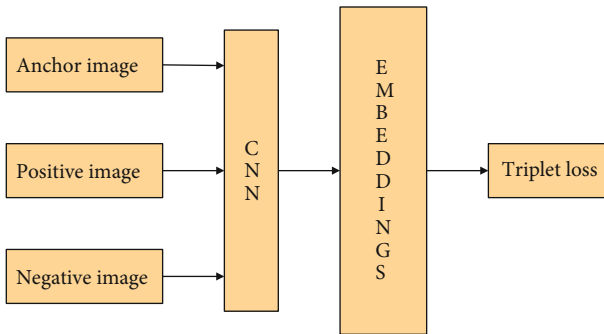


FIGURE 6: Mobile FaceNet architecture.

(iii) Figure 3 shows the primary gestures of the dataset, and Table 1 depicts the number of datasets and training data. Dataset for the proposed system can be obtained from various sources like Kaggle, which offers access to numerous free datasets. There are 24000 images in total in this collection, representing 20 distinct types of gestures. There are 900 images in each directory for training purposes and 300 images in each directory for testing purposes

3.2. Face Detection Module. The “ultralight weighed module” is designed for universally successful face recognition applications in low operating devices, including Android and iOS phones as well as Desktop computers. It is often used in limited computational devices, including Advanced RISC Machines (ARM), for continuous normal scene faces. This model is a lightweight face detection model intended for edge figuring devices.

- (i) Regarding model size, the default floating point 32 exactness (.pth) record size is 1.04 1.1 MB, and the surmising structure int8 quantization size is about 300 KB
- (ii) Regarding the estimation measure of the model, the information goal of 320×240 is about 90 109 MFlops
- (iii) There are two forms of the model, version-slim (slightly quicker) and version-remote frame buffer (RFB) (with the altered RFB module, higher exactness)
- (iv) Wider face preparing the pretraining framework with various information goals of 320×240 and 640×480 is given to work in different situations

3.3. Face Recognition Module. Figure 4 shows the flow diagram of ultralight face detector. Face recognition is majorly used in threat detection, monitoring, human-computer interaction, and interference, among other applications. The first step in facial recognition is to interpret human faces in computer-aided images, and an absolute monarch discovery model can be judged by how quickly and accurately it does so. In this proposed system, we are making use of an ultralight face detector which is a light-weighted face detector for face detection that helps to reduce the model file generated. Figure 5 shows the flowchart for face recognition system. For face recognition, we are using the Mobile FaceNet algorithm which stands out in performance.

Figure 6 shows the Mobile FaceNet architecture where FaceNet takes a picture of the individual’s face as info and yields a vector of 128 numbers which address the main high-lights of a face. In AI, this vector is called embeddings.

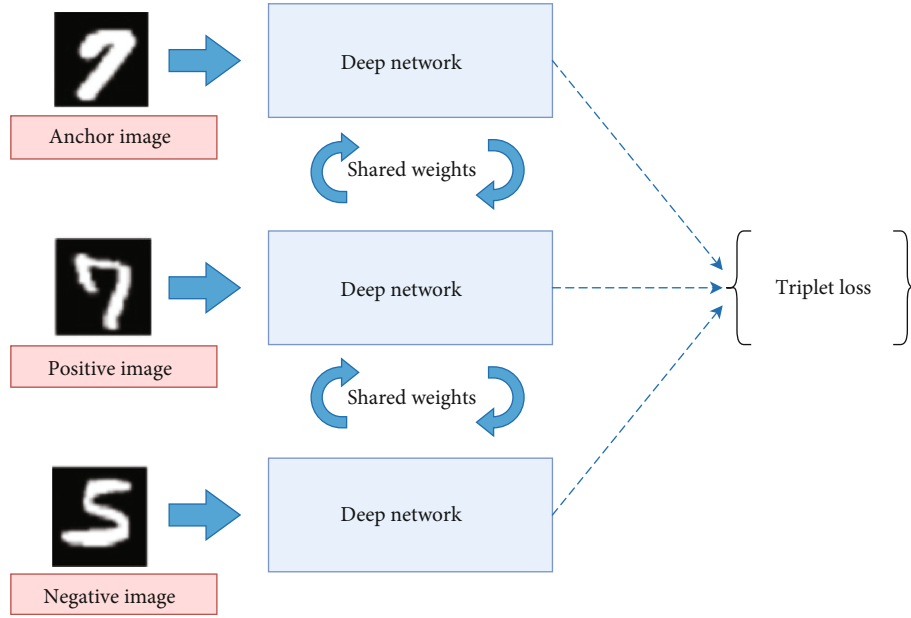


FIGURE 7: Working of triplet loss.

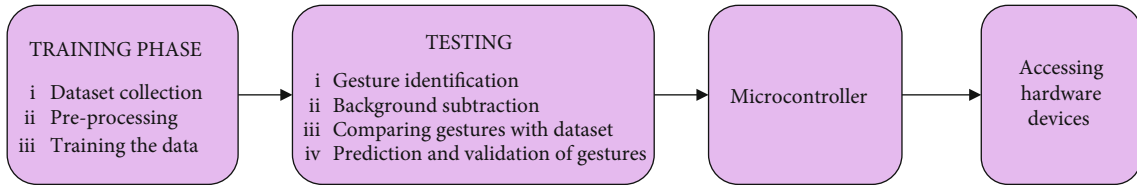


FIGURE 8: Hand gesture recognition system.

Embeddings are vectors and can decipher vectors as focused in the Cartesian facilitated framework. That implies that it can plot a picture of a face in an arranged framework utilizing its embeddings. FaceNet needs a lot of pictures of content. A similar rationale can be applied in the event that there is a huge number of pictures of various individuals. Toward the start of preparation, FaceNet produces irregular vectors for each picture which implies the pictures are dispersed haphazardly when plotted.

The FaceNet model deploys the following steps:

- (i) Arbitrarily chooses an anchor picture
- (ii) Arbitrarily chooses a picture of a similar individual as the anchor picture (positive model)
- (iii) Arbitrarily chooses a picture of an individual not the same as the anchor picture (negative model)
- (iv) Changes the FaceNet network boundaries with the goal that the positive model is nearer to the anchor than the negative model

The instinct behind the triplet loss formula is that we need our anchor (picture of a particular individual A) to be nearer to positive pictures (every one of the pictures of individual A) when contrasted with negative pictures (the wide range of various pictures). Figure 7 shows the working

of triplet loss. As such, we can say that we need the distances between our anchor picture and the embeddings of our positive pictures to be lesser when contrasted with the distances between implanting of our anchor picture and embeddings of our negative pictures.

$$\text{Loss} = \sum_{i=1}^N \left[\left\| f_i^a - f_i^p \right\|_2^2 - \left\| f_i^a - f_i^n \right\|_2^2 + \alpha \right]_+ \quad (1)$$

Using a deep convolutional neural network, FaceNet can recognize faces (CNN). It is trained such that the distance between the embeddings corresponds to face similarity by adjusting the squared L_2 distance. Scaling, transformation, and close cropping of the faces are all part of the training process. FaceNet's loss function is a key part of the overall system design. Triplet loss function is used.

Unlike previous systems, FaceNet does not need a bottleneck layer for recognition or verification. Instead, it learns the mapping from the photos and builds embeddings. All subsequent tasks, such as verification and identification, may be carried out using the freshly produced embeddings as the feature vector, using the conventional methodologies of the specific domain. When it comes to face identification, verification, and clustering, FaceNet does not come up with

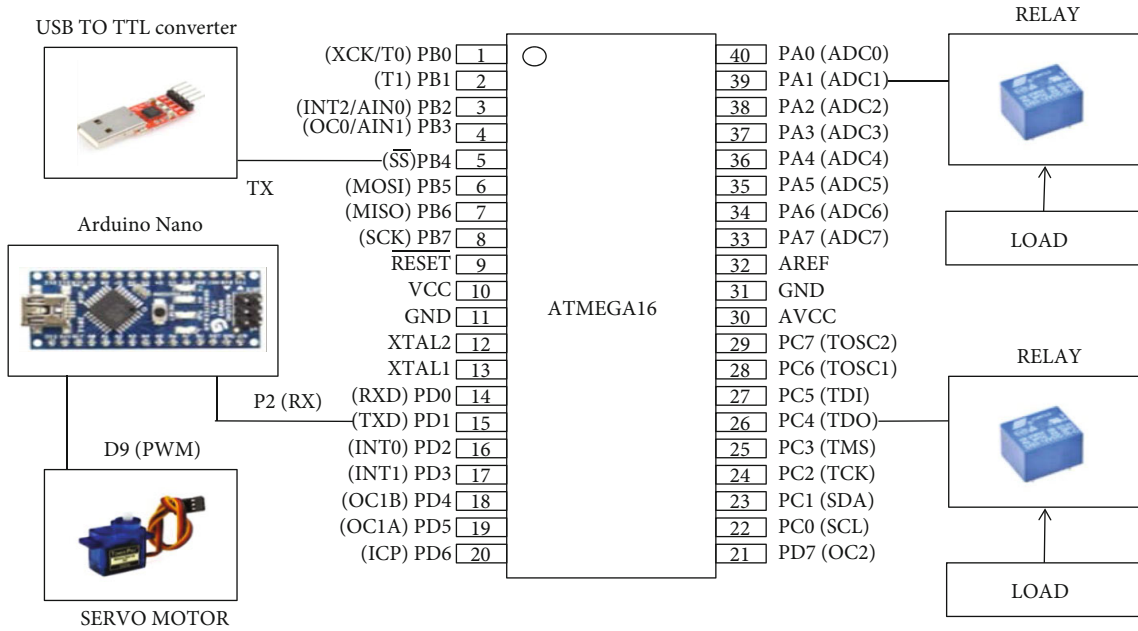


FIGURE 9: Hardware interface design.

TABLE 2: Comparison of gestures for YOLO Face.

Gesture	Precision	Recall	F1 score	Support
Four	0.93	1.00	0.96	25
Ignore	0.99	1.00	1.00	27
One	0.99	1.00	1.00	26
Three	0.99	0.92	0.96	25
Two	0.99	1.00	1.00	25

any novel algorithms, but rather provides the embeddings that may be utilized directly.

3.4. Module to Segment Hands. As a result of difficulties associated with vision-based operations such as changing lighting conditions, complex base, and skin shading analysis, shades of human skin are diverse and composition necessitated an aggressive improvement in computation for a regular interface. When it comes to identifying objects, shade is extremely useful. As a result, shading data, which is unaffected by pivot and hand math variation, was used for the division. Shader qualities like brilliance, immersion, and tint section are more noticeable to humans than the red, green, and blue levels of basic shading. Color models are useful for consistently expressing a given tone. It is a space-composed framework in which a single point corresponds to any shade preset. Three approaches for powerful hand recognition and division were provided here, each utilizing various shading areas. For the preprocessing of the hand gesture recognition (HGR) framework, the high-throughput screening (HTS) method utilizing HSV shading space is differentiated.

3.5. GestureNet Training. The camera receives inputs in the form of photographs of hand motions. Figure 8 shows the

hand gesture recognition system. Using the foundation deduction technique, it is easy and effective to separate the hand district from the initial image. The hand area can be distinguished from the rest of the moving objects using the skin tone. To forecast skin tone, HSV analysis is employed. The HSV estimations of skin color are 315, 94, and 37. The image of the recognized palm is reformatted to create the signal affirmation identical to image scores. The hand recognition produces a two-dimensional picture, with white pixels representing individuals from the hand region and dark pixels representing individuals from the foundation.

The accompanying system is then used to fragment the fingers and palm of the parallel hand image. The figuring spot is a mark in the palm's center. The distance shift method is used to investigate it. In a distance transform image, each pixel calculates the distance between itself and the nearest border pixel. The distance between the pixels and the closest limit pixels is calculated using the measured distance. The linear image's central point has the greatest distance. The location of the green tone distinguishes the detected site. We describe a novel method called GestureNet in this paper. When video streaming is available, this method is used to train diverse motions with excellent accuracy, and the model is recognized automatically.

To recognize movements, a photo of a static hand making a single movement on a white background in well-lit conditions is optimal. However, in the real world, such occurrences are very uncommon. In certain cases, it is difficult to utilize clear, solid backdrops while demonstrating motions. Some of the most severe technical issues in gesture detection may be addressed by using machine learning.

3.5.1. Backgrounds. Gesture recognition should be useful no matter where you are: on the road, at home, or just walking down the street. It is possible to educate the algorithm to

TABLE 3: Input and outputs for each specific gesture.

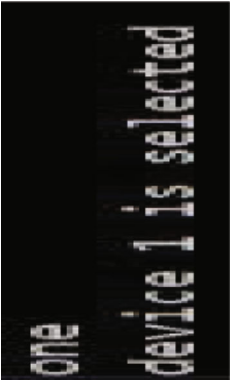
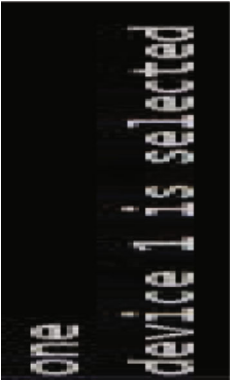


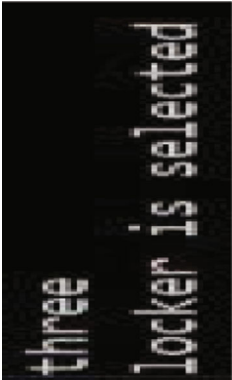
Inputs	Outputs	Working
		Gesture "1" activates device one.
		Gesture "2" activates device two.
	Gesture "3" activates device three, which is the locker.	

TABLE 3: Continued.



Inputs	Outputs	Working
	<pre>one device 1 is selected full device 1 is on</pre>	Gesture "4" or "full" serves the purpose of switching the selected/activated device on.
	<pre>three locker is selected zero locker is off</pre>	Gesture "0" serves the purpose of switching the selected/activated device off.

TABLE 3: Continued.

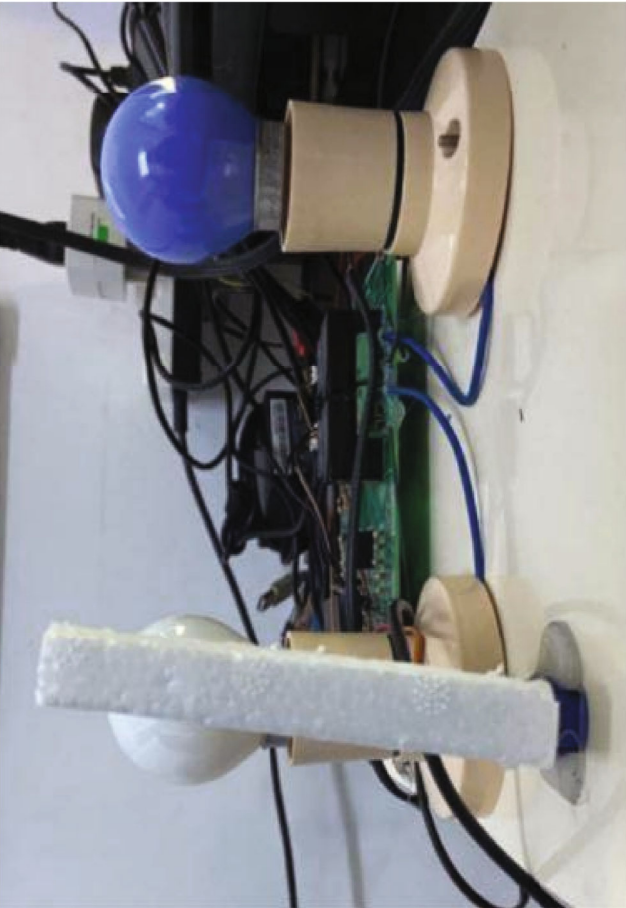
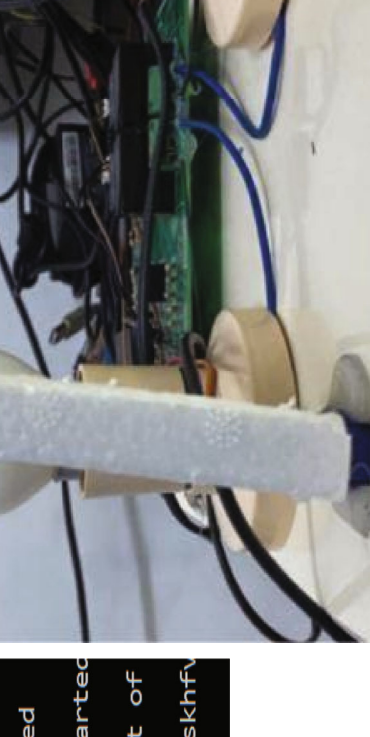
Inputs	Outputs	Working
<pre> three locker is selected full face scanning started full show face infront of 0.37572426 .SruthisdjvbsfdbvskhfV locken is open </pre>		<p>Once device three is activated and switched “on,” face recognition is initiated. When the trained person’s face is recognized, the locker opens.</p>
<pre> 0.6513307 unknownsdjvbsfdbv theft detected </pre>		<p>When an unknown face is detected, a theft alert is sent to the user.</p>

TABLE 4: Comparison of multiple face detectors with respect to time consumption.

Face detectors	Accuracy (%)	Computational time (s)	References
YOLO Face	98.9	0.105	Proposed methodology
Face recognition CNN	96	109	Moon et al. [27]
MTCNN	96.8	3.00	Chaves et al. [28]
Ultralight detector	97.2	0.102	Punnappurath et al. [16]

TABLE 5: Comparison of multiple face detectors with respect to the generated model file.

Face detectors	Model file size	References
DSFD	456 MB	Chaves et al. [28]
Pyramid	215 MB	Earp et al. [29]
Ultralight detector	8 MB	Punnappurath et al. [16]

consistently discriminate between the hand and its background using machine learning techniques.

3.5.2. Activation. A gesture, by its very nature, is not a still image but rather one that is always in motion. You may observe the wave activity and interpret it as a command to terminate the currently used software instead of recognizing an open palm image. As a result, pattern recognition should be possible for gesture recognition.

3.5.3. Movements. When a pattern like moving fingers clockwise and showing one's thumb is utilized to indicate the presence of a specified number of files or an area on a computer screen, further context and pattern identification are required, for example.

3.5.4. Diversity. There is a great deal of variation in the way people carry out certain actions. Even while we humans have a great tolerance for mistakes, this inconsistency may make it more difficult for robots to recognize and classify gestures. This is where machine learning comes in.

3.5.5. Struggling with Latency. The time it takes to classify a gesture should be as short as possible for the system that detects them. Only a constant and quick use of hand signals should be expected. If gestures do not speed up and simplify your interactions, there is simply no use in employing them. To provide feedback to the user as quickly as possible, we are aiming for a negative latency (classification even before the action is completed).

3.6. Circuit Design. In Figure 9, hardware interface design is represented. The figure gives a clear idea to assemble a connection between the gesture motions and the home devices that we need a hardware equipment interface that joins both PC and the home devices. The PC is connected to an ATmega16 microcontroller, a 40 pin MC based on enhanced RISC architecture, through a USB to TTL converter module.

Relays are used to switch on the devices according to the input gesture. Arduino Nano is connected to the microcontroller, to provide a PWM signal, for operating the servo

motor, which is used to depict the opening and closing of vaults.

4. Results and Discussions

More than 20 degrees of freedom are available for articulation in the hand. Various hand positions, locations, and orientations need the estimation of a wide range of hand characteristics. It is difficult to estimate because of the hand's large degree of freedom (DoF). Joints and finger segments that are occluded in monocular vision make it difficult to see the shape of a hand at all. These constraints necessitate the use of gestures that do not need complete hand position information in vision-based interface (VBI) applications. In unrestricted contexts, estimating an articulated hand position is still mainly an open topic. Gesture recognition faces a huge issue in dealing with occlusion. It is possible for one hand to occlude the other when making two-handed motions, as well. For example, the look of the hand might be distorted by occlusion. Viewpoint affects the look of gesturing hands in monocular vision-based gesture detection. The self-occlusion in Table 2 makes distinct hand positions seem identical from a specific perspective. An appropriate model must be used to portray a gesture in order for it to be recognized. A gesture is either static or dynamic based on its spatiotemporal features. The stance or orientation of a body component in space (such as the hand pose) characterizes a static gesture, while the temporal deformation of body parts characterizes a dynamic gesture (e.g., shape, position, and motion). The extraction and selection of appropriate characteristics to represent linked gestures are a key stage in gesture recognition after gesture modeling.

Table 3 represents all the input and corresponding output actions that are triggered by the gestures. Gestures "1," "2," and "3" are used to select the respective devices. Gesture "4" is used to switch the selected device on. Gesture "0" is used to switch the selected device off. When device 3 is switched on, face recognition is activated. When the trained face is recognized, device 3 is opened. In case an untrained face is detected, a theft alert is generated.

The following outcomes can be obtained by training the model with a newly devised algorithm. Table 4 displays the accuracy of each gesture after being trained by the module. The formula for calculating the parameters is as follows. Here, the terms "true positive" and "true negative" come from the confusion matrix, which lists the number of gestures predicted correctly and incorrectly during the testing phase of the algorithm. "True positive" represents the number of datasets accurately labeled as positive for a gesture. The number of datasets successfully analyzed as the negative

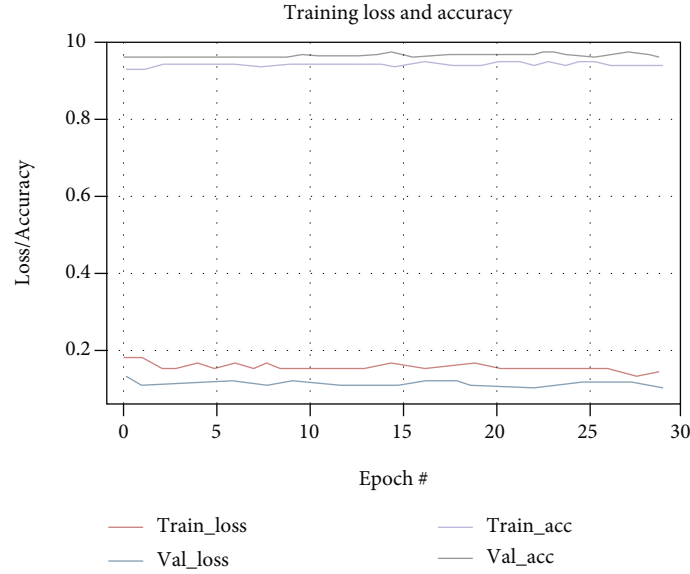


FIGURE 10: Accuracy plot for YOLO.

of a gesture is referred to as the “true negative.” The term “false positive” refers to photographs that have been mislabeled as positive when they are in fact negative.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}, \quad (2)$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}. \quad (3)$$

As a result, precision is calculated by dividing the fraction of “true positives” by the total estimated positives, both true and false. A recall is defined as the proportion of “true positives” to “real positives,” which includes both true positives and false negatives. Both these parameters are good measures to select a model’s performance and to check for under- or overfitting.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \quad (4)$$

where TP = true positive, TN = true negative, FP = false positive, and FN = false negative.

The most basic metric is accuracy, which is specified as the proportion of exactly predicted observations to total perceptions. Accuracy is a fantastic metric, but only when you have symmetric datasets with the upsides of false positives and false negatives nearly equal. As result, you must consider multiple boundaries in order to evaluate the model’s results.

$$F1 = 2 \times \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5)$$

F1 score is a function of precision and recall, which can be used to attain a balance between the mentioned parameters. Accuracy is highly influenced by the number of true positives, not taking false negatives and false positives into

account. Thus, the *F1* score comes into play when seeking a balance between true and false predictions. True response samples are counted as support for a class. A classifier’s reported scores may be affected if the training data is unbalanced or if stratified sampling or rebalancing is required. The proposed YOLO Face is compared with the existing methods such as face recognition CNN, MTCNN, and ultralight detector as shown in Table 3.

Table 4 compares the different face detector modules and the respective time consumed, in seconds, to recognize, validate, and identify the face. YOLO Face has higher accuracy and effective execution time. Ultralight detectors dominate in speed and relatively better accuracy than all other models while generating significantly low model file size. Face recognition, the primitive model, consumes the most time with good accuracy. MTCNN takes less time compared to face recognition and is very robust. It can be used in advanced applications.

From Table 5, it can be seen that the ultralight detector generates a smaller model file, thus making it ideal for mobile applications. The smaller model file results in multiple advantages like boosting speed and accuracy and thus making it much more efficient for the proposed system’s application compared to the other models.

Table 2 shows the simulated results of precision for various gestures for the YOLO Face detector. Figure 10 shows the trends of loss and accuracy of the model over the training epochs. train_loss and val_loss represent the loss occurring during training and validation of the dataset after each epoch, which remains less than 20%.

The complete dataset is divided into two sections: 75% for training and 25% for testing. Validation is set for 25% of the training set, out of a total of 75%. The graphic also shows the train_acc and val_acc values, which represent the dataset’s accuracies during training and validation over the training epochs. Throughout training and evaluation, it stays above 90%.

5. Realistic Constraints and Future Scope

The realistic constraints of this project are that at times it might take some time to recognize the gesture and perform necessary action. Video is transient and therefore difficult to review or edit and interferes significantly with other tasks. Once an error is made, the whole program must be invoked again. Another constraint is accuracy, where 100% accuracy is logically unattainable due to significant noise in the user's end. The proposed system can be further improved by integrating voice along with a camera for a holistic model, improving the accuracy of recognition by using ensemble methods, activating an alert message to the caretaker in case of emergency, integrating several other functions like fall detection.

Users must be expert and well-versed in a variety of hand gestures to utilize these software apps. Each application makes use of a separate set of hand motions since there are so many potential combinations imaginable. Hand gesture recognition is also impacted by the surrounding environment (such as light, backdrop, range, and tone color) and the hand's location and posture. Vision-based gesture recognition classifiers now available cannot keep up with all of the new classification challenges that are arising. Every one of them has a downside that reduces overall efficiency. Static and DG appearances are more difficult to represent because of the hand's highly articulated structure. The identification process is made more challenging by the variation in gesture characteristics owing to the spatiotemporal variation in hand positions.

6. Conclusion

Face validation and hand signal validation are consolidated in this framework which offers different applications and assists the differently abled to access the devices and furthermore forestall theft. The proposed system can be deployed in retirement homes owing to the simplistic nature of operations. The improved security also enables the safekeeping of valuables. The performance of the proposed technique exceptionally relies upon the consequence of hand and face identification. These applications are typically used in environmental compliance, and algorithms usually may take advantage of the inherent limitations to achieve high accuracy.

Data Availability

Data can be made available from first author on request.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgments

This work is supported by Universidad Autónoma de Zacatecas, Mexico, and CONACYT, Mexico.

References

- [1] S. Nasif and M. A. G. Khan, "Wireless head gesture controlled wheel chair for disabled persons," in *IEEE Region 10 Humanitarian Technology Conference (R10-HTC)*, pp. 156–161, Dhaka, Bangladesh, Dec 2017.
- [2] P. Vishal Bhujbal and K. K. Warhade, "Hand sign recognition-based communication system for speech-disabled people," in *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*, pp. 348–352, Madurai, India, June 2018.
- [3] L. Lamberti and F. Camastra, "Real-time hand gesture recognition using a color glove," in *International Conference on Image Analysis and Processing*, pp. 365–373, Springer, Heidelberg, Berlin, 2011.
- [4] H. P. Gupta, H. S. Chudgar, S. Mukherjee, T. Dutta, and K. Sharma, "A continuous hand gestures recognition technique for human-machine interaction using accelerometer and gyroscope sensors," *IEEE Sensors Journal*, vol. 16, no. 16, pp. 6425–6432, 2016.
- [5] M. Van den Bergh, D. Carton, R. De Nijs et al., "Real-time 3D hand gesture interaction with robot for understanding directions from humans," in *2011 Ro-Man*, pp. 357–362, Atlanta, GA, USA, July 2011.
- [6] A. Bandini and J. Zariffa, "Analysis of the hands in egocentric vision: a survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2020.
- [7] Q. Nie, J. Wang, X. Wang, and Y. Liu, "View-invariant human action recognition based on a 3D bio-constrained skeleton model," *IEEE Transactions on Image Processing*, vol. 28, no. 8, pp. 3959–3972, 2019.
- [8] P. Wei, H. Sun, and N. Zheng, "Learning composite latent structures for 3D human action representation and recognition," *IEEE Transactions on Multimedia*, vol. 21, no. 9, pp. 2195–2208, 2019.
- [9] A. Kamel, B. Sheng, P. Yang, P. Li, R. Shen, and D. D. Feng, "Deep convolutional neural networks for human action recognition using depth maps and postures," *Ieee Transactions On Systems, Man, And Cybernetics: Systems*, vol. 49, pp. 1806–1819, 2018.
- [10] Z. Tu, H. Li, D. Zhang, J. Dauwels, B. Li, and J. Yuan, "Action-stage emphasized spatio-temporal VLAD for video action recognition," *IEEE Transactions on Image Processing*, vol. 28, pp. 2799–2812, 2019.
- [11] A. A. Liu, N. Xu, W. Z. Nie, Y. T. Su, and Y. D. Zhang, "Multi-domain multi-task learning for human action recognition," *IEEE Transactions on Image Processing*, vol. 28, pp. 853–867, 2019.
- [12] S. A. W. Talha, M. Hammouche, E. Ghorbel, A. Fleury, and S. Ambellouis, "Features and classification schemes for view-invariant and real-time human action recognition," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 10, pp. 894–902, 2018.
- [13] H. Rahmani, A. Mian, and M. Shah, "Learning a deep model for human action recognition from novel viewpoints," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, pp. 667–681, 2018.
- [14] J. Cheng, H. Liu, F. Wang, H. Li, and C. Zhu, "Analysis for human action recognition based on supervised temporal t-SNE and incremental learning," *IEEE Transactions On Image Processing*, vol. 24, pp. 3203–3217, 2016.

- [15] C. Sungpill, P. Seongwook, and Y. Hoi-Jun, "A memory-efficient hand segmentation architecture for hand gesture recognition in low-power mobile devices," *Journal of Semiconductor Technology and Science*, vol. 17, pp. 473–482, 2017.
- [16] A. Punnappurath, A. N. Rajagopalan, S. Taheri, R. Chellappa, and G. Seetharaman, "Face recognition across non-uniform motion blur, illumination, and pose," *Ieee Transactions on Image Processing*, vol. 24, no. 7, pp. 2067–2082, 2015.
- [17] K. Huebner and J. Zhang, "Stable symmetric feature detection and classification in panoramic robot vision systems," in *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 3429–3434, Beijing, China, Oct 2006.
- [18] M. Mokhar Hasan and M. K. Pramod, "Features fitting using multivariate Gaussian distribution for hand gesture recognition," *International Journal of Computer Science & Emerging Technologies IJCSET*, vol. 3, no. 2, pp. 73–80, 2012.
- [19] M. M. Hasan and P. K. Mishra, "Robust gesture recognition using Gaussian distribution for features fitting," *International Journal of Machine Learning and Computing*, vol. 12, pp. 266–273, 2012.
- [20] S. Ananyaa, K. Ayush, K. Kavleen, J. Shivani, U. Richa, and P. Sameer, "Vision based static hand gesture recognition techniques," in *2017 International Conference on Communication and Signal Processing (ICCSP)*, pp. 0705–0709, Chennai, India, April 2017.
- [21] T. Gevers, S. Voortman, and F. Aldershoff, "Color feature detection and classification by learning," *IEEE International Conference on Image Processing*, 2005, Genova, Italy, September 2005, 2005.
- [22] F. Damasio and S. Musse, "Animating virtual humans using hand postures," in *Proceedings of the XV Brazilian Symposium on Computer Graphics and Image Processing*, pp. 10–10, Fortaleza, Brazil, October 2002.
- [23] S. Bhat and D. Koundal, "Multi-focus image fusion techniques: a survey," *Artificial Intelligence Review*, vol. 54, no. 8, pp. 5735–5787, 2021.
- [24] S. Bhat and D. Koundal, "Multi-focus image fusion using neurosophic based wavelet transform," *Applied Soft Computing*, vol. 106, article 107307, 2021.
- [25] H. Kaur, D. Koundal, and V. Kadyan, "Image fusion techniques: a survey," *Archives of Computational Methods in Engineering*, vol. 28, no. 7, pp. 4425–4447, 2021.
- [26] K. Bhalla, D. Koundal, S. Bhatia, M. Khalid, I. Rahmani, and M. Tahir, "Fusion of infrared and visible images using fuzzy based Siamese convolutional network," *Cmc-Computers Materials & Continua*, vol. 70, no. 3, pp. 5503–5518, 2022.
- [27] H. M. Moon, C. H. Seo, and S. B. Pan, "A face recognition system based on convolution neural network using multiple distance face," *Soft Computing*, vol. 21, no. 17, pp. 4995–5002, 2017.
- [28] D. Chaves, E. Fidalgo, E. Alegre, R. Alaiz-Rodríguez, F. Jáñez-Martino, and G. Azzopardi, "Assessment and estimation of face detection performance based on deep learning for forensic applications," *Sensors*, vol. 20, no. 16, p. 4491, 2020.
- [29] S. W. Earp, P. Noinongyao, J. A. Cairns, and A. Ganguly, "Face detection with feature pyramids and landmarks," <http://arxiv.org/abs/1912.00596>.