Design and Implementation of IOT Based Face Detection and Recognition

Vijay Gaikwad, Devanshu Rathi, Vansh Rahangdale, Rahul Pandita, Kashish Rahate, Rajendra Singh Rajpurohit

Computer Department, Vishwakarma Institute of Technology, Pune, India

Corresponding author: Vijay Gaikwad, Email: vijay.gaikwad@vit.edu

In contemporary society, the demand for accurate and efficient face detection in images and videos has grown significantly, driven by applications in surveillance, education, autonomous driving, and healthcare. Challenges such as unconstrained pose variations, occlusions, a large number of faces, and varying illumination conditions have posed formidable hurdles for existing face detection methods. In response, this study introduces a novel Depth wise Separable Convolution Block (DSCB) that not only maintains training speed but also enhances accuracy. Leveraging the proposed DSCB, a face detection model based on Multi-task Convolution Neural Network (MTCNN) is designed to tackle challenges related to occlusion, unconstrained pose variations, and numerous small targets. Compared to the original MTCNN, the proposed face detection method showcases substantial performance improvements and achievements. Furthermore, the paper delves into the realm of face recognition, an essential biometric technology that identifies individuals based on facial features. Traditional face recognition methods were plagued by slow processing speeds and lower accuracy compared to manual recognition. However, the advent of deep learning and Convolutional Neural Networks (CNNs) has revolutionized face recognition. Since the impact of pandemic moving towards physical interaction free life has become a norm. Also, with advancement in technology, recognition of individual via face has also spread its root in various daily use aspects of life like mobile phones, security safe, etc. This project aims to advance the home security systems via use of facial recognition. By addressing the associated challenges and concerns, the project aims to contribute to the ongoing development and adoption of smart doorbell IoT systems, ultimately leading to safer and more connected homes.

Keywords: MTCNN, Face Recognition, face Detection, Smart Doorbell, IoT.

2024. In Satyasai Jagannath Nanda & Rajendra Prasad Yadav (eds.), *Data Science and Intelligent Computing Techniques*, 923–933. Computing & Intelligent Systems, SCRS, India. https://doi.org/10.56155/978-81-955020-2-8-78

1. Introduction

In today's ever-evolving world, the security of personal information and the protection of our homes have taken on a new level of importance. With the continuous advancements in technology and the widespread adoption of the Internet of Things (IoT), digital door locks have emerged as a popular and convenient security solution. These locks are part of a broader trend where cutting-edge image processing techniques like Haar cascade classifiers, Multi-task Cascaded Convolutional Networks (MT-CNN), and You Only Look Once (YOLO) play a pivotal role in object detection and recognition within images and videos. These technologies are not merely tools; they mirror our human capacity to identify faces and significantly enhance the intelligence and responsiveness of IoT systems.

In response to these technological advancements and the growing need for more sophisticated security solutions, our project was conceived. We set out to create a real-time IoT application embedded with facial recognition capabilities. At its core, this system operates by comparing live scans of individuals' faces with a pre-established database. It effectively acts as a digital gatekeeper, granting access only to those individuals it recognizes as authorized. This innovative solution brings together the power of modern face detection technology and the convenience of IoT in a seamless union. Traditional access control methods, such as keys or passcodes, are rendered obsolete by our system's efficiency and user-friendliness.

Imagine a world where you can enter your home or office without fumbling for keys or recalling passcodes—a world where your face serves as the key to your security. This is the essence of our IoT-based facial recognition project. It signifies a significant leap forward toward a safer and more convenient future. By harnessing the latest in technological innovation, we are reshaping the landscape of security and access control. We are embracing the opportunities presented by a rapidly evolving digital age and enhancing the way we protect our personal spaces. Text styles are provided. The formatter will need to create these components, incorporating the applicable criteria that follow.

2. Literature Survey

With the advancement in technology in current industry several algorithms for face detection are available. As such a recent study by Kirti Dang and Shanu Sharma provide with review and comparison of existing prominent Face Detection Algorithms. Algorithm are compared based on the precision recall value calculated using a DetEval Software and the value of precision and recall is highest for Viola-Jones followed by SMQT Features [1]. Thus, stating that Viola-Jones or Haar Cascade algorithm is best whilst also supporting low computational abilities.

Another method for detection of faces is discussed in Sung and Paiggio research, face detection method is quite like previous work but in addition Multilayer perceptron classifier and distribution-based is also implemented. Each picture is vectored in 361 dimensions after each face example has been processed and normalised as a 19*19-pixel image and further grouped into six face clusters and six non-face clusters using a modified K-means algorithm. Each image is depicted using multidimensional Gaussian function with a mean image and covariance matrix and the normalised Mahala Nobis distance and the Euclidian distance are computed, then a multilayer perceptron network is employed to categorise windows with and without faces leading to face detection [2]. While the accuracy does increases so does the computational power thus making the proposed model limited to its niche.

Ishita Gupta, Varsha Patil, Chaitali Kadam, and Shreya Dumbre related works in the field follows face detection using Haar Cascade an already proven as a better alternative for such project alongside PCA for face recognition. Positive and negative pictures are produced model is trained. After getting the confidence value, the capture.pgm, positive.png, negative.png, and mean.png files are tested against an existing flag at least ten times before it can recognize the proper face [3].

While algorithms are being studies, the most powerful being Blazeface is also a well-recognized algorithm but limited application in the industry due to it computational needs reaching that of a mobile device at the minimum. In a research paper by Valentin Bazarevsky, Yury Kartynnik, Andrey Vakunov, Karthik Raveendran and Matthias Grundmann titled "BlazeFace: Sub-millisecond Neural Face Detection on Mobile GPUs", they proposed a deep learning model for real-time face detection on mobile devices with limited computational resources called Blazeface [4]. Blazeface is lightweight neural network architecture that can process images in sub-milliseconds. This model outperforms existing state-of-the-art face detection algorithms in terms of speed and accuracy, making it suitable for various applications, such as video conferencing, augmented reality, and mobile games.

Another similar work published as "EFFICIENT HUMAN IDENTIFICATION THROUGH FACE DETECTION USING RASPBERRY PI BASED ON PYTHON-OPENCV" by Lochan Basyal, Bishal Karki, Gaurav Adhikari, and Jagdeep Singh followed use of Haar cascade algorithm with OpenCV for face detection. The project was performed on Raspberry PI 3 Model B and SQLite Studio was used to store information related to user and data pertaining feature-related information used for identification of face was stored in phpMyAdmin. Dataset.py, Trainer.py and Detector.py depicts the python code and for the performance first we need to take input data samples of images and then these datasets are analyzed by trainer.py code and as a result this will be converted into its respective file in the format of Trainer.yml [5].

Q. Wu, Y. Liu et. al. [6] suggest study on Deep Learning as the deep learning exhibits strong advantages in the feature extraction, it has been widely used in the field of computer vision and among others, and gradually replaced traditional machine learning algorithms. This paper first reviews the main ideas of deep learning, and displays several related frequently-used algorithms for computer vision. Afterwards, the current research status of computer vision field is demonstrated in this paper, particularly the main applications of deep learning in the research field.

Travis Williams, Robert Li [7] explored the widespread use of machine learning, particularly Convolutional Neural Networks (CNN), for image classification in areas like business and medicine. Employing wavelet domain processing on the CIFAR-10 and KDEF databases enhances accuracy and efficiency. The method, dividing image data into subbands, facilitates feature learning across frequencies, resulting in improved detection accuracy compared to spatial domain CNN and Stacked Denoising Autoencoder (SDA). The findings underscore the effectiveness of this approach for precision in diverse applications.

Yi Wang; Xinwei Duan et. al. [8] provided a feasible face recognition algorithm based on CNN method and TensorFlow deep learning framework is proposed for face multi-pose and occlusion recognition problems. Firstly, a face detection model was established. Under the TensorFlow learning framework, the MTCNN face detection model was established and trained with Wider Face data set. Secondly, the face recognition model was established, the Facenet algorithm model was established, the CASIA-Webface dataset was used for training, and the LFW was used for model testing. The test results on LFW data set show that the face recognition accuracy of the proposed algorithm can reach 98.78 % .

P Viola, M Jones [9] his paper describes a visual object detection framework that is capable of processing images extremely rapidly while achieving high detection rates. There are three key contributions. The first is the introduction of a new image representation called the "Integral Image" which allows the features used by our detector to be computed very quickly. The second is a learning algorithm, based on AdaBoost, which selects a small number of critical visual features and yields extremely efficient classifiers [4]. The third contribution is a method for combining classifiers in a "cascade" which allows background regions of the image to be quickly discarded while spending more computation on promising object-like regions.

3. Methodology

We have taken the MALF and ImageNET dataset. MALF dataset released in 2015 is the first face detection dataset that supports fine-gained evaluation with the 5250 images with 11,931 identities. We evaluated several algorithms, including YOLO (You Only Look Once), Haar Cascade, MTCNN (Multi-task Cascaded Convolutional Networks), and YOLO on this dataset. We evaluate our method on the ImageNet 2012 classification dataset that consists of 1000 classes. The models are trained on the 1.28 million training images, and evaluated on the 50k validation images. We also obtain a final result on the 100k test images, reported by the test server. We evaluate both top-1 and top-5 error rates [10].

Our goal was to choose the algorithm that would provide the best accuracy and reliability for our system. We began by selecting the algorithms mentioned above and implementing them. We then fine-tuned these algorithms using a diverse dataset. This dataset included images captured in different lighting conditions and featured individuals from various demographics, including age, gender, and ethnicity. We made sure the dataset also covered a range of facial expressions, poses, and partial face occlusions to create a comprehensive testing environment.

YOLO, known for its speed, fell short in terms of face detection accuracy, particularly in challenging scenarios involving varying lighting conditions and occlusions. Haar Cascade, a machine learning-based method, displayed decent performance but lacked robustness in handling diverse conditions. Despite YOLO's impressive speed, it struggled with accurate face detection, especially in challenging situations with lighting variations and occlusions. While Haar Cascade showed promising results, its performance lacked robustness when faced with diverse conditions and scenarios

MTCNN, on the other hand, consistently outperformed the other algorithms in terms of accuracy. It excelled in detecting faces across different scales, orientations, and lighting situations. However, it did come with a slightly longer computation time compared to the other algorithms. When we revisited YOLO, we found that it still maintained its speed advantage but didn't significantly enhance its face detection accuracy over our initial evaluation.

Our experiments led us to conclude that MTCNN is the most suitable algorithm for our IoT-based face detection and recognition system. Its consistent accuracy and robustness, especially in challenging conditions, make it the right choice for real-world applications. While it may require a bit more computation time, this trade-off is justified by the enhanced accuracy it brings to the table. In summary, our choice of MTCNN aligns perfectly with our project's goal of achieving high accuracy and reliability in face detection, ensuring enhanced security and convenience for our users [8]. In cases where no match is found, the system classifies the individual as an unknown entity and forwards a notification, with an image of the unidentified person. Moreover, if the person remains unidentified, they have the option to engage in real-time video communication with the homeowner.

4. Implementation

Real life implementation of this algorithm which we have used is for the doorbell system where we are providing a systematic procedure for identifying visitors and granting them access to a building using facial recognition technology. Upon detecting a visitor's face, the system verifies its authenticity and proceeds to check the visitor's authorization status. If authorized, the system unlocks the door, allowing entry. In case of unauthorized access, the system notifies the building owner via a message. If the owner approves, the system initiates a live video call or a direct call with the visitor, followed by announcing the visitor's name and unlocking the door. If the owner declines, the system refrains from granting access and closes the lock. In instances where face detection or authorization fails, the system attempts to retry after a brief delay. Additionally, if unable to contact the owner, the system securely

locks the door. This flowchart demonstrates a comprehensive and secure approach to visitor identification and access control, ensuring the safety and integrity of the premises.

A. Components –

1. ESP32-CAM – The ESP32-CAM is a very small camera module with the ESP32-S chip. It is used as camera for face detection and recognition.
2. FTDI Programmer – The ESP32-CAM does not come with inbuilt a USB connector, so an FTDI programmer is needed to upload code.
3. Arduino IDE – Arduino IDE is used to program the FTDI programmer.
4. Servo motor (Lock) – Here, servo motor represent lock and rotates at 90 degrees or 180-degrees to represent opening and closing of the door.
5. Ultrasonic Sensor – Ultrasonic sensor is used to measure the distance between any object and camera to prevent any intentional damage by as an alerting using buzzer.
6. Buzzer – It is used to alter the owner with a buzz sound.
7. Arduino-It acts as an external power supply for servo motor (3.3V) as ESP-32 cam module requires high device power which leads to other components to use an external power supply.

B. Flowchart –

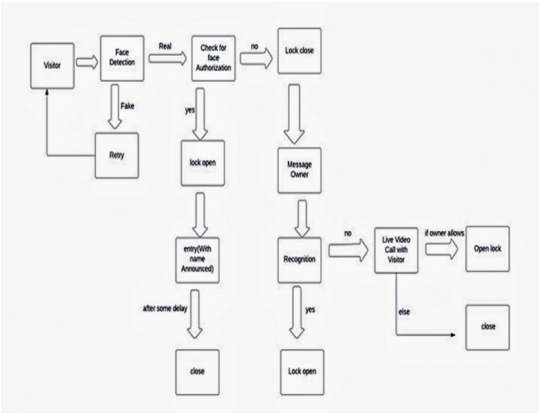

Figure 1. Flowchart of Doorbell System

Stepping into a building secured by a facial recognition visitor access control system feels like entering a familiar friend's place, only it remembers your face instead of your name. As discussed in Figure 1. a camera scans your features and instantly recognizes authorized visitors and grant them smooth entry as you approach it. For unfamiliar faces, the system adapts. If you've previously registered, a quick name confirmation unlocks the door. Newcomers have options: request a video call with the resident for approval, or wait politely while the system reverts to its secure state, keeping the door firmly locked. This automated guardian ensures authorized access while offering flexibility for unexpected guests, all with the added security of automatic re-locking for peace of mind. When the program is executed, a HTTP link is generated acting as an application interface which is entered into a browser. This interface connects user to the camera. MTCNN and Mobile Net is implemented in ESP32-CAM to recognize the faces.

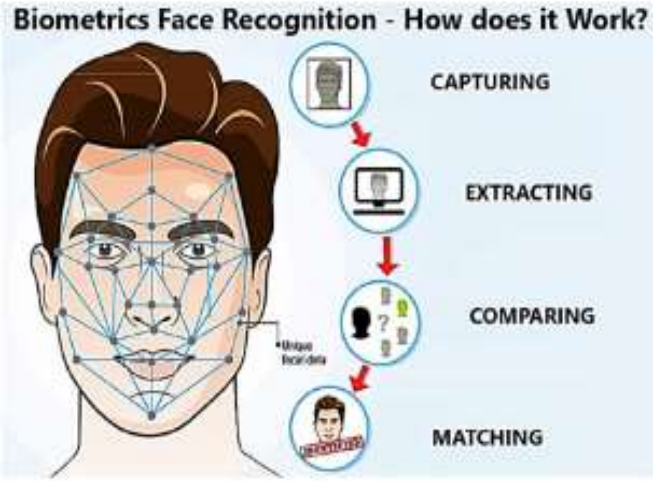

Figure 2. Biometric Face recognition

Figure 2. shows us the process of biometric face recognition involving four key steps: capturing, extracting, comparing, and matching. Firstly, a camera captures an individual's facial image. Subsequently, the image undergoes analysis to extract unique facial features, such as eye distance and jawline shape. The extracted features are then compared to a database of known faces, aiming to find a match. If a match is identified, the person's identity is confirmed. Biometric face recognition is a security technology used for personal identification, utilizing facial features. Other biometric software includes voice, fingerprint, and eye recognition. Commonly employed in security and law enforcement, facial recognition has diverse applications, including real-time identification in photos and videos [11].

5. Design

Many Algorithms are used in face detection such as Blaze Face, Har Cascade, MTCNN (Multi Convolutional Neural Network), YOLO v6, etc. This Project have implemented MT CNN which is designed to tackle challenges related to occlusion, unconstrained pose variations, and numerous small targets.

MTCNN is a deep learning architecture designed specifically for face detection tasks. Here's an elaboration on its role and significance within the context of the project. In the realm of computer vision and facial recognition, MTCNN is a pioneering technology. It is utilized as a fundamental building block in the project to enhance the accuracy and efficiency of face detection. MTCNN is particularly well-suited for scenarios where faces exhibit various challenges, such as unconstrained pose variations, occlusions, varying lighting conditions, and the presence of a large number of faces within an image or video frame. MTCNN (Multi-task Cascaded Convolutional Networks) was chosen for its exceptional accuracy in real-world face detection. It excels in recognizing faces with varying poses, occlusions, and lighting conditions, making it ideal for IoT-based facial recognition. Its real-time capabilities and efficiency in handling multiple faces ensure robust and responsive security solutions.

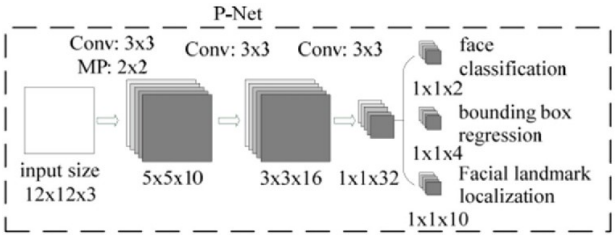

Figure 3. Architecture of MT-CNN

Deep learning is a byproduct of artificial neural network development [6]. The training begins with MLPs (Multi-layer Perceptron) as shown in Figure 3. by adding a linear layer from the input of the network connection to the output [7]. Deep learning may create a good approximation of a complicated function by increasing the number of hidden layers; hence, it can reach astonishing results in face recognition.

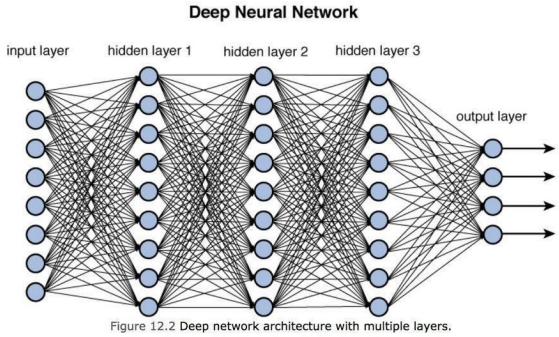

Figure 4. Architecture of Deep CNN

The proposed face detection and recognition system is a comprehensive security solution following the architecture discussed in Figure 4. and incorporating various components interconnected using jumper wires and a breadboard. To implement the facial recognition algorithm, the system generates a unique URL link for accessing the ESP32-CAM module. This device initiates the security process by detecting the presence of a human face and subsequently transmitting the image data to the recognition software. The recognition process works on a database of pre-stored facial data, employing sophisticated facial recognition techniques. When a match is successful, the system identifies the individual, granting access through the door via a 180-degree servo motor unlocking mechanism. Additionally, the system sends a notification to the homeowner, notifying them of the entry.

In cases where no match is found, the system classifies the individual as an unknown entity and forwards a notification, with an image of the unidentified person. Moreover, if the person remains unidentified, they have the option to engage in real-time video communication with the homeowner.
To fortify the security of the ESP32-CAM and associated hardware components, an ultrasonic sensor is integrated with the camera system, connected to an audible alert system in the form of a buzzer. This sensor actively monitors the proximity of individuals to the camera. Should an individual attempt to tamper with the camera or employ an unauthorized approach to unlock the door, resulting in a distance less than 20 cm from the camera, the buzzer is activated, serving as an alert mechanism to notify the homeowner of potential intrusion or tampering.

The selection of the ESP32-CAM is underpinned by its superior capability to handle image-based recognition tasks, surpassing the limitations often encountered when using pixel-based images from mobile devices, thus elevating the system's accuracy and security. The confusion matrix is used to validate the accuracy. Its evaluated using –

$$\left(\frac{(TN+T\ )}{Total}\right) \times 100\% \quad (1)$$

TN is true negative; TP is true positive.

6. Results

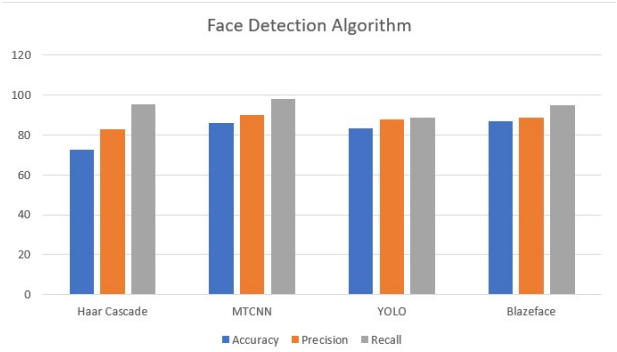

Figure 5. Comparison between Face Detection Algorithm's

The graph presented in Figure.5. compares the performance of four face detection algorithms: Haar Cascade, MTCNN, YOLO, and BlazeFace. The evaluation metrics used are accuracy, precision, and recall. Accuracy measures the proportion of correctly detected faces, precision quantifies the fraction of detections that are true faces, and recall assesses the percentage of faces that are identified. Haar Cascade exhibits the highest precision, indicating its effectiveness in minimizing false positives. However, it falls short in recall, implying that it may miss some faces. MTCNN strikes a balance between precision and recall, outperforming Haar Cascade in face detection but also being more prone to false positives. YOLO boasts the highest recall, successfully detecting most faces, but it suffers from the lowest precision, leading to a higher likelihood of false positives. BlazeFace, a newer algorithm, prioritizes speed and efficiency. It achieves a balance between precision and recall, but its accuracy is not as high as some of the other algorithms. We have gone for MT CNN as it has balanced all the parameters.

The proposed system leveraged an ESP32-CAM camera and utilized the Arduino IDE for the implementation of face detection and recognition algorithms. It demonstrated its capability to trigger alerts when unauthorized individuals attempted to gain access to a premises through any illegitimate means. The system underwent rigorous testing across a range of scenarios to assess its accuracy and effectiveness. Impressively, under optimal conditions, the system achieved facial detection in less than 3 seconds. Furthermore, the system was designed with user-friendliness in mind. It offered customizable settings, such as brightness and saturation adjustments, to cater to the specific preferences and requirements of users.

To ensure the reliability and robustness of the proposed face recognition system, a comprehensive testing protocol was devised. This protocol involved the systematic inclusion of a diverse dataset comprising multiple photographs, featuring individuals either in groups or individually. Importantly, each individual within this dataset had to be represented in the photos a minimum of 10 times. This stringent criterion was employed to guarantee a comprehensive and varied coverage of facial features and expressions, enabling a thorough assessment and validation of the system's accuracy and recognition capabilities.

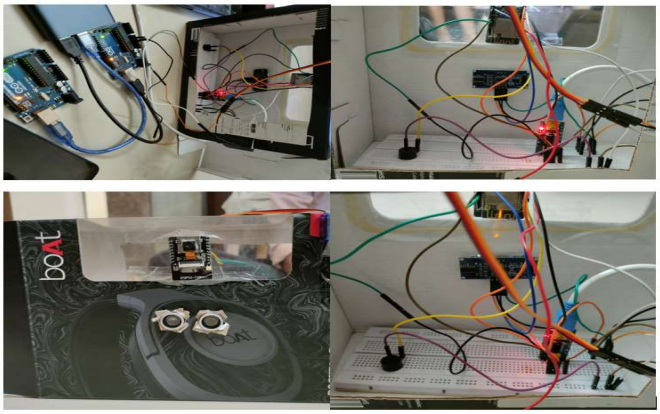

Figure 6. Assembled Hardware

Figure 6. above is just the hardware assembled using the ESP-32 Camera module, Ultrasonic Sensor, male and female wires, and breadboard. Figure 7. below is the final illustration of the model which we have made to detect any anomaly for the doorbell used at the home which detect the object in front of it and send the URL to the user to overlook for the object outside the door.

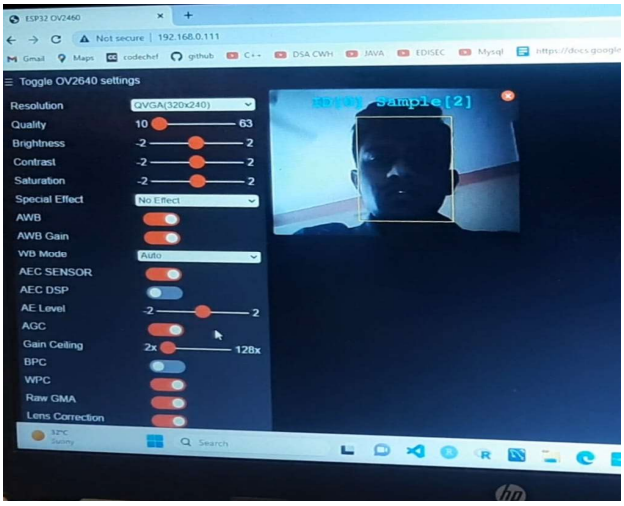

Figure 7. Camera Results

Table 1. Image Recognition Matrix

Number of Face Recognition			
	1st person	2nd person	3rd person
1st Person	8	0	2
2nd Person	1	9	0
3rd Person	1	1	8

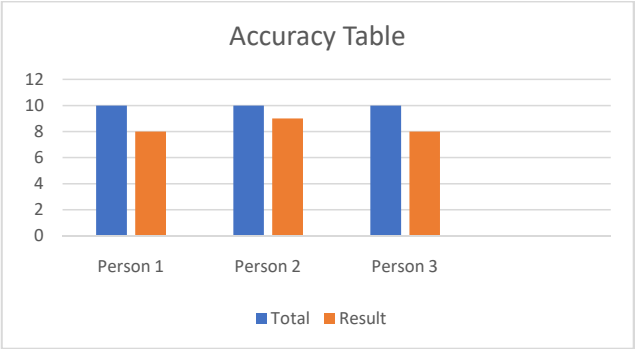

Figure 8. Accuracy Table

After deployment of this model Figure.8., Table 1. some results discussed below the bar chart shows the accuracy of three peoples after training with person database, with the average accuracy of all three people at 85%. Person 2 has the highest accuracy, with 90%, followed by Person 1 with 80% and Person 3 with 75%. This information can be interpreted in a number of ways. First, it is clear that Person 1 is the most accurate of the three people.

Second, it is also worth noting that Person 3 has the lowest accuracy. This could be a cause for concern and various conditions that matter like lighting conditions and many other factors.

7. Conclusion

We have used MT-CNN for this proposed model as it has the best-balanced parameters among all other algorithms, it also allows to do the changes it's layers according to the implementation. The combination of face detection, real face verification, and authorization checks enables a robust system that distinguishes authorized visitors from unauthorized individuals. The ability to notify and communicate with building owners further strengthens the security protocol, ensuring that only permitted individuals gain access to the premises. The system's ability to retry failed attempts and its fail-safe mechanism of locking the door in case of communication failures ensures its reliability and resilience. Overall, the presented approach demonstrates a comprehensive and secure solution for visitor identification and access control, safeguarding the safety and integrity of buildings.

Our project does have some limitations at the moment. For instance, the accuracy of face detection drops in low-light conditions, although it improves when the flash is on. In the future, we can consider adding 3D detection to prevent false recognition from images.

We can also think about introducing features like sending a video call link to the homeowner when an unknown person is at the door. This would allow the homeowner to communicate with the visitor and

enhance security. Similarly, we could implement a feature where the system greets family members when they arrive. These additions would make the system even more useful and user-friendly.

References

[1] Dang, K. and Sharma, S., 2017, January. Review and comparison of face detection algorithms. In 2017 7th International Conference on Cloud Computing, Data Science & Engineering-Confluence (pp. 629-633). IEEE.

[2] Sung, Kah & Poggio, Tomaso. (1998). Example Based Learning for View-Based Human Face Detection. Pattern Analysis and Machine Intelligence, IEEE Transactions on. 20. 39 - 51. 10.1109/34.655648.

[3] Gupta, Ishita & Patil, Varsha & Kadam, Chaitali &Dumbre, Shreya. (2016). Face detection and recognition using Raspberry Pi. 83-86. 10.1109/WIECON-ECE.2016.8009092.

[4] Bazarevsky, V., Kartynnik, Y., Vakunov, A., Raveendran, K. and Grundmann, M., 2019. Blazeface: Sub-millisecond neural face detection on mobile gpus. arXiv preprint arXiv:1907.05047.

[5] Basyal, L., Karki, B., Adhikari, G. and Aulakh, J.S., 2018. Efficient human identification through face detection using raspberry PI based on Python-OpenCV.

[6] Q. Wu, Y. Liu, Q. Li, S. Jin, and F. Li, "The application of deep learning in computer vision,"

[7] Williams, T., & Li, R. (2018). An ensemble of convolutional neural networks using wavelets for image classification. Journal of Software Engineering and Applications, 11(2), 69-88.

[8] Proc. - 2017 Chinese Autom. Congr. CAC 2017, vol. 2017–Janua, pp. 6522–6527, 2017.

[9] Viola (2001) Robust Real-Time Object Detection. International Journal of Computer Vision, 57, 87.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," IEEE Conf. Comput. Vis. Pattern Recognit., Dec. 2015

[11] P. Kasemsumran, S. Auephanwiriyakul, and N. Theera-Umpon, "Face Recognition Using String Grammar Nearest Neighbor Technique," Journal of Image and Graphics, Vol. 3, No. 1, pp. 6-10, June 2015.