A Real Time Face Recognition and Tracking Framework Using Lightweight Convolutional Neural Network

Aseel Wadood Majeed*, *Shaimaa Hameed* Shaker and *Ali Adel* Saeid

Department of Computer Science,University of Technology, Baghdad, Iraq

Abstract. Human face recognition and tracking (FRT) plays a vital role in various fields, including security, authentication, and human-computer interaction. The main modules of the FRT system are detection, feature extraction, and FRT. Using a database, these units recognize faces as well as their location, movement, and visible features. The framework aims to process large visual data in real-time, enabling accurate and fast FRT. The paper develops a real-time FRT framework using a lightweight CNN convolutional neural network to accurately match images of faces and environments with different illumination and expression differences to improve performance. This paper focused on the development of real-time facial recognition and tracking systems. The model used to achieve this is based on deep learning (DL) using a lightweight convolutional neural network (CNN) and post-feature extraction using linear discriminant analysis (LDA). Histogram of Oriented Gradients (HOG) experiments demonstrate that DL with lightweight CNN models provides a good solution for FRT tasks, even in challenging situations including changes in position, expression, illumination, and occlusion. The results of CNN-based DL were compared with several experiments. The model was also compared with many modern methods and achieved better results. The lightweight CNN model for DL outperformed it 100% of the time. When the split rate is 70:30 and the learning rate is 0.001, the epoch is 100. This demonstrates the dominance of DL over other techniques and shows how well it handles FRT tasks using lightweight and even real-time CNN methods.

1 Introduction

FRT in real-time is a technology used to identify and human FTs in live video streams or images, used in surveillance, security systems, human-computer interaction, and social media to improve accuracy and generalization capabilities. However, there are still challenges in recognizing and tracking individuals in videos due to factors like pose variations, lighting conditions, occlusions, and complex backgrounds, especially when multiple people need to be tracked simultaneously in real-time, which requires significant computational resources [1].

In a video-based FR system, there are four main modules: face detection, RT, feature extraction, and FR. Face detection determines the position and pose of faces, while FT follows their movement over time. Feature extraction locates facial features and gathers relevant information, and the FR module identifies or verifies faces using a database [2].

Real-time FT includes the task of locating and keeping up with a video stream or an image series that contains a person's forehead, represented by robust algorithms for detecting faces to deal successfully with lighting conditions, position and scale, etc. Faces are tracked over time by techniques such as optical flow or Kalman filtering to estimate their place and heading. Unlike general object tracking, FT has its own set of challenges it needs to overcome namely losing the target due to changes in appearance or occlusions, computational efficiency problems when dealing with multiple targets and difficulties related with data association over several cameras. Another requirement is that FT should be resistant to rapid head movements, have the ability simultaneously track a presentable number of faces even when they turn away and come back again [1] [3].

The paper presents the proposed system in detail in five sections. Section 2 presents related works. It then continues with Section 3 and its sub-sections, which present the preprocess for detecting the faces and algorithms used for feature

* Corresponding author: Cs.21.21@grad.uotechnology.edu.iq

extraction. Thereafter follows the DL and lightweight CNN for tracking faces that are used in real-time. Section 4 goes through testing and evaluation. Section 5: FRT Based on Real-Time Video shows the combination of several fields in a single framework. Lastly, Section 6: final Conclusion.

2 Related Works

In this section, various related works dealing with FRT. This will involve presenting existing works, to discover what is studied and considered as solution to such issues.

G. Hiten et al. in 2021 [4] they proposed face mask detection for still and real-time videos that classifies image into “with mask” and “without mask”. Their model is trained and evaluated with the “Kaggle” dataset. The combined dataset consisted of approximately 4,000 images and achieved an accuracy 98%. Their proposal model is computationally efficient and accurate in comparison to “DenseNet-121”, “MobileNet-V2”, “VGG-19” and “Inception-V3”. their proposal has been used as “a digital survey tool in schools, hospitals, banks, airports, and many other public or commercial locations”.

K. Aly et al. in 2022 [5] they suggest a FRT framework for challenging environments using lightweight CNNs for detection, alignment, and feature extraction. The framework includes RetinaFace, ArcFace, and a face tracking (FT) algorithm for improved processing time and accuracy. Tested in real-time experiments, it achieves high precision, recall, and F-score. The framework is implemented as a modular ROS package for easy integration into human-robot interaction systems.

S. Shaimaa et al. in 2022 [6] they suggest system to detect a person's gender and age using Linear-Discriminate Analysis and color facial images. The Iterative “Dichotomiser3” algorithm is used to classify individuals based on their gender and age. “The Face-Gesture-Recognition-Research-Network (EGRRN)” aging dataset is used, with facial images categorized into binary categories using k-means. The division process divides samples based on age classes for each specific sex category. The proposed method achieves an accuracy of 90.93% and an F-measure of 89.4. The study uses the EGRRN aging dataset.

B. Anil et al. in 2023 [7] suggest that this task shall be performed through the classification of masked images and feature extraction by using “Caffe-MobileNetV2 (CMNV2)” model, which involves face detection with a convolutional architecture for the fast feature embedding Caffe model, and mask identification is performed using “MobileNetV2”. In order to increase the accuracy of classification using a few training parameters for data provided related to face mask detection, they implemented five distinct layers on “MobileNetV2 architecture” within their work. Experimental results have shown that the suggested methodology was good in producing high accuracy on real-time video images and an image of 99.64% accidentally correct.

Z. Wang et al. in 2023 [8] They suggest three different kinds of masked face datasets: the “Real-world Masked FR Dataset (RMFRD)”, “the Simulated Masked FR Dataset (SMFRD)”, and “the Masked Face Detection Dataset (MFDD)”. RMFRD is now the biggest real-world masked face dataset accessible among them. The development of numerous applications for masked FR is made possible by the open access to these datasets by both academics and industry. The created multi-granularity masked FR model outperforms findings given by the industry, achieving a 95% accuracy rate.

Table 1. Summary of Related Works

Ref.	Method	Dataset	Difference to proposed work	Result
G. Hiten et al. [4]	Face mask detection model	Kaggle	The study detected faces for still and real-time videos that classify images into “with mask” and “without mask”.	accuracy rate of 98%
K. Aly et al. [5]	FR system with tracking algorithm	N/A	The study detected faces using lightweight CNN, including RetinaFace, ArcFace, and a FT algorithm in real-time.	High accuracy,
S. Shaimaa et al. [6]	Iterative Dichotomiser3	FG-NET dataset	The study uses an iterative Dichotomiser3 that is used to classify individuals based on their gender and age using k-means.	accuracy of 90.93%, and F-measure was 89.4
B. Anil Kumar et al. [7].	CMNV2	real-time video images.	The study employs the CMNV2 model for masked image classification and face detection, while the Caffe model and MobileNetV2 are used for fast feature embedding.	accuracy 99.64 % recall is 99.28 % f1-score is 99.64 % error-rate is 0.36 %
Z. Wang et al. [8]	multigranularity masked FR model	(MFDD)dataset, (RMFRD)dataset, (SMFRD)dataset	The study developed multi-granularity masked FR model using MFDD, RMFRD, and SMFRD dataset.	Accuracy of 95%

3 Methodology

This paper focuses on the methodology and algorithms used for face recognition and tracking in real-time in the proposed work. The three main stages of the proposed FRT system are face detection, feature extraction, and classification. In the classification stage, the utilization of the DL technique plays a crucial role in achieving accurate and efficient FRT systems. The proposed FRT system uses DL models and lightweight CNN. This model is trained using labeled images. After training, the model can be used for FRT in images and videos. By leveraging DL techniques, these models can effectively recognize and track faces in real-time scenarios, thereby enhancing the overall performance and reliability of FRT systems in various applications. Figure 1, depicts the main stages of the proposed system in detail.

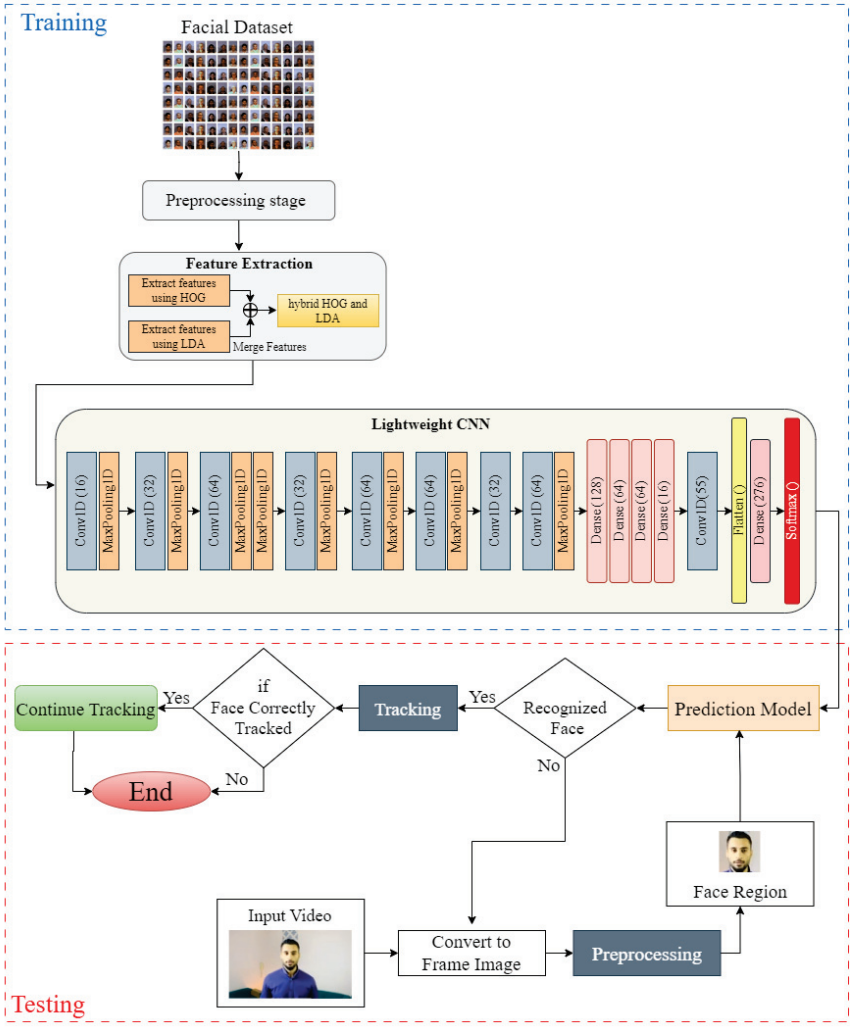

Fig. 1. Face Recognition and Tracking Model

3.1 Facial Dataset

A facial dataset is a collection of images or videos used to train and evaluate facial recognition and tracking algorithms, enhancing their accuracy, robustness, and generalization capabilities in real-time applications. The datasets utilized in the proposed system for training and testing are listed below [9].

3.1.1 Facial Images Dataset

In December 2008, a sample of the database's participants was drawn from the populace around the Lisle Social Sciences Building on the city campus of the University of Cape Town, and it contains 3755 faces. The resolution of the RGB (3 × 8 bit) image was 640 × 480 pixels. The image dataset is split into training and testing dataset. The training dataset contains 70% of the images and is intended solely for training purposes. This contains 352 images used to train the model. The testing dataset, on the other hand, contains the remaining 30% of the images. It is used as an independent set to assess the trained model's performance and generalization skills, which contain 150 images. These images are used to evaluate the model's accuracy and efficacy [10]. Figure 2 shows the samples of images for the MUCT dataset.

Fig. 2. Sample Image of The Dataset

3.2 Facial video Dataset

The study recorded real-time videos for several individuals (7 males and 2 females) of seven individuals aged 8 and above, captured at 30 frames per second. The videos were captured in a natural environment, with each frame containing 5,450 images. All the individuals faced the camera while seated; however, their orientation toward the cameras varied slightly due to variations in their posture and sitting height. The system used an HP True Vision 720p HD camera with temporal noise reduction and dual array digital microphones. The videos were saved in a dataset of face images, as shown in Figure 3.

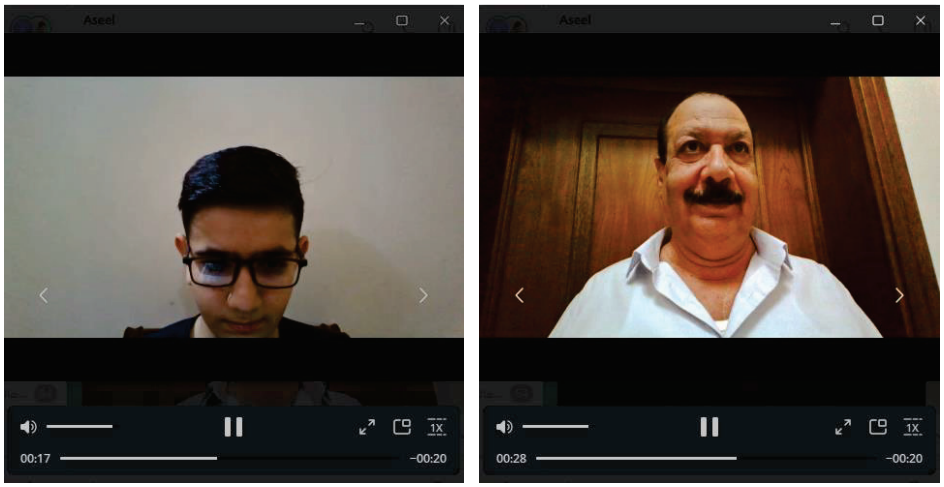

Fig. 3. Sample Video Dataset

3.3 Face Image Preprocessing and Detection Stage

The preprocessing stage enhances images for improved accuracy and effectiveness of the proposed face recognition and tracking system, using various image manipulation methods to make input face images suitable for information extraction tasks. The steps in Figure 4.

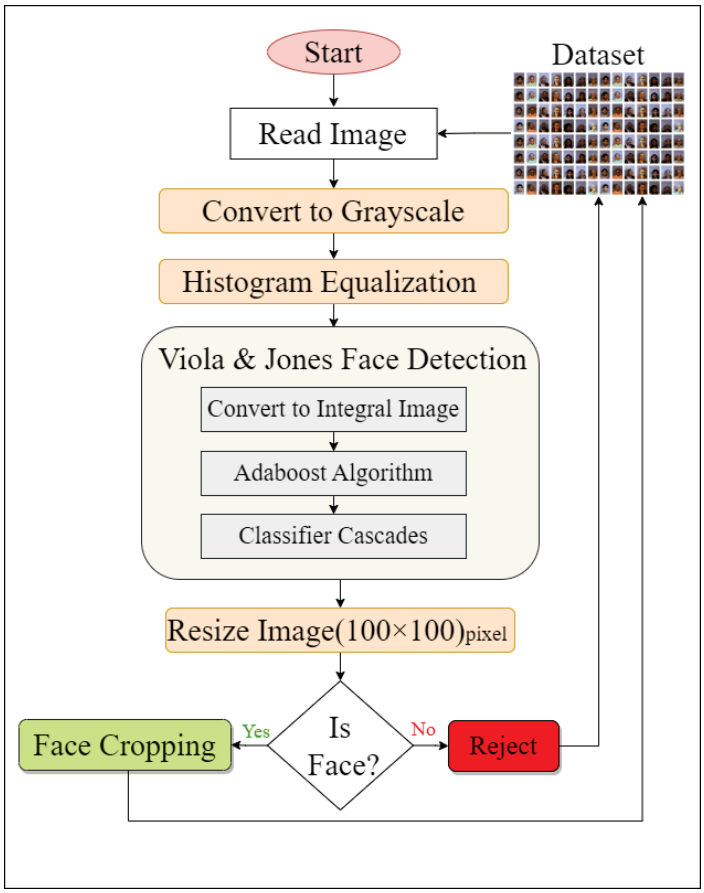

Fig. 4. Flow Diagram of Face Image Preprocessing and Detection

First, the input image is converted from RGB to grayscale, reducing the amount of data required to represent the image and improving processing speed. then enhanced the contrast by redistributing the intensity values by histogram equalization, which aims to achieve a uniform distribution of pixel intensity in the output image, improving the visibility of details. Then detect faces using "Viola & Jones." It achieves this by utilizing Haar-like features, integrated image arithmetic, AdaBoost training, and cascading classifiers, which are then used for further analysis and feature extraction to define facial features. After that, the dimensions of the face image are modified and resized to 100×100 pixels.

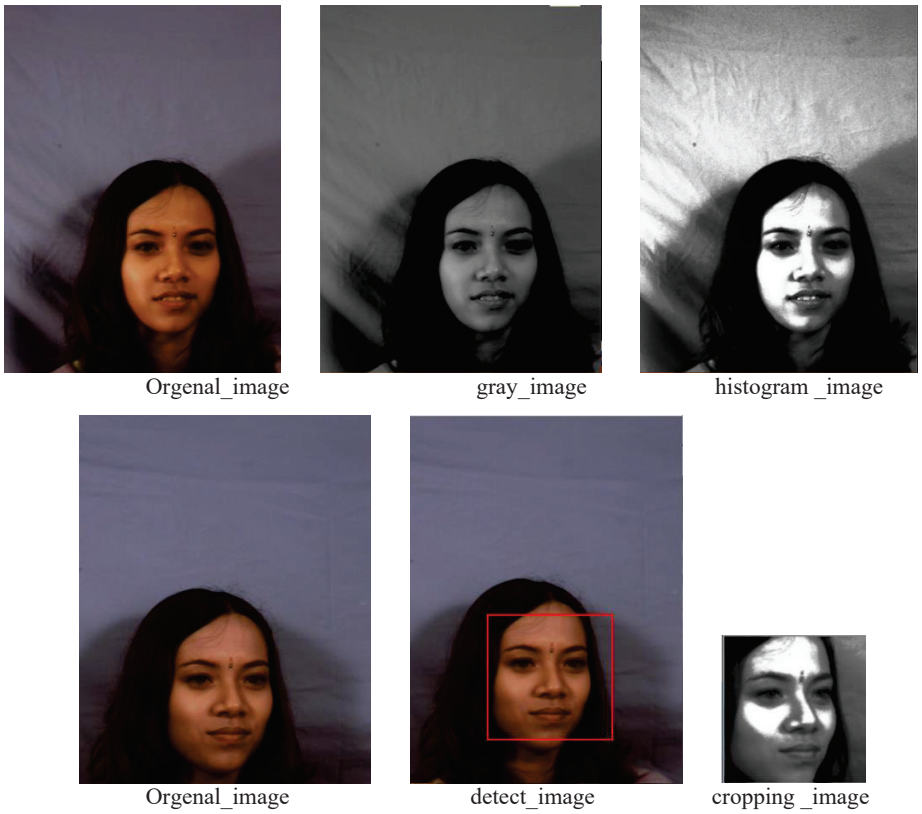

Fig. 5. Face Image Preprocessing and Detection Result

3.4 Feature Extraction Stage

This section examines the performance of the hybrid method of LDA and HOG for feature extraction, which is used in character and face recognition. The efficiency of the recognition system depends on the feature extraction steps.

Using the LDA approach, images are first converted into feature matrices in order to extract information from them. Following the LDA conversion of the images into feature matrices, the features are used to categorize the images into distinct groups. Then, to have a better understanding of the composition of features in each group, the mean and standard deviation for each class is calculated, along with the discrimination lines between classes.

After that, the images are then converted to black and white, the history of directed edges computed using the HOG algorithm, and grouped into a feature matrix.

Finally, merging feature matrices were extracted from LDA and HOG. The feature matrices extracted from LDA and HOG are combined using generalization methods. This merging process combines the features extracted from both methods to create a comprehensive feature representation from LDA and HOG.

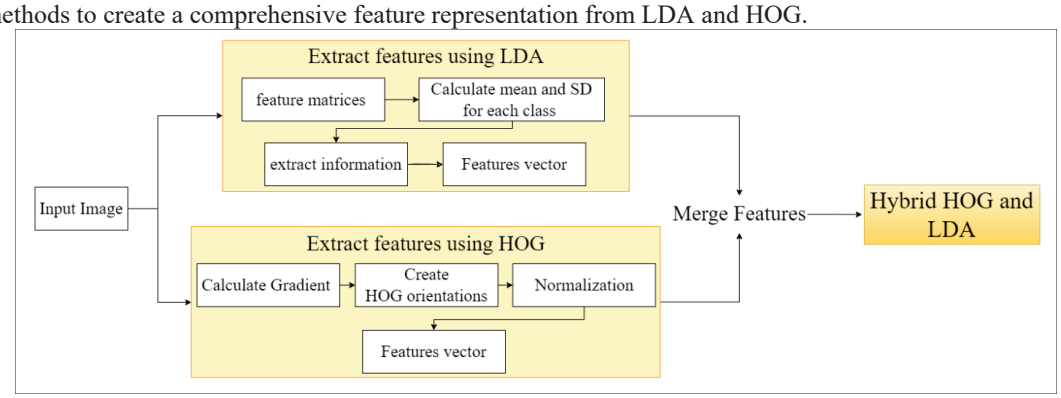

Fig. 6. Features Extraction Stage

3.5 FRT using Deep Learning (DL) Model

This section shows the implementation of the model using DL, which is a lightweight CNN. As mentioned, the input data is obtained from the hybrid feature extraction method combining HOG and LDA. Lightweight CNN architecture for face recognition and tracking purposes. The architecture consists of fewer layers and parameters, reducing computational and memory requirements. Detected faces are passed through the CNN to extract discriminative features for identification and classification. These features are then compared against a database of known faces or templates to identify the person. The face recognition algorithm calculates the similarity between the extracted features and known faces for identification. The CNN architecture's lightweight design and efficient algorithms enable real-time performance in face detection, recognition, and tracking, ensuring real-time performance.

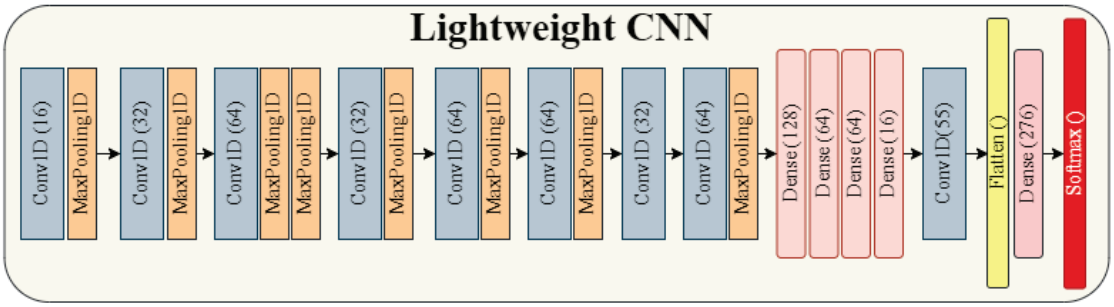

Fig. 7. Lightweight CNN Architecture

The lightweight CNN layers that are used are explained as follows:

Conv1D Layer_1: The first convolutional layer in the network has 16 filters, learning 16 different feature maps from the input. The kernel size is 3, the filter size is 3×1, the stride is 1, and Padding is set to "valid", which means no padding is added to the input.

MaxPooling1D Layer_2: is the first max pooling layer that follows the first convolutional layer. It performs max pooling operation with a pool size of 1, which means it selects the maximum value in each 1×1 region. The stride of 1 means it moves one step at a time.

Conv1D Layer _3: is the second convolutional layer in the network, consisting of 32 filters and a 3×3 kernel size, with the same stride and padding settings as the first convolutional layer and using ReLU activation function.

MaxPooling1D Layer_4: is the second max pooling layer follows the second convolutional layer. It performs max pooling with a pool size of 1 and a stride of 1.

Conv1D Layer _5: is the third convolutional layer in the network, consisting of 64 filters and a 3x3 kernel size, with the same stride and padding settings as the first convolutional layer and using ReLU activation function.

MaxPooling1D Layer_6: is the third max pooling layer follows the third convolutional layer. It performs max pooling with a pool size of 1 and a stride of 1.

MaxPooling1D Layer_7: Another max pooling layer follows the third convolutional layer. It performs max pooling with a pool size of 1 and a stride of 1.

Similar to the previous layers (Conv1D, MaxPooling1D) which will be **Layers_8,9,10, and 11**, consisting two more pairs of Conv1D and MaxPooling1D layers are added with different filter sizes (64 and 32) and activation functions (ReLU).

Dense Layers_12: Several dense layers follow the flatten layer, each with a different number of units (neurons) and activation functions (ReLU).

- **Dense_1:** number of units (neurons) = 128, and activation function is (ReLU).
- **Dense_2:** number of units (neurons) = 64, and activation function is (ReLU).
- **Dense_3:** number of units (neurons) = 64, and activation function is (ReLU).
- **Dense_4:** number of units (neurons) = 16, and activation function is (ReLU).

Conv1D Layer_13: is the final convolutional layer before the fully connected layers. t has 55 filters and a kernel size of 3×1. The padding is set to "same", which means zero-padding is added to the input to maintain the spatial dimensions. The activation function used is ReLU.

Flatten Layer_14: Preparing the output for the fully linked levels, this layer flattens it into a 1D vector from the preceding layer.

Dense Layer_15: is the final dense layer in the network. It has 276 units and uses the softmax activation function to produce probabilities for each class in a multi-class classification problem.

4 Testing and Evaluation of The Proposed System

Testing and evaluation are the final stages of the proposed approach. This data is known as the testing data, and it can be used to assess how well the algorithms are performing and how far they have come in their training, as well as to optimize or alter them for better outcomes.

The model has been trained using the MUCT dataset. To evaluate the performance of these algorithms, we also divided the MUCT dataset into two different ratios: 60:40 and 70:30. Different experiments were conducted in FR using DL, and a lightweight CNN model produced the best results. This model achieved high accuracy with a respectable splitting ratio and other evaluation metrics.

Table 2. Evaluate Performance of Proposed Lightweight CNN Model

Split Ratio	Learning Rate	Epoch	Batch Size	Precision (%)	Recall (%)	F1-Score (%)	Accuracy (%)
60:40	0.001	50	256	0.997	0.997	0.997	0.997
		100		0.998	0.998	0.998	0.998
70:30	0.001	50	256	1.000	1.000	1.000	1.000
		100		1.000	1.000	1.000	1.000

According to the table, the proposed lightweight CNN model's evaluated performance, in both situations, the model attained very high-performance metrics, demonstrating its ability to classify the data effectively.

We also compared relevant models using the proposed FRT model technique. There are several methods for FRTs. The results of three relevant studies that used three distinct approaches on the same dataset were compared to the models proposed. The results of the comparison, as given in Table 3 and Figure 8, reveal that our proposal outperformed the present ones.

Table 3. Comparison of Performance with Other Models and proposed Model

Ref.	Model	Precision (%)	Recall (%)	F1-Score (%)	Accuracy (%)
G. Hiten et al. [8] 2021	"MobileNet-V2"	0.98	0.98	0.98	0. 98
K. Aly et al. [20], 2022	lightweight CNN	0.99	0. 95	0.97	0.94
B. Anil et al. [25], 2023	CMNV2	-	99.28%	99.64%	99.64%
Our, 2024	(Lightweight CNN Model)	100%	100%	100%	100%

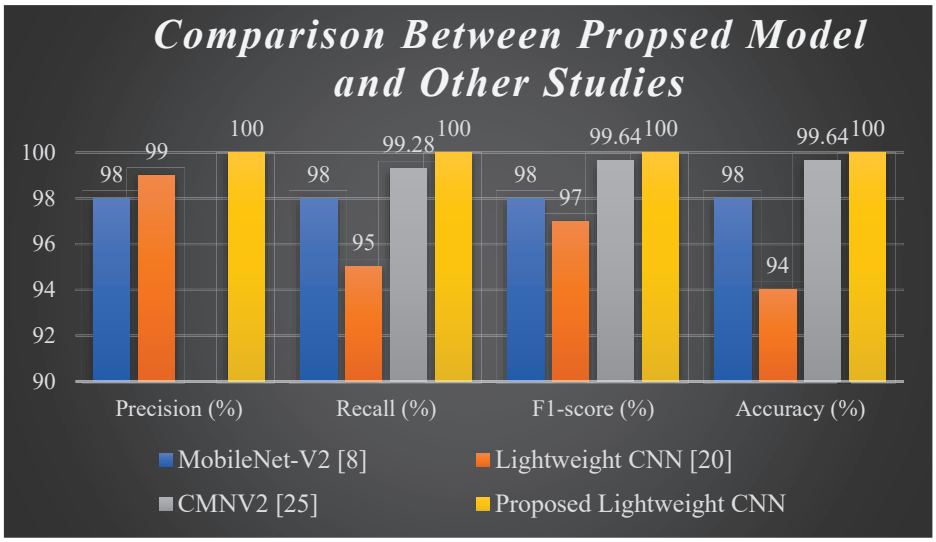

Fig. 8. Comparison of Performance with Existing Models and proposed Models

5 FRT Based on Real-Time Video

In this section, the testing process of recognition and tracking based on real-time video is explained. As mentioned in Section (3.1, B), after getting the video of the face, the video is captured, and each frame is converted into a still image. Each frame contains 5450 images, which are then saved in a dataset of face images. Each frame of the video is converted into a still image. This allows us to preprocess individual images rather than working directly with the video stream. After that, face detection was used in a preprocessed face dataset to identify and locate face regions in each image, ensuring accurate face recognition and tracking. The system compares the extracted face features or embeddings with a known face database to determine if there is a match to recognize the face. If the face in the image matches a known identity in the database, the tracking stage begins. The system then tracks the face across subsequent frames in the video clip, utilizing various tracking methods using DL, as shown in Figure 1. This allows for continuous monitoring of the identified person. If the face cannot be successfully identified or tracked in subsequent frames, the tracking process may end. By following these steps, face recognition and tracking can be performed in real-time video. Figure 9, present the window results to describe the system that has been used for FRT.

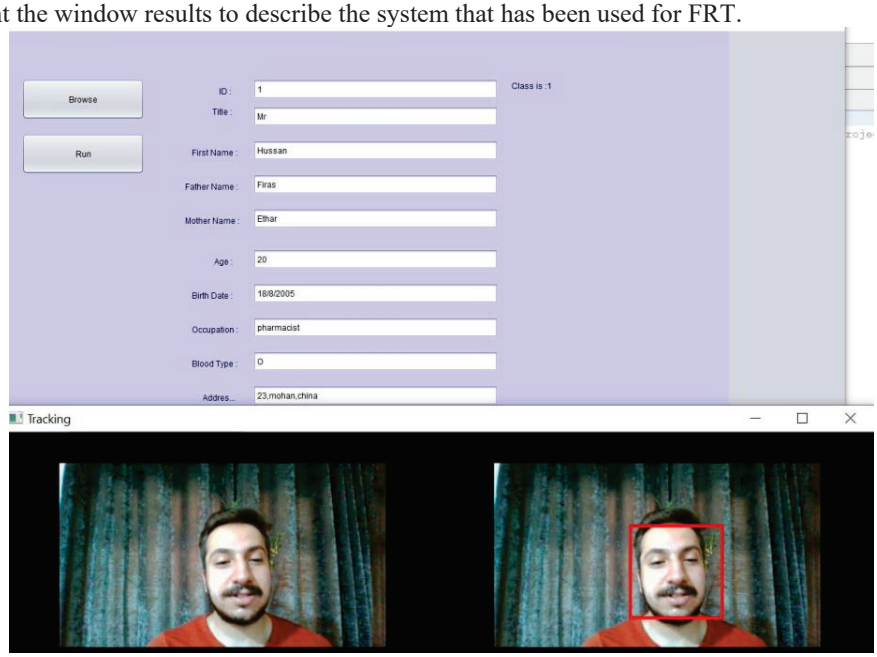

Fig. 9. Video Analysis and Retrieves Information

As shown in Figure, the system analyzes the visual features in the video frames to identify and detect faces, then uses facial recognition to match these with relevant data or a database, retrieving associated information like ID, title, name, age, gender, and occupation.

6 Conclusions

This paper offers a description of FRT technologies. The challenges associated with FR technology are also discussed, including issues related to lighting conditions, facial features, and the potential risks of false identification. These technologies optimize identification processes and offer intuitive methods for various applications in security systems and access controls.

The system starts with preprocessing, which involves converting to grayscale, histogram equalization for improved contrast, and FD using the "Viola & Jones" algorithm, which significantly improves the system. FR and image analysis, with a particular emphasis on important algorithms like LDA and HOG for extracting facial features, enhancing the performance of the system. Real-time FRT system development has been the primary goal of this study. This is achieved through the use of a model based on DL with a lightweight CNN.

The lightweight CNN model in DL achieved 100% accuracy. This shows the superiority of DL over others, and with lightweight CNN methods in mind, it tackles FRT tasks effectively. The experiments have shown that DL with

lightweight CNN models provides an effective solution for FRT tasks, even in difficult scenarios with variations in lighting conditions, pose, expression, and occlusions.

References

1. Zhang, Tong, and Herman Martins Gomes, "Technology survey on video face tracking," Imaging and Multimedia Analytics in a Web and Mobile World 2014,SPIE, vol. 9027, 2014.
2. Yan, Yan, and Yu-Jin Zhang., "State-of-the-art on video-based face recognition," Encyclopedia of Artificial Intelligence. IGI Global, pp. 1455-1461, 2009.
3. Li, Lixiang, Xiaohui Mu, Siying Li, and Haipeng Peng., "A review of face recognition technology," IEEE acces, vol. 8, pp. 139110-139120, 2020.
4. Hiten Goyal, Karanveer Sidana, Charanjeet Singh, Abhilasha Jain , swati Jindal, "A real time face mask detection system using convolutional neural network," Multimedia Tools and Applications, 2022.
5. Khalifa, A., Abdelrahman, A. A., Strazdas, D., Hintz, J., Hempel, T., & Al-Hamadi, A., "Face Recognition and Tracking Framework for Human–Robot Interaction," Appl. Sci., pp. 2-19, 30 May 2022.
6. Shaimaa Hameed Shaker, Farah Qais Al-Khalidi, "Human Gender and Age Detection Based on Attributes of Face," International Journal of Interactive Mobile Technologies (iJIM) ·, p. 176, 2022.
7. B. Anil Kumar and Mohan Bansal, "Face Mask Detection on Photo and Real-Time Video Images Using Caffe-MobileNetV2 Transfer Learning," Appl. Sci., vol. 13, p. 935, 2023.
8. Zhongyuan Wang, Guangcheng Wang, Baojin Huang, Zhangyang Xiong, Qi Hong, Hao Wu, Peng Yi, Kui Jiang,Nanxi Wang, Yingjiao Pei, Heling Chen, Yu Miao, Zhibing Huang, and Jinbi Liang, "Masked Face Recognition Dataset and Application," IEEE Transactions on Biometrics, Behavior, and Identity Science, 2023.
9. M. Mukhiddinov, O. Djuraev, F. Akhmedov, A. Mukhamadiyev, and J. Cho, "Masked Face Emotion Recognition Based on Facial Landmarks and Deep Learning Approaches for Visually Impaired People," Sensors, vol. 23, no. 3, p. 1080, Jan. 2023.
10. Rouhi, R., Amiri, M., & Irannejad, B., "A review on feature extraction techniques in face recognition," Signal & Image Processing, vol. 3, no. 6, p. 1, 2012.