INTERNATIONAL JOURNAL OF RESEARCH AND SCIENTIFIC INNOVATION (IJRSI)
ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

Deep Learning-Based Face Detection, Feature Extraction, and Face Recognition from Video: A Comprehensive Review

Patel Bhautika R.

Smt. Tanuben & Dr. Manubhai Trivedi College of Information Science, Surat

DOI: https://doi.org/10.51244/IJRSI.2026.1307000275

Received: 29 July 2026; Accepted: 03 August 2026; Published: 14 August 2026

ABSTRACT

Face recognition has become one of the most prominent biometric technologies due to its extensive applications in surveillance, access control, authentication, human-computer interaction, and intelligent security systems. Recent advances in deep learning, particularly Convolutional Neural Networks (CNNs), have significantly improved the accuracy and robustness of face detection, feature extraction, and face recognition under challenging real-world conditions. This paper presents a comprehensive review of deep learning-based techniques employed across the complete face recognition pipeline. A comparative analysis of existing studies is presented to highlight the evolution of deep learning techniques and their effectiveness in improving recognition accuracy and computational efficiency. The review also discusses widely used benchmark datasets, performance evaluation metrics, and the major challenges encountered in unconstrained environments, such as pose variation, illumination changes, occlusion, facial expressions, aging, and low-resolution imagery. Finally, emerging research directions, including lightweight deep learning models, attention mechanisms, Vision Transformers, self-supervised learning, explainable artificial intelligence, and real-time video-based face recognition, are outlined to provide insights for future research. This review serves as a comprehensive reference for researchers and practitioners seeking a thorough understanding of recent developments and future trends in deep learning-based face recognition.

Keywords: Face Detection, Face Recognition, Feature Extraction, Deep Learning, Convolutional Neural Networks (CNNs), Video-Based Face Recognition, Biometrics, Computer Vision.

INTRODUCTION

In the current era, the increasing incidence of fraud and security threats has created a growing demand for reliable and efficient biometric authentication systems. To address these challenges, various biometric traits, including the face, iris, fingerprint, palmprint, and voice, are employed for personal identification and verification. Among these, facial recognition has emerged as one of the most widely accepted non-intrusive biometric techniques due to its convenience, ease of acquisition, and user-friendly nature. Face recognition has become one of the most successful applications of computer vision and has attracted considerable research attention over the past few decades. This rapid growth is primarily driven by advances in artificial intelligence, deep learning, imaging technologies, and computational capabilities, along with the increasing demand for secure, accurate, and real-time identity verification across diverse application domains. Face recognition techniques can be classified as still-image-based or video-based, depending on whether recognition is performed on single images or video sequences. Zheng et al. [1] highlighted that video-based face recognition has gained considerable attention because of its broad range of applications, including video content analysis, intelligent surveillance, and access control systems. Compared with image-based face recognition, video-based face recognition presents greater challenges owing to significant variations in facial pose, illumination, image quality, motion blur, occlusions, and the substantial volume of video data that must be processed and analyzed efficiently.

The rapid advancement of deep learning has transformed face detection, feature extraction, and face recognition by enabling highly accurate and robust biometric systems. Numerous deep learning models, including Convolutional Neural Networks (CNNs), Residual Networks (ResNet), Vision Transformers (ViTs),

INTERNATIONAL JOURNAL OF RESEARCH AND SCIENTIFIC INNOVATION (IJRSI)
ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

and lightweight architectures, have been proposed to address challenges associated with unconstrained environments. However, the existing literature is often fragmented, with most survey papers focusing on only one aspect of the face recognition pipeline, such as face detection or face recognition, while providing limited coverage of feature extraction techniques. Furthermore, the continuous emergence of new architectures, benchmark datasets, and evaluation methodologies necessitates an updated and comprehensive review. Therefore, this review aims to provide a unified overview of deep learning-based face detection, feature extraction, and face recognition techniques, enabling researchers to understand the evolution of these methods, compare their strengths and limitations, and identify promising directions for future research.

The remainder of this paper is organized as follows. Section 2 presents the face recognition pipeline, providing an overview of the fundamental stages involved in face recognition systems. Section 3 reviews face detection techniques, including traditional and deep learning-based methods. Section 4 discusses feature extraction techniques, emphasizing the transition from handcrafted features to deep feature representations. Section 5 presents deep learning-based face recognition methods and recent advancements in the field. Section 6 provides a comparative analysis of representative studies, benchmark datasets, and performance evaluation metrics. Section 7 discusses the key challenges and future research directions in deep learning-based face recognition. Finally, Section 8 concludes the paper by summarizing the major findings and outlining potential directions for future research.

DEEP LEARNING BASED FACE RECOGNITION PIPELINE

A deep learning-based face recognition system consists of a sequence of interconnected stages that transform an input image or video into a recognized identity. Unlike traditional face recognition systems that rely on handcrafted feature descriptors, deep learning-based approaches automatically learn hierarchical and discriminative facial representations from large-scale datasets as suggested by Wang et al., [2] and Guo et al. [3]. This capability has significantly improved the robustness and accuracy of face recognition under unconstrained conditions, including variations in pose, illumination, facial expressions, occlusion, and image quality.

The typical deep learning-based face recognition pipeline comprises five major stages: face detection, preprocessing, feature extraction, feature matching, and face recognition as shown in the Figure 1. Initially, the face detection stage identifies and localizes facial regions from an input image or video frame using deep learning models such as MTCNN, RetinaFace, or YOLO. The detected face is then preprocessed through operations such as face alignment, normalization, resizing, and illumination correction to reduce variations that may affect recognition performance. Subsequently, the feature extraction stage employs deep convolutional neural networks (CNNs) to learn high-dimensional and discriminative facial embeddings. Popular architectures such as VGGFace, FaceNet, ResNet, ArcFace, and MobileFaceNet have demonstrated remarkable success in generating robust facial representations for face recognition tasks [4]-[8]. During the feature matching stage, the extracted feature vectors are compared with those stored in a database using similarity measures such as Euclidean distance or cosine similarity. Finally, the identity of the input face is determined through classification or verification based on the similarity score.

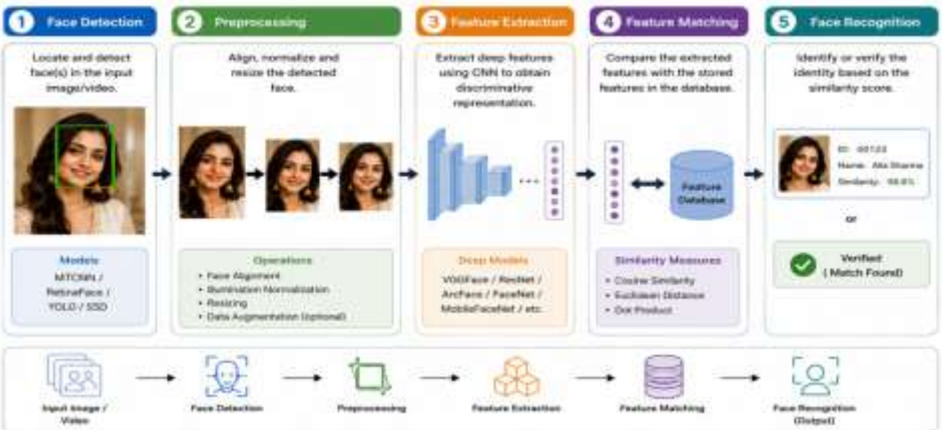

Figure 1: Deep Learning-Based Face Recognition Pipeline

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

Recent studies have demonstrated that deep learning-based face recognition systems outperform conventional machine learning methods owing to their ability to learn robust feature representations directly from raw facial images. Taigman et al. [9] introduced DeepFace, which achieved near-human-level performance by integrating deep neural networks with facial alignment techniques. Schroff et al. [5] proposed FaceNet, which learns compact facial embeddings using a triplet loss function, significantly improving face verification and clustering performance. Furthermore, Deng et al. [7] developed ArcFace, which introduced an additive angular margin loss to enhance the discriminative capability of learned facial features, leading to state-of-the-art recognition accuracy on several benchmark datasets.

Face Detection Techniques

Face detection is the first and one of the most critical stages in a face recognition system, as the overall recognition performance largely depends on the accurate localization of facial regions. The primary objective of face detection is to identify the presence and position of one or more human faces within an image or video while distinguishing them from the background. An effective face detection algorithm should be capable of handling variations in pose, illumination, facial expressions, occlusions, scale, and complex backgrounds. Face detection techniques have evolved from traditional feature-based methods to advanced deep learning-based approaches. The adoption of deep learning has significantly improved detection accuracy and robustness, making face detection a fundamental component of modern face recognition systems.

Traditional Methods

Face detection has traditionally been performed using several approaches, including knowledge-based, feature-invariant, template matching, and appearance-based methods by Yang et al. [10]. Among these, the introduction of the Eigenfaces approach by Turk and Pentland [11] marked a significant milestone in computer vision by demonstrating the feasibility of appearance-based face representation for recognition tasks. Subsequently, Viola and Jones [12] proposed the Haar feature-based cascade classifier, which enabled efficient real-time face detection and became one of the most widely adopted techniques due to its speed and accuracy. Further advancements in handcrafted feature extraction led to the development of descriptors such as Local Binary Patterns (LBP) by Ojala et al., [13], Scale-Invariant Feature Transform (SIFT) propose by Lowe [14], and Histogram of Oriented Gradients (HOG) by Dalal and Triggs [15], which significantly improved feature representation for face analysis. Despite their effectiveness in controlled environments, these handcrafted feature-based methods exhibit limited robustness under unconstrained conditions characterized by variations in pose, illumination, facial expression, occlusion, and image quality. To address these limitations, Ranjan et al. [16] suggest that deep learning-based approaches are more effective alternative that automatically learning robust and discriminative facial representations from large-scale datasets.

Deep Learning Methods

The remarkable success of deep learning has led to the development of numerous face detection algorithms capable of handling challenging real-world scenarios. Prominent deep learning-based face detection methods include the Multi-task Cascaded Convolutional Networks (MTCNN) proposed by Zhang et al. [17], Faceness-Net by Yang et al. (2015) [18], the Two-Stage Cascaded Convolutional Neural Network (Two-Stage CNN) by Yang et al. (2016) [19], Contextual Multi-Scale Region-Based Convolutional Neural Network (CMS-RCNN) by Zhu et al. [20], the Single-Shot Scale-Invariant Face Detector (S3FD) by Zhang et al. [21], the Dual Shot Face Detector (DSFD) by Li et al. [22], the Light and Fast Face Detector (LFFD) by He et al. [23], and img2pose, a six-degree-of-freedom face pose estimation framework proposed by Albiero et al. [24]. These methods employ deep convolutional neural networks to learn robust facial representations, enabling accurate detection under challenging conditions such as variations in pose, illumination, scale, occlusion, and low image quality. In particular, multi-scale face detectors have demonstrated superior performance in detecting tiny and blurred faces by effectively extracting facial features at different scales, thereby improving detection accuracy and reliability in unconstrained environments.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

Feature Extraction Techniques

Feature extraction is a fundamental stage in a face recognition system, as it transforms facial images into compact and discriminative feature representations that facilitate accurate recognition. The quality of the extracted features directly influences the performance of the recognition system, particularly under variations in pose, illumination, facial expressions, occlusion, and image quality. Over the years, feature extraction techniques have evolved from handcrafted descriptors to deep learning-based feature representations, with the latter demonstrating superior robustness and discriminative capability. Traditional approaches rely on manually designed features to characterize facial appearance, whereas deep learning methods automatically learn hierarchical feature representations from large-scale datasets, enabling improved recognition performance in unconstrained environments, a study by Wang and Deng [2].

Traditional Feature Extraction Techniques

Early face recognition systems predominantly employed handcrafted feature extraction techniques to represent facial characteristics. Among these, Principal Component Analysis (PCA) introduced by Turk and Pentland [11] projects facial images onto a lower-dimensional eigenspace, commonly known as Eigenfaces, to reduce dimensionality while preserving significant facial information. Linear Discriminant Analysis (LDA), proposed by Belhumeur et al. [25], improves class separability by maximizing between-class variance and minimizing within-class variance. Independent Component Analysis (ICA) further enhances facial representation by capturing statistically independent features that cannot be extracted using PCA, suggested by Bartlett et al. [26].

Local feature descriptors have also been widely adopted owing to their robustness against illumination and expression variations. Local Binary Patterns (LBP) encode local texture information by comparing neighboring pixel intensities, making them computationally efficient for face analysis proposed by Ojala et al., [13]. Gabor filters extract multi-scale and multi-orientation texture features that closely resemble the response of the human visual system and have demonstrated excellent performance in facial feature representation, proposed by Liu and Wechsler [27]. Furthermore, Scale-Invariant Feature Transform (SIFT), a study by Lowe [14] and Speeded-Up Robust Features (SURF) by Bay et al., [28] provide scale- and rotation-invariant local descriptors, enabling reliable feature matching under varying imaging conditions. Despite their effectiveness in controlled environments, handcrafted feature extraction techniques exhibit limited robustness when dealing with complex real-world scenarios involving significant variations in pose, occlusion, and illumination.

Deep Feature Extraction Techniques

The emergence of deep learning has revolutionized feature extraction by enabling neural networks to automatically learn hierarchical and highly discriminative facial representations directly from raw images. Convolutional Neural Networks (CNNs) have become the dominant feature extraction framework due to their ability to capture low-level features, such as edges and textures, as well as high-level semantic representations through multiple convolutional layers, presented by LeCun et al., [29]. Building upon CNNs, several deep architectures have significantly advanced face recognition performance.

AlexNet, introduced by Krizhevsky et al. [30], demonstrated the effectiveness of deep CNNs in large-scale image recognition and laid the foundation for subsequent deep face recognition models. VGGNet employed deeper architectures with small convolutional filters, enabling the extraction of more discriminative facial features, mentioned by Simonyan and Zisserman [31]. ResNet addressed the degradation problem in very deep networks through residual learning, significantly improving feature representation and recognition accuracy, proposed by He et al. [6]. Inception (GoogLeNet) introduced multi-scale convolutional modules to efficiently capture features at different receptive fields while reducing computational complexity, described by Szegedy et al. [32]. DenseNet further enhanced feature reuse by establishing dense connections between layers, improving gradient propagation and reducing the number of trainable parameters, developed by Huang et al., [33]. Evantually, MobileNet has provided lightweight yet highly efficient feature extraction suitable for real-time and mobile face recognition applications through depthwise separable convolutions, designed by Howard et al. [34].

Compared with traditional handcrafted methods, deep feature extraction techniques automatically learn robust, hierarchical, and highly discriminative facial representations from large-scale datasets. Consequently, they have become the preferred choice for modern face recognition systems operating in unconstrained environments.

Face Recognition Techniques

Face recognition is the process of identifying or verifying an individual's identity by comparing extracted facial features with those stored in a database. Over the years, face recognition techniques have evolved from traditional handcrafted feature-based approaches to advanced deep learning models that automatically learn discriminative facial representations. This evolution has substantially improved recognition accuracy and robustness, making deep learning-based methods the preferred choice for modern face recognition applications.

The emergence of deep learning has revolutionized face recognition by enabling automatic learning of robust and discriminative facial representations from large-scale datasets. Convolutional Neural Networks (CNNs) have become the dominant architecture for face recognition owing to their ability to capture hierarchical facial features with minimal manual intervention. The following subsection reviews representative deep learning-based face recognition methods and discusses their key contributions, strengths, and limitations.

Face recognition system using Convolution Neural Network (CNN) to detect the facial images and Logistic Regression Classifier (LRC) to classify the features learned by CNN has been proposed by Khalajzadeh et al., [35]. The CNN architecture is made up of four layers: input layer, two convolutional layers and one sub-sampling layer. The output layer is a fully connected layer with 15 feature maps. To evaluate system performance, the Yale Face Dataset, which includes 165 images of 15 individuals, was used. There are 11 images of each individual, and out of that 9 were used for training and 2 for testing. The accuracy of this system was 83.63%. But, dataset size can be increased to make the model more effective.

Video based face recognition framework that can handle occlusion, pose variation and blurred image has been developed by Ding et al. [36]. To address these challenges they have employed Trunk-Branch Ensemble CNN (TBE-CNN) model. The TBE-CNN model is trained using a novel Mean Distance Regularized Triplet Loss (MDRTL) function, which enhances the discriminative power of face representations. The reason behind using TBE-CNN model is that it can extract complementary information from face images and patches cropped around facial components to enhance robustness to occlusion and pose variations. The model achieves state-of-the-art performance on three popular video face datasets: PaSC, COX Face, and YouTube Faces. This paper suggests that to improve the effectiveness of face recognition in the real-world challenges including pose variation, blurred image and occlusion, the proposed TBE-CNN model can be further improved.

With the aim of reducing number of parameters to train CNN architecture for face recognition Local Binary Convolutional Neural Network (LBCNN) has been proposed by Ferraz et al. [37]. They have experiment their model on Chokepoint dataset that includes video sequences of 25 individuals in controlled scenario. The LBCNN are less prone to over-fitting because they naturally have characteristics of parameter sharing and sparsity of connections that in turn reduce the number of parameters compared to standard CNNs; and have lower model complexity due to binary-weighted convolutional filters. They have experimented various network parameters like filter size, number of LBCNN modules, LBC filters, number of intermediate channels, sparsity and hidden units in fully connected layer to accelerate the training process. To increase the accuracy of network they have utilized Sigmoid and Laplacian functions during pre-processing stage and attained notable accuracy and processing time. Here the Sigmoid normalize data and introduce non-linearity, while the Laplacian is used to reduce noise from the image. The study shows effectiveness of LBCNN in reducing processing time by adopting binary-weighted convolutional layers, while maintaining accuracy in noisy images.

As we know that the power of Artificial Intelligence is increasing in our daily life. So, real-time face detection and person recognizing system has been employed by Yi and Luo [38]. The Viola-Jones has been used for face detection and CNN for face recognition. The model incorporates techniques such as Rectified Linear Units (ReLUs) activation function, local normalization, and dropout to improve generalization and reduce complex co-adaptations. The algorithm is experimented on CASIA-Webface and Labelled Faces in the Wild (LFW) dataset and achieves accuracy of ~92%. The authors suggest investigating the success of the proposed model on other face recognition datasets and exploring the implmentation of more advanced techniques, such as transfer learning and multi-task learning, to improve the model's generalization and robustness. Additionally, it highlights research opportunities for future exploration, using unsupervised pretraining to further enhance the performance of the proposed neural network model.

Zhai and He [39] developed a real-time video-based face recognition model. To improve recognition accuracy upon existing methods they have incorporated spatial pyramid pooling and center loss to handle images of varying sizes and minimize intra-class variance. The model achieves state-of-the art performance on the YouTube Faces dataset, outperforming existing FaceNet method. So, the authors have used a deep CNN to handle images of varying sizes and improve the effectiveness of face recognition. Although the model is evaluated on YouTube dataset its performance can be tested on other datasets in real-world scenarios.

A real-time face recognition system using different methods has been proposed by Acuña-Escobar et al. [40] to achieve a reliable real-time recognition rate. For face recognition they have used Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Elastic Bunch Graph Matching (EBGM), and Deep Learning. After conducting a comparative analysis by testing and evaluating the functionality of these methods, the CNN approach was ultimately selected as the best solution. For face feature extraction the Multitask Cascaded Convolutional Networks (MTCNN) was employed. The environment was tested with 32 volunteers, in 4 different environments: front face, smiling, wearing a cap, and glasses and obtained 94%, 91%, and 90% in metrics true-positive-rate, precision, and accuracy, respectively. In future, other deep learning architectures and techniques can be integrated and dataset can also be extended to get better accuracy and robustness of the architecture.

Face recognition domains have been attracting researchers. The cluster-matching based face recognition method that uses unsupervised learning to cluster faces has been developed by Mendes et al. [41]. MTCNN has been used for face detection, Labeled Clusters Generation and Cluster Matching is used for face recognition from the video. They have used CNNs like ResNet-50, StNet-50 and VGG16 and clustering algorithms k-Means, affinity propagation, and agglomerative clustering to recognize face from the videos of Brazilian Chamber of Deputies and 513 images from the Labelled Faces in the Wild (LFW) dataset. The authors also evaluate the quality of the clustering generated in the faces clustering step using the V-Measure and Silhouette Score. This paper contributes to existing knowledge by proposing a novel cluster-matching-based approach for face recognition in videos, which is scalable and adaptable to new classes. They achieve best homogeneity, completeness, and V-Measure as 0.9862, 0.9833, 0.9847 respectively using a combination of SetNe50 and agglomerative clustering. Despite of this randomly only one image of each deputy is used for validation.

Deep learning has received state-of-the-art performance to recognize faces from still images; unconstrained video based face recognition is a difficult task due to pose variation, illumination, occlusion etc. So, Zheng et al. [1] proposed robust and efficient face recognition model from the unconstrained video. In this system, face detection is carried out using Single Shot MultiBox Detector (SSD) and Deep Pyramid Single Shot Face Detector (DPSSD). Deep features are represented using face association and finally subspace-to-subspace similarity measure is applied for recognition. The system is assessed on four datasets namely Multiple Biometric Grand Challenge (MBGC), Face and Ocular Challenge Series (FOCS), IARPA Janus Surveillance Video Benchmark (IJB-S) and IARPA JANUS Benchmark B (IJB-B) and achieved good accuracy. However, they have used very high rank to predict most relevant class. The IJB-B and IJB-S datasets are used to get face association results whereas verification results are obtained on MBGC and FOCS datasets using different scenario with different rank.

Jiwei et al [42] presented video based face retrieval method using deep learning and key frame. They have extracted key frame by the inter-frame difference method, MTCNN model is utilized for face detection and the

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

FaceNet model serves as face recognition. The experimental data includes training sets, validation sets, and test sets of 10 actors. The authors received good performance for face recognition but the dataset size is too small only 500 photos for training and 81 photos for testing were used.

Patel and Desai [43] proposed a deep learning-based framework for real-time face identification from unconstrained videos using Convolutional Neural Networks (CNNs). The proposed system consists of four major stages: video frame extraction, face detection using Multi-task Cascaded Convolutional Networks (MTCNN), feature extraction, and face recognition using a CNN model. The model was trained and evaluated on the YouTube Faces (YTF) dataset, which contains videos captured under unconstrained conditions with significant variations in pose, illumination, facial expressions, occlusion, and image quality. The proposed CNN architecture comprises two convolutional layers, two max-pooling layers, and two fully connected layers, and was trained using the Adam optimizer with categorical cross-entropy loss. Experimental results demonstrated a test accuracy of 93.11%, with an average precision of 92.39%, recall of 91.59%, and F1-score of 91.93%, indicating the effectiveness of the proposed approach for video-based face recognition in challenging real-world environments. However, the proposed framework primarily relies on spatial feature learning from individual video frames and does not explicitly exploit temporal information across consecutive frames, which could further improve recognition performance in unconstrained video sequences.

Comparative Analysis

To provide a comprehensive understanding of the evolution of deep learning-based face recognition, the representative studies reviewed in this paper are comparatively analyzed based on their methodologies, datasets, performance metrics, strengths, and limitations. This comparative analysis highlights various deep learning architectures, emphasizing their effectiveness in addressing challenges such as pose variation, illumination changes, facial expressions, occlusion, and low-resolution images. Furthermore, it enables the identification of prevailing research trends, commonly used benchmark datasets, and existing limitations, thereby providing valuable insights into the current state of deep learning-based face detection, feature extraction, and face recognition techniques. The comparative summary of the reviewed studies is presented in Table 1.

Table 1 : Summary of Existing Face Recognition Approaches

Ref. No	Year	Face Detection Method	Feature Extraction	Face Recognition Approach	Dataset(s)	Accuracy / Performance	Limitation
[36]	2014	Not Specified	CNN	Logistic Regression Classifier (LRC)	Yale face	83.63%	The dataset is developed under changes in illumination, expression, and occlusion, itis not much challenging to handle real-time images. Tested only on 165 face images of 15 individuals only
[37]	2017	Not Mentioned	Trunk-Branch Ensemble CNN (TBE-CNN)	depending on dataset and conditions	PaSC, COX Face, and YouTube	~90% (depending on dataset and conditions)	Although these datasets are source for unconstrained video-based face recognition, we can improve the

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

							recognition rate
[38]	2018	Not specified	Local Binary Convolutional Neural Network (LBCNN)	LBCNN	Chokepoint	Not provided	Sensitive to pose, illumination, occlusion because dataset designed for controlled surveillance scenario
[39]	2018	Viola-Jones	CNN	CNN	CASIA-Webface and LFW	~92%	The system is complex because of combining object detection, person recognition, and tracking in one pipeline, not a generalize model, sensitive to illumination, occlusion. However, quantitative accuracy metrics like precision, recall, or F1 score are also missing. Face detection is also sensitive
[41]	2019	MTCNN	Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Elastic Bunch Graph Matching (EBGM), CNN	CNN	32 Volunteers	90%	The system is tested in four controlled environments: front-facing, smiling, wearing a cap, and wearing glasses, the system was trained and tested on a specific group of 32 volunteers
[42]	2020	MTCNN	ResNet-50 StNet-50 VGG-16	Unsupervised learning (k-Means, affinity propagation, and agglomerative clustering)	Brazilian Chamber of Deputies and 513 images from the (LFW)	Vary depending of CNN architecture	The clustering process can become computationally intensive as the dataset size increases, performance is primarily evaluated on a specific dataset, sensitive to clustering

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

							parameters and only one image of each deputy is used for validation, limited images of LFW is used
[43]	2020	Single Shot MultiBox Detector (SSD) and Deep Pyramid Single Shot Face Detector (DPSSD)	face association	subspace-to-subspace similarity measure	MBGC, FOCS, IJB-S and IJB-B	Better	Face association is sensitive to rapid motion, significant pose changes, or when faces are partially occluded, leading to incorrect associations, computationally complex because it has to process multiple-shot videos and low-quality frames and used very high rank to predict most relevant class
[44]	2020	MTCNN	FaceNet	FaceNet	500 photos of 10 actors	Good	It does not provide explicit numerical accuracy metrics including precision, recall, or F1-score that makes it challenging to quantitatively assess the method's performance. Also FaceNet performance drops with poor lighting, pose variation, occlusion etc. due to limited images there is a chance of overfitting and bias

The comparative analysis represents various deep learning-based face recognition approaches, which have significantly enhanced the accuracy and robustness. CNN-based architectures dominate recent research due to their ability to learn hierarchical and discriminative facial representations automatically. Moreover, benchmark datasets such as LFW, YouTube Faces, CASIA-WebFace, MegaFace, and IJB are widely adopted for evaluating model performance. Despite remarkable progress, challenges including pose variations, illumination changes, occlusion, aging effects, and computational complexity continue to limit the deployment of face recognition systems in unconstrained real-world environments. These observations underscore the need for more efficient, lightweight, and robust deep learning models capable of achieving reliable real-time face recognition under diverse operating conditions.

Challenges And Future Research Directions

Despite the remarkable progress achieved by deep learning-based face recognition techniques, several challenges remain that limit their effectiveness in real-world applications. Although modern architectures have demonstrated high recognition accuracy on benchmark datasets, their performance often degrades in unconstrained environments due to variations in pose, illumination, facial expressions, occlusion, motion blur, aging, and low-resolution images or videos. Furthermore, large-scale deployment of face recognition systems requires significant computational resources, extensive annotated datasets, and efficient model optimization, making real-time implementation on resource-constrained devices a challenging task. Concerns related to data privacy, security, algorithmic bias, fairness, and adversarial attacks have also emerged as critical issues that must be addressed to ensure the reliability and ethical use of face recognition systems.

Future research should focus on developing more robust, efficient, and trustworthy face recognition frameworks capable of operating reliably under diverse real-world conditions. Emerging research directions include lightweight deep learning architectures for edge and mobile devices, self-supervised and unsupervised learning to reduce dependence on labeled datasets, Vision Transformers (ViTs) and hybrid CNN–Transformer models for improved feature representation, and attention mechanisms for enhanced discriminative learning. In addition, explainable artificial intelligence (XAI) can improve the transparency and interpretability of deep learning models, while privacy-preserving techniques such as federated learning and secure model training can address data security concerns. Further investigation into multimodal biometric systems, video-based face recognition, domain adaptation, and continual learning is expected to enhance the robustness, scalability, and adaptability of future face recognition systems.

CONCLUSION

Face recognition has witnessed significant advancements with the emergence of deep learning, leading to substantial improvements in the accuracy, robustness, and efficiency of biometric recognition systems. This review presented a comprehensive overview of deep learning-based techniques for face detection, feature extraction, and face recognition, highlighting the evolution from traditional handcrafted feature-based methods to modern deep neural network architectures. Representative deep learning models and their contributions were discussed, along with benchmark datasets, performance evaluation metrics, and recent developments in the field. A comparative analysis of existing studies demonstrated that deep learning approaches consistently outperform conventional methods by learning robust and discriminative facial representations capable of handling challenging real-world conditions.

Despite these advancements, several challenges, including pose variation, illumination changes, occlusion, aging, low-resolution imagery, computational complexity, and privacy concerns, continue to hinder the widespread deployment of face recognition systems in unconstrained environments. Addressing these challenges requires the development of more efficient, interpretable, and privacy-preserving algorithms. Emerging technologies such as lightweight deep learning architectures, Vision Transformers, self-supervised learning, attention mechanisms, and multimodal biometric systems offer promising opportunities for improving the reliability and scalability of future face recognition systems. It is anticipated that continued research in these areas will facilitate the development of intelligent, secure, and real-time face recognition solutions for a wide range of practical applications.

This review provides a consolidated reference for researchers and practitioners by summarizing the current state of deep learning-based face recognition and identifying key challenges and future research directions that can guide the development of next-generation face recognition systems.

REFERENCES

1. Zheng, J., Ranjan, R., Chen, C. H., Chen, J. C., Castillo, C. D., & Chellappa, R. (2020). An automatic system for unconstrained video-based face recognition. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, *2*(3), 194-209.

2. Wang, M., & Deng, W. (2021). Deep face recognition: A survey. *Neurocomputing*, *429*, 215-244.
3. Guo, G., & Zhang, N. (2019). A survey on deep learning based face recognition. *Computer vision and image understanding*, *189*, 102805.
4. Parkhi, O., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition. In *BMVC 2015-Proceedings of the British Machine Vision Conference 2015*. British Machine Vision Association.
5. Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 815-823).
6. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).
7. Deng, Jiankang, JiaGuo, NiannanXue, and Stefanos Zafeiriou. "Arcface: Additive angular margin loss for deep face recognition." In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4690-4699. 2019.
8. Chen, S., Liu, Y., Gao, X., & Han, Z. (2018, August). Mobile Facenets: Efficient CNNs for accurate real-time face verification on mobile devices. In *Chinese conference on biometric recognition* (pp. 428-438). Cham: Springer International Publishing.
9. Taigman, Y., Yang, M., Ranzato, M. A., & Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1701-1708).
10. Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, *415*, 295-316.
11. Turk, M., & Pentland, A. (1991). Eigenfaces for recognition. *Journal of cognitive neuroscience*, *3*(1), 71-86.
12. Viola, P., & Jones, M. (2001, December). Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001* (Vol. 1, pp. I-I). IEEE.
13. Ojala, T., Pietikäinen, M., & Harwood, D. (1996). A comparative study of texture measures with classification based on featured distributions. *Pattern recognition*, *29*(1), 51-59.
14. Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, *60*(2), 91-110.
15. Dalal, N., & Triggs, B. (2005, June). Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)* (Vol. 1, pp. 886-893). IEEE.
16. Ranjan, R., Sankaranarayanan, S., Bansal, A., Bodla, N., Chen, J. C., Patel, V. M., & Chellappa, R. (2018). Deep learning for understanding faces: Machines may be just as good, or better, than humans. *IEEE Signal Processing Magazine*, *35*(1), 66-83.
17. Zhang, K., Zhang, Z., Li, Z., &Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, *23*(10), 1499-1503.
18. Yang, S., Luo, P., Loy, C. C., & Tang, X. (2015). From facial parts responses to face detection: A deep learning approach. In *Proceedings of the IEEE international conference on computer vision* (pp. 3676-3684).
19. Yang, S., Luo, P., Loy, C. C., & Tang, X. (2016). Wider face: A face detection benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5525-5533).
20. Zhu, C., Zheng, Y., Luu, K., & Savvides, M. (2017). CMS-RCNN: contextual multi-scale region-based cnn for unconstrained face detection. In *Deep learning for biometrics* (pp. 57-79). Cham: Springer International Publishing.
21. Zhang, S., Zhu, X., Lei, Z., Shi, H., Wang, X., & Li, S. Z. (2017). S3fd: Single shot scale-invariant face detector. In *Proceedings of the IEEE international conference on computer vision* (pp. 192-201).
22. Li, J., Wang, Y., Wang, C., Tai, Y., Qian, J., Yang, J., & Huang, F. (2019). DSFD: dual shot

face detector. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5060-5069).

23. He, Y., Xu, D., Wu, L., Jian, M., Xiang, S., & Pan, C. (2019). LFFD: A light and fast face detector for edge devices. *arXiv preprint arXiv:1904.10633*.
24. Albiero, V., Chen, X., Yin, X., Pang, G., & Hassner, T. (2021). img2pose: Face alignment and detection via 6dof, face pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7617-7627).
25. Belhumeur, P. N., Hespanha, J. P., & Kriegman, D. J. (1996, April). Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. In *European conference on computer vision* (pp. 43-58). Berlin, Heidelberg: Springer Berlin Heidelberg.
26. Bartlett, M. S., Movellan, J. R., & Sejnowski, T. J. (2002). Face recognition by independent component analysis. *IEEE Transactions on neural networks*, *13*(6), 1450-1464.
27. Liu, C., & Wechsler, H. (2002). Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. *IEEE Transactions on Image processing*, *11*(4), 467-476.
28. Bay, H., Tuytelaars, T., & Van Gool, L. (2006, May). Surf: Speeded up robust features. In *European conference on computer vision* (pp. 404-417). Berlin, Heidelberg: Springer Berlin Heidelberg.
29. LeCun, Y., Bottou, L., Bengio, Y., &Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278-2324.
30. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*.
31. Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
32. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., & Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1-9).
33. Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4700-4708).
34. Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., & Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.
35. Khalajzadeh, H., Mansouri, M., & Teshnehlab, M. (2013, November). Face recognition using convolutional neural network and simple logistic classifier. In *Soft computing in industrial applications: proceedings of the 17th online world conference on soft computing in industrial applications* (pp. 197-207). Cham: Springer International Publishing.
36. Ding, C., & Tao, D. (2017). Trunk-branch ensemble convolutional neural networks for video-based face recognition. *IEEE transactions on pattern analysis and machine intelligence*, *40*(4), 1002-1014.
37. Ferraz, C. T., & Saito, J. H. (2018, October). A comprehensive analysis of local binary convolutional neural network for fast face recognition in surveillance video. In *Proceedings of the 24th Brazilian Symposium on Multimedia and the Web* (pp. 265-268).
38. Yi, X., & Luo, X. (2018, November). A system for real-time detecting and recognizing object person. In *Proceedings of the 4th International Conference on Communication and Information Processing* (pp. 67-71).
39. Zhai, Y., & He, D. (2019, February). Video-based face recognition based on deep convolutional neural network. In *Proceedings of the 2019 International Conference on Image, Video and Signal Processing* (pp. 23-27).
40. Acuña-Escobar, D., Ibarra-Fiallo, J., & Intriago-Pazmiño, M. (2019, December). Real-time face identification from video surveillance cameras. In *Proceedings of the Second International Conference on Data Science, E-Learning and Information Systems* (pp. 1-4).
41. Mendes, P. R. C., Busson, A. J. G., Colcher, S., Schwabe, D., Guedes, Á. L. V., & Laufer, C. (2020, November). A cluster-matching-based method for video face recognition. In *Proceedings*

of the Brazilian Symposium on Multimedia and the Web (pp. 97-104).

42. Jiwei, X., Jiyuan, X., Yi, F., &Dongfang, C. (2020, June). Research on video face retrieval method based on deep learning and key frame. In *Proceedings of the 2020 4th International Conference on Digital Signal Processing* (pp. 75-80).

43. Patel, B. R., & Desai, A. A. (2026). *Real-time face identification from video in an uncontrolled environment using CNN*. International Journal of Innovative Technology and Exploring Engineering, 15(3), 1–7. https://doi.org/10.35940/IJITEE.A3710.15030226