SSRG International Journal of Electrical and Electronics Engineering
ISSN: 2348-8379/ https://doi.org/10.14445/23488379/IJEEE-V11I8P126

Volume 11 Issue 8, 295-308, August 2024

Original Article

Implementing Deep Learning: A Novel Approach in CNNs for Face Recognition

Zubin C. Bhaidasna[1], Priya R. Swaminarayan[2], Hetal Z. Bhaidasna[3]

[1]*Computer Science & Engineering Dept, Parul University, Gujarat, India.*
[2]*Dean of Faculty of IT & CS, Parul University, Gujarat, India.*
[3]*CSE Dept, PIET-DS, Parul University, Gujarat, India.*

[1]*Corresponding Author : zbhaidas@gmail.com*

Received: 12 June 2024 Revised: 15 July 2024 Accepted: 12 August 2024 Published: 31 August 2024

Abstract - *Considering the vast amount of data available, it is clear that facial recognition and related technologies have made great progress in recent years. Facial recognition is especially important for law enforcement and forensic science in verifying someone's identity. Many researchers are working on using deep learning and machine learning to identify and classify people accurately based on their facial features. The first part of this review focuses on deep learning methods for facial identification and matching. The second part looks at a new technique that improves the accuracy of facial recognition algorithms by using large datasets for training. The paper discusses different ways to identify facial features from images and videos and proposes methods to improve accuracy, achieving 99.50% and 96.75% on the LFW and YTF datasets, respectively.*

Keywords - *Deep Learning, CNN, Face recognition, LetNet, AlexNet, ZFNet, Google Net, ResNet, R CNN, YOLO.*

1. Introduction

The use of facial recognition systems is ready to become a breakthrough innovation in the sphere of computer science, and it is the technology of the future that can identify and differentiate people in pictures or films. Uses include ATMs, health sector usage, license issuing, train booking, and security surveillance, among others.

However, the issue of facial images remains challenging, especially when working with large databases. It is explicitly clear that in the contemporary technological world, people are unique and can, therefore, be characterized by a variety of biometric characteristics, including fingerprints, palm signatures, hand geometry, iris scans, voice, and other special characteristics. Therefore, the ultimate goal of all these biometric applications is complementary to the concept of smart cities.

In different parts of the world, many researchers and scientific engineers are dedicated to developing and improving algorithms and methods that are accurate and robust enough to be used in real-life situations. The simplest type of recognition is password-based authentication, and due to the high frequency of their use, they require effective protection of personal data.

One of the main issues in current-day authentication systems is acquiring data for fingerprint scanning, speech recognition, iris scanning, or any such process. People have to align their biological features, for example, thumb, face or eye, in a way that would allow for an effective biometric scan. On the other hand, facial image acquisition is more friendly to the user, and the user may not necessarily know that the facial image is capturing him. Since they apply to most of the populations, facial features have important implications in research and present a solution to numerous problems connected with object identification.

Facial recognition systems primarily encompass two key aspects when dealing with facial data extracted from images or videos: Facial recognition systems primarily encompass two key aspects when dealing with facial data extracted from images or videos:

- Face Verification it can also be called authentication.
- Face Identification or Recognition Due to better technology.

It is also known as recognition for better results. Closely relating contemporary neural networks to the working of the human brain, deep learning and machine learning offer ways to solve the mentioned challenges. These are sub-disciplines of artificial neural networks that attempt to model the highly complex structures of the human brain. In order to enhance the odds of procuring a better result, the principles of deep learning are used profitably. Within the sphere of

surveillance systems and social networks like Facebook, the key function is delivered by deep learning, particularly in relation to the identification of a person. However, the most compelling problem for the present study concerns the issue of how to recognize and then re-identify the same person under conditions of appearance transformation like growing facial hair, wearing face masks, aging, changes in illumination, etc. This goes a long way in emphasizing the need to develop further robust algorithms for the use of deep learning.

2. Related Work

Facial recognition systems have been occupying a central place in research for more than ten years. Facial recognition is a field of study that is quite vast and covers areas such as machine learning, neural networks, imaging, computer vision, and pattern recognition. Many methods and strategies have been developed to address the solutions applied in the case of face recognition in videos. Here, based on the methods, a list of facial recognition algorithms and strategies is as follows:

2.1. Face Recognition with Deep Convolutional Neural Networks (CNNs)

The author over here addresses the problem of video-based face recognition, specifically focusing on associating face and body information for improved recognition accuracy [1]. The authors have used a new method that uses both face and body curves to improve the performance of a video-based face recognition system. They argue by stating that information from the body can provide more context that will help in identifying faces, especially in those scenarios where faces are partially occluded.

There are two main steps in this method: face detection and its association with the body. The face detection method is used to identify each face in individual video frames. The next step, association, will connect these detected faces with corresponding body regions, which is used to determine based on the relative spatial position of faces and their bodies, which helps create meaningful features that are used to capture both facial and body information.

The problem that lies with template adaption for face verification and identification tasks is template is the process of updating a pre-trained face recognition model to perform better to accommodate variations in pose, lighting, and other face factors that can affect the face appearance [2]. The authors have proposed a new method that uses both labeled and unlabeled face data for template adaptation by introducing an approach where the new method is iteratively updated by using a combination of labeled and unlabeled examples. This will help the model to become more robust to various variations that are commonly encountered in real world situations. A detailed study of a novel architecture, Local Binary Convolutional Neural Network (LBCNN), proposed for the task of fast face recognition in surveillance videos [3]. Deciding what makes one face look more similar to another is essential for real-time face recognition under limitations on the precision and storage capacity of computers, making it a research focus as well as an issue that the authors explore. A certain feature of LBCNN is that it is optimized for lesser computation and memory and still achieves the desired level of recognition. Local binary patterns and convolutional layers are used to obtain facial features efficiently. After that, the authors thoroughly analyze the potential of the presented method from the surveillance video databases and check the effectiveness of the developed approach compared to other existing methods.

The task of unrestricted face verification deals with cases when facial images are captured in still poses or in videos with rotations, illumination changes, and various facial expressions. The authors use a Deep Convolutional Neural Network (DCNN) to perform this task, as suggested in [4]. It contains more than one convolutional and fully connected layer that learns different abstract features of face images from raw data. The authors stress that face verification should be trained in large-scale datasets to obtain the robustness of algorithms and their high accuracy in an unconstrained environment. They furnish tremendous experimental results to substantiate the proposed method.

The early works inaugurated the view of using Convolutional Neural Networks (CNNs) for face recognition. The authors present an idea of using CNNs to learn hierarchical features from face images [5] automatically. It has a convolutional layer, a pooling layer, and a fully connected layer, thus exhibiting a network architecture. It is, however, relevant to point out that when this paper was written, the use of CNNs in face recognition was relatively new, and lots of improvements in architectures and training algorithms have been made that have resulted in better performance.

Face recognition using a Convolutional Neural Network (CNN) with a newly augmented dataset. The authors acknowledge knowledge that data augmentation plays a valuable role in the enhancement of the robust ability of CNN models and its invariance towards input data variations. Based on the above analysis, the authors recommend CNN architecture with an inclusion of DATA AUGMENTATION METHOD [6]. Augmentation techniques that are used are rotation, scaling, and flipping of face images so as to have a rich set of images to use in training. The outcomes of their experiments argue the benefits of data augmentation for face recognition.

Other loss functions previously in use are unsuccessful in achieving the best results, and therefore, ArcFace proposes a new additive angular margin loss to improve face recognition [7]. The authors pay special attention to the enhancement of the feature discrimination with the help of the margin that is applied in the angular space of the different classes. They

compute several novel softmax losses, and in particular, they generalize the traditional softmax loss to include an angular margin that brings the nearest classes of the embeddings and moves away the distant classes' embeddings. This is done by including an angle-based margin in the loss function so as to enhance intra-class compactness as well as inter-class separability. ArcFace also undergoes a cosine similarity transformation to normalize the embeddings, which enables the margin computation. For every experiment, the authors also provide extensive analysis and benchmarking on different datasets and show that ArcFace yields superior performance over other methods typical for facial recognition.

The authors propose SphereFace, a new model of face recognition that assumes the learning of discriminative features via mapping of the face images onto the hypersphere [8]. The authors stress that it is important to bear in mind that conventional softmax loss functions are rather ineffective considering the face recognition setting; they do not perform well in maximizing the angular margin between different classes in the feature space. This is done to optimize the angular distance between the feature vectors of two different classes directly, which is a strategy adopted by SphereFace. The novel contribution of SphereFace is the proposed angular softmax loss, which tries to make the learned features maximally discriminative and, at the same time, invariant to illumination and pose variance. The authors test SphereFace on several well-known face recognition databases. The experimental results show that SphereFace doesn't only offer the best performance compared to several other approaches that have been proposed previously.

FaceNet also estimates angular distances between faces and clusters of faces in a unified embedding space whose distances correspond to their similarities [9]. The authors use a deep convolutional neural network architecture CNN to model the metric directly for the same person's face images such that the output of comparable images of the face is closer than other ones and the outputs of dissimilar face images of different individuals maximum. However, they propose a triplet loss function whereby an anchor image is paired with a positive image that is from the same person as the anchor image and a negative image from another person to form informative training triplets. These triplets are used to provide the network minimal guidance on what to learn in order to distinguish the two classes. The authors also highlight triplet mining for online, where, approaching the convergence, hard negative and positive samples are selected. FaceNet is shown to provide embeddings that are favorable in terms of face verification/congest likelihood and clustering on multiple datasets.

2.2. Face Recognition with Specialized Architectures and Techniques

The paper also seeks to enhance deep face recognition by enhancing inter-class differences and intra-class compactness and presents the CosFace loss function [10]. Classical softmax loss works with equal importance of every class, which is rather disadvantageous in the case of face recognition, where the classes are intrinsically similar.

The CosFace loss intervenes with the angular margin of the embedding space in a way that involves the cosine value of angles made with reference vectors for each class. Concretely, in order to generalize the notion of angular margin, the authors introduce a scalar margin per class. They then scale the feature vector by the cosine of the margin-adjusted angle, which brings the classes apart with greater angular measure. The paper's contributions are the CosFace loss that is proposed in order to increase the discriminative ability of the features learned by the network and boost the generalization capability across face identity space. As the proposed Cosface loss seems to find its best application in specific datasets, it may not be very effective to cover other kinds of face recognition scenarios or non-standard faces. The model could, first of all, be less accurate in relation to variations in lighting, pose, movement, and other features that pertain to real-world scenarios.

The author discusses networks for face recognition under NIR and VIS spectra, and the work here is not easy since the characteristics of these spectra are dissimilar. For cross-spectrum face recognition, the authors have put forward a Wasserstein CNN to learn invariant features [11]. Wasserstein distance may also be called Earth Mover's Distance when it is intended to determine how close two probability distributions are. The model adopts NIR-VIS architecture, where the proposed model receives multiple domains as input and then outputs a common representation of shared latent space of NIR and VIS spectra. The distance between distributions of different spectra is restrained by the Wasserstein distance, which makes it robust to spectral variations and yet discriminative for face recognition. Some downsides of this model are as follows: it can be ineffective in dealing with traditional visual data; it works in near-infrared data but does not function well in conventional visible-light data.

In this case, this paper deals with video face recognition, where the objective is to perform face recognition across a video frame sequence. To overcome the difficulties of variations in illumination, pose and expression, the authors put forward a methodology that integrates adversarial learning and variational Aggregation [12]. The paper presents the use of a generator and a discriminator network in a problem-solving framework. Another detected augmentation involves the use of a generator to create additional face features for the model to cater to variations. The real and the generated features are distinguished by the discriminator, hence helping the generator. In addition, a variational aggregation process is used to merge features of the video frames by considering the dynamic changes in the facial expressions of the subjects in the video frames. The two approaches, such as adversarial

training and variational aggregation, may be sensitive to training data, thereby requiring a huge amount of data that may not be easily available when working with a small dataset. It is also worth noting that adversarial training can sometimes require careful tuning of parameters, and getting to convergence could, at times, be a problem. The issue of facial identity across different poses of the face under uncontrolled conditions. Another factor that is a great challenge in face recognition is pose variation because it induces non-linear transformation of the face. The authors have also designed a pose-aware deep learning model that learns a classifier for identity and regresses facial pose [13]. The model consists of several Convolutional Neural Networks (CNNs) to obtain the identity-related features and pose-related cues. As a result, integrating pose information into the training stage brings about good face recognition performance that consists of mirror or turned poses. The contributions of this paper are centred on establishing a multi-task learning framework that can cope with pose-invariant face recognition in the real environment.

As for the issue of face recognition, the author provides a new loss function known as CosFace. The authors also introduce the CosFace loss with the purpose of improving the discriminative ability of face recognition models [10]. CosFace loss thus modifies the traditional softmax loss in the sense that it introduces an angular margin for the classes. It makes all the features of various classes to be placed at a distance described by the cosine of the angle on the hypersphere. The method is useful in the sense that intra-class variations are significantly reduced while inter-class variances are maximized. The authors also present a degree of freedom they call the scale factor to adjust the size of the angular margin. Based on a number of face recognition datasets, the proposed CosFace loss has demonstrated higher accuracy in comparison with the softmax loss and the other methods.

This is particularly problematic for deep face recognition, which the author seeks to tackle through presentations of ArcFace loss, which seeks to increase the discriminative capability of the learned features [7]. The ArcFace loss function is based on the softmax loss, but it includes an angular margin between the classes. This margin is added to the cosine of the angle between the feature representations of the input image and the class-specific weight vector in the last fully connected layer of the network. The angular margin helps features from the same class to be tight and different from the features of the other classes. The authors also suggest a new approach to determining the class-specific weight vectors through the angular centres of given classes. Experimental results show that the deep face recognition models' accuracy is enhanced by using the ArcFace loss function.

This work views the problem of face verification as the problem of deciding whether two images belong to the same face or not. Authors have provided a deep discriminative feature learning approach that helps to improve discriminative characteristics of feature descriptors [14]. They try to formulate a novel Siamese network structure whereby the network learns discriminative features of a pair of face images. Learnt with contrast loss that helps the distance between similar faces to be close while that of different faces to be far apart. The authors present a feature fusion method to integrate both global and local features. The results of the experiments presented prove that employing the proposed approach enhances face verification compared to other traditional approaches.

This work of the author focuses on face recognition difficulty and proposes a new loss function known as CosFace. In order to improve the discriminative ability of face recognition models, the authors put forward the CosFace loss [10]. CosFace loss is a form of softmax loss where an angular margin between classes has been incorporated. It brings the features of different classes by their margin on the hypersphere in which the cosine similarity is used as distance. For this reason, the method is effective in ensuring that intra-class scattering is reduced and, at the same time, inter-class scattering is improved. The authors also bring another factor in to act as the scaling factor of the angular margin as well. CosFace loss, being efficient in enhancing the intra-class distance and reducing the inter-class distance, yields better accuracy of face recognition than softmax loss and other benchmark techniques through experimental results on different face datasets.

The author discusses the challenges of increasing deep face recognition by presenting the ArcFace loss used in an effort to increase the discriminative ability of the learned features [7]. The ArcFace loss is based on the softmax loss, which was further developed by adding an angular margin between classes. This margin is added to the angle between the feature representation of the input image and the class-specific weight vector in the final fully connected layer. It also keeps the angular margin from being large, which makes features of the same class to be compact and dissimilar to features of other classes. The authors also provide a new way of computing the class-specific weight vectors in terms of the angular centers of this class. Experimental outcomes have shown that ArcFace loss enhances the accuracy of deep-face models.

Face verification, the application that will be discussed in this paper, distinguishes if two given face images belong to the same identity. The authors combine a deep discriminative feature learning solution that focuses on the promotion of the discriminative ability of feature representations [14]. They put forward a Siamese network architecture that allows the learning of discriminative features for pairs of face images. The network itself is trained using a contrastive loss in order to have faces of the same identity closer in terms of the

distance between their features while faces of different identities are further apart. Also, the authors present a feature fusion technique that is used to merge between the features possessed by such a Global model and features possessed by such Local models. The results of the experiments indicate that the newly developed approach is more efficient than the prior methods of face verification.

The work over here presents DeepFace, a face verification model with substantially enhanced performance which depends on deep learning. The authors provide a deep CNN architecture for learning the hierarchical representation of face images [15]. The network is trained with a large dataset of images containing facial information and performs well in face verification. DeepFace takes several convolutional layers along with several fully connected layers to learn features that can be spatially and semantically distinctive. Compared to previous methods, it offers greater accuracy and gain in performance, bringing the performance of the machine near or slightly more than human performance in face verification.

2.3. Face Recognition with Metric Learning and Riemannian Geometry

The authors have presented a method that can be used to create binary hash video representations for face retrieval only [16]. The method aims to quantize face videos into binary codes and, at the same time, ensure that the quantization preserves structural features that are essential in the retrieval process. They do this using deep Convolutional Neural Networks (CNNs) to learn discriminative features from the video frames. Some of the parts of their approach include the feature representations that are extracted from the deep CNN from each frame of the video.

All of these frame-level features are then passed to a binary hash layer that quantifies the continuous features space into a binary search space. This makes the hash layer try to make the distances between the original feature vectors as close as possible while, at the same time, the binary codes used in the storage and retrieval must be compact. In order to test their method, the authors carry out experiments on face retrieval tasks employing benchmark datasets. They then evaluate whether they are more accurate and efficient binary hash video representations in contrast to the existing methods. This underscores the effectiveness of the proposed method for face retrieval in real-life scenarios.

For metric learning in face recognition in video data, the author discussed a technique that acts as a connection between the Euclidean space and the Riemannian space [17]. Riemannian geometry should be applicable to work with non-Euclidean structured data, such as covariance matrices, which are involved in face recognition issues. The authors have come up with another approach that aims to learn a metric on the Riemannian manifold to improve face recognition from video sequences. According to the two authors, their methodology comprises two stages. In the first stage, they extract discriminative features from the video frames using the deep network. These features are then used to compute the covariance matrices which live on the Riemannian manifold. In the second stage, they learn a distance measure, which in some way optimizes this manifold in such a way that points belonging to different classes are far apart. In order to support the analysis, the authors performed experiments on face recognition tasks using videos. They compare their method with other metric learning methods, and in many cases, especially with few samples for training, they have shown how effective their approach is. By mastering the correlation between the Euclidean and the Riemannian space and by mastering this relationship to create a concrete strategy for the implementation thereof, the paper makes its contribution.

2.4. Face Recognition and Deep Learning Based on Survey

The author pays great attention to the ethnicity recognition of people with the help of facial recognition, with the Deep Convolutional Neural Network (CNN) method [18]. This is the biggest challenge being encountered in society. Therefore, the goal of the study is to find an effective technique for identifying ethnicity so as to help in the following fields: security, sociology, and marketing, among others. The general idea of the proposed solution is based on the use of a deep CNN model trained on a dataset containing face images, the ethnic background of which is diverse. Through iteration, the ‘filter’ learns to find features that represent a given ethnicity in the best way. This study will most definitely contain processes such as data pre-processing, architecture design, and training protocols, all of which are specific to ethnicity recognition. The paper may present the used dataset, the metrics of performance and the possible use of the proposed approach.

This paper is a survey focused on the most current and innovative architectures of deep CNNs. It includes different CNN architectures created to improve the quality of activities, such as image recognition, bounding box detection, and segmentation [19]. The survey possibly enlists the works done on these styles of architectures, including the number of layers, the application of skip connections, normalization and other features. Such architectures may be discussed in the paper, as well as their purposes and the changes they introduce in comparison to previous systems. CNN structure can also be talked about as the current issue and prospects for the further development of CNN architecture.

In the following paper, the author examines several methods of face recognition and describes the difficulties associated with them. They probably include both conventional approaches used in face recognition and more recent approaches such as deep learning approaches [20]. The review may mention the limitations, such as pose variation, illumination, and occlusions. Maybe it will give some understanding of how various methods try to rise above these

challenges. The paper could also describe the results that can be obtained with different techniques on standard datasets, what kind of problems each of them is good for, and what kind of problems they are not so good for. In addition, it could insist on continuing research to address these aspects.

The current paper presents a semantic face parsing method to cater for occlusions in facial images [21]. The proposed model presumably employs a deep architecture that is well-tuned for segmenting multiple facial areas and deciphering the semiotics of each of them. It may employ some novel approaches with a particular reference to parsing face-masked areas and improve the face-parsing result. The paper will probably provide information regarding the network structure, the dataset employed for training and testing, and proof of the method's efficiency.

The special topic of interest of this paper is deep discriminative feature learning for face verification. As mentioned above, the main objective is to find highly discriminative representations of faces in the Facenets to distinguish between real and fake faces [22]. The authors might design a deep learning structure that would be able to encode necessary features for face verification problems. It may provide information about the training procedures and the loss functions that have been applied to detect the performance of the model. Presumably, experimental results of the proposed approach are demonstrated on the database of benchmark face verification to demonstrate the efficiency of the given approach.

First, it is worthwhile to introduce the method presented in this paper, called Deep Mutual Learning, whereby multiple deep neural networks are trained cooperatively in order to enhance each of their performance. The proposed approach may, therefore, involve sharing information or gradients between the networks during training to improve their joint learning [23]. This learning cooperation process is intended to ensure the effectiveness of the joint solution through the diversification of the separate networks. The paper could describe the motivation, the process of deep mutual learning, and the advantages of this approach, and it could probably show the results of experiments confirming the efficiency of the analyzed method on certain tasks.

This paper reviews various deep learning methods used in the face recognition field [24]. It possibly entails various forms of deep learning like Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs) and others. These techniques may include fading, time delay, chaining, successive approximations, and other such methods that may be older or newer in their usage. It may prescribe those complications discussed by these methods, such as dealing with fluctuations in facial expressions, pose, and lighting. The paper could give an idea of what currently exists on the topic of deep learning face recognition and what could be done in the future. The method described in this paper deals with face representation in the realm of deep learning for both identification and verification objectives [25]. The authors probably introduce an architecture of the neural network that has to be trained to make the facial images mapped into the feature space, the measures of which correspond to identification and verification. Such a joint learning process may also improve the discriminative capability of learned features for both tasks. The architecture design, the training process, and the evaluation of the proposed joint learning strategy for the text-to-speech task may be described in the paper.

2.5. General Deep Learning Techniques

This paper is a review of the available literature on the subject of how deep learning has been implemented in recommender systems. Recommender systems must perform well on user modelling and must have a clear concept of what is likely to be of preference to a particular user [26]. Matters like data scarcity, cold start issues, issues of scaling, and explainability features are among the challenges discussed in the paper regarding the use of deep learning in such systems. The authors also delineate possible solutions for these issues, including the utilization of side information, the development of models with a blend of schema updates, the use of attention mechanisms, and the address of dynamic user preference. The paper that has been developed outlines the present ideas on the use of deep learning for recommender systems and offers useful tips for potential advancements.

This paper aims to provide a brief exploration of the latest developments in None, which is a category of deep learning algorithms supervised to address image processing tasks. The authors then introduce the traditional architectures, LeNet and AlexNet; then, they feature the advanced architectures, VGG, GoogLeNet, and ResNet. Some of the developments that were noted are skip connections, residual networks, and inception modules [27]. It also includes methods like data augmentation, transfer learning and ways to understand the workings of CNN through visualization. The paper provides a state-of-the-art analysis of CNN's advances until 2018, which can be beneficial for researchers and practitioners.

CNNs play an essential role in image classification, hence, the subject of this survey paper. CNNs have extended the emphasis on the analysis of images and have provided a high level of accuracy in image classification [28]. This paper also discusses the basics of CNN, including the convolutional layer, the pooling layer, and the fully connected layer. It also covers types or architectures such as LeNet, AlexNet, VGG, GoogLeNet, ResNet, and DenseNet, addressing the contribution to image classification. The authors examine data augmentation, transfer learning, and fine-tuning methodologies. The survey is useful in answering questions about image classification with CNNs and, as such, is a good point of reference in the area.

This paper is a conceptual work; its goal is to define Convolutional Neural Networks (CNNs) and explain their functioning by using the deep learning approach. These items include the convolutional layer, pooling layer, activation function, and fully connected layer, which are important in constructing CNNs. Regarding the image, the authors describe the function of each part in terms of extracting hierarchical features from images [29]. Weight sharing and local receptive fields are important in CNN designs, as pointed out in the paper. The site also covers other topics that may be important during training, such as backpropagation and gradient descent. It is rather helpful to provide a conceptual overview of the content described in the paper for newcomers to the field of deep learning and CNNs.

In this paper, a comparison of several deep learning techniques when used for image detection. Image detection involves the task of locating items inside images. The comparisons of the state-of-the-art deep learning architectures such as CNNs, YOLO and Faster R-CNN in terms of accuracy, speed and computational cost have been presented [30]. Finally, the paper compares and contrasts the merits and demerits of each approach and shares comparative results of the strategies on various datasets. It allows researchers and practitioners to specify comparative characteristics of various deep-learning image detection algorithms.

3. Research Gap

Based on the above literature review, the following are the major research gaps that are identified as below:

- Handling Unconstrained Environments: Most existing face recognition systems struggle with the complexities of unconstrained environments despite varying backgrounds, lighting conditions, and occlusions.
- Handling Variations and Occlusions: There is a need to handle extreme variations such as occlusions and low-resolution images.
- Illumination Invariance and Pose Invariance: Varying lighting conditions pose a significant challenge for facial expression recognition systems. Current models struggle with recognizing faces at different angles, which is crucial for real-world applications where subjects may not always face the camera directly.
- Efficiency and Scalability of Training Methods: Exploring more efficient training methods and developing new architectures to enhance performance further.
- Exploration of Deeper Network Architectures: Future work to explore the potential of very deep architectures on larger-scale training datasets.

4. Proposed System

The technique called Deep Learning has been shown to be extremely robust for the last two decades because of its ability to manage large volumes of data and mathematical calculation. Especially in the field of pattern recognition, face recognition, etc., the usage of hidden layers has overtaken more interest in more conventional methods. Convolution Neural Networks (CNN), also called ConvNet, are among the most widely used deep neural networks in deep learning that deal with computer vision applications.

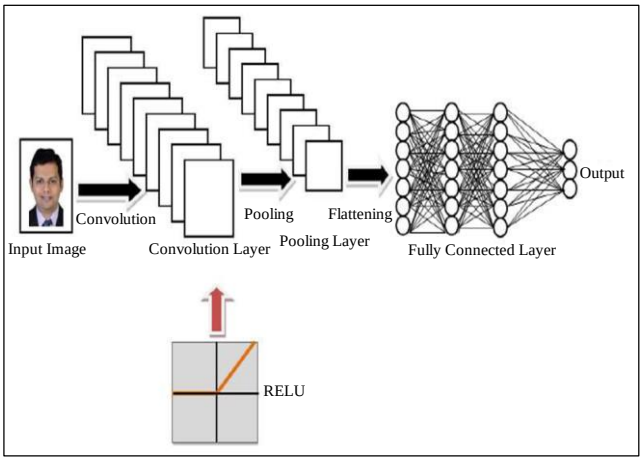

Fig. 1 Traditional convolution neural network

The fundamental mathematical function used widely in deep neural networks is called “Convolution” in CNN. It is one kind of linear operation that allows multiplying two functions to produce a third function that will represent the way shape of one function, which the other function can alter. In simple terms, a process involves multiplying two matrices that are images to give output in the form of feature extraction. CNN uses many such layers to process and extract necessary information from the images and automatically learn these features. Also, these layers apply a set of filters to the input image, sliding the filters over the entire image to detect different patterns like edges, corners, textures, etc.

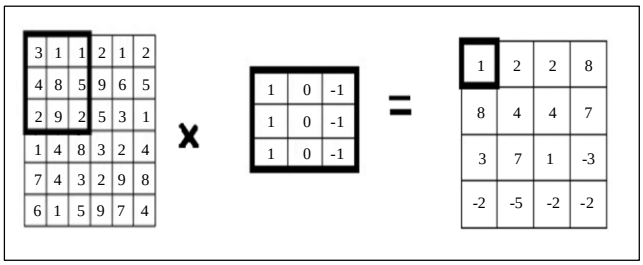

Fig. 2 Convolution operation

Following the convolution operations comes the pooling layers that are used to down sample the feature maps produced by the convolution layers, reducing the spatial dimensions and retaining the most salient information. There are two types of pooling operations, i.e. Max pooling and average pooling techniques. The max pooling takes the maximum value from the filter, while the average pooling takes the average values of all pixels in the filter.

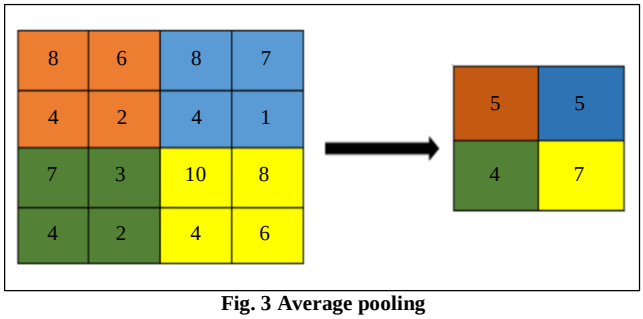

Fig. 3 Average pooling

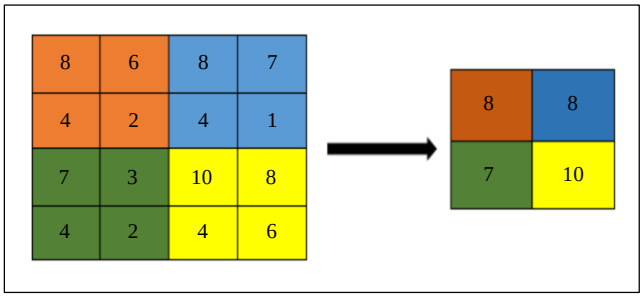

Fig. 4 Max pooling

After the pooling layer, the output sample is fed to the last layer, also called the fully connected layer. In this layer, the first operation is the flattening operation. All the resulting 2D matrix arrays from the pooled feature maps are flattened into a single, long and continuous linear vector, also called a 1D array. To categorize the image, the fully connected layer will receive the flattened 1D matrix as the input. After feature extraction, CNN typically has one or more fully connected and dense hidden layers for the final classification and regression tasks.

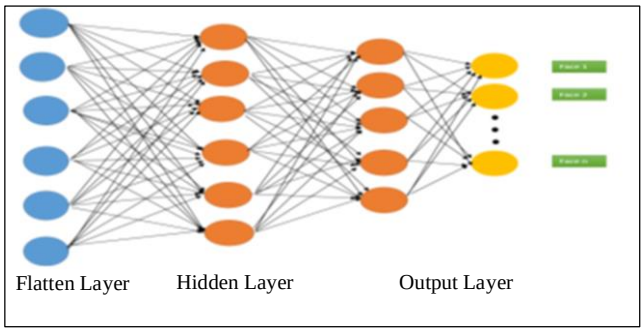

Fig. 5 Fully connected CNN layer

In the preceding chapters, various research articles on face recognition systems using deep neural networks have been studied, and the present level of technology is reviewed and summarized. The study and findings of these publications have helped us to create a face recognition system. As per the comprehensive survey conducted, many approaches have been studied, and combinations of these approaches can be used to create a new facial recognition system. In order to develop a state-of-the-art architecture for face recognition, a new architecture is proposed, one that is customized CNN architecture capable of face recognition. The below-mentioned figure below depicts the proposed architecture.

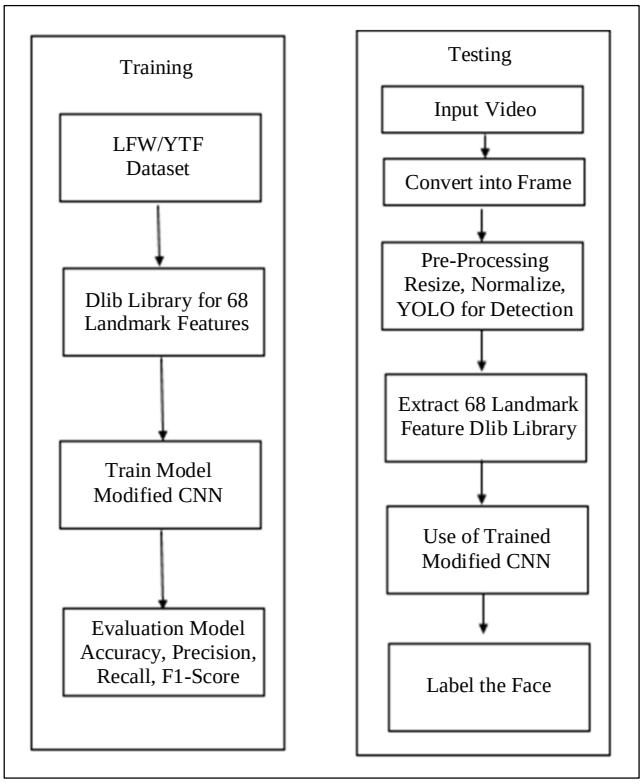

Fig. 6 Proposed novel modified CNN model

The above figure shows the proposed system for face recognition. The research aims to improve the performance of the face recognition system by making new modifications to the CNN architecture. The new proposed system aims to advance in the field of Deep learning computer vision in the areas of Face recognition from video by developing a customized and new CNN architecture that is capable of face recognition. Based on the above model depicted in the figure, In order to implement the execution of the new proposed modified CNN Model, there is a need to perform the following tasks:

Step1: Preprocessing

In the proposed work, preprocessing is first applied to the image, where preprocessing plays a pivotal role in optimizing data for the model. The input is a video for face recognition, and frames are extracted from these videos. These frames undergo preprocessing operations like reducing complexity while preserving necessary details. Each frame is resized to 96x96 pixels, ensuring uniformity across all input frames. By doing this, there is a need to improve the efficiency of this model.

Step 2: Face Detection

A CNN known as You Only Look Once V3 or YOLO V3 is used for face detection, and the preprocessed image is fed

to it. The network gives the regions of an image containing faces together with a confidence level of the same regions in the form of a score. These detections can be thought of as being represented by a bounding box. Cropping to the corresponding region is done by using the bounding box coordinates of the image are obtained. After the face has been detected and aligned using YOLO V3, the next process will be the use of Dlib's facial landmark detection function. This makes it possible to detect particular facial features, such as the eyes, the nose, and the mouth. These landmarks represent markers for further investigation and the recognition of faces.

Step 3: Proposed CNN Architecture

Here is what the modified CNN model looks like in the proposed work. In the proposed CNN, an attempt is made to improve the performance of convolutional layers that extract features, introduce new layers, and optimize the filter size. This is with the aim of achieving the best stability in face recognition.

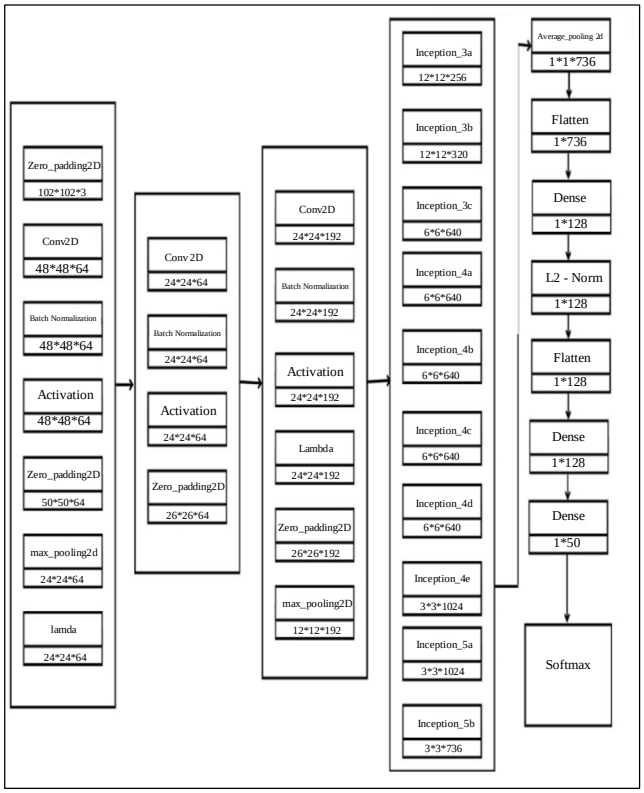

Fig. 7 Proposed novel modified CNN model

5. Result Analysis

The experiments were done on two sets, namely Labeled Faces in the Wild (LFW) and YouTube Faces (YTF) Dataset, to make sure that it works well in as many situations as possible. These datasets are different in terms of image resolution, illumination conditions, facial expressions, and poses, which gives a good benchmark for evaluating the performance of the proposed face recognition system.

5.1. LFW Dataset

Labeled Faces in the Wild is an easily available image dataset on the web that contains face images which are specifically procured for study and research only in the context of the problems associated with the 'in the wild' face recognition. The number of classes in the dataset is approximately 5751, and it contains more than 13000 images of different faces, all fetched from the web world. The images in this data set are named with the name of the person in the image. The LFW dataset is derived from the website and comprises the faces of people, with each image accompanied by the name of the person in question. This causes an unequal distribution of images in the different classes, with the number of images depending on the number of separate images the individual comprises.

Fig. 8 Random images of identity from the LFW dataset

5.2. YTF Dataset

YTF, Also known as the YouTube Faces Dataset, is a processed set of the YFD, which consists of short YouTube videos of celebrities that are publicly available. For each celebrity, there are several videos. YTF is composed of videos sampled from YouTube, and each video has only one face. All in all, the dataset contains 3425 videos of 1595 different participants. The YTF dataset was largely developed for training and testing the identification of face algorithms. It is reported that it has been employed in the research to construct new advanced facial recognition models and evaluate these models.

Table 1. Labeled faces in the wild and youtube face dataset data

Datasets	Identity	Image
LFW	5,751	13,267
YTF	1,595	621316

In this section, the results of the experiments on face recognition are presented using the proposed model. The process starts by giving out the measured performance of the proposed model on LFW and YTF datasets. Out of the dataset of LFW & YTF, data of 50 people has been used for testing on the proposed model.

By creating a graph of sorts, in which the image distribution of each class (person) in the LFW & YTF dataset will be represented through a bar plot, one will have a clear view of how the dataset is set up. It is seen from below figure that the distribution of the classes (persons) along with the number of images is as follows: The '|' can be interpreted to mean a new record, and the '-' represents different variables pertaining to another different person in the same population. Bar height is proportional to the quantity of images provided for the particular person.

The proposed model for facial recognition comprises several vital parts that are needed to settle for the right identification. First, YOLOv3 is used to locate faces in each of the input images; second, Dlib guarantees alignment for further operation. After alignment, a Convolutional Neural Network (CNN) is used to perform 128-dimensional embedding, resulting in L2 normalization and containing the unmatched facial signature of every person. These embeddings are important because, unlike the previous embeddings, faces can be compared as vectors in the embedding space using Euclidean distance.

In order to do face recognition, the input embedding is matched to labeled ones that exist in the database. This comparison is made using Support Vector Machine (SVM) classifiers with the labeled embedding used to train the classifiers. These classifiers are useful in predicting the identity of the new inputs that are used in the face recognition process. This proposed model is hoped to improve the performance and accuracy of facial recognition tasks by having proper preprocessing and proper methods of applying deep learning algorithms.

Approaching the model for face recognition proposed in this paper toward LFW and YTF datasets justifies their viability. It is easy to assess the level of accuracy of the model under consideration with the help of a confusion matrix. The confusion matrix gives clear information about the performance of the classification models with true positive, false positive, true negative and false negative. It is used in the assessment of the efficiency and correctness of the classification model as it relates to various classes.

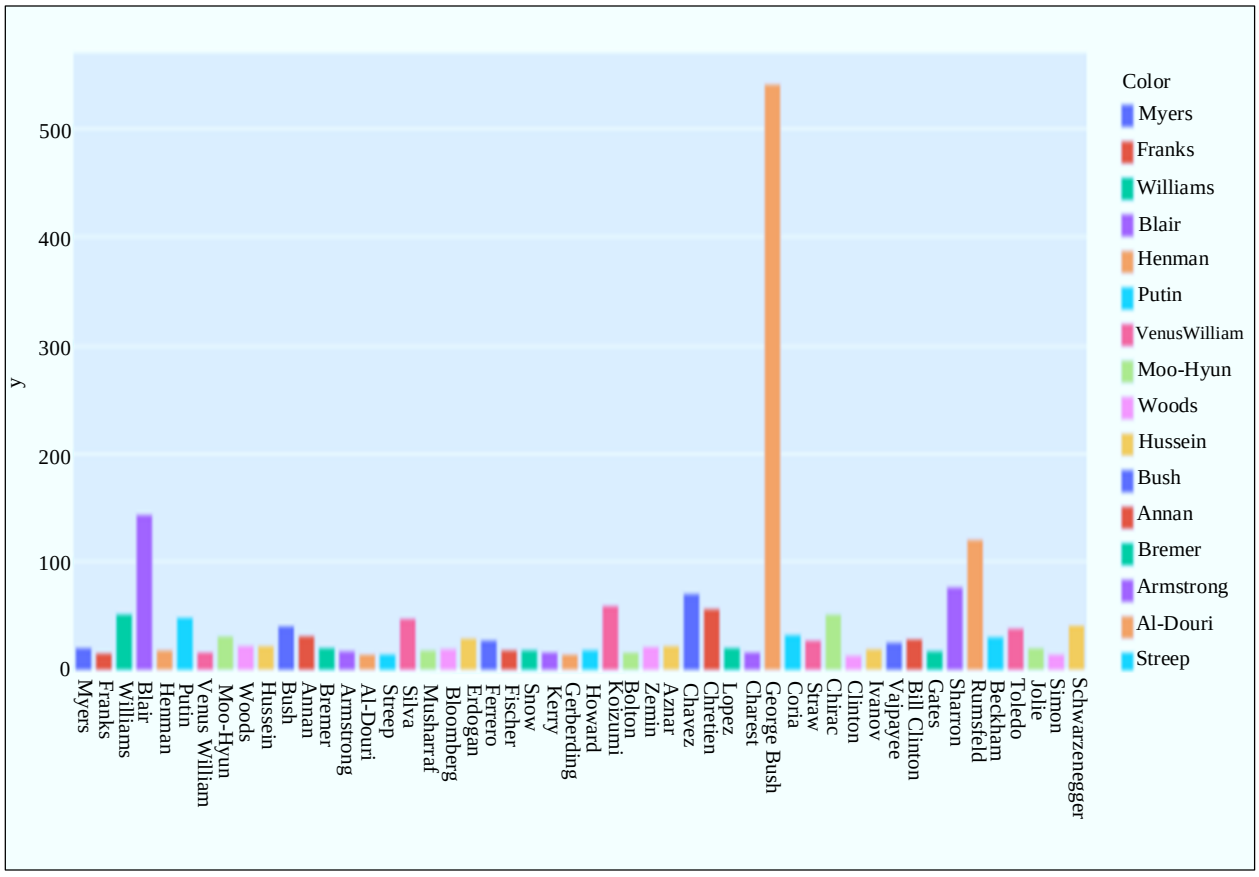

Fig. 9 Distribution of classes with a number of images (LFW)

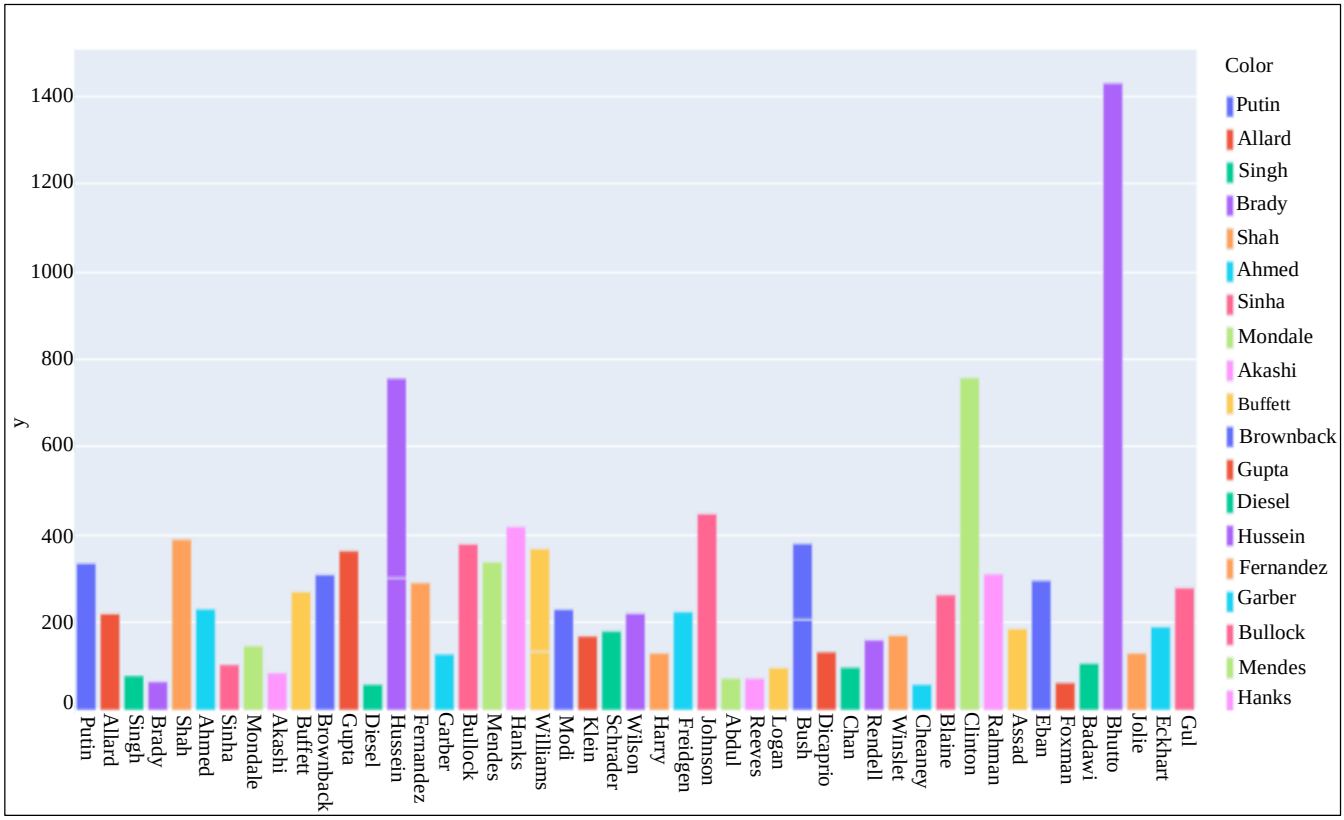

Fig. 10 Distribution of classes with number of images (YTF)

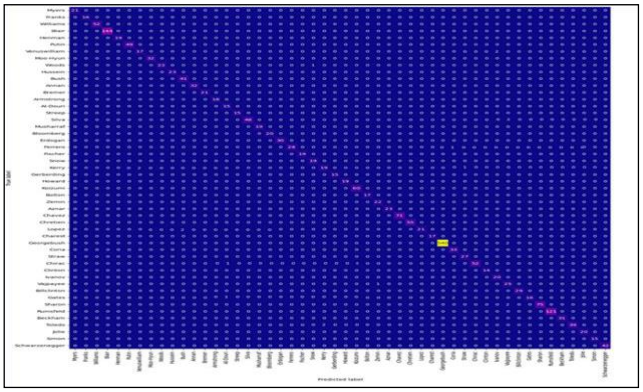

Fig. 11 Confusion matrix for LFW

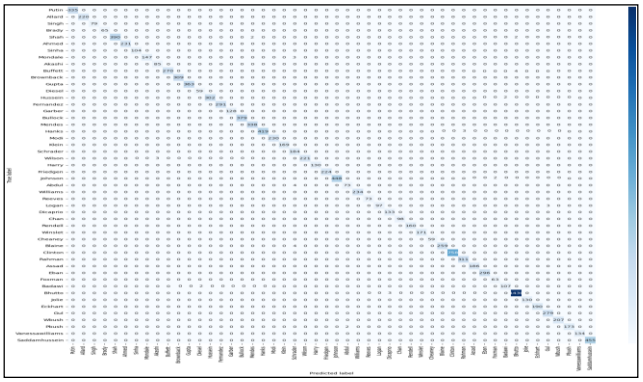

Fig. 12 Confusion matrix for YTF

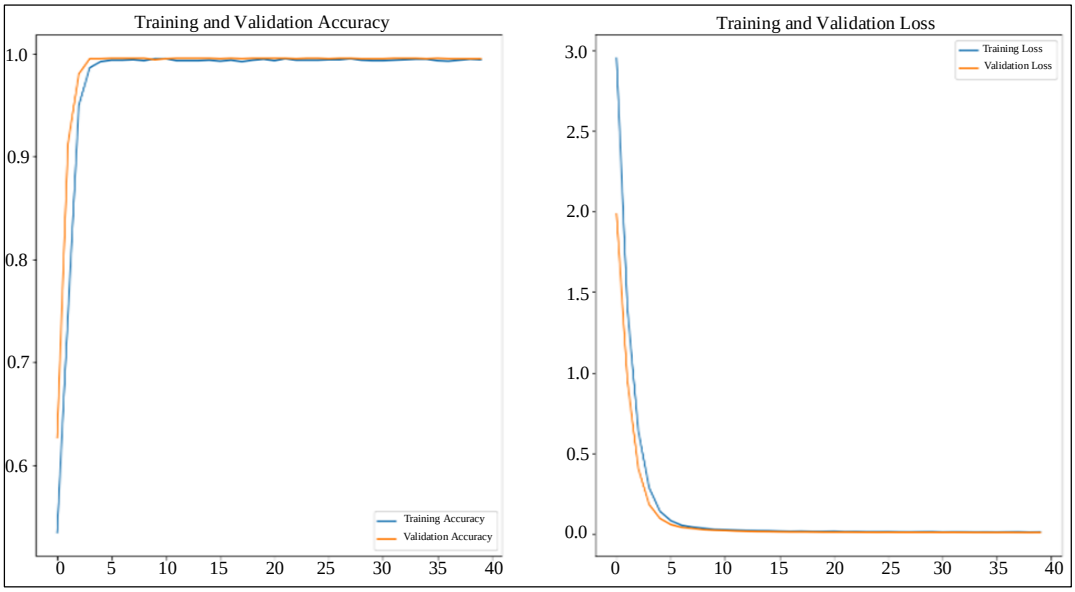

Fig. 13 Accuracy and loss on LFW dataset

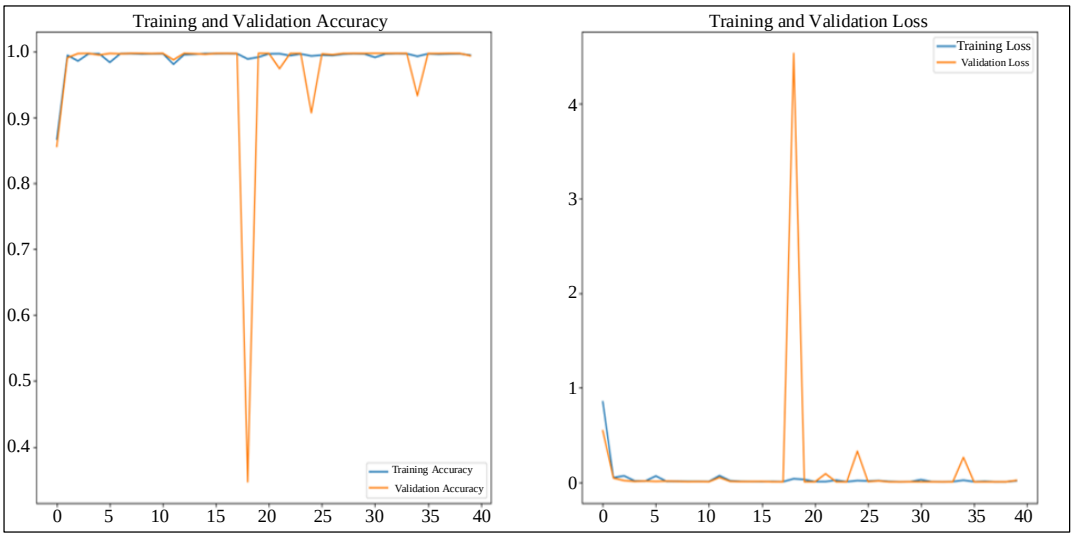

Fig. 14 Accuracy and loss on YTF Dataset

Table 2. Comparison of the proposed method on the YTF and LFW dataset

Method	YTF	LFW
coseFace [10]	96.1	99.33
FaceFilter [31]	94.02	99.70
HiReST-9+ [32]	95.40	99.03
Proposed Method	96.75	99.50

As shown in Table 2, the proposed model achieves state-of-the-art results of 99.50% on LFW and 96.75% on YTF Dataset. After training and testing the face recognition model, the model was used on two datasets, namely YouTube Faces Dataset (YTF) and Labeled Faces in the Wild (LFW). A face was also detected and recognized in the videos with the help of the model.

In the below figure, the result shows the frame that has face recognition from the video of the YTF and LFW Dataset.

Fig. 15 Face recognition from video in LFW dataset

Fig. 16 Face recognition from video in YTF dataset

6. Conclusion

In numerous research papers exploring face recognition using deep learning techniques on datasets like YouTube Faces (YTF) and Labeled Faces in the Wild (LFW), a prevailing approach combines Convolutional Neural Networks (CNNs) for feature extraction and SVM for classification.This hybrid model harnesses CNN's ability to learn discriminative features from facial images, capturing complex patterns and representations. Studies consistently highlight the effectiveness of this combined methodology in achieving notable accuracies on challenging face recognition tasks. Moreover, the utilization of CNNs for extracting hierarchical features followed by SVM's non-parametric approach showcases promising results, providing a robust and interpretable solution for face recognition.

The proposed method achieves different ways to identify facial features from images and videos and proposes methods to improve accuracy, achieving 99.50% and 96.75% on the LFW and YTF datasets.

References

[1] Kanggeon Kim et al., "Face and Body Association for Video-Based Face Recognition," *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Lake Tahoe, NV, USA, pp. 39-48, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[2] Nate Crosswhite et al., "Template Adaptation for Face Verification and Identification," *Image and Vision Computing*, vol. 79, pp. 35-48, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[3] Carolina Todedo Ferraz, and Jose Hiroki Saito, "A Comprehensive Analysis of Local Binary Convolution Neural Network for Fast Face Recognition in Surveillance Video," *WebMedia '18: Proceedings of the 24th Brazilian Symposium on Multimedia and the Web*, pp. 265-268, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[4] Jun-Cheng Chen et al., "Unconstrained Still/Video-Based Face Verification with Deep Convolutional Neural Networks," *International Journal of Computer Vision*, vol. 126, pp. 272-291, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[5] Steve Lawrence et al., "Face Recognition: A Convolutional Neural-Network Approach," *IEEE Transactions on Neural Networks*, vol. 8, no. 1, pp. 98-113, 1997. [CrossRef] [Google Scholar] [Publisher Link]

[6] Peng Lu, Baoye Song, and Lin Xu, "Human Face Recognition Based on Convolutional Neural Network and Augmented Dataset," *Systems Science & Control Engineering*, vol. 9, no. 2, pp. 29-37, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[7] Jiankang Deng et al., "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4690-4699, 2019. [Google Scholar] [Publisher Link]

[8] Weiyang Liu et al., "SphereFace: Deep Hypersphere Embedding for Face Recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 212-220, 2017. [Google Scholar] [Publisher Link]

[9] Florian Schroff, Dmitry Kalenichenko, and James Philbin, "FaceNet: A Unified Embedding for Face Recognition and Clustering," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 815-823, 2015. [Google Scholar] [Publisher Link]

[10] Hao Wang et al., "Cosface: Large Margin Cosine Loss for Deep Face Recognition," *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5265-5274, 2018. [Google Scholar] [Publisher Link]

[11] Ran He et al., "Wasserstein CNN: Learning Invariant Features for NIR-VIS Face Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 7, pp. 1761-1773, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[12] Yibo Ju et al., "Adversarial Embedding and Variational Aggregation for Video Face Recognition," *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, Redondo Beach, CA, USA, pp. 1-8, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[13] Iacopo Masi et al., "Learning Pose-Aware Models for Pose-Invariant Face Recognition in the Wild," *IEEE Transaction Pattern Analysis Machine Intelligence*, vol. 41, no. 2, pp. 379-393, 2017. [CrossRef] [Google Scholar] [Publisher Link]
[14] X. Wang et al., "Deep Discriminative Feature Learning for Face Verification," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
[15] Yaniv Taigman et al., "DeepFace: Closing the Gap to Human-Level Performance in Face Verification," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1701-1708, 2014. [Google Scholar] [Publisher Link]
[16] Zhen Dong et al., "Deep CNN Based Binary Hash Video Representations for Face Retrieval," *Pattern Recognition*, vol. 81, pp. 357-369, 2018. [CrossRef] [Google Scholar] [Publisher Link]
[17] Zhiwu Huang et al., "Cross Euclidean-to-Riemannian Metric Learning with Application to Face Recognition from Video," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 12, pp. 2827-2840, 2018. [CrossRef] [Google Scholar] [Publisher Link]
[18] Ahmed Jawad A. AlBdairi, Zhu Xiao, and Mohammed Alghaili, "Identifying Ethnics of People through Face Recognition: A Deep CNN Approach," *Scientific Programming*, vol. 2020, no. 1, 2020. [CrossRef] [Google Scholar] [Publisher Link]
[19] Asifullah Khan et al., "A Survey of the Recent Architectures of Deep Convolutional Neural Networks," *Artificial Intelligence Review*, vol. 53, pp. 5455-5516, 2020. [CrossRef] [Google Scholar] [Publisher Link]
[20] Shilpi Singh, and S.V.A.V. Prasad, "Techniques and Challenges of Face Recognition: A Critical Review," *Procedia Computer Science*, vol. 143, pp. 536-543, 2018. [CrossRef] [Google Scholar] [Publisher Link]
[21] C. Liu et al., "Semantic Face Parsing with Occlusion Aware Network," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
[22] X. Wang et al., "Deep Discriminative Feature Learning for Face Verification," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
[23] R. Zhang, J. Tang, and J. Zhang, "Deep Mutual Learning," *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.
[24] H. Kharroubi, E. Granger, and A. Hadid, "Deep Learning for Face Recognition: A Comprehensive Review," *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2130-2134, 2017.
[25] Yi Sun et al., "Deep Learning Face Representation by Joint Identification-Verification," *Advances in Neural Information Processing Systems (NIPS)*, pp. 1-9, 2014. [Google Scholar] [Publisher Link]
[26] Zeynep Batmaz et al., "A Review on Deep Learning for Recommender Systems: Challenges and Remedies," *Artificial Intelligence Review*, vol. 52, pp. 1-37, 2018. [CrossRef] [Google Scholar] [Publisher Link]
[27] Jiuxiang Gu et al., "Recent Advances in Convolutional Neural Networks," *Pattern Recognition*, vol. 77, pp. 354-377, 2018. [CrossRef] [Google Scholar] [Publisher Link]
[28] Ahmed A. Elngar et al., "Image Classification Based on CNN: A Survey," *Journal of Cybersecurity and Information Management*, vol. 6, no. 1, pp. 18-50, 2021. [CrossRef] [Google Scholar] [Publisher Link]
[29] Sakshi Indolia et al., "Conceptual Understanding of Convolutional Neural Network- A Deep Learning Approach," *Procedia Computer Science*, vol. 132, pp. 679-688, 2018. [CrossRef] [Google Scholar] [Publisher Link]
[30] Shrey Srivastava et al., "Comparative Analysis of Deep Learning Image Detection Algorithms," *Journal of Big Data*, vol. 8, no. 66, 2021. [CrossRef] [Google Scholar] [Publisher Link]
[31] Mohammed Alghaili, Zhiyong Li, and Hamdi A.R. Ali, "Facefilter: Face Identification with Deep Learning and Filter Algorithm," *Scientific Programming*, vol. 2020, no. 1, 2020. [CrossRef] [Google Scholar] [Publisher Link]
[32] Wanglong Wu et al., "Recursive Spatial Transformer (REST) for Alignment-Free Face Recognition," *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 3772-3780, 2017. [Google Scholar] [Publisher Link]