International Journal of Engineering &Technology, 7 (4) (2018) 6181-6188

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET
doi: 10.14419/ijet.v7i4.23127
Research paper

Face recognition using deep learning methods a review

Othman. I. Hammadi [1] *,[2] Abdulkarim Dawah Abas [3], Khaled Hammad Ayed [3]

[1] College of Education for Humanities Department of English, University of Anbar, Ramadi, Iraq
[2] College of Computer Science and Information Technology, University of Anbar, Ramadi, Iraq
[3] Al-maaref University College, Department of Engineering of Computer Technology Ramadi, Iraq
**Corresponding author E-mail: ed.osman.ibrahim@uoanbar.edu.iq*

Abstract

Face recognition is one of the most challenging field of image analysis and computer vision due to its wide practical applications in the areas of biometrics, information security, law enforcement and surveillance systems. It has been a topic of active research proposing solutions to several practical problems giving rise to the significant amount of research in recent times aimed at addressing the challenges of face recognition attributed to the following factors such as illumination, emotion, occlusion, facial expressions and poses, which greatly affect the performance in achieving efficient and robust face recognition systems. In this field, many researchers adopted different techniques that solely rely on extracting handcrafted features to achieve better results. Recent development in deep learning and neural networks have made it possible to achieve promising results in numerous fields including pattern recognition and image processing. Deep learning methods boost up the learning process and facilitates the data creation task. Many algorithms have been developed to use deep learning architectures to get maximum result and achieve the state-of-the art accuracy. Some algorithms design their architectures from scratch and others fine-tuned the existing models to get maximum efficiency of generalization power. Algorithm complexity, data augmentation and loss minimization are the main concern of deep learning paradigms. We have reviewed these architectures in relation to algorithm complexity and experimental results on benchmark dataset. In this paper, we presented a literature survey of latest advances in researches on machine learning for face recognition and their experimental results on public databases.

Keywords: *Face Recognition; Deep Learning; Face Identification; Face Verification.*

1. Introduction

Face recognition has been one of the most actively studied topic in the computer vision community. With the advancement in technology and the high usage of multimedia application in smartphone, the challenge for face detection and efficient recognition is greater than before. There has been great advancement in face recognition, starting with the Viola Jones as pioneer work for detecting frontal-face in real time along with low computational complexity. This was followed by different approaches that involved basic image processing techniques that would extract various facial features from the face images and were fed to different classifiers for training and recognition. Other than this most of the initial approaches in the area of face recognition used up-right images without much variations in pose, illumination, occlusions etc.

Although the initial approaches worked well for the front-face images but failed with different angles or illuminations. The various other classifiers used lack the ability to classify multiview facial features. This led to other approaches for multi-view face recognition and approaches that would eliminate the hand crafted facial features. Face recognition can be broken down into two steps stated below:

Step 1: Identification

Recognizing individual by locating their faces in a given image is the first step in Face recognition system. The identification tep ensures that the algorithm identifies the image as a facial image and then utilizes this information to identify the faces in the image. The identification step checks for the face in the image against the other faces to look for the identity of the face in the image, which makes this a multiclass classification problem.

Step 2: Verification

The verification is concerned with validation of identity based on the input image of a face. It performs a one-to-one matching by either accepting or rejecting the identity which makes this a binary classification problem.

2. Deep learning

It is one of the machine learning methods that is inspired by the neural networks. It involves different models based on the neural networks and it uses multiple layers for feature extractions. Its layer sequence or architecture is such that each layer input is an output from the previous layer. It learns in a supervised as well as unsupervised manner[1]. Deep learning has many models like Deep Neural network, recurrent neural network and convolutional neural networks.These models have improved the ability of classification, recognition, detection and localization. Deep learning is now being advancing by the development of new machine learning approaches, new versions of neural networks and increasing computational powers like GPU.

Applications of deep learning

1) Image recognition: Deep learning has been widely used in computer vision and it has shown promising results in this area. Practical problems and many computer vision challenges are now being addressed with deep learning producing better results than humans.

Copyright © 2018 Othman. I. Hammadi et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

2) Speech Recognition: Deep learning has shown convincing results on speech recognition. For this purpose, recurrent neural networks are used. Today all commercial speech recognition system like Baidu, Google now and Skype translator are based on deep learning.
3) Natural Language Processing: LSTMs a recurrent network has been using for many natural processing tasks like sentiment analysis, word translation and many more and they are producing outstanding results on these problems.
4) Recommendation Systems: Recommendation systems are using deep learning to learn user interest and preferences and to recommend their interests.

Why deep learning for face recognition

Traditionally features have been extracted from the images using different image descriptors like SIFT, HOG or a hybrid descriptor. In this method, we explicitly convolve different filters on the image to detect features (edges, line, shapes etc.) that are discriminant enough in recognition tasks. In deep learning, it's a lot easier. CNN has convolutional layers that convolve the filters on the image to get the feature maps, these feature maps are passed in sequence to other convolutional layers until the CNN predicts the label, the error gets back propagated and the model learns weights that best fit the model using gradient descent as an optimization algorithm.

Deep learning has allowed the automation of the process of selecting the filters that extracts the best features from the image to give the best accuracy on the dataset. It has been observed that deep learning shows better accuracy on face recognition problem as they have more parameters to learn the details of the dataset, but it has a drawback of overfitting the model for the dataset but this has been handled by using the dropout and regularization techniques.

3. Literature review

This paper [2] focuses on the detection of Multi-view faces. Despite many extensive studies, many techniques have been used but still require annotations of the facial landmarks in addition to multiple trained models to learn the faces in different orientations. This paper proposes a Deep Dense Face Detector that does not need to annotate the faces and it can detect faces in many different orientations by just using a single model based on deep convolutional neural network. Additionally, it does not use any segmentation, bounding box, or regression like[3]. It does not use any classifiers like SVM as used in [4]. A. Krizhevsky et al proposed model that uses a pre-trained [5]. It is a CNN model that took part in ILSVRC- ImageNet large scale visual recognition challenge; they just fine tuned it for their face detection problem. They trained the model on Annotated Facial Landmarks in the Wild (AFLW) dataset that contains 21K images with 24K face annotations. To increase the positive examples in the dataset a random sub indow of the images were created and used as positive examples in the dataset. Other techniques used to increase the positive examples include the following:

- Image Transformation: flipping the images to generate more examples
- Resizing images: Images were resized to 227 x 227 and used to fine tune the pre-trained AlexNet model.

The Modelled architecture has similarity to the architecture of AlexNet which contains 8 layers namely:

- 5 convolutional layers and
- 3 fully connected layers.

They made few changes inside the layers by converting the fully connected layers to the convolutional layers which is achieved by reshaping the layer parameters; in addition, obtained heat-maps (heat-map is the class activation map, highlighting the importance of the image region in the prediction. They localize the interested regions in the image) out of it. The regions detected by the heat-maps were further processed my non-maximal suppression for accurate localization of the faces. (Non-maximal suppression is a technique that eliminates the multiple detections of the same face and finds the accurate one.

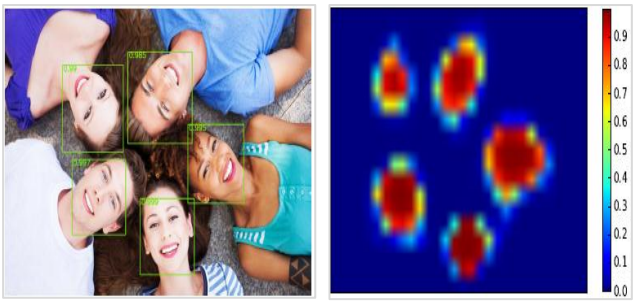


Fig. 1: (Left) Faces with Different Rotations and the Confidence Score with Each Detection (Right) Heat-Map for the Scores of the Image on the Left.

Error! Reference source not found. (Left) shows that the confidence of the detector on the faces in different orientations is very high which is close to 1. And it can also be seen that heat map shows the regions other than the faces have very low scores that are close to zero in **Error! Reference source not found.** (Right). It's seen that upright faces have high score of 0.999 while faces with in-plane rotation have less score. We hypothesize this trend not because of difficulties in detecting rotated faces but due to lack of training of model with rotated faces.

The proposed method has been compared with R-CNN and other methods; the former shown better results even without annotations of face landmarks. Results can be further improved by using better sampling and data augmentation techniques [6] made an improvement on faster RCNN framework to get promising results on Face Detection Dataset and Benchmark (FDDB). [7] presents a unified system-that achieves Face verification, Recognition and Clustering. The method is based on finding the Euclidean distances which directly correspond to a measure of face similarity. Once the Euclidean distance is created, face verification, recognition and clustering can be successfully achieved.

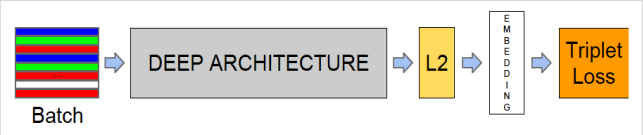


Fig. 2: Structure of the Model with an Architecture and the Distance Embedding Passed to the Triplet Loss During Training.

Other approaches include training the network over a dataset and then taking an intermediate layer for general recognition. The dimension of this intermediate layer is usually very high. However, FaceNet trains the output to be compact with the dimension being 128. To achieve this a triplet loss function is being used. The triplets include:

Two matching face thumbnails and

A non-matching face thumbnail.

The objective is to create a distance margin between the positive (similar) pair and the negative (dissimilar) thumbnail.

For training two architectures are being used. The models adds 1 x 1 x d convolutional layers in the standard layers used by [8] architecture – making it a 22 layer deep model. A total of 140 million parameters. The Inception model of GoogLeNet is being used. The trained model is then evaluated on four datasets using the following technique:

- Give a pair of images and a threshold
- Determine the classification of being same or different based on the squared distance between the pair of images given.

A hold out test set of 1million images is kept, with the same distribution as the training set, but disjoint identities. This set is then split into 5 disjoint sets (200K images each) and the standard error is reported for the 5 splits. Inception models reduce the size dra-

matically. All the models perform well, however the inception-based model NN3 performs significantly well. The figure below shows the performance of the models on the personal photos set which consists of 12K images, which has clean labels and is manually verified and consists of three personal photo collections.

Another approach proposes a method to learn high level feature representation called Deep hidden IDentity features called as DeepID for face verification task [9, 10]. They argue that face verification task can be done by learning DeepID through solving challenging multi-class identification task. The description of their dataset is given below:

Learning DeepID

Training set: 80% CelebFace (4349 people) (randomly select)

Validation set: 10% images of each training person (randomly select)

Learning Joint Bayesian

Dataset	People	Images
LFW	5749	13233
CelebFace	5436	87628
CelebFace	10177	202599

Training set: remaining 20% CelebFace (1400 people)

Testing set: all LFW pairs (6000 pairs)

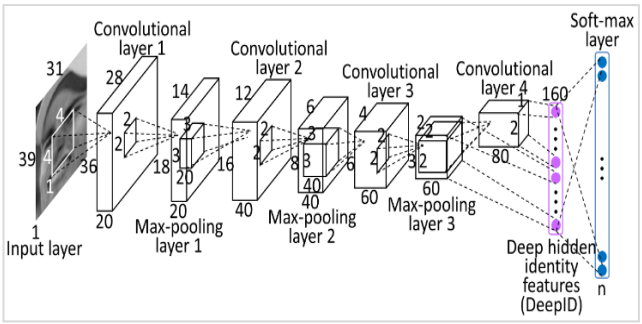


Fig. 3: Structure of the Convolutional Neural Network Used by This Paper.

4. Deep conv nets

Proposed model contains 4 convolutional layers for feature extraction which are fully connected to DeepID layer and soft-max layer that predicts the class. The below ConvNet (Fig. 3) takes an image of size 39 x 31 x 1 and predicts n identity classes (n can be equal to 10,000). Below steps are utilized:

- Step1: Feature extraction from 60 face patches with ten regions, three scales, and RGB or gray channels. They trained 60 ConvNets with same architecture used for the face identification, each of which extracts two 160-dimensional DeepID vectors from a particular patch and its horizontally flipped counterpart. The total length of DeepID becomes 19, 200 (160 x 2 x 60), which is ready to input to the classifier for face verification.

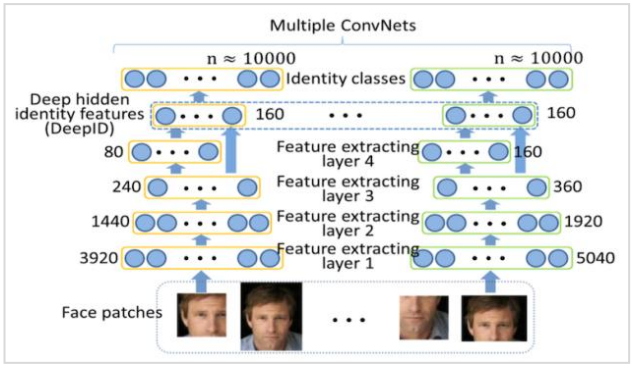


Fig. 4: Structure Overview with Dimensions

- Step 2: Feature recognition. Joint Bayesian technique based on the DeepID was used.

The author also trained a neural network for verification and compared it to Joint Bayesian to see if the model performs well in learning the extracted features. The feature dimension is reduced to 150 by PCA before learning the Joint Bayesian model. Joint Bayesian based on our DeepID features achieves 97.45% test accuracy on LFW, which is competitive with the human-level performance of 97.53%.

[11] Proposed another approach and offered two main contributions to the study:

1) To show how a large-scale datasets can be assembled.
2) To avoid the complexities of deep leaning network training by presenting procedures to achieve result that can be compared with the start of art results.

Previously local descriptors were used for feature extraction like SIFT, LBR etc. followed by some pooling method to create a local face descriptor and passed it to the state of the art machine learning algorithms for classification.

Data Collection comprises of 5 steps as discussed below:

1) [Bootstrapping and filtering a list of candidates for data creation]: For data creation they use google search image of 2.5k male and 2.5k female public figures from IMDB.
2) [Collection of each entity]: Around 2000 images per entity were collected.
3) [Auto filtering]: SVM (One vs all) is trained over top 50 google searched entities are used as ground truth and all other entities results are labelled as negatives. So erroneous values are discarded from search.
4) [Duplication removal]: VLAD descriptor used to remove duplications regarding data as they were collected from multiple sources. VLAD descriptor is then clustered with tight threshold and based on its results, images were removed which lie in the same cluster.
5) [Final Manual Filtering]: This step adopts CNN's Alex architecture based ranking model to minimize human intensive annotation work.

The classification process is described as follows:

- Input size must be 224x224 Dimensional face image.
- 11 block Convolutional neural network is then adopted for classification.
- First 8 blocks are convolutional layer followed by Rectified Linear Unit (RELU) and last 3 layers are fully connected.
- Last layer (for classification) has N = 2622 which is the total number of entities in the database for classification. 'A' Conv-Net configuration from "Very Deep Convolutional Network for large scale image data set" used in their work which is then modified by replacing soft-max layer.

For data enrichment in training they perform data augmentation. Augmentation involves flipping of images horizontally and/or vertically and no other colour based augmentation was performed.

Hand-crafted methods like Local Binary Pattern (LBP) and Local Phase Quantisation (LPQ) are well recognized for face recognition. excellent performance has been achieved by[12] but fails in unconstrained environments. Intra-personal variations make it very complex for recognition. So it's an open problem to get an ideal facial features under face recognition in unconstrained environments (FRUE).

Deep learning features are more robust to these intra-personal variations due to generalizing power of CNN i.e., larger data set for training, GPU implementations and more effective regularization strategies like dropout. LFW is database for FRUE. They have proposed 3 architectures for training and a local patch fusion method named as C Fusion and A-Fusion. Feature fusion outperforms as compared to no-fusion and score-fusion metric. LFW dataset was used for training. The dataset is small to avoid overfitting they proposed three architectures named as CNN-S, CNN-M and CNN-L.

CNN-S and CNN-M has 3 convolutional layer while CNN-L has 4 convolutional layer followed by 1 FC layer and a Soft-max layer.

CNN-S: 1st convolutional layer contains 12 filters of size 5x5, 2nd conventional layers contains 24 filters of dimension 4x4 then 32 filters of dimension 3x3.

CNN-M: 1st convolutional layer contains 16 filters of size 5x5, 2nd conventional layers contains 32 filters of dimension 4x4 then 48 filters of dimension 3x3.
CNN-L: 1st convolutional layer contains 16 filters of size 3x3, 2nd conventional layers contains 16 filters of dimension 3x3, 32 filters of dimension 3x3 and then 48 filters of dimension 3x3.
FC layer with 160 neurons followed by 5000 classes in soft-max layer. Their aim is to pull images on the basis of having similar intra-personal variations and push images on the basis of inter-personal variations. Joint Bayesian (JB) model is used for that purpose. They have a small dataset so data augmentation is introduced. Flipping and mirroring images are mainly used as augmentation metric. They applied their network on grey vs color images of depth three. Gray images shown performance upgrade as compared to color images. Features from top layers are more discriminative then bottom layer so they consider top 3 layers and FC and soft-max layer for architecture for testing. To reduce features dimensionality PCA is applied on feature vector to boost up computational power with comparable results.
They get 3 single layer and combine it with FC, Soft-max, Soft-max + FC, Soft-max + FC + JB. And fusion of (Soft-max + FC + JB) outperforms in accuracy measure. JB improves further face recognition.
[13] focus on hard negative mining and iteratively update Faster RCNN for better performance. Challenges in face detection are luminance, occlusion, facial expression and expression change. Performance can be increased by reducing number of false positives. These false positives are reduced by hard negative mining as proposed in this paper. Region proposal selects background (bg) and foreground (fg). With background proposals having Intersection over Union (IoU) minimum overlap with ground truth bounding box threshold value is 0.5 and less than that considered as background proposal while greater than that is considered as foreground (RoI) region. In an image 100:1 region are extracted but to reduce computational cost and data imbalance they use bg/fg ratio as 3:1.
The architecture followed is same as the RCNN the only change is in its re-training of negative image who has Intersection over Union (IoU) less than 0.5. Proposals with IoU less 0.5 are reconsidered in training process maintaining same bg/fg ratio i.e., 3:1.
Faster RCNN have 2 modules. 1st is Region Proposal Network (RPN) which selects RoI and then passed it to Fast RCNN for recognition. And here false positive images are then re-trained to network for better performance. This method outperforms over previous state of the art algorithms.
[14] presents a face representation scheme MultiModal Deep Face Representation (MM-DFR). This scheme uses CNN to extract features of a face. The extracted features are then concatenated as a raw feature vector, and the dimension is reduced using 3-layer stacked auto-encoder (SAE).
There are 9000 subjects for the face dataset. The proposed Multimodal face recognition method extracts multimodal features from:holistic face image, rendered frontal face by 3D face modeluniformly sampled image patches.
The 3D model is used to sample small patches of the face and render a frontal face in 3D domain. The SAE is used to compress the high dimensional deep feature into a compact face signature. SAE can learn non-linear transformations for reduction. The MMDFR consists of 2 steps first being the Multimodal feature extraction using a set of CNNs and second being the feature-level fusion of the set of CNN features using SAE.
Single CNN Architecture: The face images are normalized to 230 x 230 pixels with an affine transformation according to the five facial feature points: both eye centres, nose tip, mouth corners (both ends). The normalized image gives one holistic image of size 165 x 120 pixels, and six image patches of size 100 x 10.
There are seven CNNs that extract features from holistic images and from the small image patches as shown in the figure below:



Fig. 5: Images for Feature Extraction Using Cnns, (Left) Original Image on Top and Holistic Image Bottom, (Right) Small Image Patches.

CNN-H1: Extracts features from the holistic face image. NN1: It has 10 convolutional layers, 4 max-pooling layers, 1 mean-pooling layer, and 2 fully-connected layers. NN2: This CNN has 12 convolutional layers compared to NN1. This is deeper than NN1 and is more robust to nonlinear features compared to NN1, so NN2 is applied to the holistic networks. NN1 is small and efficient, and is applied to the rest of the networks for the small patches. The filter size is small i.e. 3x3. The activation function being used is ReLU (Except for the last layer, this helps in generating dense features). The second last layer (Fc6) maps convolutional features from dense to dense, reducing them (this acts similar to PCA or LDA) – this favours sparse features. The output of this layer is treated as face representation. The dimension of last layer (Fc7) is set to 9000 – same as the number of subjects in the train set. Dropout is being used for the first fully connected layer at a value of 0.4.
CNN-H2 is used to extract features from the other holistic image rendered by OpenGL with the help of 3D generic face model. CNN-P1, CNN-P2 up to CNN-P6, extract features from the six image patches. These CNNs have the same structure. For sampling small image patches uniformly, the pose invariant face recognition approach [15] is used. 3D landmarks are manually labelled on a generic 3D face model (9 landmarks in total). The 3D landmarks are uniformly spread across the face.
The technique for patch sampling is as follows:
Using the 2D face image, it is aligned to the 3D face model with the help of the five features using orthogonal projection.
The 3D landmarks manually labelled are then projected on to the 2D image.
A patch of size 100 x 100 is then cropped, which is centred around the landmarks.
The figure below shows the detected 2D landmarks. The patches are uniformly sampled semantically regardless of the pose variation of the face in the image.

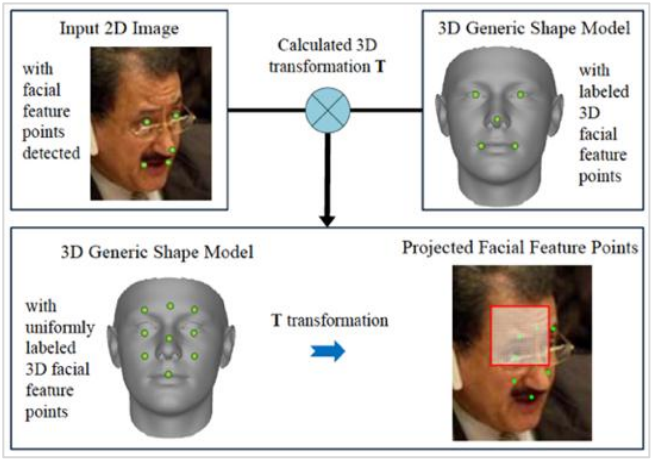


Fig. 6: Process for Getting Landmarks onto 2D Images (Left).

[16]focus on how big data impacts face recognition task and proposed Megvii Face Recognition System that uses a simple deep convolutional network without any tuning tricks for the purpose of face recognition. It collects a large amount of labelled web data

called Megvii Face Classification (MFC) database. Two critical observations were made; first, data size and its distribution affects the recognition accuracy and secondly performance of many good methods has been decreased by increasing the size of the data.

The paper developed a very simple deep neural network architecture that is used for multi-class classification. The deep neural network contains ten convolutional and pooling layer pairs with the last layer of softmax classifier in the training phase for learning purposes. The output of the last layer before the softmax layer is considered as the features of the input image. It has been observed that large amount of data significantly enhances the performance of the system. It has been observed by training the same network with different number of training samples from 4000 to 16000. Four face regions have been cropped and sent as an input to the model and their output has been combined that is a representation of the image and it is passed directly to a softmax multi-class classifier in the training phase, but for the testing phase PCA has been applied on these combined features for feature reduction, after that a L2 norm has been used to measure the pair of testing faces.

The Network has surprisingly achieved 99.50% accuracy of LFW dataset by using labelled web data during training phase that surpassed other state of the art methods. Paper tested the system on another benchmark called Chinese ID (CHID), it is a real security environment data that has been collected to test the system's generalization. Unfortunately, it does not perform well on the real environment data CHID with only 66% accuracy.

[17] proposed an object localization task which is achieved using FAST R-CNN with the help of region proposals. Generating proposals is also a time computational process as detection. But if region proposal task is neglected then FAST RCNN reduces detection time for object detection as well. So here bottleneck is region proposal method. Moreover, detection task is done via GPU implementation pre-trained are available to ease this task so this process is speedy due to GPU's computational power and generic pre-trained models, VGG16 reported 5 frames/second processing time. While region proposal was achieved via CPU implement so there is a lot of room for improvement.

To implement region proposal with GPU implementation detection part will be neglect and will lose the essence of shared network. So to minimize region proposal task's computation this paper came up with a shared network based approach called Region Proposal Network (RPN). This approach's implementation takes negligible time for region proposal as it uses shared convolutional features of same architecture they are using for detection by FAST RCNN. RPN produces high quality proposal because they introduce learning mechanism into it. They use back propagation and Stochastic Gradient decent for optimization. Advancement in this paper is about introduction of RPN along with FAST RCNN.

VGG has 13 sharable layer while ZF has 5 sharable layers. They are pre-trained model and is used for both RPN and FAST RCNN in this paper. Architecture differs after last convolutional layer. 2 convolutional layers were proposed on top of last convolutional layer of pre trained model. A sliding window is introduced on top of last convolutional layer, on each pixel point it produces 256d feature vector for ZF model while 512d feature vector for VGG model. Based on that vector next layer have 2k score classification layer and 4k regression layer. Here k is number of proposals.

For classification each point is a binary class problem i.e., object or non-object. If Intersection over Union (IoU) value is greater than 0.7 than RPN consider it as a positive object and it IoU value is less than 0.3 than RPN consider it as negative object or non object. In Regression layer anchors are considered. Anchors are proposed bounding box it has variations. 3 scales and 3 aspect ratios are considered while proposing anchors. Scales are 1282, 2562, 5122 while aspect ratios are 1:1, 1:2 and 2:1. 10k proposal are generated on 1000x600 image. These proposals have overlap in it. So to remove excessive proposal cross boundary proposal are not considered then these proposals are clip down on basis of Non-Maximal Suppression (NMS) method and in last they consider top N ranked proposal. These proposals are then fed into FAST RCNN for more refinement.

[18, 19] proposes two very deep neural network architectures referred to as DeepID3 for face recognition. DeepID3 model comprises of 2 networks based on modifications of VGG and Google LeNet models. Joint Face Verification and Recognition's supervisory verification signals are added to both intermediate and final feature extraction layers of these models. The ensemble of this DeepID3 model showed an accuracy of 99.53 in face verification while 96% accuracy in face identification on LFW dataset. These two nets of DeepID3 reduces the error rate on average to 0.81% and 0.26% respectively [18].

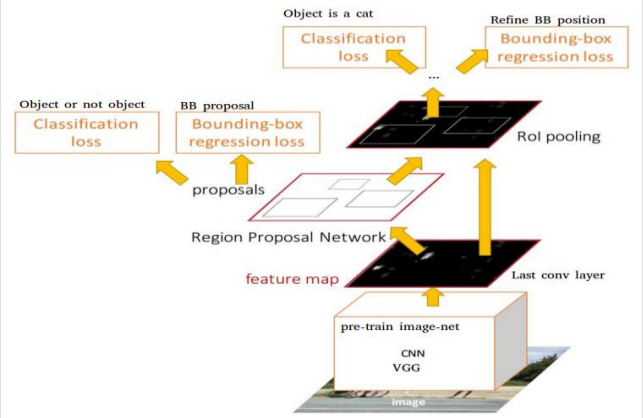


Fig. 7: Flow of the Architecture Used.

Before DeepID3's architecture DeepID2+ architecture achieve 99.47% accuracy which was highest of all in DeepID Series. But its architecture was much shallower than the highest accuracy achieving model in field of object recognition i.e., VGG and GoogleLeNet models are the most accurate in ILSVRC 2014. So they proposed a much deeper network for face recognition based on these two models. DeepID2+ has 4 convolution and pooling layer along with supervisory verification signal. But for this model they proposed two networks, one is based on VGG network and the other is based on Google LeNet.

These models are trained on original data or horizontally flipped but not on both to reduce redundancy. In testing stage, feature extraction takes 50 times of the forward propagation with half features collected from net1 and half from net2. These collective features are concatenated to form feature set of 30,000 dimensionality. To reduce dimensionality they use PCA to get 300 dimensional data on which Joint Bayesian Network is learned for face recognition.

As there are a lot of public available implementations of the CNN but there are no public large datasets available for face recognition tasks, as CNN needs large dataset to not to overfit the model. To solve this problem Dong [20]proposed a method to collect a large dataset from the World Wide Web, that contains 10,000 subjects and 500,000 images called CASIA-WebFace [20]. Based on this dataset proposed, network has obtain the state of the art accuracy on LFW (Labelled faces in the Wild) and YTF (YouTube face database). Input layer's dimension is 100x100x1 channel i.e. grey image. The proposed network includes 10 convolutional layers, 5 pooling layers and 1 fully connected layer. The filters size has been kept to 3x3 throughout the whole network. In order to reduce the network parameter for the efficiency, the network used small filters that approximates the large filters, redundant fully connected layers have also been removed. [20]Combines the tricks from propose network with 10 convolutional layers and just 1 fully connected layer with 3x3 filters for convolution. Pool5 layer is used as face representation as the output from the pool5 layer would be the input to the fully connected layer Fc6, and the dimension of face representation is equal to the number of channels of Conv52, that is 320. Lastly the output from the Fc6 would be used as input of Softmax cost function for classification of the face.

[21] designed a high-performance deep convolutional network (DeepID2+) for face recognition with Sparsity, Selectiveness and Robustness as its main contextual point for DeepID2+ net. DeepID2+ inherited from DeepID2. Last layer has 512 neuron which shows different neurons sets are activated for different faces that leads to selectiveness for face verification and identification. This net is moderately sparse which shows high discriminative power for net. Their architecture is more robust towards occlusion although occlusion is not learned during training.Their architecture was inspired from DeepID2 architecture. DeepID2's architecture has 4 convolutional layers of feature maps 20, 40, 60 and 80 respectively. Supervisory signal layer contains 160 dimensional features on 3rd and 4th layer supervised by both face verification. Their dataset contains 8000 identities. Improvements made by DeepID2+ are 512 dimensional feature layer instead of 160.Dataset is increased from 1, 60,000 training examples to 2, 90,000 examples and the last thing is supervisory signal is added to all convolutional layers instead of last convolutional layer.Dataset used are constructed form LFW and WDRrf Dataset. 25 face regions are selected for training these 25 regions are selected by DeepID2. Results for DeepID2+ surpasses all the related net along with DeepID2.

Face recognition usually consists of four stages, detect face align represent classify.[18, 22] revisited the align and represent stages by doing 3D face modelling and applying transformations and rotations to align the face better, and then it represents the face using nine-layer deep neural network.

Given below are the stages for face recognition employed by the paper.

Input an image

Detect the face and crop it.

Align the face using 3D modelling and affine transformations

Represent the face using feature vector

Classify the face.

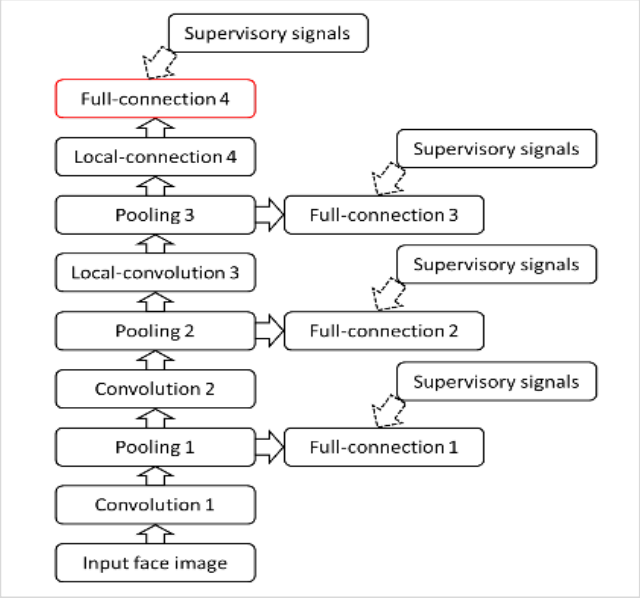


Fig. 8: DeepID2+.

In order to align the face, paper has used 3D model of the face. Initially it has detected six reference points in the image and on the basis of these reference points it has cropped the face, then it has marked 67 reference points and on the basis of it 3D mesh has been created. From 3D mesh 2D image has been created using affine transformations. Result (g) would be the input to the network. 3D-algined image (152 by 152 pixels) is an input to the network which is passed through a convolutional layer, a max-pooling layer, and a convolutional layer. They did not add many pooling layers because they believe that pooling layers remove information about the face. Next three layers in the network are locally connected layers. Network has two fully connected layers and finally it has a softmax classifier for classification of an image to K labels.

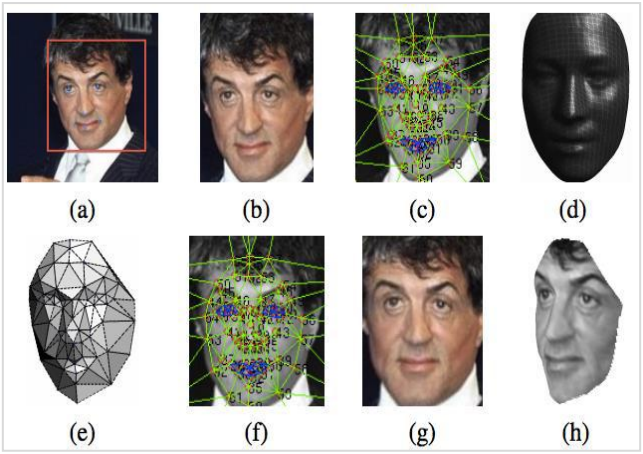


Fig. 9: 3D Model of the Face Image.

It has achieved 97% with single model and 97.35% with ensemble model on the LFW dataset that beats all the previous state of the art models.

The proposed network uses 4 convolutional layers with max pooling layers (for first three layers) [23] Weight Sharing: in the third layer the neuron weights are locally shared for 2 x 2 local regions. For the fourth layer there is no weight sharing.

The network extracts 160 dimensional vector at the last feature extraction layer. DeepID2 layer (fully connected layer) are connected to the third and fourth layer. The fourth layer extracts more global features than the third one so DeepID2 takes multi-scale features as input. Activation function being used is ReLU units for all layers (including DeepID2). Identification is achieved by following the DeepID2 layer with an n -way softmax layer, which outputs a probability distribution over the n classes. The network minimizes the cross-entropy loss (identification loss):

$$\text{ident}(f, t, \theta_{id}) = -\sum_{i=1}^{n} -p_i \log \hat{p}_i = -\log \hat{p}_t$$

Where f is the DeepID2 vector, t is the target class, and θ_{id} denotes the softmax layer parameters. p_i is the target probability distribution, where $p_i = 0$ for all i except $p_t = 1$ for the target class t. $\hat{p}_i$ is the predicted probability distribution. The goal is to learn the parameters in feature extraction function. 21 face landmarks are detected using SDM algorithm. Based on similarity transformation of the detected landmarks the face images are globally aligned. According to the globally aligned faces, 400 face patches were cropped (these varied in positions, scaling, colour channel and horizontal flipping). 400 DeepID2 feature vectors were extracted from a total of 200 deep ConvNets, each of which extracts two 160-Dimensional deepID2 vector for one face patch (and for its horizontally flipped face too).

Forward backward greedy algorithm is used to reduce redundancy in the DeepID2 features to select a small number of effective and complementary DeepID2 vectors. 25 patches that are selected are then used for extracting the 160 dimensional DeepID2 vectors and are concatenated to a 4000-dimensional DeepID2 vector. The 4000 dimensional vector is then compressed for face verification using PCA. Joint Bayesian is learnt for the face verification step. The results are compiled on the LFW dataset [23]

5. Analysis

After analyzing the studies discussed in this review paper, it can be claimed that the deep learning methods for face recognition are efficient and accurate in identifying and verifying the faces in a given image. The highest accuracy achieved on the LFW dataset is more than 99%, by three of the reviewed approaches. The network used by one of the approaches[16] combines multiple CNNs to

analyze four different patches of the image, the resulting image is combined and uses PCA to reduce the features. The second approach [14] to score the highest accuracy of the LFW dataset uses 3D modelling of the face and sampling small patches from the face image, the final vector from the CNN is reduced as well by using a stacked auto-encoder. The approach [7, 24] score the highest i.e. 99.63% on the LFW dataset, utilizes the distance between similar images and non-similar images as the loss function. The structure of their CNN model is based on the architecture used by [8] which leads to a better accuracy due to the understanding of the working at each layer in the model. Through the analysis of these three approaches, the commonality among them is the dimensionality reduction of the output vector from the CNN. Each of the approaches have either used a method for reducing the dimensions or output a compact vector from the CNN. Additionally, two of the approaches make use of sampling image patches and passing them to the network. However, one of the approach [11] out of the three, claims to work poorly on a real world data, although achieving 99.5% on LFW dataset.

Table 1: Accuracy for the Papers in Review with the Dataset Used

Ref	LFW	PASCAL	AFW	FDDB	Youtube Faces
1	-	91.79	96.26	-	-
6	99.63	-	-	-	95.12
8	97.45	-	-	-	-
9	97.35	-	-	-	91.4
10	88.70	-	-	-	-
12	99.0	-	-	-	-
14	99.50	-	-	-	-
15	96.0	-	-	-	-
16	97.73	-	-	-	90.60
17	95	-	-	-	-
18	97.5	-	-	-	92.5
21	99.15	-	-	-	-

Ref	Architecture	Data Augmentation	Dataset	Improvements	Shortcomings
1	5 CL, 3 FCL	Sub-window, flipping	Annotated Facial Landmarks in the Wild (AFLW)	Fine-tuned pre-trained model is used.	Better results with R-CNN even without using annotations of face landmarks.
5	13 CL, 3 FCL	-	WIDER FACE, FDDB dataset		
6	12 CL, 4 PL, 3 FCL	-	LFW, Youtube Face DB	3 images used for loss function as distance between similar and dissimilar image from.	Inception models reduce the size dramatically. All the models perform well, however the inception based model NN3 performs significantly well.
8	4 CL, 1 PL, 1 FCL, 1 Softmax	Horizontally flipped, image patches (regions)	LFW, CelebFace, CeleFace+	Trained 60 convNets and extracts two 160 dim feature vectors used for verification	Joint Bayesian based on our DeepID features achieves 97.45% test accuracy on LFW, which is competitive with the human-level performance of 97.53%.
9	8 FC, 3 FCL	Horizontally and vertically flipped	Google search IMDB image 2.5k male and 2.5k female images. 2000 images per entity collected	Adopts CNN's Alex architecture based ranking model to minimize human intensive annotation work.	-
10	CNN-S [3 CL, 1 SL] CNN-M [4CL, 1 SL] CNN-L [5CL, 1 SL]	Cropping [58x58 eye region]	LFW dataset	Fusion of (Soft-max + FC + JB) outperforms in accuracy measure.	-
11	Faster RCNN	-	FDDB dataset	false positive images are then re-trained to network for better performance	Obtain results for FRCNN R50, trained without hard negative mining.
12	NN1 [10 CL, 4 PL, MPL, 2 FCL], NN2 [12 CL as NN1], CNN-H2	Cropping eyes, nose centered.	LFW dataset	Patches extracted for the recognition of face image and dimensionality reduced using stack auto encoder.	-
14	10 CL, 1 PL, 1 SL	-	Megvii Face Classification (MFC), LFW dataset	Trained simple CNN on 4 face regions of the same face and combined their features	High performance on LFW does not perform well on the real environment data CHID with only 66% accuracy.
15	13 CL, 3 FC	-	PASCAL VOC 2007, PASCAL VOC 2012		Learned RPN improves region proposal quality and accuracy.
16	DeepID3 net1 (VGG16), net2 (Inception)	horizontally flipped	LFW dataset	Supervisory verification signal to intermediate and final feature extraction layers.	-
17	5 CL, 1 FCL	Yes	CASIA-WebFace. [LFW (Labelled faces in the Wild) and YTF (YouTube face database)]	Image size retained by convolution and pooling layer.	Used simple deep neural net without novel methods
18	4 CL, 1 LL, 1 FCL	25 face regions are selected	LFW and WDRrf Dataset	Supervisory signal in between 3rd and 4th layer.	-
21	3 CL, 3 LL, 1 SL	Nope	LFW dataset	Align the face using 3D mod-	-

19	4 CL, 1 FCL, 1 SL.	400 patches cropped	LFW dataset	elling and affine transformations Vector of 160-D extracted for features. Joint Bayesian model used.	-

6. Conclusion

This paper has reviewed the latest studies to provide a good knowledge of successful growth of deep learning in the field of face recognition. We have seen that different deep learning techniques has performed outstandingly on benchmark datasets like LFW and YouTube Faces (YTF), but the results of "FaceNet: A Unified Embedding for Face Recognition and Clustering" [7] has outperformed other studies with its outstanding architecture. It has been seen that deep learning models perform better when they are trained with a large dataset. In future face recognition can further be improved using deep learning by tweaking the best studies and by using different data augmentation techniques that will generalize the face recognition model.

References

[1] S. Awang and N. M. A. N. Azmi, "Vehicle Counting System Based on Vehicle Type Classification Using Deep Learning Method," in IT Convergence and Security 2017: Springer, 2018, pp. 52-59. https://link.springer.com/chapter/10.1007/978-981-10-6451-7_7

[2] S. S. Farfade, M. J. Saberian, and L.-J. Li, "Multi-view face detection using deep convolutional neural networks," in Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, 2015, pp. 643-650: ACM. https://doi.org/10.1145/2671188.2749408

[3] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," arXiv preprint arXiv:1312.6229, 2013.

[4] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," IEEE transactions on pattern analysis and machine intelligence, vol. 38, no. 1, pp. 142-158, 2016. https://doi.org/10.1109/TPAMI.2015.2437384

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in neural information processing systems, 2012, pp. 1097-1105.

[6] X. Sun, P. Wu, and S. C. Hoi, "Face detection using deep learning: An improved faster rcnn approach," arXiv preprint arXiv:1701.08289, 2017. https://doi.org/10.1016/j.neucom.2018.03.030.

[7] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 815-823. https://doi.org/10.1109/CVPR.2015.7298682

[8] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in European conference on computer vision, 2014, pp. 818-833: Springer. https://doi.org/10.1007/978-3-319-10590-1_53

[9] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1891-1898. https://doi.org/10.1109/CVPR.2014.244

[10] M. A. Talab, S. N. H. S. Abdullah, and M. H. A. Razalan, "Edge direction matrixes-based local binary patterns descriptor for invariant pattern recognition," in 2013 International Conference on Soft Computing and Pattern Recognition (SoCPaR), 2013, pp. 13-18: IEEE. https://doi.org/10.1109/SOCPAR.2013.7054123

[11] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep Face Recognition," in BMVC, 2015, vol. 1, no. 3, p. 6. https://doi.org/10.5244/C.29.41

[12] G. Hu et al., "When face recognition meets with deep learning: an evaluation of convolutional neural networks for face recognition," in Proceedings of the IEEE international conference on computer vision workshops, 2015, pp. 142-150. https://doi.org/10.1109/ICCVW.2015.58

[13] S. Wan, Z. Chen, T. Zhang, B. Zhang, and K.-k. Wong, "Bootstrapping face detection with hard negative examples," arXiv preprint arXiv:1608.02236, 2016.

[14] C. Ding and D. Tao, "Robust face recognition via multimodal deep face representation," IEEE Transactions on Multimedia, vol. 17, no. 11, pp. 2049-2058, 2015. https://doi.org/10.1109/TMM.2015.2477042

[15] D. Yi, Z. Lei, and S. Z. Li, "Towards pose robust face recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3539-3545. https://doi.org/10.1109/CVPR.2013.454

[16] E. Zhou, Z. Cao, and Q. Yin, "Naive-deep face recognition: Touching the limit of LFW benchmark or not?," arXiv preprint arXiv:1501.04690, 2015.

[17] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," IEEE transactions on pattern analysis and machine intelligence, vol. 39, no. 6, pp. 1137-1149, 2017. https://doi.org/10.1109/TPAMI.2016.2577031

[18] Z. Wu et al., "3d shapenets: A deep representation for volumetric shapes," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1912-1920.

[19] S. Awang and N. M. A. N. Azmi, "Automated Toll Collection System based on Vehicle Type Classification using Sparse-Filtered Convolutional Neural Networks with Layer-Skipping Strategy (SF-CNNLS)," in Journal of Physics: Conference Series, 2018, vol. 1061, no. 1, p. 012009: IOP Publishing. https://iopscience.iop.org/article/10.1088/1742-6596/1061/1/012009/meta

[20] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," arXiv preprint arXiv:1411.7923, 2014.

[21] W.-S. T. WST, "Deeply learned face representations are sparse, selective, and robust," perception, vol. 31, pp. 411-438, 2008.

[22] M. M. H. Mohammed Jabbar Mohammed, "Hybrid Tow Feature Extraction Descriptor forShape Pattern Recognition," Australian Journal of Basic and Applied Sciences, vol. 12, no. 7, pp. 32-39, 2018 July.

[23] Z. Chai, Z. Sun, H. Méndez-Vázquez, R. He, and T. Tan, "Gabor ordinal measures for face recognition," IEEE transactions on information forensics and security, vol. 9, no. 1, pp. 14-26, 2014. https://doi.org/10.1109/TIFS.2013.2290064

[24] M. A. Talab, H. Tao, and A. A. M. Al-Saffar, "Review on Deep Learning-Based Face Analysis," Advanced Science Letters, vol. 24, no. 10, pp. 7630-7635, 2018. https://doi.org/10.1166/asl.2018.12991