International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2014): 5.611

Face Recognition Revisited on Pose, Alignment, Color, Illumination and Expression-PyTen

Mugdha Tripathi

Computer Science, BIT Noida, India

Abstract: ***Growing interest in intelligent human computer interactions has motivated a recent surge in research on problems such as pose estimation, illumination variation, color differences, alignment distinction and expression variation. Human faces are highly non-rigid objects with high degree of variability in pose, color, expression, alignment angles and illumination conditions and most face recognition algorithms (not all), are designed to work best with well aligned, well illuminated, and frontal pose face images. An optimal face representation should be discriminative, robust, compact and easy to implement. The conventional pipeline of face representation consists of image pre-processing, extraction, alignment, representation and classification. Our approach is based on feature sharing structure of deep network called Pyramid CNN (Pyramid Convolutional Neural Network) which has known to adopt a greedy filter and down sampling approach for a fast and computation efficient training procedure. CNN learns representation of the face utilized by recognition algorithms in later stages. The color values of face images are normalized to RGB color space to reduce the lightning effect in normalization process. We use Field proposed Log Gabor filters for feature extraction which allows more information to be captured in high frequency domains with desirable high-pass characteristics. Using feature sharing Pyramid CNN we are able to achieve competitive accuracy on LFW database***

Keywords: Face recognition, Pyramid CNN, Deep network, Pose, Alignment, Color, Illumination, Expression

1. Introduction

Basic goal of human computer interaction system is to improve interactions between users and computers by enhancing user friendliness features in machines. Face recognition in unconstrained images is at the forefront of algorithmic perception revolution. The social and cultural implications of face recognition technologies are far reaching, yet the current performance gap in this domain between machines and human visual system serves as a buffer which provides further improvement scope. The very first step in most face recognition systems is to represent facial images as feature vectors after which various learning algorithms are applied to perform task of classification , verification etc.

In this paper, we present a face recognition approach that implements a unified and fast learning framework of deep convolutional neural network called Pyramid CNN which uses supervised learning signals in the form of face pairs. With specially designed operation of filter and down sample, the network can be trained fast and computation efficient and also it achieves high recognition accuracy with compact features. In addition, Pyramid CNN can naturally incorporate feature sharing across multi-scale face representations thus increasing the discriminative ability of resulting representation appreciably. The proposed approach differs from several recognition contributions in the field in that it uses Deep Learning framework in lieu of well-engineered features. Deep learning is a part of machine learning based on set of algorithms that attempt to model high level abstractions in data by using multiple processing layers with complex structures or otherwise composed of multiple non-linear transformations. A variant of Deep Learning architecture is Deep Neural network which is an artificial neural network with multiple hidden layers of units between input and output layers. Deep learning is very efficient for dealing with large training sets, specifically with faces; the success of the learned net in capturing facial appearance in robust manner is highly dependent on a very rapid 3D alignment step.

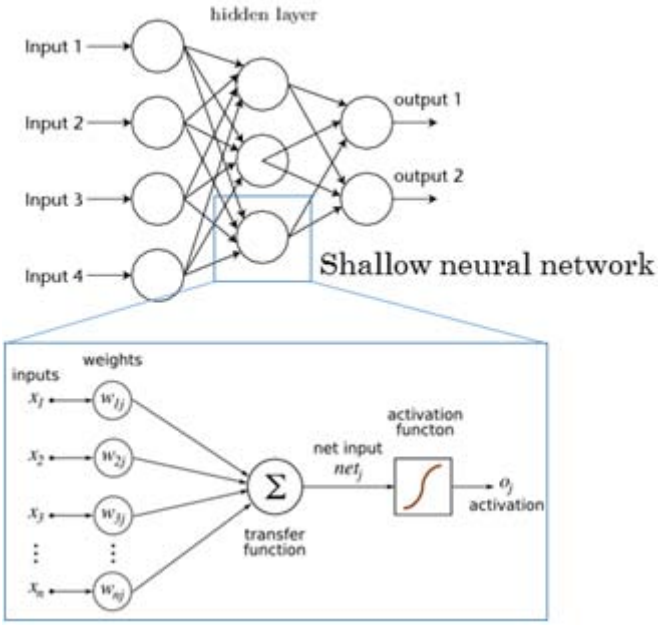


Figure 1: Deep Learning Achitecture

In a traditional pipeline, the image is preprocessed, encoded and transformed to representations with various levels of semantics. Existing methods focus on improving building blocks in the pipeline (shown in Fig. 2) to close the semantic gap between image pixels and identity.

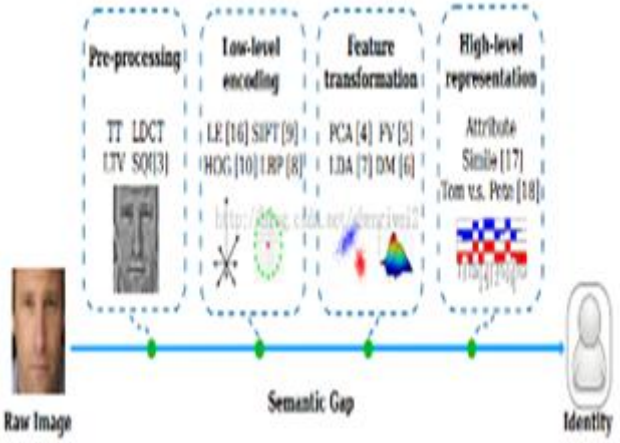


Figure 2: Traditional face representation pipeline

Pyramid CNN unifies the traditional pipeline with deep learning techniques and neural network directly works on image pixels. Inside the network, signal undergoes multiple layers of non-linear transformations which resemble manually designed multi-stage pipeline.

Use of facial expression for measuring people's emotion has dominated psychology since late 1960's and research study by Mehrabian has indicated that 7% of the communication information is transferred by linguistics, 38% by paralanguage and 55% by facial expressions in human face-to-face communication. This therefore is a proof that facial expressions provide large amount of information in communication. Also recent research proves that color information improves face recognition and image retrieval performance. Hence in our face recognition approach we normalize the color values of face images with respect to RGB values of the images to reduce lightning effect.

We also use weighted chi square similarity on Deep-face feature vector of Pyramid CNN and Siamese network to compare two input images and detect whether they belong to the same identity.

Multiple hand crafted feature extractors have been used in face recognition and a large number of pictures have been crawled by search engines and social networks upload and all these images include number of constrained objects like scenes, faces etc. LBP, Gabor Filter, SIFT and other descriptors have been proposed on various heuristics and many of these aim to be an extension or improvement or changed approach of the existing work. Large volume of image data on internet has enabled the use of increasingly powerful statistical models which have efficiently improved robustness of vision systems to several important variations such as clutter, occlusion, illumination etc.

2. Proposed Method

The new architecture of Deep Learning pushes further the limit of what is achievable by Deep neural networks by incorporating 3D alignment, customizing the architecture for aligned inputs, scaling the network by almost two order of magnitudes and demonstrating a simple knowledge transfer method once the network has been trained on large datasets.

The representation of face image can be considered as a function map from image pixel to numeric vector:

$$f: \mathbb{R}^{h*w} \rightarrow \mathbb{R}^{m} \quad (1)$$

There are some natural criterions for good face representation:

1) **Preserving identity:** Influence of irrelevant factors should be minimized for which distance in the mapped space should closely reflect the semantic distance of face image identity
2) **Compact and abstract representation:** Short length of representation (m) is appreciated by recognition models, hence to keep discriminative power in low dimensional space, the representation should encode abstract and high level information of face identity
3) **Uniform and automatic procedure:** Design of representation is automated instead of using hand-wired and hand crafted methods wherever possible, for which the ideal method is to close the semantic gap within a single uniform model. Thus one way to obtain desired representation is to learn it from data which involves parameterizing function family and also we use an object function L to choose from representation extractor

$$\Theta_0 = \text{argmin } L(f_\Theta, I_{data})$$

The function family f should contain enough complexity to express complex and high level computation required and also to guarantee ID-preserving property; identity information should be used in objective function L. This provides a supervised representation learning method which highly influences factors like illumination, alignment, pose and expression. This method extracts and exploits the property of facial images to improve the training of deep neural networks.

2.1 Face Alignment

Aligning faces in unconstrained scenario is still considered a difficult problem that has to account for many factors like pose, non-rigid expressions that are hard to decouple from identity bearing facial morphology. We employ the analytical method of 3D modeling of face based on fiducial points, that is used to warp a detected facial crop to a 3D frontal mode (*frontalization*). A simple point detector is applied in several iterations to refine the image output and at each iteration, fiducial points are extracted by Support Vector Regressor which is trained to predict point configurations from an image descriptor.

2D Alignment

6 fiducial points are detected inside the detection crop which are centered at the center of eyes, tip of nose and mouth (Fig. 3).

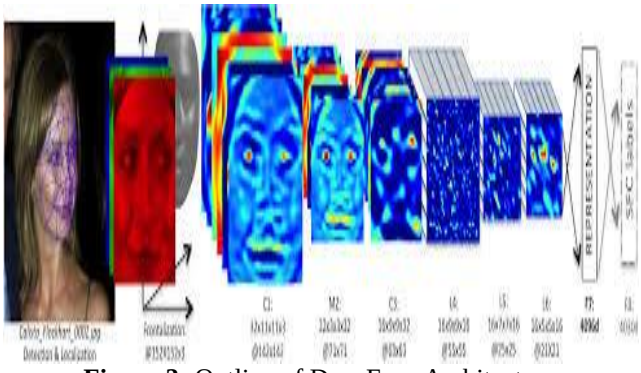


Figure 3: Outline of DeepFace Architecture

Selected fiducial points are used to approximately scale, rotate and translate image into six anchor locations by fitting:

$T^i_{2d} = (s_i , R_i, t_i)$; T is induced similarity matrix (2)

xjanchor = si[Ri|ti] * xjsource for points j=1 to 6 (3)

Iteration is done on the warped image until no substantial change is observed, eventually composing the final 2D similarity transformation as:

$$T_{2d} = T^1_{2d} * \ldots\ldots * T^k_{2d}$$

This finally obtains 2D aligned crops in Fig. 4 (b):

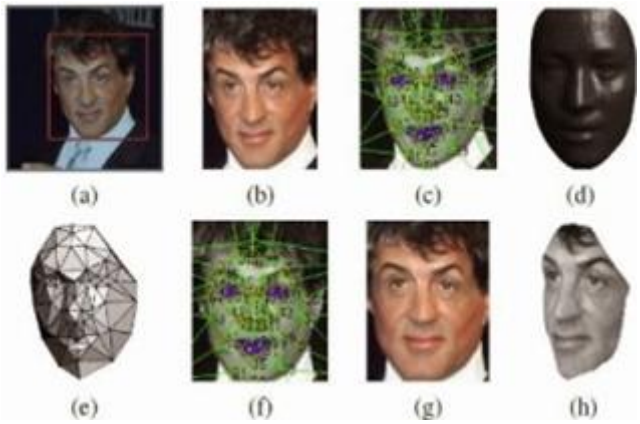


Figure 4: Alignment Pipeline

3D Alignment

Faces undergoing out-of-plane rotations are aligned in 3D version using a generic 3D shape model which generates 3D aligned version of the crop as shown in Fig. 4 (g). The 67 fiducial points x_{2d} are localized in the 2D-aligned crop using second SVR. As a 3D generic shape model, average of the 3D scans from the USF Human-ID database is taken which were post-processed to be represented as aligned vertices $v_i = (x_i; y_i; z_i)^n_i=1$. 67 anchor points are then manually placed on the 3D shape and full correspondence between the 67 detected fiducial points and their 3D references is achieved. An affine 3D-to-2D camera P is then fitted using the generalized least squares solution to the linear system $x_{2d} = X_{3d}$ ~P with a known covariance matrix $\sum$, that is, ~P that minimizes the following loss: loss(~P) = $r^T\sum^{-1}r$ where r = (x_{2d} - X_{3d} ~P) is the residual vector and X_{3d} is a (67 * 2) x 8 matrix composed by stacking the (2*8) matrices [x^T_{3d} (i); 1;~0;~0; x^T_{3d} (i), 1], with ~0 denoting a row vector of four zeros, for each reference fiducial point x_{3d}(i). The affine camera P of size 2*4 is represented by the vector of 8 unknowns ~P. The loss can be minimized using the Cholesky decomposition of $\sum$ that transforms the problem into ordinary least squares.

2.2 Face Normalization

The aim of this module is to obtain face images, which have normalized intensity, are uniform in size and shape. The face area of an image is detected using the Viola–Jones method based on the Haar-like features and the AdaBoost learning algorithm. The Viola and Jones method is an object detection algorithm providing competitive object detection rates in real-time. It was primarily designed for face detection. The features used by Viola and Jones are derived from pixels selected from rectangular areas imposed over the picture, and exhibit high sensitivity to the vertical and the horizontal lines. After face detection stage, the face images are scaled to the same size (e.g., 64×64 pixels). The color values of face images are then normalized with respect to RGB values of the image. The purpose of color normalization is to reduce the lighting effect because the normalization process is actually a brightness elimination process.

Given an input image N_1 * N_2 pixels, represented in RGB color space:

$x = \{X^{n3} [n_1, n_2] \mid 1<=n_1<=N_1, 1<=n_2<=N_2, 1<=n_3<=N_3\}$ (4)

The normalized values $X^{n3}_{norm} [n_1, n_2]$ are given as:

$$X^{n3}_{norm} [n_1, n_2] = x^n_3[n_1, n_2]/ \sum^3_{n3=1} x^n_3[n_1, n_2] \quad (5)$$

When X^{n3}_{norm} for n_3=1, 2, 3 correspond to red, green and blue components of the image x respectively, it is found that summation of all these three components is equal to 1.

2.3 Weighted χ^2 distance test

Fitting a model to a relatively small space or dataset reduces its generalization to other datasets. Hence learning an unsupervised metric that generalizes well to almost every dataset is preferable than limiting the scope of dataset. The unsupervised similarity implemented is the inner product of two normalized feature vectors. The DeepFace feature vector implemented has several features: (1) Is very sparse, (2) Contains non negative values, (3) Values vary between [0,1]. Using Weighted χ^2 similarity we have:

$$\chi^2(f1, f2) = \sum_i w_i(f_1(i)-f_2(i))^2/ (f_1(i)+f_2(i)) \quad (6)$$

here f_1 and f_2 are DeepFace representations.

After the implementation of chi square test, we use supervised learning ahead on Pyramid CNN as fitting of model has been implemented using unsupervised learning already.

2.4 Pyramid CNN and its architecture

Neural networks are applied to image patches and their last layer activation is taken as representation. The architecture of Pyramid CNN takes advantage of multi scale structure of face and training is highly accelerated in Pyramid CNN as it is a group of CNN's divided into multiple levels. Network here is composed of several shared layers and an unshared part which has the same structure at all levels. First layer is shared across all levels and second layer is shared by networks from the second level and this sharing scheme is repeated throughout till the last layer of Pyramid CNN. As down-sampling operation is followed in shared layers, therefore, input size of network at higher level is larger than

the lower levels. In Pyramid CNN it is allowed that more than one network exists in the same level and they work on different region while sharing the first layer parameters. Greedy training of the Pyramid CNN is done to train networks on part of the face and image thereby producing agreeable output.

Network in the first level is first trained on the part of face and after training it, its first layer is fixed and that first layer is used to down sample and filter the training images. Second level networks are then trained on the processed image and in this way the input size of the network that is actually trained does not increases as the level increases. This Greedy Layer wise training has been conducted till the final network with extra depth is obtained.

On the point of utilizing Pyramid CNN's multi-scale feature extraction architecture, image patches of different sizes are fed to networks at corresponding scale levels. Pyramid takes advantage of multi scale structure by using deeper networks for larger input region and the increase in depth allows higher level networks to undertake more complex and abstract computation on larger image patches.

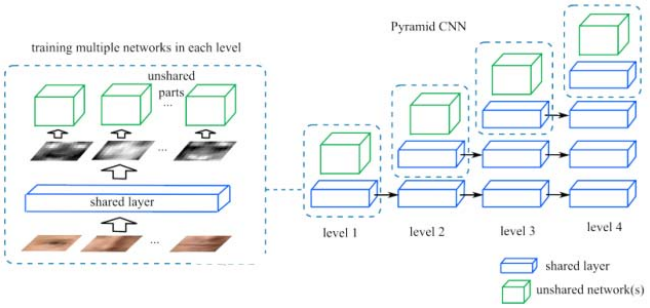


Figure 5: Pyramid CNN, to train which Siamese network is used

Two images are fed to the same CNN and the outputs are compared by the output neuron which predicts whether two faces are of the same identity. Pyramid CNN consists of several level of networks which have different depths and input sizes and they share some of the layers. Pyramid is trained in greedy manner of which first network is trained on part of the face, and its first layer is fixed.

2.5 ID preserving representation learning using Siamese Network

To train an individual network, Siamese network is used which compares two input images and detects whether they belong to same identity. Siamese neural networks consist of twin networks which accept distinct inputs but are joined by an energy function at the top which ensures consistency and a symmetric network. In order to prevent over fitting on the face verification task only two topmost layers are enabled for training. The known induced distance in Siamese Network is:

$$d(f_1, f_2) = \sum_i \alpha_i |f_1[i]-f_2[i]| \quad (7)$$

here α_i is the trainable parameter

For ID preserving property, same CNN is applied to the two images to produce the corresponding representations and one output neuron uses a distance function to compare the representation and predict whether the face pair belongs to the same person.

Loss function is:

$$L = \sum_{I_1,I_2} log\ (1 + exp(\delta(I_1, I_2)D(I_1, I_2))) \quad (8)$$

If prediction is perfectly accurate, loss function is Zero otherwise we receive penalty in Loss function.

$$\text{Also, } D(I_1, I_2) = \alpha\ .\ d\ (f_\theta(I_1), f_\theta(I_2)) - \beta \quad (9)$$

$\delta(I_1, I_2)$ indicate whether two images belong to same person, f_θ indicates computation done by neural network and d is a function to measure distance between two vectors. Θ is for weights in the network and α and β are trainable parameters. Factors corresponding to intra person variation will be suppressed by the network.

2.6 Advanced tensor application:

Color images are similar to 3D data array with three components: Horizontal, Vertical and Color. For these 3D data array we generate a tensor which is a higher order generalization of vector and matrix and then apply filtering operation on this tensor for accurate image recognition in varying color conditions. We can say that a color image T is a tensor of order 3 which can be represented as $T \in R\Pi^3{}_1 N_n$ where N_1 is height of image, N_2 is width and N_3 is number of color channels. We experiment on variations of N_1 and N_2 from 32 to 128 and N_3 is kept as 3. Tensors can be unfolded to n-mode mathematical objects for feature extraction and classification. 3D color image is unfolded to obtain 2D tensors based on multi linear analysis criteria. We horizontally unfold the image from $X^{N1xN2xN3}$ to $X^{N1x\ (N2xN3)}$ which is efficiently employed to detect images in varying color conditions.

3. Experiments and Results

LFW dataset contains more than 13000 images acquired from the web and our experiment protocol is to evaluate the accuracy of image in different conditions of illumination, expression, pose, alignment and color and we compared our results with other existing methods

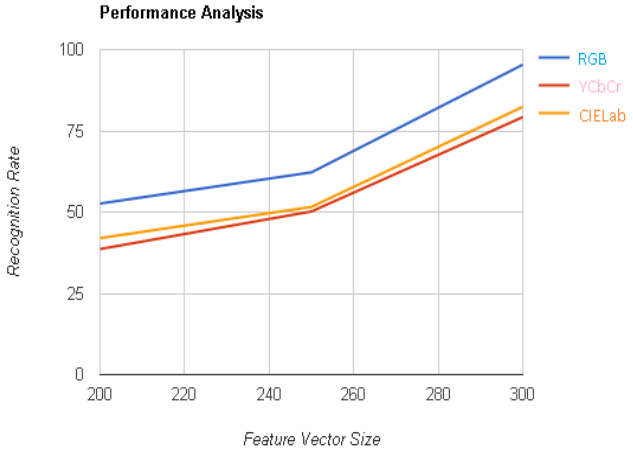
Table 1: Comparison of average recognition rates for different color spaces

Size	Gray	RGB	YCbCr	CIELab
144x144	49.75 57.68 92.02	52.63 62.23 95.35	38.63 50.20 79.21	42 51.60 82.38
256x256	36.05 52 91	39.3 49.33 89	29 36.24 65.32	30.21 38.68 61
1024x1024	28 31 80.67	20.26 30.65 70	18.61 20.34 32.27	19 20.61 39.51

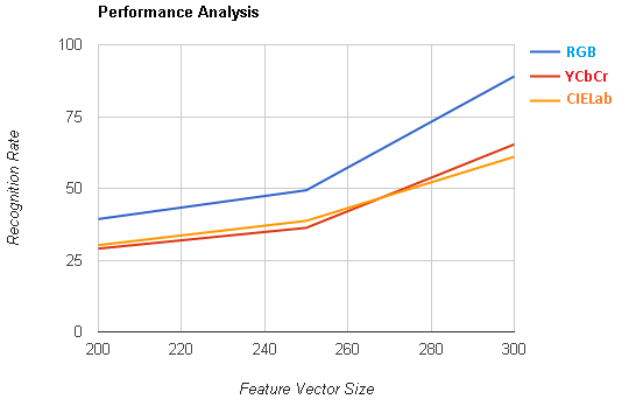
From Table 1 it can be seen that average recognition rate is improved as well when the resolution is increased from 144 x 144 to 1024 x 1024. Since color images are more sensitive to illumination than gray scale images, the testing set under slight illumination variation is used to test system robustness which has proved to be highly efficient.

The graphs show the improvement in recognition rate which are plotted against size of feature vector. These are plotted

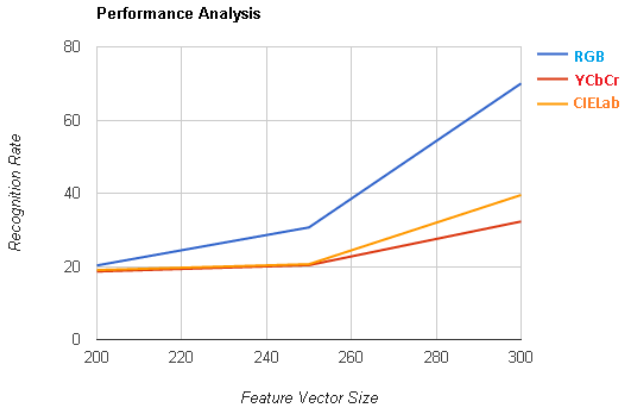
for different col or spaces and performance analysis is inspected.



(a) Size of 144 x 144 image



(b) Size of 256 x 256 image



(c) Size of 1024 x 1024 image

Figure 6: Comparative evaluation of performance in different color spaces: RGB, YCbCr, CIELab

Compact Single Feature: 5-level Pyramid CNN is run on the whole face image and output of the last level network is taken as final representation. Fig. 7 shows accuracy of this feature at different number of dimensions. Compared to LBP and PCA baselines, performance of our proposed system deteriorates slowly as the dimension is reduced. Except for the last level network all other levels are constructed for training. The representation is tested on LFW benchmark. This representation is the output of the highest level CNN in the pyramid. Performance of LBP baseline highly deteriorates in low dimensional setting whereas our compact learning representation deteriorates slowly in low dimensional settings as well.

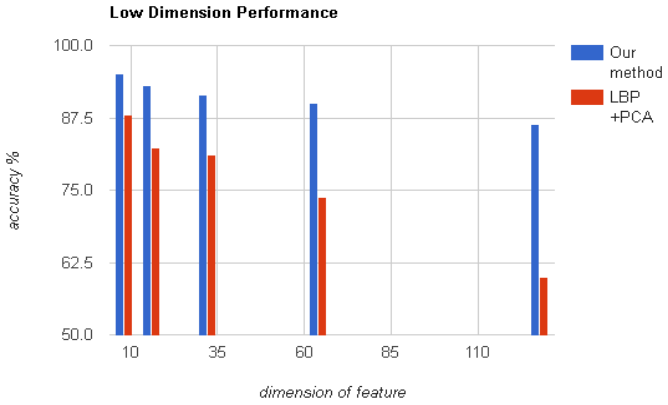
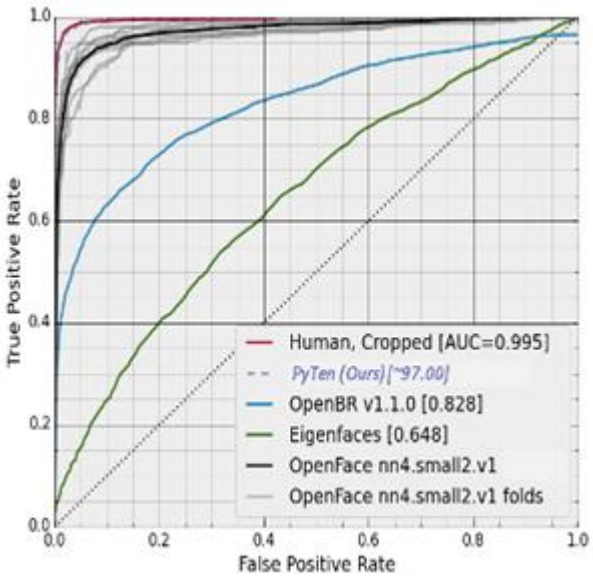
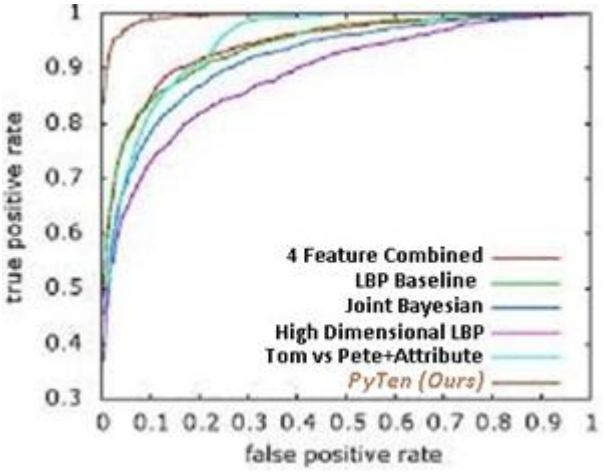


Figure 7: Performance of representation when dimension is low

High Dimensional Representation: Fig. 8 represents performance on LFW benchmark when there is no constraint on the length of representation. As experimented, our system has achieved efficient results when compared with LBP baseline, Joint Bayesian, EigenFaces method and others. Also the results were comparable with DeepFace method which has achieved state of the art performance.



(a)



(b)

Figure 8: Both (a) and (b) represent performance of PyTen (our) method on LFW benchmark

LFW dataset consists of 13323 web photos of 5749 celebrities which are divided into 6000 face pairs in 10 splits. Performance is measured by mean recognition accuracy using, a) restricted protocol in which only some and not same labels are available in training, b) unrestricted protocol in which no training is performed on LFW images.

Results on LFW dataset have been competitive with that of DeepFace representation. Our system has achieved an accuracy of 97.2% thus significantly improving the face recognition efficiency. This system has achieved the mentioned accuracy across pose, alignment, color, illumination and expression.

4. Conclusion

For now the method has been applied only to human faces but further improvements will include other areas also. For object classification, greater input resolution and higher number of scale levels are typically used which require higher performance improvement of implemented Pyramid CNN and also object identification does not has a relatively fixed configuration like that of a face, hence it is more challenging. The underlying face descriptor has been invariant to pose, illumination, expression, alignment and color and it may be applied to varying population with fewer modifications.

References

[1] D. Yi, Z. Lei, and S. Z. Li. *Towards pose robust face recognition. In CVPR, 2013*
[2] Y. Bengio, *Learning deep architectures for AI. Foundations and Trends in Machine Learning, 2009*
[3] Yi Sun, Xiaogang Wang, Xiaoou Tang, *Deep Learning Face Representation by Joint Identification-Verification, June 2014*
[4] J. van de Weijer and Th. Gevers, *Tensor Based Feature Detection for Color Images, 2014*
[5] Ahmadreza Baghaie and Zeyun Yu, *Structure Tensor Based Image Interpolation Method, 2014*
[6] H. Li, G. Hua, Z. Lin, J. Brandt, and J. Yang, *Probabilistic elastic matching for pose variant face verification. In CVPR, 2013*
[7] T. Hassner, *Viewing real-world faces in 3D. In International Conference on Computer Vision (ICCV), Dec. 2013*
[8] Y. Sun, X. Wang, and X. Tang, *Deep convolutional network cascade for facial point detection. In CVPR, 2013*
[9] Voruganti Ravi Kumar, Sk Subhan, Devireddy Venkatarami Reddy, *A Novel Approach for Facial Expression Recognition Rate (FER) By Using Tensor Perceptual Color Framework, 2014*
[10] T. Berg and P. N. Belhumeur. *Tom-vs-pete classifiers and identity preserving alignment for face verification. In BMVC, 2012*
[11] L. Wolf, T. Hassner, and I. Maoz, *Face recognition in unconstrained videos with matched background similarity. In CVPR, 2011*
[12] Dean, J., Corrado, G., Monga, R., Chen, K., Devin, M., Le, Q.V., Mao, M.Z., Ranzato, M., Senior, A.W., Tucker, P.A., et al.: *Large scale distributed deep networks. In: NIPS 2012*
[13] Megvii: Face++. *http://www.faceplusplus.com Accessed 2014-3-7*
[14] Y. Sun, X. Wang, and X. Tang, *Deep convolutional network cascade for facial point detection. In CVPR, 2013*
[15] Sun, Y., Wang, X., Tang, X., et al.: *Hybrid deep learning for face verification, ICCV 2013*
[16] G. E. Dahl, T. N. Sainath, and G. E. Hinton. *Improving deep neural networks for LVCSR using rectified linear units and dropout. In ICASSP, 2013*
[17] Gregory Koch, *Siamese Neural Networks for One-Shot Image Recognition, 2015*
[18] Nitish Srivastava, *Improving neural networks with dropout, 2013*
[19] Karen Simonyan and Andrew Zisserman. *Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556, 2014*
[20] O. Barkan, J. Weill, L. Wolf, and H. Aronowitz, *Fast high dimensional vector multiplication face recognition. In ICCV, 2013*

Author Profile

Mugdha Tripathi received the B.Tech. degree from Banasthali University and currently pursuing M.Tech and working as Lead Engineer in Software Organization. The total work experience is of 5 years which includes proficient work in 3 different IT organizations as a Software Developer. Currently working in IBM as Senior System Engineer and M.Tech student of BIT Noida, Ranchi.