One to Many Face Recognition with Bilinear CNNS

Sonal Mishra, Pratyush Tripathi

***Abstract*— In programmed face recognition we longing to either recognize or check at least one people in still or video pictures of a scene by methods for a put away database of faces. One of the imperative components of face recognition is its non-meddlesome and non-contact property that recognizes it from different biometrics like iris or unique finger impression recognition that require subjects' participation. The current development in convolutional neural network (CNN) investigate has created an assortment of new structures for profound learning. One interesting new engineering is the bilinear CNN (B-CNN), which has indicated sensational execution picks up on certain fine-grained recognition issues.**

We apply this new CNN to the testing new face recognition benchmark, the IARPA Janus Benchmark An (IJB-A). It highlights faces from an expansive number of personalities in testing genuine conditions. Since the face pictures were not distinguished naturally utilizing an electronic face recognition framework, it doesn't have the inclination characteristic in such a database. We exhibit the execution of the B-CNN shows starting from an AlexNet-style arrange pre-prepared on Image Net. We then show comes about for calibrating utilizing a direct measured and open outside database. We likewise give comes about extra adjusting on the constrained preparing information given by the convention. In each case, the fine-tuned bilinear model shows substantial improvements over the standard CNN. Finally, we demonstrate how a standard CNN pre-trained on a large face database, the recently released VGG-Face model can be converted into a B-CNN without any additional feature training.

***Index Terms*— Face Recognition, Bio-metric Identification, Convolutional neural network (CNN)**

I. Introduction

Face recognition is one of the most relevant applications of image analysis. It's a true challenge to build an automated system which equals human ability to recognize faces. Although humans are quite good identifying known faces, we are not very skilled when we must deal with a large amount of unknown faces. The computers, with an almost limitless memory and computational speed, should overcome human's limitations.

Affective state plays a fundamental role in human interactions, influencing cognition, perception and even rational decision making. This fact has inspired the research field of "affective computing" which aims at enabling computers to recognize, interpret and simulate affects [1]. Such systems can contribute to human computer communication and to applications such as learning environment, entertainment, customer service, computer games, security/surveillance, and educational software as well as in safety critical application such as driver monitoring [2]. To make human-computer interaction (HCI) more natural and friendly, it would be beneficial to give computers the ability to recognize affects the same way a human does. Since speech and vision are the primary senses for human expression and perception, significant research effort has been focused on developing intelligent systems with audio and video interfaces.

We introduce comes about for IJB-An utilizing the bilinear convolutional neural system (B-CNN) of Lin et al. [3] in the wake of adjusting it to our requirements and rolling out some minor specialized improvements. Keeping in mind the end goal to make utilization of pictures from different points of view, we additionally research a strategy proposed by Su et al. [4] that pools pictures at the element level, instead of pooling characterization scores. We take after the open-set 1:N convention and report both total match trademark (CMC) and choice blunder exchange off (DET) bends, taking after the best practice portrayed in the 2014 Face Recognition Vendor Test [5] and recommended in the IJB-A paper [6].

We report results on a baseline network, a network fine tuned with a publicly available face database (FaceScrub) and also a network further fine-tuned using the IJB-A trainset. Since IJB-A contains multiple images per probe, we explore two pooling strategies as well. We show that for the fine-tuned networks, and for both pooling strategies, the B-CNN architecture always outperforms the alternative, often by a large margin.

At last, we exhibit how a pre-prepared CNN can be changed over into a B-CNN with no extra tweaking of the model. The "VGG-Face" CNN from Parkhi et al. [7] was prepared on a huge face informational collection which, at the season of distribution, was not openly accessible. Be that as it may, the effortlessness of the bilinear design permits the production of a B-CNN from the pre-prepared CNN engineering without the requirement for retraining the system.

II. Literature Review

Firstly discuss some of the representative works for facial expression recognition and then move our discussion on existing audio-visual affect recognition approaches to highlight the challenges lies in the integration of the two modalities. For an overview of audio only, visual only and audio-visual affect recognition, readers are encouraged to study a recent survey by Zeng et al. [7]. Because of the importance of face in emotion expression and perception, most of the vision-based affect recognition studies focus on facial expression analysis. A large amount of existing facial expression recognizers employ various pattern recognition approaches and are based on 2D spatiotemporal facial features: geometric features or appearance based features. Geometric based approaches track the facial geometry information over time and classify expressions based on the deformation offacial feature. Chang et al. [4] defined a set of

Sonal Mishra, Department of Electronics and Communication Engineering, M.Tech Scholar, Kanpur Institute of Technology, Kanpur, India

PratyushTripathi, Assistant Professor, Department of Electronics and Communication Engineering, , Kanpur Institute of Technology, Kanpur, India.

points as the facial contour feature, and an Active Shape Model (ASM) is learned in a low dimensional space. Lucey et al. [6] employed Active Appearance Model (AAM)-derived representation while Valtar, Patras, and Pantic [8] tracked 20 fiducial facial points on raw video using a particle filter.

On the other hand, appearance-based approaches emphasize on describing the appearance of facial features and their dynamics. Zhao and Pietikaninen [9] employed the dynamic Local Binary Pattern (LBP) which is able to extract information along the time axis. Bartlett et al. used a bank of Gabor wavelet filter to decompose the facial texture.

The system has been successfully tested on a set of origami objects. Recognition time is expected to grow only logarithmically with the number of objects stored. (Xiong, 2010).Dyar. C.R (1994) argued that there are two basic behaviors that allow reconstruction of a patch around any point in a reconstruction surface region. These behaviors rely only on information extracted directly from face images, and are simple enough to be executed in real time. Global surface reconstruction can be provably achieved by integrating these behaviors to iteratively "grow" the reconstructed regions integrating these behaviors to iteratively "grow" the reconstructed regions (Dyar. C.R.,1994) (PIter, alet, 2008).

One of the challenging tasks of the visual tracking systems is to deal with changes in the shape of the mouth caused due to speech. In order to deal with this situation, Datcu et al. [3] proposed a data fusion technique where they rely only on the visual data in the silent phase of the video sequence and the fused audiovisual data during non-silent segments. The visual modality during non-silent segments only focused on the upper half of the facial region to eliminate the effects caused by changes in the shape of the mouth. However, the result shows that full face based model performs superior than partial face. Hence an alternative strategy is requiring filtering out the influence of phonemes.

An important audio visual fusion scheme which aim at making use of the correlation between audio and visual data streams and relaxing the requirement of synchronization of these streams, is that of model-level fusion. Zeng et al. [10] presented a Multistream Fused HMM to build an optimal connection among multiple streams from audio and visual channels according to the maximum entropy and the maximum mutual information criterion. Author, however, considered tightly coupled HMMs. Song et al. [11] proposed an approach for multimodal emotion recognition which was specifically focused on temporal analysis of three sets of features: 'audio only features', 'visual only features' (upper half of facial region) and 'visual speech features' (lower half of facial region) using a triple HMM, i.e., one HMM for each of the information modes. This model was proposed to deal with state asynchrony of the audio-visual features while maintaining the original correlation of these features over time. On the other extreme is the model that allows complete asynchrony between the streams. This is, however, infeasible due to the exponential increase in the number of state combinations possible due to the asynchrony.

The new IARPA Janus Benchmark A (IJB-A) is designed to satisfy this need. The IJB-Ais presented in a CVPR paper that describes the database,gives detailed information about proper protocols for use,and establishes initial baseline results for the defined protocols[12].

III. Convolutional Neural Network

Convolutional neural networks (CNNs) are widely used in pattern- and image-recognition problems as they have a number of advantages compared to other techniques. This white paper covers the basics of CNNs including a description of the various layers used. Convolutional neural networks (CNNs) are composed of a hierarchy of units containing a convolution, pooling and non-linear layer. In recent years deep CNNs typically consisting of the order of 10 or so such units and trained on massive labeled datasets such as Image Net have yielded generic features that are applicable in a number of recognition tasks ranging from image classification [12], object detection [13], semantic segmentation [14] to texture recognition [15].

A CNN consists of one or more convolutional layers, often with a sub sampling layer, which are followed by one or more fully connected layers as in a standard neural network. The design of a CNN is motivated by the discovery of a visual mechanism, the visual cortex, in the brain. The visual cortex contains a lot of cells that are responsible for detecting light in small, overlapping sub-regions of the visual field, which are called receptive fields. These cells act as local filters over the input space, and the more complex cells have larger receptive fields. The convolution layer in a CNN performs the function that is performed by the cells in the visual cortex [3]. A typical CNN for recognizing traffic signs is shown in Figure 1. Each feature of a layer receives inputs from a set of features located in a small neighborhood in the previous layer called a local receptive field. With local receptive fields, features can extract elementary visual features, such as oriented edges, end-points, corners, etc., which are then combined by the higher layers. In the traditional model of pattern/image recognition, a hand-designed feature extractor gathers relevant information from the input and eliminates irrelevant variability. The extractor is followed by a trainable classifier, a standard neural network that classifies feature vectors into classes.

In a CNN, convolution layers play the role of feature extractor. But they are not hand designed. Convolution filter kernel weights are decided on as part of the training process. Convolutional layers are able to extract the local features because they restrict the receptive fields of the hidden layers to be local.

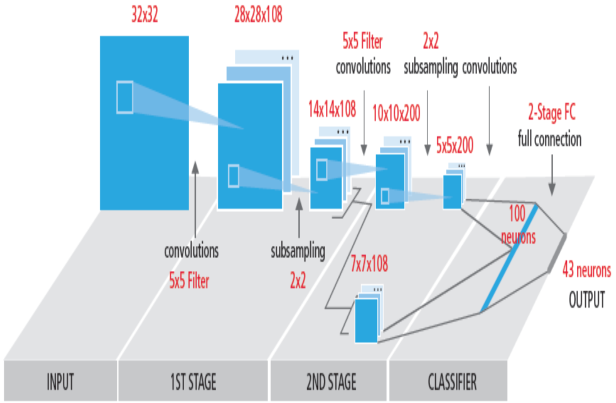


Figure 1: Typical block diagram of a CNN [14]

CNNs are used in variety of areas, including image and pattern recognition, speech recognition, natural language processing, and video analysis. There are a number of reasons that convolutional neural networks are becoming important. In traditional models for pattern recognition, feature extractors are hand designed. In CNNs, the weights of the convolutional layer being used for feature extraction as well as the fully connected layer being used for classification are determined during the training process. The improved network structures of CNNs lead to savings in memory requirements and computation complexity requirements and, at the same time, give better performance for applications where the input has local correlation (e.g., image and speech). CNN-based face recognition schema is given first is take a input face and next step is face detection This requires the algorithm to locate a single face with a known scale and orientation [12]. Face detection at frame rate is an impressive goal that has an apparent application to practical face tracking and real time face recognition. The input images acquired via still or video cameras might suffer from noise, bad illumination or unrealistic color. Therefore, noise removal might be a necessary block in some cases. Histogram equalization is the most common method used for image enhancement when images have illumination variations [4]. Even for images under controlled illumination, histogram equalization improves the recognition results by flattening the histogram of pixel intensities of the images. Then CNN training are allows and after completing all process at last Face Verification process are doing

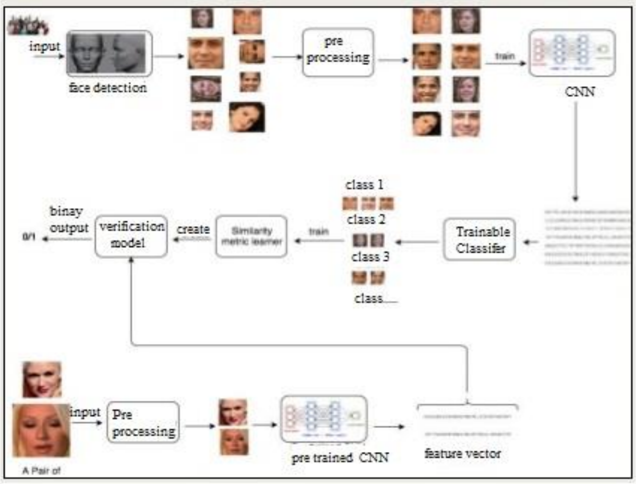


Figure 2: General face recognition pipelines

This method proposes a novel approach for recognizing the human faces. The recognition is done by comparing the characteristics of the new face to that of known faces. It has Face localization part, where mouth end point and eyeballs will be obtained. In feature Extraction, Distance between eyeballs and mouth end point will be calculated. The recognition is performed by Neural Network (NN) using Back Propagation Networks (BPN) and Radial Basis Function (RBF) networks. Back propagation can train multilayer feed-forward networks with differentiable transfer functions to perform function approximation, pattern association, and pattern classification. The BPN is designed with one input layer, one hidden layer and one output layer. The input layer consists of six neurons the inputs to this network are feature vectors derived from the feature extraction method in the previous section. The network is trained using the right mouth end point samples. The Back propagation training takes place in three stages: Feed forward of input training pattern, back propagation of the associated error and Weight adjustment. During feed forward, each input neuron receives an input value and broadcasts it to each hidden neuron, which in turn computes the activation and passes it on to each output unit, which again computes the activation to obtain the net output. During training, the net output is compared with the target value and the appropriate error is calculated. From this, the error factor has been calculated which is used to distribute the error back to the hidden layer. The weights are updated accordingly. In a similar manner, the error factor is calculated for a single unit. After the error factors are obtained, the weights are updated simultaneously. The output layer contains one neuron. The result obtained from the output layer is given as the input to the RBF. RBF uses the gaussian function for approximation. For approximating the output of BPN, it is connected with RBF. The Radial Basis Function neural network is found to be very attractive for the engineering problems. They have a very compact topology, universal approximations; their learning speed is very fast because of their locally tuned neurons. The RBF neural network has a feed forward architecture with an input layer, a hidden layer and an output layer. A RBF neural network is used as recognizer in face recognition system and the inputs to this network are the results obtained from the BPN. This neural network model combined with BPN and RBF networks is developed and the network is trained and tested.

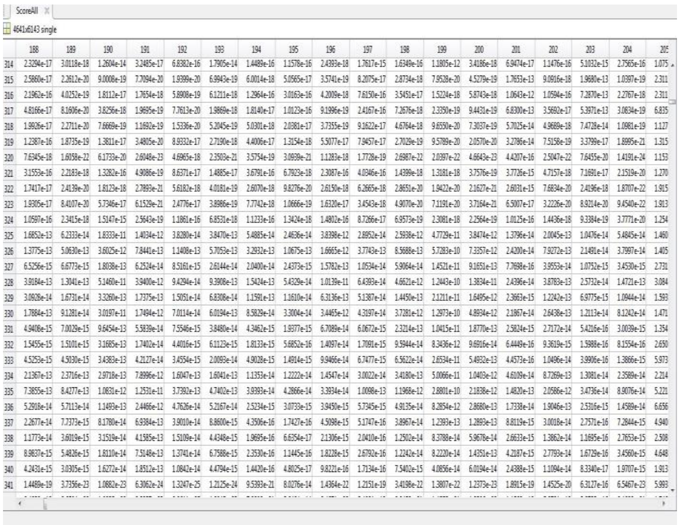
A key advantage is that the bilinear CNN model can be trained using only image labels without requiring ground-truth part-annotations. Since the resulting architecture is a directed acyclic graph (DAG), both the networks can be trained simultaneously by back-propagating the gradients of a task-specific loss function. This allows us to initialize generic networks on ImageNet and then fine-tune them on face images. Instead of having to train a CNN for face recognition from scratch, which would require both a search for an optimal architecture and a massive annotated database, we can use pre-trained networks and adapt them to the task of face recognition.

When using the symmetric B-CNN (both the networks are identical), we can think of the bilinear layer being similar to the quadratic polynomial kernel often used with Support Vector Machines (SVMs). However, unlike a polynomial-kernel SVM, this bilinear feature is pooled over all locations in the image and can be trained end-to-end.

IV. Dataset and Protocols

The IJB-A face recognition protocol [12] provides three sets of data for each of its 10 splits. Models can be learned on the train set, which contains 333 persons with varying number of images, including video frames, per person. The gallery set consists of 112 persons. The probe set is comprised of imagery from 167 persons, 55 of whom are not present in the gallery (known as “distractors” or “impostors”). It follows the open-set protocol in its identification task.

The IJB-A benchmark comprises both a one-to-many recognition task (identification) and a verification task. We focus in this paper on the identification task. The details of the identification protocol and reporting of results can be found in the NIST report by Grother et al. [16]. To evaluate the performance of a system in correctly matching a probe template to its identity (from among the identities present in the gallery set), the Cumulative Match Characteristic (CMC) curve is used. This summarizes the accuracy on probe templates that have a match among the gallery identities at various ranks. The rank-1 and rank-5 values are individual points on this curve, which usually reports recall from ranks 1 to 100.In the open-set protocol, two particular scenarios may arise as follows: firstly, the "non-match" or "impostor" templates might be wrongly classified as a gallery identity, if the classifier score from the one-versus-rest SVM for that identity is above some threshold (false alarms). Secondly, a template that is genuinely from among the gallery identities may be wrongly rejected if all the SVM scores for it are below some threshold (misses).



ScoreAll

4641x6143 single

	188	189	190	191	192	193	194	195	196	197	198	199	200	201	202	203	204	205
314	2.3294e-17	3.0118e-18	1.2604e-14	3.2485e-17	6.8382e-16	1.7905e-14	1.4489e-16	1.1578e-16	2.4393e-18	1.7617e-15	1.6349e-16	1.1805e-12	3.4186e-18	6.9474e-17	1.1476e-16	5.1032e-15	2.7565e-16	1.075
315	2.5860e-17	2.2612e-20	9.0008e-19	7.7094e-20	1.9399e-20	6.9943e-19	6.0014e-18	5.0565e-17	3.5741e-19	8.2075e-17	2.8734e-18	7.9528e-20	4.5279e-19	1.7653e-13	9.0916e-18	1.9680e-13	1.0397e-19	2.311
316	2.1962e-16	4.0252e-19	1.8112e-17	1.7654e-18	5.8908e-19	6.1211e-18	1.2964e-16	3.0163e-16	4.2009e-18	7.6150e-16	3.5451e-17	1.5224e-18	5.8743e-18	1.0643e-12	1.0594e-16	7.2870e-13	2.2767e-18	2.311
317	4.8166e-17	8.1606e-20	3.8256e-18	1.9695e-19	7.7613e-20	1.9869e-18	1.8140e-17	1.0123e-16	9.1996e-19	2.4167e-16	7.2676e-18	2.3350e-19	9.4431e-19	6.8300e-13	3.5692e-17	5.3971e-13	3.0834e-19	6.835
318	1.9926e-17	2.2711e-20	7.6669e-19	1.1692e-19	1.5336e-20	5.2045e-19	5.0301e-18	2.0381e-17	3.7355e-19	9.1622e-17	4.6764e-18	9.6550e-20	7.3037e-19	5.7025e-14	4.9689e-18	7.4728e-14	1.0981e-19	1.127
319	1.2387e-16	1.8735e-19	1.3811e-17	3.4805e-20	8.9332e-17	2.7190e-18	4.4006e-17	1.3154e-18	5.5077e-17	7.9457e-17	2.7029e-19	9.5789e-20	2.0570e-20	3.2786e-14	7.5158e-19	3.3799e-17	1.8995e-21	1.315
320	7.6345e-18	1.6058e-22	6.1733e-20	2.6048e-23	4.6965e-18	2.3503e-21	3.5754e-19	3.0939e-21	1.1283e-18	1.7728e-19	2.6987e-22	2.0397e-22	4.6643e-23	4.4207e-16	2.5047e-22	7.6455e-20	1.4191e-24	1.153
321	3.1553e-16	2.2183e-18	1.3282e-16	4.9086e-19	8.6371e-17	1.4885e-17	3.6791e-16	6.7923e-18	2.3087e-16	4.0346e-16	1.4399e-18	1.3181e-18	3.7576e-19	3.7726e-15	4.7157e-18	7.1691e-17	2.1519e-20	1.270
322	1.7417e-17	2.4139e-20	1.8123e-18	2.7893e-21	5.6182e-18	4.0181e-19	2.6070e-18	9.8276e-20	2.6150e-18	6.2665e-18	2.8651e-20	1.9422e-20	2.1627e-21	2.6031e-15	7.6834e-20	2.4196e-18	1.8707e-22	1.915
323	1.9305e-17	8.4107e-20	5.7346e-17	6.1529e-21	2.4776e-17	3.8986e-19	7.7742e-18	1.0666e-19	1.6320e-17	3.4543e-18	4.9070e-20	7.1191e-20	3.7164e-21	6.5007e-17	3.2226e-20	8.9214e-20	9.4540e-22	1.913
324	1.0597e-16	2.3415e-18	1.5147e-15	2.5643e-19	1.1861e-16	6.8531e-18	1.1233e-16	1.3424e-18	1.4802e-16	8.7266e-17	6.9573e-19	2.3081e-18	2.2564e-19	1.0125e-16	1.4436e-18	9.3384e-19	3.7771e-20	1.254
325	1.6852e-13	6.2333e-14	1.8333e-11	1.4034e-12	3.8280e-14	3.8470e-13	5.4885e-14	2.4636e-14	3.8398e-12	2.8952e-14	2.5938e-12	4.7729e-11	3.8474e-12	1.3796e-14	2.0045e-13	1.0476e-14	5.4845e-14	1.460
326	1.3775e-13	5.0630e-13	3.6025e-12	7.8441e-13	1.1408e-13	5.7053e-13	3.2932e-13	1.0675e-13	1.6665e-12	3.7743e-13	8.5688e-13	5.7283e-10	7.3357e-12	2.4200e-14	7.9272e-13	2.1491e-14	3.7997e-14	1.405
327	6.5256e-15	6.6773e-15	1.8038e-13	6.2524e-14	8.5161e-15	2.6144e-14	2.0400e-14	2.4373e-15	1.5782e-13	1.0534e-14	5.9064e-14	1.4521e-11	9.1651e-13	7.7698e-16	3.9553e-14	1.0752e-15	3.4530e-15	2.731
328	3.9184e-13	1.3041e-13	5.1460e-11	3.9400e-12	9.4294e-14	9.3908e-13	1.5424e-13	5.4329e-14	1.0139e-11	6.4393e-14	4.6621e-12	1.2443e-10	1.3834e-11	2.4396e-14	3.8783e-13	2.5732e-14	1.4721e-13	3.084
329	3.0928e-14	1.6731e-14	3.3260e-13	1.7375e-13	1.5051e-14	6.8308e-14	1.1591e-13	1.1610e-14	6.3136e-13	5.1387e-14	1.4450e-13	2.1211e-11	1.6495e-12	2.3663e-15	1.2242e-13	6.9775e-15	1.0944e-14	1.593
330	1.7884e-13	9.1281e-14	3.0197e-11	1.7494e-12	7.0114e-14	6.0194e-13	8.5829e-14	3.3004e-14	3.4465e-12	4.3197e-14	3.7281e-12	1.2973e-10	4.8934e-12	2.1867e-14	2.6438e-13	1.2113e-14	8.1242e-14	1.471
331	4.9408e-15	7.0029e-15	9.6454e-13	5.5839e-14	7.5546e-15	3.8480e-14	4.3462e-15	1.9377e-15	6.7089e-14	6.0672e-15	2.3214e-13	1.0415e-11	1.8770e-13	2.5824e-15	2.7172e-14	5.4216e-16	3.0039e-15	1.354
332	1.5455e-15	1.5101e-15	3.1685e-13	1.7402e-14	4.4016e-15	6.1123e-15	1.8133e-15	5.6852e-16	1.4097e-14	1.7091e-15	9.5944e-14	8.3436e-12	9.6916e-14	6.4449e-16	9.3619e-15	1.5988e-16	8.1554e-16	2.650
333	4.5253e-15	4.5030e-15	3.4383e-13	4.2127e-14	3.4554e-15	2.0093e-14	4.9028e-15	1.4914e-15	9.9466e-14	6.7477e-15	6.5622e-14	2.6534e-11	5.4932e-13	4.4573e-16	1.0496e-14	3.9906e-16	1.3866e-15	5.973
334	2.1367e-13	2.3716e-13	2.9718e-13	7.8996e-12	1.6047e-13	1.6041e-13	1.1353e-14	1.2222e-14	1.4547e-14	3.0022e-14	3.4180e-13	5.0066e-11	1.0403e-12	4.6109e-14	8.7269e-13	1.3081e-14	2.3589e-14	2.214
335	7.3855e-13	8.4277e-13	1.0831e-12	1.2531e-11	3.7392e-13	4.7402e-13	3.9393e-14	4.2866e-14	3.3934e-14	1.0098e-13	1.1968e-12	2.8801e-10	2.1838e-12	1.4820e-13	2.0586e-12	3.4736e-14	8.9076e-14	5.221
336	5.2918e-14	5.7113e-14	1.1493e-13	2.4466e-12	4.7626e-14	5.2167e-14	2.5234e-15	3.0733e-15	3.9450e-15	5.7345e-15	4.9135e-14	8.2854e-12	2.8680e-13	1.7338e-14	1.9046e-13	2.5316e-15	1.4589e-14	6.656
337	2.2677e-14	7.7373e-15	8.1780e-14	6.9384e-13	3.9010e-14	8.8600e-15	4.3506e-16	1.7427e-16	4.5098e-15	5.1747e-16	3.8967e-14	1.2393e-13	1.2893e-13	8.8119e-15	3.0018e-14	2.7571e-16	7.2844e-15	4.940
338	1.1773e-14	3.6019e-15	3.1519e-14	4.1585e-13	1.5109e-14	4.4348e-15	1.9695e-16	6.6354e-17	2.1306e-15	2.0410e-16	1.2502e-14	8.3788e-14	5.9678e-14	2.6633e-15	1.3862e-14	1.1695e-16	2.7653e-15	2.508
339	8.9837e-15	5.4826e-15	1.8110e-14	7.5148e-13	1.3741e-14	6.7588e-15	2.3530e-16	1.1445e-16	1.8228e-15	2.6792e-16	1.2242e-14	8.2220e-14	1.4351e-13	4.2187e-15	2.7793e-14	1.6729e-16	3.4560e-15	4.648
340	4.2431e-15	3.0305e-15	1.6272e-14	1.8512e-13	1.0842e-14	4.4794e-15	1.4420e-16	4.8025e-17	9.8221e-16	1.7134e-16	7.5402e-15	4.0856e-14	6.0194e-14	2.4388e-15	1.1094e-14	8.3340e-17	1.9707e-15	1.913
341	1.4489e-19	3.7356e-23	1.0882e-23	6.3062e-24	1.3247e-25	1.2125e-24	9.5393e-21	8.0276e-14	1.4364e-22	1.2151e-19	3.4198e-22	1.3807e-22	1.2373e-23	1.8915e-19	1.4525e-20	6.3127e-16	6.5467e-23	5.993

Figure 3: Data Set Matrix

The Decision Error Trade-off (DET) curve plots false alarm rate or false positive identification rate (FPIR) and miss rate or false negative identification rate (FNIR) by varying this threshold, capturing both of the scenarios mentioned above. As specified in the IJB-A benchmark, we report FNIR at FPIR's of 0.1 and 0.01. The Face Scrub dataset [17] is an open-access database of face images of actors and actresses on the web, provided as hyperlinks from where the actual images can be downloaded.

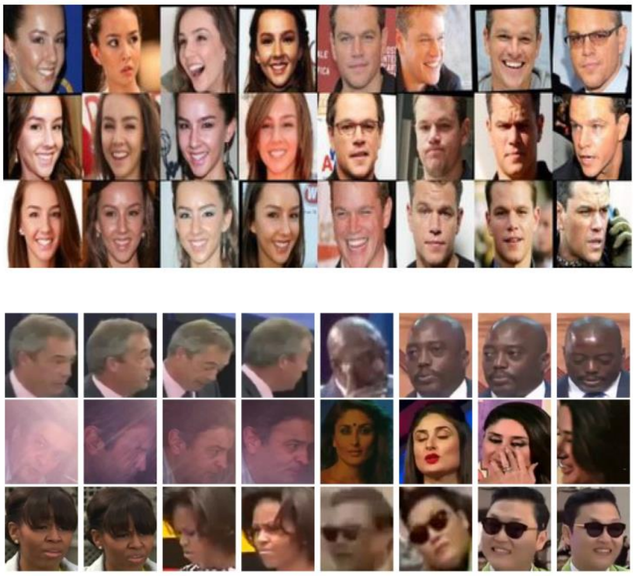


Figure 4: Data Set (Image Data Example)

It contains 530 persons with 107,818 still images in total. There are on average 203 images per person. In an additional experiment, we use this external data to first fine-tune the networks, before subsequent fine-tuning on the IJB-A train data. All overlapping identities between the two datasets are removed from Face-Scrub before training the networks on it. As some of the download links provided in Face-Scrub were broken (and after overlap removal) we finally train the networks on 513 identities, having a total of 89,045 images. We keep a third of the images in each class as validation sets and use the rest for training the networks.

V. Proposed Method

a) *Pre-processing*

The bounding boxes provided in the IJB-A metadata were used to crop out faces from the images. The images are resized according to the normalization parameters specified in the architecture details of a particular network (see below). This resizing does not maintain the aspect ratio.

b) *Network architectures*

As a baseline for deep models, we use the Imagenetpretrained "M-net" model from VGG's MatConvNet [26]. All results using this network architecture are hereafter referred to as "CNN". We consider the network outputs of the fully-connected layer after rectification, i.e. layer-19 ('fc7' + 'relu7') to be used as the face descriptor. An input image is resized to 224 * 224 following the way the network had been initially trained on Imagenet, resulting in a 4096- dimensional feature vector.

We use a symmetric bilinear-CNN model, denoted from now on as "B-CNN" that has both Network A and Network B set to the "M-net" model. Similar to the procedure followed in [15], the bilinear combination is done by taking the rectified outputs of the last convolutional layer in each network, i.e. layer-14 ('conv5' + 'relu5'). We chop off both the networks at layer 14, add the bilinear combination layer, a layer each for square-root and L2 normalization, and then a soft-max layer for classification. For this architecture, the image is up-sampled to be 448 X 448, resulting in a 27 X 27 X 512 output from each network at layer-14 (27 X 27 are the spatial dimensions of the response map and 512 denotes the number of CNN filters at that layer).

The bilinear combination results in a 512 X 512 output, and its vectorization (followed by the normalization layers mentioned earlier) gives us the final face descriptor.

c) *Network fine-tuning*

The models described in this set of experiments were trained initially for large-scale image classification on the Imagenet dataset. Fine-tuning the networks for the specific task of face recognition is expected to significantly boost performance. We consider three different scenarios with respect to fine-tuning:

no-ft: No fine-tuning is done. We simply use the Imagenet-pretrained model without any retraining on face images as a baseline for our experiments.

FaceScrub: The Imagenet-pretrained network is finetuned on the FaceScrub dataset by replacing the last layer with a softmax regression and running backpropagation with dropout regularization of 0.5 for 30 epochs. We begin fine-tuning with a learning rate of 0.001 for the lower layers and 0.01 for the last layer and divide them both by 10 if the validation error rate does not change. The stopping time is determined when the validation error remains constant even after learning rate is changed. The B-CNN is similarly finetuned for 70 epochs on FaceScrub data.

FaceScrub+Train: The FaceScrub data provides a good initialization for the face identification task to the networks, following which we fine-tune on the IJB-A train set for 30 epochs in case of the regular CNN and 50 epochs for the B-CNN. The fine-tuning on FaceScrub gives us a single model each for CNN and B-CNN. In the current setting, we take this network and further fine-tune it on each of the train sets provided in the 10 splits of IJB-A. This setting considers fine-tuning the network on images that are closer in appearance to the images it will be tested upon, i.e., the train set of IJB-A, being drawn from the same pool of imagery as the probe and gallery sets, is more similar to the test images than FaceScrub images.

d) Classifiers and pooling

One-versus-rest linear SVM classifiers are trained on the gallery set for all our experiments. We do not do any form of template-pooling at this stage and simply consider each image or video frame of a person as an individual sample. The weight vectors of the classifiers are rescaled such that the median scores of positive and negative samples are +1 and -1. Since all evaluations on this protocol are to be reported at the template level, we have the following straightforward strategies for pooling the media (images and video-frames) within a template at test time:

Score pooling: We use the max operation to pool the SVM scores of all the media within a probe template. This is done after SVM scores are computed for each individual image or frame that comprises a template.

Feature pooling: The max operator is applied on the features this time to get a single feature vector for a template. Thus, the SVM is run only once per probe template.

VI. Result and Analysis

We report two main sets of results. The first set of experiments shows that B-CNNs consistently outperform standard CNNs when fine-tuned on generic face data and also when adapted further to the specific data set at hand. The second set of results, using the pre-trained VGG-Face network, shows that even in a scenario when such massive training data is not readily available to the end user, a B-CNN can be constructed, without further network training that improves the final accuracy of the network. For both sets of results, we evaluate the task of open-set identification on this dataset, also referred to as the “1: N identification task”. We include the ROC curve plots which summarize the accuracy of the model.

Face Verification and ROC Analysis

The proposed method is tested as a face verification/authentication system. Face verification is to either verify or reject a claimed identity by comparing an unknown input face image of the person with other images belonging to the individual in a database. A similarity metric is needed to verify the claimed identities. If this metric or score is more than a specific threshold, the input identity is approved, otherwise rejected. In our work for sample *i*, the total of the votes of the classifiers for class *j* shows how similar sample *i* is to the stored images of class *j*, therefore this value can be used as the similarity score. The chance of a correct acceptance of sample *i* with claimed identity of class *j* increases with more classifiers voting sample *i* to belong to class *j*. this will result in a similarity matrix that describes how close one test sample is to each class. The size of the similarity matrix is number of test samples by number of classes.“Min error” is where the total of false accepts and false rejects is minimum. Equal error rate is the point where false accepts ratio (FAR) and false reject ratio (FRR) are equal. *Vth* can be selected based on the FAR (False Acceptance Rate) and FRR (False Recognition Rate) than can be tolerated in the desired application.

False Acceptance Rate (FAR)

The false acceptance rate, or FAR, is the measure of the likelihood that the biometric security system will incorrectly accept an access attempt by an unauthorized user. A system’s FAR typically is stated as the ratio of the number of false acceptances divided by the number of identification attempts.

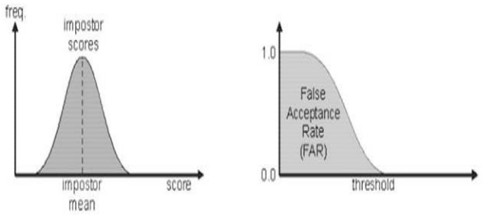


Figure 5 False Acceptance Rate (FAR)

False Recognition Rate

The false recognition rate, or FRR, is the measure of the likelihood that the biometric security system will incorrectly reject an access attempt by an authorized user. A system’s FRR typically is stated as the ratio of the number of false recognitions divided by the number of identification attempts.

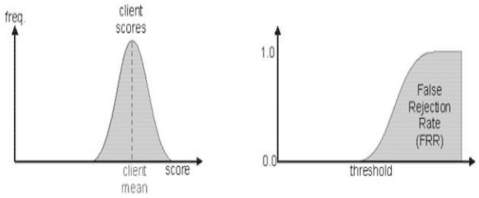


Figure 6 False Recognition Rate (FRR)

The False Reject Rate (FRR) measures the probability that an individual who has enrolled into the system is not identified by the system, It Occurs when the system says that the sample does not match any of the entries in the gallery, but the sample in fact does belong to someone in the gallery. The proportion of genuine or authentic attempts, whose, HD exceeds a given threshold. The rate at which a matching algorithm incorrectly fails to determine that a genuine sample matches an enrolled sample. It is also known as Type-I error.

FRR can be calculated as:

$$\text{FRR(n)}=\frac{\text{Number of rejected verification attempts for a qualified individual n}}{\text{Total number of verification attempts for that qualified individual n}}$$

And $\text{FRR}=\frac{1}{N}\sum_{n=1}^{N} FRR(n)$

Where ‘n’ is the total number of enrolments.

A. Result under Our Approach

In our approach, we are using Resnet50 deep neural network architecture with face alignment, In our approach, we fine retrieval rate of our model, ROC Curve of AUS=0.9990, False Acceptance Rate, False Reject Rate respectively show in figure 7, figure 8, figure 9 and figure 10.

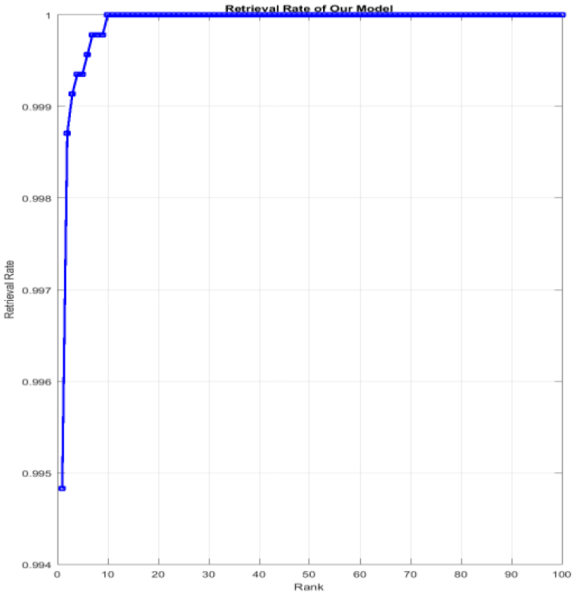


Figure 7 Retrieval Rate of Our Model

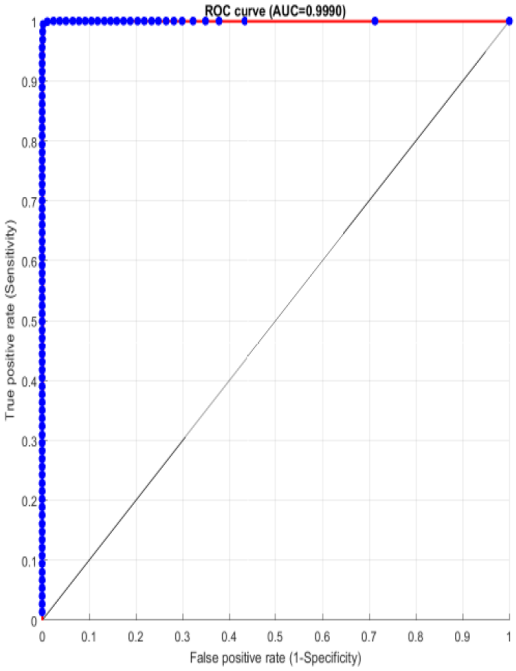


Figure 8 ROC Curve (AUC=0.9990)

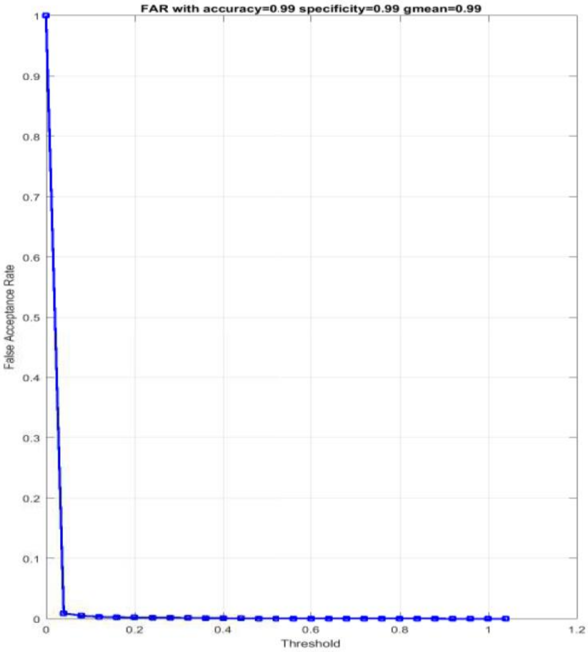


Figure 9 FAR with accuracy=0.99 specificity=0.99 gmean=0.99

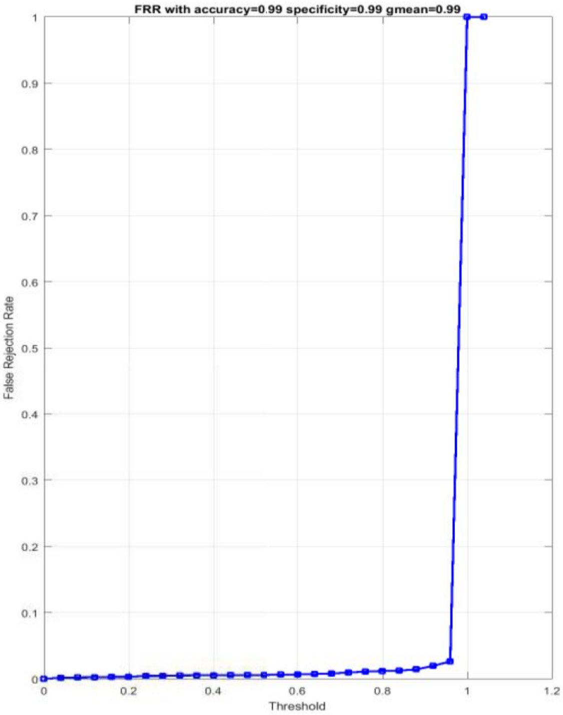


Figure 10 FRR with accuracy=0.99 specificity=0.99 gmean=0.99

B. Result under VGGNet 16 Fine Tune All Layers

VGGNet 16 is standard deep neural network architecture, we use this architecture for two ways comparison. Here, we use first architecture way fine tune all layers.

In VGGNet 16 we find retrieval rate of VGGNet 16 Fine Tuned all Layers, ROC Curve of AUS=0.7423, False Acceptance Rate, False Reject Rate respectively show in figure 11, figure 12, figure 13, figure 14.

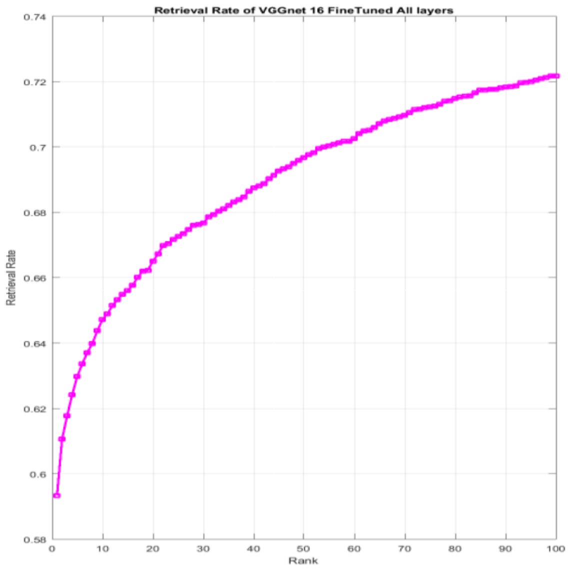


Figure 11 Retrieval Rate of VGGNet 16 Fine Tuned All Layers

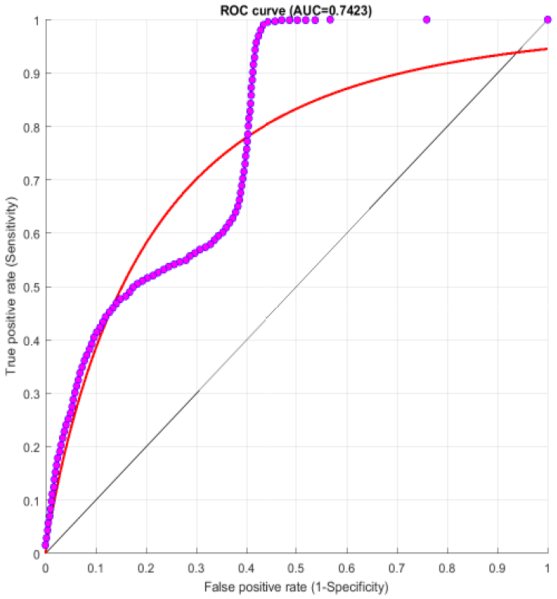


Figure 12 ROC Curve (AUC=0.7423)

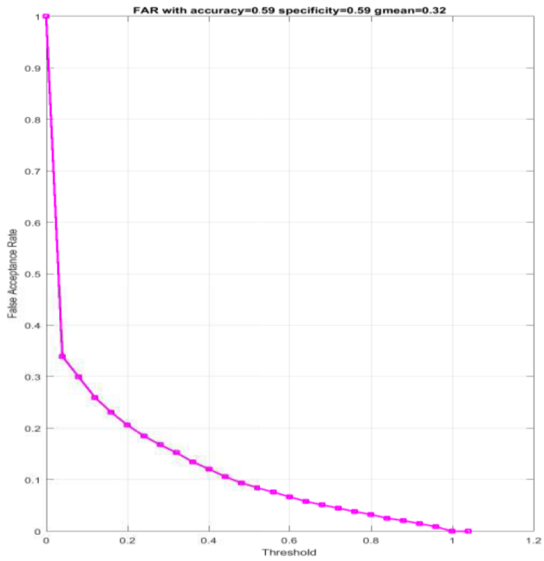


Figure 13 FAR with accuracy=0.59 specificity=0.59gmean=0.32

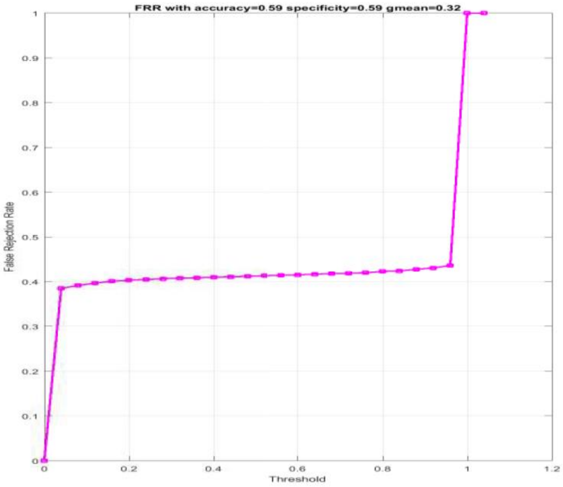


Figure 14 FRR with accuracy=0.59 specificity=0.59gmean=0.32

C. Result under VGGNet 16 FT last layers

VGGNet 16 is standard deep neural network architecture; we use this architecture for two ways comparison. Here, we use second architecture way fine tune last layers of this network over our database.

In VGGNet 16 we find retrieval rate of VGGNet 16 Fine Tuned last Layers, ROC Curve of AUS=0.7059, False Acceptance Rate, False Reject Rate respectively show in figure 15, figure 16, figure 17, figure 18

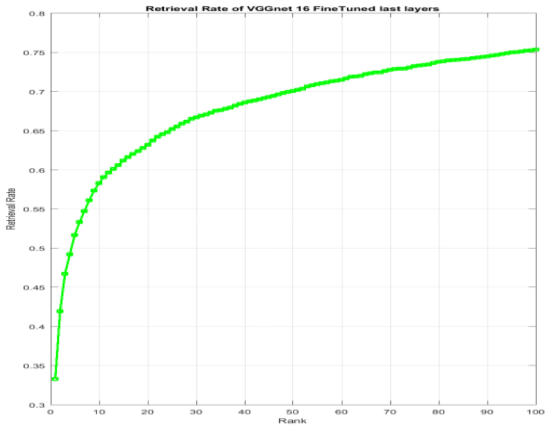


Figure 15 Retrieval Rate of VGGNet 16 Fine Tuned last Layers

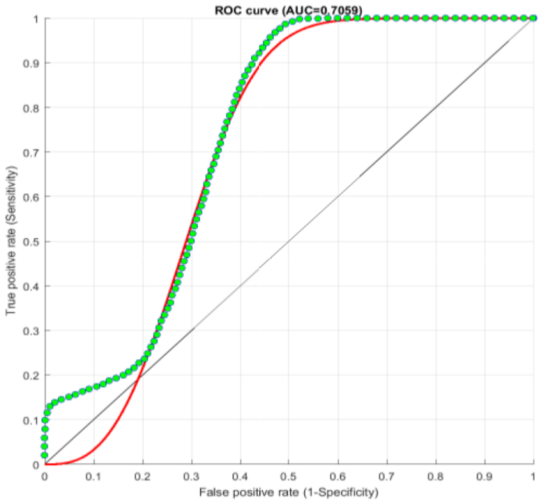


Figure 16 ROC Curve (AUC=0.7059)

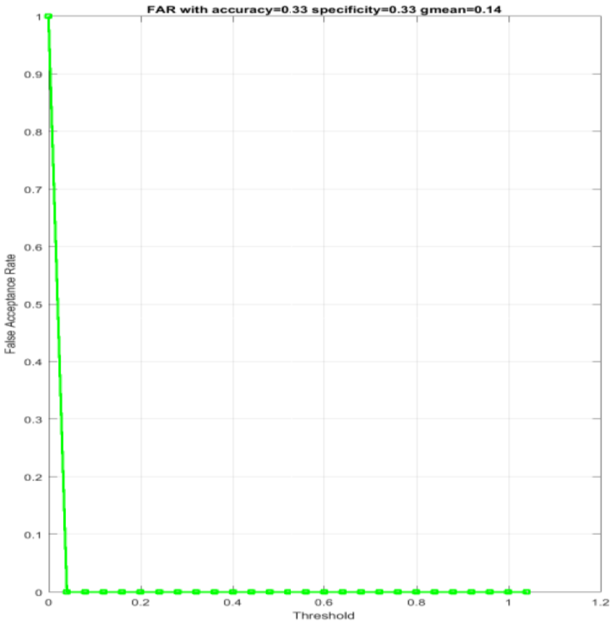


Figure 17 FAR with accuracy=0.33 specificity=0.33gmean=0.14

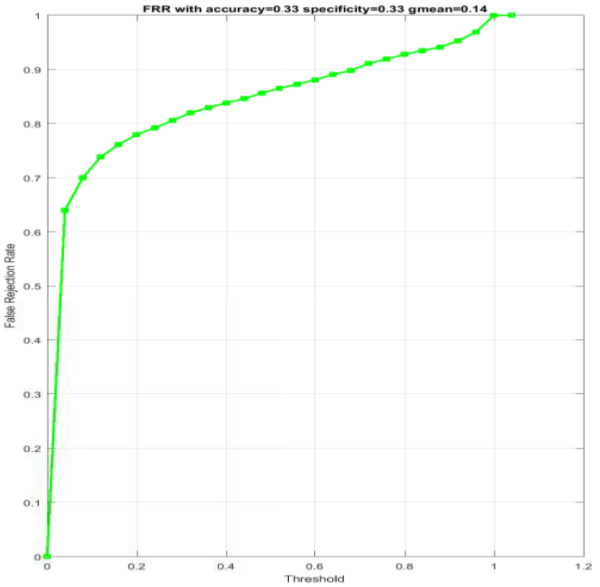


Figure 18 FRR with accuracy=0.33 specificity=0.33 gmean=0.14

VII. Conclusion

Automatic analysis of human affective behavior has been extensively studied in past several decades. Face expression recognition systems, in particular, have matured to a level where automatic detection of small number of expressions in posed and controlled displays can be done with reasonably high accuracy. Detecting these expressions in less constrained settings during spontaneous behavior, however, is still a challenging problem. In recent years, increasing number of efforts has been made to collect spontaneous behavior data in multiple modalities. The research shift towards this direction suggests utilizing the multimodal data analysis approaches.

CNN architectures that have succeeded on other fine-grained recognition problems also do well at face identification, both after fine-tuning and also as a simple modification to pre-trained models. There are a number of directions to continue exploring these models:

- Re-training the entire model using an objective similar to the Multi-view CNN objective in which the parameters of the network are learned under the assumption that the max will be taken across the multiple images in a template.
- Using datasets much larger than Face-Scrub, such as the CASIA Web-Face (with 10,000 identities and half-a-million images) to train the network, should further improve the performance.
- Training a very deep architecture from scratch on a sufficiently large face dataset, instead of fine-tuning pre-trained networks.

We believe the success of the B-CNN relative to non-bilinear architectures makes them a sensible starting point for a wide variety of continued experiments.

References

[1] H. Fujita and H. Akagi, “The unified power quality conditioner: the integration of series and shunt-active filters,” IEEE Trans. Power Electronics, vol. 13, no. 2, pp. 315-322, Mar. 1998.

[2] H. Schneiderman, T. Kanade, A statistical method for 3D object detection applied to faces and cars, Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, CVPR2000 (2000) 746-751 vol.1.

[3] C. Liu, H. Wechsler, Comparative Assessment of Independent Component Analysis (ICA) for Face Recognition, International Conference on Audio and Video Based Biometric Person Authentication. (1999) 22-24.

[4] J. Yang, A.F. Frangi, J.Y. Yang, D. Zhang, Z. Jin, KPCA plus LDA: a complete kernel Fisher discriminant framework for feature extraction and recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence, 27 (2005) 230-244.

[5] E. Osuna, R. Freund, F. Girosit, Training support vector machines: an application to face detection, Proceedings of IEEE Computer Society Conference on. Computer Vision and Pattern Recognition (1997) 130-136.

[6] P. Viola, M.J. Jones, Robust Real-Time Face Detection, International Journal of Computer Vision, 57 (2004) 137-154.

[7] F.R. Bach, M.I. Jordan, Kernel Independent Component Analysis, Journal of Machine Learning Research, 3 (2002) 1-48.

[8] S.S. Reddi, Radial and Angular Moment Invariants for Image Identification, IEEE Transactions on Pattern Analysis and Machine Intelligence, 3 (1981) 240-242.