Face image quality assessment for face selection in surveillance video using convolutional neural networks

3 authors, including:

Sumohana Channappayya
Indian Institute of Technology Hyderabad
96 PUBLICATIONS **1,017** CITATIONS

Some of the authors of this publication are also working on these related projects:

Photoacoustic Imaging View project

Stereoscopic image and video quality assessment View project

Face Image Quality Assessment for Face Selection in Surveillance Video using Convolutional Neural Networks

Vignesh S, Manasa Priya K. V. S. N. L., Sumohana S. Channappayya, *Member, IEEE*

Dept. of Electrical Engineering

Indian Institute of Technology Hyderabad

`{ee11b033, ee12m1020, sumohana}@iith.ac.in`

Abstract—**Automated Face Quality Assessment (FQA) plays a key role in improving face recognition accuracy and increasing computational efficiency. In the context of video, it is very common to acquire multiple face images of a single person. If one were to use all the acquired face images for the recognition task, the computational load for Face Recognition (FR) increases while recognition accuracy decreases due to outliers. This impediment necessitates a strategy to optimally choose the good quality face images from the pool of images in order to improve the performance of the FR algorithm. Toward this end, we propose a FQA algorithm that is based on mimicking the recognition capability of a given FR algorithm using a Convolutional Neural Network (CNN). In this way, we select those face images that are of high quality with respect to the FR algorithm. The proposed algorithm is simple and can be used in conjunction with any FR algorithm. Preliminary results demonstrate that the proposed method is on par with the state-of-the-art FQA methods in improving the performance of FR algorithms in a surveillance scenario.**

Index Terms—**Convolutional Neural Networks, Face Image Quality Assessment, Face Recognition, Surveillance.**

I. INTRODUCTION

In the past few decades, Face Recognition (FR) has received great attention not only due to its numerous applications, including video surveillance, access control, entertainment and law enforcement but also to understand the FR process in humans. Since FR is the natural way of identification and verification, this field is rich with excellent literature [1], [2], [3]. In the last two decades various algorithms have been proposed for FR based on still images and video sequences. However, in realistic scenarios, FR is limited by low quality images and variation in pose, illumination, occlusion and expression in the acquired face image [2]. Such problems are even more severe in surveillance systems where users may be uncooperative and the environment is uncontrolled. Since poor quality images in the surveillance video sequences offer very little information for FR. They not only increase the computational load due to complex processes such as feature extraction and matching, it also reduce the recognition accuracy because of outliers. To address this problem, many algorithms have been proposed in recent years to select the subset of high quality faces and avoiding outliers.

To the best of our knowledge, Berrani et al. [4] were among the first to address this issue by using statistical approaches to remove outliers. However, this approach is not suited to the surveillance scenario where a majority of the images are of poor quality. There are several algorithms for face quality assessment that are based on estimating facial properties such as estimating the pose [5], calculating the asymmetry of the face by estimating out of plane rotation, and non-frontal illumination to quantify the degradation of the quality [6], [7], [8]. These methods consider only a subset of factors affecting FR and hence not suitable for robust image selection. Instead of considering the factors affecting FR and fusing the scores, Wong et al. proposed the definition of "standard" face as frontal faces with uniform illumination and there by simultaneously considering the variations in pose, sharpness and alignment errors [9]. However, the applications are limited to the above algorithms since these are based on hand-picked factors affecting the face image and assumptions on "standard" face. Also, these algorithms do not leverage the strengths of FR algorithms i.e., the above algorithms do not fully utilize the abilities of FR algorithms that may be good at recognizing faces with occlusions, pose variations or non-uniform illumination.

Chen et al. [10] proposed an algorithm based on multiple feature fusion and learning to rank to address the above issue. In this algorithm, they considered three databases with faces acquired in a controlled environment, an uncontrolled environment, and with non-face images respectively. Then they ranked the databases based on the recognition performance and assumed the faces in the same database to have equal rank. Their algorithm is implemented in two levels. In the first level, they learned the weights for the feature vector of face images from the above-mentioned datasets by using a linear kernel such that the sum of weighted feature resembles the ranking of databases. For this, they considered five different feature vectors and learned the corresponding weights. In the second level, they combined five first level scores with respect to each feature vector by using a second order polynomial kernel to give the final quality score for the given probe face image.

The motivation of our algorithm is fundamentally similar to learning to rank based algorithm, however it uses a different and novel approach that is based on modelling the system response of an FR algorithm using CNN.

The paper proceeds as follows: Section II describes our proposed algorithm in the framework of FR System. Section

III discusses the experimental set-up of FQA. Section IV discusses our experiments on ChokePoint dataset [9] and compares our algorithm with other face selection algorithms. Section V concludes the discussion and gives the direction for future work.

II. FACE RECOGNITION SYSTEM

A typical FR system comprises of face detection, face localization, face subset selection (optional), face feature extraction and face matching components as shown in the Fig.1 [2]. In the proposed algorithm, the face subset selection component is analysed and the best subset of faces are selected for further processing in order to enhance the performance of FR. The details of each component is presented below.
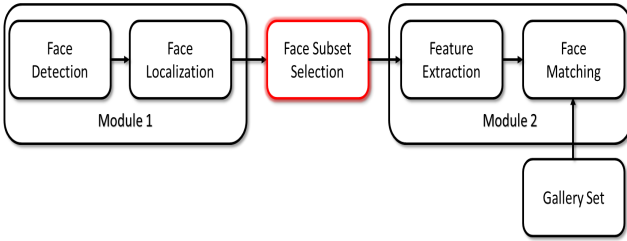

Fig. 1: Face Recognition System

A. Face Detection and Localization

To detect and localize the face in each frame, we used Viola Jones Haar feature based cascade classifier [11]. We omitted the faces where the Haar cascade classifier is not able to detect the face. After detecting the facial region from the aforementioned classifier, localization is done by fixing the center and cropping the facial region by omitting borders. The final cropped images are then resized to 64×64 pixels.

B. Face Subset Selection

It is quite common in a video surveillance scenario to acquire multiple face images of same person. Selecting the subset of faces with high quality improves the performance of recognition algorithm by removing the outliers. It also reduces the complexity of FR algorithm, considering the fact that face feature extraction process is computationally expensive and complex. As discussed in section I, it is difficult to define the quality of a face image. Several researchers have defined the quality in different ways. In this paper, we define the quality with respect to the FR algorithm.

The motivation behind our definition of the quality is explained as follows by considering an example. If an FR algorithm is good at recognizing the faces with pose variations but not able to recognize the faces with non-uniform illumination, then the faces with pose variations should be considered high quality and faces with non-uniform illumination should be considered low quality. Since different FR algorithms work better in different aspects such as occlusion, pose variations, illumination variations, hence fixing the definition of quality doesn't take the full advantage of the FR algorithm under

consideration. So, in this paper, we propose a novel take on face quality definition and choose to define the quality of the face image with respect to the FR algorithm.

For convenience, we categorize the FR system into two modules. The first module consists of face detection and face localization. The second module is the FR algorithm that consists of face feature extraction and face matching. As an FQA algorithm, we need to predict the face images that performs best in the second module of the given FR algorithm and this is not a trivial problem. Toward this end, we considered the second module as an unknown system (or a black box) and have attempted to model its system response using a CNN. To achieve this, we first used a training set of images to evaluate the system performance of the FR algorithm. Thus, by knowing the input and output to the black box, we model its performance such that it is able to predict the quality of test images a priori. By this, we could select the high quality face images that are best recognized by the FR algorithm.

A CNN is used to model the performance of the FR algorithm due to its strengths over other modelling techniques. Since a CNN accepts the entire 2D image as input, there is no need for complicated image transformation or feature extraction. This feature is particularly useful in our case where the definition of quality of face image is not fixed. So, the CNN learns its parameters and defines the quality of the face image depending on the FR algorithm. The proposed algorithm is depicted in figure 3 and framework of CNN is explained in section III.

C. Face Feature Extraction

After selecting the high quality faces, features are extracted from each face image to form the feature vector and is used as input for face matching. For this purpose we use two feature extraction techniques, Local binary patterns (LBP) [12] and Histogram of Oriented Gradients (HOG) [13], to show that our FQA algorithm works across different FR algorithms. LBP makes use of both shape and texture information of the face while HOG makes use of the distribution of intensity gradient or edge directions to describe a image. Though the feature extraction techniques that we used are not exhaustive. However, they form the reasonable subset to prove that our FQA algorithm works across the different FR algorithm and still able to select high quality images with respect to the given FR algorithm.

D. Image Set Matching

We make use of Mutual Subspace Method (MSM) for face image set matching [14]. The two image sets are considered similar if the canonical angle between two image sets is within the threshold. For each video sequence in the gallery set, feature vectors are calculated and compared with feature vectors of probe sequence using MSM. In MSM, the probe sequence and each video sequence in gallery set is considered as separate subspaces and similarity between the subspace is measured by calculating the mean canonical angle between the two subspaces.
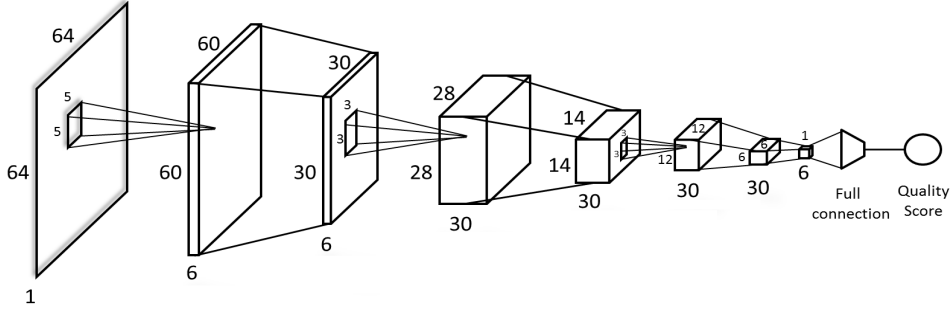
Fig. 2: Our CNN Architecture

Let the probe sequence be N dimensional subspace and each video sequence in the gallery set be M dimensional subspace. The canonical angle for each element of the probe sequence is defined as the maximum angle between the given element and all the elements of the M dimensional subspace for a given video sequence in the gallery. The final similarity score is calculated by taking the mean of canonical angle of all the elements of the probe sequence. The score is then compared to a threshold and final decision is made whether the probe and the gallery sequence pair is matched/mismatched pair. The threshold is obtained from a labeled set at which the total number of false positives and false negatives is minimum. This is referred as Minimum Error Rate (MER).

III. EXPERIMENTAL SETUP FOR FQA

The proposed algorithm uses CNN for face image quality estimation. Given the face image, it is resized to 64×64 pixels and PCA whitening is done to make it less redundant such that face image is less correlated and then input to the CNN to estimate the quality.

As a preliminary implementation, our CNN has four convolution layers with sub sampling. The first convolutional layer has 6 kernels with each of size 5×5 and produces 6 feature maps with size 60×60, followed by sub sampling layer. The second convolutional layer has 30 kernels with size 3×3 and produces 30 feature maps with size 28×28 and then followed by sub sampling layer. The third convolutional layer has 30 kernels each of size 3×3 and produces 30 feature maps with size 12×12 and then followed by sub sampling layer. The final convolutional layer has 6 kernels with each of size 6×6 and then followed by linear regression with one dimensional output that gives the quality score of the face image. The architecture of our CNN is shown in Fig. 2.

In our experiments, we use the ChokePoint dataset [9] that is ideally suited for face recognition/verification in the surveillance scenario. The dataset consists of 25 subjects (19 male and 6 female) with 64,204 face images. We divide the images into training and testing sets. Set 1 contains image sequences of 13 subjects for training the CNN and Set 2 contains the rest of the images sequences to evaluate the performance of the FR algorithm.

A. Training

For training, we assign each face image with a quality score. This score is evaluated based on how the FR algorithm is able to recognize the input face image given the faces in the gallery set. Without loss of generality, we consider the FR algorithm with LBP/HOG for feature extraction and MSM for face matching. Since the MSM score depicts the performance of recognition algorithm, we assign the quality of the given input face image with MSM score to train the CNN.

Let I_n be the preprocessed input face image, Q_n be the quality/MSM score and $f(I_n, W)$ be the predicted score of the input face image where W is the weight matrix of the network. The network parameters/weight matrix is obtained by solving the following objective function.

$$S = \frac{1}{N} \sum_{n=1}^{N} \|f(I_n, W) - Q_n\|_2^2,$$

$$W' = \underset{W}{\operatorname{argmin}}\ S$$

where N is the total number of images in the training set. Back propagation and Stochastic gradient descent method are used to minimize the above cost function.

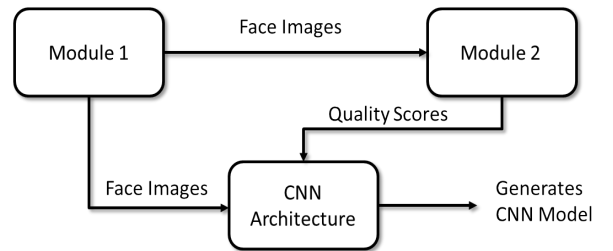We used a batch size of 50 and ran the CNN for 500 epochs. The entire training process is depicted in Fig. 3.

Fig. 3: Training the CNN for predicting face quality.

B. Testing

The preprocessed face images in the probe sequence is given as the input to the trained CNN and quality scores are predicted for each face image in the sequence. The CNN model is trained such that these scores resemble the MSM scores of the FR algorithm. Subset selection is done by sorting the predicted

quality scores of face image sequence and taking the top N images from the sorted list. Then these selected N images are given to the FR algorithm for further processing. The subset selection process is illustrated in Fig. 4.
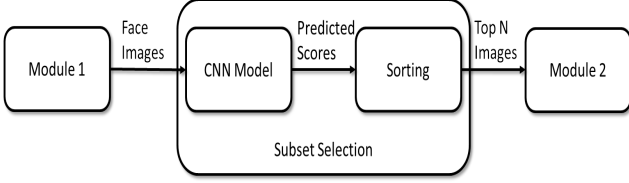

Fig. 4: Testing the CNN face quality model.

IV. RESULTS AND DISCUSSION

Any FR algorithm can be used in conjunction with proposed face quality assessment algorithm for improved accuracy and computational efficiency. The test set for evaluating the performance of FR algorithm is split further into two sets $G1$ and $G2$ where each dataset plays the role of development and evaluation sets respectively. The subjects that are used for training the CNN model are not used here to evaluate the performance of the FR algorithm.

We performed our experiment in two phases. In first phase, $G1$ is considered as the development set and $G2$ as the evaluation set and roles of $G1$ and $G2$ are reversed in the second phase. By considering one group as development set (labeled set), we calculated matched and mismatched scores. Then, we find the threshold where the sum of False Acceptance Rate (FAR) and False Rejection Rate (FRR) is minimum i.e., Minimum Error Rate. By applying this threshold on the scores of pairs of evaluation sets, recognition rate (R_{G2}) is calculated as follows

$$R_{G2} = 0.5 \times [(1 - FAR) + (1 - FRR)].$$

In the second phase, recognition rate(R_{G1}) is calculated in a similar fashion and final recognition rate (R_{avg}) is calculated as follows

$$R_{avg} = 0.5 \times (R_{G1} + R_{G2}).$$

For subset selection, we selected N high quality images for face recognition based on different selection metrics and characterized how different metrics improve the recognition performance.

We compare the performance of the proposed FQA algorithm with four other selection methods: (1) sequential selection, (2) random selection, (3) quality assessment based on patch-based probabilistic approach [9], (4) Learning to Rank based quality assessment [10]. After face selection, the aforementioned protocol is used to compare the results by varying N from 4 to 16. Results using LBP and HOG for feature extraction and MSM for face matching with different selection methods are shown in Table I and II. From the results, we can infer that high verification performance is achieved by the proposed method which in turn implies that it is able to select the best subset of faces from the sequence of faces. Further, it can be observed that the proposed method outperforms the random, sequential methods and the patch-based

probabilistic approach and is on par with the learning to rank based approach. In the patch-based probabilistic approach, the subset selection assumes that the FR algorithm can recognize face images only when they resemble standard faces. This assumption may not always be true. The rank based approach makes use of five feature extraction algorithms which is an expensive step (the step which motivated the researchers to work on subset selection). Also, they considered the same rank for all the images in a single database which may not be the true in all cases.

In our experiments, we use MSM for face matching which does not consider the relationship among the acquired face images. Thus there might be the chances that the best subset of faces might have identical information. If we choose the FR algorithm that considers the relationship among the acquired face images, then our FQA implementation can be modified to take multiple images as input and will be able to select the best subset of faces that doesn't have identical information for the FR algorithm under consideration. Hence, we can use our FQA implementation in conjunction with any FR algorithm and fully leverage the ability of the FR algorithm.

TABLE I: Video-based face verification performance on the ChokePoint dataset, using LBP and MSM (higher is better).

Subset Selection Method	N=4	N=8	N=16
Sequential	0.6114	0.6174	0.6278
Random	0.6825	0.691	0.704
Probabilistic based [9]	0.6995	0.7181	0.7252
Rank based [10]	0.7328	0.7511	0.7645
Proposed Method	0.7226	0.7564	0.7786

TABLE II: Video-based face verification performance on the ChokePoint dataset, using HOG and MSM (higher is better).

Subset Selection Method	N=4	N=8	N=16
Sequential	0.6419	0.6504	0.6669
Random	0.7329	0.7552	0.7706
Probabilistic based [9]	0.7603	0.7753	0.7876
Rank based [10]	0.7843	0.787	0.7857
Proposed Method	0.7589	0.7775	0.7917

V. CONCLUSIONS AND FUTURE WORK

In this paper, we presented a novel FQA algorithm. We were motivated by the fact that different FR algorithms have different capabilities in recognition, and therefore the quality of the image should be defined with respect to the FR algorithm. To achieve this, a CNN is used to define the quality by training its network with the score/value that depicts the performance of the FR algorithm in consideration. By using this quality measure to sort the input sequence and taking only high quality images we successfully demonstrated that it not only increases the recognition accuracy but also reduces the computational complexity. From the initial results, we strongly believe that the proposed algorithm is promising and has attractive features. As part of future work, we plan to improve the performance of the algorithm by fine-tuning the parameters.

REFERENCES

[1] J. R. Barr, K. W. Bowyer, P. J. Flynn, and S. Biswas, "Face recognition from video: A review," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 26, no. 05, 2012.

[2] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *Acm Computing Surveys (CSUR)*, vol. 35, no. 4, pp. 399–458, 2003.

[3] Y. Wong, M. T. Harandi, and C. Sanderson, "On robust face recognition via sparse coding: the good, the bad and the ugly," *IET Biometrics*, vol. 3, no. 4, pp. 176–189, 2014.

[4] S.-A. Berrani and C. Garcia, "Enhancing face recognition from video sequences using robust statistics," in *Advanced Video and Signal Based Surveillance, 2005. AVSS 2005. IEEE Conference on*, pp. 324–329, IEEE, 2005.

[5] Z. Yang, H. Ai, B. Wu, S. Lao, and L. Cai, "Face pose estimation and its application in video shot selection," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, vol. 1, pp. 322–325, IEEE, 2004.

[6] X. Gao, S. Z. Li, R. Liu, and P. Zhang, "Standardization of face image sample quality," in *Advances in Biometrics*, pp. 242–251, Springer, 2007.

[7] J. Sang, Z. Lei, and S. Z. Li, "Face image quality evaluation for iso/iec standards 19794-5 and 29794-5," in *Advances in Biometrics*, pp. 229–238, Springer, 2009.

[8] G. Zhang and Y. Wang, "Asymmetry-based quality assessment of face images," in *Advances in Visual Computing*, pp. 499–508, Springer, 2009.

[9] Y. Wong, S. Chen, S. Mau, C. Sanderson, and B. C. Lovell, "Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on*, pp. 74–81, IEEE, 2011.

[10] J. Chen, Y. Deng, G. Bai, and G. Su, "Face image quality assessment based on learning to rank," *Signal Processing Letters, IEEE*, vol. 22, no. 1, pp. 90–94, 2015.

[11] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, no. 2, pp. 137–154, 2004.

[12] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face recognition with local binary patterns," in *Computer vision-eccv 2004*, pp. 469–481, Springer, 2004.

[13] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1, pp. 886–893, IEEE, 2005.

[14] O. Yamaguchi, K. Fukui, and K.-i. Maeda, "Face recognition using temporal image sequence," in *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, pp. 318–323, IEEE, 1998.