Face Image Quality Vector Assessment for Biometrics Applications

Nima Najafzadeh, Hossein Kashiani, Mohammad Saeed Ebrahimi Saadabadi, Niloufar Alipour Talemi,
Sahar Rahimi Malakshan, and Nasser M. Nasrabadi
West Virginia University
{nn00008, hk00014, me00018, na00027, sr00033}@mix.wvu.edu, nasser.nasrabadi@mail.wvu.edu

Abstract

In this paper, we propose a multi-task convolutional neural network which produces an image quality vector for an input face image. This image quality vector contains the face quality score and the information about the nuisance factors (i.e., pose, illumination, blurriness and expression) that caused the predicted image quality score. Our multi-task network utilizes a pretrained ResNet-50 as its stem. Employing different data augmentation techniques, we create a huge and diverse dataset. We ground truth this dataset and use it to fine-tune our multi-task neural network. Our Multi-task learning framework enables us to learn a shared and beneficial feature representation among the relevant tasks to achieve a better performance. Moreover, the proposed multi-task neural network provides useful information about the nuisance factors.

Nuisance factors information is useful for applications like face image quality assessment during the automatic enrollment process where user's photo should comply with a standardized criteria. In this case, if the user upload a low-quality image which does not comply with a predefined standard, the system provides feedback about the nuisance factors associated with the captured image. Therefore, the user can resolve the problem, and upload a new image that can remedy the issue. Although an extensive research has been done on face image quality assessment, non of them have addressed face image quality vector effectively. To the best of our knowledge, our method is the first method that uses deep learning approach to generate a face image quality vector. The evaluation of results demonstrates that our method gets higher or comparable accuracy for face image quality assessment in comparison to the state-of-the-art methods and analyzes the input image to provide detailed information about the nuisance factors.

1. Introduction

Among different biometrics modalities, face images are one of the most utilized traits, and face recognition is the most active research area in the field of biometrics [1][2][3].The performance of a face recognition system depends on the utility (usefulness) of its input to a large extent. That is, the accuracy of a face recognition system should depend on how good are its input images. This utility of a face image for biometrics systems is measured in terms of image quality. Therefore, if the input image of a face recognition system is of a low quality, the accuracy of that system would be low, and its identification score would be inaccurate. In contrast, if the input image is of the high quality, the accuracy of the face recognition system would be high.

Due to the significant effect of the quality of the input image on the accuracy of the biometrics systems specially face recognition systems, recently, face image quality assessment has received a large amount of attention in the research community. As a result, developing a model that can predict the quality of a face image would lead to an improvement in the overall performance of biometrics systems. To be more specific, the assessment of face image quality for face recognition systems is very important because in many real-world scenarios face recognition systems work under unconstrained environments. For example, video surveillance systems capture face images under a huge variation in illumination, pose, expression, occlusion, and many other nuisance factors. In this situation, developing a Face Image Quality Assessment (FIQA) model can help the overall performance of a face recognition system to be more stable by filtering out the very low-quality images and keeping the moderately high-quality ones.

An FIQA model has utility in photo acceptance applications where a single scalar value that represents the quality of a face image can be used to make acceptance or rejection decision. For example, when a user uploads a low-quality image, the system rejects the image due to its quality score is below a pre-specified threshold, and asks the user to upload a new image. Another application of FIQA is in photo selection where the receiving system only accepts one image of the subject, however, there are more than one image available from that subject; therefore, the system can select the best image with the highest image quality. For example, when the goal is to recognize the identity of a suspicious person in a video sequence which is captured by a surveillance camera, FIQA can be used to select the frame with the highest image quality among the different frames of the video sequence. Also, quality summarization is another application of FIQA to be mentioned. In some enterprises, face images are captured from many subjects, by different staff, from different sites, and under different conditions. This is where a scalar image quality value can be used to assess the effectiveness of the collection. For instance, when multiple samples exists from a frequent traveler that were captured at different locations, and different times, the quality score of images can represent the effectiveness of collection at different sites.

Due to the paramount importance of face image quality, extensive research has been done in this area; consequently different methods for measuring the quality of a face image have been introduced. These methods can be categorized into two different categories the analytic-based [4][5][6] and learning-based methods [7][8][9]. The analytic methods in the first category [6] evaluate face image quality by using handcrafted features like illumination intensity, blurriness, odd skin color, vertical edge density, and so on. These methods require to manually extract features, but extracting all these features that degrade face image quality is unrealistic. Learning-based methods are based on the deep learning approach which is the most utilized approach for different computer vision tasks from image segmentation [10][11] and action recognition [12] to biometric-related tasks like face recognition [13][14] and morph detection [15]. Learning-based methods for FIQA obtain a quality score which is comparable to the recognition model outcome. In fact learning-based methods establish a mapping between the image quality and a recognition model.

Although a comprehensive research has been done on assessing face image quality, none of the current FIQA models provide sufficient information about the nuisance factors associated with a low-quality face image. This work utilize information about the nuisance factors that can be useful in several applications such as automatic verification of a person identity at a border kiosk, or assessing face image quality during the automatic enrollment process (e.g., visa application) where the user's photo must comply with a special criteria. In these cases, if the user receive a feedback from the system about the nuisance factors associated with his captured face image, or in case of uploading an image which does not comply with a predefined specification (standard), he can resolve the problem about his face image and upload a new image to comply with the required standard.

In this paper, we propose a new face image quality vector assessment model which is based on a multi-task convolutional neural network architecture. Our model receives a face image as an input and returns a quality vector. This quality vector not only contains the quality score of the input face image, but also it provides information about the pose, illumination, and expression of the input face. A sample of applying our model on an input face image is shown in Figure 1. We evaluated our method across two different face recognition systems and on two different datasets to assess its performance under different conditions. The results illustrate that our proposed method obtains higher or comparable accuracy in comparison with other state-of-the-art methods for predicting the face image quality. Also, our method obtains around 90% accuracy for estimating pose, illumination, and expression of the face image.

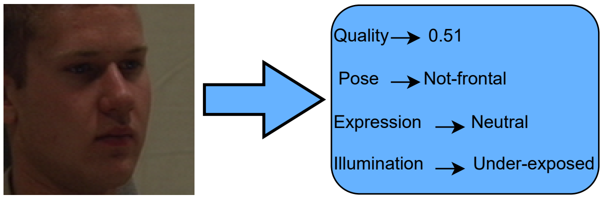


Figure 1: Example of a quality vector. The output of applying our proposed method on an input image is a quality vector. The quality vector contains quality, pose, expression, and illumination of the face image.

2. Related Works

In this section, we present some of the most relevant methods related to our work, which are divided into multi-task learning for face-related attributes, and face image quality assessment.

The concept of multi-task learning was first discussed in detail in [16]. Since that time it has been widely used in the computer vision area specially for extracting face-related attributes. Transferring and sharing knowledge among various tasks is the main reason that multi-task learning algorithms achieve a good performance in computer vision face-related problems. [17] was one of the earliest methods that used mixture of tree models with shared pool for jointly learning landmark localization, pose estimation, and face detection. HyperFace [18] fused intermediate layers of CNN to extract better features for face detection, landmark localization, pose and gender estimation. [19] trained a multi-task neural network for head-pose estimation, facial attribute inference and landmark localization. [20] proposed a multi-task learning framework that regularizes the shared parameters of CNN and builds a synergy among different domains and tasks for the purpose of simultaneously face detection, face alignment, pose estimation, gender recognition, smile detection, age estimation and, face recognition.

Although a lot of multi-task learning methods for face-related attributes have been proposed, none of these methods address the face image quality problem and its nuisance factors. Also, our approach uses asymmetric multi-task learning approach where the network is trained first for face image quality assessment as the primary task and then fine tuned for pose estimation, facial expression recognition, and illumination classification as auxiliary tasks that regularize and improve the performance of the primary task.

Face image quality assessment methods can be categorized into the analytic-based and learning-based approaches. Early face image quality assessment methods focused on analytic image quality factors. Most of these methods are inspired from the ISO/IEC 19794-5 [21] standard which is a series of guidelines for capturing high quality images. These methods evaluate face image characteristics like illumination, pose, expression, and occlusion and analyze their impacts on face image quality. [4] assesses the effect of variation in illumination on face image quality. [5] proposed a method which extracts some hand-craft features from face images with the purpose of assessing the impact of improper lightning and facial pose on face image quality. [6] uses vertical edge density as a quality metric that can estimate pose variation and improves the quality estimation of face images. [22] combines five quality factors (illumination, brightness, sharpness, contrast, and focus) to compute face quality index. The main problem with analytic-based face image quality assessment methods is that they consider just a limited number of features for assessing the image quality which leads to an inaccurate quality estimation.

The second category of face image quality assessment methods are learning-based approaches. These approaches are completely different from analytic-based approaches which are focused on directly measuring factors that affect the face image quality like illumination, pose, blurriness, and occlusion. However, the learning-based approaches try to make a relationship between the image quality and face recognition performance. To be more specific, these approaches correlate image quality of a sample image to the accuracy of a face recognition system when applied on that sample. [7] is the first learning-based method, where a multi-dimensional scaling approach is used to map space characterization features to its score. [8] proposed a method which first label images for quality by calculating the Euclidean distance between each query image with the best quality image of its respective subject, then trains a regression model on that labeled dataset to estimate the quality score of a desired query image. Furthermore, they extended their method [9] by taking the advantage of using four different face recognition systems for labeling the ground truth which makes it unbiased toward a specific face recognition system.

[23] uses a similarity distribution distance for face image quality assessment. This method utilize the Wasserstein distance between the intra-class similarity distribution and inter-class similarity distribution for generating quality labels. Then, it uses these labels to train a regression network for quality prediction. In [24] the authors proposed an unsupervised face quality assessment method base on a face recognition model that is trained with using dropouts. This method uses various subnetworks of a face recognition model which are generated by applying dropouts to measure the robustness of a sample image representation. This robustness of the sample representation is considered as the quality of that sample. [25] estimates the variance of element-wise embedding features and then uses the variance for face image quality. [1] calculates the face image quality score as the magnitude of the sample encoding. In this model, a face recognition system is trained with a loss that adapts the penalty margin loss based on this magnitude. In fact, it links the closeness of a sample to its class center to the magnitude of the embedding vector.

All of these learning-base face image quality assessment methods return a scalar value as the quality score which describes the overall quality of the face image. Although this scalar quality score can be used to improve the performance of face recognition systems, it does not give us any information about nuisance factors related to that image. In some applications like automatic enrollment for visa application or automatic identity authentication at border kiosks, it is useful to give some information about the factors that affect the image quality to the user. Knowing about these factors is beneficial for the user when the quality of the captured image is not good enough. In this situation, user can receive feedback from the system identifying which nuisance parameters have affected the quality of the image negatively. Therefore, the subject can resolve the problem and capture another image. The only work that gives some information about the nuisance factors related to images is [26] which uses 25 individual tests to check image compliance with ISO/IEC 19794-5 standard, but it is not efficient since these 25 tests are independent from each other and its accuracy for some critical factors which have a significant effect on the image quality like pose estimation is low.

3. Proposed Method

This section presents our face image quality vector assessment method. We propose a multi-task convolutional neural network for simultaneous face image quality assessment, head pose estimation, facial expression recognition, and illumination classification. Each component of the system will be discussed in detail in the following subsections. The architecture of the proposed method is presented in Figure 2.

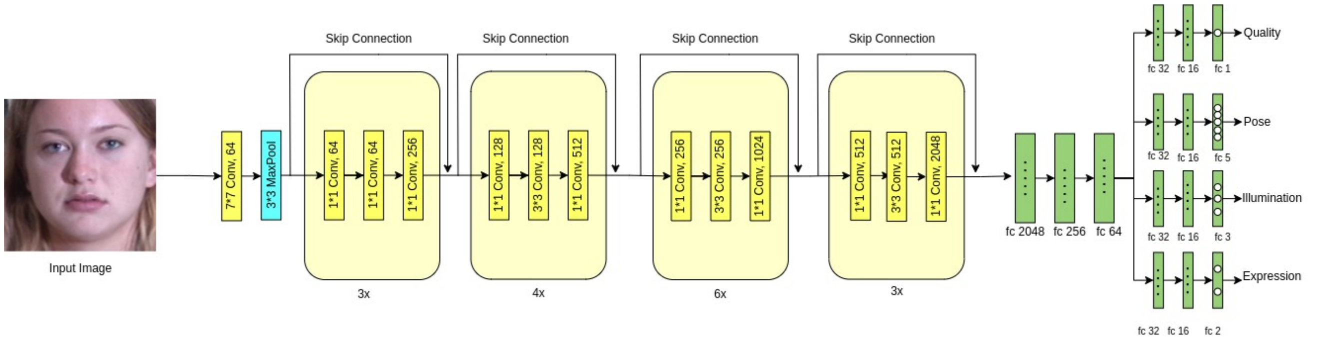


Figure 2: The architecture of the proposed method. A pretrained ResNet-50 model which its fully connected layers are eliminated is used as the network stem. The new fully connected layers are added to the stem and represented in color green. The output dimension of each fully connected layer is shown below it. Three fully connected layers are in common among branches and at the end of the third fully connected layer the network divides to four branches for quality estimation, pose estimation, illumination classification, and expression recognition. Each branch has three fully connected layers. The last fully connected layer of the quality estimation branch is a regression layer, so the output dimension of this layer is one. The last fully connected layer of other branches are classification layers, so the output dimension of them depends on the number of classes for that special task in data. For example, data has three classes for illumination(well-exposed, over-exposed, and under-exposed), so the output dimension of the last layer of this branch is three.

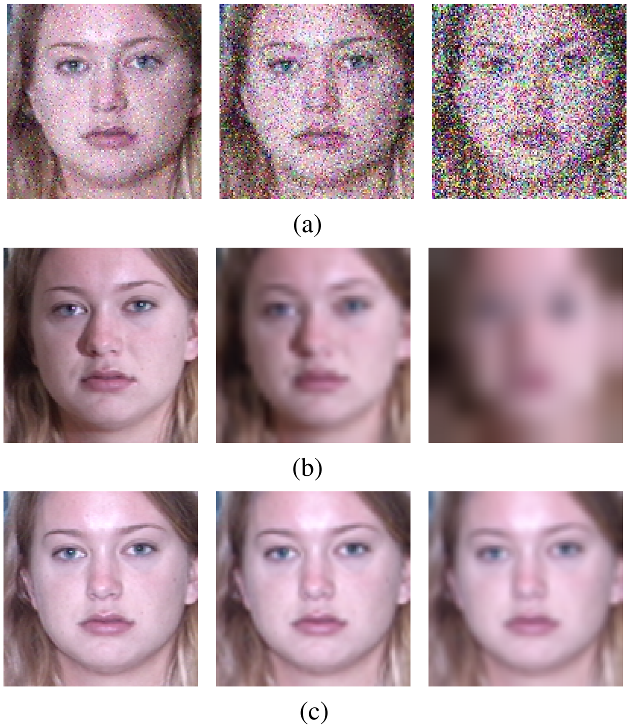
3.1. Architecture

Our proposed method utilize the ResNet-50 architecture [27] as its stem. We eliminate the last classification layer of the network and replace it with three fully connected layers which are shared among four heads of the network. The output of the last convolutional layer is a 100,352 dimensional feature vector. The output dimension of the first fully connected layer is 2,048 which is followed by another fully connected layer with 256 dimension, and the third fully connected layer output dimension is 64. At the end of those three fully connected layers, we split the network to four separate branches corresponding to different tasks. Each branch is dedicated for a specific task and has three fully connected layers. The first branch is the quality estimator branch that consists of three fully connected layers. The output dimension of the first two fully connected layers are 32. The third fully connected layer is a regression layer; thus the output dimension of this layer is one. The second branch is the pose estimator branch. Similar to the quality estimator branch it has three fully connected layers. The output dimension of two first fully connected layers are 32 and 16, respectively. The third fully connected layer is a classifier. As the number of pose classes is five, the output dimension of this layer is five. The third branch is the expression recognition branch. Same as other branches we use three fully connected layers for this branch where the output dimension of the first two layers are 32, and 16 respectively. The third layer is the classifier which should classify expression. As for our application, we classified expressions to neutral and non-neutral classes, the output dimension of this layer is two. The last branch is the illumination classification class. Like other branches it has three fully connected layers with the output dimension of 32, and 16 for the first two layers. The third layer is the illumination classification layer. We classified illumination intensity to three different classes well-exposed, over-exposed, and under-exposed. Hence, the output dimension of the illumination classification layer is three. In the proposed architecture after every fully connected layer, we deploy the Rectified Linear Unit (ReLU) activation function [28].

3.2. Choosing Datasets and Data Augmentation

Although a lot of face image datasets are available, the number of face image datasets which are tagged for illumination, expression, and pose are limited. After surveying different datasets, we selected the CMU Multi-PIE face dataset [29] as our training dataset. This dataset contains more than 750K images from 337 subjects recorded in four different sessions over the span of five months. Images are captured under 19 illumination conditions and 15 different view points while displaying a range of facial expressions.

The CMU Multi-PIE dataset has about 750k images; however, our model has more than 200 million trainable parameters; Consequently more images are needed for training our model. Additionally, all of the CMU Multi-PIE images are high quality images without any distortions (i.e., noise, blurriness and low-resolution). To make our model robust to low quality images, we need to integrate these pertubation into our training set. Therefore, we used data augmentation techniques to not only increase the number of


(a)

(b)

(c)

Figure 3: Augmented images. Figure 3a shows augmented images which are generated by adding Gaussian noise to the original image at three different levels. Figure 3b shows augmented images which are generated by applying Gaussian blur filter on the original image using three different window sizes. Figure 3c shows augmented images which are generated by downsampling and then upsampling the original image with the goal of generating the low-resolution images at three different levels.

our training data samples, but also make our model more robust to noisy, blurry, and low-resolution data.

For generating noisy images, Gaussian noise is added to the original images at three different levels of noise. We utilize 50, 100, and 150 respectively for the standard deviation and zero for the mean to generate three different levels of noisy images. To generate low-resolution images, we use downsampling and upsampling techniques. Likewise adding noise, we generate the low-resolution images at three different levels to make our training dataset more diverse, and our model more robust. Also, for generating blurry images we apply Gaussian blur filter with 3, 5, and 7 as the window size to blur images at three different blurry levels. Therefore, for each image in the original dataset nine more images are generated, and added to the dataset. Some examples of noisy, low resolution and blurry augmented images are shown in Figure 3.

3.3. Creating the Ground truth Face Quality

As it is mentioned before, the training dataset need to be labeled for quality, illumination, pose, and expression, but none of the face images in the dataset have quality labels. The question which arises is that how can we assign quality label to each face image. Quality is a relative concept; thus a face image can be defined as a low-quality image when it is compared to a high-quality one. In biometrics applications some criteria are defined for a high quality face image in the ISO/IEC 19794-5 standard [21]. Base on this standard a high-quality face image is a frontal, well-exposed with neutral expression face image.

To label face images for each subject, we select the image with the highest quality of that subject and call it gallery image. We assign 1.0 as the quality score to the galley image. By considering the ISO/IEC 19794-5 standard criteria and the CMU Multi-PIE dataset conditions, for each subject we choose the image which is captured in zero degree (frontal) with the highest level of illumination and neutral expression as the gallery image. Then, we score other images for the same subject (that we call them probe images) based on their relative difference with the gallery image. As the gallery image for each subject is captured at the same condition (with the same camera, at the same pose, and with the same illumination condition), it is reasonable that all of the galley images have the same quality score. Furthermore, since the quality score for probe images of each subject is computed based on its differences with the gallery image, the quality score for probe images from different subjects would also be consistent.

To be more specific, we use the FaceNet model [2], a CNN which is pretrained for face recognition, as a feature extractor to get embedding for all of the images in the database. Afterwards, we take the advantage of this embedding for calculating the quality score. For this purpose, first we extract a 128-dimensional feature vector from the last fully-connected layer of FaceNet. Given this embedding, we compute the distance between the gallery image and each probe image. Then, we normalize this similarity value to the [0,1] range and use it as the quality score for the probe image. The value close to 1 represents a high-quality image, and the value close to 0 represents the low-quality one. Figure 3 demonstrates the process of creating the ground truth which is explained in this section.

3.4. Training the Network

For the training phase, the weights of a pretrained ResNet-50 on VGGFace2 [30] dataset is used as the stem of our network. During the training phase, we freeze all of the parameters of the convolutional layers and only train the parameters of the fully connected layers. Although we only train the fully connected layers, our network still has more than 200 million parameters. Therefore, a huge dataset is

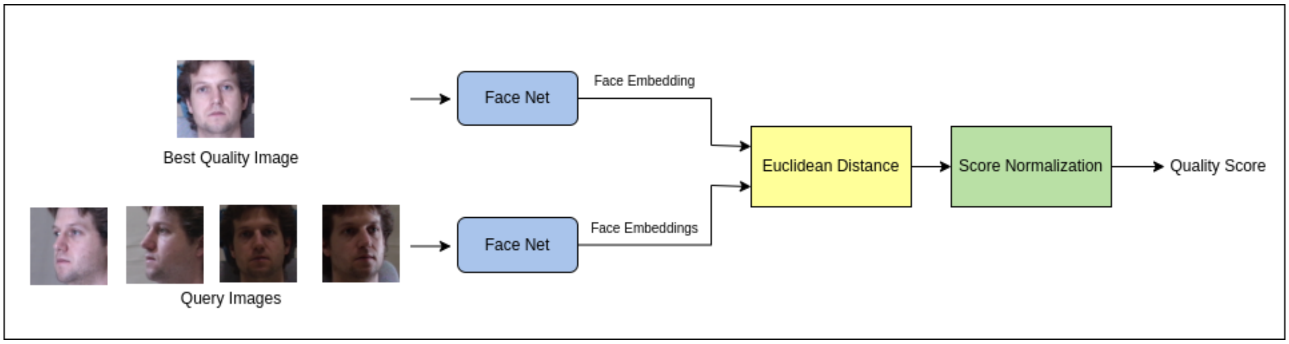


Figure 4: Shows the process of generating the ground truth quality for face images in the training dataset. For each subject in the dataset the image with frontal pose, neutral expression and highest illumination is selected as the best quality image for that subject. This selection is done based on the pose, illumination and expression labels of the dataset. The best quality image of the subject and other images of that subject are fed to FaceNet and the embedding vector for each face image is calculated. The Euclidean distances between the embedding vector of the best quality image of the subject and embedding vectors of other images of the same subject are calculated and normalized in the range of zero to one. This normalized value is assigned to each image as its quality score.

needed for training our network. Also, the training dataset need to have quality, pose, expression, and illumination labels.

Transfer learning has shown very good performance in deep learning. Fine-tuning deep learning models, training the model for a specific task with huge amount of data and then retraining it for a different but closely related tasks with limited amount of data, has been successfully tested in face related problems where the network is initially trained for recognizing the identity and then has been used for other attributes related to face like age, gender [31], emotion [32], and race [20]. Since the accuracy of a face recognition system is closely related to the quality of the face images and the face quality itself is related to factors like illumination, facial expression, and head pose, it is expected that the feature vector which contains discriminative information of faces, also contains information of their quality, pose, illumination and expression. Moreover, the amount of face quality training data is limited, but huge amount of training data for face recognition is available. Therefore, we take the advantage of transfer learning and, use a pretrained ResNet-50 model which is already trained on the VGGFace2 dataset. In order to use the pretrained ResNet-50 model which is trained for face recognition for quality estimation, we need to extract quality related information from the face embedding and this process is done by fine-tuning the model by using the ground truth.

The predicted quality score represents the information about the nuisance factors in the image, and using the same model which is already trained for quality estimation would increase the accuracy for predicting the nuisance factors. First, we only train the quality branch, after training the quality branch we add three other branches and train them simultaneously. Training other branches on a stem which is already trained for quality helps us to take the advantage of using middle layers' feature vector which contains quality information for training other branches. This process is similar to using transfer learning which increases the performance and accuracy in comparison to training the network from the scratch.

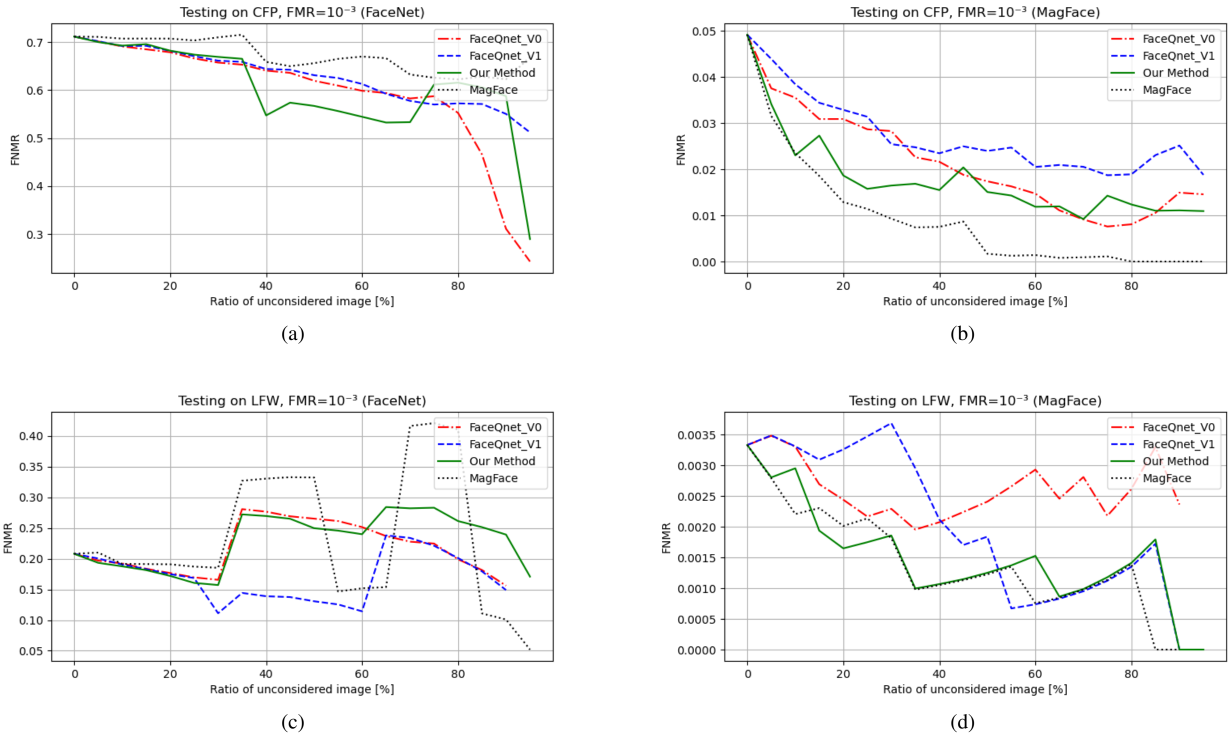
For the training process, we freeze all the weights of the pretrained ResNet-50 stem (convolutional layers parameters) and just train the fully connected layer parameters for the quality branch. The cost function J_q that we used for the quality estimation task is the Mean Squared Error (MSE) given as $J_q = \frac{1}{N}[\sum_{i=1}^{N}(\hat{Y} - Y)^2]$, where N is the number of sample images, $\hat{Y}$ and Y are the predicted and ground truth quality, respectively.

The equation for obtaining network optimal parameter values is shown as

$$(\theta_s^*, \theta_q^*) = \arg \min_{(\theta_s, \theta_q)} J_q(\theta_s, \theta_q; I), \tag{1}$$

where I is the input data, θ_s and θ_q are the shared and task-specific (quality estimation) parameters, respectively. Also, θ_s^* and θ_q^* denotes the optimal network parameters.

After training the quality branch, we add three other branches and train four branches simultaneously. The reason that we continue to train quality branch while training other branches is that training other branches would change the middle layers weights which are shared among all of the branches and this can affect the quality branch negatively; Hence, the quality branch parameters should not be frozen during training of other branches to be adjusted proportionally. The cost function that we use for pose estimation, and



(a)

(b)

(c)

(d)

Figure 5: Face verification performance on the predicted face image quality scores. The curves show the effectiveness of rejecting low-quality face images on the verification error in terms of FNMR at a threshold of FMR= 10^{-3} . Figure 5a and 5b show the results for the FaceNet [2] and MagFace [1] (MagFace is both a face recognition and a FIQA method) embeddings on the CFP dataset, respectively. Figure 5c and 5d show the the same on the LFW dataset.

illumination classification tasks is multi-class cross entropy cost function. This cost function for pose J_p is shown as

$$J_p = \sum_{c=0}^{C} -y_c.log(p_c). \tag{2}$$

In the above equation C illustrates the number of classes. $y_c = 1$ if the sample belongs to class c, otherwise 0. Additionally, p_c shows the probability that a sample belongs to class c. The cost function for illumination classification task is the same just number of classes is different.

Also, we use the binary cross entropy cost function for the expression recognition task in that it has two classes. This cost function for expression J_e is shown as

$$J_e = -(1 - y_e).log(1 - p_e) - y_e.log(p_e). \tag{3}$$

In the above equation y_e=1 for neutral expression and 0 otherwise. p_e shows the probability that the facial expression is neutral.

The overall cost function is the weighted sum of the individual cost functions. Therefore, to obtain network optimal parameter values for the desired task t_i in this multi-task convolutional neural network the weighted sum of cost functions for all tasks should be minimized. The equation for obtaining the optimal network parameters is shown as

$$(\theta_s^*, \theta_{t_i}^*) = \arg \min_{(\theta_s, \theta_{t_i})} \sum_{i}^{n} \lambda_i J_i(\theta_s, \theta_{t_i}; I), \tag{4}$$

where θ_s and θ_{t_i} are the shared and task-specific parameters, respectively. J_i denotes the cost function for the task t_i, I is the input data, n is the number of tasks, and λ_i is the weight associated with each task's cost function. Also, θ_s^* and $\theta_{t_i}^*$ represent the optimal network parameters.

4. Experiments

The proposed multi-task network is evaluated for each of the four different tasks on which it was trained separately.

4.1. Quality Estimation

Evaluation Method: In this paper, we use the Error versus Reject Curve (ERC) which is the most widely accepted metric for evaluating performance of face image quality assessment algorithms [33] and was adapted in the approved ISO working item [34]. This metric has been used widely

[35] [36]. The ERC demonstrates effect of discarding a fraction of face images with the lowest quality on face verification performance which is denoted by False Not Match Rate (FNMR). In fact, an ERC curve shows the relationship between FNMR and reject rates. It explains how FNMR changes when the increasing amount of data with the lowest quality is being discarded. Therefore, for plotting ERC the FNMR value should be calculated while an increasing amount of data with the lowest quality is being discarded. In the ERC curve for an ideal face quality assessment algorithm FNMR should decrease consistently with increasing amount of discarded low-quality images. Plotting ERC curves is a fair method to compare performance of different face image quality assessment algorithms because it is independent of the absolute and range of the quality score values.

For plotting the ERC curve a face recognition system and a face dataset with subject identity labels is required. To compare face image quality assessment algorithms on different face recognition systems, the effect of face recognition system on the results should be eliminated. To this end, FNMR is computed at a fixed False Match Rate (FMR). In this paper, we calculate FNMR at FMR=10^{-3} which is recommended for border control by Fronex [37].

Datasets: For plotting the ERC curve the FNMR values need to be computed. Therefore, the dataset which is used should contain pair images. Here we perform experiments on two publicly available datasets which contain verification pair images for assessing performance of our method and comparing it with other face image quality assessment algorithms. The purpose of using two different datasets for evaluating and comparing performance of different FIQA algorithms is to have variations in quality of images and show generalization of our approach on multiple datasets. The datasets that are used in this experiment are introduced below.

The Labeled Faces in the Wild (LFW) [38] is an unconstrained face verification dataset which contains 13,233 face images of 5,749 identities collected from the web. The Celebrities in Frontal-Profile in the Wild (CFP-FP) [39] is a dataset for comparison between frontal and profile faces. This dataset contains 7,000 images from 500 identities. For each identity 10 frontal and 4 profile images exists in the dataset.

Face Recognition Systems: As it was mentioned previously for plotting the ERC curve, a face recognition system is required to calculate the FNMR values. Here we report the verification performance at different quality rejection rates over two state-of-the-art face recognition systems to show generalizability of face image quality assessment over different face recognition systems. The face recognition systems that are used in this study are the MagFace [1] and FaceNet [2].

The ERC results are shown in Fig. 5. As we can see, our method generally outperforms other state-of-the-art methods specially for the ratio of unconsidered image in the range of 0 to 20 percent. The reason that we emphasize on comparing the performance of different algorithms within the range of 0 to 20 percent of unconsidered images is that in most of the real world scenarios only the very low-quality images which may result in failure in face recognition systems are discarded. These very low-quality images are results of extreme illumination and pose condition, low resolution, blurriness, and occlusion. Therefore, to consider real world scenarios we compare the performance of different methods for the ratio of unconsidered images in the range of 0 to 20 percent.

4.2. Pose Estimation, Facial Expression Recognition, and Illumination Classification

For testing our algorithm, we selected 50 different identities from the CMU Multi-PIE face dataset [29] which are completely separate from the 200 identities which were used for the training phase. For the pose estimation task, we obtained 92.71% accuracy on the test dataset. For the facial expression recognition task we reached to 93.15 % accuracy. Finally, for the illumination classification task we reached to 87.35 % accuracy.

5. Conclusion

In this work, we proposed a multi-task convolutional neural network which returns a quality vector for an input probe face image. This quality vector contains a scalar quality score and information about the nuisance factors like pose, illumination, and expression. This information is useful for some applications like automatic enrollment and identity authentication at border kiosks. For this purpose, first we created a huge and diverse dataset by using data augmentation techniques, and use it to create a ground truth. Then, we trained our multi-task neural network on that ground truth. We evaluated our proposed method across two different face recognition systems and on two different datasets to assess its generalizablity. The results indicated our method obtained higher or comparable accuracy in comparison with the state-of-the-art methods for the task of quality estimation. Moreover, our method reached to near 90 % accuracy for pose estimation, facial expression recognition, and illumination classification tasks. As for future improvements, we are going to add more features like occlusion and sun glass detection to our quality vector estimator.

References

[1] Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. Magface: A universal representation for face recognition and

quality assessment. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14220–14229, 2021.

[2] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015.

[3] Mohammad Saeed Ebrahimi Saadabadi, Sahar Rahimi Malakshan, Sobhan Soleymani, Moktari Mostofa, and Nasser M. Nasrabadi. Information maximization for extreme pose face recognition. *arXiv preprint arXiv:2209.03456*, 2022.

[4] J. Ross Beveridge, David S. Bolme, Bruce A. Draper, Geof H. Givens, Yui Man Lui, and P. Jonathon Phillips. Quantifying how lighting and focus affect face recognition performance. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 74–81, 2010.

[5] Xiufeng Gao, Stan Z. Li, Rong Liu, and Peiren Zhang. Standardization of face image sample quality. In Seong-Whan Lee and Stan Z. Li, editors, *Advances in Biometrics*, pages 242–251, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.

[6] Pankaj Shivdayal Wasnik, K. Bommanna Raja, Ramachandra Raghavendra, and Christoph Busch. Assessing face image quality for smartphone based face recognition system. *2017 5th International Workshop on Biometrics and Forensics (IWBF)*, pages 1–6, 2017.

[7] Gaurav Aggarwal, Soma Biswas, Patrick J. Flynn, and Kevin W. Bowyer. Predicting performance of face recognition systems: An image characterization approach. In *2011 Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 52–59, 2011.

[8] J. Hernandez-Ortega, Javier Galbally, Julian Fierrez, Rudolf Haraksim, and Laurent Beslay. Faceqnet: Quality assessment for face recognition based on deep learning. *2019 International Conference on Biometrics (ICB)*, pages 1–8, 2019.

[9] J. Hernandez-Ortega, Javier Galbally, Julian Fierrez, and Laurent Beslay. Biometric quality: Review and application to face recognition with faceqnet. *ArXiv*, abs/2006.03298, 2020.

[10] Gani Rahmon, Imad Eddine Toubal, and Kannappan Palaniappan. Extending u-net network for improved nuclei instance segmentation accuracy in histopathology images. In *2021 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pages 1–7, 2021.

[11] Amirhossein Rasoulian, Soorena Salari, and Yiming Xiao. Weakly supervised intracranial hemorrhage segmentation using hierarchical combination of attention maps from a swin transformer. In *Machine Learning in Clinical Neuroimaging*, pages 63–72, 2022.

[12] Hojat Asgarian Dehkordi, Ali Soltani Nezhad, Hossein Kashiani, Shahriar Baradaran Shokouhi, and Ahmad Ayatollahi. Multi-expert human action recognition with hierarchical super-class learning. *Knowledge-Based Systems*, 250:109091, 2022.

[13] Moktari Mostofa, Mohammad Saeed Ebrahimi Saadabadi, Sahar Rahimi Malakshan, and Nasser M. Nasrabadi. Pose attention-guided profile-to-frontal face recognition. *arXiv preprint arXiv:2209.07001*, 2022.

[14] Pedro C. Neto, Fadi Boutros, Joao Ribeiro Pinto, Mohsen Saffari, Naser Damer, Ana F. Sequeira, and Jaime S. Cardoso. My eyes are up here: Promoting focus on uncovered regions in masked face recognition. In *2021 International Conference of the Biometrics Special Interest Group (BIOSIG)*, 2021.

[15] Hossein Kashiani, Shoaib Meraj Sami, Sobhan Soleymani, and Nasser M Nasrabadi. Robust ensemble morph detection with domain generalization. *arXiv preprint arXiv:2209.08130*, 2022.

[16] Rich Caruana. *Multitask Learning*, pages 95–133. Springer US, Boston, MA, 1998.

[17] Xiangxin Zhu and Deva Ramanan. Face detection, pose estimation, and landmark localization in the wild. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2879–2886, 2012.

[18] Rajeev Ranjan, Vishal M. Patel, and Rama Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41:121–135, 2019.

[19] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In *2014 European Conference on Computer Vision (ECCV)*, pages 94–108, 2014.

[20] Rajeev Ranjan, Swami Sankaranarayanan, Carlos D. Castillo, and Rama Chellappa. An all-in-one convolutional neural network for face analysis. In *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, pages 17–24, 2017.

[21] *ISO/IEC 19794-5: 2011 Information technology — Biometric data interchange formats — Part 5: Face image data*. Standard.

[22] Ayman A. Abaza, Mary Ann F. Harrison, and Thirimachos Bourlai. Quality metrics for practical face recognition. *Proceedings of the 21st International Conference on Pattern Recognition (ICPR)*, pages 3103–3107, 2012.

[23] Fu-Zhao Ou, Xingyu Chen, Ruixin Zhang, Yuge Huang, Shaoxin Li, Jilin Li, Yong Li, Liujuan Cao, and Yuan-Gen Wang. SDD-FIQA: Unsupervised face image quality assessment with similarity distribution distance. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7666–7675, 2021.

[24] Philipp Terhorst, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. SER-FIQ: Unsupervised estimation of face image quality based on stochastic embedding robustness. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5650–5659, 2020.

[25] Yichun Shi and Anil Jain. Probabilistic face embeddings. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6901–6910, 2019.

[26] Javier Hernandez-Ortega, Julian Fierrez, Luis F. Gomez, Aythami Morales, Jose Luis Gonzalez-de Suso, and Francisco Zamora-Martinez. Faceqvec: Vector quality assessment for face biometrics based on iso compliance. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pages 84–92, 2022.

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[28] Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning (ICML)*, page 807–814. Omnipress, 2010.

[29] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-PIE. In *2008 8th IEEE International Conference on Automatic Face Gesture Recognition (FG)*, pages 1–8, 2008.

[30] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. VGGFace2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG)*, pages 67–74, 2018.

[31] Aythami Morales, Julian Fierrez, Ruben Vera-Rodriguez, and Ruben Tolosana. Sensitivenets: Learning agnostic representations with application to face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(6):2158–2164, 2021.

[32] Alejandro Peña, Julian Fierrez, Aythami Morales, and Agata Lapedriza. Learning emotional-blinded face representations. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 3566–3573, 2021.

[33] Patrick Grother and Elham Tabassi. Performance of biometric quality measures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(4):531–543, 2007.

[34] *ISO/IEC WD 24357: Performance evaluation of face image quality algorithms*. Standard.

[35] Patrick Grother, Mei Ngan, and Kayee Hanaoka. *Face recognition vendor test - face recognition quality assessment concept and goals.* 2019.

[36] Elham Tabassi and Patrick Grother. *Biometric Sample Quality*, pages 194–206. Springer US, Boston, MA, 2015.

[37] Frontex. *Best practice technical guidelines for automated border control (abc) systems*. 2015.

[38] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[39] C.D. Castillo V.M. Patel R. Chellappa D.W. Jacobs S. Sengupta, J.C. Cheng. Frontal to profile face verification in the wild. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9, 2016.