

2019

Evaluation and Understandability of Face Image Quality Assessment

Mohammad I. Nouyed
West Virginia University, monouyed@mix.wvu.edu

Follow this and additional works at: <https://researchrepository.wvu.edu/etd>



Part of the [Artificial Intelligence and Robotics Commons](#), [Information Security Commons](#), and the [Other Computer Sciences Commons](#)

Recommended Citation

Nouyed, Mohammad I., "Evaluation and Understandability of Face Image Quality Assessment" (2019). *Graduate Theses, Dissertations, and Problem Reports*. 7422.
<https://researchrepository.wvu.edu/etd/7422>

This Thesis is protected by copyright and/or related rights. It has been brought to you by the The Research Repository @ WVU with permission from the rights-holder(s). You are free to use this Thesis in any way that is permitted by the copyright and related rights legislation that applies to your use. For other uses you must obtain permission from the rights-holder(s) directly, unless additional rights are indicated by a Creative Commons license in the record and/ or on the work itself. This Thesis has been accepted for inclusion in WVU Graduate Theses, Dissertations, and Problem Reports collection by an authorized administrator of The Research Repository @ WVU. For more information, please contact researchrepository@mail.wvu.edu.

2019

Evaluation and Understandability of Face Image Quality Assessment

Mohammad I. Nouyed

Follow this and additional works at: <https://researchrepository.wvu.edu/etd>



Part of the [Artificial Intelligence and Robotics Commons](#), [Information Security Commons](#), and the [Other Computer Sciences Commons](#)

Evaluation and Understandability of Face Image Quality Assessment

Mohammad Iqbal Nouyed

Thesis submitted to the
Benjamin M. Statler College of Engineering and Mineral Resources
at West Virginia University

in partial fulfillment of the requirements for the degree of

Masters in
Computer Science

Guodong Guo, Ph.D., Chair
Donald Adjero, Ph.D.,
Hong-Jian Lai, Ph.D.

Lane Department of Computer Science and Electrical Engineering

Morgantown, West Virginia
2019

Keywords: Face image quality, biometric quality, face recognition, face biometrics, FIQA
© 2019 Mohammad Iqbal Nouyed

ABSTRACT

Evaluation and Understandability of Face Image Quality Assessment.

Mohammad Iqbal Nouyed

Face image quality assessment (FIQA) has been an area of interest to researchers as a way to improve the face recognition accuracy. By filtering out the low quality images we can reduce various difficulties faced in unconstrained face recognition, such as, failure in face or facial landmark detection or low presence of useful facial information. In last decade or so, researchers have proposed different methods to assess the face image quality, spanning from fusion of quality measures to using learning based methods. Different approaches have their own strength and weaknesses. But, it is hard to perform a comparative assessment of these methods without a database containing wide variety of face quality, a suitable training protocol that can efficiently utilize this large-scale dataset. In this thesis we focus on developing an evaluation platform using a large scale face database containing wide ranging face image quality and try to deconstruct the reason behind the predicted scores of learning based face image quality assessment methods. Contributions of this thesis is two-fold.

Firstly, (i) a carefully crafted large scale database dedicated entirely to face image quality assessment has been proposed; (ii) a learning to rank based large-scale training protocol is developed. Finally, (iii) a comprehensive study of 15 face image quality assessment methods using 12 different feature types, and relative ranking based label generation schemes, is performed. Evaluation results show various insights about the assessment methods which indicate the significance of the proposed database and the training protocol.

Secondly, we have seen that in last few years, researchers have tried various learning based approaches to assess the face image quality. Most of these methods offer either a quality bin or a score summary as a measure of the biometric quality of the face image. But, to the best of our knowledge, so far there has not been any investigation on what are the explainable reasons behind the predicted scores. In this thesis, we propose a method to provide a clear and concise understanding of the predicted quality score of a learning based face image quality assessment. It is believed that this approach can be integrated into the FBI's understandable template and can help in improving the image acquisition process by providing information on what quality factors need to be addressed.

Acknowledgements

It is a pleasure for me to express my sincere gratitude to Dr. Guodong Guo, for giving me the opportunity to work under his supervision, for his patience and guidance. I would also like to thank my academic examination committee members Dr. Donald Adjero, and Dr. Hong-Jian Lai for agreeing to evaluate my thesis and providing valuable feedback on how to improve it further. I would also like to thank all of my course instructors here at WVU and my colleagues at WVUCVL, past and present, whose varied contribution, big and small, have ultimately lead me to complete this arduous and challenging work.

Last but not least, I would like to thank my parents for their silent encouragement and support throughout my studies.

Contents

Acknowledgements	iii
List of Figures	vi
List of Tables	vii
1 Introduction	1
1.1 Summary of contributions	2
1.2 Organization of Thesis	2
2 Evaluation of Face Image Quality Assessment Methods on a Large Dataset	3
2.1 Introduction	3
2.2 Related Work	4
2.3 Face Image Quality Database	8
2.4 Quality Labels	9
2.4.1 Relevance levels from identification rank	10
2.5 Quality Representation	12
2.5.1 Heuristics (HR) based features	12
2.5.2 Weighted sum of heuristic measures (WSHRM)	13
2.5.3 Face recognition (FR) features	13
2.5.4 Natural Scene Statistics (NSS) features	15
2.5.5 Other features	19

2.6	Evaluation results	19
2.6.1	EvR based evaluation: Observations and findings	21
2.6.2	Bin vs Recognition rate: Evaluation measures	23
2.6.3	Bin vs Recognition rate: Observations and findings	24
2.6.4	Observation from inspecting sample face images	24
2.6.5	Discussion	26
2.7	Conclusion	27
3	Understandable Face Image Quality	28
3.1	Introduction	28
3.2	Related Works	29
3.3	The understandable face image quality method	29
3.3.1	What is Understandable face image quality ?	29
3.3.2	Database description	31
3.3.3	The quality measures	33
3.3.4	The statistical approach	36
3.4	Experiments	38
3.4.1	Limitations	40
3.5	Conclusions	40
4	Conclusion & Future work	42
	Bibliography	43

List of Figures

2.1	Quadrature face match graph showing cosine similarity scores between arrow points and quality scores above/below face image. Cropped faces shown inset.	9
2.2	Error vs. Reject curves of the all face image quality assessment methods used in our experiment. X-axis indicates the percetage of low quality images rejected. Y-axis indicates FNMR@FMR=0.01%	20
2.3	Demonstrating the effect of different types of biometric quality degradation on the predicted scores. Cropped faces are shown inset. Quality scores are generated by (FR+HR)+SVR method	22
2.4	Correlation plot of all face image quality assessment methods investigated in this study, grouped by feature type. X-axis indicates the bin number (quality level). Y-axis indicates the rank-100 identification rate for each bin (quality level).	25
3.1	Example of obtaining understandable information for a test image.	30
3.2	An example face image consisting of composite biometric quality. Which quality measure(s) has ultimately caused the quality score to be 5?	31
3.3	Sample images from the database with predicted scores, shown in the top-right corner, demonstrating the variety of face image qualities.	33
3.4	Face image with 68 face landmarks detected using Openface toolkit.	35
3.5	We can use are under the gaussian distribution for a quality bin to obtain the probability of any quality measure belong to that disribution.	36
3.6	Histogram for each quality bins for sharpness measure and head pose yaw measure (euler_y). bin=1 has the lowest quality and bin=10 is the bin with highest quality images.	37
3.7	Mean vs Standard deviation plot for contrast and focus measure. We can observe the noisy distriubtions are not following the order of the quality level.	39

3.8	Sample images with understandable information regarding quality. At right side of the images we have: 1) the predicted quality score 2) quality measure having non-optimal value 3) Probability of the value belonging to bin-N 4) Bin number with the max probability 5) understandable label category	41
-----	---	----

List of Tables

2.1	A brief overview of FIQA methods. (<i>Details of the databases can be found in [1]</i>) . .	6
2.1	A brief overview of FIQA methods. (<i>Details of the databases can be found in [1]</i>) (<i>continued.</i>)	7
2.1	A brief overview of FIQA methods. (<i>Details of the databases can be found in [1]</i>) (<i>continued.</i>)	8
2.2	Results for FIQA methods.	26

Chapter 1

Introduction

It has been well agreed upon by researchers [2–4] that, biometric sample quality is defined as a measure of a samples utility to automatic matching. It has been referred as an intrinsic physical data content. National Institute of Standards and Technology (NIST) defined biometric quality scores as - “the accuracy with which physical characteristics are represented in a given biometric data” [5,6]. Note that, the term quality here, is not just limited to image size, resolution, dimension, color depth and any acquisition parameters. We know that, in ideal scenario, the image acquisition process should produce high quality images that are ideal for feature extraction and matching in later steps. But in real-world cases, live samples are obtained under unconstrained environments which necessitates preprocessing the non-ideal real world image to either enhance quality and remove different types noises present in the image. For example, face images obtained in non-ideal conditions can have blurriness, large pose variations and poor illumination. Therefore, quality of the biometric data, such as face images, is a very important factor in ensuring robustness of any security system. By using sophisticated sensor technologies and by applying advanced noise removal and image enhancement methods the condition of the low quality face images can be improved. Thus, assessment of quality (and enhancement afterwards if necessary) is crucial in any biometric modality, especially in face biometrics. But, till today, face image quality assessment is a non-trivial problem because of a multitude of behavioral and extraneous conditions that can simultaneously affect the face appearance. Moreover, there is a major difference between quality assessment of face images and the traditional image and video quality assessment in terms of its multiple objectives, e.g., to ensure its fidelity to the human visual system (HVS), generating feedback to image acquisition system, predict face matching performance, securing the image enrollment process and in case of multimodal biometrics, provided a weight for deciding merger.

Face image quality assessment (FIQA) has been an area of interest to researchers as a way to improve the face recognition accuracy. By filtering out the low quality images we can reduce various difficulties faced in unconstrained face recognition, such as, failure in face or facial landmark detection or low presence of useful facial information. In last decade or so, researchers have proposed different methods to assess the face image quality, spanning from fusion of quality measures to using learning based methods. Different approaches have their own strength and weaknesses. But, it is hard to perform a comparative assessment of these methods without a database containing wide variety of face quality, a suitable training protocol that can efficiently utilize this large-scale dataset. In the first part of this thesis, (i) a carefully crafted large scale

database dedicated entirely to face image quality assessment has been proposed; (ii) a learning to rank based large-scale training protocol is developed. Finally, (iii) a comparative study of representative face image quality assessment methods is conducted using 12 different feature types is performed. Evaluation results show various insights about the assessment methods which indicate the significance of the proposed database and the training protocol.

Recent studies (including our work in the first part of this thesis) have shown that a learning-based paradigm can do better than the traditional heuristic methods [7]. But, these methods usually provides a quality score summary or a bin label. This single value prediction does not provide use much information to understand what are the underlying factors regarding the quality change. Since it has been found that, biometric quality of face image is defined in terms of automatic face recognition performance, human visual perception of image quality may not be well correlated with recognition performance [2, 3, 7]. Therefore, intuitive judgement by visually inspecting the predicted face image is not enough to understand how the learning based method has learned to label quality itself. In the second part of this thesis, we propose a novel method, which provides human understandable information for face image quality assessment, which can help address the issues in quality assessment of face images. We believe that this novel approach can give a better understanding about the characteristics of learning based quality assessment method in consideration.

1.1 Summary of contributions

- Developing of a large scale quality database containing a wide range of face image qualities.
- Establishing a learning to rank based face image quality label generation method for the large scale data set.
- A comparative evaluation of a 15 different learning based representative FIQA methods along with a traditional FIQA approach, which use 12 different feature types for quality assessment.
- Define the understandable face image quality (UFIQ) paradigm, and how a mapping from score summary to heuristic attributes can provide understanding regarding quality change.
- Establish the understandable face quality method using the help of statistical measures.
- Provide experimental evaluation of understandable face image quality.

1.2 Organization of Thesis

The thesis has been divided into two different parts based on the handled problems. In chapter 2, we present the comparative assessment of face image quality assessment method using a large database. We describe the organization of the large scale face database focused on face image quality, training protocol development and comparative analysis of different representative face image quality assessment methods. In chapter 3, we describe the problem of understandable face image quality assessment, how we approach to solve it, the statistical method to understandability using heuristic attributes and finally experimental evaluation and discussions are provided. Finally, we conclude and mention plans regarding future works in Chapter 4.

Chapter 2

Evaluation of Face Image Quality Assessment Methods on a Large Dataset

2.1 Introduction

In the field of biometrics research, the term biometric quality is defined as a measure of a sample’s utility to automatic matching [2–4]. National Institute of Standards and Technology (NIST) has defined biometric quality score as the accuracy with which physical characteristics are represented in a given biometric data [5, 6]. The quality of biometric data is an important issue to ensure the robustness of the biometric system. In various real world scenarios, samples may be collected in non-ideal conditions, which needs a quality assessment before recognition. Estimation of the quality of face images is a non-trivial problem. The reason is, there can be a large number of behavioral and extraneous conditions that can impact on the face appearance in images, such as, different facial variations, image acquisition devices, and environmental conditions. The quality of the input samples can vary from one system to another. Moreover, face image quality assessment is different from the traditional image or video quality assessment, because it also has to ensure several other criteria as well, e.g., it has to provide a reliable prediction on face matching performance, generate feedback on the quality of the image acquisition, contribute to the robustness of the face registration process, and provide a weight for each modality, in the case of multimodal biometrics. Researchers have worked on developing different biometric quality assessment methods for face images, however, all these methods may have either used a single, or a combination of few public databases, privately collected face images, and surveillance videos for training and testing (see Table 2.1). Even the recently proposed, deep learning based FIQA methods have used relatively small datasets, e.g., [8] used video frames of 37,213 images of 93 subjects, [9] used ChokePoint dataset of 64,204 images of 25 subjects, and [10] used combination of Color FERET and Kinect Face database containing a total of 11,338 of 994 subjects to train their deep models. So far, there is no large scale database specifically built for benchmarking face image quality assessment methods, to the best of our knowledge. Also, lack of a reliable quality label generation strategy pose further problem for such large databases. In this chapter, a large-scale database named “Face Image Quality Database (FIQDB)” is presented, containing more than 500K images of more than

14K subjects. To demonstrate its usefulness, a comprehensive evaluation is conducted on a set of representative face image quality assessment methods with specially selected feature types. Our main contributions include:

- Developing of a large scale quality database containing a wide range of face image qualities.
- Establishing a learning to rank based face image quality label generation method for the large scale data set.
- A comparative evaluation of a 15 different learning based representative FIQA methods along with a traditional FIQA approach, which use 12 different feature types for quality assessment.

2.2 Related Work

There have been various approaches to assess the face image quality [11]. Table 2.1 provides a brief review of face image quality assessment approaches developed in last decades or so. A traditional approach to assess face image quality is to use various facial attributes measures, and fuse these values as quality score. Hsu et al. [12] presented a quality assessment framework that employs a classification based score normalization process for various quality metrics and techniques to fuse those quality scores into an overall quality score. Fourney and Laganier [13] proposed a quality assessment method based on 6 different criteria, then performed a weighted sum of the measures to get the overall quality score. Abaza et al. [14] developed an image quality assessment for face recognition based on five quality factors, then integrated into a generic face quality index (FQI). They also proposed a face image quality index that combines multiple quality measures [15]. Omidiora et al. [16,17] developed a facial image verification and quality assessment framework (FVQA) using different algorithms and methods to extract quality measures. Chen and Li [18] presented an image quality assessment model which produces a noise score for a face image using several quality measures. Bagdanov et al. [19] proposed two quality measures for face images based on symmetry and pose. Zhang and Wang [20] proposed three asymmetry-based face quality measures by utilizing SIFT feature points on face images. Xiong and Jaynes [21] introduced an intrinsic quality measure using bilateral symmetry, color, resolution and aspect ratio, and performed a weighted summation to get the quality score. Yao et al. [22] developed an adaptive face image quality measure based on image sharpness measures. Bhattacharjee et al. [23] proposed a quality metric based on sharpness, noise, contrast, luminance and eye detection ability. Anantharajah et al. [24] presented a framework that used face symmetry, sharpness, contrast, closeness of mouth, brightness and openness of the eye, and employed a neural network to fuse the normalized feature scores. In another work, they also considered a fusion of all four measures using a weighted summation [25]. Nasrollahi and Moeslund did several studies [26–29] where they used a four parameter based face quality assessment method: head-pose, resolution, sharpness and brightness. They tried different fusion methods, such as, using fuzzy inference engine [26], or, combining them into one quality score using weighted sum of normalized values [27, 29]. Nasrollahi et al. [30] also proposed a face quality assessment system based on ten facial attributes fed to a multilayer perceptron to produce the quality score. Haque et al. [31] employed a face quality assessment method using the out-of-plan face rotation (pose), sharpness, brightness and resolution, and calculated the final quality score by linearly combining the parameters with empirically assigned weight factors. Lin et al. [32] employed a fuzzy inference engine based data fusion method to integrate four quality criteria. Wei et al. [33] used pose,

brightness, and face size as quality features, and used a weighted averaging scheme for fusion. Gao et al. [34] developed a face symmetry based quality score generation method. Long and Li [35] presented a five feature based near infrared face image quality assessment system. They combined the features into a general score using a weighted sum, where weights were empirically set. Sang et al. [36] presented methods using Gabor feature based facial symmetry and sharpness measure. De Marisco et al. [37] proposed new quality indices based on pose and illumination distortion and face symmetry. Abboud et al. [38] presented two no-reference image quality measures for face recognition, called symmetrical adaptive local quality index (SALQI) and middle halve (MH).

Another traditional approach for face quality assessment is to use a standard face template, and use the discrepancy from this template to the query image as the measure of face quality. Kryszczuk et al. [39] used a combination three quality measures: the difference of mean of query image from the mean of the normalized images of the reference set, normalized cross-correlation between query image and the average face template, and block based likelihood estimation. Later they modeled the quality of the face images by creating an average face template out of reference images, they also incorporate sharpness estimation for quality assessment [40]. Kryszczuk and Drygajlo [41,42] used absolute distance between the log-likelihood ratio and the decision threshold, sum of log-likelihoods, correlation with average face template and image sharpness estimation as quality measures. Truong et al. [43] used a quality measure that reflects the difference in quality between a template image and quality image.

Researchers have also shown interest in learning based face image quality assessment methods. Liao et al. [44] employed a hierarchical binary decision tree classifier based on support vector machines (SVM) to categorize the face images into five quality levels. Bharadwaj et al. [45] used Gist and sparsely pooled Histogram of Oriented Gradient (HoG) features and a one-vs-all multi-class SVM to classify face images into four quality categories. Bhatt et al. [46,47] presented a quality assessment algorithm which computes a quality vector comprising no-reference quality, edge spread, spectral energy, and pose, and then trained SVMs for decision making. Ozay et al. [48] proposed a unified probabilistic framework to simultaneously predict the quality of the facial image samples and perform quality-based face recognition by exploiting these relationships. Wong et al. [7] proposed a patch-based face image quality assessment algorithm which quantifies the similarity of a face image to a probabilistic face model. El-Abed et al. [49] used a no-reference based image quality metric called BLINDS, SIFT keypoints, DC coefficient, and, mean and standard deviation of scales as features and then used a SVM to predict the quality. Chen et al. [50] proposed a learning to rank based framework for assessing the face image quality. Kim et al. [51] proposed a learned FIQ assessment method that considers visual quality and mismatch between training and test face images for quality assessment. Recently, deep learning based quality assessment methods have also been introduced by different researchers. Liu et al. [8] proposed a non-reference face image assessment algorithm based on the deep features extracted from the VGG network. Vignesh et al. [9] proposed a FQA algorithm based on mimicking the recognition capability of a given face recognition algorithm by using a Convolutional Neural Network (CNN). Pan et al. [10] trained the VGG-16 deep CNN to output a general face quality metric which considers various quality factors such as bright, contrast, blurriness, occlusion, pose etc. Thorsten et al [52] used a type of Recurrent Neural Network (RNN) called Long Short-Term Memory (LSTM) to perform a binary quality classification.

Table 2.1: A brief overview of FIQA methods. (*Details of the databases can be found in [1]*)

Paper	Database	Features / Quality metrics	Method / Model
Hsu et al. [12]	FRGC 2.0, passport database (private)	Detectable Eyes, Face Geometry, Good exposure and contrast, well focused, proper lighting, head pose, wearing eyeglasses, presence of face, high resolution skin texture, liveness, natural expression etc.	Fusion of quality metrics
Fourney et al. [13]	Videos (private)	Pose, illumination, sharpness, presence of human skin, image resolution	Fusion of quality metrics
Abaza et al. [14, 15]	CAS-PEAL, Yale, FMTC (FERET+MBGC), FOCS, QFIRE	Contrast, brightness, focus, sharpness, illumination	Fusion of quality metrics
Omidiora et al. [16, 17], Abayomi et al. [53]	SCFace, BFSC, local database (private)	Faceness, pose, contrast, illumination, similarity	Fusion of quality metrics
Chen et al. [18]	CAS-PEAL	Occlusion, face-to-camera distance, pose, uneven illumination	Fusion of quality metrics
Bagdanov et al. [19]	Videos (private)	Image symmetry, face pose	Quality metric thresholding
Bhattacharjee et al. [23]	IITK	Eye detection, sharpness, noise, contrast, luminance	Fusion of quality metrics
Anantharajah et al. [24, 25]	Honda/UCSD	Face symmetry, sharpness, contrast, brightness, expression neutrality (mouth closeness), eye openness	Fusion of quality metrics
Nasrollahi et al. [26–29], Haque et al. [31], Lin et al. [32]	CVL, AT&T (ORL), FERET, Face96, Video database (private)	Pose, sharpness, brightness, image resolution	Fusion of quality metrics
Nasrollahi et al. [30]	FERET, CVL, Face96, local database (private)	Pose, brightness, sharpness, resolution, eye openness, gaze, mouth closeness	Fusion of quality metrics
Wei et al. [33]	ICP Workshop Evaluation Data	Pose, sharpness, brightness, face size	Fusion of quality metrics
Long et al. [35]	CBSR (private)	Pose, brightness, sharpness, resolution, eye openness, mouth closeness	Fusion of quality metrics
De Marsico et al. [37]	FERET, LFW, SCFace	Pose, illumination, and symmetry distortion	Quality metric thresholding
Abboud et al. [38]	YaleB	Universal Image Quality Index (UIQI), symmetrical adaptive local quality index (SALQI), middle halve (MH)	Quality metric thresholding

Table 2.1: A brief overview of FIQA methods. (*Details of the databases can be found in [1]*)
(continued.)

Kryszczuk et al. [39, 41, 42]	BANCA	Distance between log-likelihood and decision threshold, Sum of log-likelihoods, Correlation with an average face template, sharpness	Gaussian Mixture Model (GMM)
Truong et al. [43]	Local database (private)	brightness, contrast, focus, illumination	Fusion of quality metrics
Raghavendra et al. [54]	Video (private)	Pose and texture features extracted using GLCM	GMM Classifier
Vatsa et al. [55]	Local database (private)	Redundant DWT based quality metrics	Multiclass SVM classifier trained on fusion score
Liao et al. [44]	Local database (private)	Gabor feature	SVM
Bharadwaj et al. [45]	CAS-PEAL, SCFace	Gist, HoG	Multi-class SVM
Bhatt et al. [46], Bharadwaj et al. [47]	WVU Multimodal Database, MBGCv2, Multi-PIE, AR, LEA (private)	Energy spectrum, Edge spread, Blockiness, Activity, ZC-rate, Pose	SVM
Ozay et al. [48]	IMM Face Database	Appearance coefficient of an Active Appearance Model (AAM)	Maximum Likelihood Estimation (MLE)
Wong et al. [7]	FERET, PIE, Chokepoint	2D Discrete Cosine Transform (DCT) feature	Probabilistic face model to measure distance from an ‘ideal’ face
El-Abed et al. [49]	Faces94, ENSIB, FERET, AR	BLIINDS, SIFT keypoints, DC coefficients, mean of scale, standard deviation of scale	Multiclass SVM
Chen et al. [50]	LFW, ALFW, FRGC, FERET, card photo db (private)	Hough, Gabor, Gist, LBP, CNN	Ranking SVM
Kim et al. [51]	FRGC 2.0	Pose/alignment, blurriness, brightness, mismatch in metric, color mismatch	AdaBoost
Berrani et al. [56]	Asian Face Image DB, FDB15	Image vectors processed using classical PCA	RobPCA
Sellahewa et al. [57]	Extended YaleB, AT&T (ORL)	Universal image quality index (UIQI)	Quality metric compared against a reference image
Abdel-Motaleb et al. [58]	WVU face database, FERET	Sharpness, illumination, pose, expression assessment	GMM-UBM classifier
Kim et al. [59, 60]	FRGC 2.0, Local database (private)	Pose, illumination, sharpness, eye openness, contrast, resolution	Fusion of quality metrics
Liu et al. [8]	Surveillance video db (private)	VGG deep feature	support vector regression (SVR)

Table 2.1: A brief overview of FIQA methods. (*Details of the databases can be found in [1]*)
(*continued.*)

Vignesh et al. [9]	ChokePoint dataset	Mutual Subspace Method (MSM) based match scores from a face recognition method using LBP and HoG features	Custom defined CNN
Best-Rowden and Jain [61]	LFW, IJB-A	Deep-320 (deep features), Feat-5 (Hog, Gabor, Gist, LBP, CNN)	SVR, RankSVM
Pan et al. [10]	Color FERET, Kinect-FaceDB	Deep features	VGG-16 network
Yu et al. [62]	CASIA-Webface, LFW, YoutubeFaces	Deep features	LightCNN
Thorsten et al. [52]	AR, FRGC, NCKU face, Yale Face, CASIA Face V5	Deep features	LSTMs

2.3 Face Image Quality Database

Almost all works reviewed in previous section use small datasets for face quality assessment (See also Table 2.1). To develop a large-scale face image quality database, we have made some preliminary observations: 1) The database should contain a wide range of quality, hopefully with all possible face image qualities; 2) If face matching, is a step towards quality assessment then, the subjects should have at least two images; 3) It could be useful to have face images with many different qualities for each subject. For the publicly available face databases, we can find basically two groups of datasets: 1) Controlled: constructed with sets of images taken in several environmental conditions, such as lighting, expression, accessories, pose, indoor/outdoor, etc. Each of these variations are usually grouped into data subsets; 2) Real-world: collected from the internet, where images are taken in unconstrained environments, and then annotated according to the identity. We utilized both data types by regrouping the controlled subsets each subject, which gives us subjects with varied face image quality. For, real-world face images, the main concern was to avoid identity noise images. Manual cleaning was done for some of the comparatively larger real-world datasets as much as possible.

We selected 40 public databases from the list of face databases available in the face recognition homepage [1]. Our composite database contains face images taken in controlled and unconstrained real-world scenarios. The resulted FIQDB database has a total number of 545,684 images of 14,373 subjects. How to ensure that the database is large enough to contain all possible qualities? It is not trivial because there is no universal consensus about all factors that affect the face image quality. Controlled face recognition databases are usually constructed with emphasis on some specific set of imaging conditions, so there is less chance for these databases to cover a wide range of qualities. Even for datasets with real-world images, there could be a large imbalance among the image numbers of different quality levels. We assume that by aggregating as many datasets as possible with varied conditions, we may acquire all possible qualities potentially.

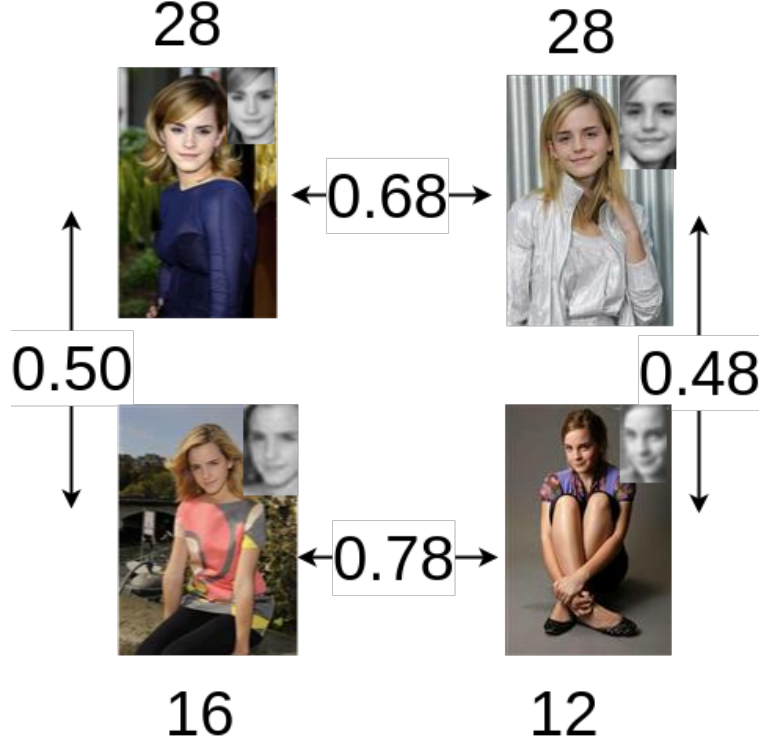


Figure 2.1: Quadrature face match graph showing cosine similarity scores between arrow points and quality scores above/below face image. Cropped faces shown inset.

2.4 Quality Labels

There is no clearly defined method for generating face image quality labels. Rowden and Jain [61] identified three possible approaches for quality label generation: (1) using a fusing scheme with different face image quality measures to generate a single score representing overall face image quality, (2) using human assessors to estimate face image quality, and (3) generating labels from match scores coming from face recognition methods. A major limitation of using fusion of quality measures is that there are an overwhelmingly large number of quality factors that can influence the performance of a face recognition system and their exact count is still unknown. Furthermore, accurate measurement of image quality measures is still an open problem. Kryszczuk et al. [63] provided theoretical and experimental results to show that the mapping of quality measures into one quality score inadvertently causes a loss of information and a reduction of the classifier’s degree of freedom. Moreover, fusion-based approaches only perform as well as their individual classifiers.

Since the biometric sample quality is typically defined in terms of automatic recognition performance, human visual perception of image quality may not be well correlated with recognition performance [2, 3, 45]. Adler and Dembinsky [64] found very low correlation between human and algorithm measurements of the face image quality. It is evident that human perception and computer processing are not always consistent. For example, a human may assess a face image to have good quality because of its sharpness measure, but a recognition algorithm that works in low frequencies will be inconsistent with the human statement of quality. Human inspection can improve with adequate training on quality assessment, but it is expensive and time consuming. In

addition, incorporating a human quality assessor could create other problems, such as inaccuracy due to the tiredness, boredom, or lack of motivation [65]. These problems will increase more as the database gets larger. Hsu et al. [66] found some correlation between human assessment and recognition based measures of face image quality, but also showed that it is more difficult to separate middle quality images from the low quality than to separate high quality from the middle quality. Recently, Best-Rowden and Jain [61] did a comparison between match score and human assessment based quality labeling. The results based on assessing false non-match rate (FNMR) against “percentage of probe images removed” showed that, match score based labels are much more efficient in reducing FNMR than human assessment based quality labels. But, directly using face match scores as labels might not be optimal solution for training learning based face image quality assessment methods. Beveridge et al. [67] showed that, it is much more common to find relationships in which two images that are hard to match to each other can be easily matched with other images of the same person. Figure 2.1 shows an example of low quality image pair of the same subject matching with high similarity scores, but high and low quality pairs produce lower similarity scores.

2.4.1 Relevance levels from identification rank

It is relatively easier to evaluate the quality difference between two face images, than individually assess the absolute quality of a face image. This information can be used to train a ranking function in a semi-supervised manner to generate quality labels. Rowden and Jain [61] used human assessment to partially obtain preference information for the training data and then used a matrix completion technique to generate the preference for the rest of the face images. This kind of semi-supervised approach is based on the assumption that the human assessments are consistent, which is actually difficult to satisfy because it is dependent on the expertise of the humans on assessing the biometric quality. Chen et al. [50] assumed the same quality for all images in one dataset. This assumption may be unrealistic, because databases constructed using real world images have different biometric qualities. Another way to generate the preference pairs is from relevance levels, specially when the number of images is large. In their implementation of large scale linear RankSVM (denoted as LSLRSVM in this chapter), Lee and Jin [68] used relevance levels to construct the preference pair matrix. But for their work they used web-search engine ranks which is widely different than our case where we used different types of image features. In our work, the image with the higher preference will have a higher relevance level than the other, which can be represented by a rank or score.

In this work we propose to generate the relevance levels from the identification rank. Usually the face quality is related to the face recognition method in consideration when the matching scores are used to compute the quality values. Using a suitable face recognition method, we find out the match rank of each probe image and use it as its relevance level. We try to resolve the ambiguity issue in match scores, by ensuring that the gallery is constructed with high quality face images. Note that, we cannot use the match scores directly as relevance levels because it is not always the case that the true positive image will have the highest match score. Moreover, human based relative ranking of each preference pair is also not feasible due to the huge number of pairs that can be generated for our FIQDB training set.

In order to compute the relevance levels from the identification rank, the gallery images are sorted in a descending order of similarity score, and then the position of the true positive

(TP) image is located. If the TP image for the i -th probe image is at the r -th position of the sorted gallery, then the relevance level $y_i = |G| - r$, where G is the gallery set. If K is the set of relevance levels, then we can write, $y_i \in K \subset \mathbf{R}$ and the extracted feature vectors as $\mathbf{x}_i \in \mathbf{R}^n$, where $i = 1, 2, \dots, l$, and l is the number of samples. From the relevance levels, we can obtain the set of preference pairs as,

$$P = \{(i, j) | y_i > y_j\} \quad (2.1)$$

If $p = |P|$ then, we can construct a preference pair matrix A of size $p \times l$ such that,

$$A = \begin{matrix} & \vdots & & & & & & & & \\ & & \dots & & i & & \dots & & j & \dots \\ (i, j) & \left[\begin{array}{cccccccc} 0 & \dots & 0 & +1 & 0 & \dots & 0 & -1 & 0 & \dots \end{array} \right] \\ & \vdots & & & & & & & & \end{matrix}$$

That is for a preference pair (i, j) , s.t. $y_i > y_j$ the corresponding row in A in the i -th column is $+1$, and j -th column is -1 . Then, according to Lee and Jin [68], an the objective function of L_2 -loss linear RankSVM can be written as,

$$\min_{\mathbf{w}} \frac{1}{2} \mathbf{w}^T \mathbf{w} + C(\mathbf{e} - \mathbf{A}\mathbf{X}\mathbf{w})^T \mathbf{D}_w(\mathbf{e} - \mathbf{A}\mathbf{X}\mathbf{w}) \quad (2.2)$$

where $C > 0$ is a regularization parameter, \mathbf{w} is the weight vector, $X = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_l\}$ is the set of feature vectors, $\mathbf{e} \in \mathbf{R}^{p \times 1}$ is a vector of ones, and D_w is a $p \times p$ diagonal binary matrix following a conditional property. In prediction, for any test feature vector \mathbf{x} , a larger $\mathbf{w}^T \mathbf{x}$ implies that \mathbf{x} should be ranked higher.

Algorithm 1 Quality label generation algorithm

Input: G, P // probes, $P = \{p_i\}_{i=0}^{M-1}$, gallery, $G = \{g_j\}_{j=0}^{N-1}$

Input: I_G, I_P // Identity of gallery and probe sets.

Output: predicted scores $\{\hat{y}_i\}_{i=0}^{m-1}$

- 1: $R = \{\}$ // R stores the relevance levels
 - 2: $k = 0$
 - 3: **for** $i = 0, \dots, M-1$ **do**
 - 4: $S = \{\}$ // S stores the match scores.
 - 5: **for** $j = 0, \dots, N-1$ **do**
 - 6: $S_i \leftarrow \text{Cosine_Similarity}(p_i, g_j)$ // Get match score
 - 7: **end for**
 - 8: $\text{Sort}(S, \text{Descend})$ // Sort in descending order
 - 9: $r = \text{FindTPRank}(S, I_G, I_P)$ // Get position of the TP feature
 - 10: $R_k \leftarrow |G| - r$
 - 11: $k = k + 1$
 - 12: **end for**
 - 13: construct preference pair matrix A using R
 - 14: Divide P into non-overlapping training and test set P_{train} and P_{test}
 - 15: Divide P_{train} into subsets P_{tr1}, P_{tr2}
 - 16: optimize Eq. (2.2) using P_{tr1} and A
 - 17: $\hat{y} \leftarrow \text{predict}(P_{tr2})$ // P_{tr2} is used for training the FIQA methods
 - 18:
 - 19: **return** predicted quality scores $\{\hat{y}_i\}_{i=0}^{m-1}$
-

Addressing the large number of pair generation problem: If on an average, l/k instances are with the same relevance level where $k = |K|$, the number of pairs in P is $\binom{k}{2} \times O((\frac{l}{k})^2) = O(l^2)$. This large number of pairs is the main difficulty to train the RankSVM, because of the memory issue. For a database, such as our FIQDB with over 500K images, this becomes a huge problem. Airola et al. [69] showed that it is possible to avoid the $O(l^2)$ complexity of going through all the pairs in calculating the objective function, gradient or other information needed in the optimization procedure, by employing an order-statistic tree. The complexity can be reduced to $O(l\bar{n} + l\log l + l\log k + n)$, where $O(l\bar{n})$ cost is for calculating $\mathbf{w}^T \mathbf{x}_i$, $\forall i$; \bar{n} is the average number of non-zero features per training instance; $O(l\log l + lk)$ is for the sum of training losses in Eq. (2.2); and $O(n)$ is for the regularization term $\mathbf{w}^T \mathbf{w}/2$.

2.5 Quality Representation

Based on the existing FIQA approaches (Table 2.1, three different feature categories were selected for the study. They are: 1) a set of traditional face image quality measures, which are referred to as “heuristic measures”; 2) face recognition features; and 3) natural scene statistics based image quality features. A total of 12 different feature types are explored for the quality representation: 4 facial, 7 natural scene statistics, and 1 set of heuristic features.

2.5.1 Heuristics (HR) based features

We use the following 14 commonly used quality measures that can be directly calculated from the image in consideration:

1) *Brightness*: The average value of the illumination component of all pixels in the face region is considered as the brightness measure. For a grayscale face image I , the brightness measure B is calculated as [27, 31],

$$B = \frac{1}{MN} \sum_{x=1}^M \sum_{y=1}^N [I(x, y)] \quad (2.3)$$

where the image size is I is $M \times N$.

2) *Contrast*: Image contrast is the difference in color intensities that makes a face distinguishable. The face image contrast C can be measured using the following [15]

$$C = \sqrt{\frac{\sum_{x=1}^M \sum_{y=1}^N [I(x, y) - \mu]^2}{MN}} \quad (2.4)$$

where μ is the mean of pixel values in image I .

3) *Focus*: Edge density measures the average magnitude of the gradient over the face image. The assumption is that when images are in focus, the average gradient magnitude will be higher than when the image is out of focus [3, 70–72].

4) *Illumination*: Spectral energy is used as the illumination measure, which describes abrupt changes in illumination and specular reflection. The image is tessellated into several non-overlapping blocks and the spectral energy is computed for each block. The value is computed as the magnitude of Fourier transform components in both horizontal and vertical directions [13, 46].

5) *Illumination Symmetry*: This is calculated as the absolute difference between the mean intensity

of left and right sides of the face image.

6) *Sharpness*: The sharpness measure, denoted by S , is computed by [27]:

$$S = \frac{1}{MN} \sum_{x=0}^M \sum_{y=0}^N \left(|I(x, y) - H(I(x, y))| \right) \quad (2.5)$$

where $I(x, y)$ is the intensity of image I at location (x, y) , $H(\cdot)$ is a lowpass Gaussian filter, and $M \times N$ is the size of the image.

7) *Compression*: A no-reference based JPEG image quality measure [73] is used to measure the compression quality.

8-10) *Pose estimation (yaw, pitch and roll)*: Three euler angles in radian are estimated for the three poses w.r.t the camera center. The angles are in the range $[-\pi, \pi]$.

11-13) *Eyes and mouth openness*: Eyes and mouth openness is calculated based on the rectangular region detected around the eyes and mouth measured from facial landmarks. The ratio of the width and height of the detected region is considered. The higher the ratio, the more open the eyes or mouth. For two eye regions the mean of the ratios is considered.

14) *Face symmetry*: The mean difference between the original and horizontally flipped images is considered as the measure of how symmetrical are both sides of the face. The lower the difference, the more symmetrical [25]. This can also assess the pose and illumination variance of the face.

2.5.2 Weighted sum of heuristic measures (WSHRM)

From Table 2.1 it is evident that most of the traditional fusion based methods use a limited number of quality measures to calculate their fusion based quality score. We propose to use a quality measure fusion based approach by aggregating the above mentioned 14 heuristic measures for quality. To calculate WSHRM, for each image, all heuristic measures are collected and then normalized. A weighted sum of the normalized measures is used to map the heuristic measures to a quality score.

$$QS = \frac{\sum_{i=0}^n S_i * W_i}{\sum_{i=0}^n W_i} \quad (2.6)$$

We use the spearman correlation between the individual heuristic measure and the identification ranks as the summation weights W . We use the training set P_{tr1} to calculate the correlation.

2.5.3 Face recognition (FR) features

It is quite natural to use face recognition features for quality assessment. Several researches have used face recognition features for FIQA [11, 45, 50, 61]. We chose, four different face recognition features were chosen, e.g., HoG [74], Gist [75], Gabor [76], and LBP [77], to characterize the face quality from different aspects.

HOG descriptor [74]: This feature representation is based on local histograms of image gradient orientations in a dense grid. This is implemented by dividing the image window into small spatial regions called “cells”. Each cell accumulates a local 1-D histogram of gradient directions or edge

orientations over the pixels of the cell. The combined histogram of all the cell entries form the feature representation. To improve robustness against illumination variance, contrast-normalization is performed on the local responses.

A cell size of 10×10 is adopted to capture the spatial information. A 2×2 block size was used to hold the cells with 1×1 block overlap to ensure adequate contrast normalization. 8 orientation histogram bins were used where orientation values were evenly spaced. The resulted feature vector size is 1440 with an image size of 100×60 .

Gabor filters [78]: Due to the biological relevance and computation properties, Gabor wavelets were introduced to image analysis. As a feature generator, Gabor filters have been widely used in face recognition. The kernels of Gabor wavelets are similar to the 2D receptive field profiles of the mammalian cortical simple cells, exhibiting desirable characteristics of spatial locality and orientation selectivity. The Gabor wavelets (kernels, filters) can be defined as following:

$$\psi_{\mu,v} = \frac{\|k_{\mu,v}\|^2}{\sigma^2} e^{-\|k_{\mu,v}\|^2 \frac{\|z\|^2}{2\sigma^2}} [e^{ik_{\mu,v}z} - e^{-\frac{\sigma^2}{2}}] \quad (2.7)$$

where v and μ define the scale and orientation of the Gabor kernel, $z = (x, y)$ denotes the pixel location, $\|\cdot\|$ denotes the Euclidean norm operator, and the wave vector $k_{\mu,v} = k_v e^{i\phi_\mu}$. Here, $\phi_\mu = \frac{\pi\mu}{8}$ is the orientation parameter and $k_v = \frac{k_{max}}{f^v}$, where f is the spacing factor between filters in the frequency domain. Given an input face image I , its convolution with a Gabor wavelet $\psi_{\mu,v}$ can be defined as

$$G_{\mu,v} = I(z) * \psi_{\mu,v}(z) \quad (2.8)$$

where $*$ denotes the convolution operator. For each Gabor kernel, at every image pixel z , a complex number containing two parts; real $Re(\cdot)$ and imaginary $Im(\cdot)$, can be obtained. Based on these two parts, the magnitude $|G_{\mu,v}(z)|$ is computed as follows:

$$|G_{\mu,v}| = \sqrt{Im(G_{\mu,v}(z))^2 + Re(G_{\mu,v}(z))^2} \quad (2.9)$$

A Gabor filter bank with filter size 29×29 at 3 scales and 8 orientations was used to convolve with the cropped face image of size 100×60 . Then the magnitude part of Gabor response was extracted, down sampled by a factor of 10 and normalized. The size of the feature vector is $10 \times 6 \times 3 \times 8 = 1440$.

GIST descriptor [75]: GIST summarizes the gradient information (scales and orientations) for different parts of an image, which provides a rough description, “the gist”, of the scene. This descriptor is computed by convolving the image with 32 Gabor filters at 4 scales, 8 orientations, producing 32 feature maps of the same size of the input image. Each feature map is divided into 36 regions by a 6×6 grid and then we average the feature values within each region. Concatenating the 36 averaged values of all 32 feature maps results in a $36 \times 32 = 1152$ size GIST descriptor. Before convolving, the image is usually pre-filtered using the local contrast scaling method.

Local Binary Pattern (LBP) [77]: LBP is a type of visual descriptor that encode local texture information. It is a powerful feature for texture extraction. To create LBP feature, the image is divided into cells. For each pixel in a cell, it is compared with the pixel to each of its 8 neighbors by following in a clockwise or counter-clockwise direction. If the center pixel’s value is greater than

the neighbor's value, a value of 1 is assigned, otherwise, 0 is assigned. This gives an 8-digit binary number, which is usually converted to decimal for convenience.

An input image of size 100×60 is scaled to half its size of 50×30 , then convolved with $8 \ 3 \times 3$ LBP filters. The response matrices are then converted to binary based on the positive or negative values, and combined by multiplying with corresponding positional weights. Resultant LBP Feature size is $50 \times 30 = 1500$.

2.5.4 Natural Scene Statistics (NSS) features

Natural scene statistics (NSS) models seek to capture the statistical properties of natural scenes that hold across different contents. Presence of distortions in natural images alters the natural statistical properties of images, thereby rendering them and their statistics unnatural. Image quality assessment methods based on NSS capture this "unnaturalness" in the distorted image and relate it to the perceived quality. Several researchers have tried to incorporate NSS based IQA methods for FIQA: El-Abel et al. [49] used BLINDS features [79] combined with others. Bharadwaj et al. [11] used BRISQUE index [80] as a quality metric to study its behavior with respect to match scores obtained from face recognition systems. Recently, Liu et al. [81] have done a performance evaluation of different no-reference measures for face biometrics. A comparative study on the NSS features with other popular feature types could tell us more about the usefulness of this feature type for FIQA. We selected seven NSS feature types from well known no-reference image quality assessment (NR-IQA) algorithms for our study. These are: 1) Spatial and spectral entropy features [82], 2) BRISQUE features [80], 3) BLINDS-II features [79], 4) DIIVINE features [83], 5) Curvelet features [84], 6) NIQE features [85] and 7) TMIQA features [86]. Sizes of the NSS features are: Spatial and spectral entropy features 12, BRISQUE features 36, BLINDS-II features 24, DIIVINE features 88, Curvelet features 12, NIQE features 36 and TMIQA features 36.

1) *Spatial and spectral entropy features [82]*: Natural photographic images are highly structured in the sense that their pixels exhibit strong dependencies in space and frequency. These dependencies carry important information about the visual scene. Localized image entropy features can capture the degree of local image structure and the entropy degree can denote the dependence level between pixels.

To extract these features, input image is first downsampled by a factor of 2, enabling simple multiscale analysis. The image is decomposed into 3 scales: low, middle and high, yielding 3 scale responses. These responses are partitioned in 8×8 blocks, called local image patches. For each image blocks spatial and frequency entropies are computed. In final step, a percentile feature pooling is performed, the spatial and spectral entropies are sorted in non-decreasing order. This provides two ordered sets $S = \{s_1, s_2, \dots, s_n\}$, and $F = \{f_1, f_2, \dots, f_n\}$, where s_i and f_i are local spatial and spectral entropies, respectively. 60% of the central elements are extracted from S , which produces the set $S_c = \{s_{\lfloor 0.2m \rfloor}, s_{\lfloor 0.2m \rfloor + 1}, \dots, s_{\lfloor 0.8m \rfloor}\}$, also 60% of the central elements are extracted from F , which produces $F_c = \{f_{\lfloor 0.2m \rfloor}, f_{\lfloor 0.2m \rfloor + 1}, \dots, f_{\lfloor 0.8m \rfloor}\}$. The following formula produces the features from each scale:

$$f = (\overline{S_c}, S_\gamma, \overline{F_c}, F_\gamma) \quad (2.10)$$

where $\overline{S_c}$ and $\overline{F_c}$ are mean of S_c and F_c , S_γ and F_γ are skewness of S and F , respectively. For three

scales, the combined final feature size becomes $2 \times 2 \times 3 = 12$.

2) *BRISQUE features* [80]: Blind/Referenceless Image Spatial QUality Evaluator (BRISQUE) approach is based on the assumption that when image distortions are introduced statistical regularities of natural images are disturbed. It uses scene statistics of locally normalized luminance coefficients to quantify possible losses of “naturalness” in the image due to the presence of distortions. The features used are derived from the empirical distribution of locally normalized luminances and products of locally normalized luminances under a spatial natural scene statistic model. Given an input image, locally normalized luminances are computed via local mean subtraction and divisive normalization. Local mean (μ) and contrast (σ) are computed by.

$$\mu(i, j) = \sum_{k=-K}^K \sum_{l=-L}^L w_{k,l} I_{k,l}(i, j), \quad (2.11)$$

$$\sigma(i, j) = \sqrt{\sum_{k=-K}^K \sum_{l=-L}^L w_{k,l} (I_{k,l}(i, j) - \mu(i, j))^2}, \quad (2.12)$$

where w is 2D Gaussian weighting function. Then a zero mode asymmetric generalized Gaussian distribution (AGGD) is utilized to produce the features ($\gamma, \sigma_l^2, \sigma_r^2$):

$$f(x; \gamma, \sigma_l^2, \sigma_r^2) = \begin{cases} \frac{\gamma}{(\beta_l + \beta_r) \Gamma(\frac{1}{\gamma})} e^{-(\frac{-x}{\beta_l})^\gamma} & \forall x \leq 0 \\ \frac{\gamma}{(\beta_l + \beta_r) \Gamma(\frac{1}{\gamma})} e^{-(\frac{x}{\beta_r})^\gamma} & \forall x > 0 \end{cases} \quad (2.13)$$

where

$$\beta_l = \sigma_l \sqrt{\frac{\Gamma(\frac{1}{\gamma})}{\Gamma(\frac{3}{\gamma})}} \quad (2.14)$$

$$\beta_r = \sigma_r \sqrt{\frac{\Gamma(\frac{1}{\gamma})}{\Gamma(\frac{3}{\gamma})}} \quad (2.15)$$

and $\Gamma(\cdot)$ is the gamma function

$$\Gamma(a) = \int_0^\infty t^{a-1} e^{-t} dt, a > 0 \quad (2.16)$$

The parameter γ controls the shape of the distribution, while σ_l^2 and σ_r^2 are scale parameters that control the spread on each side of the mode, respectively. Mean of the distribution is also used as a feature, which is defined as below:

$$\eta = (\beta_r - \beta_l) \frac{\Gamma(\frac{2}{\gamma})}{\Gamma(\frac{1}{\gamma})} \quad (2.17)$$

16 parameters are calculated by computing $(\gamma, \sigma_l^2, \sigma_r^2, \eta)$ along four orientations, which gives a total of 18 features (including local mean and contrast) per scale. Two scales are used to compute all features, by low pass filtering and downsampling by a factor of 2, to capture multiscale behavior. This gives the final feature of size 36.

3) *BLIINDS-II features [79]*: BLind Image Integrity Notator using DCT Statistics - II (BLIINDS-II) is a blind/ no-reference image quality assessment (IQA) algorithm which uses a natural scene statistics (NSS) model of discrete cosine transform (DCT) coefficients. The features are based on an NSS model of the image DCT coefficients. The input image, first subjected to local 2-D DCT coefficient computation. The image is partitioned into equally sized 5×5 blocks, called local image patches. Then on each of these blocks, a local 2-D DCT is computed. 4 model based features are computed: 1) Shape parameter (γ), 2) coefficient of frequency variation (ζ), 3) energy subband ratio measure (R), and 4) orientation feature. The shape parameter (γ) is a model-based feature obtained by using the generalized Gaussian density function:

$$f(x|\alpha, \beta, \gamma) = \alpha e^{(-\beta|x-\mu|)^\gamma}, \quad (2.18)$$

where μ is the mean, and α and β are the normalizing and scale parameters. Coefficient of frequency variation feature is defined as:

$$\zeta = \sqrt{\frac{\Gamma(1/\gamma)\Gamma(3/\gamma)}{\Gamma^2(2/\gamma)} - 1}, \quad (2.19)$$

The energy subband ratio measure is computed by the following function:

$$R_n = \frac{|E_n - \frac{1}{n-1} \sum_{j < n} E_j|}{E_n + \frac{1}{n-1} \sum_{j < n} E_j}, \quad (2.20)$$

where R_n is defined for $n = 2, 3$, and the means of R_2 and R_3 are used.

For orientation features, DCT coefficients are collected along three orientation bands. A generalized Gaussian model is fitted to the coefficients and the variance of ζ is computed along each of the three orientations.

All features are computed for all blocks in the image. Then the feature is pooled by averaging over the highest 10th percentile and over all (100th percentile) of the local block scores across the image. This provides 2 numbers for each feature per image, resulting in a $2 \times 4 = 8$ features. The feature extraction is repeated 3 times after low-pass filtering of the image, and subsampling it by a factor of 2, which gives a final feature of size $3 \times 8 = 24$.

4) *DIIVINE features [83]*: Distortion Identification-based Image Verity and INtegrity Evaluation (DIIVINE) index is a NSS based QA algorithm that assesses the quality of a distorted image without the need for a reference image. The input image goes through a wavelet decomposition using a steerable pyramid decomposition, over two scales and six orientations. The resulting decomposition results in 12 subbands across orientations and scales. The obtained subband coefficients are then utilized to extract a series of statistical features, stacked to form a vector, which is a statistical description of the distortion in the image. The features extracted are: 1) Scale and orientation selective statistics ($f_1 - f_{24}$), 2) orientation selective statistics ($f_{25} - f_{31}$), 3) correlations across scales ($f_{32} - f_{43}$), 4) spatial correlation ($f_{44} - f_{73}$), and 5) across orientation statistics ($f_{74} - f_{88}$).

$f_1 - f_{12}$ correspond to variance σ^2 accross subbands, $f_{13} - f_{24}$ correspond to the shape parameter γ across subbands. $f_{25} - f_{30}$ correspond to γ from the statistics across scales over different orientations, while f_{31} corresponds to γ from the statistics across subbands. The structural correlation is computed as

$$\rho = \frac{2\sigma_{xy} + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \quad (2.21)$$

where σ_{xy} is the cross-covariance between the windowed regions from the band-pass and high-pass bands, σ_x^2, σ_y^2 are their windowed variances, respectively, and C_2 is a stabilizing constant. Using 12 subbands, 12 such correlations are computed, creating the features $f_{32} - f_{43}$. In order to capture spatial correlation statistics, for each chess-board distance $\tau \in \{1, 2, \dots, 25\}$ the joint distribution is computed. The joint distribution for a value of τ can be considered as the joint distribution $p_{XY}(x, y)$ between the random variables X and Y . To estimate the correlation between the two variables, we can compute as follows,

$$\rho(\tau) = \frac{E_{p_{XY}(x,y)}[(X - E_{p_X(x)}[X])^T(Y - E_{p_Y(y)}[Y])]}{\sigma_X \sigma_Y}, \quad (2.22)$$

where $E_{p_X(x)}[X]$ is the expectation of X with respect to the marginal distribution $p_X(x)$, τ is the distance at which the estimate of ρ is computed. After calculating $\rho(\tau)$ the obtained curve is parameterized by fitting it with a 3rd order polynomial. The coefficients of the polynomial and the error between the fit and the actual $\rho(\tau)$ form the features $f_{44} - f_{73}$. Statistical correlations across orientations are computed by utilizing windowed structural correlation between all possible pairs of subbands at the coarsest scale. The lowest 5% of the structural correlation values for each pair form the features $f_{74} - f_{88}$. The final feature representation is constructed by concatenating all 88 features $f_1 - f_{88}$.

5) *Curvelet features [84]*: These are intermediate-level image features, extracted from the curvelet image transform. They capture regularities arising in low-level NSS models in a localized way, and consequently capture perceptual image distortions in a content independent way. The input image is divided into blocks of size 256×256 , and curvelet features [84] are extracted from each block, yielding a set of feature vectors. Then, the mean feature vectors are calculated to create the final feature vector f . During the curvelet transform from each block 5 layers of curvelet coefficients can be obtained at 5 different scales. AGGD fitting (Eqn (2.13)) is deployed over the fine scale to create a 4-dimensional feature vector $f_{CNSS} = (\gamma, \mu, \sigma_l, \sigma_r)$, consisting of amplitude (γ), mean (μ), and standard deviations σ_l, σ_r .

Curvelet transform provides 64 orientation information, which can be divided into two halves. The mean magnitude of the coefficient of the prior 32 orientation matrices provides the orientation energy. 2 peaks occur near the cardinal (horizontal and vertical) orientations. The mean kurtosis mk around these cardinal peaks is used as a feature. The coefficient of variation of the non-cardinal orientation energy, $cv = \sigma_{so}/\mu_{so}$ is another feature, where μ_{so} and σ_{so} are the sample mean and standard deviation of the non-cardinal orientation energies.

The mean of the logarithm of the magnitude of the curvelet coefficients in all scales are utilized as scalar energy measure to calculate the energy differences between the adjacent layers and interval layers. Energy difference can be calculated as, $e_j = E(\log_{10}(|\theta_j|))$, where θ is a set of coefficients of the orientation matrix, and $j = 1, 2, 3, 4, 5$. This provides a six-dimensional feature group that describe the scalar energy distribution $f_{SED} = (d_1, d_2, d_3, d_4, d_5, d_6)$, where $d_1 = e_5 - e_4, d_2 = e_4 - e_3, d_3 = e_3 - e_2, d_4 = e_2 - e_1, d_5 = e_5 - e_3, d_6 = e_4 - e_2$. Thus, the final feature vector (f) is concatenation of all these features, $f = (f_{CNSS}, f_{SED}, mk, cv)$ of size 12.

6) *NIQE features [85]*: Natural Image Quality Evaluator (NIQE) features are based on the construction of a quality aware collection of statistical measures. The features used are similar to those used in BRISQUE [80]. To extract NIQE features, the input image is partitioned into 96×96 image patches. Specific NSS features are then computed from the coefficients of each patch. The

4 parameters $(\gamma, \beta_l, \beta_r, \eta)$, from Eqn. (2.13), (2.14) and (2.17), respectively, are computed along the four orientations which yields 16 parameters. Combined with the γ and $(\beta_l + \beta_r)/2$ computed from original coefficients, it yields 18 overall features. All features are computed at two scales to capture multiscale behavior, by low pass filtering and downsampling by a factor of 2, yielding a final feature set of size 36, extracted from each patch.

7) *TMIQA features [86]*: Topic model based Image Quality Assessment (TMIQA) uses the NSS features introduced in BRISQUE [80]. Given an input image, it is at first divided into overlapping patches of size 64×64 , with an overlap of 8×8 between neighboring patches, and local BRISQUE features are computed from each patch. This gives a set of 36 features per patch.

2.5.5 Other features

Several researchers have used deep features for FIQA. Liu et al. [8] used features extracted from a pretrained VGG model [87]. Best-Rowden and Jain [61] used ConvNet CNN, and Yu et al. [62] used LightCNN, both trained on CASIA-Webface containing 10,575 subjects of 494,414 images (an average of about 47 images per person). For a deep face model to be effective, it requires a large number of images, with varied images per subject, and these images are usually different qualities. Datasets, such as CASIA-Webface, that are constructed using images in the wild do not have the same quality images per subject. Moreover, the architectures of deep networks need intra-face variations. So it seems contradictory to use a deep neural network face model to extract features for assessing the face image quality, from a theoretical point of view. Recently, Na and Guo [88] found that the deep features have the difficulty in matching face images with large quality gap (in a quality score range of 0-100, score difference > 60), but matching with smaller gaps can get very high accuracies. This reduced sensitivity to quality changes make deep features less preferable for learning the quality differences.

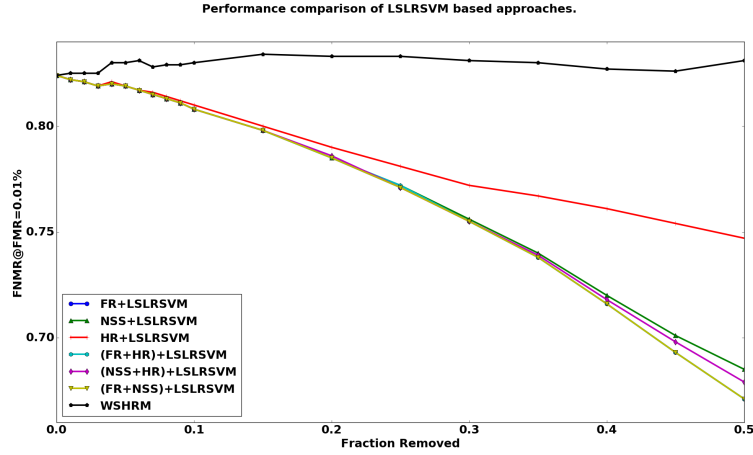
2.6 Evaluation results

In this section, we discuss about the evaluation settings, present evaluation results, detail the observations and discuss the significance of the comparative study of the FIQA methods.

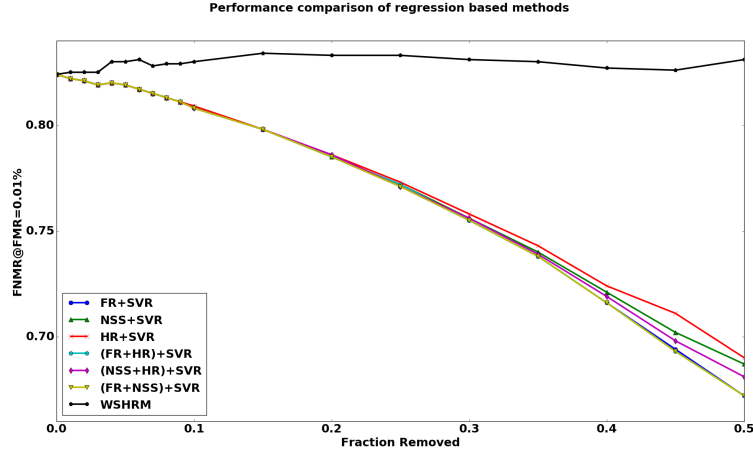
Rank labels for classification and regression: To train the regressors and classifiers, at first LSLRSVM based methods are trained and performance comparison is made using the test set. Then the best method is used to generate training label for the SVR and SVM based FIQA methods. To train the classifiers, the LSLRSVM predicted scores (ranging 0-100) are converted into 10 classes $\{1, \dots, 10\}$, such that score 0-10 is converted to label 1, 11-20 converted to label 2 and so on.

Gallery and probe sets: For each subject, the image with highest quality is manually selected for the gallery set, and, rest of the images are assigned to the probe set. This results a probe set of 531,311 images and a gallery set containing 14,374 images (equal to the number of available subjects).

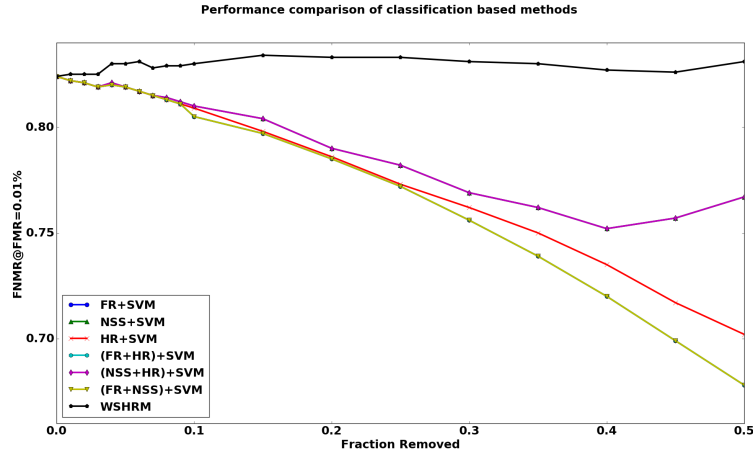
Training and test sets: The probe set is split in 25%-50%-25% ratio, with no overlap between training and test sets. In order to make sure that the training set has all quality variations, the probe set was split into 10 quality groups and then from each group images were taken from randomly



(a) Error vs. Reject curves for LSLRSVM based methods.



(b) Error vs. Reject curves for SVR based methods.



(c) Error vs. Reject curves for SVM based methods.

Figure 2.2: Error vs. Reject curves of the all face image quality assessment methods used in our experiment. X-axis indicates the percentage of low quality images rejected. Y-axis indicates FNMR@FMR=0.01%

selected non-overlapping subjects. The first training set is used to train the LSLRSVMs using the relevance levels. Then, those trained LSLRSVMs are used to generate training label for the second set. This second training set is used to train the regressor and classifiers.

Selection of face recognition method: There has to be a trade-off between the strength of the face recognition method and its ability to assess quality variance. The stronger the algorithm, the better is the recognition of low quality face images. Therefore, a robust and advanced face recognition method, such as those using deep features, can be counter productive for the quality assessor. The face recognition method should be sensitive enough to produce high variance in match rank. Na and Guo [88] have tested using 4 different types of popular deep face features and found that unless there is a very large quality gap between the probe and gallery set (quality score difference > 30 in a $0 - 100$ range) then the quality variance does not significantly affect the recognition performance. Moreover, most commercial off-the-shelf (COTS) face recognition systems still uses traditional face recognition features, so if we want to train FIQA models that can accurately assess the quality of the face images for majority of the COTS systems, it is more preferable to use the similar traditional features. Similar approaches have been taken by other researchers, e.g., Bharadwaj et al. [45] and Best-Rowden and Jain [61] used match scores from two commercial off-the-shelf (COTS) face recognition systems instead of choosing the state-of-the-art. Vignesh et al. [9] used HoG and LBP feature based face recognition methods. We used the fusion of the four different face recognition features, i.e, HoG, Gabor, LBP and Gist for face matching. Cosine distance is used to find the match score.

Score level fusion: In order to investigate the effect of the different feature combinations, we applied score fusion to the predicted scores to generate the final score. For LSLRSVM and SVR, this was done by averaging the predicted scores. For SVM, majority vote was used to select the final predicted class. The different types of FR and NSS features are individually combined using this strategy. Moreover, FR, HR and NSS features are also combined, which are denoted as FR+NSS and FR+HR.

2.6.1 EvR based evaluation: Observations and findings

Error versus Reject (EvR) curves evaluate the efficiency of rejecting low quality samples for reducing error rates. EvR curve plots an error rate (FNMR or FMR) versus the fraction of images removed/rejected, where the error rates are computed using a fixed threshold (e.g. $FMR=0.01\%$). We measure performance based on how well (in percentage) the method reduces FNMR after 50% of low quality images are rejected. Based on this evaluation criteria, from Figure 2.2 several important observations have been made about the representative FIQA methods:

- 1) For LSLRSVM based methods we observed that, FR+LSLRSVM is top performing, followed by NSS+LSLRSVM and HR+LSLRSVM. Same is true for their SVR based counterparts. Though for SVM, NSS+SVM performs poorly compared to HR+SVM. This observation indicate that FR features are most favorable for FIQA, and for classification scenario NSS features could be more discriminative than the HR features.
- 2) (FR+HR)+LSLRSVM and (FR+NSS)+LSLRSVM do not improve upon the performance of the FR+LSLRSVM based method. Which is also true for SVR and SVM based counterparts. This further indicates that fusing FR feature with NSS and HR feature does not improve the performance.
- 3) Fusion (NSS+HR)+LSLRSVM performs better than the HR+LSLRSVM and NSS+LSLRSVM

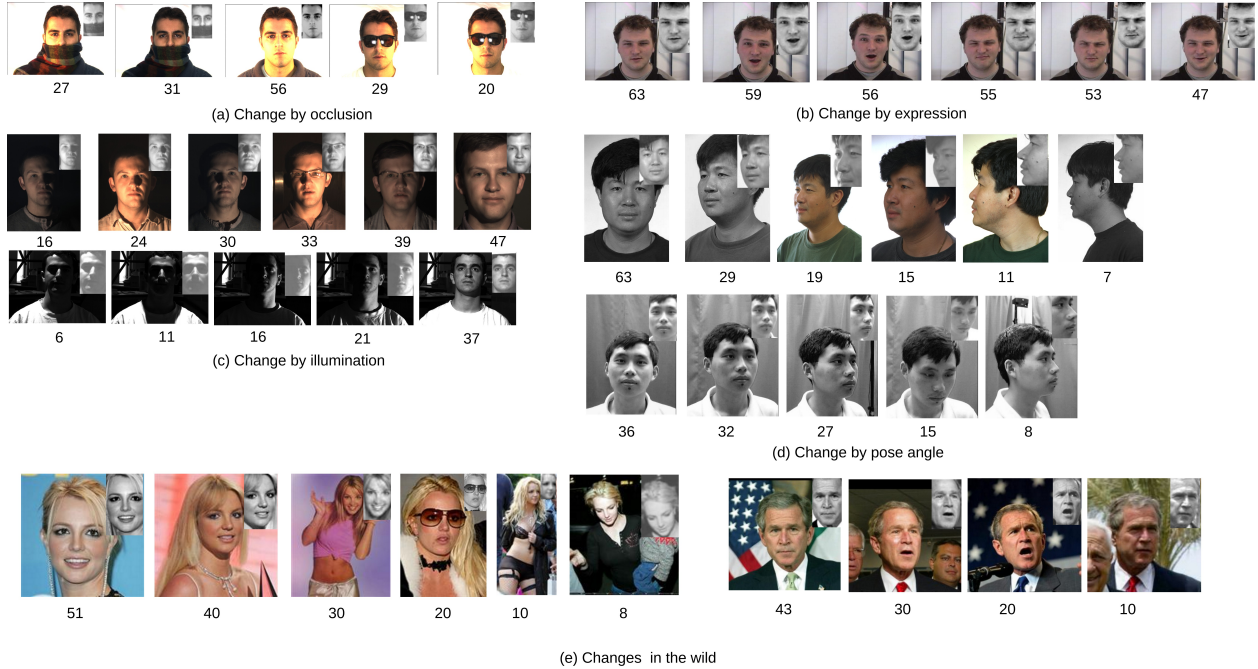


Figure 2.3: Demonstrating the effect of different types of biometric quality degradation on the predicted scores. Cropped faces are shown inset. Quality scores are generated by (FR+HR)+SVR method

individually. Which is also true for (NSS+HR)+SVR and HR+SVR. Though these methods are not top performing their feature size is very small compared to FR feature based method, which can be useful when computational cost is very high.

4) For Figure 2.2(a) and (b) we can see, HR+SVR performs slightly better than HR+LSLRSVM. This may be indicative of SVR has generalized well when using LSLRSVM generated training labels to train the model. Though the better performing FR features using SVR and SVM, do not improve upon their LSLRSVM counterparts, more sophisticated models using the rank scores as training labels with FR features may be able to achieve that.

5) Though HR+LSLRSVM, HR+SVR and HR+SVM reduces FNMR close to 15% at 50% reject, when the same 14 heuristic measures are used in a traditional approach (WSHRM) there is no FNMR reduction. This provides evidence that learning based methods are more effective than the traditional quality measure fusion based approach.

From the Figure 2.2 we can observe that, the range of FNMR drops from 0.824 to at most 0.672 (difference 0.152). At 50% removal, most of the learning based methods reduces FNMR from 12% to 15%. This narrow difference in rate drop indicate that for benchmarking FIQA methods on a large dataset, there is a scope of using more effective evaluation strategy that can show more information to better compare the FIQA methods. Note that, we could have gone to reject more images, but as discussed by Grother and Tabassi [5] rejecting such large percentage of images from the test data is not a operational possibility. In following section we use a quality bin versus face recognition rate based evaluation method that may provide further insight.

2.6.2 Bin vs Recognition rate: Evaluation measures

It has been widely established that biometric sample quality is defined as a measure of a sample's utility to automatic matching [2–5]. Assessment of face image quality should be an indicator of face recognition performance. In this work, we evaluate the different FIQA methods based on the relation between quality level and recognition performance. We divide the test images based on their predicted scores into different quality bins (or levels), and use these bins (or quality levels) as probe sets and match them with the previously constructed gallery. The correlation between binned predicted scores of a FIQA method and the recognition accuracy of the corresponding face matching algorithm should produce a increasing monotonic function, such that higher values of quality, correspond to higher similarity scores for the same subject. This approach is similar to Chen et al's [50] evaluation using identification accuracy w.r.t. face quality ordering and bin to rank-1 identification approach taken in [45], but we further incorporate three evaluation measures to assess performance among different FIQA methods. Note that, in our case, instead of rank-1 we perform rank-100 recognition in order to get a higher than 0% recognition rate in the lower quality bins so that a meaningful curve is obtained. The three evaluation criteria are described in following:

1) *Linear correlation (ρ)*: For the FIQA method to closely predict the recognition performance, there has to be a strong correlation between the predicted quality scores and recognition rates. Therefore, if the quality scores are binned by partitioning them into equal intervals, the face images in high quality bins should perform better than the lower quality ones. In ideal scenario the bin numbers and their corresponding identification rate should form a linearly increasing relation between the two lists demonstrating maximum separation of the identification performance between consecutive quality bins. Pearson's correlation coefficient (ρ) can provide us the measure of the relation, which is defined as:

$$\rho(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_x \sigma_y}, \quad (2.23)$$

where X and Y are two variables and σ_x and σ_y are their standard deviations, respectively.

2) *Nonmonotonicity (N)*: It is reasonable to assume that a FIQA and the corresponding face matching algorithm are bound by an increasing monotonic function, such that higher values of quality, correspond to higher recognition rate. Therefore, higher quality bins should produce higher accuracy than the lower quality ones. We measure nonmonotonicity using the Kendall tau rank distance, this metric can give us a count of the number of pairwise disagreements between the bin ranking and their corresponding identification rates. Larger distance signifies more order-wise dissimilarity between the two lists. This metric is also known as the bubble-sort distance, since it is equivalent to the number of swaps that the bubble sort algorithm would require to place one list in the same order as the other list. Kendall tau ranking distance N between two lists τ_1 and τ_2 is defined as follows [89]:

$$N(\tau_1, \tau_2) = |\{(i, j) : i < j, (\tau_1(i) < \tau_1(j) \wedge \tau_2(i) > \tau_2(j)) \vee (\tau_1(i) > \tau_1(j) \wedge \tau_2(i) < \tau_2(j))\}| \quad (2.24)$$

where $\tau_1(i)$ and $\tau_2(i)$ are the rankings of the element i in τ_1 and τ_2 respectively and $|\tau_1| = |\tau_2| = n$. $N(\tau_1, \tau_2)$ will be equal to 0 if the two lists are identical in order and $n(n-1)/2$ if one list is the reverse of the other.

3) *Uniformity (U)*: The training and test sets were constructed by randomly sampling equal number of images from 10 quality bins. And after prediction, again 10 quality bins were created by partitioning the test set images according to their predicted score. We argue that this newly binned

test images also need to show a tendency towards equal distribution among all bins. A method that cannot differentiate the quality well, will have the tendency to clump together all the images to few number of bins and thereby produce a high variance of number of images in bins, whereas, a good quality assessor will produce a uniform distribution (low variance) of images among the bins. We define uniformity of the bins U as:

$$U = \frac{\sum_{i=1}^{|B|} |b_i - \bar{b}|}{\sum_{i=1}^{|B|} b_i}, \quad (2.25)$$

where $B = \{b_1, b_2, \dots, b_n\}$, b_i denotes the number of images in the i -th bin and $|B| = n$. U is 0 when all the bins have same number of images.

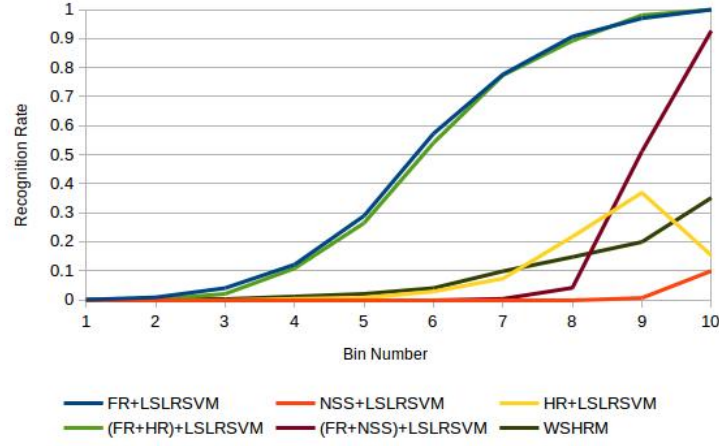
2.6.3 Bin vs Recognition rate: Observations and findings

From Table 2.2 and Figure 2.4, the following observations were made:

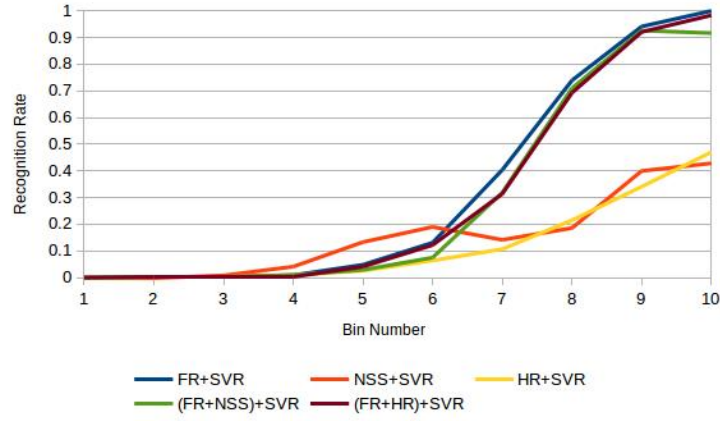
- 1) 6 out of the 15 learning-based methods show linear correlations $\rho \geq 0.9$. In descending order of correlation they can be represented as follows: FR+LSLR SVM (0.971) > (FR+HR)+LSLR SVM (0.969) > NSS+SVR (0.927) > FR+SVM (0.916) > FR+SVR (0.910) > (FR+HR)+SVR (0.900).
- 2) Among the 6 methods with $\rho \geq 0.9$, 4 of them have a correlation curve which is increasingly monotonic ($N = 0$). They are: FR+LSLR SVM, (FR+HR)+LSLR SVM, FR+SVR, (FR+HR)+SVR
- 3) Out of the 4 methods with ($\rho \geq 0.9, N = 0$), 2 of them have uniformity measure, $U < 1.0$. In ascending order of bin uniformity, they are: (FR+HR)+SVR (0.854) < FR+SVR (0.932).
- 4) 9 out of the 15 learning based methods have better bin-to-recognition rate correlation than the traditional approach WSHRM. Also, the traditional approach fails to produce a monotonic correlation curve, where 7 out of the 15 learning based methods produce monotonic correlations.
- 5) Based on the combined assessment of the 3 evaluation metrics, it can be concluded that (FR+HR)+SVR ($\rho = 0.900, N = 0, U = 0.854$) FR+SVR ($\rho = 0.910, N = 0, U = 0.932$) are the top performing methods among all 15 learning based methods evaluated. Also, all top performing methods used FR feature to get their results.
- 6) Table 2.2 shows all the relative rank label generation methods for training the classifiers and regressors. Both of the top performing methods were trained using labels generated by FR+LSLR SVM and (FR+HR)+LSLR SVM, respectively. They are the two methods that also show high bin number-to-identification rate correlation with $\rho > 0.9$ and nonmonotonicity $N = 0$, with FR+LSLR SVM showing highest correlation measures. This indicates the effectiveness of FR+LSLR SVM generated quality labels for training the learning based FIQA methods.

2.6.4 Observation from inspecting sample face images

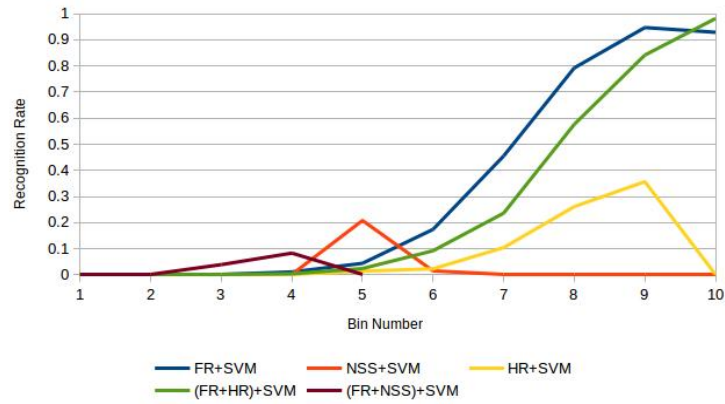
In Figure 2.3(b) it can be observed how expression change affecting the quality score for a subject whose images were taken in a controlled environment with just expression variation. It can be observed that the neutral face has the highest score which deteriorates as the expression changes. This shows that the method has been able to capture the affect of expression change on quality. Figure 2.3(d) shows the affect of pose change on quality on two different subjects. Even though frontal faces have high scores, they drastically reduce as the pose angle increases. Figure 2.3(c)



(a) Correlation plot for LSLRSVM based methods.



(b) Correlation plot SVR based methods.



(c) Correlation plot for SVM based methods.

Figure 2.4: Correlation plot of all face image quality assessment methods investigated in this study, grouped by feature type. X-axis indicates the bin number (quality level). Y-axis indicates the rank-100 identification rate for each bin (quality level).

Table 2.2: Results for FIQA methods.

FIQA Method	ρ	N	U
FR+LSLR SVM	0.971	0	1.056
NSS+LSLR SVM	0.512	0	1.600
HR+LSLR SVM	0.774	2	0.900
(FR+NSS)+LSLR SVM	0.659	0	1.473
(FR+HR)+LSLR SVM	0.969	0	1.054
FR+SVR	0.910	0	0.932
NSS+SVR	0.927	2	1.163
HR+SVR	0.886	1	0.680
(FR+NSS)+SVR	0.889	1	1.020
(FR+HR)+SVR	0.900	0	0.854
FR+SVM	0.916	1	0.978
NSS+SVM	-0.113	9	1.412
HR+SVM	0.586	7	0.911
(FR+NSS)+SVM	-0.108	12	1.587
(FR+HR)+SVM	0.884	0	0.924
WSHRM	0.879	1	0.518

shows the affect on quality score for low or uneven illumination for two different subjects. Right most image has the highest score with comparatively better illumination of the face region, while other faces either have partial illumination or very low illumination, which is reflected in the quality scores. Figure 2.3(a) demonstrates that the FIQA method has been able to capture quality variance due to occlusion such as sun-glass or scarf etc. It can also be observed that, illumination variation also affects over occlusion and reduce the score in the right most two, and left most two images. This is an example of combined quality attribute change and this method successfully captures this. In real world scenario, where images are taken from the wild, faces are affected by multiple issues of quality distortion simultaneously, such as pose, expression, illumination changes occurring for the same face image. Figure 2.3(e) shows two examples of this using two different subjects, for the subject in the top row, it can be seen that, face size, sun-glass and pose variation affecting the quality. For the subject in the bottom row, it is shown how pose, illumination and expression change affecting the quality score. It can be noticed that, the quality score reflecting those deteriorations.

2.6.5 Discussion

From both EvR and Bin vs Recognition rate based evaluation results, it can be observed that, the top performing methods use face recognition features to get high correlation. One reason for this can be face recognition features contain much more information regarding facial characteristics than the heuristic measures. Natural scene statistics can predict image distortion, but not all aspects of biometric quality is affected by image quality distortion. Poor performance of heuristic features

confirmed the previously stated argument that the traditional quality measures about face image quality is not optimally predictive of the recognition performance.

Observations made on Table 2.2 show that the semi-supervised approach using ranking models can improve performance of FIQA methods. This is because relative ranks are devoid of the problem of “contrary images”, where face same image can produce high and low match scores with different images of the same subject [67]. It is also evident that, as a quality label generation scheme, relative ranking can perform better than traditional approach.

The weighted sum of heuristic measures (WSHRM) method was outperformed by most of the learning-based methods. Moreover, learning based approaches using heuristic features outperforms the traditional approach using the same set of features. This shows the efficacy of the learning-based approaches over the traditional approach.

2.7 Conclusion

Biometric quality assessment for face images is quite a challenging topic. In this chapter, a large scale database, solely for studying face image quality assessment is introduced, a learning to rank based approach for quality label generation is presented, and a comparative evaluation of different representative face quality assessment methods, using commonly used feature types, is performed to demonstrate the usefulness of this database and the training protocol. The comparative study has shown that, the ranking based quality scores can help improve results, especially for, regressor based methods when used with face recognition features. Investigation on the predicted quality scores by the top performing method show that the method can capture various face quality changes such as illumination, pose, occlusion, expression etc. With this new database, we wish to encourage the research community to further investigate the challenges of face image quality assessment. We believe that, by having a platform for benchmarking the state-of-the-art will move community further in face biometric quality research.

Chapter 3

Understandable Face Image Quality

3.1 Introduction

Face recognition performance is affected significantly by the face image quality, especially in real-world applications [90–92], where images are taken from unconstrained environment. Face image quality varies significantly because of the use of different imaging sensors, compression techniques, video frames, and image acquisition conditions. It is very challenging to assess face image qualities automatically, quickly and precisely in real world images.

There have been a lot of learning-based face image quality assessment methods proposed in last few years. These approaches usually provide a single quality score (or a quality bin label) as the output. However, this “single-value quality score” cannot provide enough information to communicate effectively with the human assessors. Furthermore, many issues regarding face image quality still have not been addressed yet, such as, what does a quality score mean? How to interpret a quality score with imaging conditions? Why a face image has a quality score of 50 rather than 60? How well the quality scores characterize the real face image qualities? Can more useful cues (e.g., levels of details) be acquired to develop a complete representation for face image quality assessment? In this chapter, a new paradigm is investigated which provides human understandable information for face image quality assessment. It is motivated by the understandable template of FBI, where some detailed information of faces is included during face template extraction. The proposed new face image quality representation has the potential to be integrated into the FBI’s understandable template. Following list of contributions are made in this chapter:

- Define the understandable face image quality (UFIQ) paradigm, and how a mapping from score summary to heuristic attributes can provide understanding regarding quality change.
- Establish the understandable face quality method using the help of statistical measures.
- Provide experimental evaluation of understandable face image quality.

3.2 Related Works

In last few decades, quite a few learning based face image quality assessment methods have been proposed by researchers. Vatsa et al. [55] proposed a biometric image quality assessment algorithm that uses redundant discrete wavelet transform to compute the approximation of horizontal, vertical and diagonal bands, which are used as quality factors. These quality factors are combined using a weighted sum to get the quality score and SVM based multiclass classification is performed to determine the class of the score. Liao et al. [44] proposed a face image quality assessment scheme that employs a hierarchical binary decision tree classifier based on support vector machines (SVM) to categorize the face images into five quality levels. Bharadwaj et al. [45] used Gist and sparsely pooled Histogram of Oriented Gradient (HoG) features and a one-vs-all multi-class SVM to classify face images into four different quality categories. Bhatt et al. [46, 47] presented a quality assessment algorithm which computes a quality vector comprising no-reference quality, edge spread, spectral energy, and pose, and then trained Support Vector Machines (SVM) for decision making. Ozay et al. [48] proposed a unified probabilistic framework to simultaneously predict the quality of the facial image samples and perform quality-based face recognition. Wong et al. [7] proposed a patch-based face image quality assessment algorithm which quantifies the similarity of a face image to a probabilistic face model. El-Abed et al. [49] used no-reference based image quality metric called BLIINDS, SIFT keypoints, DC coefficient, and, mean and standard deviation of scales as features and then used a SVM to predict the quality. Chen et al. [50] proposed a learning to rank based framework for assessing the face image quality. Kim et al. [51] proposed a learned FIQA method that considers visual quality and mismatch between training and test face images for quality assessment.

Recently, deep learning based quality assessment methods have also been introduced by different researchers. Liu et al. [8] proposed a no-reference face image assessment algorithm based on the deep features extracted from VGG network [93]. Vignesh et al. [9] proposed a FIQA algorithm which is based on mimicking the recognition capability of a given face recognition algorithm by using a Convolutional Neural Network (CNN). Yu et al. [94] proposed a novel FIQA method where five common homogeneous distortion categories in video surveillance applications were considered. A lightweight CNN was trained to simultaneously predict the categories and degrees of the degradation in a face image.

3.3 The understandable face image quality method

3.3.1 What is Understandable face image quality ?

From the early days of face image quality assessment research, researchers have tried to identify a particular set of quality metrics that can effectively measure the face image quality and tried to set an optimal range of values for these metrics [34]. Several standards are available that describe what a top quality face image should consist of. E.g. what the value of the brightness should be, what is the optimum intra-ocular distance, optimal range of head pose etc. [95, 96]. All these estimations are based on human visual perception. Because human visual processing is still considered to perform the best for face identification, researchers considered that expert human assessors should be able to provide the best set of measures and their optimal range required for top quality face

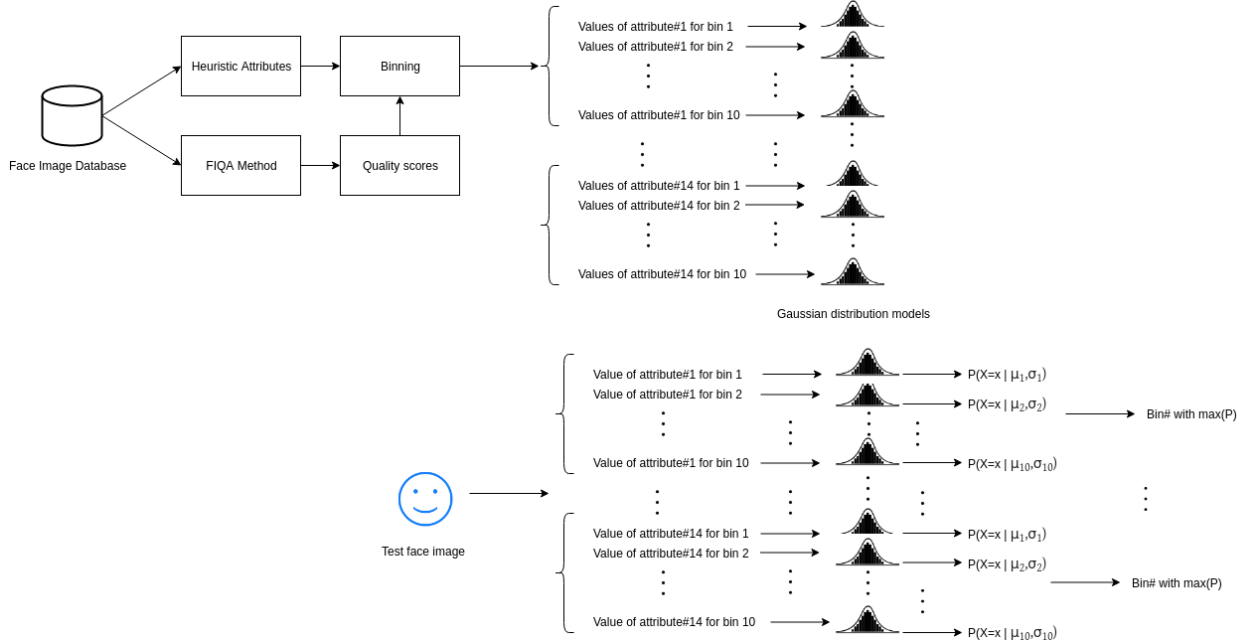


Figure 3.1: Example of obtaining understandable information for a test image.

images. But later on, it was found that, because biometric quality of face image is defined in terms of automatic face recognition performance, human visual perception of image quality may not be well correlated with recognition performance [2, 3, 7]. Investigation were made on finding learning based methods [7, 9, 49, 50, 55]. Instead of identifying the quality degradation factors, these methods provided either a quality score or a quality bin label to establish the face image quality. But since these methods are based on image features the reason for their performance went hidden from human understanding. Moreover, most image quality assessment techniques strive to achieve the perceptual understanding of the image, and therefore consider human annotation of quality as the gold standard for comparison and testing of automated face image quality assessment algorithms. A high correlation between the predicted quality and the human measure of quality is considered as good assessor of image quality. But, there is no conclusive evidence that human interpretation of quality has any correlation with the biometric quality in terms of performance of a face recognition method [3]. For these reasons, face image quality has become almost an abstract concept.

In this work, it is proposed that, for any reliable score summary producing learning based FIQA method, it is possible to trace back to a set of quality measures and find out what quality measures and what range of values of these measures, are optimal. By trying to find out the statistical distribution for each of the heuristic attributes for a number of quality levels, this method gives back that “understandability” that was lacking due to the use of learning based face image quality assessment. But, in order to do that, at first the relationship between each of the quality measures (or heuristic attributes) and the predicted score of the FIQA method in consideration has to be mapped. Using a carefully crafted database consisting of wide ranging quality, this relationship between a quality measure and a quality score can be learned. By using a simple statistical model, it is possible to obtain a stochastic measure of the heuristic attribute for all quality levels and then we can categorize the values in a meaningful way that can clearly communicate what the value of the quality measure mean in terms of the predicted quality score. By providing such information for all the quality metrics, we believe, it is possible to provide an overall picture for at least some

of the major reasons behind the face image achieving predicted quality score to a certain extent.

Based on the above description, the different stages of the method can be divided as following: 1) Construct a large scale face database with enough width (in terms of number of subjects) and depth (in terms of number of images per subject) and with wide ranging quality variance. 2) Select a set of quality measures that are commonly used for FIQA and can be directly extracted from face images, 3) Select a reliable learning based FIQA method whose predicted scores will be investigated. 5) Generate quality scores using the learning based FIQA method for all images of the database. 6) Obtain the heuristic features for all images of the database, 7) Apply a statistical analysis to obtain distribution models, 8) Using these statistical models, for a test image, we obtain a probabilistic decision for the quality measures. This decision tells us which quality level the quality measure is highest probable to belong. Using this we can produce a helpful picture for a human observer to understand whether the quality measure is degraded or not. 9) Finally, through aggregating all these categorized quality measure values, a human understandable information about the quality score is provided. Figure 3.1 provides an overview of this process. More details about developing the UFIQ method is provided in the following sections.



Figure 3.2: An example face image consisting of composite biometric quality. Which quality measure(s) has ultimately caused the quality score to be 5?

3.3.2 Database description

In the following sections details about how to create a suitable database for learning UFIQ is provided.

For a face image with low quality score, the problem of identifying the affected quality measure (or measures) for the non-optimal quality score is a hard problem. Because all the quality measures in the low quality bin may not be responsible the low quality range. One quality metric with low quality value is enough to change the low quality score, e.g., a face image with optimal

pose angle, sharpness, no occlusion can still have low quality score if the brightness is too low or too high, which will hamper the face recognition performance. One approach to ensure the collection of right value range for a quality measure, is to create or obtain a database that has collected the quality scores for every quality measure variation possible.

Currently, there is no readily available public database that can satisfy this requirement. So, any one of the following two choices can be taken: 1) to synthetically change the quality measure of a high quality image to create images at different quality levels and collect subjects with largest possible variation in facial variations (e.g. pose, expression etc). But the problem is you can only generate a limited number of images per subject. Moreover, not for all cases it is possible to produce synthesized faces without altering the facial identity, such as expression and pose. The other option is, 2) to construct a large database by assembling datasets constructed both in controlled and real-world scenario such that a wide range of face quality can be obtained. It is reasonable to assume that by aggregating datasets from different sources where image acquisition process had different environmental parameters and by making the assembled database large enough, we can achieve a large variance in face quality. And by analyzing a large database like that a general trend for the heuristic attributes at different quality levels can be identified. This is not an optimal solution, but it is something we can work on.

If we study the publicly available face databases, we can find basically two groups of datasets: 1) constructed with sets of images taken in a specific number of different environmental conditions such as, lighting, expression, accessories, pose, indoor/outdoor etc. Each of these variations are usually grouped into data subsets. 2) constructed using images collected from the internet, where images are taken in unconstrained environments, and then grouped according to each identity. For our database, we can utilize both of these scenario by regrouping the controlled (or semi-controlled) subsets into each subject, which gives us subjects with varied face image quality. For, datasets containing real-world images, the main concern was to avoid noisy images as much as possible. Since, no data-driven cleaning approach were applied for these datasets, manual cleaning was done for some of the comparatively larger real-world datasets as much as possible.

Based on the above mentioned criteria, a database, named “Face Image Quality Database (FIQDB)” is assembled. We selected 40 public databases from the list of face databases available in the face recognition homepage [1]. The database contains images taken in controlled, semi-constrained and unconstrained real-world scenarios. The resulted FIQDB database has a total number of 545,684 images of 14,373 subjects.

How to make sure the database is large enough to contain all possible qualities ? It is not a trivial problem because there is no universal consensus about all factors that affect face image quality. We make a simplistic assumption that by applying a reliable FIQA method on the dataset and inspecting whether the generated scores cover all possible quality scores produced by the method can ensure it to some extent. To demonstrate this a third party learning based face image quality assessment method [50] is applied on FIQDB. The FIQA method provides quality scores ranging from 0 to 100. Figure 3.3 shows some sample images with their corresponding quality scores. It can be noticed that, the images have scores as low as 2 to as high as 97 which indicates that the database has a wide quality range.



Figure 3.3: Sample images from the database with predicted scores, shown in the top-right corner, demonstrating the variety of face image qualities.

3.3.3 The quality measures

A major limitation of using quality measures as a performance prediction feature is that there are an overwhelmingly large number of quality factors that can influence the performance of a face recognition system and their exact count is still unknown. Furthermore, accurate measurement of image quality measures is still an unsolved problem. But the goal of this chapter is not to evaluate the quality using these quality measures, rather to provide a kind of understandability to the predicted score. Very few studies have been done on comparative assessment of different measuring method of the quality measures. Abaza et al [14,15] did a investigation on the image quality measures and found several useful formulas. Some of these measures are selected for this work and add additional measures that have been used for traditional face quality assessment. Description of all these quality measures are given below -

Brightness

Mean of intensity has been a common way to measure brightness found in several traditional face quality assessment. The average value of the illumination component of all of the pixels in the face region can be considered as the brightness of that region. Therefore, for a grayscale image $I(x, y)$, brightness B is defined as [15, 27, 31]:

$$B = \frac{1}{MN} \sum_{x=1}^M \sum_{y=1}^N I(x, y) \quad (3.1)$$

Sharpness

The sharpness measure, denoted by S , is computed by [27]:

$$S = \frac{1}{MN} \sum_{x=0}^M \sum_{y=0}^N \left(|I(x, y) - H(I(x, y))| \right) \quad (3.2)$$

where $I(x, y)$ is the intensity of the image I at location (x, y) , $H(\cdot)$ is a lowpass Gaussian filter, and $M \times N$ is the size of the image.

Focus

Edge density measures the average magnitude of the gradient over the face image. The assumption is that when images are in focus the average gradient magnitude will be higher than when the image is out of focus [3, 70–72].

Contrast

Image contrast is the difference in color intensities that makes a face distinguishable. The face image contrast C can be measured using the following equation [15]

$$C = \sqrt{\frac{\sum_{x=1}^M \sum_{y=1}^N [I(x, y) - \mu]^2}{MN}} \quad (3.3)$$

where μ is the mean of image I .

Illumination

Spectral energy is used as the illumination measure, which describes abrupt changes in illumination and specular reflection. The image is tessellated into several non-overlapping blocks and the spectral energy is computed for each block. The value is computed as the magnitude of Fourier transform components in both horizontal and vertical directions [13, 46].

Illumination Symmetry:

This is calculated as the absolute difference between the mean intensity of left and right sides of the face image.

Compression

A no-reference perceptual quality assessment method for JPEG compressed images proposed by Wang et al. [97] was used to measure the compression quality. This method is able to effectively capture the artifacts introduced by JPEG compression.

Eye openness

The Openface toolkit [98] was used for this measure, Figure 3.4 shows the detected 68 landmarks. It detects 6 points for each eye region, these points are used to calculate the convex hull. The mean of the two areas is considered as the eye openness measure.

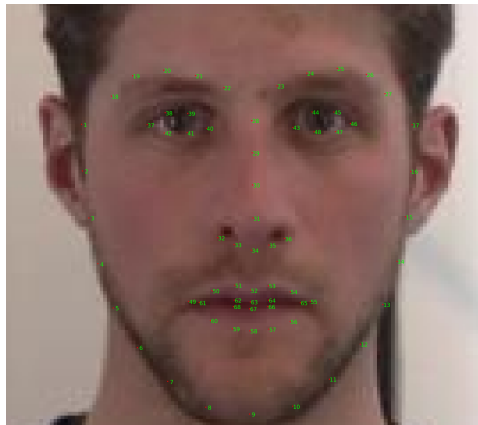


Figure 3.4: Face image with 68 face landmarks detected using Openface toolkit.

Mouth closeness

Openface toolkit [98] detects 8 points around the boundary of the inner lip region. Similar to eye openness measure, the area of the convex hull of the region is calculated by connecting these points, which is used as the measure of the mouth closeness.

Pose measure: Roll, Yaw and pitch

a headpose angle detection method provided by Openface was used. This provides three different Euler angles for pitch, yaw and roll, $euler_z$, $euler_y$ and $euler_x$, respectively. The angles are in radian and ranges from $[-\pi, \pi]$.

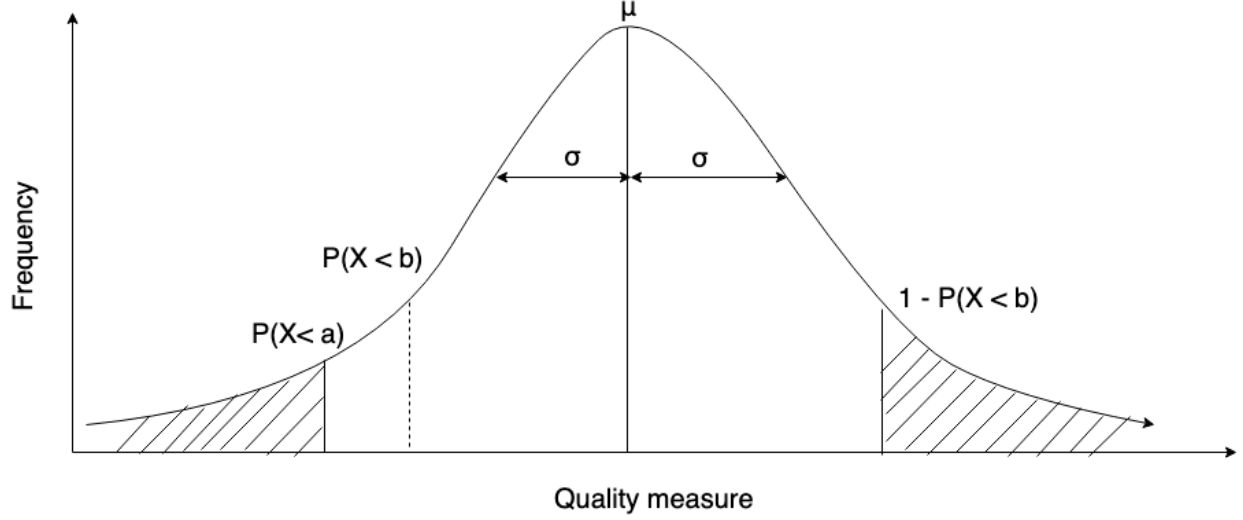


Figure 3.5: We can use area under the gaussian distribution for a quality bin to obtain the probability of any quality measure belong to that distribution.

Face symmetry

The mean difference between the original and horizontally flipped images is considered as the measure of how symmetrical are both sides of the face. The lower the difference, the more symmetrical is the face [25]. This can also assess the pose and illumination variance of the face.

3.3.4 The statistical approach

From our observation of the histogram of the quality bins of each of the heuristic attributes, all of them form the unimodal Gaussian distribution (See Figure 3.6). Therefore, for each heuristic attribute and quality bin pair we can obtain a Gaussian model $\mathcal{N}(x|\mu, \sigma)$. During test, we can use $\mathcal{N}(x|\mu, \sigma)$ to measure the probability of x belonging to any of the quality bin distributions. We know

$$P(X \leq c|\mu, \sigma) = \int_{-\infty}^c \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx \quad (3.4)$$

where μ and σ are the mean and standard deviations; and X is a normally distributed random variable. Because the probabilities $P(X \leq x)$ and $P(X > x)$ span the entire sample space ($-\infty < x < \infty$), therefore, $P(X \leq x) + P(X > x) = 1$. Then, we can also write, $P(X > x) = 1 - P(X \leq x)$.

Then, we can formulate the probability formula of x belonging to a continuous distribution $\mathcal{N}(x|\mu, \sigma)$ as a simple conditional function as below,

$$P(X = x|\mu, \sigma) = \begin{cases} 2 * P(X \leq x|\mu, \sigma), & \text{if } x < \mu \\ 1 - 2 * P(X \leq x|\mu, \sigma), & \text{otherwise} \end{cases} \quad (3.5)$$

Closer the value of x to μ the closer the probability is to 1. In Figure 3.5 we depict how our probabilistic measure is applied for a generic normal distribution.

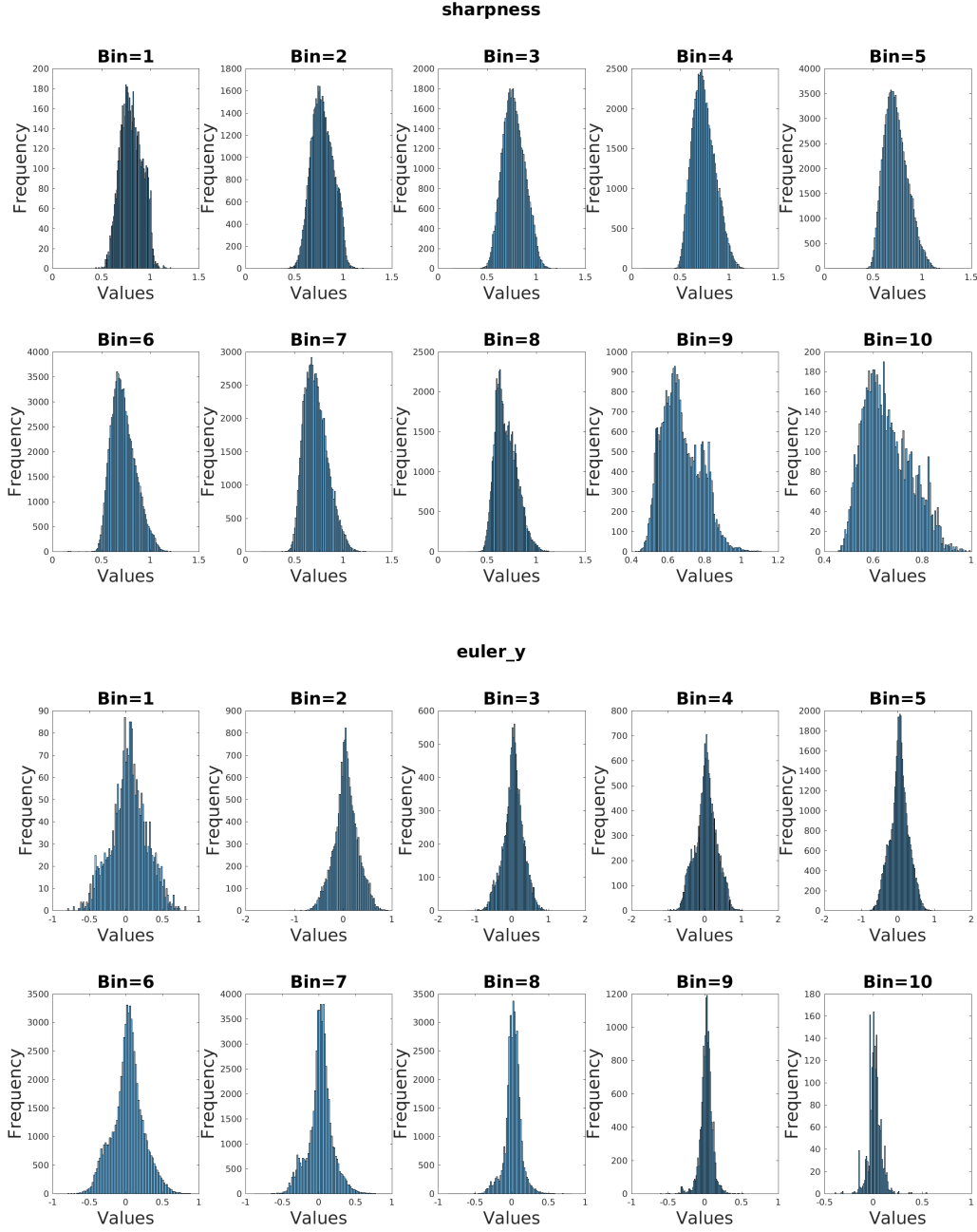


Figure 3.6: Histogram for each quality bins for sharpness measure and head pose yaw measure (euler_y). bin=1 has the lowest quality and bin=10 is the bin with highest quality images.

3.4 Experiments

Using the large database containing wide range of face quality, at first we obtain the quality scores. We selected Chen et al. [50] learning to rank based FIQA method for this work. The authors have generously made their implementation online which provides a score between 0 to 100 for each face, where higher value indicates higher quality. After that, we obtain the set of 14 different heuristic attributes for each image. We divide the images into 10 such that image scoring 0..10 goes to bin 1, image scoring 11..20 goes to bin 2, and so on. This creates 10 quality bins, where each image has 14 different heuristic attribute values. For each attribute for each quality bin, we generate the histograms, so 140 histograms in total. A few of them are shown in Figure 3.6.

Several observations have been made from this histogram plots: 1) All of the distributions following the unimodal Gaussian distribution pattern with one mean peak and symmetric distribution. 2) For illumination symmetry we have a truncated Gaussian distribution because there is no value below 0, but we can still model it using Gaussian function. 3) For eye openness and mouth closeness ratio, there were a lot of default values that we needed to ignore in order to get the original distribution that why the frequency is so low. Same situation for some bins in the pose angle based attributes.

The mean (μ) and standard deviation (σ) from these histograms are used to create individual Gaussian model $\mathcal{N}(x|\mu, \sigma)$ for each quality bin for each heuristic attributes.

Now, from the previous discussion it is evident that, even with a large scale face database with wide ranging quality is not possible to guarantee that each distribution calculated above represents the values responsible for the quality scores in that quality bin. Because of the presence of a composite of different biometric qualities in the face images, it is very difficult to separately identify which quality measure (or measures) has ultimately caused its predicted quality score. But, we can assume that the reasonable distribution of any one of the attributes will have 10 unique mean values, otherwise, it will not be possible to use the statistical measure devised above. In order to obtain unique distributions, we have to consider the distributions not following the order of the bin quality level are the noisy distributions and disregard them during probability calculation. Figure 3.7 shows a mean and standard deviation scatter plot for contrast and focus measure. We can observe, bin 8 and bin 10 are out of order for contrast measure, and, bin 1 is out of order for focus measure. We manually select the maximum number of bin distributions for all attributes that can be kept without destroying the quality bin order.

We kept a 100 randomly selected face image for manually investigating how much the provided information is relevant to the face image. Figure 3.8 provides some sample from the outputs we received for the test set. Each figure shows the original image, estimated quality score and the values of the quality metrics that are found to be sub-optimal. The estimated quality score is calculated using the learning based FIQA mentioned earlier. Note that, the degree of deterioration is not measured in terms of the quality metric or general human perception of that quality metric, but how the quality metric has affected the quality score. In other words, it is mainly describes how far from the optimum range of values is the measure. For some cases where the ranges are very narrow, even a small change in quality metric can cause high deviation in quality score. Moreover, intuitively we might think that all measures in the low quality images will be in low quality range, but that is not the case because low quality of any one or few of the quality metrics are enough to cause a overall quality degradation, and thus a low quality score.

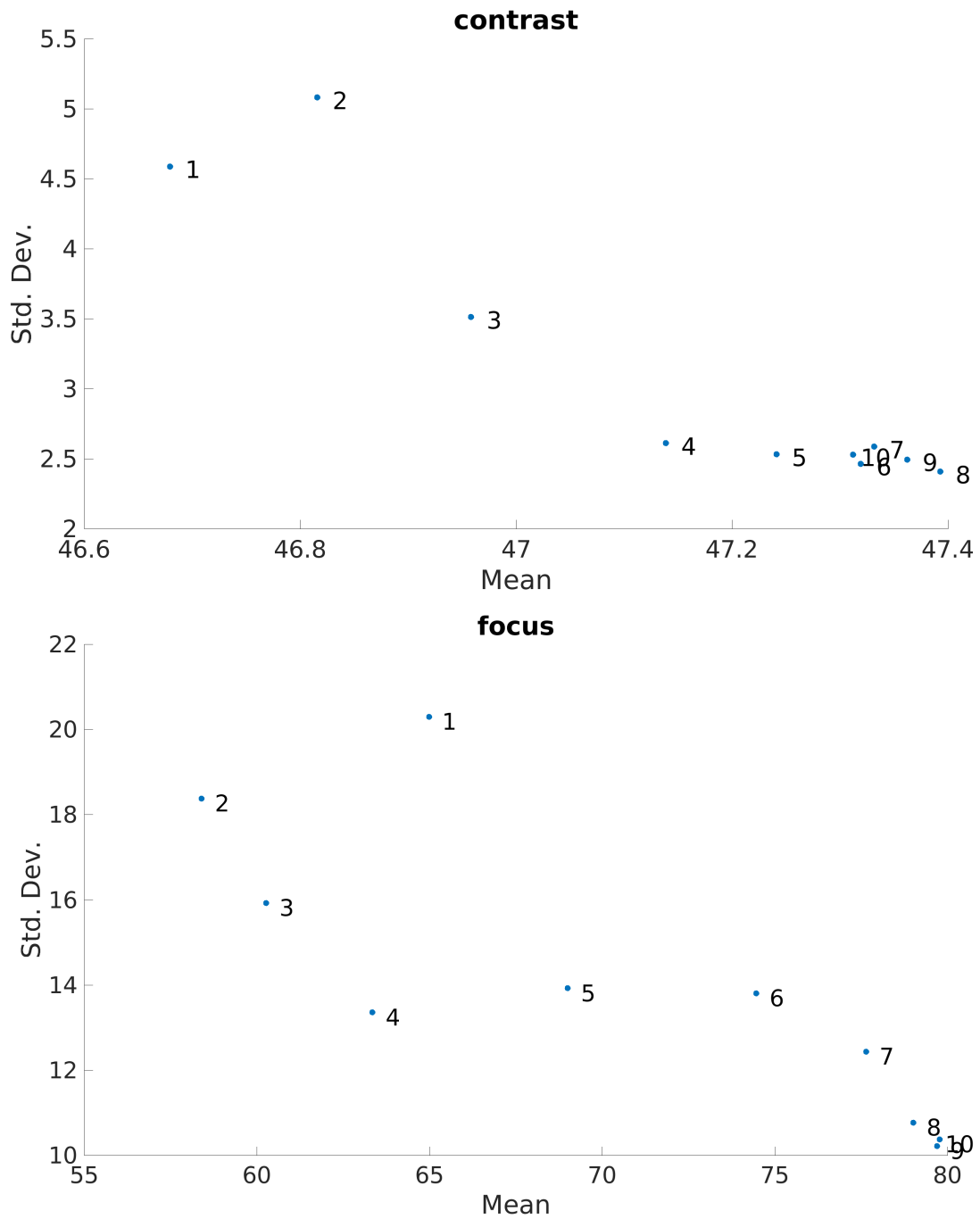


Figure 3.7: Mean vs Standard deviation plot for contrast and focus measure. We can observe the noisy distributions are not following the order of the quality level.

But, we need all of the quality metrics to provide optimal to near optimal measure values to get a high quality score. For this reason, quantitative evaluation of “understandability” is a very hard problem. Because, in order to calculate that we need to study how change in one quality metric affects the rest of the metrics, which is a whole new research topic.

3.4.1 Limitations

Some of the limitation we faced while performing the experiments are: 1) Noise in data hampers getting better results, as we have seen because of the composite effect of different quality factors, it is very difficult to figure out which factors are ultimately responsible for the quality degradation, especially for lower quality images. It is possible that composite effect of two or more quality measures, how ever small, can have bigger effect than a single measure, this could be something needed to be investigated in future work , 2) even though the database we have used has wide ranging quality, it is possible that it has not covered all possible quality values for every quality measures considered, 3) there are more ways quality measures which can affect the face quality outside of what we have considered. 4) Some quality measures may have bimodal distributions .e.g. brightness, sharpness, focus, illumination etc. Because quality can degrade both higher and lower values of these measures, but from our database it seems that in reality the occurrence of such cases is not common, but with more data it can be further looked into.

3.5 Conclusions

There are number of reasons that can affect the quality of a face image. These reasons can range from presence of different image sensors, compression algorithms, video or image acquisition conditions, time of acquisition etc. For these varied reasons, automatic face image quality assessment is a very challenging subject. In recent years a number of learning based FIQA methods have been proposed which provides good prediction of face recognition performance based on the face image quality score. But, providing just a single score or a quality bin for the face image cannot provide much information for the human assessor. From the end-user perspective, it is important that there should more explanation provided for the predicted score and some form of interpretation of it. This auxiliary information can help provide useful hints that can help to develop a improved image acquisition process or set up suitable environment that can ensure high face recognition accuracy. In this chapter, we have proposed a new paradigm, which can provide human understandable information for face image quality assessment. We believe this can help address the lack of explanation for a learning based quality assessment of face images. Our experiments and provided results show the effectiveness of the proposed method in providing important information regarding different quality factors related for face images.



Figure 3.8: Sample images with understandable information regarding quality. At right side of the images we have: 1) the predicted quality score 2) quality measure having non-optimal value 3) Probability of the value belonging to bin-N 4) Bin number with the max probability 5) understandable label category

Chapter 4

Conclusion & Future work

In this work we have presented, a large scale database for benchmarking face image quality assessment methods, proposed a learning to rank based approach for quality label generation, and a comparative study of different representative face quality assessment methods, using commonly used feature types, to demonstrate the effectiveness of this database and the training protocol as a FIQA benchmarking platform. The results found in comparative study has shown that, the ranking based quality scores can help improve results, especially for, regressor based methods when used with face recognition features. Face recognition features has been found to be the best suitable for FIQA. Visual inspection of the scores generated by the top performing method show that the method can capture various face quality changes such as illumination, pose, occlusion, expression etc. With this new database and training protocol, we wish to invite the biometric research community to further investigate the challenges of face image quality assessment. We believe that, by having a platform for benchmarking the state-of-the-art will move the biometric research community further in face biometric quality research. In future work we plan to further increase the database both in width and depth, to include much more quality variance. Deep feature based facial quality assessment is another aspect that needs to be investigated.

In second part of this thesis, we have shown providing just a single score or a quality bin for the face image cannot provide much information for the human assessor. From an end-user point-of-view, it is important that there more explanation should be provided for the predicted score. This auxiliary information can help provide useful hints that can help to develop a improved image acquisition process or set up robust enrollment system that can ensure high confidence in face recognition accuracy. In our work, we have proposed a novel stochastic method which can provide human understandable information for face image quality assessment. We believe this can help address those issues. Our experiments and provided results show promising aspects of the proposed method. In future work, we have plans to devise methods that can be devised to enhance those degraded quality measures, detected using UFIQ, so that the quality score can be improved.

Bibliography

- [1] “Face recognition homepage, databases section.” [Online]. Available: <http://www.face-rec.org/databases>
- [2] P. Grother and E. Tabassi, “Performance of biometric quality measures,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 4, pp. 531–543, 2007.
- [3] S. Bharadwaj, M. Vatsa, and R. Singh, “Biometric quality: a review of fingerprint, iris, and face,” *EURASIP Journal on Image and Video Processing*, vol. 2014, no. 1, p. 34, 2014.
- [4] F. Alonso-Fernandez, J. Fierrez, and J. Ortega-Garcia, “Quality measures in biometric systems,” *IEEE Security Privacy*, vol. 10, no. 6, pp. 52–62, Nov 2012.
- [5] E. Tabassi and P. Grother, *Fingerprint Image Quality*. Boston, MA: Springer US, 2009, pp. 482–490.
- [6] A. Hicklin and R. Khanna, “The role of data quality in biometric systems,” Mitretek Systems, Tech. Rep., 2006.
- [7] 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 *CVPR 2011 WORKSHOPS*, June 2011, pp. 74–81.
- [8] G. Liu, Y. Xu, and J. Lan, “No-reference face image assessment based on deep features,” in *2016 Applications of Digital Image Processing*, vol. 9971, 2016, pp. 99 711S–99 711S–7.
- [9] S. Vignesh, K. M. Priya, and S. S. Channappayya, “Face image quality assessment for face selection in surveillance video using convolutional neural networks,” in *2015 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Dec 2015, pp. 577–581.
- [10] C. Pan, B. Ni, Y. Xu, and X. Yang, “Recognition oriented facial image quality assessment via deep convolutional neural network,” in *Proceedings of the International Conference on Internet Multimedia Computing and Service*, 2016, pp. 160–163.
- [11] S. Bharadwaj, M. Vatsa, and R. Singh, “Biometric quality: a review of fingerprint, iris, and face,” *EURASIP Journal on Image and Video Processing*, vol. 2014, no. 1, p. 34, 2014.
- [12] R. L. V. Hsu, J. Shah, and B. Martin, “Quality assessment of facial images,” in *2006 Biometrics Symposium: Special Session on Research at the Biometric Consortium Conference*, Sept 2006, pp. 1–6.

- [13] A. Fourney and R. Laganier, "Constructing face image logs that are both complete and concise," in *Computer and Robot Vision, 2007. CRV '07. Fourth Canadian Conference on*, May 2007, pp. 488–494.
- [14] A. Abaza, M. A. Harrison, and T. Bourlai, "Quality metrics for practical face recognition," in *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, Nov 2012, pp. 3103–3107.
- [15] A. Abaza, M. A. Harrison, T. Bourlai, and A. Ross, "Design and evaluation of photometric image quality measures for effective face recognition," *IET Biometrics*, vol. 3, no. 4, pp. 314–324, 2014.
- [16] E. Omidiora, S. Olabiyisi, J. Ojo, A.-A. Adebayo, O. Abayomi-Alli, and K. Erameh, "Facial image verification and quality assessment system-faceivqa," *International Journal of Electrical and Computer Engineering*, vol. 3, no. 6, p. 863, 2013.
- [17] E. O. Omidiora, S. O. Olabiyisi, J. A. Ojo, R. A. Ganiyu, and A. Abayomi-Alli, "Enhanced face verification and image quality assessment scheme using modified optical flow technique," *Proc. WCECS*, vol. 14, 2014.
- [18] X. h. Chen and C. z. Li, "Image quality assessment model based on features and applications in face recognition," in *2011 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC)*, Sept 2011, pp. 1–4.
- [19] A. D. Bagdanov, A. D. Bimbo, F. Dini, G. Lisanti, and I. Masi, "Posterity logging of face imagery for video surveillance," *IEEE MultiMedia*, vol. 19, no. 4, pp. 48–59, Oct 2012.
- [20] G. Zhang and Y. Wang, *Asymmetry-Based Quality Assessment of Face Images*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 499–508.
- [21] Q. Xiong and C. Jaynes, "Mugshot database acquisition in video surveillance networks using incremental auto-clustering quality measures," in *Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance, 2003.*, July 2003, pp. 191–198.
- [22] Y. Yao, B. R. Abidi, N. D. Kalka, N. A. Schmid, and M. A. Abidi, "Improving long range and high magnification face recognition: Database acquisition, evaluation, and enhancement," *Computer Vision and Image Understanding*, vol. 111, no. 2, pp. 111 – 125, 2008.
- [23] D. Bhattacharjee, S. Prakash, and P. Gupta, *No-Reference Image Quality Assessment for Facial Images*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 594–601.
- [24] K. Anantharajah, S. Denman, S. Sridharan, C. Fookes, and D. Tjondronegoro, "Quality based frame selection for video face recognition," in *2012 6th International Conference on Signal Processing and Communication Systems*, Dec 2012, pp. 1–5.
- [25] K. Anantharajah, S. Denman, D. Tjondronegoro, S. Sridharan, C. Fookes, and X. Guo, "Quality based frame selection for face clustering in news video," in *2013 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, Nov 2013, pp. 1–8.
- [26] K. Nasrollahi and T. B. Moeslund, "Complete face logs for video sequences using face quality measures," *IET Signal Processing*, vol. 3, no. 4, pp. 289–300, July 2009.

- [27] —, *Face Quality Assessment System in Video Sequences*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 10–18.
- [28] —, “Hybrid super resolution using refined face logs,” in *2010 2nd International Conference on Image Processing Theory, Tools and Applications*, July 2010, pp. 435–440.
- [29] —, “Finding and improving the key-frames of long video sequences for face recognition,” in *2010 Fourth IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, Sept 2010, pp. 1–6.
- [30] K. Nasrollahi, T. B. Moeslund, and M. Rahmati, “Summarization of surveillance video sequences using face quality assessment,” *International Journal of Image and Graphics*, vol. 11, no. 02, pp. 207–233, 2011.
- [31] M. A. Haque, K. Nasrollahi, and T. B. Moeslund, “Real-time acquisition of high quality face sequences from an active pan-tilt-zoom camera,” in *2013 10th IEEE International Conference on Advanced Video and Signal Based Surveillance*, Aug 2013, pp. 443–448.
- [32] K. Lin, X. Wang, S. Cui, and Y. Tan, “Heterogeneous feature fusion-based optimal face image acquisition in visual sensor network,” in *2015 IEEE International Instrumentation and Measurement Technology Conference (I2MTC) Proceedings*, May 2015, pp. 1078–1083.
- [33] Z. Wei, X. Li, and L. Zhuo, “An automatic face log collection method for video sequence,” in *Proceedings of the Fifth International Conference on Internet Multimedia Computing and Service*, ser. ICIMCS ’13. New York, NY, USA: ACM, 2013, pp. 376–379.
- [34] X. Gao, S. Z. Li, R. Liu, and P. Zhang, *Standardization of Face Image Sample Quality*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 242–251.
- [35] J. Long and S. Li, “Near infrared face image quality assessment system of video sequences,” in *2011 Sixth International Conference on Image and Graphics*, Aug 2011, pp. 275–279.
- [36] J. Sang, Z. Lei, and S. Z. Li, “Face image quality evaluation for iso/iec standards 19794-5 and 29794-5,” in *Proceedings of the Third International Conference on Advances in Biometrics*, ser. ICB ’09. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 229–238.
- [37] M. De Marsico, M. Nappi, and D. Riccio, “Measuring measures for face sample quality,” in *Proceedings of the 3rd International ACM Workshop on Multimedia in Forensics and Intelligence*, ser. MiFor ’11. New York, NY, USA: ACM, 2011, pp. 7–12.
- [38] A. J. Abboud, H. Sellahewa, and S. A. Jassim, “Quality based approach for adaptive face recognition,” in *Mobile Multimedia/Image Processing, Security, and Applications 2009*, 2009.
- [39] K. Kryszczuk, J. Richiardi, P. Prodanov, and A. Drygajlo, “Error handling in multimodal biometric systems using reliability measures,” in *2005 13th European Signal Processing Conference*, Sept 2005, pp. 1–4.
- [40] —, “Reliability-based decision fusion in multimodal biometric verification systems,” *EURASIP Journal on Advances in Signal Processing*, vol. 2007, no. 1, p. 086572, 2007.
- [41] K. Kryszczuk and A. Drygajlo, “On combining evidence for reliability estimation in face verification,” in *2006 14th European Signal Processing Conference*, Sept 2006, pp. 1–5.

- [42] —, “On face image quality measures,” in *Proceedings of the 2nd Workshop on Multimodal User Authentication*, 2006.
- [43] Q. C. Truong, T. K. Dang, and T. Ha, *Face Quality Measure for Face Authentication*. Cham: Springer International Publishing, 2016, pp. 189–198.
- [44] P. Liao, H. Lin, P. Zeng, S. Bai, H. Ma, and S. Ding, “Facial image quality assessment based on support vector machines,” in *2012 International Conference on Biomedical Engineering and Biotechnology*, May 2012, pp. 810–813.
- [45] S. Bharadwaj, M. Vatsa, and R. Singh, “Can holistic representations be used for face biometric quality assessment?” in *2013 IEEE International Conference on Image Processing*, Sept 2013, pp. 2792–2796.
- [46] H. S. Bhatt, S. Bharadwaj, M. Vatsa, R. Singh, A. Ross, and A. Noore, “A framework for quality-based biometric classifier selection,” in *2011 International Joint Conference on Biometrics (IJCB)*, Oct 2011, pp. 1–7.
- [47] S. Bharadwaj, H. S. Bhatt, R. Singh, M. Vatsa, and A. Noore, “Qfuse: Online learning framework for adaptive biometric system,” *Pattern Recognition*, vol. 48, no. 11, pp. 3428 – 3439, 2015.
- [48] N. Ozay, Y. Tong, F. W. Wheeler, and X. Liu, “Improving face recognition with a quality-based probabilistic framework,” in *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, June 2009, pp. 134–141.
- [49] M. El-Abed, C. Charrier, and C. Rosenberger, “Quality assessment of image-based biometric information,” *EURASIP Journal on Image and Video Processing*, vol. 2015, no. 1, p. 3, 2015.
- [50] J. Chen, Y. Deng, G. Bai, and G. Su, “Face image quality assessment based on learning to rank,” *IEEE Signal Processing Letters*, vol. 22, no. 1, pp. 90–94, Jan 2015.
- [51] H. I. Kim, S. H. Lee, and Y. M. Ro, “Face image assessment learned with objective and relative face image qualities for improved face recognition,” in *2015 IEEE International Conference on Image Processing (ICIP)*, Sept 2015, pp. 4027–4031.
- [52] T. Thorsen, P. Wasnik, C. Busch, R. Raghavendra, and K. Raja, “Assessing face image quality with lstms,” *NISK 2018*, p. 53, 2018.
- [53] A. Abayomi-Alli, E. Omidiora, S. Olabiyisi, and J. Ojo, “Adaptive regression splines models for predicting facial image verification and quality assessment scores,” *Balkan Journal of Electrical and Computer Engineering*, vol. 3, no. 1, 2015.
- [54] R. Raghavendra, K. B. Raja, B. Yang, and C. Busch, “Automatic face quality assessment from video using gray level co-occurrence matrix: An empirical study on automatic border control system,” in *2014 22nd International Conference on Pattern Recognition*, Aug 2014, pp. 438–443.
- [55] M. Vatsa, R. Singh, and A. Noore, *SVM Based Adaptive Biometric Image Enhancement Using Quality Assessment*. Springer Berlin Heidelberg, 2008, pp. 351–371.
- [56] S. A. Berrani and C. Garcia, “Enhancing face recognition from video sequences using robust statistics,” in *IEEE Conference on Advanced Video and Signal Based Surveillance, 2005.*, Sept 2005, pp. 324–329.

- [57] H. Sellaheewa and S. A. Jassim, "Image-quality-based adaptive face recognition," *IEEE Transactions on Instrumentation and Measurement*, vol. 59, no. 4, pp. 805–813, April 2010.
- [58] M. Abdel-Mottaleb and M. H. Mahoor, "Application notes - algorithms for assessing the quality of facial images," *IEEE Computational Intelligence Magazine*, vol. 2, no. 2, pp. 10–17, May 2007.
- [59] H. I. Kim, S. H. Lee, and Y. M. Ro, "Investigating cascaded face quality assessment for practical face recognition system," in *2014 IEEE International Symposium on Multimedia*, Dec 2014, pp. 399–400.
- [60] Y. G. Kim, W. O. Lee, K. W. Kim, H. G. Hong, and K. R. Park, "Performance enhancement of face recognition in smart tv using symmetrical fuzzy-based quality assessment," *Symmetry*, vol. 7, no. 3, p. 1475, 2015.
- [61] L. Best-Rowden and A. K. Jain, "Learning face image quality from human assessments," *IEEE Transactions on Information Forensics and Security*, vol. PP, no. 99, pp. 1–1, 2018.
- [62] J. Yu, K. Sun, F. Gao, and S. Zhu, "Face biometric quality assessment via light cnn," *Pattern Recognition Letters*, 2017.
- [63] K. Kryszczuk, J. Richiardi, and A. Drygajlo, "Impact of combining quality measures on biometric sample matching," in *2009 IEEE 3rd International Conference on Biometrics: Theory, Applications, and Systems*, Sept 2009, pp. 1–6.
- [64] A. Adler and T. Dembinsky, "Human vs. automatic measurement of biometric sample quality," in *2006 Canadian Conference on Electrical and Computer Engineering*, May 2006, pp. 2090–2093.
- [65] K. E. Wertheim, "Human factors in large-scale biometric systems: A study of the human factors related to errors in semiautomatic fingerprint biometrics," *IEEE Systems Journal*, vol. 4, no. 2, pp. 138–146, 2010.
- [66] R. L. V. Hsu, J. Shah, and B. Martin, "Quality assessment of facial images," in *2006 Biometrics Symposium: Special Session on Research at the Biometric Consortium Conference*, 2006, pp. 1–6.
- [67] J. R. Beveridge, P. J. Phillips, G. H. Givens, B. A. Draper, M. N. Teli, and D. S. Bolme, "When high-quality face images match poorly," in *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, 2011, pp. 572–578.
- [68] C. P. Lee and C. J. Lin, "Large-scale linear ranksvm," *Neural Computation*, vol. 26, no. 4, pp. 781–817, April 2014.
- [69] A. Airola, T. Pahikkala, and T. Salakoski, "Training linear ranking svms in linearithmic time using red-black trees," *Pattern Recogn. Lett.*, vol. 32, no. 9, Jul. 2011.
- [70] J. R. Beveridge, G. H. Givens, P. J. Phillips, B. A. Draper, and Y. M. Lui, "Focus on quality, predicting frvt 2006 performance," in *2008 8th IEEE International Conference on Automatic Face Gesture Recognition*, Sept 2008, pp. 1–8.

- [71] P. J. Phillips, J. R. Beveridge, D. S. Bolme, B. A. Draper, G. H. Givens, Y. M. Lui, S. Cheng, M. N. Teli, and H. Zhang, "On the existence of face quality measures," in *2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, 2013, pp. 1–8.
- [72] G. Aggarwal, S. Biswas, P. J. Flynn, and K. W. Bowyer, "Predicting good, bad and ugly match pairs," in *2012 IEEE Workshop on the Applications of Computer Vision (WACV)*, Jan 2012, pp. 153–160.
- [73] Z. Wang, H. R. Sheikh, and A. C. Bovik, "No-reference perceptual quality assessment of jpeg compressed images," in *Proceedings. International Conference on Image Processing*, vol. 1, 2002, pp. I-477–I-480 vol.1.
- [74] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, 2005, pp. 886–893 vol. 1.
- [75] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001.
- [76] L. Wiskott, N. Krger, N. Kuiger, and C. von der Malsburg, "Face recognition by elastic bunch graph matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 775–779, July 1997.
- [77] T. Ojala, M. Pietikinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognition*, vol. 29, no. 1, pp. 51 – 59, 1996.
- [78] J. G. Daugman, "Two-dimensional spectral analysis of cortical receptive field profiles," *Vision Research*, vol. 20, no. 10, pp. 847 – 856, 1980.
- [79] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind image quality assessment: A natural scene statistics approach in the dct domain," *IEEE Transactions on Image Processing*, vol. 21, no. 8, pp. 3339–3352, Aug 2012.
- [80] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, Dec 2012.
- [81] L. Xinwei, P. Marius, C. Christophe, M., and B. Patrick, "Performance evaluation of no-reference image quality metrics for face biometric images," *Journal of Electronic Imaging*, vol. 27, pp. 27 – 27 – 24, 2018.
- [82] L. Liu, B. Liu, H. Huang, and A. C. Bovik, "No-reference image quality assessment based on spatial and spectral entropies," *Signal Processing: Image Communication*, vol. 29, no. 8, pp. 856 – 863, 2014.
- [83] A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Transactions on Image Processing*, vol. 20, no. 12, pp. 3350–3364, Dec 2011.

- [84] L. Liu, H. Dong, H. Huang, and A. C. Bovik, “No-reference image quality assessment in curvelet domain,” *Signal Processing: Image Communication*, vol. 29, no. 4, pp. 494 – 505, 2014.
- [85] A. Mittal, R. Soundararajan, and A. C. Bovik, “Making a completely blind image quality analyzer,” *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, 2013.
- [86] A. Mittal, G. S. Muralidhar, J. Ghosh, and A. C. Bovik, “Blind image quality assessment without human training using latent quality factors,” *IEEE Signal Processing Letters*, vol. 19, no. 2, pp. 75–78, Feb 2012.
- [87] O. M. Parkhi, A. Vedaldi, and A. Zisserman, “Deep face recognition,” in *British Machine Vision Conference*, 2015.
- [88] G. Guo and N. Zhang, “What is the challenge for deep learning in unconstrained face recognition?” in *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, 2018, pp. 436–442.
- [89] R. Kumar and S. Vassilvitskii, “Generalized distances between rankings,” in *Proceedings of the 19th International Conference on World Wide Web*, ser. WWW ’10, 2010, pp. 571–580.
- [90] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, H. Mobahi, and Y. Ma, “Toward a practical face recognition system: Robust alignment and illumination by sparse representation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 2, pp. 372–386, Feb. 2012. [Online]. Available: <http://dx.doi.org/10.1109/TPAMI.2011.112>
- [91] H. I. Kim, S. H. Lee, and Y. M. Ro, “Investigating cascaded face quality assessment for practical face recognition system,” in *2014 IEEE International Symposium on Multimedia*, Dec 2014, pp. 399–400.
- [92] A. F. Abate, M. Nappi, D. Riccio, and G. Sabatino, “2d and 3d face recognition: A survey,” *Pattern Recogn. Lett.*, vol. 28, no. 14, pp. 1885–1906, Oct. 2007. [Online]. Available: <http://dx.doi.org/10.1016/j.patrec.2006.12.018>
- [93] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” *ArXiv e-prints*, Sep. 2014.
- [94] J. Yu, K. Sun, F. Gao, and S. Zhu, “Face biometric quality assessment via light cnn,” *Pattern Recognition Letters*, 2017. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0167865517302477>
- [95] I. 19794-5, “Information technology – Biometric data interchange formats – Part 5: Face image data,” International Organization for Standardization, Geneva, CH, Standard, Jun. 2005.
- [96] D. 9303, “Machine readable travel documents,” International Civil Aviation Organization, Standard, Aug. 2006.
- [97] Z. Wang, H. R. Sheikh, and A. C. Bovik, “No-reference perceptual quality assessment of jpeg compressed images,” in *Proceedings. International Conference on Image Processing*, vol. 1, 2002, pp. I-477–I-480 vol.1.
- [98] T. Baltruaitis, P. Robinson, and L. P. Morency, “Openface: An open source facial behavior analysis toolkit,” in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2016, pp. 1–10.