Learning based Image Quality Assessment

K. V. S. N. L. Manasa Priya

A Thesis Submitted to

Indian Institute of Technology Hyderabad

In Partial Fulfillment of the Requirements for

The Degree of Master of Technology

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Department of Electrical Engineering

June 2015

Declaration

I declare that this written submission represents my ideas in my own words, and where ideas or words of others have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be a cause for disciplinary action by the Institute and can also evoke penal action from the sources that have thus not been properly cited, or from whom proper permission has not been taken when needed.

K. Maanasa priya

(Signature)

K.V.S.N.L. Manasa priya

(K. V. S. N. L. Manasa Priya)

EE12M1020

(Roll No.)

Approval Sheet

This Thesis entitled Learning based Image Quality Assessment by K. V. S. N. L. Manasa Priya is approved for the degree of Master of Technology from IIT Hyderabad

B. Vineeth

(Dr. Vineeth N Balasubramanian)(.) Examiner
Dept. of Computer Science Eng
IITH

(K. SRIRAMAMURTY).

(SOUMYA JANA)(.) Examiner
Dept. of Electrical Eng
IITH

(Dr. Sumohana Channappayya) Adviser
Dept. of Electrical Eng
IITH

(Dr. KETAN DETROJA) (.) Chairman
Dept. of Electrical Eng
IITH

Acknowledgements

Firstly, I would like to express my deep and sincere gratitude to my advisor, Dr. Sumohana Channappayya, for his inspiration, guidance and constructive comments throughout my studies. His constant support and insightful suggestions were instrumental in completing my thesis.

I would like to thank all my professors who taught me and motivated me in every aspect. I would also like to extend my gratitude to department of Electrical Engineering, Indian Institute of Technology, Hyderabad for providing financial support.

My sincere thanks also goes to my lab-mates in the LFOVIA Lab for the inspirations, countless arguments and all the fun we have had in the last three years. I couldn't have achieved what I did without all of you.

Finally, I am forever indebted to my parents and my brother for their understanding, endless patience and encouragement from the beginning.

Abstract

In this thesis, we present an abstract view of image quality assessment algorithms. Most of the research in the area of image quality assessment is focused on the scenario where the end-user is a human observer and therefore commonly known as perceptual image quality assessment. However, we believe that we should extend the field of image quality assessment to the task specific scenario where the end-user is not a human observer. The quality of image/video should be assessed based on the end-system/user that we call task-based image quality assessment. In this thesis, we discuss both perceptual image quality assessment and task-based image quality assessment with respect to face recognition task.

In the case of perceptual image quality assessment, we present image quality assessment algorithms in a full-reference setting and a no-reference setting. In both the settings, our algorithm is inspired by the sparse representation of natural images in the human visual system (HVS). The hypothesis behind the proposed method is that the properties of natural images that afford their sparse representation are altered in the presence of distortion. We attempt to quantify this change in sparsity and show that it is indeed a measure of the unnaturalness or distortion in an image. We show that the proposed algorithms consistently correlate well with subjective scores over several popular image databases.

In the case of task-specific image quality assessment, we present an image quality assessment method that is aimed at the face recognition task. Face image Quality Assessment (FQA) plays a key role in improving face recognition accuracy and increasing computational efficiency. In the context of video, it is very common to acquire multiple face images of a single person. If one were to use all the acquired face images for the recognition task, the computational load for Face Recognition (FR) increases while recognition accuracy decreases due to outliers. This impediment necessitates a strategy to optimally choose the good quality face images from the pool of images. To address this, we propose two algorithms. One is based on the hypothesis that sparseness of the probe face will be altered if the probe face is not similar to ideal face and the other is based on mimicking the recognition capability of a given FR algorithm using a Convolutional Neural Network (CNN). Preliminary results demonstrate that the proposed method is on par with the state-of-the-art FQA methods in improving the performance of FR algorithms in a surveillance scenario.

Contents

Chapter 1

Introduction

There is a massive and ubiquitous role of digital multimedia-based applications in our day-to-day life ranging from medical diagnosis to security to entertainment. Often, this data passes through different processing stages before it reaches the end-user/system. At each processing stage, data is subjected to different distortions which degrades the quality. Hence, the efficient and reliable evaluation of multimedia quality assessment has gained importance - especially given the massive scale of multimedia data. Quality assessment is one of the basic and challenging problems in the field of image and video processing as well as many practical applications, such as process evaluation, bench marking of algorithms, optimization, testing and monitoring. The approach to evaluate the quality of image/video is task dependent i.e., depends on end-user/system. There is extensive research in the area of quality assessment which is largely focused on perceptual based quality since in most of the cases end-users are human observers. In addition to perceptual image quality assessment, I also worked on face quality assessment in the surveillance scenario where the end-user/end-system is a face recognition algorithm.

Having understood the significance of quality assessment in various fields and the need to evaluate the quality of image/video based on the end-user/system; the main contributions of my research are in the areas of perception based quality assessment and face quality assessment in surveillance settings. The approaches are briefly discussed in the following sections.

1.1 Perceptual based quality assessment

The most accurate and reliable way for perception based quality assessment is through subjective evaluations. But these evaluations depend on environmental settings, time of the day and evaluation, utmost care needs to be taken that the ideal settings should not alter with time, proper representative set of subjects has to be considered so that evaluations wont be biased. All these above factors have made them expensive and time-consuming; therefore, these evaluations are highly impractical in real-time settings. Hence, there is a necessity to design objective algorithms for estimating the quality score of an image/video which needs to be highly correlated with the score given by average human observer. The effectiveness of quality assessment algorithm is evaluated based on the correlation of predicted scores with the subjective evaluations. When the correlation is high, algorithm is able to mimic the average human observer with high probability. Based on the availability of reference image which is considered to be pristine and distortion-free, image quality assessment (IQA) algorithms are broadly classified into three categories viz., full-reference, reduced-reference, and no-reference algorithms. In this work, I have primarily focused on the first and third categories.

1.1.1 Full Reference (FR) - IQA algorithm

Complete reference image of same scene is available and the quality of test image is evaluated with respect to the reference image. The underlying principles of the state-of-the-art FR IQA algorithms have ranged from attempting to model the physiology of the human visual system [1] to using abstract notions from information theory [2]. An excellent exposition of these principles can be found in [3]. The success of these varied principles leads one to believe that there could either be several different approaches to solving the FR IQA problem or that these approaches are yet to converge to the true solution. Recent works by Guha et. al. [4, 5, 6] provide yet another approach to measuring image similarity that is based on sparse representations of natural images. This is a promising approach given its close analogy with sparse representations in the human visual system [7]. In this work, I used this approach to demonstrate the several useful properties to make it an attractive FR IQA algorithm and will be discussed in detail in subsequent chapters.

1.1.2 Reduced Reference (RR) - IQA algorithm

Complete reference image is not available; however partial information like features of reference image is available to evaluate the quality of test image.

1.1.3 No Reference (NR) - IQA algorithm

In most of the practical settings, the reference image is not available for quality assessment. Hence, there is a need to evaluate the quality solely based on the test image. These algorithms generally follow one or a combination of three approaches:

- Distortion-specific approaches: In this approach, algorithms quantify the distortions such as blur[8], ringing effect [9], or blockiness [10] and evaluate the image accordingly.
- Training-based approaches: In this approach, the algorithm predicts the quality of an image by training a model from the features extracted [11, 12].
- Natural scene statistics (NSS) approaches: These algorithms assume that undistorted/pristine images occupy a small subspace of the space of all possible images and estimates the quality of test image by calculating the distance between the test image and subspace of pristine images [13].

In this work, I used a combination of the second and third approaches to predict the quality of images. The proposed NR-IQA algorithm will be discussed in detail in subsequent chapters.

1.2 Face Quality Assessment in Surveillance Settings

In the past few decades, face recognition has received great attention not only due to its numerous applications, including video surveillance, access control, entertainment and law enforcement but also to understand the face recognition process in humans. Since face recognition is the natural way of identification and verification, this field is rich with excellent literature [14, 15, 16]. In the last two decades, various algorithms have been proposed for face recognition based on still images and video sequences. However, in realistic scenarios, face recognition is limited by low quality images and variation in pose, illumination, occlusion and expression in the acquired face image [15]. Such problems are even more severe in surveillance systems where users may be uncooperative and the environment is uncontrolled. Since poor quality images in the surveillance video sequences offer very little information for face recognition, they not only increase the computational load because of complex processes such as feature extraction and matching, but also reduce the recognition accuracy because of outliers. Therefore, there is a need to develop an automated face quality assessment in improving face recognition accuracy and the computational efficiency.

In the context of video, it is very common to acquire multiple shots of a single person. If we consider all the shots, computational complexity of face recognition algorithm increases while recognition accuracy decreases due to outliers. So, we need a strategy to optimally choose the good quality faces from the pool of shots in order to improve the performance of recognition algorithms.

The approach to choose good quality faces from video sequence will be discussed in detail in subsequent chapters.

Chapter 2

Background Theory

2.1 Overview

In this chapter we provide the related theory used in this thesis. We first define the concept of sparse representation, followed by the techniques used for obtaining sparse solution and dictionary learning. The concept of sparse representation and dictionary learning is used in the work related to perceptual quality assessment and face quality evaluation. Then, we define the concept of Convolutional Neural Networks (CNN) that are used for face quality evaluation.

2.2 Sparse Representation of Signals

The objective of sparse representation of signals is to represent a signal with a few number of representative elements. Using an overcomplete dictionary matrix $D \in \mathbb{R}^{n \times K}$ that contains K representative signal-atoms, a signal $y \in \mathbb{R}^n$ can be represented with linear combination of fewer atoms.

The representation of y w.r.t dictionary may be exact $y = Dx$ in a noiseless scenario [17]. But in real-life situations, the representation can be approximated as $\|y - Dx\|_p \leq \epsilon$. The typical norms used for measuring deviation are the l^p norms where p varies from 1 to ∞ [18]. But in most of the cases p is preferably considered as 2.

As D is overcomplete, there is a possibility of an infinite number of solutions to represent the signal y and hence there is a need to set the constraints on solution set. To acquire a sparse representation, the solution with fewest number of coefficient is certainly an appealing constraint on solution set. This sparse representation is solution of either

$$\min_{x} \quad \|x\|_0 \quad \text{subject to} \quad y = Dx \tag{2.1}$$

or

$$\min_{x} \quad \|x\|_0 \quad \text{subject to} \quad \|y - Dx\|_2 \leq \epsilon \tag{2.2}$$

where the operator $\|.\|_0$ counts the number of non-zero elements. A similar objective as mentioned in (2.2) is alternately met by considering either

$$\min_{x} \quad \|y - Dx\|_2 \quad \text{subject to} \quad \|x\|_0 \leq L \tag{2.3}$$

or

$$\min_{x} \quad \frac{1}{2}\|y - Dx\|_2 + \lambda\|x\|_0 \tag{2.4}$$

where the parameter $L \geq 1$ controls the degree of sparsity and $\lambda \geq 0$ balances the residual and degree of sparsity.

In the following, we review the Orthogonal Matching Pursuit(OMP) algorithm [19] to solve for approximate sparse solution and then K-SVD algorithm [20] for dictionary learning.

2.3 OMP algorithm for approximate sparse solution

Solving (2.1) and (2.2) is NP hard and computationally expensive. There are several algorithms that have tried to approximate the objective function with alternatives and one such algorithm is the OMP algorithm. It is a simple and effective approximation method among greedy pursuit methods. It finds the locally optimum solution at each iteration by searching the basis that most resembles a residual. Considering an overcomplete dictionary A and a compressible sample b, the OMP algorithm is stated as follows:

Task: $(P_0) : min_x\|x\|_0$ subject to $Ax = b$.

Parameters: We are given the matrix A, the vector b, and the error threshold ϵ_0.

Initialization: Initialize $k = 0$, and set solution as $x^0 = 0$, residual as $r^0 = b - Ax^0 = b$ and solution support as $S^0 = Support\{x^0\} = 0$

Main Iteration: Increment k by 1 and perform the following steps:

- **Sweep:** Compute the errors $\epsilon(j) = min_{z_j}\|a_j z_j - r^{k-1}\|_2^2$ for all j using the optimal choice ${z_j}^* = {a_j}^T r^{k-1}/{\|a_j\|_2}^2$

- **Update Support:** Find a minimizer, j_0 of $\epsilon(j) : \forall j \notin S^{k-1}, \epsilon(j_0) < \epsilon(j)$, and update $S^k = S^{k-1} \cup \{j_0\}$.
- **Update Provisional Solution:** Compute x^k, the minimizer of $\|Ax - b\|_2{}^2$ subject to $Support\{x\} = S^k$.
- **Update Residual:** Compute $r^k = b - Ax^k$.
- **Stopping Rule:** Stop, if stopping criteria holds. Otherwise, apply another iteration.

Output: The proposed solution is x^k obtained after k iterations.

Generally, there are two stopping criteria for the OMP algorithm. First, the iterative process is performed for a fixed number of iterations. Second, the OMP algorithm stops when the bounded noises are within the predefined thresholds ($\|r^k\|^2 < \epsilon_0$).

2.4 K-SVD algorithm for dictionary learning

An overcomplete dictionary is used to approximate the given signal as a sparse signal. The types of dictionary can be classified into analytic dictionary and learned dictionary. The atoms of analytic dictionary is formulated analytically and is supported by optimally proofs and error rate bounds[21]. Examples of the analytic formulation include the Fast Fourier Transform (FFT) [22], the Discrete Cosine Transform (DCT) [23] and the Gabor transform [24]. In contrast, the learned dictionary tends to learn the dictionary from the given training data, which benefits from the finer adaptation to the nature of the problem on hand. In general, learned dictionaries often demonstrate state-of-the-art results in many of the applications[21]. In the past decade, a large volume of work [25, 20] has been devoted to dictionary training methods focusing on l_0 norm and l_1 norm sparsity measurements, which lead to the development of more efficient sparse coding algorithms[21]. Among the dictionary training methods, the K-SVD algorithm[20], which takes its name from the Singular Value Decomposition (SVD) process in the dictionary update stage, aims to train a dictionary D with a faster and more efficient algorithm. It first initialises a random dictionary D with l_2 normalised atoms and performs the iterative two-stage process until convergence is stated as follows:

Task: Find the best dictionary to represent the data samples $\{y_i\}$, $i = 1$ to N as sparse compositions, by solving

$$\min_{D,X} \|Y - DX\|_F^2 \text{ subject to } \forall i, \|x_i\|_0 \leq T_0$$

Initialization: Set the dictionary matrix $D^{(0)} \in \mathbb{R}^{n \times K}$ with l_2 normalized columns.

Repeat until convergence(stopping rule):

Step 1: Sparse Coding stage: Use any pursuit algorithm to compute the representation vectors x_i for each example y_i,by approximating the solution of

$$min_{x_i} \|y_i - Dx_i\|_2^2 \text{ subject to } \|x_i\|_0 \leq T_0$$

Step 2: Codebook Update Stage: For each column k = 1 to K in D,update it by

- Define the groups of examples that use this atom $\omega_k = \{i|1 \leq i \leq N, x_T^k(i) \neq 0\}$.
- Compute the overall representation error matrix, E_k, by

$$E_k = Y - \sum_{j \neq k} d_j x_T^j$$

- Restrict E_k by choosing only the columns corresponding to ω_k,and obtain E_k^R.
- Apply SVD decomposition $E_k^R = U\Delta V^T$. Choose the updated dictionary column d_k to be the first column of U. Update the coefficient vector x_k^R to be the first column of V multiplied by $\Delta(1,1)$.

Stopping Rule: Stop, if stopping criteria holds. Otherwise, apply another iteration.

2.5 Convolutional Neural Network (CNN)

A CNN comprises of one or more convolutional layers with subsampling layer as optional layer (the purpose is to reduce the computational complexity) and then followed by standard multi-layer neural network (NN) [26]. The architecture of a CNN is designed to take the advantage of the two dimensional structure of an image. Also, a CNN is easier to train and have fewer parameters compared to a fully connected NN with the same number of hidden units.

2.5.1 Architecture

A CNN consists of a number of convolutional and subsampling layers optionally followed by fully connected layers. The input to a convolutional layer is a $m \times m \times r$ image where m is the height and width of the image and r is the number of channels. The convolutional layer will have k filters of size $n \times n \times q$ where n is smaller than the dimension of the image and q can either be the same as the number of channels r or smaller and may vary for each kernel. The size of the filters gives rise to the locally connected structure which are each convolved with the image to produce k feature maps of size $m - n + 1$. Each map is then subsampled typically with mean or max pooling. Either before or after the subsampling layer an additive bias and sigmoidal nonlinearity is applied to each feature map. The figure 2.1 illustrates a CNN consisting of convolutional and subsampling sublayers followed by fully connected layers.

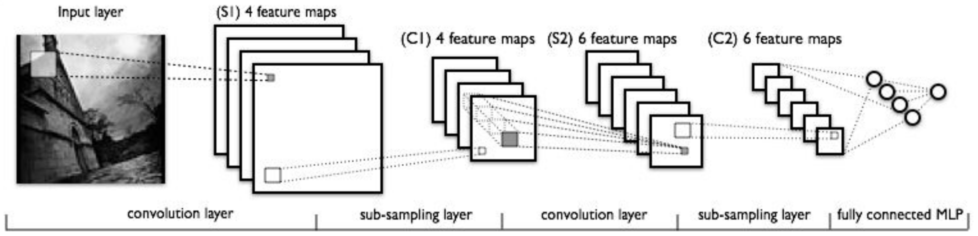


Figure 2.1: Architecture of CNN.

A CNN typically has three types of layers as defined above. The forward and backward propagations differs depending on what layer data is propagating through.

2.5.2 Forward Propagation

The output of a neural network unit is the output of the last layer L. We use the following terminology:

u_i^l : i_{th} neuron in layer l.

x_i^l : The input to neuron u_i^l.

y_i^l : The output of neuron u_i^l.

The input values to the neurons u_i^0 are fixed by input data. The neural network learns by adjusting a set of weights, w_{ij}^l, where w_{ij}^l is the weight from some neuron u_i^l's output to other neuron in next layer u_j^{l+1}.

Fully connected Network

1. Compute activations for layers with known inputs: $y_i^l = f(x_i^l) + b_i^l$
2. Compute inputs for the next layer from these activations: $x_i^l = \sum_j w_{ji}^{l-1} y_j^{l-1}$.
3. Repeat steps 1 and 2 until we reach the output layer and know the values of y^L

where f is the activation function and applies nonlinearity. This function is typically a sigmoid or tanh or RELU function. b_i^l is bias unit connected externally to the network

Convolutional Layer Suppose that we have a $N \times N$ square input to the convolutional layer. If we use an $m \times m$ filter ω, the convolutional layer output will be of size $(N - m + 1) \times (N - m + 1)$. The output of the convolution layer is computed as given below

$$y^l = x^l * \omega$$

where $*$ is convolution operator and then activation function is applied on the output.

Pooling/Sub Sampling Layer The pooling layers take $k \times k$ region from the input and output a single value, which is the maximum or average value in that region. For instance, if their input layer is a $N \times N$ layer, they will then output a $N/k \times N/k$ layer, as each $k \times k$ region is reduced to just a single value.

2.5.3 Back Propagation

In order to update the weights, error $E(y^L)$ is computed. This error can be computed in different ways such as cross-entropy or sum of squared residuals and selection of error computation technique depends on the application. The purpose of being able to compute the error is to be able to optimize the weights to minimize the error; that is, the process of learning. We learn via an algorithm known as back propagation, which we can derive in a similar manner to forward propagation.

Fully connected Network

1. Compute errors at the output layer L: $\frac{\partial E}{\partial y_i^L} = \frac{d}{dy_i^L} E(y^L)$
2. Compute partial derivative of error with respect to neuron input at layer l that has known errors: $\frac{\partial E}{\partial x_j^l} = f^{'}(x_j^l) \frac{\partial E}{\partial y_j^l}$
3. Compute errors at the previous layer (back propagate errors): $\frac{\partial E}{\partial y_i^l} = \sum w_{ij}^l \frac{\partial E}{\partial x_j^{l+1}}$
4. Repeat steps 2 and 3 until errors are known at all but the input layer
5. Compute the gradient of the error (derivative with respect to weights): $\frac{\partial E}{\partial w_{ij}^l} = y_i^l \frac{\partial E}{\partial x_j^{l+1}}$

Pooling Layer The pooling layers do not actually do any learning themselves. Instead, then reduce the size of the problem by introducing sparseness. In forward propagation, $k \times k$ blocks are reduced to a single value. Then, this single value acquires an error computed from backwards propagation from the previous layer. This error is then just forwarded to the place where it came from. Since it only came from one place in the $k \times k$ block, the back propagated errors from pooling layers are rather sparse.

Convolutional Layer It is almost identical to the back propagation algorithm for fully connected network. The only difference to take into account is the weight sharing in the convolution layer. If the l^{th} layer is convolutional layer, then the error is computed as follows given that error for next layer is known:

$$\frac{\partial E}{\partial x_{ij}^{l}} = f^{'}(x_{ij}^{l})\frac{\partial E}{\partial y_{ij}^{l}}$$

Since we know the errors at the current layer, we now have everything we need to compute the gradient with respect to the weights used by this convolutional layer. In addition to compute the weights for this convolutional layer, we need to propagate errors back to the previous layer.

$$\frac{\partial E}{\partial y_{ij}^{l-1}} = \frac{\partial E}{\partial x_{ij}^{l}} * \omega$$

Knowing the errors at all layers, error gradient with respect to weights is calculated.

Chapter 3

Full Reference Quality Assessment

3.1 Introduction

The vital role that digital multimedia plays in our lives (ranging from medical diagnosis to security to entertainment to online society) is no longer a question for debate but rather a well-accepted norm. This change to our lifestyle has led to a rapid proliferation of multimedia content that needs to be managed (compressed, stored, and communicated) efficiently and effectively. The role of automated or objective multimedia quality assessment to manage multimedia cannot be overemphasized - especially given the cost of subjective evaluation and the massive scale of multimedia data.

Automated or objective image and video quality assessment algorithms have made giant strides in the past decade. The invention of the Structural SIMilarity (SSIM) index [27] heralded a wave of significant improvements in the automatic assessment of image quality and in turn video quality as well. Several excellent full-reference (FR) [28, 2], reduced-reference (RR) [29], and no-reference (NR) [30] image quality assessment (IQA) algorithms have since been proposed. Each of these algorithms take us a step closer to the ultimate goal of being able to mimic the human visual system's assessment of image quality. Given the context of the proposed work, we will restrict our focus to full-reference IQA algorithms.

The underlying principles of the state-of-the-art FR IQA algorithms have ranged from attempting to model the physiology of the human visual system [1] to using abstract notions from information theory [2]. An excellent exposition of these principles can be found in [3]. The success of these varied principles leads one to believe that there could either be several different approaches to solving the FR IQA problem or that these approaches are yet to converge to the true solution. Recent work by

Guha et. al. [4, 31, 32] provide yet another approach to measuring image similarity that is based on sparse representations of natural images. This is a promising approach given its close analogy with sparse representations in the human visual system [7].

In this chapter, we consider one flavor of the sparsity-based similarity measure – the SDM, and attempt to determine its efficacy as an FR IQA algorithm. A preliminary evaluation of the SDM as an FR IQA has been carried out by Guha et. al. [4]. The main contributions of this work are: (i) a comprehensive statistical performance evaluation of the SDM on the LIVE image database [33, 34], and (ii) a demonstration of several useful properties of the SDM that make it an attractive FR IQA algorithm. This work is appeared in [35].

3.2 Sparsity-Based Distance Measure (SDM)

An interesting trend seen in image similarity measurement is the use of Kolmogorov complexity-inspired [36] formulations. An information distance between the two strings x and y can be defined as $max\{K(x|y), K(y|x)\}$ where $K(x|y)$ is the Kolomogorov complexity of x relative to y and vice-versa for $K(y|x)$. To convert it to a normalized symmetric metric, a novel normalized information distance (NID) measure was defined by Li et. al. [37] as follows:

$$NID(x, y) = \frac{max\{K(x|y), K(y|x)\}}{max\{K(x), K(y)\}} \tag{3.1}$$

where $K(x|y)$ is the conditional Kolmogorov complexity of x relative to y. While NID has nice analytical properties, it is not practical since computing the Kolmogorov complexity is an NP-hard problem. Recent methods attempt to approximate Kolmogorov complexity using quantities that can be computed using fast algorithms. To the best of our knowledge, the first such approach to measure image similarity was introduced by Nikvand et. al. [38] where the size of the encoded bitstream from a lossless image coder was used to approximate Kolmogorov complexity.

Guha et.al. [4] related sparsity and Kolomogorov complexity based on the inference that the number of components required to represent a signal increases with signal complexity. The SDM was then defined to measure image similarity as follows.

$$SDM(X, Y) = \frac{N(X|Y) + N(Y|X)}{N(X) + N(Y)}. \tag{3.2}$$

where X is the reference image; Y is the test image; $N(X)$ and $N(Y)$ represents the number of components required to represent the image from the dictionary learnt from the patches of X and Y respectively. $N(X|Y)$ and $N(Y|X)$ represent the number of components required to represent the image from the dictionary learnt from the patches of Y and X respectively. $N(X) < N(X|Y)$; since number of components required to represent the current image X from the dictionary learnt from the patches of X is always less than the dictionary learnt from the patches of Y. Hence lower values of SDM indicate better similarity between images under consideration and is always greater than or equal to one.

In this work, the SDM has been implemented using the K-SVD algorithm [39] to find $N(X), N(Y)$ and the cross term $N(X|Y), N(Y|X)$. A randomly chosen set of 3000 8×8 images patches were used for learning a dictionary containing 128 atoms.

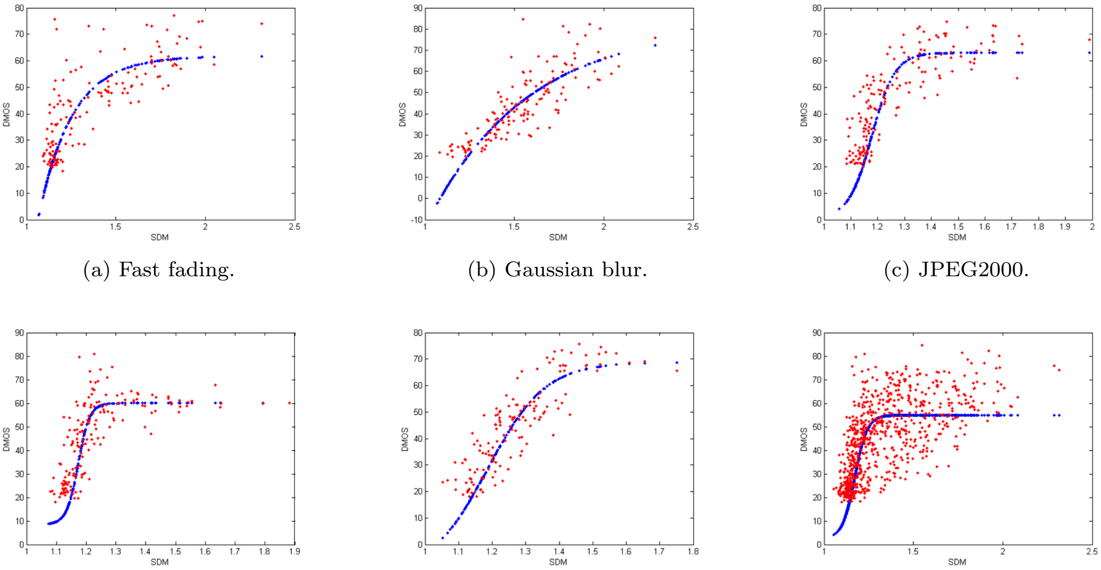
3.3 Statistical Evaluation

One of the main contributions of this work is to perform a statistical evaluation of the SDM as an FR IQA algorithm. The results of the statistical evaluation and an intuitive explanation of the performance are presented in the following subsections.

3.3.1 Evaluation

The SDM was evaluated over the LIVE database [40] that consists of 779 images covering a range of 5 types of distortions. There are 29 reference images and distortion types include fast fading, white noise, JPEG, JPEG 2000 and gaussian blur. SDM is compared with the state of art full reference algorithms such as SSIM [27], MSSSIM [28] and VIF [2]. The SDM scores were fit to the subjective scores (DMOS) using the four parameter exponential logistic function specified in [41].

The results of the statistical evaluation are presented in Fig. 3.1 and Table 3.1. Fig. 3.1 shows the scatter plots for each of the distortion types in the database along with an overall scatter plot. It is clear that the SDM performs best when the distortion type is either blur or additive white noise and performance drops for JPEG and fast fading distortions. We present an intuitive explanation for this performance in the following subsection. From Table 3.1, we see that the SDM performs fairly when compared to the state-of-the-art using Spearman Rank Ordered Correlation Coefficient (SROCC). However, we show in Section 3.4 that the SDM has several useful properties that the state-of-the-art IQA algorithms lack. These properties make the SDM a very promising IQA algorithm.



(a) Fast fading.

(b) Gaussian blur.

(c) JPEG2000.

(d) JPEG.

(e) White noise.

(f) All distortions.

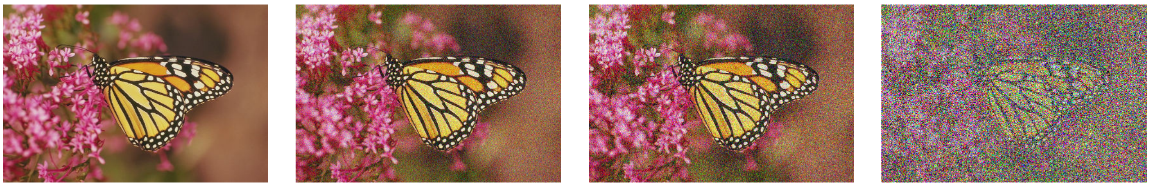
Figure 3.1: Scatter plots of SDM vs DMOS for various distortions. The blue line represents the best fit function.

	FF	Blur	JPEG	JP2K	AWGN	All
SSIM	0.9629	0.9481	0.9266	0.8711	0.9903	0.9298
MSSSIM	0.8499	0.9274	0.9445	0.962	0.9865	0.924
VIF	0.9587	0.976	0.9025	0.9355	0.8852	0.8677
SDM	0.8277	0.9102	0.8188	0.843	0.8913	0.7885

Table 3.1: Performance of the SDM on the LIVE image database measured using SROCC. Also shown are state-of-the-art IQA algorithms.

3.3.2 Intuition

We present an intuitive explanation for the performance of the SDM using images distorted with white noise. Fig. 3.2 and Table 3.2 corroborate the inference made in [4] about requiring a large number of dictionary elements to represent complex signals (for e.g., images corrupted with noise). The loss in sparsity is clearly seen in Table 3.2. As the noise variance increases, $N(Y)$, $N(Y|X)$, and $N(X|Y)$ increase suggesting (expectedly) that noise cannot be sparsely represented. We also observed the opposite effect for blurred images i.e., a decrease in the aforementioned quantities. The other distortion types (fast fading, JPEG and JPEG2000) do not bring about changes to the images that significantly affect their sparsity, there explaining the average performance. These qualitative observations combined with the statistical evaluation in the previous subsection suggest that the SDM is able to quantify departure of images from "naturalness" that correlates fairly well with subjective evaluation.



(a) Original. (b) $\sigma = 0.117$. SDM $= 1.38$. (c) $\sigma = 0.187$. SDM $= 1.43$. (d) $\sigma = 1.0$. SDM $= 1.75$.

Figure 3.2: Intuition behind SDM's performance.

Noise σ	$N(X)$	$N(Y)$	$N(X\|Y)$	$N(Y\|X)$	SDM
0.117	8.441	23.3747	15.2877	43.1803	1.8377
0.187	8.4407	23.9843	16.6767	46.654	1.9531
1.0	8.2673	24.0303	26.8163	50.3037	2.3878

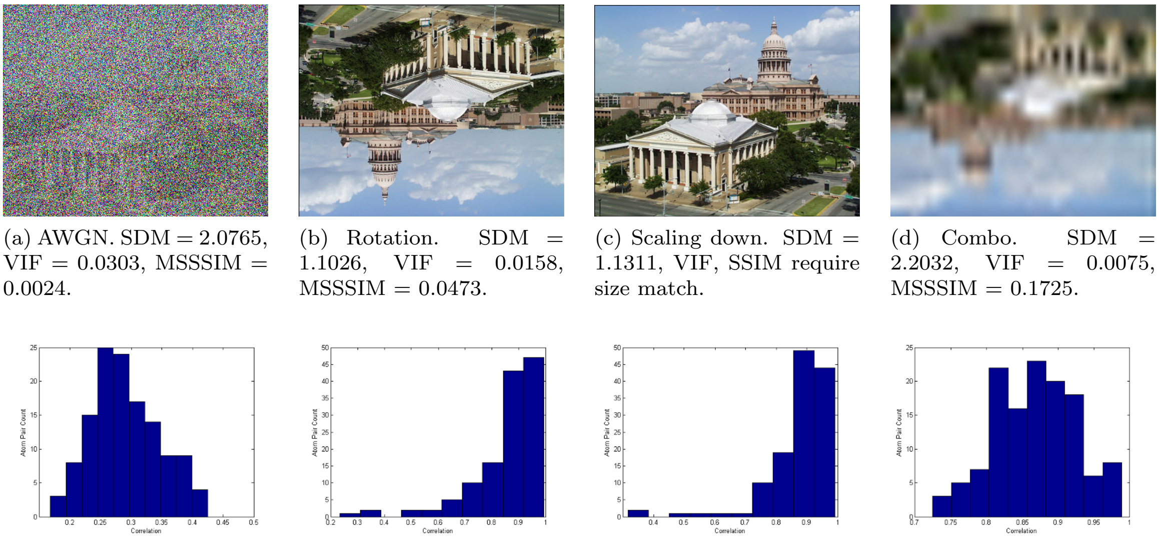
Table 3.2: An intuitive explanation of the SDM's ability to measure image similarity. These values correspond to the images in Fig. 3.2.

3.4 Salient Properties of the SDM

In this section, we demonstrate salient properties of the SDM that make it a very attractive IQA and distinguish it from the state-of-the-art IQAs. Specifically, we demonstrate SDM's robustness to rotation, scaling, and combinations of distortions. The top row of Fig. 3.3 shows various distortions and corresponding SDM, VIF, and MSSSIM scores. It is clear from Figs. 3.3b, 3.3c, and 3.3d that the SDM outperforms both VIF and MSSSIM for the mentioned distortion types. Fig. 3.3 is an illustrative example. This robustness has been consistently observed over a much larger dataset. From Fig. 3.3c, it is worth highlighting that unlike MSSSIM and VIF, the SDM does not require the reference and distorted image sizes to match. It is to be noted that a score close to 1 means low distortion for all the algorithms considered.

We present an empirical explanation of the robustness of SDM to rotation, scaling, and combinations of distortions. The bottom row of Fig. 3.3 shows the histogram of the maximum pairwise correlation between the atoms of the reference and distorted image dictionaries. Let D_R and D_D be the reference and distorted dictionaries respectively. Let $D_R = [a_1^R, a_2^R, \ldots, a_{128}^R]$, $D_D = [a_1^D, a_2^D, \ldots, a_{128}^D]$ where a_i^R is the column vector of size 64 representing the i^{th} atom of the reference dictionary D_R, and a_j^D is the column vector of size 64 representing the j^{th} column of D_D. We construct the correlation matrix R where R_{ij} is the correlation between the a_i^R and a_j^D. The maximum value of row i in R represents the best matching atom in D_D to a_i^R. The histograms in the bottom row of Fig. 3.3 correspond to the row-wise maximum correlation for each distortion type. Note that the histograms in Fig. 3.3 correspond to the images directly above them.

From these histograms, we see that there are a large number of atom-pairs with high correlation (> 0.9) for the distortions where SDM is robust. This can be interpreted as the dictionaries D_R and D_D being composed of atoms that are "similar". This in turn implies that the cross terms in the SDM index $(N(X|Y), N(Y|X))$ would be small and therefore the robustness of the SDM.



(a) AWGN. SDM = 2.0765, VIF = 0.0303, MSSSIM = 0.0024.

(b) Rotation. SDM = 1.1026, VIF = 0.0158, MSSSIM = 0.0473.

(c) Scaling down. SDM = 1.1311, VIF, SSIM require size match.

(d) Combo. SDM = 2.2032, VIF = 0.0075, MSSSIM = 0.1725.

(e) White noise.

(f) Rotation by 180^o.

(g) Scaled down by 0.8.

(h) FF and rotation by 180^o.

Figure 3.3: Robustness of the SDM to rotation, scaling, and a combination of distortions. Top row showing various distortion types. Bottom row showing histogram of maximum correlation between atom pairs formed from reference and distorted image dictionaries.

3.5 Conclusions and Future Work

We have presented a statistical evaluation of the SDM and shown that it performs fairly when compared to the state-of-the-art. However, we have shown that the SDM possesses several useful properties such as robustness to rotation, scaling and distortion combinations that make it appealing in a wider variety of applications than most popular full-reference image quality assessment algorithms. The strength of the SDM as an objective function has already been demonstrated in image classification, clustering and retrieval applications [4].

We believe that the SDM opens up interesting avenues for further investigation in the measurement of image similarity with potential extensions to video similarity as well. As future work, we plan to explore these avenues with a particular emphasis on video similarity measurement.

Chapter 4

No Reference Quality Assessment

4.1 Introduction

The role of automated or objective measurement of image and video quality in today's multimedia-centric society cannot be overemphasised. Algorithms that accurately predict the subjective quality of multimedia data can be used to improve the performance of a wide gamut of multimedia systems ranging from codecs to cross-layer optimization techniques to display design, to name a few. In a majority of settings, the pristine or undistorted content is unavailable for comparison. Blind or no-reference (NR) quality assessment algorithms attempt to estimate the perceptual quality of multimedia content in such a setting. Specifically, we focus on opinion-unaware and distortion-unaware algorithms. By opinion-unaware, we mean algorithms that do not use mean opinion scores (MOS) of subjective evaluation for training. By distortion-unaware, we mean algorithms that are not tailored to specific (known) distortion types.

BIQA algorithms have received a lot of attention in the recent past and several excellent algorithms have been proposed. A non-exhaustive list of the current state-of-the-art methods includes Quality Aware Clustering (QAC) [42], Sparse representation for blind image quality assessment (SRNSS) [43], Natural Image Quality Evaluator (NIQE) [44], Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) [45], probabilistic latent semantic analysis (pLSA) [46], BLind Image Integrity Notator using DCT Statistics-II (BLIINDS-II) , [47], and Distortion Identification-based Image Verity and INtegrity Evaluator (DIIVINE) [48]. The performance of several of these BIQA algorithms is comparable to the state-of-the-art full-reference IQA (FRIQA) algorithms. The most common approach to BIQA relies on constructing "reference" features of images using a training set consisting of either pristine images or a mixture of pristine and distorted images – for e.g., NIQE [44] (distortion agnostic), QAC [42], BRISQUE [45], pSLA [46], DIIVINE [48], BLIINDS [49] (all distortion aware). Features derived from a test image are compared with these "reference" features to compute a quality score. Several BIQA approaches also attempt to learn the relationship between the "reference" features and corresponding subjective scores during the training process – for e.g. in BRISQUE [46], DIIVINE [48], BLIINDS [49], SRNSS [43]. In such methods, the learnt relationship is used to predict the score of a test image given its features.

We briefly discuss two recent state-of-the-art opinion-unaware methods in order to place our algorithm in context. QAC [42] is an opinion-unaware and distortion-aware BIQA algorithm that achieves its opinion-unawareness by replacing subjective scores with FSIM [50], a state-of-the-art FRIQA algorithm. FSIM is applied on overlapping blocks of a small set of pristine images along with their distorted versions to assign quality scores to the distorted blocks. The distorted blocks are clustered into different quality levels and difference of Gaussian features extracted for each block. These features are then clustered in a quality aware manner and their centroids saved in a lookup-table. The quality of a test image is inversely proportional to the Euclidean distance of its features vectors with the centroids of the quality aware clusters. We would like to note that the structure of quality aware clusters corroborate the local oriented receptive fields of area V1 of the visual cortex. QAC also provides a coarse quality map of the image – a first among BIQA methods (to the best of our knowledge).

NIQE [44] is an opinion-unaware and distortion-unaware algorithm that attempts to quantify the unnaturalness in an image. It is based on the hypothesis that the pixel statistics of natural scenes are altered in the presence of distortion. A generalized gaussian density (GGD) is used to model the statistics of mean-subtracted-contrast-normalized pixels of a set of pristine images (chosen from a source that is completely different from the test datasets). The GGD parameters are the features that are in turn modeled using a multivariate gaussian (MVG) model to form the "reference" fit. The quality of a test image is computed by comparing its MVG fit to the "reference" MVG fit. NIQE can be classified as a truly blind BIQA algorithm.

While our proposed algorithms are similar in philosophy to these techniques, it uses a fundamentally different approach to quantify unnaturalness that is based on the sparse representation of natural images.

We proposed two algorithms based on the hypothesis that the HVS perceives distortions when it detects a change in the sparse representation pattern of an image relative to the average sparse representation pattern of pristine natural images. Further, we attempt to quantify this change in the sparse representation pattern of test images. The hypothesis for both the algorithms is same, however the method to quantify the change in sparse representation differs. First we start with a brief discussion on the sparse representation of natural images and then we discuss the proposed algorithm, Sparsity based Image Quality Evaluation-1(SBIQE-1) followed by SBIQE-2 (improvized version of SBIQE-1) in subsequent sections.

4.2 Sparse Representation of Natural Images

The role of natural scene statistics in the understanding of the visual system has been studied by several researchers. It has been conclusively shown that the receptive fields of V1 neurons are tuned to the statistics of natural scenes [51] (and references therein). It is now well-accepted that the primary visual cortex adopts a sparse-coding strategy to represent visual stimulus. The coefficients of such sparse representations of natural scenes are typically uncorrelated, thereby maximizing the amount of information they convey. In their seminal paper, Olshuasen and Field [52] proposed an algorithm to construct overcomplete linear codes or dictionaries that sparsely represent natural scenes. They showed that that the primary visual cortex is modeled well using such overcomplete dictionaries and is used in Full Reference IQA – for e.g. Sparsity based Distance Metric (SDM) [53, 54] and BIQA – for e.g SRNSS [43] algorithms. While SRNSS is an opinion aware method, we focus on an opinion unaware algorithm.

4.3 SBIQE-1

4.3.1 Dictionary Construction

As mentioned in the previous section, we attempt to mimic the behavior of the HVS to measure the amount of unnaturalness or distortion in an image. The first step in our algorithm is the construction of an overcomplete dictionary to sparsely represent natural scenes. While the overcomplete dictionary construction technique in [52] gives good results, we chose to work with the more recent K-SVD [55] algorithm for dictionary construction. The K-SVD algorithm has been shown to outperform other overcomplete signal representations such as wavelets in terms of reconstruction error. Further, the efficacy of the K-SVD algorithm has been demonstrated in a myriad of applications including pattern recognition, denoising and restoration, super-resolution, to name a few.

We construct an overcomplete dictionary consisting of 162 atoms with each atom being an 81-dimensional vector (corresponding to 9×9 image patches). The number of atoms was chosen to be

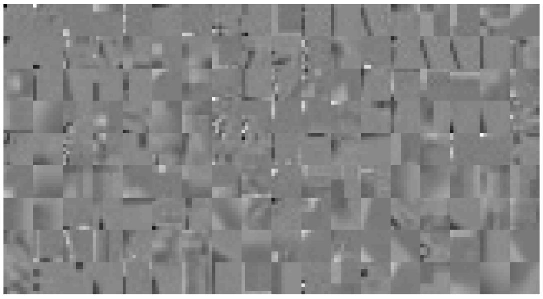


Figure 4.1: The dictionary of atoms constructed from pristine images. Each atom is of size 9×9, with 18 atoms per row and 9 rows in all.

twice the dimensionality of the atoms. The dictionary is constructed using all the patches chosen from 70 pristine images. The patch size (9×9) is chosen to avoid the standard block size of 8×8 used in popular image and video codecs. Each of the 70 images is divided into overlapping patches of size 9×9 with an overlap of 3 pixels in each dimension. Also, these images are chosen from a dataset [56] that has no overlap with any of the popular datasets used for testing (LIVE [33], CSIQ [57], TID [58]). The dictionary construction happens once and can be performed offline. The dictionary with all atoms concatenated is shown in Fig. 4.1. We clearly see the oriented nature of a majority of the atoms along with a few atoms containing lower spatial frequencies.

4.3.2 "Reference" Feature Extraction

The next step in the algorithm is the extraction of "reference" features that are representative of pristine natural images. To this end, we choose images from a set of pristine image at http://live.ece.utexas.edu/research/quality/pristinedata.zip. Again, this source is chosen so as to avoid any overlap with the datasets used for algorithm evaluation. The chosen images are sparsely represented using the constructed overcomplete dictionary. The orthogonal matching pursuit (OMP) [59] algorithm is used for generating the sparse representations. As with dictionary construction, 9×9 patches (with overlap of 3 pixels in both dimensions) from the pristine images are sparsely represented.

The feature vector $\mathbf{f}_r$ is constructed as follows. A histogram of the atoms is constructed and divided by the total number of patches. In other words, we count how many times every atom in the dictionary occurs in the sparse representation of the pristine patches and divide by the total

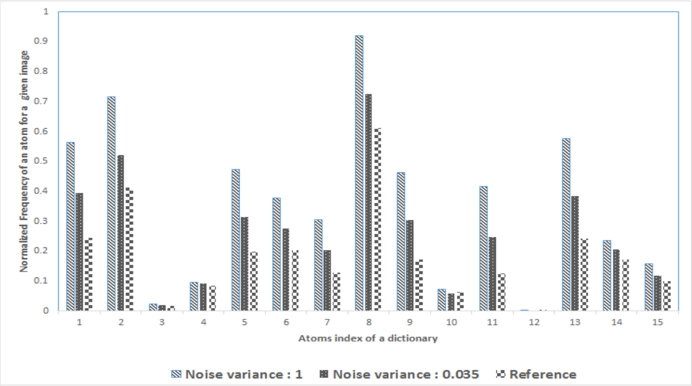


Figure 4.2: Motivation behind SBIQE. Feature vectors of pristine images and images distorted with AWGN. The y-axis represents the magnitude of the feature point while the x-axis represents feature point index. Three noise levels ($\sigma = 1, \sigma = 0.035, \sigma = 0$) are shown to illustrate the discriminability of the feature choice.

number of patches used in this training stage. The feature vector is normalized to get

$$\mathbf{n}_r = \frac{\mathbf{f}_r - \mu_r}{\sigma_r}, \tag{4.1}$$

where μ_r and σ_r are the mean and variance of $\mathbf{f}_r$ respectively. As with the dictionary construction, the "reference" feature vector extraction happens only once and can therefore be done offline as well. The motivation for this choice of feature is that it provides a pattern of natural image representation in the HVS. Further, it possesses good distortion discriminability as illustrated in Fig. 4.2. The figure shows the magnitude of a subset of feature points (i.e., normalized histogram points) from images subject to different levels of additive white Gaussian noise. In this illustration, noise standard deviation σ is chosen to be 1, 0.035 and 0. It is clear that our choice of features is able to differentiate noise levels – especially so when the magnitude of the feature vector element is high. We found this to be true for other common distortions including blur, and compression-induced artifacts.

4.3.3 Image Quality Measurement

Given a test image, it is divided into overlapping blocks (as in Section 4.3.2) and each block is represented using the dictionary constructed in Section 4.3.1. The test feature vector $\mathbf{f}_t$ is constructed from the sparse representation of the overlapping blocks by counting the number of occurrences of each of the dictionary atoms in them. This count is divided by the total number of patches in the images. As with the "reference" feature vector, $\mathbf{f}_t$ is normalized to get

$$\mathbf{n}_t = \frac{\mathbf{f}_t - \mu_t}{\sigma_t}, \tag{4.2}$$

where μ_t and σ_t are the mean and variance of $\mathbf{f}_t$ respectively. Finally, the quality score is computed as

$$Q_t = 1 - \frac{||\mathbf{n}_t - \mathbf{n}_r||_2}{||\mathbf{n}_t||_2 + ||\mathbf{n}_r||_2}. \tag{4.3}$$

The error norm between the test and reference vectors is normalized by the sum of norm of the reference and test vectors. From the triangle inequality, it follows that

$$0 \leq Q_t \leq 1. \tag{4.4}$$

Higher values of Q_t (close to 1) correspond to better quality while lower values (close to 0) reflect poor quality or high distortion. We would also like to note from the definition of our feature that each of the feature points is non-negative.

4.4 SBIQE-2

4.4.1 Dictionary Construction

Same as dictionary construction in SBIQE-1.

4.4.2 "Reference" Parameters

The next step in the algorithm is the evaluation of "reference" parameters that are representative of pristine natural images. To this end, we choose images from a set of pristine image at http://live.ece.utexas.edu/research/quality/pristinedata.zip. Again, this source is chosen so as to avoid any overlap with the datasets used for algorithm evaluation. The chosen images are sparsely represented using the constructed overcomplete dictionary. The orthogonal matching pursuit (OMP) [59] algorithm is used for generating the sparse representations. As with dictionary construction, 9×9 patches (with overlap of 3 pixels in both dimensions) from the pristine images are sparsely represented. Then the empirical distribution of sparse representation evaluated from the pristine images corresponding to each atom is constructed and we found that a generalized Gaussian distribution (GGD) can be used to effectively capture the statistics of sparse representation corresponding to each atom, which often exhibit changes in the kurtosis of the empirical coefficient distributions [60] where the GGD with zero mean is given by

$$p(x;\alpha,\sigma^2) = \frac{\alpha}{2*\beta*\Gamma(1/\alpha)}\, exp\Big(- \Big(\frac{|x|}{\beta}\Big)^{\alpha}\Big)$$

where

$$\beta = \sigma \sqrt{\frac{\Gamma(1/\alpha)}{\Gamma(3/\alpha)}}$$

and $\Gamma(.)$ is the gamma function:

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

The shape parameter α controls the shape of the distribution while σ^2 control the variance. We choose the zero mean distribution, since sparse coefficients are symmetric centered around zero. The parameters of the GGD $(\alpha_i, {\sigma_i}^2)$,where $i = 1...M$ and M is the number of atoms in the dictionary, are estimated using the moment-matching based approach proposed in [60]. As with the dictionary construction, the "reference" parameters evaluation happens only once and can therefore be done offline as well.

4.4.3 Image Quality Evaluation

Given a test image, it is divided into overlapping blocks (as in Section 4.4.2) and each block is sparsely represented using the dictionary constructed in Section 4.3.1. Probability of this sparse representation corresponding to each atom is calculated given the reference parameters obtained as discussed in Section 4.4.2. Then the final probability of each block in the image is the product of individual probabilities corresponding to each atom and is given as below.

$$pb_j = \prod_{i=1}^{M} p_{ij}$$

where pb_j is final probability for each block in the image, p_i is probability corresponding to each atoms and M is number of atoms in the dictionary.

Quality is evaluated as mean of probabilities of all overlapping blocks in the image as given below.

$$Q = \frac{1}{N} \sum_{j=1}^{N} pb_j$$

where N corresponds to number of overlapping blocks in the given test image.

Higher values of Q correspond to better quality while lower values reflect poor quality or high distortion. We hypothesize that if the sparse representation of the block is drawn from the probability distribution of "reference" space, then the probability of drawing that block from the "reference" space is higher.

With this algorithm, we can have block level quality map. To the best of our knowledge, we are the first one to evaluate the quality map in NR setting in opinion and distortion unaware scenario.

Local Quality Estimation

Our algorithm measures the quality on image patches, so it can be used to detect low/high quality local regions as well as giving a global score for the entire image. To demonstrate, we select an image from LIVE dataset [33] and the corresponding distorted images including WN, BLUR, JPEG and JP2K. We then perform local quality estimation on these images using our algorithm. Figure 4.3 shows estimated quality map on these images. We can see that our algorithm distinguishes the clean and the distorted parts of each image.

4.5 Results and Discussion

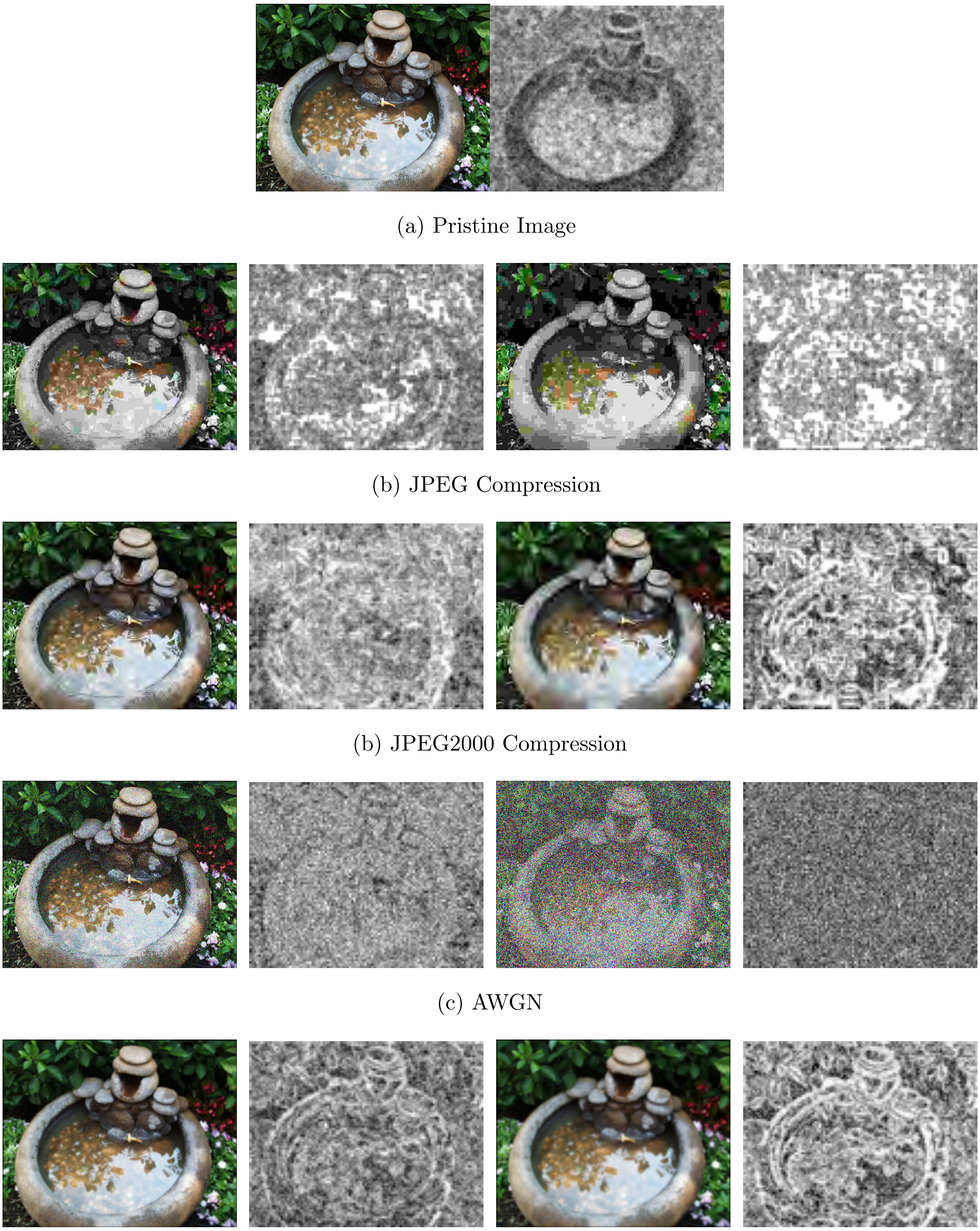
We present the results of our algorithms and compare it with state-of-the-art BIQA methods. BRISQUE [46], an opinion-aware distortion-aware method, QAC [42] an opinion-unaware distortion-aware method and NIQE [44] an opinion and distortion unaware method are used as the benchmarks for our comparison. The numbers for BRISQUE are quoted for the case of 80% samples used training and the rest used for testing.

The performance of the algorithm on the LIVE [33], CSIQ [57], TID [58] datasets are presented in Tables 4.1, 4.2, 4.3 respectively. For brevity, we only present Spearman rank ordered correlation coefficient (SROCC) values.

	AWGN	Blur	JPEG	JP2K	All
BRISQUE	0.99	0.98	0.92	0.94	0.94
QAC	0.96	0.91	0.94	0.85	0.88
NIQE	0.97	0.93	0.94	0.91	0.91
SBIQE-1	*0.96*	*0.89*	*0.73*	*0.83*	*0.76*
SBIQE-2	*0.98*	*0.94*	*0.82*	*0.86*	*0.87*

Table 4.1: Performance (SROCC) on the LIVE database.

From these results we see that the proposed method compares reasonably with the current state-of-the-art. We also studied the effect of the atom size on performance and found no significant change when atom size was varied from 9×9 to 15×15. At this point, we would like to note (again) that we excluded the block/atom size of 8×8 on purpose so as to avoid overlapping with the typical block size used in standard image codes and therefore attempting to capture quantization artifacts at block boundaries.



(a) Pristine Image

(b) JPEG Compression

(b) JPEG2000 Compression

(c) AWGN

(d) Gaussian blur

Figure 4.3: Local Quality Estimation Results. Brighter pixels indicate distortion.

	AWGN	Blur	JPEG	JP2K	All
BRISQUE	0.92	0.87	0.88	0.86	0.87
QAC	0.86	0.85	0.91	0.87	0.86
NIQE	0.81	0.87	0.86	0.89	0.88
SBIQE-1	*0.96*	*0.89*	*0.73*	*0.83*	*0.76*
SBIQE-2	*0.93*	*0.84*	*0.80*	*0.85*	*0.84*

Table 4.2: Performance (SROCC) on the CSIQ database.

	AWGN	Blur	JPEG	JP2K	All
BRISQUE	0.92	0.79	0.88	0.90	0.85
QAC	0.70	0.85	0.89	0.88	0.87
NIQE	0.78	0.82	0.86	0.90	0.78
SBIQE-1	*0.96*	*0.89*	*0.73*	*0.83*	*0.76*
SBIQE-2	*0.69*	*0.83*	*0.71*	*0.92*	*0.81*

Table 4.3: Performance (SROCC) on the TID database.

We would like to highlight features of the proposed method that make it an interesting and promising direction for exploration. Firstly, the proposed method is both opinion-unaware and distortion-unaware and is inspired by the sparse representation of natural scenes in the HVS.

4.6 Conclusions and Future Work

We presented a novel sparsity-based blind image quality assessment algorithm that is inspired by the sparse representation of natural scenes in the HVS. We hypothesized that a change in the sparse representation pattern of a given image relative to pristine image sparse patterns is a measure of unnaturalness or distortion. We quantified this change in sparsity and showed that it is indeed a measure of the perceptual quality of an image. We do recognize that in its current form, our algorithm is subpar (overall) relative to the state-of-the-art. However, we strongly believe that the initial results are promising and the proposed method has a number of attractive features.

As future work, we plan to improve the performance of the algorithm by fine-tuning the features and score computation metrics. Also, we intend to extend this hypothesis to no-reference video quality assessment.

Chapter 5

Face Quality Assessment

5.1 Introduction

Face Recognition has received substantial attention due to its value both in understanding how the face recognition process works in humans as well as in addressing many applications, including access control, video surveillance, entertainment and law enforcement. Since face recognition is the natural way of identification and verification, this field is rich with excellent literature [14, 15, 16]. In the last two decades various algorithms have been proposed for face recognition based on still images and video sequences. However, in realistic scenarios, face recognition is limited by low quality images and variation in pose, illumination, occlusion and expression in the acquired face image [15]. Such problems are even more severe in surveillance systems where users may be uncooperative and the environment is uncontrolled. Since poor quality images in the surveillance video sequences offer very little information for face recognition, they not only increase the computational load because of complex processes such as feature extraction and matching, but also reduce the recognition accuracy because of outliers. To address this problem, many algorithms have been proposed in recent years to select the subset of high quality faces and avoiding outliers.

To the best of our knowledge, Berrani et al. [61] were the first one to address this issue by using statistical approaches to remove outliers. However, this approach does not work when all the images are of poor quality and is a common scenario in surveillance. There are many algorithms proposed based on the facial properties such as estimating the pose [62] to evaluate the quality of face, calculating the asymmetry of the face by estimating out of plane rotation and non-frontal illumination to quantify the degradation of the quality [63, 64, 65]. The above methods consider only a subset of factors affecting face recognition and hence not suitable for robust image selection. Instead of considering the factors affecting face recognition and fusing the scores, Wong et al. proposed the definition of ideal face as frontal faces with uniform illumination and there by simultaneously considering the variations in pose, sharpness and alignment errors [66]. Inspired by this definition we propose a sparsity-based face quality assessment algorithm to quantify the quality of faces, that allows us to select the best subset of high quality images for further processing. But we realized this definition of quality has limited applications since it is based on assumptions on standard face. Also, these algorithms do not leverage the strengths of FR algorithms i.e., the above algorithms do not fully utilize the abilities of FR algorithms that may be good at even recognizing faces with occlusions, pose variations or non-uniform illumination.

Chen et al. [67] proposed an algorithm based on multiple feature fusion and learning to rank were the first one to address the above issue. In this algorithm, they considered three databases with faces acquired in a controlled environment, an uncontrolled environment, and with non-face images respectively. Then they ranked the databases based on the recognition performance and assumed the faces in the same database to have equal rank. Their algorithm is implemented in two levels. In the first level, they learned the weights for the feature vector of face images from the above-mentioned datasets by using a linear kernel such that the sum of weighted feature resembles the ranking of databases. For this, they considered five different feature vectors and learned the corresponding weights. In the second level, they combined five first level scores with respect to each feature vector by using a second order polynomial kernel to give the final quality score for the given probe face image.

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

The chapter proceeds as follows: Section 5.2 describes our proposed algorithms(Sparsity-based, CNN based FQA) in the framework of FR System. Section 5.3 discusses our experiments on Choke-Point dataset [66] and compares our algorithm with other face selection algorithms. Section 4.6 concludes the discussion and gives the direction for future work.

5.2 Face Recognition System

A typical FR system comprises of face detection, face localization, face subset selection (optional), face feature extraction and face matching components as shown in the Fig.5.1 [15]. In the proposed algorithm, the face subset selection component is analysed and the best subset of faces are selected for further processing in order to enhance the performance of FR. The details of each component is presented below.

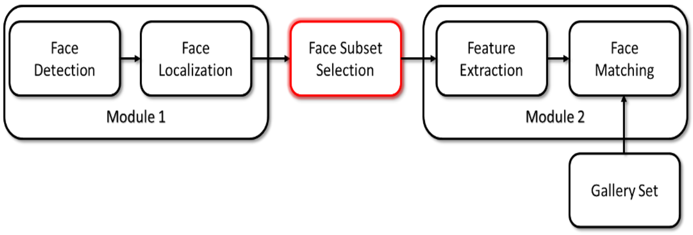


Figure 5.1: FR System

5.2.1 Face Detection and Localization

To detect and localize the face in each frame, we used Viola Jones Haar feature based cascade classifier [68]. We omitted the faces where the Haar cascade classifier is not able to detect the face. After detecting the facial region from the aforementioned classifier, localization is done by fixing the center and cropping the facial region by omitting borders. The final cropped images are then resized to 64 × 64 pixels.

5.2.2 Face Subset Selection

It is quite common in a video surveillance scenario to acquire multiple face images of same person. Selecting the subset of faces with high quality improves the performance of recognition algorithm by removing the outliers. It also reduces the complexity of FR algorithm, considering the fact that face feature extraction process is computationally expensive and complex. As discussed in section 5.1, it is difficult to define the quality of a face image. Several researchers have defined the quality in different ways. In this chapter we consider two definitions of quality for subset selection. One is based on standard/ideal face assumption and we called it as sparsity based FQA. Another one is adaptive i.e with respect to the FR algorithm under consideration and we called it as CNN based FQA. We show the supremacy of performance of later definition to former one.

A. Sparsity based FQA

For the face in a sequence to be selected for further processing, the face must be of high quality i.e the face should be comparable with “ideal” face. Here, “ideal” face means the frontal face with uniform illumination and neutral expression. The given test face is compared with “ideal” face as specified above and assigned a rank of face in a video sequence. Finally the faces with high rank are chosen for feature extraction and matching. In order to achieve this, we propose a sparsity-based face quality assessment algorithm.

The proposed algorithm is based on the hypothesis that average sparseness of the probe face will be altered if the probe face is not similar to “ideal” face. This hypothesis is inspired from our previous work sparsity-based image quality assessment algorithm [69] . We describe our algorithm in the following subsections starting with a discussion of construction of “ideal” face and “reference” feature. These two steps are one time processes and can be done offline. For this, we make use of ’fa’ subset of FERET database which contains frontal faces with neutral expression and uniform illumination [70]. This subset is further divided into two sets with 80% of faces for construction of “ideal face” and remaining faces for construction of “reference” feature. The motivation for this choice of feature vector is to represent a pattern for “ideal” face.

Construction of “ideal” face: The first step in our algorithm is to construct the “ideal” face. To construct the “ideal” face, faces of ’fa’ subset of FERET database are properly aligned using the eye coordinates and are closely cropped and then resized to 64×64 pixels. All faces are log transformed to reduce the dynamic range of pixel intensities. In order to remove the person specific content and to get the holistic information of “ideal” face, high frequency component needs to be removed. To achieve this, we used Daubechies db1 wavelet to decompose the image and consider the low frequency region of the face for further processing. Then the low frequency region of the face is divided into overlapping patches of size 8×8 pixels with overlap of 5 pixels in each dimension. By considered patches from all faces, an overcomplete dictionary consisting of 128 atoms of dimension 64 (corresponding to 8×8 patches) is constructed to sparsely represent the patches and hence in total we have N overcomplete dictionaries for N patches. For this, we chose to work with the K-SVD algorithm [71] which performs well in many applications including super-resolution, image quality assessment, denoising, pattern recognition, to name a few. The whole process of constructing dictionaries related to “ideal” face is shown in the Fig.5.2.

Construction of “reference” feature: The next step in the algorithm is the extraction of “reference” features that are representative of “ideal” face and the whole process is shown in the Fig. 5.3. To this end, we choose the remaining 20% of faces from fa subset of FERET database. The chosen images are log transformed and considered low frequency region after wavelet decomposition. As with dictionary construction, $8{\times}8$ patches (with overlap of 5 pixels in both dimensions) from

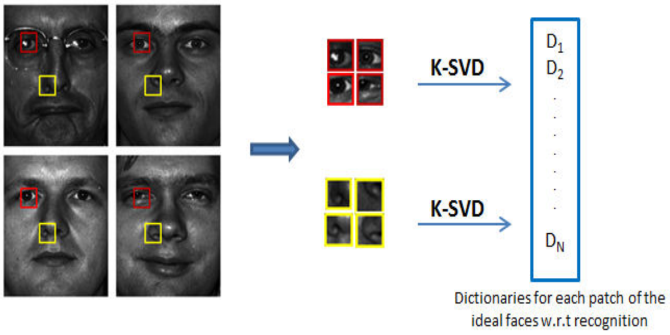

K-SVD
K-SVD
D1
D2
DN
Dictionaries for each patch of the ideal faces w.r.t recognition

Figure 5.2: Construction of “ideal” face

the chosen images are sparsely represented using the orthogonal matching pursuit (OMP) algorithm [19] with the corresponding constructed dictionary. Then the feature vector (f_{rn}, where $n = 1 \dots N$ and N is number of patches) for each patch is constructed by averaging the sparse coefficients over all the chosen images and in total we have N “reference” feature vectors.

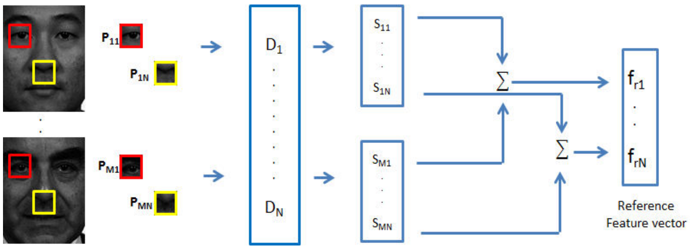


Figure 5.3: Construction of “reference” feature vector

Face Quality Evaluation: This is the final step in subset selection. Here, we make use of the N dictionaries and N “reference” feature vectors obtained from the training to assign the rank of face in a probe video sequence. The whole process of quality evaluation is shown in the Fig. 5.4. A given face image is log transformed first followed by a wavelet decomposition. The low frequency subbands are divided into overlapping patches are mentioned previously. Each patch in the face is sparsely represented using OMP with corresponding constructed dictionary and in total we have N test feature vectors (f_{tn}) where $n = 1 \dots N$ and N is number of patches. The assumption is that if the patch is sparsely represented using dictionary of “ideal” face, then the corresponding patch is similar to corresponding patch of “ideal” face. Finally, the quality score is computed as

$$Q = \frac{1}{N} \sum_{n=1}^{N} \|f_{rn} - f_{tn}\|_2^2.$$

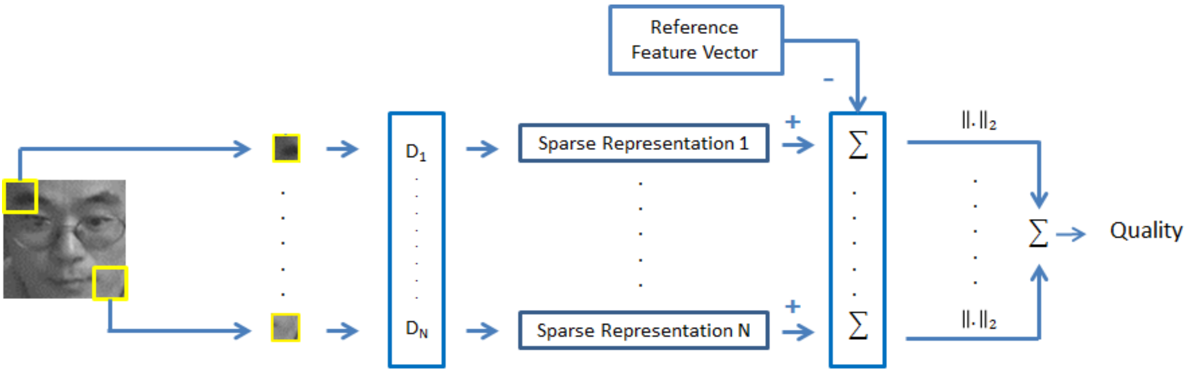


Figure 5.4: Overview of Face Quality Evaluation Algorithm

Image with low quality score indicates high quality image and vice versa. Then, according to the quality scores of each face in the video sequence, ranking will be given. After ranking the faces in video sequence, we select only top M faces for further processing and is shown in the Fig. 5.5.

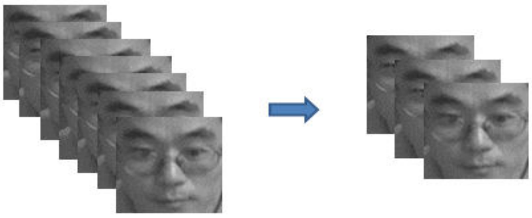


Figure 5.5: Face Selection based on quality of face w.r.t "ideal" face

B. CNN based FQA

As discussed in section 5.1, it is difficult to define the quality of a face image. We define the quality with respect to the FR algorithm. The motivation behind our definition of the quality is explained as follows by considering an example. If an FR algorithm is good at recognizing the faces with pose variations but not able to recognize the faces with non-uniform illumination, then the faces with pose variations should be considered high quality and faces with non-uniform illumination should be considered low quality. Since different FR algorithms work better in different aspects such as occlusion, pose variations, illumination variations, hence fixing the definition of quality doesn't take the full advantage of the FR algorithm under consideration. So, we propose a novel take on face quality definition and choose to define the quality of the face image with respect to the FR algorithm.

For convenience, we categorize the FR system into two modules. The first module consists of face detection and face localization. The second module is the FR algorithm that consists of face feature extraction and face matching. As an FQA algorithm, we need to predict the face images that performs best in the second module of the given FR algorithm and this is not a trivial problem. Toward this end, we considered the second module as an unknown system (or a black box) and have attempted to model its system response using a CNN. To achieve this, we first used a training set of images to evaluate the system performance of the FR algorithm. Thus, by knowing the input and output to the black box, we model its performance such that it is able to predict the quality of test images a priori. By this, we could select the high quality face images that are best recognized by the FR algorithm. A CNN is used to model the performance of the FR algorithm due to its strengths over other modelling techniques. Since a CNN accepts the entire 2D image as input, there is no need for complicated image transformation or feature extraction. This feature is particularly useful in our case where the definition of quality of face image is not fixed. So, the CNN learns its parameters and defines the quality of the face image depending on the FR algorithm. The proposed algorithm is depicted in figure 5.7 and framework of CNN is explained as follows.

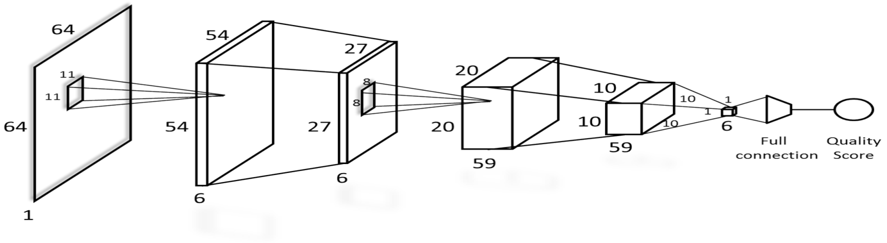


Figure 5.6: Our CNN Architecture

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

As a preliminary implementation, our CNN has three convolution layers with sub sampling. The first convolutional layer has 6 kernels with each of size 11 × 11 and produces 6 feature maps with size 54 × 54, followed by sub sampling layer. The second convolutional layer has 59 kernels with size 8 × 8 and produces 59 feature maps with size 20 × 20 and then followed by sub sampling layer. Third convolutional layer has 6 kernels each of size 10 × 10 and then followed by linear regression with one dimensional output that gives the quality score of the face image. The architecture of our CNN is shown in the figure 5.6.

In our experiments, we use the ChokePoint dataset [66] that is ideally suited for face recognition/verification in the surveillance scenario. The dataset consists of 25 subjects (19 male and 6 female) with 64,204 face images. We divide the images into training and testing sets. Set 1 contains image sequences of 13 subjects for training the CNN and Set 2 contains the rest of the images sequences to evaluate the performance of the FR algorithm.

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

Let I_n be the preprocessed input face image, q_n be the quality/MSM score and $f(I_n, W)$ be the predicted score of the input face image where W is the weight matrix of the network. Then, back propagation and stochastic gradient descent is used to minimize the following cost function:

$$S = \frac{1}{N} \sum \left\| f(I_n, W) - Q_n \right\|_2,$$

where N is the total number of images in the training set. We used a batch size of 50 and ran the CNN for 50 epochs. The entire training process is depicted in Fig. 5.7.

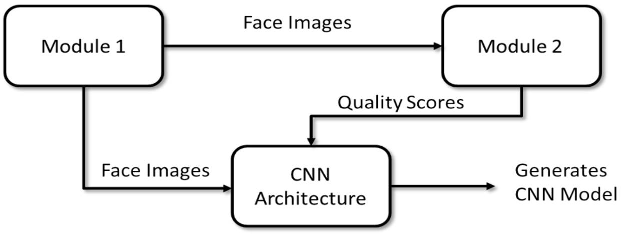


Figure 5.7: Training

2. Testing: The preprocessed face images in the probe sequence is given as the input to the trained CNN and quality scores are predicted for each face image in the sequence. The CNN model is trained such that these scores resemble the MSM scores of the FR algorithm. Subset selection is done by sorting the predicted quality scores of face image sequence and taking the top N images from the sorted list. Then these selected N images are given to the FR algorithm for further processing. The subset selection process is illustrated in Fig. 5.8.

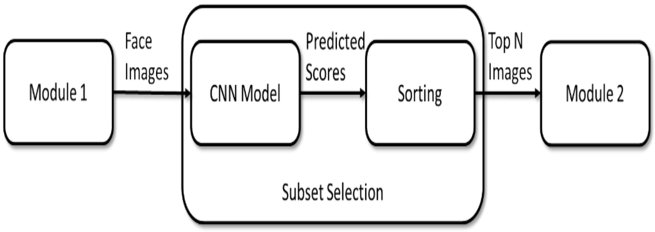


Figure 5.8: Testing

5.2.3 Face Feature Extraction

After the subset of faces are selected, features are extracted for each face using Local Binary Pattern (LBP) [72], which makes use of both shape and texture information of the face. Each face region is sub-divided into small regions and LBP features are extracted for each region and is concatenated to form the feature vector of the face. This feature vector is able to achieve three levels of locality by considering patterns at pixel level, features at regional levels and concatenated feature vector at global level.

5.2.4 Image Set Matching

We make use of Mutual Subspace Method (MSM) for face image set matching [73]. The two image sets are considered similar if the canonical angle between two image sets is within the threshold. For each video sequence in the gallery set, feature vectors are calculated and compared with feature vectors of probe sequence using MSM. In MSM, the probe sequence and each video sequence in gallery set is considered as separate subspaces and similarity between the subspace is measured by calculating the mean canonical angle between the two subspaces. Let the probe sequence be N dimensional subspace and each video sequence in the gallery set be M dimensional subspace. The canonical angle for each element of the probe sequence is defined as the maximum angle between the given element and all the elements of the M dimensional subspace for a given video sequence in the gallery. The final similarity score is calculated by taking the mean of canonical angle of all the elements of the probe sequence. The score is then compared to a threshold and final decision is made whether the probe and the gallery sequence pair is matched/mismatched pair. The threshold is obtained from a labeled set at which the total number of false positives and false negatives is minimum. This is referred as Minimum Error Rate (MER).

5.3 Results and Discussion

In this section, we present the performance of proposed algorithms for good quality subset selection to improve the performance of face recognition algorithm. As a preliminary experiment, we used LBP as a facial feature and MSM for face matching. To reiterate, any face recognition algorithm can be used in conjunction with proposed face quality assessment algorithm for improved accuracy and the computational efficiency. For FR algorithm, the test set for evaluating the performance of FR algorithm is split further into two sets G1 and G2 where each dataset plays the role of development and evaluation set. The subjects which are used for training the CNN model is not used here to evaluate the performance of the FR algorithm. We performed our experiment in two phases. In first phase, G1 is considered as development set and G2 as evaluation set and roles of G1 and G2 are reversed in second phase. By considering one group as development set (labeled set), we calculated matched and mismatched scores. Then, we find the threshold where the sum of False Acceptance Rate (FAR) and False Rejection Rate (FRR) is minimum i.e Minimum Error Rate. By applying this threshold on the scores of pairs of evaluation set, recognition rate(R_{G2}) is calculated as follows

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

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

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

For subset selection, we selected a N high quality images for face recognition based on different selection metric and characterized how different metrics improve the recognition performance. Along with the proposed selection methods, we have considered four selection methods to compare the performance: (1) sequential selection, (2) random selection, (3) quality assessment based on patch-based probabilistic approach [66], (4) Learning to Rank based quality assessment[67], (5) Sparsity based FQA (6) CNN based FQA. After face selection, the aforementioned protocol is used to compare the results by varying N from 4 to 16 and the results are presented in Table 5.1.

From the results, we can infer that high verification performance is still achieved with subset of faces which means that the proposed method is able to select the best subset of faces from the sequence of faces. In patch-based probabilistic approach and Sparsity based FQA, the subset selection assumes the face images that can be recognized by FR algorithm only when they resembles standard faces which may not be true in all cases. Rank based approach makes use of five feature extraction algorithm which is an expensive step (the step which motivated the researchers to work on subset selection). Also, they considered the same rank for all the images in a single database which may not be the true in all cases. In the proposed CNN based FQA algorithm, we can use our FQA implementation in conjunction with any FR algorithm and fully leverage the ability of the FR algorithm.

Table 5.1: Video-based FR performance on the ChokePoint dataset, using LBP and MSM (higher is better).

Subset Selection Method	N=4	N=8	N=16
Sequential	0.6114	0.6174	0.6278
Random	0.6825	0.691	0.704
Probabilistic based [66]	0.7027	0.7139	0.7234
Rank based [67]	0.7328	0.7511	0.7645
Sparsity based FQA	0.701	0.7144	0.726
CNN Method FQA	0.7231	0.7519	0.7601

5.4 Conclusions and Future Work

In this chapter, we presented a novel sparsity based face quality assessment algorithm. In this, We hypothesized that the face images that are not similar to “ideal” face loses its sparseness when represented by dictionary constructed from “ideal face”. We quantified this change in sparsity and assigned the ranks to faces. But we realized that quality definition should be adpative and we were motivated by the fact that since different algorithm have different capabilities in recognition, quality of the image should be defined with respect to the FR algorithm. To achieve this, a CNN is used to define the quality by training its network with the score/value that depicts the performance of the FR algorithm in consideration. By using this quality measure to sort the input sequence and taking only high quality images we successfully demonstrated that it not only increases the recognition accuracy but also reduces the computational complexity. From the initial results, we strongly believe that the proposed algorithm is promising and has attractive features. As part of future work, we plan to improve the performance of the algorithm by fine-tuning the parameters.

References

[1] J. Lubin. A human vision system model for objective picture quality measurements .

[2] H. R. Sheikh and A. C. Bovik. Image information and visual quality. *Image Processing, IEEE Transactions on* 15, (2006) 430–444.

[3] Z. Wang and A. C. Bovik. Mean squared error: love it or leave it? A new look at signal fidelity measures. *Signal Processing Magazine, IEEE* 26, (2009) 98–117.

[4] T. Guha and R. K. Ward. Image similarity using sparse representation and compression distance. *arXiv preprint arXiv:1206.2627* .

[5] T. Guha, R. K. Ward, and T. Aboulnasr. Image similarity measurement from sparse reconstruction errors. In Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, 2013 1937–1941.

[6] T. Guha, E. Nezhadarya, and R. K. Ward. Sparse representation-based image quality assessment. *Signal Processing: Image Communication* 29, (2014) 1138–1148.

[7] B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision research* 37, (1997) 3311–3325.

[8] Z. P. Sazzad, Y. Kawayoke, and Y. Horita. No reference image quality assessment for JPEG2000 based on spatial features. *Signal Processing: Image Communication* 23, (2008) 257–268.

[9] X. Feng and J. P. Allebach. Measurement of ringing artifacts in JPEG images. In Electronic Imaging 2006. International Society for Optics and Photonics, 2006 60,760A–60,760A.

[10] Z. Wang, A. C. Bovik, and B. Evan. Blind measurement of blocking artifacts in images. In Image Processing, 2000. Proceedings. 2000 International Conference on, volume 3. Ieee, 2000 981–984.

[11] M. Jung, D. Le, M. Gazalet et al. Univariant assessment of the quality of images. *Journal of Electronic Imaging* 11, (2002) 354–364.

[12] C. Charrier, G. Lebrun, and O. Lezoray. A machine learning-based color image quality metric. In Conference on Colour in Graphics, Imaging, and Vision, volume 2006. Society for Imaging Science and Technology, 2006 251–256.

[13] T. Brandão and M. P. Queluz. No-reference image quality assessment based on DCT domain statistics. *Signal Processing* 88, (2008) 822–833.

[14] J. R. Barr, K. W. Bowyer, P. J. Flynn, and S. Biswas. Face recognition from video: A review. *International Journal of Pattern Recognition and Artificial Intelligence* 26.

[15] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *Acm Computing Surveys (CSUR)* 35, (2003) 399–458.

[16] Y. Wong, M. T. Harandi, and C. Sanderson. On robust face recognition via sparse coding: the good, the bad and the ugly .

[17] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM journal on scientific computing* 20, (1998) 33–61.

[18] J. A. Tropp and S. J. Wright. Computational methods for sparse solution of linear inverse problems. *Proceedings of the IEEE* 98, (2010) 948–958.

[19] J. A. Tropp and A. C. Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *Information Theory, IEEE Transactions on* 53, (2007) 4655–4666.

[20] M. Aharon, M. Elad, and A. Bruckstein. K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation. *Signal Processing, IEEE Transactions on* 54, (2006) 4311–4322.

[21] R. Rubinstein, A. M. Bruckstein, and M. Elad. Dictionaries for sparse representation modeling. *Proceedings of the IEEE* 98, (2010) 1045–1057.

[22] J. W. Cooley and J. W. Tukey. An algorithm for the machine calculation of complex Fourier series. *Mathematics of computation* 19, (1965) 297–301.

[23] C. M. Bishop et al. Pattern recognition and machine learning, volume 4. springer New York, 2006.

[24] T. S. Lee. Image representation using 2D Gabor wavelets. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 18, (1996) 959–971.

[25] S. Lesage, R. Gribonval, F. Bimbot, and L. Benaroya. Learning unions of orthonormal bases with thresholded singular value decomposition. In Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on, volume 5. IEEE, 2005 v–293.

[26] C. M. Bishop et al. Neural networks for pattern recognition .

[27] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *Image Processing, IEEE Transactions on* 13, (2004) 600–612.

[28] Z. Wang, E. P. Simoncelli, and A. C. Bovik. Multiscale structural similarity for image quality assessment. In Signals, Systems and Computers, 2003. Conference Record of the Thirty-Seventh Asilomar Conference on, volume 2. IEEE, 2003 1398–1402.

[29] R. Soundararajan and A. C. Bovik. RRED indices: reduced reference entropic differencing for image quality assessment. *Image Processing, IEEE Transactions on* 21, (2012) 517–526.

[30] A. Mittal, A. Moorthy, and A. Bovik. No-reference image quality assessment in the spatial domain .

[31] T. Guha and R. K. Ward. IMAGE SIMILARITY MEASUREMENT FROM SPARSE RECONSTRUCTION ERRORS. In Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on. 2013 .

[32] T. Guha, E. Nezhadarya, and R. K. Ward. Sparse Representation-based Image Quality Assessment. *arXiv preprint arXiv:1306.2727* .

[33] H. Sheikh, Z. Wang, L. R. Cormack, and A. C. Bovik. LIVE Image Quality Assessment Database Release 2.

[34] H. R. Sheikh, M. F. Sabir, and A. C. Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *Image Processing, IEEE Transactions on* 15, (2006) 3440–3451.

[35] K. Manasa Priya, K. Manasa, and S. S. Channappayya. A statistical evaluation of Sparsity-based Distance Measure (SDM) as an image quality assessment algorithm. In Acoustics, Speech

and Signal Processing (ICASSP), 2014 IEEE International Conference on. IEEE, 2014 2789–2792.

[36] M. Li. An introduction to Kolmogorov complexity and its applications. Springer, 1997.

[37] M. Li, X. Chen, X. Li, B. Ma, and P. M. Vitányi. The similarity metric. *Information Theory, IEEE Transactions on* 50, (2004) 3250–3264.

[38] N. Nikvand and Z. Wang. Generic image similarity based on Kolmogorov complexity. In Image Processing (ICIP), 2010 17th IEEE International Conference on. IEEE, 2010 309–312.

[39] M. Aharon, M. Elad, and A. Bruckstein. K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation. *Signal Processing, IEEE Transactions on* 54, (2006) 4311–4322.

[40] H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik. LIVE image quality assessment database release 2 2005.

[41] A. M. Rohaly, P. J. Corriveau, J. M. Libert, A. A. Webster, V. Baroncini, J. Beerends, J.-L. Blin, L. Contin, T. Hamada, D. Harrison et al. Video quality experts group: Current results and future directions. In Visual Communications and Image Processing 2000. International Society for Optics and Photonics, 2000 742–753.

[42] W. Xue, L. Zhang, and X. Mou. Learning without Human Scores for Blind Image Quality Assessment. In Computer Vision and Pattern Recognition (CVPR), 2013. IEEE Conference on. IEEE, 2013 995–1002.

[43] L. He, D. Tao, X. Li, and X. Gao. Sparse representation for blind image quality assessment. In Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE, 2012 1146–1153.

[44] A. Mittal, R. Soundararajan, and A. Bovik. Making a Completely Blind Image Quality Analyzer. *Signal Processing Letters, IEEE* 20, (2013) 209 – 212.

[45] A. Mittal, A. Moorthy, and A. Bovik. No-reference image quality assessment in the spatial domain. *Transactions on Image Processing, IEEE* 21, (2012) 4695 – 4708.

[46] A. Mittal, G. S. Muralidhar, J. Ghosh, and A. C. Bovik. Blind image quality assessment without human training using latent quality factors. *Signal Processing Letters, IEEE* 19, (2012) 75–78.

[47] M. A. Saad, A. C. Bovik, and C. Charrier. Blind image quality assessment: A natural scene statistics approach in the DCT domain. *Image Processing, IEEE Transactions on* 21, (2012) 3339–3352.

[48] A. Moorthy and A. Bovik. Blind Image Quality Assessment: From Natural Scene Statistics to Perceptual Quality. *Image Processing, IEEE Transactions on* 20, (2011) 3350 –3364.

[49] M. Saad, A. Bovik, and C. Charrier. A DCT Statistics-Based Blind Image Quality Index. *Signal Processing Letters, IEEE* 17, (2010) 583 –586.

[50] L. Zhang, L. Zhang, X. Mou, and D. Zhang. FSIM: a feature similarity index for image quality assessment. *Image Processing, IEEE Transactions on* 20, (2011) 2378–2386.

[51] J. M. Bower. 20 Years of Computational Neuroscience. Springer, 2013.

[52] B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* 381, (1996) 607–609.

[53] T. Guha and R. K. Ward. On image similarity, sparse representation and kolmogorov complexity .

[54] K. Manasa Priya, K. Manasa, and S. S. Channappayya. A statistical evaluation of Sparsity-based Distance Measure (SDM) as an image quality assessment algorithm. In Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on. IEEE, 2014 2789–2792.

[55] M. Aharon, M. Elad, and A. Bruckstein. K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation. *Signal Processing, IEEE Transactions on* 54, (2006) 4311–4322.

[56] J. Yang, J. Wright, T. Huang, and Y. Ma. Image super-resolution as sparse representation of raw image patches. In Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. IEEE, 2008 1–8.

[57] E. C. Larson and D. M. Chandler. Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of Electronic Imaging* 19, (2010) 011,006–011,006.

[58] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti. Tid2008-a database for evaluation of full-reference visual quality assessment metrics. *Advances of Modern Radioelectronics* 10, (2009) 30–45.

[59] J. A. Tropp and A. C. Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *Information Theory, IEEE Transactions on* 53, (2007) 4655–4666.

[60] K. Sharifi and A. Leon-Garcia. Estimation of shape parameter for generalized Gaussian distributions in subband decompositions of video. *Circuits and Systems for Video Technology, IEEE Transactions on* 5, (1995) 52–56.

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

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

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

[64] J. Sang, Z. Lei, and S. Z. Li. Face image quality evaluation for ISO/IEC standards 19794-5 and 29794-5. In Advances in Biometrics, 229–238. Springer, 2009.

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

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

[67] J. Chen, Y. Deng, G. Bai, and G. Su. Face Image Quality Assessment Based on Learning to Rank .

[68] P. Viola and M. J. Jones. Robust real-time face detection. *International journal of computer vision* 57, (2004) 137–154.

[69] K. Manasa Priya and S. S. Channappayya. A Novel Sparsity-inspired Blind Image Quality Assessment Algorithm. In accepted to GlobalSIP, 2014 IEEE International Conference on. IEEE, 2014 .

[70] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss. The FERET evaluation methodology for face-recognition algorithms. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 22, (2000) 1090–1104.

[71] M. Aharon, M. Elad, and A. Bruckstein. k-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *Signal Processing, IEEE Transactions on* 54, (2006) 4311–4322.

[72] T. Ahonen, A. Hadid, and M. Pietikäinen. Face recognition with local binary patterns. In Computer vision-eccv 2004, 469–481. Springer, 2004.

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