

Sparsity Based Spatio-Temporal Video Quality Assessment

Pochimireddy CharanTej Reddy

A Thesis Submitted to
Indian Institute of Technology Hyderabad
In Partial Fulfillment of the Requirements for
The Degree of Master of Technology



Department of Electrical Engineering

June 2016

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.

P. Charantej Reddy .

(Signature)

POCHIMIREDDY CHARANTEJ REDDY

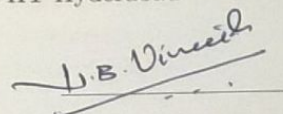
(Pochimireddy CharanTej Reddy)

EE14MTECH11008

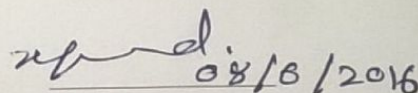
(Roll No.)

Approval Sheet

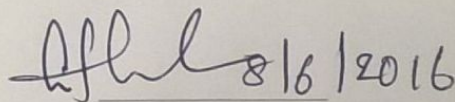
This Thesis entitled Sparsity Based Spatio-Temporal Video Quality Assessment by Pochimireddy Charantej Reddy is approved for the degree of Master of Technology from IIT Hyderabad



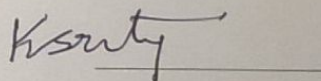
(Dr. Vineeth N Balasubramanian) Examiner
Dept. of Computer Science Eng
IITH


08/08/2016

(Dr. Lakshmi Natarajan) Examiner
Dept. of Electrical Eng
IITH


8/6/2016

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



(Dr. K Sri Rama Murty) Chairman
Dept. of Electrical Eng
IITH

Acknowledgements

I sincerely thank my advisor Dr. Sumohana Channappayya. I benefited a lot from discussions with him. He had always believed in me and supported my ideas. I also thank all my professors from the department and other disciplines who inspired me through their presence.

Dedication

I dedicate this thesis to my parents and to all my friends.

Abstract

In this thesis, we present an abstract view of Image and Video quality assessment algorithms. Most of the research in the area of quality assessment is focused on the scenario where the end-user is a human observer and therefore commonly known as perceptual quality assessment. In this thesis, we discuss Full Reference Video Quality Assessment and No Reference image quality assessment.

With the massive increase in video content being generated and viewed, it is very important to be able to objectively measure the quality of the content. This would in turn allow for better content management and for the provision of quality aware services. We are addressing the problem of objective video quality assessment in full reference setting by quantifying the change in sparsity of natural video sequences. This algorithm is inspired by sparse representation of videos in human visual system (HVS). It is well known that primary visual cortex adopts sparse coding strategy to visual stimulus. By using the over complete dictionaries we can represent the natural scenes sparsely. Primary visual cortex can be modelled well using such over complete dictionaries. Our hypothesis is that the sparse representation of natural videos are altered in the presence of distortion. We plan to measure this deviation in Sparsity of distorted video with respect to undistorted video in order to quantify perceptual quality.

In the case of no reference image quality assessment, we tried to improve the performance of Sparstiy-based Blind Image Quality Evaluation (SBIQE) using salience features.

Contents

Declaration	ii
Approval Sheet	iii
Acknowledgements	iv
Abstract	vi
Nomenclature	viii
1 Introduction To Quality Assessment	1
1.1 Subjective Quality Assessment	1
1.2 Objective Quality Assessment	2
1.2.1 Full Reference Quality Assessment	2
1.2.2 Reduced Reference Quality Assessment	3
1.2.3 No Reference Quality Assessment	3
2 Backgournd Theory	5
2.1 Overview	5
2.2 Sparse Representation of Signals	5
2.3 OMP Algorithm for Approximate Sparse Solution	6
2.4 K-SVD Algorithm For Dictionary Learning	6
2.5 Salience Based Visual Attention	7
2.6 Gaussian Mixture Models	8
3 Sparsity Based No Reference Image Quality Assessment	11
3.1 Introduction	11
3.2 Sparse Representation of Natural Images	12
3.3 SBIQE-1	12
3.3.1 Dictionary Construction	12
3.3.2 “Reference” Feature Extraction	13
3.3.3 Image Quality Measurement	14
3.4 SBIQE-2	15
3.4.1 A Model of Saliency-Based Visual Attention For Rapid Scene Analysis	15
3.4.2 Saliency based SBIQE	19
3.5 Results and Discussion	20
3.6 Conclusions and Future Work	20

4	Sparsity Based Video Quality Assessment	22
4.1	Introduction	22
4.2	Sparse Representation of videos	24
4.3	Metric	25
4.3.1	Dictionary Construction	25
4.3.2	Sparse Decomposition	26
4.3.3	approach-1	27
4.3.4	approach-2	27
4.4	Results and Discussion	27
4.5	Conclusions and Future Work	29
	References	30

Chapter 1

Introduction To Quality Assessment

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.

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 image (no reference) and video (full reference) quality assessment and face quality assessment. Quality assessment can be done in two ways. they are Subjective quality assessment and Objective quality assessment. The approaches are briefly discussed in the following sections.

1.1 Subjective 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.

1.2 Objective Quality Assessment

In objective quality assessment we write algorithms to evaluate the quality of image/video. The goal of objective image and video quality assessment research is to design the quality metrics that can predict perceived image and video quality automatically. Generally speaking, an objective image and video quality metric can be employed in three ways [1]:

1. It can be used to monitor image quality for quality control systems. For example, an image and video acquisition system can use the quality metric to monitor and automatically adjust itself to obtain the best quality image and video data. A network video server can examine the quality of the digital video transmitted on the network and control video streaming.
2. It can be employed to benchmark image and video processing systems and algorithms. If multiple video processing systems are available for a specific task, then a quality metric can help in determining which one of them provides the best quality results.
3. It can be embedded into an image and video processing system to optimize the algorithms and the parameter settings. For instance, in a visual communication system, a quality metric can help optimal design of the pre filtering and bit assignment algorithms at the encoder and the optimal reconstruction, error concealment and post filtering algorithms at the decoder.

Based on the availability of reference signal (image/video) which is considered to be pristine and distortion-free, objective quality assessment 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.

1.2.1 Full Reference Quality Assessment

In Full Reference Quality Assessment both test and reference signal are available to evaluate the perceived quality. Quality of test signal is evaluated with respect to the reference signal. The most widely used FR objective image and video distortion/quality metrics are mean squared error (MSE) and peak signal-to-noise ratio (PSNR), which are defined as:

$$MSE = \frac{1}{N} \sum_{i=1}^N (x_i - y_i)^2 \quad (1.1)$$

$$PSNR = 10 \log_{10} \frac{L^2}{MSE} \quad (1.2)$$

where N is the number of pixels in the image or video signal, and x_i and y_i are the i -th pixels in the original and the distorted signals, respectively. L is the dynamic range of the pixel values. For an 8bits/pixel monotonic signal, L is equal to 255. MSE and PSNR are widely used because they are simple to calculate, have clear physical meanings, and are mathematically easy to deal with for optimization purposes (MSE is differentiable, for example). However, they have been widely criticized as well for not correlating well with perceived quality measurement [2]. There are a number of reasons why MSE may not correlate well with the human perception of quality [1]:

1. Digital pixel values on which the MSE is typically computed, may not exactly represent the light stimulus entering the eye.

2. The sensitivity of the HVS to the errors may be different for different types of errors, and may also vary with visual context. This difference may not be captured adequately by the MSE.

3. Two distorted image signals with the same amount of error energy may have very different types of errors.

4. Simple error summation, like the one implemented in the MSE formulation, may be markedly different from the way the HVS and the brain arrives at an assessment of the perceived distortion.

In the last three to four decades, a great deal of effort has been made to develop objective image and video quality assessment methods (mostly for FR quality assessment), which incorporate perceptual quality measures by considering human visual system (HVS) characteristics.

VQA Algorithm

Initial VQA methods attempted to extend image quality metrics (like SSIM, VIF) to videos, where videos are divided into frames and then image metrics are applied on frames and frame wise scores are used to generate the final quality score for video. More recent approaches (like MOVIE [3], FLOSIM [4]) use temporal information (motion information) along with the spatial information. In these approaches spatial and temporal scores are calculated separately and combined generate the final quality score.

In our approach we are trying calculate the perceived quality score by quantifying the change in sparsity of natural video sequences. This algorithm is inspired by sparse representation of videos in human visual system (HVS). It is well known that primary visual cortex adopts sparse coding strategy to visual stimulus. By using the over complete dictionaries we can represent the natural scenes sparsely. Primary visual cortex can be modeled well using such over complete dictionaries. our hypothesis is that the sparse representation of natural videos are altered in the presence of distortion. We plan to measure this deviation in Sparsity of distorted video with respect to undistorted video in order to quantify perceptual quality. In our approach final spatio-temporal score is calculated directly instead of calculating spatial scores and temporal scores and pooling them to find a final score. The proposed FR-VQA algorithm will be discussed in detail in subsequent chapters.

1.2.2 Reduced Reference Quality Assessment

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

1.2.3 No Reference Quality Assessment

In most of the practical settings, the reference signal is not available for quality assessment. Hence, there is a need to evaluate the quality solely based on the test signal.

IQA Algorithm

These algorithms generally follow one or a combination of three approaches:

1. Distortion-specific approaches: In this approach, algorithms quantify the distortions such as blur [5], ringing effect [6], or blockiness [7] and evaluate the image accordingly.

2. Training-based approaches: In this approach, the algorithm predicts the quality of an image by training a model from the features extracted [[8], [9]].

3. 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 [10].

In SBIQE a combination of the second and third approaches used to predict the quality of images. we tried to improve the performance of SBIQE using salience features. The proposed NR-IQA algorithm 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. Then, we define the concepts Saliency Based Visual Attention and Gaussian Mixture Models and are used for perceptual quality assessment.

2.2 Sparse Representation of Signals

The objective of sparse representation of signals is to represent the 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 [11]. 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 ∞ [12]. But in most of the cases p is preferably considered as 2.

As D is overcomplete matrix, there is possibility of 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

$$\min_x \|x\|_0 \quad \text{subject to} \quad y = Dx \quad (2.1)$$

$$\min_x \|x\|_0 \quad \text{subject to} \quad \|y - Dx\|_2 \leq \epsilon \quad (2.2)$$

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

$$\min_x \|y - Dx\|_2 \quad \text{subject to} \quad \|x\|_0 \leq L \quad (2.3)$$

$$\min_x \frac{1}{2} \|y - Dx\|_2 + \lambda \|x\|_0 \quad (2.4)$$

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

In the subsequent subsections, we first review the Orthogonal Matching Pursuit(OMP) algorithm [13] to solve for approximate sparse solution and then K-SVD algorithm [14] for dictionary learning.

2.3 OMP Algorithm for Approximate Sparse Solution

Solving the equations 2.1 and 2.2 is NP hard and computationally expensive. There are many algorithms tried to approximate the objective function with alternatives and one such algorithm is OMP algorithm, a simplest and effective approximation method among greedy pursuit methods. It finds the locally optimum solution at each iteration by searching the basis which 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 $S^0 = \text{Support}\{x^0\} = \emptyset$

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 $\text{Support}\{x\} = S^k$.
- **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 [15]. Examples of the analytic formulation include Fast Fourier Transform (FFT) [16], Discrete Cosine Transform (DCT) [17] and Gabor transform [18]. 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 [15]. In the past decade, a large number of papers [14] have 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 [15]. Among the dictionary training methods, the K-SVD algorithm [14], 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

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

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

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

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 Saliency Based Visual Attention

It is intuitively obvious that each region in an image may not bear the same importance as others. Visual importance has been explored in the context of visual saliency, fixation calculation, and foveated image and video compression. However, region-of-interest based image quality assessment remains relatively unexplored. It is the furtherance of this area of quality assessment that motivates this work. Under the hypothesis that certain regions in an image may be visually more important than others, methods used to spatially pool the quality scores from the SSIM maps are an appealing possibility for improving SSIM scores. The effect of using different pooling strategies was evaluated, including local quality-based pooling. It was concluded that the best possible gains could be achieved by using an information-theoretic approach deploying information ‘‘content-weighted pooling’’. In this paper, we further investigate quality based pooling and also consider pooling based on predicted human gaze behavior.

There are two hypotheses which may influence human perception of image quality. The first is visual attention and gaze direction“where” a human looks. The second hypothesis is that humans tend to perceive “poor” regions in an image with more severity than the “good” ones and hence penalize images with even a small number of “poor” regions more heavily. Existing IQA algorithms, on the contrary, do not attempt to compensate for this prejudice. By weighting more heavily quality scores from lower scoring regions, such a compensation can be achieved. This idea of heavily weighting the lower scoring regions is a form of visual importance. However, their approach, which involved weighting quality scores as a monotonic function of quality, led to the conclusion that quality-weighted pooling yields incremental improvement, at best. By contrast, in our approach, we show that quite significant gains in performance can be obtained using the right strategy.

Research into the general area of how humans deploy eye movements in visual tasks has received significant attention for many decades. Competing theories for gaze selection can be broadly classified into two general categories: top-down (cognitive/high-level) and bottom-up (precognitive/low-level). Top-down approaches for gaze prediction emphasize a high-level understanding of the scene and has been popular in task-specific experiments. Yarbus, in his pioneering work on eye movements, demonstrated that human eye movements are strongly influenced by high-level mechanisms such as the specific visual task given to the observer. Top-down implementations of gaze-selection have incorporated spatial relationships of object and scene schema representations and shown significant improvements in search times in visual search tasks. While such top-down implementations provide possible directions of exploration in gaze selection, cognitive interpretation of scenes is far from being sufficiently mature to generalize to natural viewing tasks. Given the rapidity and sheer volume of saccades during search tasks (over 15,000 each hour), it is reasonable to suppose that there is a significant bottom-up, computationally inexpensive component to selecting fixation locations. The goal of this paper is to investigate bottom-up, image-based mechanisms that guide eye fixations. Moreover, we believe that the development of future high-level visual search systems may benefit from the insights gained from successful low-level search strategies.

2.6 Gaussian Mixture Models

One of the most important task when working with mixture model-based clustering is precisely, selecting the type of function which offers a better adjust to the data field and the type of task we face up to. Between the different types of mixture model-based clustering, one of the most commonly used is clustering based in Gaussian Mixture Models (GMMs) [17]

Gaussian mixture model as a simple linear superposition of Gaussian components, aimed at providing a richer class of density models than the single Gaussian. Techniques based on GMM are applied to many different tasks. Some of the most common applications are speaker identification, speech recognition, image segmentation, biometric verification or detection of image color and texture.

Gaussian Mixture Model (GMM) is a parametric probability density function represented as a weighted sum of Gaussian component densities. GMMs are commonly used as a parametric model of the probability distribution of continuous measurements or features in a biometric system, such as vocal-tract related spectral features in a speaker recognition system. GMM parameters are estimated from training data using the iterative Expectation-Maximization (EM) algorithm or Maximum A Posteriori (MAP) estimation from a

well-trained prior model.

Fig. 2.1 illustrates the joint density capturing capabilities of GMM, using 2-dimensional data uniformly disturbed along a circular ring. The red ellipses, superimposed on the data (blue) points, correspond to the locations and shapes of the estimated Gaussian mixtures.

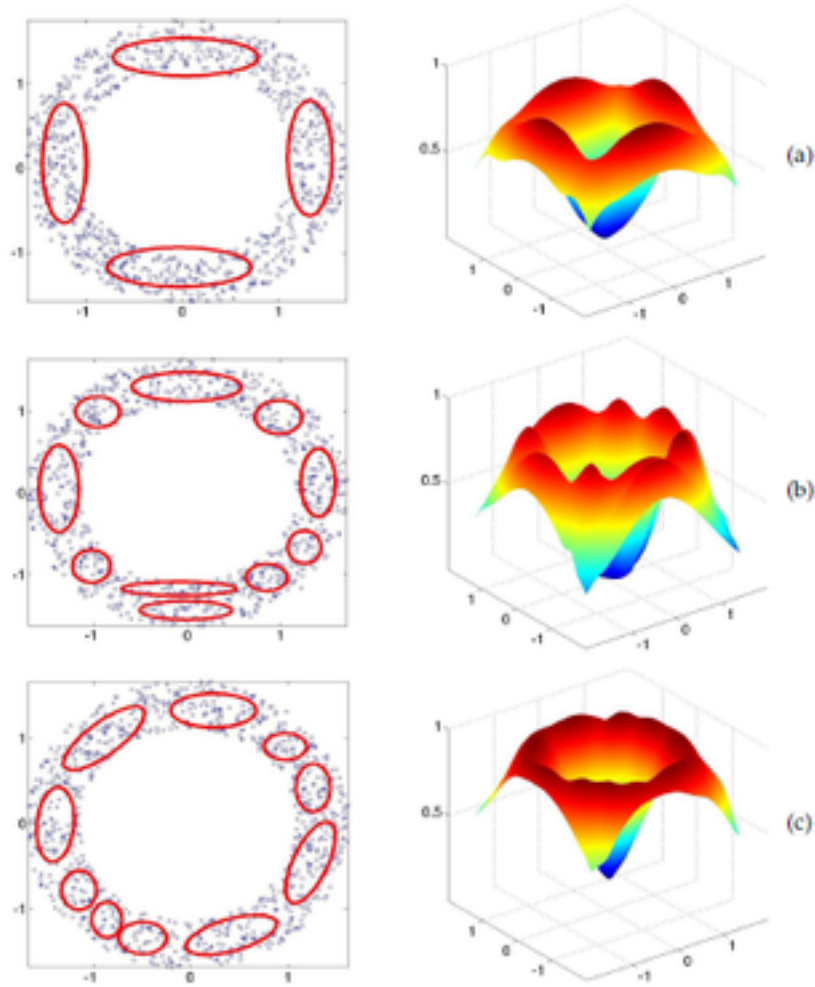


Figure 2.1: Illustration of distribution capturing capability of GMM. GMM trained with diagonal covariance matrices (a) 4-mixtures (b) 10-mixtures and (c) 10-mixture GMM trained with full covariance matrices

A Gaussian mixture model is a weighted sum of M component Gaussian densities as given by the equation,

$$p(x|\lambda) = \sum_{i=1}^M w_i g(x|\mu_i, \Sigma_i) \quad (2.5)$$

where x is a D -dimensional continuous-valued data vector (i.e. measurement or features), w_i , $i = 1, \dots, M$, are the mixture weights, and $g(x|\mu_i, \Sigma_i)$, $i = 1, \dots, M$, are the component Gaussian densities. Each component

density is a D-variate Gaussian function of the form,

$$g(x|\mu_i, \Sigma_i) = \frac{1}{(2\pi)^{D/2} |\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(x - \mu_i)' \Sigma_i^{-1} (x - \mu_i)\right\} \quad (2.6)$$

with mean vector μ_i and covariance matrix Σ_i . The mixture weights satisfy the constraint that $\sum_{i=1}^M w_i = 1$.

The complete Gaussian mixture model is parameterized by the mean vectors, covariance matrices and mixture weights from all component densities. These parameters are collectively represented by the notation,

$$\lambda = \{w_i, \mu_i, \Sigma_i\}, \quad i = 1, 2, \dots, M \quad (2.7)$$

EM Algorithm for Gaussian Mixture Models:

Given a Gaussian mixture model, the goal is to maximize the likelihood function with respect to the parameters (comprising the means and covariances of the components and the mixing coefficients).

1. Initialize the means μ_i , covariances Σ_i and mixing coefficients w_i , and evaluate the initial value of the log likelihood.
2. **E-Step:** Evaluate the responsibilities using the current parameter values

$$\gamma(z_{ni}) = \frac{w_i g(x_n|\mu_i, \Sigma_i)}{\sum_{j=1}^M w_j g(x_n|\mu_j, \Sigma_j)} \quad (2.8)$$

3. **M-Step:** Re-estimate the parameters using the current responsibilities

$$\mu_i^{new} = \frac{1}{N_k} \sum_{n=1}^N \gamma(z_{nk}) x_n \quad (2.9)$$

$$\Sigma_i^{new} = \frac{1}{N_k} \sum_{n=1}^N \gamma(z_{nk}) (x_n - \mu_i^{new})(x_n - \mu_i^{new})^T \quad (2.10)$$

$$w_i^{new} = \frac{N_k}{N} \quad (2.11)$$

where,

$$N_k = \sum_{n=1}^N \gamma(z_{nk}) \quad (2.12)$$

4. Evaluate the log likelihood

$$\ln p(X|\mu, \Sigma, w) = \sum_{n=1}^N \ln \left\{ \sum_{i=1}^M w_i g(x|\mu_i, \Sigma_i) \right\} \quad (2.13)$$

and check for convergence of either the parameters or the log likelihood. If the convergence criterion is not satisfied return to step 2.

Chapter 3

Sparsity Based No Reference Image Quality Assessment

3.1 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) [19], Sparse representation for blind image quality assessment (SRNSS) [20], Natural Image Quality Evaluator (NIQE) [21], Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) [22], probabilistic latent semantic analysis (pLSA) [23], BLind Image Integrity Notator using DCT Statistics-II (BLIINDS-II) , [24], and Distortion Identification-based Image Verity and INtegrity Evaluator (DIIVINE) [25]. 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 [21] (distortion agnostic), QAC [19], BRISQUE [22], pSLA [23], DIIVINE [25], BLIINDS [26] (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 [23], DIIVINE [25], BLIINDS [26], SRNSS [20]. 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 [19] is an opinion-unaware and distortion-aware BIQA algorithm that achieves its opinion-

unawareness by replacing subjective scores with FSIM [27], 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 [21] 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.

3.2 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 [28] (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, Olshausen and Field [29] 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) [30, 31] and BIQA – for e.g SRNSS [20] algorithms. While SRNSS is an opinion aware method, we focus on an opinion unaware algorithm.

3.3 SBIQE-1

3.3.1 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 over complete dictionary to sparsely represent natural scenes. While the over complete dictionary construction technique in [29] gives good results, we chose to work with the more recent K-SVD [14] algorithm for dictionary construction. The K-SVD algorithm has been shown to outperform other over complete 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,

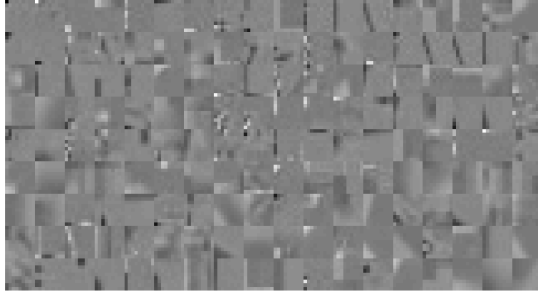


Figure 3.1: The dictionary of atoms constructed from pristine images. Each atom is of size 9×9 , with 18 atoms per row and 9 rows in all.

super-resolution, to name a few.

We construct an over complete 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 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 that has no overlap with the popular data set used for testing (LIVE). The dictionary construction happens once and can be performed offline. The dictionary with all atoms concatenated is shown in Fig. 3.1. We clearly see the oriented nature of a majority of the atoms along with a few atoms containing lower spatial frequencies.

3.3.2 “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) [13] 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 number of patches used in this training stage. The feature vector is normalized to get

$$\mathbf{n}_r = \frac{\mathbf{f}_r - \mu_r}{\sigma_r}, \quad (3.1)$$

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. 3.2. The figure shows the magnitude

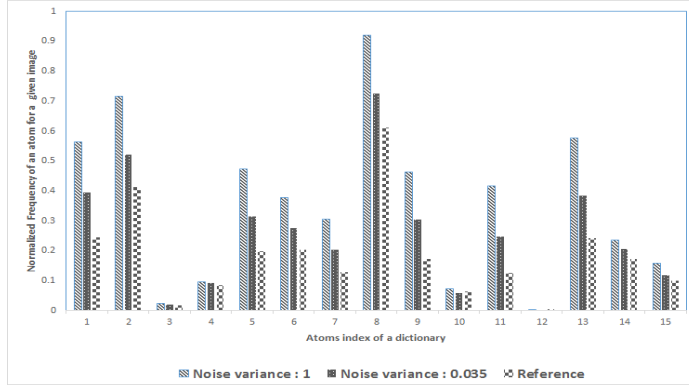


Figure 3.2: Motivation behind SBIQE. Feature vectors of pristine images and images distorted with AWGN. The y-axis represents the magnitude of the feature point while the x-axis represents feature point index. Three noise levels ($\sigma = 1, \sigma = 0.035, \sigma = 0$) are shown to illustrate the discriminability of the feature choice.

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.

3.3.3 Image Quality Measurement

Given a test image, it is divided into overlapping blocks (as in Section 3.3.2) and each block is represented using the dictionary constructed in Section 3.3.1. 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}_r is normalized to get

$$\mathbf{n}_t = \frac{\mathbf{f}_t - \mu_t}{\sigma_t}, \quad (3.2)$$

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

$$Q_t = 1 - \frac{\|\mathbf{n}_t - \mathbf{n}_r\|_2}{\|\mathbf{n}_t\|_2 + \|\mathbf{n}_r\|_2}. \quad (3.3)$$

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

$$0 \leq Q_t \leq 1. \quad (3.4)$$

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.

3.4 SBIQE-2

3.4.1 A Model of Saliency-Based Visual Attention For Rapid Scene Analysis

Primates have a remarkable ability to interpret complex scenes in real time, despite the limited speed of the neuronal hardware available for such tasks. Intermediate and higher visual processes appear to select a subset of the available sensory information before further processing, most likely to reduce the complexity of scene analysis. This selection appears to be implemented in the form of a spatially circumscribed region of the visual field, the so-called “focus of attention,” which scans the scene both in a rapid, bottom-up, saliency-driven, and task-independent manner as well as in a slower, top-down, volition-controlled, and task-dependent manner. Models of attention include “dynamic routing” models, in which information from only a small region of the visual field can progress through the cortical visual hierarchy. The attended region is selected through dynamic modifications of cortical connectivity or through the establishment of specific temporal patterns of activity, under both top-down (taskdependent) and bottom-up (scene-dependent) control. It is related to the so-called “feature integration theory”, explaining human visual search strategies. Visual input is first decomposed into a set of topographic feature maps. Different spatial locations then compete for saliency within each map, such that only locations which locally stand out from their surround can persist. All feature maps feed, in a purely bottom-up manner, into a master “saliency map”, which topographically codes for local conspicuity over the entire visual scene. In primates, such a map is believed to be located in the posterior parietal cortex as well as in the various visual maps in the pulvinar nuclei of the thalamus. The model’s saliency map is endowed with internal dynamics which generate attentional shifts. This model consequently represents a complete account of bottom-up saliency and does not require any top-down guidance to shift attention. This framework provides a massively parallel method for the fast selection of a small number of interesting image locations to be analyzed by more complex and time consuming objectrecognition processes. Extending this approach in “guided-search”, feedback from higher cortical areas (e.g., knowledge about targets to be found) was used to weight the importance of different features, such that only those with high weights could reach higher processing levels.

Model

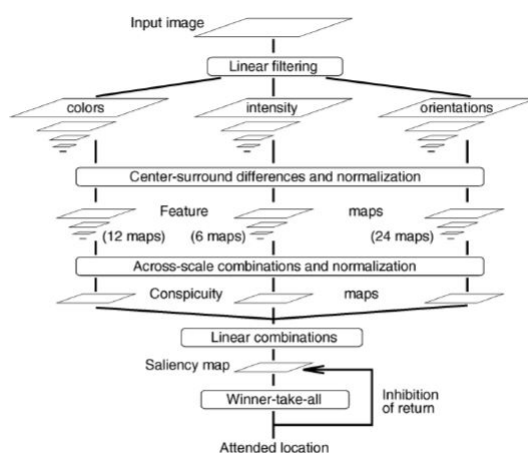


Figure 3.3: General architecture of the model

Input is provided in the form of static color images, usually digitized at 640X480 resolution. Nine spatial scales are created using dyadic Gaussian pyramids, which progressively low-pass filter and subsample the input image, yielding horizontal and vertical image-reduction factors ranging from 1 : 1 (scale zero) to 1 : 256 (scale eight) in eight octaves. Each feature is computed by a set of linear “centersurround” operations akin to visual receptive fields. Typical visual neurons are most sensitive in a small region of the visual space (the center), while stimuli presented in a broader, weaker antagonistic region concentric with the center (the surround) inhibit the neuronal response. Such an architecture, sensitive to local spatial discontinuities, is particularly well-suited to detecting locations which stand out from their surround and is a general computational principle in the retina, lateral geniculate nucleus, and primary visual cortex. Center-surround is implemented in the model as the difference between fine and coarse scales: The center is a pixel at scale $c \in \{2, 3, 4\}$ and the surround is the corresponding pixel at scale

$$s = c + \delta, \text{ with } \delta \in \{3, 4\} \quad (3.5)$$

The across-scale difference between two maps, denoted “*” below, is obtained by interpolation to the finer scale and point-by-point subtraction. Using several scales not only for c but also for $d = s - c$ yields truly multi-scale feature extraction, by including different size ratios between the center and surround regions (contrary to previously used fixed ratios).

Extraction of Early Visual Features:

With \mathbf{r} , \mathbf{g} , and \mathbf{b} being the red, green, and blue channels of the input image, an intensity image \mathbf{I} is obtained as

$$I = (r + g + b)/3$$

\mathbf{I} is used to create a Gaussian pyramid $I(\sigma)$, where $\sigma \in [0.8]$ is the scale. The \mathbf{r} , \mathbf{g} , and \mathbf{b} channels are normalized by \mathbf{I} in order to decouple hue from intensity. However, because hue variations are not perceivable at very low luminance (and hence are not salient), normalization is only applied at the locations where \mathbf{I} is larger than 1/10 of its maximum over the entire image (other locations yield zero \mathbf{r} , \mathbf{g} , and \mathbf{b}). Four broadly-tuned color channels are created:

$$R = r - (g + b)/2 \text{ for red}$$

$$G = g - (r + b)/2 \text{ for green}$$

$$B = b - (r + g)/2 \text{ for blue}$$

$$Y = (r + g)/2 - |r - g|/2 - b \text{ for yellow (negative values are set to zero)}$$

Four Gaussian pyramids $R(s)$, $G(s)$, $B(s)$, and $Y(s)$ are created from these color channels. Centersurround differences (“*” defined previously) between a “center” fine scale c and a “surround” coarser scale s yield the feature maps. The first set of feature maps is concerned with intensity contrast, which, in mammals, is detected by neurons sensitive either to dark centers on bright surrounds or to bright centers on dark surrounds. Here, both types of sensitivities are simultaneously computed (using a rectification) in a set of six maps, (c, s) , with $c \in \{2, 3, 4\}$ and $s = c + \delta$, with $\delta \in \{3, 4\}$:

$$I(c, s) = |I(c) * I(s)| \quad (3.6)$$

A second set of maps is similarly constructed for the color channels, which, in cortex, are represented using a so-called color double-opponent system: In the center of their receptive fields, neurons are excited by one color (e.g., red) and inhibited by another (e.g., green), while the converse is true in the surround. Such spatial and chromatic opponency exists for the red/green, green/red, blue/yellow, and yellow/blue color pairs in human primary visual cortex. Accordingly, maps $RG(c, s)$ are created in the model to simultaneously account for red/green and green/red double opponency and $BY(c, s)$ for blue/yellow and yellow/blue double opponency:

$$RG(c, s) = |(R(c) - G(c)) * (G(s) - R(s))| \quad (3.7)$$

$$BY(c, s) = |(B(c) - Y(c)) * (Y(s) - B(s))| \quad (3.8)$$

Local orientation information is obtained from I using oriented Gabor pyramids $O(\sigma, \theta)$, where $\theta \in [0, \pi]$ represents the scale and $\theta \in \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}$ is the preferred orientation. (Gabor filters, which are the product of a cosine grating and a 2D Gaussian envelope, approximate the receptive field sensitivity profile (impulse response) of orientation-selective neurons in primary visual cortex.) Orientation feature maps, $O(c, s, \theta)$, encode, as a group, local orientation contrast between the center and surround scales:

$$O(c, s, \theta) = |O(c, \theta) * O(s, \theta)| \quad (3.9)$$

In total, 42 feature maps are computed: six for intensity, 12 for color, and 24 for orientation.

saliency map

The purpose of the saliency map is to represent the conspicuity or ‘‘saliency’’ at every location in the visual field by a scalar quantity and to guide the selection of attended locations, based on the spatial distribution of saliency. A combination of the feature maps provides bottom-up input to the saliency map, modeled as a dynamical neural network. One difficulty in combining different feature maps is that they represent a priori not comparable modalities, with different dynamic ranges and extraction mechanisms. Also, because all 42 feature maps are combined, salient objects appearing strongly in only a few maps may be masked by noise or by less-salient objects resented in a larger number of maps. In the absence of top-down supervision, we propose a map normalization operator, $N(\cdot)$, which globally promotes maps in which a small number of strong peaks of activity (conspicuous locations) is present, while globally suppressing maps which contain numerous comparable peak responses $N(\cdot)$:

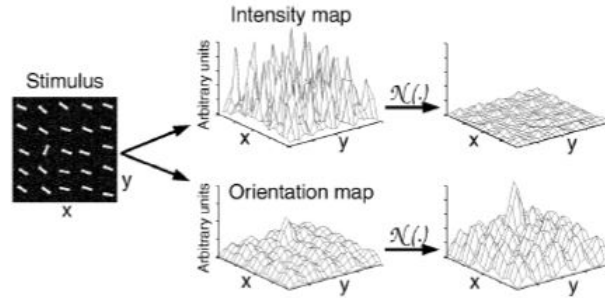


Figure 3.4: General architecture of the model

1. Normalizing the values in the map to a fixed range $[0, \dots M]$, in order to eliminate modality-dependent amplitude differences.
2. Finding the location of the map's global maximum M and computing the average m of all its other local maxima.
3. Globally multiplying the map by $(M - \bar{m})^2$.

Only local maxima of activity are considered, such that $N(\cdot)$ compares responses associated with meaningful "activation spots" in the map and ignores homogeneous areas. Comparing the maximum activity in the entire map to the average overall activation measures how different the most active location is from the average. When this difference is large, the most active location stands out, and the map is strongly promoted. When the difference is small, the map contains nothing unique and is suppressed. The biological motivation behind the design of $N(\cdot)$ is that it coarsely replicates cortical lateral inhibition mechanisms, in which neighboring similar features inhibit each other via specific, anatomically defined connections.

Feature maps are combined into three "conspicuity maps," \bar{I} , for intensity & \bar{C} for color (6), and \bar{O} for orientation, at the scale ($\sigma = 4$) of the saliency map. They are obtained through across-scale addition " \oplus " which consists of reduction of each map to scale four and point-by-point addition:

$$\bar{I} = \bigoplus_{c=1}^4 \bigoplus_{s=c+3}^{c+4} N(I(c, s)) \quad (3.10)$$

$$\bar{C} = \bigoplus_{c=2}^4 \bigoplus_{s=c+3}^{c+4} [N(RG(c, s)) + N(BY(c, s))] \quad (3.11)$$

For orientation, four intermediary maps are first created by combination of the six feature maps for a given q and are then combined into a single orientation conspicuity map

$$\bar{O} = \sum_{\theta \in \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}} N \left(\bigoplus_{c=2}^4 \bigoplus_{s=c+3}^{c+4} N(O(c, s, \theta)) \right) \quad (3.12)$$

The motivation for the creation of three separate channels, \bar{I} , \bar{C} and \bar{O} , and their individual normalization is the hypothesis that similar features compete strongly for saliency, while different modalities contribute independently to the saliency map. The three conspicuity maps are normalized and summed into the final input 6 to the saliency map:

$$S = \frac{1}{3} (N(\bar{I}) + N(\bar{C}) + N(\bar{O})) \quad (3.13)$$

At any given time, the maximum of the saliency map (SM) defines the most salient image location, to which the focus of attention (FOA) should be directed. We could now simply select the most active location as defining the point where the model should next attend. However, in a neuronally plausible implementation, we model the SM as a 2D layer of leaky integrate-and-fire neurons at scale four. These model neurons consist of a single capacitance which integrates the charge delivered by synaptic input, of a leakage conductance, and of a voltage threshold. When the threshold is reached, a prototypical spike is generated, and the capacitive charge is shunted to zero. The SM feeds into a biologically plausible 2D "winner-take-all" (WTA) neural network at scale $\sigma = 4$, in which synaptic interactions among units ensure that only the most active location remains, while all other locations are suppressed.

The neurons in the SM receive excitatory inputs from 6 and are all independent. The potential of SM neurons at more salient locations hence increases faster (these neurons are used as pure integrators and do not fire). Each SM neuron excites its corresponding WTA neuron. All WTA neurons also evolve independently of each other, until one (the “winner”) first reaches threshold and fires. This triggers three simultaneous mechanisms:

1. The FOA is shifted to the location of the winner neuron.
2. The global inhibition of the WTA is triggered and completely inhibits (resets) all WTA neurons
3. Local inhibition is transiently activated in the SM, in an area with the size and new location of the FOA; this not only yields dynamical shifts of the FOA, by allowing the next most salient location to subsequently become the winner, but it also prevents the FOA from immediately returning to a previously-attended location.

Such an “inhibition of return” has been demonstrated in human visual psychophysics. In order to slightly bias the model to subsequently jump to salient locations spatially close to the currently attended location, a small excitation is transiently activated in the SM, in a near surround of the FOA (“proximity preference” rule of Koch and Ullman).

Since we do not model any top-down attentional component, the FOA is a simple disk whose radius is fixed to one sixth of the smaller of the input image width or height. The time constants, conductances, and firing thresholds of the simulated neurons were chosen so that the FOA jumps from one salient location to the next in approximately 30 – 70ms (simulated time), and that an attended area is inhibited for approximately 500 – 900ms, as has been observed psychophysically. The difference in the relative magnitude of these delays proved sufficient to ensure thorough scanning of the image and prevented cycling through only a limited number of locations. All parameters are fixed in our implementation, and the system proved stable over time for all images studied.



Figure 3.5: Original image

3.4.2 Saliency based SBIQE

In SBIQE, we took all the image patches for quality evaluation. But in this method instead of taking all the patches we took only 3000, visually important patches and evaluated quality scores of images.



Figure 3.6: Saliency result

3.5 Results and Discussion

We present the results of our algorithms and compare it with state-of-the-art BIQA methods. BRISQUE [23], an opinion-aware distortion-aware method, QAC [19] an opinion-unaware distortion-aware method and NIQE [21] 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 dataset are presented in Table. 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
<i>SBIQE-1</i>	<i>0.96</i>	<i>0.89</i>	<i>0.73</i>	<i>0.83</i>	<i>0.76</i>
<i>SBIQE-2</i>	<i>0.9196</i>	<i>0.7830</i>	<i>0.7430</i>	<i>0.7998</i>	<i>0.7452</i>

Table 3.1: 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.

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.

3.6 Conclusions and Future Work

We tried to improve the results of SBIQE using salience features. 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 4

Sparsity Based Video Quality Assessment

4.1 Introduction

The objective assessment of the perceptual quality of natural videos is a challenging and open research problem. This statement is true for all the flavors of video quality assessment (VQA) - full reference (FR), reduced-reference (RR), and no-reference (NR). The evidence for this claim is the fact that the state-of-the-art methods for all three flavors have only recently been approaching a reasonable level of correlation (in the range (0.75, 0.9)) with subjective scores on a moderately complex video database (LIVE) [32] [33]. In the following, we non-exhaustively review recent and relevant FR-VQA algorithms in order to place our work in context.

FR-VQA is a challenging task due to several reasons of which we opine the primary ones are the highly non-stationary nature of video signals, and an incomplete understanding of the human visual system (HVS). FR-VQA is a well-studied problem with several approaches having been explored. One classification of FR-VQA algorithms could be based on their domain of operation as either compressed-domain methods or uncompressed-domain methods. We briefly review the compressed-domain FR-VQA literature. The compressed domain FR-VQA approaches operate with limited compressed bit stream information and are primarily geared toward addressing artifacts arising out of communication errors such as bit-errors and packet losses. The effect of these packet losses on the videos have been estimated using mean squared error estimation techniques that are based on bitstream parsing [34], a spatio-temporal technique by Yang et al. [35], a weighted combination of blockiness, blurriness and noise by Farias and Mitra [36], temporal quality estimation based on frame drop by Yang et al. [37], generalized linear models for packet loss visibility [38], [39], to name a few.

Uncompressed-domain FR-VQA approaches on the other hand utilize all the information available in the spatio-temporal domain. Both image and video quality measures based on human visual perception [40] [41] have been explored from the very beginning. The properties of the HVS that have been employed include modeling the visual sensitivity to predict the visibility of the error [40], modeling the human visual sensitivity to spatial and chromatic signals [42], incorporating aspects of early visual processing, such as light adaptation, luminance and chromatic channels, spatial and temporal filtering, spatial frequency channels, contrast masking, and probability summation [43], measurement of sensory scales based on Bayesian estimation [41]. A

significant amount of I/VQA approaches in the past were based on the error sensitivity philosophy [44] [45] motivated from psychological vision science research where the distorted signal is the sum of a reference signal and an error signal, the amount of perception of the error signal by HVS is determined. The error sensitivity based quality measurement is as follows: the original and test signals are subject to preprocessing procedures, such as alignment, luminance transformation, and color transformation. A channel decomposition method (wavelet transforms, discrete cosine transform (DCT), and Gabor decompositions) is then applied to these two preprocessed signals.

The errors between the two signals in each channel are calculated and weighted, usually by a Contrast Sensitivity Function (CSF). The weighted error signals are adjusted by a visual masking effect model, which reflects the reduced visibility of errors presented on a background reference signal. The Minkowski error pooling of the weighted and masked error signal is then employed to obtain a single quality score.

Wang et al. [46] proposed a structural distortion estimation approach to FR-VQA which hypothesized that amount of distortion in the structure is a measure of perceived distortion. It is an extension of the Structural SIMilarity (SSIM) index [47] with two adjustments including local spatial and temporal weighting based on the luminance and global motion respectively. Another popular approach to the quality assessment problem is to cast it in an information communication framework [48], [49], [50], where the HVS is modeled as an error-prone communication channel. Yet another popular approach is based on the observation that the HVS does not perceive all the distortions equally [3] [32], [46], [51], [52].

Wang and Li [48] incorporated the model proposed by Stocker and Simoncelli [53] for human visual speed perception. Spatio-temporal weighting is done based on the motion information content and the perceptual uncertainty.

Seshadrinathan and Bovik employed an approach to tune the orientation of a set of 3D Gabor filters according to local motion based on optical flow [54], [55]. The adapted Gabor filter responses are then incorporated into the SSIM and the visual information fidelity (VIF) [49] measures for the purpose of VQA.

Ninassi et al. [56] hypothesize that the temporal distortion is the temporal evolution of the spatial distortion and is closely linked to the visual attention mechanisms. Hence they resort to short-term temporal pooling and long-term temporal pooling. In the short-term evaluation of the temporal distortions, the spatio-temporal perceptual distortion maps are computed from the spatial distortion maps (wavelet based quality assessment metric (WQA) [57]) and the motion information. In the long term evaluation, the quality score for the whole video sequence is computed based on the concepts of perceptual saturation and asymmetric behavior of the humans.

The Motion-tuned Video Integrity Evaluator (MOVIE) index [3] is a HVS-inspired algorithm where the response of the visual system to video stimulus is modeled as a function of linear spatio-temporal bandpass filter outputs. The central idea is that distortions cause the optical flow plane of the distorted video to move away from the reference videos optical flow plane. The filters close to the reference videos optical flow plane are given excitatory weights and those away from it are given inhibitory weights.

A video quality assessment model based on Most Apparent Distortion (MAD) [58] called Spatio-temporal MAD (ST-MAD) [59], is designed on the assumption that motion artifacts will manifest as spatial artifacts and the appearance-based model of MAD can measure these changes to agree with human perception. MAD has two stages; a detection-based stage, which computes the perceived degradation due to visual detection of distortions and an appearance based stage, which computes the perceived degradation due to visual appearance changes.

Park et al. [32] hypothesize that non-uniform local distortion is perceptually more annoying than distortion that is more or less uniform both spatially and temporally. This hypothesis is demonstrated with a strategy for pooling local quality scores into a global score. Local quality scores are sorted in ascending order and higher weights are assigned to patches that occupy the steepest ascent region in the sorted quality curve.

Wolf and Pinson [60] present a VQA algorithm called VQM_VFD to quantify the effects of temporal distortion due to frame delay on perceptual video quality. This is an extension to their previous work called video quality metric (VQM) [51]. In VQM_VFD, the authors extract hand crafted spatio-temporal features primarily using edge detection filters. A neural network is trained using these features extracted from a training set composed of a varied collection of video data. The VQM_VFD was evaluated on the LIVE Mobile database in [61] and shown to have the best performance across all VQA algorithms.

Manasa and Sumohana presented a simple yet effective FR-VQA algorithm called FLOSIM [4] based on local optical flow statistics and a robust FR-IQA algorithm, and also proposed a perceptually inspired pooling strategy and demonstrate the efficacy of our approach on popular SD and HD video databases.

In this work we present simple FR-VQA algorithm based on the change in sparse represent of the videos in the presence of distortion. By quantifying this change we are measuring the quality of the video.

4.2 Sparse Representation of videos

A small image patch $I(x, y)$ is modeled as a linear superposition of basis functions, $\phi_i(x, y)$, multiplied by coefficients, a_i :

$$I(x, y) = \sum_i a_i \phi_i(x, y) \quad (4.1)$$

When a set of basis functions is sought such that the coefficients are as sparse and statistically independent as possible, averaged over many natural images, the basis functions that emerge are localized, oriented, and bandpass (selective to structure at different spatial scales). These properties are similar to the receptive fields of neurons in mammalian primary visual cortex (area V I), thus suggesting that the cortex has evolved according to a similar coding principle.

van Hateren and Ruder [62] extended this idea to the time domain and showed that the basis functions that emerge have similar spatial properties and translate as a function of time, similar to the non-separable (direction selective) receptive fields of cortical simple cells. However, their image model relies upon blocking the image stream into a small number of frames and treating time simply as another dimension:

$$I(x, y, t) = \sum_i a_i \phi_i(x, y, t) \quad (4.2)$$

An image sequence is then represented by simply computing inner products between a set of bi-orthogonal functions and a block of image frames.

Olshausen [63] modeled time-varying images without blocking by assuming time-in variance in the basis functions, so that each function can be applied at all points in time. Importantly, the basis set is over complete, so that there are multiple ways to describe a given image sequence. When a sparse representation is selected via matching pursuit, then one obtains a recording of the image in terms of sparse, punctate events in time, similar to neural spike trains. The suggestion is that the spike trains of V1 neurons themselves serve as a sparse code in time, and that V1 receptive fields have been adapted to represent images in this way.

4.3 Metric

Generally in these sparsity based quality evaluation approaches three steps are there. They are

1. Dictionary construction
2. Sparse decomposition
3. Quality evaluation

4.3.1 Dictionary Construction

Similar to the NR-IQA work here also we used K-SVD for dictionary construction. We divided video into spatio temporal patches and made them into vectors. These vectors are used for dictionary generation. Our hypothesis is indexing in vectors is helping to capture both spial and temporal variations in video. We constructed dictionaries at different scales from reference videos patches using K-SVD algorithm. Along spatial dimensions we considered patch sizes of $3 * 3$, $5 * 5$, $7 * 7$, $9 * 9$, $16 * 16$. Along the temporal direction we considered groups of 3, 5, 7, 9, 11, 16, 25, 31 we get spatio temporal patches with different sizes like $3*3*3$, $3*3*5$, ...

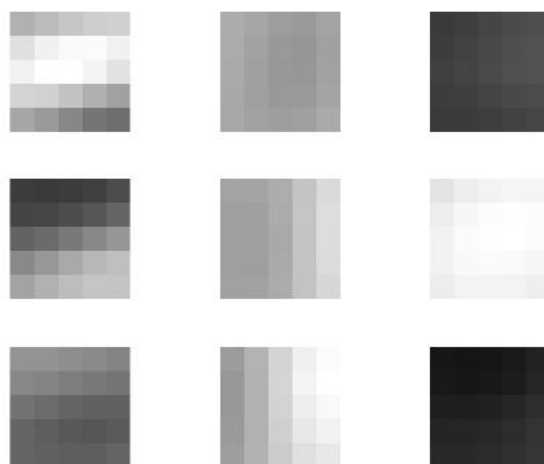


Figure 4.1: $5*5*3$ Dictionary atoms (column wise). Left: Atom 1, Middle: Atom 75, Right: 150.

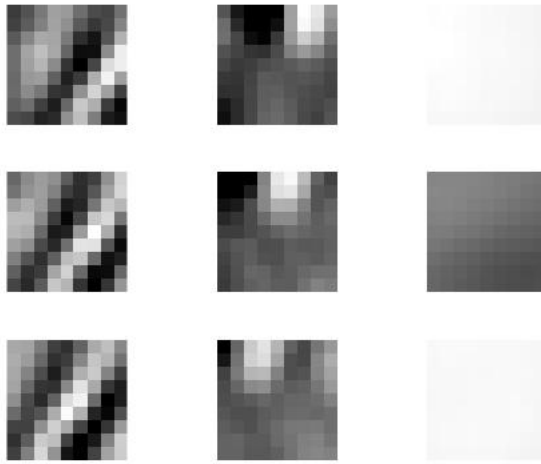


Figure 4.2: $9 \times 9 \times 3$ Dictionary atoms (column wise) Left: Atom 1, Middle: Atom 243, Right: 486.

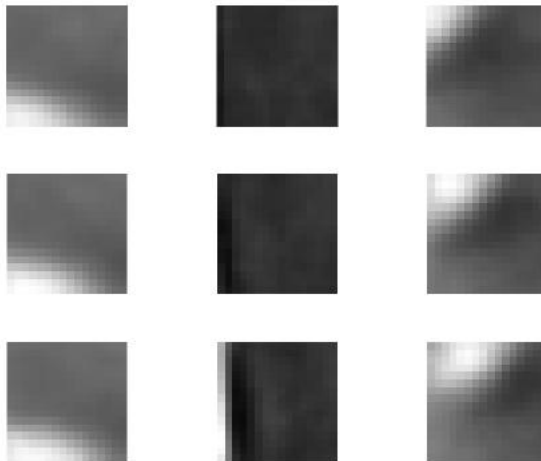


Figure 4.3: $16 \times 16 \times 3$ Dictionary atoms (column wise) Left: Atom 1, Middle: Atom 768, Right: 1536

In the above figures (Fig. 4.1, Fig. 4.2, Fig. 4.3) first, middle and last atoms (column wise) of dictionaries $5 \times 5 \times 3$, $9 \times 9 \times 3$, $16 \times 16 \times 3$ are shown respectively. Clearly we can observe the variations in the atoms both spatially and temporally. We tried to reconstruct the videos from the sparse representations of videos. Reconstructions were pretty good. Visually original and reconstructed video look same. PSNR between the original and reconstructed videos are high.

4.3.2 Sparse Decomposition

We used OMP to generate sparse representation of spatio-temporal patches of video. We generated a sparse matrix using these. Each column in this matrix is sparse representation of a patch. From here we have two

approaches two generate the quality score.

4.3.3 approach-1

From the sparse matrix we generated feature vector by talking $L1$ norm row wise. Each dimension in the feature vector represents absolute strength of the atom coefficient in representing the video sparsely. we follow the same procedure for both test and reference videos. For each scale (dictionary) we get a reference and test feature vectors (f_t and f_r). BY using these we are calculating quality score per scale Q_s .

$$Q_s = \| f_t - f_r \|_2 \quad (4.3)$$

we took weighted combination of these scores to generate final score.

final Quality

$$Q = \sum_{i=1}^{38} W_i Q_i \quad (4.4)$$

We used linear regression to find the weights. We didn't use $16 * 16 * 25$ and $16 * 16 * 31$ dictionaries in calculating final score because of two reasons. 1) Time complexity and 2) reconstruction is not good with these dictionaries.

4.3.4 approach-2

From the sparse matrix we generated feature vector by talking $L1$ norm column wise. Each dimension in the feature vector represents absolute strength of the sparse representation of spatio-temporal patch. we follow the same procedure for both test and reference videos. For each scale (dictionary) we get a reference and test feature vectors (f_t and f_r). By using these we are calculating quality score per scale Q_s .

$$Q_s = \| f_t - f_r \|_2 \quad (4.5)$$

we took weighted combination of these scores to generate final score.

final Quality

$$Q = \sum_{i=1}^{38} W_i Q_i \quad (4.6)$$

We used linear regression to find the weights. We didn't use $16 * 16 * 25$ and $16 * 16 * 31$ dictionaries in calculating final score because of two reasons. 1) Time complexity and 2) reconstruction is not good with these dictionaries.

4.4 Results and Discussion

We used LIVE data set in our work. The LIVE Video Quality Database uses ten uncompressed high-quality videos with a wide variety of content as reference videos. A set of 150 distorted videos were created from these reference videos (15 distorted videos per reference) using four different distortion types. (MPEG-2 compression, H.264 compression, IP-Distortions and Wireless Distortions). Each video in the LIVE Video Quality Database was assessed by 38 human subjects. The mean and variance of the Difference Mean Opinion Scores (DMOS) obtained from the subjective evaluations.

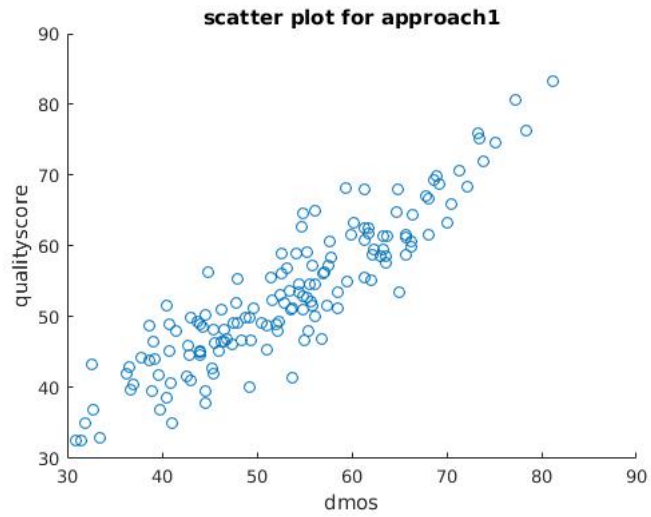


Figure 4.4: Scatter plot of score vs dmos for approach -1

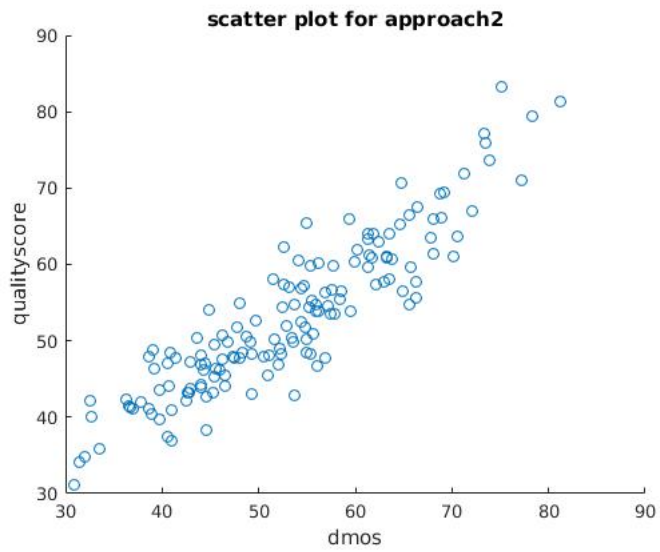


Figure 4.5: Scatter plot of score vs dmos for approach -2

From the above figures we can see that scatter plots for the two approaches are almost linear.

Metrics	LCC	SROCC
Only PSNR	0.4035	0.3684
Only SSIM	0.5498	0.5381
Only VIF	0.5721	0.574
Only FSIM	0.7376	0.7278
Only MS-SSIM	0.7642	0.7482
MOVIE	0.8116	0.789
MOVIE with VQ pooling	0.8611	0.8427
FLOSIM (with MS-SSIM)	0.859	0.8537
<i>Approach-1</i>	<i>0.9051</i>	<i>0.8934</i>
<i>Approach-2</i>	<i>0.9082</i>	<i>0.9007</i>

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

we presented both Spearman rank ordered correlation coefficient (SROCC) and linear correlation coefficient (LCC). From the results table we can say that the proposed approaches performs better than the existing state of art techniques.

4.5 Conclusions and Future Work

we did not use any Spatial metrics (pre existing) in this work. Directly evaluated spatio-temporal scores. If we observe from the literature people evaluated spatial and temporal scores separately and combined them to get spatio-temporal scores. We generated spatio-temporal dictionaries using K-SVD. Finally, we demonstrated the competitive performance of the algorithm on the LIVE SD.

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.

References

- [1] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* 13, (2004) 600–612.
- [2] 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.
- [3] K. Seshadrinathan and A. C. Bovik. Motion tuned spatio-temporal quality assessment of natural videos. *Image Processing, IEEE Transactions on* 19, (2010) 335–350.
- [4] K. Manasa and S. S. Channappayya. An Optical Flow-Based Full Reference Video Quality Assessment Algorithm. *IEEE Transactions on Image Processing* 25, (2016) 2480–2492.
- [5] 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.
- [6] 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.
- [7] 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.
- [8] M. Jung, D. Le, M. Gazalet et al. Univariant assessment of the quality of images. *Journal of Electronic Imaging* 11, (2002) 354–364.
- [9] 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.
- [10] T. Brandão and M. P. Queluz. No-reference image quality assessment based on DCT domain statistics. *Signal Processing* 88, (2008) 822–833.
- [11] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM journal on scientific computing* 20, (1998) 33–61.
- [12] J. A. Tropp and S. J. Wright. Computational methods for sparse solution of linear inverse problems. *Proceedings of the IEEE* 98, (2010) 948–958.
- [13] 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.

- [14] 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.
- [15] R. Rubinstein, A. M. Bruckstein, and M. Elad. Dictionaries for sparse representation modeling. *Proceedings of the IEEE* 98, (2010) 1045–1057.
- [16] J. W. Cooley and J. W. Tukey. An algorithm for the machine calculation of complex Fourier series. *Mathematics of computation* 19, (1965) 297–301.
- [17] C. M. Bishop et al. Pattern recognition and machine learning, volume 4. springer New York, 2006.
- [18] T. S. Lee. Image representation using 2D Gabor wavelets. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 18, (1996) 959–971.
- [19] 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.
- [20] 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.
- [21] A. Mittal, R. Soundararajan, and A. Bovik. Making a Completely Blind Image Quality Analyzer. *Signal Processing Letters, IEEE* 20, (2013) 209 – 212.
- [22] 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.
- [23] 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.
- [24] 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.
- [25] 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.
- [26] M. Saad, A. Bovik, and C. Charrier. A DCT Statistics-Based Blind Image Quality Index. *Signal Processing Letters, IEEE* 17, (2010) 583 –586.
- [27] 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.
- [28] J. M. Bower. 20 Years of Computational Neuroscience. Springer, 2013.
- [29] 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.
- [30] T. Guha and R. K. Ward. On image similarity, sparse representation and kolmogorov complexity .
- [31] 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.

- [32] J. Park, K. Seshadrinathan, S. Lee, and A. C. Bovik. Video quality pooling adaptive to perceptual distortion severity. *Image Processing, IEEE Transactions on* 22, (2013) 610–620.
- [33] M. Saad, A. C. Bovik, and C. Charrier. Blind Prediction of Natural Video Quality and H. 264 Applications. In *Seventh International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VQPM)*. 2013 47–51.
- [34] A. R. Reibman, V. A. Vaishampayan, and Y. Sermadevi. Quality monitoring of video over a packet network. *Multimedia, IEEE Transactions on* 6, (2004) 327–334.
- [35] F. Yang, S. Wan, Y. Chang, and H. R. Wu. A novel objective no-reference metric for digital video quality assessment. *Signal Processing Letters, IEEE* 12, (2005) 685–688.
- [36] M. C. Farias and S. K. Mitra. No-reference video quality metric based on artifact measurements. In *Image Processing, 2005. ICIP 2005. IEEE International Conference on*, volume 3. IEEE, 2005 III–141.
- [37] K.-C. Yang, C. C. Guest, K. El-Maleh, and P. K. Das. Perceptual temporal quality metric for compressed video. *IEEE Transactions on Multimedia* 9, (2007) 1528–1535.
- [38] T.-L. Lin, S. Kanumuri, Y. Zhi, D. Poole, P. C. Cosman, and A. R. Reibman. A versatile model for packet loss visibility and its application to packet prioritization. *Image Processing, IEEE Transactions on* 19, (2010) 722–735.
- [39] Y.-L. Chang, T.-L. Lin, and P. C. Cosman. Network-based H. 264/AVC whole-frame loss visibility model and frame dropping methods. *Image Processing, IEEE Transactions on* 21, (2012) 3353–3363.
- [40] S. A. Karunasekera and N. G. Kingsbury. A distortion measure for image artifacts based on human visual sensitivity. In *Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on*. IEEE, 1994 V–117.
- [41] A. B. Watson and L. Kreslake. Measurement of visual impairment scales for digital video. In *Photonics West 2001-Electronic Imaging*. International Society for Optics and Photonics, 2001 79–89.
- [42] A. B. Watson. Toward a perceptual video-quality metric. In *Photonics West’98 Electronic Imaging*. International Society for Optics and Photonics, 1998 139–147.
- [43] A. B. Watson, Q. J. Hu, J. F. McGowan III, and J. B. Mulligan. Design and performance of a digital video quality metric. In *Electronic Imaging’99*. International Society for Optics and Photonics, 1999 168–174.
- [44] Z. Wang. Rate scalable foveated image and video communications. Ph.D. thesis, University of Texas at Austin 2001.
- [45] W. Zhou, R. Hamid, and A. Bovik. *The handbook of video databases: Design and Applications* 2003.
- [46] Z. Wang, L. Lu, and A. C. Bovik. Video quality assessment based on structural distortion measurement. *Signal processing: Image communication* 19, (2004) 121–132.
- [47] 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.

- [48] Z. Wang and Q. Li. Video quality assessment using a statistical model of human visual speed perception. *JOSA A* 24, (2007) B61–B69.
- [49] H. R. Sheikh and A. C. Bovik. A visual information fidelity approach to video quality assessment. In *The First International Workshop on Video Processing and Quality Metrics for Consumer Electronics*. Citeseer, 2005 23–25.
- [50] H. R. Sheikh and A. C. Bovik. Image information and visual quality. *Image Processing, IEEE Transactions on* 15, (2006) 430–444.
- [51] M. H. Pinson and S. Wolf. A new standardized method for objectively measuring video quality. *Broadcasting, IEEE Transactions on* 50, (2004) 312–322.
- [52] D. M. Chandler and S. S. Hemami. VSNR: A wavelet-based visual signal-to-noise ratio for natural images. *Image Processing, IEEE Transactions on* 16, (2007) 2284–2298.
- [53] A. A. Stocker and E. P. Simoncelli. Noise characteristics and prior expectations in human visual speed perception. *Nature neuroscience* 9, (2006) 578–585.
- [54] K. Seshadrinathan and A. C. Bovik. A structural similarity metric for video based on motion models. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 1. IEEE, 2007 I–869.
- [55] K. Seshadrinathan and A. C. Bovik. An information theoretic video quality metric based on motion models. In *Proc. Third International Workshop on Video Processing and Quality Metrics for Consumer Electronics*. Citeseer, 2007 25–26.
- [56] A. Ninassi, O. Le Meur, P. Le Callet, and D. Barba. Considering temporal variations of spatial visual distortions in video quality assessment. *Selected Topics in Signal Processing, IEEE Journal of* 3, (2009) 253–265.
- [57] A. Ninassi, O. Le Meur, P. Le Callet, and D. Barba. On the performance of human visual system based image quality assessment metric using wavelet domain. In *SPIE Conference Human Vision and Electronic Imaging XIII*, volume 6806. 2008 680,610–1.
- [58] 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.
- [59] P. V. Vu, C. T. Vu, and D. M. Chandler. A spatiotemporal most-apparent-distortion model for video quality assessment. In *Image Processing (ICIP), 2011 18th IEEE International Conference on*. IEEE, 2011 2505–2508.
- [60] S. Wolf and M. Pinson. Video quality model for variable frame delay (VQM-VFD). *US Dept. Commer., Nat. Telecommun. Inf. Admin., Boulder, CO, USA, Tech. Memo TM-11-482* .
- [61] M. H. Pinson, L. K. Choi, and A. C. Bovik. Temporal video quality model accounting for variable frame delay distortions. *Broadcasting, IEEE Transactions on* 60, (2014) 637–649.
- [62] J. H. van Hateren and D. L. Ruderman. Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex. *Proceedings of the Royal Society of London B: Biological Sciences* 265, (1998) 2315–2320.

- [63] B. A. Olshausen. Learning sparse, overcomplete representations of time-varying natural images. In Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on, volume 1. IEEE, 2003 I-41.