Image Information Distance Analysis and Applications

by

Nima Nikvand

A thesis
presented to the University of Waterloo
in fulfillment of the
requirement for the degree of
Doctor of Philosophy
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2014

© Nima Nikvand 2014

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Image similarity or distortion assessment is fundamental to a broad range of applications throughout the field of image processing and machine vision. These include image restoration, denoising, coding, communication, interpolation, registration, fusion, classification and retrieval, as well as object detection, recognition, and tracking. Many existing image similarity measures have been proposed to work with specific types of image distortions (e.g., JPEG compression). There are also methods such as the structural similarity (SSIM) index that are applicable to a wider range of applications. However, even these "general-purpose" methods offer limited scopes in their applications. For example, SSIM does not apply or work properly when significant geometric changes exist between the two images being compared.

The theory of Kolmogorov complexity provides solid groundwork for a generic information distance metric between any objects that minorizes all metrics in the class. The Normalized Information Distance (NID) metric provides a more useful framework. While appealing, the challenge lies in the implementation, mainly due to the non-computable nature of Kolmogorov complexity. To overcome this, a Normalized Compression Distance (NCD) measure was proposed, which is an effective approximation of NID and has found successful applications in the fields of bioinformatics, pattern recognition, and natural language processing. Nevertheless, the application of NID for image similarity and distortion analysis is still in its early stage. Several authors have applied the NID framework and the NCD algorithm to image clustering, image distinguishability, content-based image retrieval and video classification problems, but most reporting only moderate success. Moreover, due to their focuses on specific applications, the generic property of NID was not fully exploited.

In this work, we aim for developing practical solutions for image distortion analysis based on the information distance framework. In particular, we propose two practical approaches to approximate NID for image similarity and distortion analysis. In the first approach, the shortest program that converts one image to another is found from a list of available transformations and a generic image similarity measure is built on computing the length of this shortest program as an approximation of the conditional Kolmogorov complexity in NID. In the second method, the complexity of the objects is approximated using Shannon entropy. Specifically, inspired by the Visual Information Fidelity (VIF) approach a wavelet domain Gaussian Scale Mixture (GSM) model is adopted for entropy computation. Moreover, we propose a tone mapping operator parameter selection scheme for High Dynamic Range (HDR) images. The scheme attempts to find tone mapping parameters that minimize the NID of the HDR image and the resulting Low Dynamic

Range (LDR) image, and thereby minimize the information loss in HDR to LDR tone mapping. The resulting images created by minimizing NID exhibit enhanced image quality.

When applying image information distance framework in real-world applications, we find information distance measures often lead to useful features in many image processing applications. In particular, we develop a photo retouching distortion measure based on training a Gaussian kernel Support Vector Regression (SVR) model using information theoretic features extracted from a database of original and edited images. It is shown that the proposed measure is well correlated with subjective ranking of the images.

Acknowledgements

This thesis would not have been possible without the help and mentorship of Dr Zhou Wang. His knowledge and creative mind and his supportive supervision had great impact on my understanding of subjects and research methodology, and I am very grateful for the opportunity I had to be a member of his research team.

University of Waterloo is a great place to study and research. I am very grateful to have had the chance to study in such a great environment and enjoy companionship of great researchers. In particular I would like to thank Dr Ming Li for his invaluable support and cooperation in the field of Kolmogorov complexity, and my colleagues Mr Hojatallah Yeganeh and Dr Abdul Rehman for their help on various problems I encountered during my research.

Last but not the least, I would like to thank my parents. My mother Forough and my father Jahansha, whose unconditional love and support has been the main source of inspiration and confidence in my life. This thesis is thankfully dedicated to them.

Dedication

This thesis is dedicated to my parents, Forough and Jahansha.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Image similarity and quality measurement are fundamental problems in everyday image processing applications. It is a basic task of Human Visual System (HVS) to note the similarities and differences between images, perceive the quality of images or rank a set of images based on similarity; however, this is not as easy with computers. Image similarity measures account for the structural similarity information held in one image in relation to another. An accurate image similarity measure can be used to compare images and retrieve images based on their content. Furthermore, a perceptual similarity measure can be used to estimate perceived image quality by measuring the similarity between the reference and distorted images [1]. The most accurate method of measuring image quality is to obtain opinions of human subjects which is time consuming and expensive and cannot be incorporated into automatic algorithms. The goal of an automated (objective) perceptual Image Quality Assessment (IQA) method is to accurately predict the quality of an image as evaluated by human subjects. Thus, it is important to evaluate the performance of an objective quality assessment method against existing subjective studies. Based on the availability of a reference image (distortion-less image) for comparison with the distorted image, IQA methods are divided into Full Reference (FR), Reduced Reference (RR) and No Reference (NR) methods. The FR quality metrics are used when a reference image is available for comparison. RR quality metrics attempt to predict the quality of the image when only partial reference image information is available. This method can be useful in reducing the required bandwidth of communication systems, by transmitting only certain information from the reference image, which can be used to predict the perceptual quality of the received images and to track visual degradations to control streaming resources [2]. The NR image quality assessment methods provide a blind assessment of the quality of the image when no reference image is available. The main focus of this work is on FR image

similarity methods.

Traditionally, the most widely used form of image quality assessment was Mean Squared Error (MSE) and its associated Peak-signal to Noise Ratio (PSNR). These are appealing since they are easy to calculate and optimize and have clear physical meaning in most signal processing applications. However, it has been long known that MSE and PSNR are inconsistent with the HVS and that the MSE and PSNR scores do not reflect the quality of an image as perceived by the HVS [3]. In the past few decades, there has been a great deal of effort to develop an IQA method which is consistent with the HVS and several successful methods have been proposed, most notably, the Structural Similarity Index Measure (SSIM) [3] and, the Visual Information Fidelity (VIF) [4]. The SSIM is currently the most highly cited algorithm for IQA used in most applications, and is based on the hypothesis that natural images are highly structured, in the sense that the samples of natural image signals have strong neighbor dependencies, and these dependencies contain important information about the structure of an image. In this context, the HVS is assumed to be an ideal information extractor that is sensitive to structural distortions and compensates for non structural distortions by default and, thus, SSIM mimics the HVS by capturing structural distortions such as additive noise or blur and lossy compression that exists in distorted image compared to the reference image and compensates for non-structural distortions such as luminance change, contrast change, Gamma distortion and spatial shift which are usually caused by environmental conditions during the process of acquiring the image [5]. Visual Information Fidelity (VIF) is another highly cited method which provides a novel information theoretic approach towards the IQA problem. Visual Information Fidelity uses Natural Scene Statistics (NSS) in conjunction with distortion models and quantifies the Shannon information shared between the reference image and the distorted image as seen by the HVS. Note that VIF is a fraction of the reference image information that could be extracted from the distorted image [4].

Our work is largely inspired by the theory of Kolmogorov Complexity (KC). Kolmogorov complexity based Normalized Information Distance (NID) provides a universal similarity metric framework which is applicable to any two objects regardless of their type [6]. Kolmogorov complexity is non-computable and is usually approximated using a practical data compressor. Normalized Compression Distance (NCD) is an approximation of NID using a data compressor and is proved to hold metric properties and universality of NID up to a constant. Normalized Compression Distance has many successful applications in the fields of bioinformatics, pattern recognition and natural language processing [7]. Applications of NID or NCD to image processing, image similarity measurement and image quality assessment are still in their early stages, with several authors reporting limited success in applying NCD to image retrieval [8], image clustering [9], and video classification [10].

1.1 Motivations

The motivation of our work is to imitate the HVS in capturing similarities among images using a simple hypothesis: Visual scenes are highly redundant, the optic nerve is a limited channel, and humans need to react quickly to changes in visual images. Therefore, an efficient coding scheme is required in the HVS [11]. Based on the same hypothesis, Barlow's sensory coding principle asserts that information is encoded in the neural system such that a minimum number of spikes are required to transmit a visual or audio observation to the brain [12]. In a separate series of work relating this hypothesis to experimental data and trying to model how the brain works, Attneave showed that the retina's role is mainly to get rid of redundancy in the visual world [11,13]. This hypothesis, Barlow's efficient coding principle, and the experimental results by Attneave draw some connections between the biology of the HVS and description of images in their shortest form, in other words a non-redundant code for describing the image as Shannon would call it, or an optimal statement, the length of which is called its Kolmogorov complexity. This might be exemplified by an image of 8×8 checkerboard and another image of the same checkerboard but with the white and black squares exchanged. Humans quickly notice this change and find the two image similar. However most existing image similarity measurement methods based on pixel differences fail to recognize this similarity. When asked to describe the difference between the two images, a simple description is usually given, to flip the color of all squares. Kolmogorov complexity is a tool that provides a similar explanation. Given the first image, Kolmogorov complexity looks for the shortest program to go to the second one on the Universal Turing Machine (UTM), and one possible answer could be by simply flipping the bits in the first checkerboard image.

The theory of Kolmogorov complexity provides a solid groundwork to build a universal and generic information distance metric between any objects that minorizes all metrics in the class [6]. Applications of the information distance metric to image processing are still in the early stages and although several authors have done pioneering work on NID and have applied it to image clustering [9], image distinguishability [14], content-based image retrieval [8] and video classification [10] problems, most of these authors have reported moderate success. Despite its success in bioinformatics applications, the metric remains relatively undiscovered in the image similarity measurement field and by applying it to images, improvements in image similarity and perceptual image quality assessment can be reached.

1.2 Objective

The objective of this research is to explore potential applications of information distances in the field of image similarity and image quality assessment, and develop novel frameworks and algorithms for practical applications based on Normalized Information Distance.

1.3 Contributions

In this thesis we aim to introduce information distance framework into the field of image similarity and image quality assessment, The major contributions of this thesis are as follows:

- A practical framework for the approximation of NID is developed. Normalized Conditional Compression Distance (NCCD), is a flexible and expandable framework that uses a list of possible transformations that convert one image to another along with an image compressor to approximate the non-computable Kolmogorov complexity terms in NID, and in effect finds the simplest transformation of one image to another in accordance with visual sensory coding principle. Two different implementations of NCCD based on image compression and video compression are proposed and it is shown that the NID based similarity framework has a wide range of applicability in problems where most other image similarity measures fail.
- A NID based perceptual distortion analysis method is developed based on approximating Kolmogorov complexity using Shannon's entropy, and it is shown that the method is competitive with state-of-the-art image quality assessment algorithms when tested using subject-rated image databases.
- A parameter selection scheme based on entropy approximation of NID is proposed for tone-mapping of High Dynamic Range (HDR) images, and it is shown that the Low Dynamic Range (LDR) images tone-mapped using this scheme have enhanced quality.
- A set of information distance based features which quantify a higher bound for the average number of bits required to describe modifications carried out on a pair of original and edited images is proposed, and it is shown that the features can be used to predict perceptual scores for quantifying the impact of modifications carried out on images when used along with a Support Vector Regression (SVR) tool that is trained by human observer data.

1.4 Thesis Organization

The organization of the rest of the thesis is as follows: Chapter 2 discusses the background of Kolmogorov complexity, normalized information distance and its applications in signal and image processing and existing objective image quality assessment methods. Chapter 3 introduces Normalized Conditional Compression Distance (NCCD) framework for the approximation of NID, proposes two implementations for it, and tests the applicability of the framework on a wide range of applications including texture and face recognition. Chapter 4 introduces two NID based frameworks by using Shannon entropy approximation of Kolmogorov complexity. The first framework extends application of NID to perceptual image quality assessment, and the second framework presents a parameter selection scheme for tone-mapping operators which minimizes information loss in tone-mapping by finding the optimum parameters for the operator of choice. Chapter 5 presents a set of information distance based features which can be used by machine learning algorithms to quantify the impact of modifications carried out on images. Finally, Chapter 6 concludes the thesis, and discusses future research directions.

Chapter 2

Literature Review

2.1 Kolmogorov Complexity

There are many ways to describe a given object by a finite string and a finite alphabet. Assuming that all the strings are describing exactly the same object, a natural preference is given to the shortest string, which is associated with the simplest description of the object. Objects are simple if they have relatively short descriptions and are complex if they have relatively long descriptions. To measure the relative simplicity or complexity of the objects accurately, a framework is needed in which each string describes, at most, one object with finite descriptions, providing for a countable set of descriptions and objects. A descriptional complexity can be defined as the least number of bits required to transmit a message x using a specification function D through a communication channel. However, this definition relies on the specification function and the complexity can vary depending on the choice of D. To be able to objectively compare descriptional complexity of an object, we need the definition to be independent of any parameter other than the object itself. The complexity can be defined under a general specification function which is assumed to be universal and is optimal if a lower complexity is not achieved by any other method. To have a useful specification method, D, we need this method to be executed in an effective manner, which means it must be performable by mind or machine. The notion of effective computation is characterized by partially recursive functions in mathematics [15]. The optimal partially recursive function minimizes the description of every other function in the set of partially recursive functions. If this function is denoted by D_0, for any other partially recursive function D there is an x which has a shorter description y under D_0 than any description of x such as z under D. The length of the description of object x

under D_0 is defined as algorithmic complexity of x and is independent of the choice of the specification function D_0 up to a constant. Partial recursive functions are intended to model any function that can be calculated using a mechanical calculation device (or, alternatively, a human mind) given infinite time and storage space. Alternatively, any function that can be implemented through an algorithm is also considered as computationally effective and partially recursive.

Kolmogorov complexity of an object is defined to be the length of the shortest program that produces that object on a Universal Turing Machine (UTM) and halts [15]:

$$K(x) = \min_{U(p)=x} l(p) \tag{2.1}$$

where U is UTM and p is the shortest program that produces the object x. Thus, $K(x)$ is the number of bits of information that can be retrieved from object x computationally. The Kolmogorov complexity used throughout this text will be the "prefix" version, meaning that none of the programs ran on the UTM to find the shortest program will be a prefix to another. Note that we need to have object x produced from the shortest form x^*, but we do not need to have a general compressor that compresses x into x^*. It is proven that such a compressor does not exist [15], and Kolmogorov complexity is a non-computable function. Hence, Kolmogorov complexity is a lower-bound on all computable functions of all computable functions that compress object x into a shorter program.

The conditional Kolmogorov complexity $K(x|y)$ of an object x given y is similarly defined as the length of the shortest program p that produces the object x on a UTM, if object y is given to the machine as side information:

$$K(x|y) = \min_{U(p,y)=x} l(p) \tag{2.2}$$

Although both notations $K(x)$ and $K(x|y)$ are defined based on the UTM, both concepts are machine independent up to an additive constant based on Church's thesis, which shows that the UTM is capable of simulating any machine that can be effectively computed [15]. This means that the lengths of the shortest prefix program to produce an object x in two programming languages, such as LISP and C++, are equal up to a constant length, where the constant number of bits are used to implement one programming language using the other.

An upper semi-computable function is defined to be a real-valued function $f(x, y)$ such that there exists a rational-valued recursive function $g(x, y, t)$ which decreases in time and has a limit as the time approaches infinity equal to $f(x, y)$, and is defined to be lower

semi-computable if $-f(x, y)$ is upper semi-computable. The definition for upper-semi computable functions may be written as:

- $g(x, y, t + 1) \leq g(x, y, t)$
- $\lim_{t\to\infty} g(x, y, t) = f(x, y)$

It is shown that functions $K(x)$ and $K(y|x^*)$ are upper semi-computable and that they are non-computable [6].

The information content of x, contained in y is defined to be [15]:

$$I(y; x) = K(x) - K(x|y^*) \tag{2.3}$$

The joint Kolmogorov complexity of two objects, $K(x, y)$, is defined to be the length of the shortest program which prints out objects x and y and also the way to distinguish between the two. It is shown that the joint complexity is symmetric up to a constant [16]:

$$K(x, y) = K(x) + K(y|x^*) = k(y) + k(x|y^*) \tag{2.4}$$

and hence with equality up to a constant "c" [1], we have:

$$I(x; y) = I(y; x) \tag{2.5}$$

Interestingly, both of these relationships resemble those in Shannon information theory, which is built upon statistical frameworks.

2.1.1 Information Distance

Much research has been done in similarity measurement among two objects. Depending on the application, we are interested in clustering objects or ranking raw data, be it of any type, based on their similarity and classifying them in groups that reflect their common characteristics. However, to define a measure of similarity, we need an application independent framework that formulates all the existing similarities between the two objects

[1] It is notable that in Kolmogorov complexity literature, most of the equalities and inequalities are said to be true up to an additive constant which accounts for the length of a fixed binary program that is independent of the variables, hence it is customary to show this constant with $O(1)$, which resembles a term of order one in complexity.

while uncovering on dominant similar features. Such a framework is of course based on inherent characteristics of objects and independent of the object type, and, hence, applicable to similarity measurent of objects from different classes. A natural choice would be to compare the binary strings that define the two objects on a UTM. The strings could have different lengths since they are representing objects of a different nature. In our quest for a similarity measure, we are interested in a framework which compares the information content of the two binary strings; hence, it is apparent that the similarity measure would be based on Kolmogorov complexity of the objects, carrying the essence of the objects in a compressed form. Formally, there are several definitions for the information distance based on Kolmogorov complexity. A preliminary form of information distance is defined as follows [17]:

$$E(x, y) = \max\{K(y|x), K(x|y)\} \tag{2.6}$$

It follows from the definition that the information distance, E, is also upper semi-computable, since we can run all the prefix programs p such that $U(p, x) = y$ and p' such that $U(p', y) = x$ and find shorter programs until we obtain a limit $|p| = E(x, y)$, with the potential of spending an infinite amount of time in the process.

For the information distance to be appealing in applications, it is required to be a metric. For a non-negative distance function D and a space X it is required that for every $x, y, z \in X$ we have:

- $D(x, y) = 0$ iff $x = y$ *(identity Axiom)*
- $D(x, y) + D(y, z) \geq D(x, z)$ *(triangle inequality)*
- $D(x, y) = D(y, x)$ *(the symmetry axiom)*

A set X which is provided with a distance function D that satisfies metric properties is called a metric space. It was shown in [17], that $E(x, y)$ is a metric up to an additive fixed constant.

This definition of information distance has the property that it minorizes every admissible distance up to an additive constant [17], where an admissible distance is a function D: $\Omega \times \Omega \to \mathbb{R}^+$ which is upper semi-computable, and symmetric. For every pair of $x, y \in \Omega$ the distance $D(x, y)$ is the length of a binary prefix codeword that is a program that computes x from y, and vice versa, in the reference programming language, and $\Omega = \{0, 1\}^*$ [6].

2.1.2 Normalized Information Distance

Unnormalized information distance is not an accurate measure of similarity among objects of different types, most importantly due to the fact that it cannot adequately compare objects of different lengths. As an example, in the information distance previously introduced, $E(x, y)$, two strings of length 10^6 bits that are different by 10^3 bits are more distant than two strings of the length 10^3 bits that are different by 10^3 bits. Thus, there needs to be a type of normalization to compensate for the length of the two objects being compared. Several attempts have been made to formulate a normalized version of information distance with similar properties to $E(x, y)$. Most notably Ming Li, et al, proposed two versions of normalized distance in [6], one of which, widely known as Normalized Information Distance (NID) is the central inspiration in our work.

The first attempt to normalize information distance is based on normalizing the summation of conditional complexities of the two objects by the joint complexity [6]:

$$d_s(x, y) = \frac{K(x|y^*) + K(y|x^*)}{K(x, y)} \tag{2.7}$$

Since it is known that $I(x; y) = I(y; x)$ up to an additive constant, the above notation is simplified to:

$$d_s(x, y) = 1 - \frac{I(x; y)}{K(x, y)} \tag{2.8}$$

It is shown that this distance satisfies the triangle inequality, up to a small error term and maintains universality within a factor [6].

Another attempt for the normalization of information distance by the same authors is shown to be more successful. It is formally known as Normalized Information Distance (NID) and is shown to be precisely symmetrical and satisfies the identity axiom and triangle inequality up to a precision of $O(1/K(x))$. Note that NID is defined as follows:

$$\mathrm{NID}(x, y) = \frac{\max\{K(x|y^*), K(y|x^*)\}}{\max\{K(x), K(y)\}} \tag{2.9}$$

A thorough proof of these properties may be found at [6] and [15].

2.1.3 Applications of NID

All the nice properties of NID necessitate non-computability of the Kolmogorov complexity terms which are inherent to it. Although non-computable, Kolmogorov complexity is upper semi-computable, meaning that given time, an implementation of UTM can run many possible programs and find one of the short programs that produce the object x and use its length as an approximation of the Kolmogorov complexity of the object. If the object consists of a sequence of random coin flips, obviously the short form program x^* that we can find is not much shorter than the original sequence, but if it is supposedly the first 10^{10} bits of $\pi = 3.1415\cdots$, we have a very short program to produce the object. Since most of the strings of bits we deal with are random, there are precisely very few strings which contain the kind of regularities that we are interested in, and these regularities may be captured using natural compressors [7]. We may replace the unconditional Kolmogorov complexity terms with compressors. As for the conditional Kolmogorov complexity terms, they can be simply replaced by joint complexity using equation 2.4:

$$\mathrm{NCD}(x, y) = \frac{C(xy) - \min\{C(x), C(y)\}}{\max\{C(x), C(y)\}} \tag{2.10}$$

This approximation of NID is called Normalized Compression Distance (NCD) and is shown to have properties similar to NID with some margin for error [7]. Several successful applications of NCD have been reported in DNA sequence analysis and in creating phylogeny trees for animals and languages [7].

Normalized Compression Distance has also been applied to audio processing applications and it has proven successful for genre classification of MIDI files. Four groups of MIDI files with different genres were used to test the performance of the NCD. The Lempel Ziv 78 (LZ78) compression algorithm has been used to approximate NCD due to its simplicity of implementation. For better classification, the authors remove number of bits required to represent the codewords for blocks and use a simple measure of compressed object size without unnecessary details. The four groups of MIDI files were fed into compressors and were ranked according to NCD as well as a Support Vector Machine (SVM) and a Statistical Language Model (SLM), both driven by bigram and trigram features. There results show that NCD performed better than SLM and SVM in classification of the MIDI files according to their genre [18]. Criticism of this work may be categorized into three important groups. First, their choice of MIDI files is not balanced as they have a different number of files in different groups. Second, their tendency to work with a simple coding algorithm prevents them from using generally higher compression ratio compressors for music, namely block-sorting compressors such as bzip2 and finally their choice of melody

contours such as their decision on not including melody duration features in MIDI files would also be of general importance.

Several authors have also used NID and NCD in applications such as clustering images, content-based image retrieval and image distinguishability using different test sets and different compression algorithms. In [9] the authors use bzip2 compressor to obtain a distance matrix by calculating pairwise distances among objects of data set. The authors have three different test sets, one basic test set consisted of images of three different letters (A,B,C) varying position of the letter inside the image from top left, top right, bottom left and bottom right with an image size of 256×256. Another test set was sixty images of the three letters, twenty of each, with four rotations at ninety degree intervals. The image size used in this test was 512×512. The last test set consisted of fifty images of the three letters (approximately fifteen of each) with unusual font styles and same image size as the previous set. For all three cases, the authors tried Z-ordering as well as bit interleaving and direct concatenation of images to estimate the NID. In all tests, direct concatenation turns out with the best results. For the first test set, the experiment is quite successful considering both minimum distances and average distance for clustering algorithm. In the second test set also authors report an interesting success using average distances and direct concatenation. The third set however, is a failure considering any type of NCD estimation as well as using either minimum or average distances for clustering.

In [8], the authors explored possibility of using NID on content-based image retrieval (CBIR). The idea in CBIR is to rank images based on their similarity to a query image. To further increase possibility of detecting objects of interest located in different spatial locations in two images, the authors propose partitioning the image into n blocks of equal size, and then interleaving the blocks such that every pair of block appear next to each other once. Using this partitioning, the distance between each measure is defined to be as the following:

$$d(x,y) = \frac{\min_{j=1,2,\cdots,n}\{\max\{(|c(xy_j)| - |c(y)|), (|c(yx_j)| - |c(x)|)\}\}}{\max\{|c(x)|, |c(y)|\}}. \tag{2.11}$$

NID was tested against uniform random retrieval as well as several feature based methods. Five different image libraries were tested which consisted of different types of images. The image sets included Texture, Letter, GroundTruth, IAPR-12 and Corel. Each image set was tested using a special compression algorithm and images were ranked according to their distance defined in equation 2.11 with the query image. The authors report very promising results for detection of texture and letters via NCD.

In [14], the author tests NID with the famous Goldmeier test set and compares the

Image	NCD(x,y)	NID(d)
Watermarked	0.64	0.71
Graffitized	0.57	0.61
Rotated	0.97	0.86

Table 2.1: Baboon compared to its watermarked, Graffitized and rotated version using NCD

performance of NCD with several different compression schemes. He concludes that NCD is not compatible with perceptual visual similarity. The author also conducts a simple experiment that we reproduce in this paper. The idea is to compare the performance of NID by directly estimating the numerator from subtraction of the two images or by using NCD for an image that is rotated, watermarked, and graffitized. The original paper uses a 512×512 image of lena to test the performance. We have regenerated the results using a "bmp" type 512×512 photo of Baboon. Using gzip compression algorithm in MATLAB, we estimated NID by subtracting the original image from the test images and compressing the resulted one-dimensional string as well as using NCD. The image was watermarked using a free watermarking program called Invisible Ink and rotated via simple MATLAB functions. Our general results seem to disagree with that of [14]. In both cases, NCD distinguishes the watermarked and graffitized image. This maybe because the author in [14] uses a different watermarking scheme or because the graffitized image is more affected by the modification. The results are shown in Table 2.1.

In [19] the authors propose to use complexity based analysis in Earth Observation Imagery as an alternative to classical methods that require a priori data model and report successful clustering of earth observation images as well as image artifact detection and mining satellite image time series using NCD. In [10], the authors use NCD to classify video based on genre and report a hit ratio of 95% for the method in searching compressed video databases.

Normalized Information Distance was also proposed to be used in compression-based image registration [20]. This work is based on the conjecture that the optimally registered image would achieve the highest compression given the fixed image. Figure 2.2 shows the main components of this process.

The metric in the registration process is of crucial importance and is a tool to quantify the quality of alinement of the two image at each iteration. NID is used as a metric

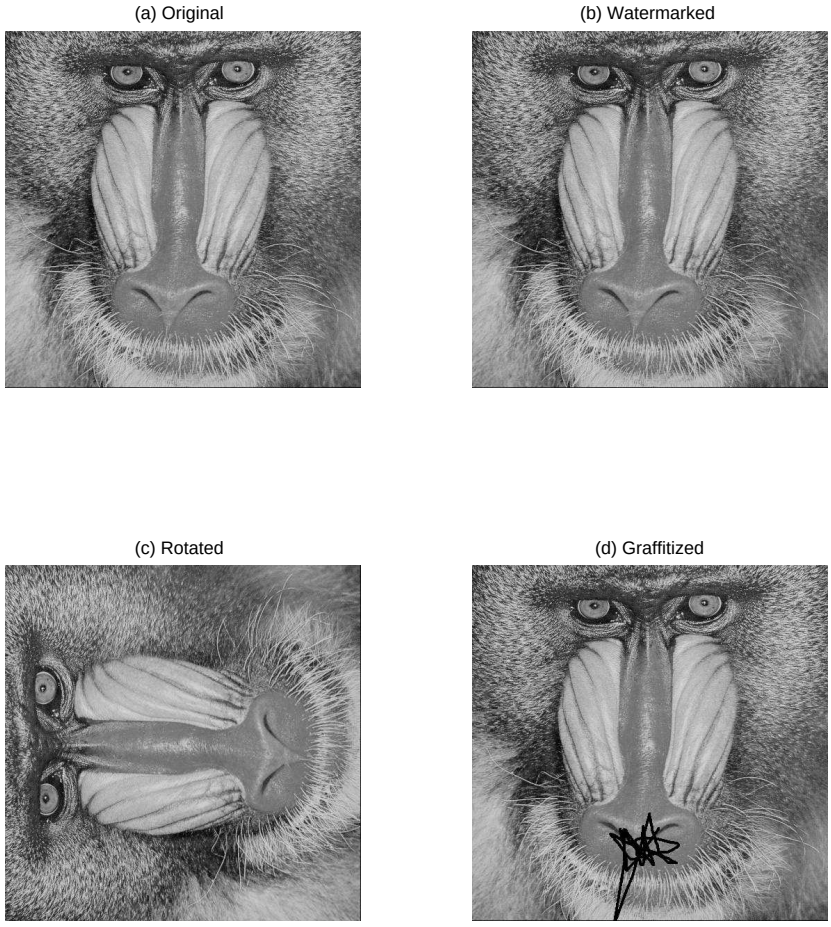


Figure 2.1: NCD tests using images with geometric and compound distortions.

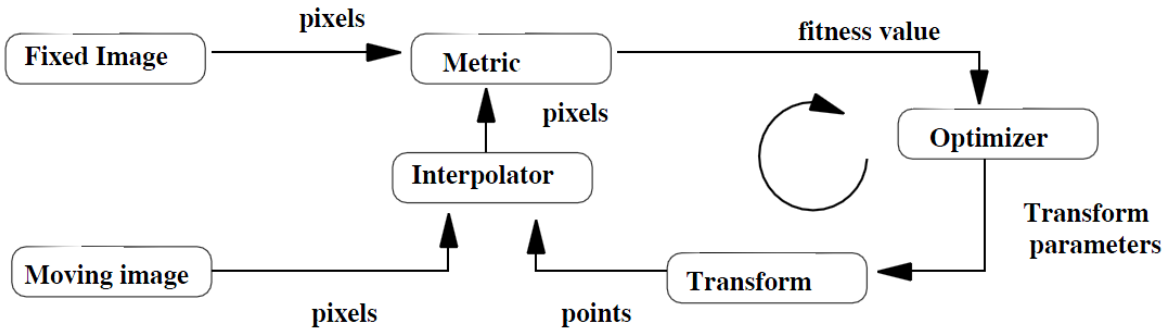


Figure 2.2: Image registration using Normalized Information Distance [20].

in this process as opposed to Normalized Mutual Information (NMI) or the more commonly used Mutual Information [20]. The authors iteratively compress a moving image given a fixed image using three different compression schemes. One scheme is based on a JPEG2000 compression of the biggest rectangle that is common between the two images. Another is based on a bzip2 compression of common sequences of the two images and, finally, an approximation of Kolmogorov complexity using Shannon entropy. The paper concludes that registration based on Shannon entropy is more accurate than the other two compressor-based methods.

Other compression-based distances have also been proposed for texture classification and texture similarity applications, notably in [21]. The main advantage of using such distances over the commonly used methods is that these distances do not require domain specific parameters and avoid problems such as over fitting, which are frequent in other methods such as Gabor filtering. In [21], the authors propose a MPEG-based distance which shows properties such as symmetry and positive definiteness and test the measure against a variety of comprehensive data sets and report promising results.

2.1.4 Remarks

We reviewed the basics of Kolmogorov complexity and complexity based information distances. NID was studied as a metric that will be central to our work. Several applications of NID and its practical estimation tool, Normalized Compression Distance (NCD), were also reviewed. The broad range of applications of NID in DNA sequence analysis, linguistics, audio and visual information processing, including image and video similarity, encouraged us to further research applications of this framework in image similarity and

quality assessment. Although many authors have reported moderately successful use of NCD in classifying, clustering, and measuring similarity of objects, there is a research opportunity in this area for formulating NID as a similarity index for images and by improving NCD approximation. One possible criticism to the available literature in this field is the approximation of NCD by replacing conditional complexity in its numerator with a concatenation term. Although the relationship stands up to a constant in terms of Kolmogorov complexity, the replacement is not necessarily the best option for estimating the conditional Kolmogorov complexity when a practical compressor is applied. More accurate approximation of the conditional Kolmogorov Complexity may also be achieved by using the relationship between Kolmogorov complexity and Shannon's entropy, as authors in [22] and [23] have suggested.

2.2 Image Quality Assessment

Objective quality assessment of videos and images is of fundamental importance to a broad range of applications in telecommunication, machine vision, and image and video processing. The goal of an objective quality assessment algorithm is to automatically predict the quality of an image or a video clip in accordance with subjective human quality assessment. Over the past decade, many algorithms have been proposed that have had significant contribution to this field and many subjective image and video quality assessment databases have been created based on the opinion of human subjects. Subjective image quality databases are created to validate the performance of the objective algorithms based on human judgment and compare the performance of different methods with each other. Subjective studies are cumbersome and expensive since they usually involve a great number of human subjects rating images which are corrupted by different types and levels of distortion by comparing them to reference images. The IQA algorithms are generally divided into three categories of Full Reference (FR), Reduced Reference (RR) and No Reference (NR) methods based on the availability of a reference image or partial information of the reference image, for the assessment of the quality.

The quality of an image signal is traditionally evaluated based on the assumption that it is the summation of an undistorted reference signal with an error signal. It is also assumed that visibility of error signal has direct effect on the loss of perceptual quality. The simplest implementation of this assumption is Mean Squared Error (MSE), which quantifies the strength of the error signal. However images with different types and levels of visible distortions could have the same MSE. In other words, MSE is not consistent with perceived quality of the HVS, and despite being an easy tool for optimization, it cannot be used as

an effective metric in IQA. To overcome this issue, traditional approaches in the literature propose to weight different aspects of the error signal based on different assumptions on error visibility, which are usually determined by psychophysical measurements in human subjects or physiological measurements in animals [3].

A typical IQA method based on error weighting assumption is shown in Figure 2.3. A pre-processing stage is common among such frameworks to eliminate known distortions from the images being compared, and properly scale and align them with each other. The IQA methods might also need to transform the reference and distorted images into a color space which is more suitable for the HVS, or simulate point spread function of the eye optics using a low pass filter. All of these functions are usually embedded in the pre-processing stage [3]. The Contrast Sensitivity Function (CSF) is used to account for the sensitivity of human visual system to different spatial and temporal frequencies that are present in the visual stimulus. The images are then seperated into subbands which are selective for spatial and temporal frequency as well as orientation, this stage is called Channel Decomposition and is intended to simulate the neural responses in the primary visual cortex [3]. The difference between the decomposed reference and distorted images are then calculated and normalized according to a certain masking model, taking into account many factors such as proximate in spatial or temporal location, spatial frequency, or orientation. This is done at the Error Normalization block, where the error signal is weighted by a space-varying visibility threshold, which is calculated based on the energy of reference or distorted coefficients in a certain neighborhood and base-sensitivity for that channel. This normalization converts the error into units of just noticeable difference (JND). The last step of the quality assessment combines the normalized error signals over the spatial domain of the image, and across different channels, this is called Error Pooling, and usually involves calculating a Minkowski norm over the spatial domain or different channels of error signal as follows:

$$E(\{e_{l,k}\}) = (\sum_{l}\sum_{k} |e_{l,k}|^{\beta})^{\frac{1}{\beta}} \tag{2.12}$$

where $e_{l,k}$ is the normalized error of the $k-th$ coefficient in the $l-th$ channel, and β is a constant exponent which ranges between 1 to 4 [3].

This generic framework however, has many intrinsic limitations which make it unable to predict a quality score consistent with the HVS. The most fundamental problem of this framework is the definition of image quality itself [5]. There is no scientific evidence that shows error visibility and loss of quality are directly correlated. Furthermore, many examples can be found that show that some types of distortions (e.g. contrast enhancement)

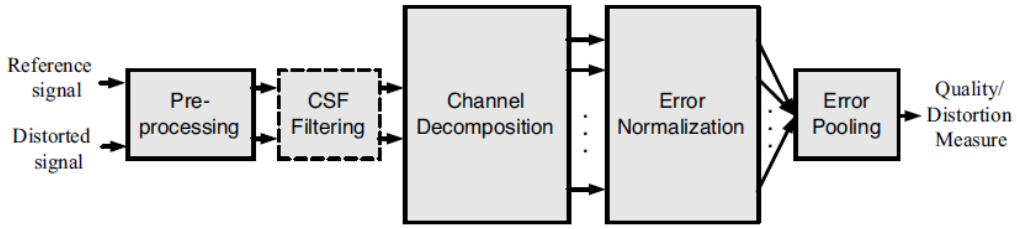


Figure 2.3: Generic image quality assessment framework based on error sensitivity [3].

may be clearly visible but not so objectionable to the human eye. Another issue is that error sensitivity models are designed to estimate the threshold at which a stimulus becomes visible, however the assumption that such thresholds may be generalized to characterize perceptual distortions are not generally supported by psychophysical studies [3]. Moreover, most of the psychophysical studies conducted in the area are based on simple images such as spots, or sinusoidal gratings, which cannot effectively model the complexity of natural images. The use of Minkowski error in such frameworks also raises a problem since it comes with the implicit assumption that the visibility of error at different locations is statistically independent, where we know that in practice, strong correlation exists among different subbands and in fact such correlations (and often redundancies) are used in modern compression techniques to achieve higher compression ratios in image applications, compared to traditional data compression approaches.

All of these intrinsic limitations of error sensitivity approaches towards the IQA problem call for a need of different approaches which do not have such inconsistencies with the HVS and are based on new assumptions which are more adapted to the psychophysical studies carried out in this area.

In this chapter, we review two important objective quality assessment algorithms based on new assumptions which are shown to be more consistent with the HVS and perform consistently closer to subjective studies in predicting a quality score for different distorted images we also review six widely cited subjective quality assessment databases, all of which are to be used to validate our proposed method in the next chapters.

2.2.1 Structural Similarity Index Measure

Structural Similarity Index Measure (SSIM) is the most highly cited FR IQA method which is based on the assumption that human visual perception is highly adapted for extracting

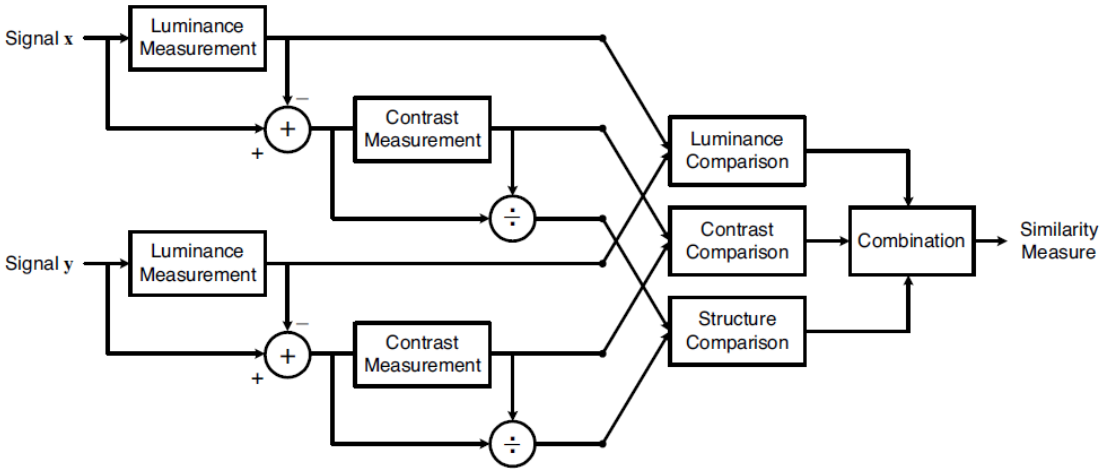


Figure 2.4: Generic framework for SSIM measurement [3].

structural information from a scene [3, 5], and the structural dependencies embedded in natural images carry a lot of important perceptual information for the HVS, especially when it comes to spatial proximate and dependencies among different neighborhoods of a natural image. Contrary to the error sensitivity approach, SSIM is a top-down approach which tries to model the overall functionality of HVS by measuring the structural distortions to achieve an image fidelity measure. To explore the structural information in an image, SSIM separates the effect of luminance information from structural information since luminance of the surface of an object which is being observed in the image is only the product of the illumination and the reflectance, and does not carry any structural information. Similarly, contrast information of the reference and distorted images are removed to provide a better judgment of the structural information embedded in the images. Figure 2.4 shows a generic framework for SSIM quality measurement.

Similarity measurement in this system is separated into comparison of luminance, contrast and structural information of images [3]. The luminance of the two images are estimated by as means of intensity:

$$\mu_x = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{2.13}$$

The luminance information of the image is then removed by subtracting the mean from

the image $\mathbf{x} - \mu_x$, and a luminance comparison function l(x,y) is defined:

$$l(x, y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \tag{2.14}$$

where C_1 is a constant to avoid situations where $\mu_x^2 + \mu_y^2$ is close to zero and is chosen to be $C_1 = (K_1 L)^2$, where L is the dynamic range of the images and $K_1 \ll 1$ is a small constant [3].

Similarly, the contrast information of the image may be obtained using an unbiased standard deviation estimator:

$$\sigma_x = (\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu_x)^2)^{\frac{1}{2}}. \tag{2.15}$$

The contrast comparison function is similar to the luminance comparison function:

$$c(x, y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \tag{2.16}$$

where $C_2 = (K_2 L)^2$ and $K_2 \ll 1$. The images are then normalized by their corresponding contrast information and the two normalized and luminance removed images ($(\mathbf{x} - \mu_x)/\sigma_x$ and $(\mathbf{y} - \mu_y)/\sigma_y$), are used for comparison of the structure. The structure comparison function is defined as:

$$s(x, y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \tag{2.17}$$

where σ_{xy} is the cross correlation between images x and y and is estimated as:

$$\sigma_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y) \tag{2.18}$$

and C_3 is a constant similar to C_2 and C_1 [3].

Finally, the Structural Similarity Index Measure is defined as a combination of (2.14), (2.16) and (2.17) as follows:

$$\text{SSIM(x,y)} = [l(\mathbf{x}, \mathbf{y})]^{\alpha}.[c(\mathbf{x}, \mathbf{y})]^{\beta}.[s(\mathbf{x}, \mathbf{y})]^{\gamma} \tag{2.19}$$

where $\alpha > 0$, $\beta > 0$, $\gamma > 0$ are parameters which are used to adjust relative importance of the corresponding components and are usually taken to be one. If we take $C_3 = C_2/2$, we

can simplify the results as follows [3]:

$$\text{SSIM(x,y)} = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{2.20}$$

The SSIM framework also has several qualities which make it consistent with well-known properties of the HVS. For example it is shown that (2.14) is philosophically consistent with Weber's law, which is widely used to model light adaptation in HVS and also (2.16) is consistent with contrast-masking features of HVS [3].

The SSIM framework also satisfies some conditions such as symmetry, boundedness, and unique maximum which makes it an effective tool in similarity measurement [24]:

1. symmetry: SSIM(x,y) = SSIM(y,x)
2. boundedness: SSIM(x,y) ≤ 1
3. unique maximum: SSIM(x,y) = 1 if and only if x = y

In practice the SSIM index is calculated locally due to the fact that image statistics are non-stationary in spatial domain, and the distortion type and strength might change in different locations of the image. Another reason that makes local calculation of SSIM and other IQA measures interesting is the fact that at typical viewing distances, only a local area in the image can be seen with high resolution by the human observer, due to the foveation feature of the HVS [25,26]. SSIM has been calculated using a 8×8 sliding square window in [27] and [24], and an 11×11 circular-symmetric Gaussian weighting function with standard deviation of 1.5 samples and normalized to unit sum in [3], and results in both case are shown to be consistent with subjective evaluations of images. In most cases a single overall quality score is calculated by averaging the local SSIM index maps:

$$\text{MSSIM(X,Y)} = \frac{1}{M}\sum_{j=1}^{M} \text{SSIM}(\mathbf{x}_j, \mathbf{y}_j) \tag{2.21}$$

where M is the number of local windows of the image, X and Y are the reference and distorted images, and x_j and y_j are the image contents at the jth local window [3].

To incorporate image details at different resolutions, multi-scale framework for SSIM has been proposed in [28]. In the multi-scale scheme, low-pass filtering and downsampling by a factor of 2 is iteratively applied to the reference and distorted images, and the images are indexed from scale 1 to M accordingly. Figure 2.5 shows this process in details. At

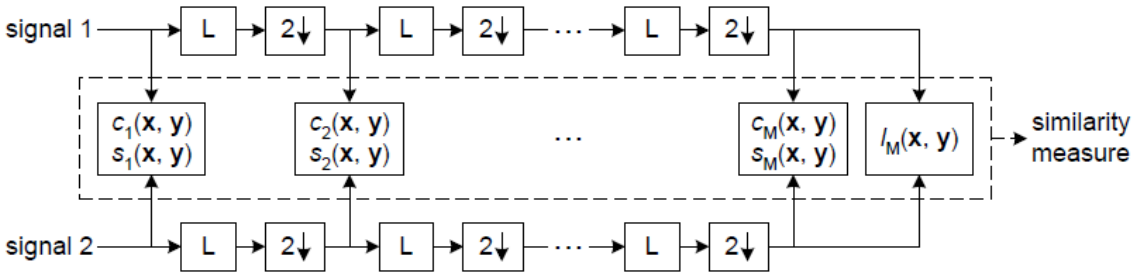


Figure 2.5: Multi-scale SSIM measurement framework. L: Low-pass filtering, 2 ↓: down-sampling by 2 [28].

each scale, the contrast comparison (2.16) and structure comparison (2.17) are calculated and shown by $c_j(x, y)$, $s_j(x, y)$. The luminance comparison is only done at the Mth scale and is denoted as $l_M(\mathbf{x}, \mathbf{y})$. The overall SSIM score is then calculated by:

$$\mathrm{SSIM}(\mathbf{X}, \mathbf{Y}) = [l_M(\mathbf{x}, \mathbf{y})]^{\alpha_M} . \prod_{j=1}^{M} [c_j(\mathbf{x}, \mathbf{y})]^{\beta_j} [s_j(\mathbf{x}, \mathbf{y})]^{\gamma_j}. \tag{2.22}$$

Similar to (2.19), the exponents α_M, β_j and γ_j determine the relative importance of luminance, contrast and structure components in different scales. For simplicity, it is assumed that $\alpha_j = \beta_j = \gamma_j$ and the exponent are normalized $\sum_{j=1}^{M} \gamma_j = 1$. The relative values of the parameters across different scales are then determined by subjective testing based on a table of images corrupted by twelve different types of distortions and decomposed at five scales. It is concluded that an optimum fine-to-coarse scale weight may be achieved by selecting the weights to be $\beta_1 = \gamma_1 = 0.0448$, $\beta_2 = \gamma_2 = 0.2856$, $\beta_3 = \gamma_3 = 0.3001$, $\beta_4 = \gamma_4 = 0.2363$ and $\alpha_5 = \beta_5 = \gamma_5 = 0.1333$, for a five scale decomposition respectively [28].

Inspired by the fact that local phase of natural images contains more structural information that magnitude, a Complex Wavelet form of SSIM index (CW-SSIM) is also proposed in [1]. The CW-SSIM has the advantage that it is more robust towards small geometric distortions, which are mainly considered as non-structural distortions, due to the fact that rigid translations of image structures leads to consistent phase shifts [1]. In complex wavelet domain, if $c_x = \{c_{x,i} | i = 1, 2, \cdots, N\}$ and $c_y = \{c_{y,i} | i = 1, 2, \cdots, N\}$ represent two sets of coefficients corresponding to a local patch of the reference image and

the distorted image respectively, then local CW-SSIM is defined as follows:

$$\begin{aligned}\tilde{S}(c_x, c_y) &= \tilde{m}(c_x, c_y).\tilde{p}(cx, cy) \\ &= \left(\frac{2\sum_{i=1}^{N}|c_{x,i}||c_{y,i}| + K}{\sum_{i=1}^{N}|c_{x,i}|^2 + \sum_{i=1}^{N}|c_{y,i}|^2 + K}\right) \cdot \left(\frac{2|\sum_{i=1}^{N} c_{x,i}c_{y,i}^{*}| + K}{2\sum_{i=1}^{N}|c_{x,i}c_{y,i}^{*}| + K}\right).\end{aligned} \tag{2.23}$$

The first component, $\tilde{m}(c_x, c_y)$, is determined by the magnitude of the coefficients and the maximum value is achieved if and only if for all i we have $|c_{x,i}| = |c_{y,i}|$, which implies the uniqueness. The second component $\tilde{p}(c_x, c_y)$ is determined by the consistency in the phase changes between c_x and c_y. When the phase difference between $c_{x,i}$ and $c_{y,i}$ is constant for all i, it achieves its maximum value, one. The functionality of this phase component is to capture image structural information since local image structure is maintained by the relative phase patterns of local image frequencies and a constant shift in the phase of all coefficients will not change the structure of the image [1, 5].

It has been shown that the CW-SSIM scheme is robust with respect to small scalings, rotations, luminance shifts, and contrast changes and provides low scores to images containing structural distortions [5], which is consistent with the overall SSIM index philosophy and is an improvement to the traditional SSIM schemes, which showed little robustness towards such geometric distortions.

2.2.2 Visual Information Fidelity

Visual Information Fidelity (VIF) proposed in [4], is an information theoretic approach towards the IQA problem. It is based on modeling the reference image as the output of a random process, called the natural source. It then assumes that the output of this random process is passed through a distortion channel which is intended to model the HVS. The information content of the reference image is then quantified as being the mutual information between the natural source and the output of the HVS channel. This is assumed to be a quantification of the amount of information that the brain could ideally extract from the reference image. The same measure is then quantified for the output signal of natural source, which is passed through a distortion channel before being processed by the HVS, and is similarly assumed to be the amount of information the brain could ideally extract from the distorted image. The two information measures are then combined to form a VIF score, which relates visual quality to relative image information and is shown to be consistent with subjective scores [4]. Figure 2.6 demonstrates these elements in a generic VIF framework.

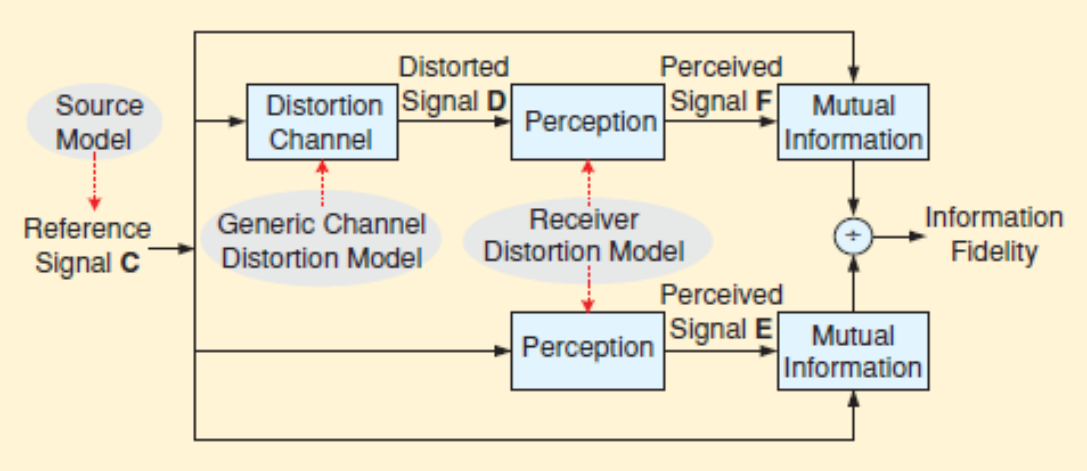


Figure 2.6: Generic Visual Information Fidelity framework [5].

Images and videos of three dimensional visual environment come from natural scenes, which represent a small subset of all possible images. The VIF approach uses the Gaussian Scale Mixture (GSM) model to account for the Natural Scene Statistics (NSS), which has been shown to be an effective model for representing non-Gaussian marginal distributions of the wavelet coefficients of natural images [29]. If c is a collection of M neighboring wavelet coefficients from a local patch in a wavelet subband, it is modeled as a mixture of random Gaussian vectors which have the same covariance structure $\mathbf{C_u}$, but have different weight scales. The mathematical representation of the vector c would be:

$$c = \sqrt{z}\mathbf{u}, \tag{2.24}$$

where $\mathbf{u}$ is a zero-mean Gaussian vector and $\sqrt{z}$ is an independent scalar random variable [29].

VIF also assumes a generic and simple image distortion model which is used to represent all types of distortions that may occur to the reference signal. The model describes distortions as local combination of a uniform wavelet domain energy attenuation and an independent additive noise:

$$d = gc + \nu, \tag{2.25}$$

where c and d are random vectors from a local patch in the reference and distorted images, g represents a deterministic gain factor used to model for distortions such as blur, and ν is an independent additive zero-mean white Gaussian noise with covariance $\mathbf{C}_\nu = \sigma_\nu^2\mathbf{I}$. The distortion model provides a simple, yet effective approximation for all types of distortions and has been shown to be very successful in VIF index over a wide range of distortion [4,30]. A stationary zero-mean additive white Gaussian noise process is used to model for the internal neural noise of the HVS, thereby adding to both the reference coefficients c and distorted coefficients d:

$$f = d + n \tag{2.26}$$

$$e = c + n, \tag{2.27}$$

where e and f denote the vectors of coefficients c and d in as perceived by the HVS and n is an independent white Gaussian noise with covariance matrix $\mathbf{C_n} = \sigma_n^2\mathbf{I}$ [4].

Based on the statistical model of the NSS, distortion channel, and human visual system channel, the mutual information quantities are calculated as follows:

$$\begin{aligned} I(c;e|z) &= \frac{1}{2}\log_2\frac{|zC_\mathbf{u} + \sigma_n^2\mathbf{I}|}{|\sigma_n^2\mathbf{I}|} \\ &= \frac{1}{2}\sum_{j=1}^{M}\log_2(1+\frac{z\lambda_j}{\sigma_n^2}); \end{aligned} \tag{2.28}$$

$$\begin{aligned} I(c;f|z) &= \frac{1}{2}\log_2\frac{|g^2zC_\mathbf{u} + (\sigma_\nu^2+\sigma_n^2)\mathbf{I}|}{|(\sigma_\nu^2+\sigma_n^2)|} \\ &= \frac{1}{2}\sum_{j=1}^{M}\log_2(1+\frac{g^2z\lambda_j}{\sigma_\nu^2+\sigma_n^2}). \end{aligned} \tag{2.29}$$

In simplifying these expressions, an orthogonal eigenvalue decomposition of the covariance matrix $C_\mathbf{u}$ has been employed, such that $\mathbf{C_u} = \mathbf{Q}\Lambda\mathbf{Q}^T$ and $\mathbf{\Lambda}$ is a diagonal matrix whose entries are the eigenvalues $\lambda_1, \lambda_2, \cdots, \lambda_M$. Local mutual information is computed at each patch of each subband of the image, by estimating z, g and σ_ν using maximum likelihood estimators. Based on the assumption that the local patches are independent, the overall mutual information is then computed by summing across these patches, and the VIF index is defined as:

$$\text{VIF} = \frac{I(C;F|z)}{I(C;E|z)} = \frac{\sum_{i=1}^{N} I(c_i;f_i|z_i)}{\sum_{i=1}^{N} I(c_i;e_i|z_i)} \tag{2.30}$$

Figure 2.7: Generic IQA framework with the pooling stage [32].

VIF has been validated using a wide variety of distortion types, and has been shown to be an effective measure of visual fidelity for both video and image quality measurement, which is superior to other state-of-the-art image quality/fidelity measures [4, 5, 31].

2.2.3 Information Content Weighting Approach

Perceptual image quality assessment algorithms commonly share a two stage structure which is based on local image quality/distortion assessment, and weighting and summing the local quality/distortion into a single index which is usually called pooling. The pooling stage is an important part of every IQA algorithm which has limited improvement in the recent years, and is usually carried out using ad-hoc algorithms. Figure 2.7 shows a generic two-stage framework of an IQA system with the pooling stage.

An optimal pooling strategy based on the information content of images is proposed in [32]. The method uses the total perceptual information content estimated from the reference image and the distorted image as a weight index for the quality score calculated on local patches of the distorted image, and may be used with most IQA algorithms. The framework for computing the information content is shown in figure 2.8, where R represents the reference image and E represents the reference image as seen by the HVS, and similarly D represents the distorted image, and F represents the distorted image as seen by the HVS. The sum of mutual information between R, and E, and the mutual information between D and F, minus the shared information between E and F is taken as an information content weight for a local patch of the image:

$$\omega = I(R;E) + I(D;F) - I(E;F) \tag{2.31}$$

Assuming that $R = sU$ is a K column vector of GSM random fields with U being a zero-mean Gaussian vector with covariance matrix $\mathbf{C_U}$, and a given scalar factor $S = s$, which represents a group of K neighboring transform coefficients, and $D = gR + V$, $E = R + N_1$,

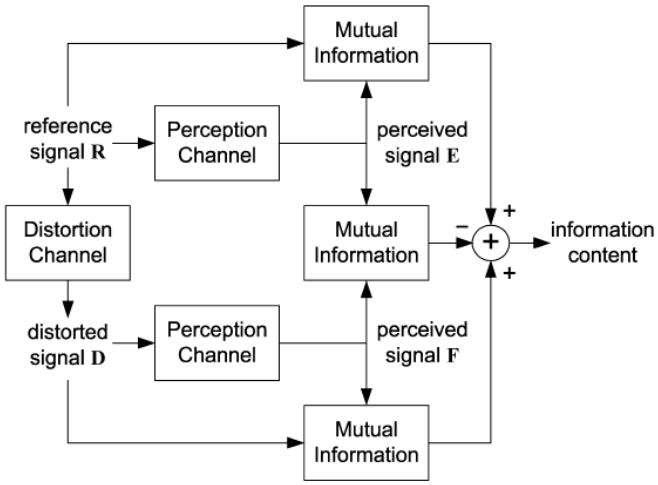


Figure 2.8: Diagram for computing information content [32].

and $F = D + N_2$, where g is a scaling factor used to model distortions and V, N_1 and N_2 are independent white Gaussian noise vectors, used to model the lumped uncertainty of the visual system and distortion respectively, the mutual information evaluations may be simplified based on the determinants of the covariances [32]:

$$I(R;E) = \frac{1}{2}\log_2\left[\frac{|\mathbf{C}_R||\mathbf{C}_E|}{|\mathbf{C}_{(R,E)}|}\right], \tag{2.32}$$

$$I(D;F) = \frac{1}{2}\log_2\left[\frac{|\mathbf{C}_D||\mathbf{C}_F|}{|\mathbf{C}_{(D,F)}|}\right], \tag{2.33}$$

$$I(E;F) = \frac{1}{2}\log_2\left[\frac{|\mathbf{C}_E||\mathbf{C}_F|}{|\mathbf{C}_{(E,F)}|}\right], \tag{2.34}$$

where $\mathbf{C}$ is the covariance matrix, and the information content weight simplifies into the following expression:

$$\omega = \frac{1}{2}\log_2\left[\frac{|\mathbf{C}_{(E,F)}|}{\sigma_n^{4K}}\right], \tag{2.35}$$

and $|\mathbf{C}_{(E,F)}|$ is:

$$|\mathbf{C}_{(E,F)}| = |[(\sigma_\nu^2 + \sigma_n^2)s^2 + \sigma_n^2 g^2 s^2]\mathbf{C}_\mathbf{U} + \sigma_n^2(\sigma_\nu^2 + \sigma_n^2)\mathbf{I}|. \tag{2.36}$$

(4.25) can be further simplified using an eigenvalue decomposition of the covariance matrix $\mathbf{C_U} = \mathbf{Q}\Lambda\mathbf{Q}^T$, where $\mathbf{Q}$ is an orthogonal matrix and λ is a diagonal matrix with entries $\lambda_1, \cdots, \lambda_K$:

$$|\mathbf{C}_{(E,F)}| = \prod_{k=1}^{K}\{[\sigma_\nu^2 + (1+g^2)\sigma_n^2]s^2\lambda_k + \sigma_n^2(\sigma_\nu^2 + \sigma_n^2)\}. \tag{2.37}$$

By plugging (2.37) into (2.35), we have:

$$\omega = \frac{1}{2}\sum_{k=1}^{K}\log_2\left\{1 + \frac{\sigma_\nu^2}{\sigma_n^2} + \left(\frac{\sigma_\nu^2}{\sigma_n^4} + \frac{1+g^2}{\sigma_n^2}\right)s^2\lambda_k\right\}. \tag{2.38}$$

The weight ω is completely based on evaluation of local information content, and the covariance matrix $\mathbf{C_U}$, and other parameters involved in (2.38) like s and g are estimated using a sliding window that runs across each subband [32].

The proposed weighting scheme can be incorporated into IQA algorithms with small modifications and significantly improve the results [32]. As the simplest signal fidelity measure, results of MSE are significantly improved by including the information weight content in the algorithm. If $x_{i,j}$ and $y_{i,j}$ represent the $i-th$ transform coefficients at the $j-th$ scale, and $w_{i,j}$ is the information content weight computed at the corresponding location, IW-MSE is defined as follows:

$$\text{IW-MSE} = \prod_{j=1}^{M}\left[\frac{\sum_i \omega_{i,j}(x_{j,i} - y_{j,i})^2}{\sum_i \omega_{j,i}}\right]^{\beta_j}, \tag{2.39}$$

where M is the number of scales and β_j is the weight given to the $j-th$ scale in the fine-to-coarse weighting of subbands as proposed in [28]. Similarly IW-PSNR is defined as follows:

$$\text{IW-PSNR} = 10\log_{10}\left(\frac{L^2}{\text{IW-MSE}}\right) \tag{2.40}$$

The information content weighting scheme has also been applied to the multi-scale SSIM with consistent improvements in results [32]. If $w_{j,i}$ is the information content weight computed at the $i-th$ spatial location at the $j-th$ scale, the $j-th$ scale IW-SSIM is defined as follows:

$$\text{IW-SSIM}_j = \frac{\sum_i \omega_{j,i}c(x_{j,i}, y_{j,i})s(x_{j,i}, y_{j,i})}{\sum_i \omega_{j,i}}, \tag{2.41}$$

for $j = 1, \cdots, M-1$, and

$$\text{IW-SSIM}_j = \frac{1}{N_j} \sum_i l(x_{j,i}, y_{j,i}) c(x_{j,i}, y_{j,i}) s(x_{j,i}, y_{j,i}) \tag{2.42}$$

for $j = M$. The final overall IW-SSIM measure is then computed using the fine-to-coarse scale weights introduced in [28] as follows:

$$\text{IW-SSIM} = \prod_{j=1}^{M} (\text{IW-SSIM}_j)^{\beta_j}, \tag{2.43}$$

where $\{\beta_1, \cdots, \beta_M\}$ are selected according to [28].

The information content weighting scheme has been validated using six publicly available subject-rated image databases and the results of IW-PSNR and IW-SSIM are compared to 13 other algorithms, all of which show significant improvement to the results [32].

2.2.4 Image Quality Databases

Human subjects can provide the best judgement for quality assessment of images. However, subjective quality assessment of images is very cumbersome and often requires expensive laboratory experiments which are time consuming and cannot be used to predict quality of images in real time applications. In order to overcome these issues, the goal of objective IQA algorithms is to predict the quality of images in agreement with subjective opinion of human observers. Therefore to calibrate QA algorithms and validate their performance, subject-ranked image databases are required. These databases can be used to train the objective QA methods, or compare the QA methods' performance with that of an average human subject. There are currently six widely-cited subjective image quality assessment studies which are available to the public for free, all of which validate the performance of our proposed algorithm in the following chapters.

The Laboratory for Image and Video Engineering (LIVE) database consists of 29 original reference images contaminated by five types of distortions at different distortion levels. The distortion types include JPEG compression, JPEG2000 compression, white noise, Gaussian blur and fast fading channel distortion of JPEG2000 compressed bitstream. A total of 982 subject-rated images are created from these distortions and the subjective scores of all images are adjusted according to an alignment process in which a cross-comparison of mixed images from all distortion types was done [32, 33].

The Tampere Image Database 2008 (TID2008) database consists of 25 original reference images contaminated by 17 distortion types at 4 different distortion levels. A total of 1700 distorted images are generated and rated by subjects. The distortion types include additive Gaussian noise, additive noise where the noise in color components is more intensive than the noise in luminance components, spatially correlated noise, masked noise, high frequency noise, impulse noise, quantization noise, Gaussian blur, image denoising, JPEG compression, JPEG2000 compression, JPEG transmission errors, JPEG2000 transmission errors, non eccentricity pattern noise, local block-wise distortions of different intensity, mean shift and contrast change [32, 34].

The Categorical Image Quality (CSIQ) database consists of 866 distorted images created from 30 original reference images using six types of distortions at four to five distortion levels. CSIQ images are subjectively rated base on a linear displacement of the images across four calibrated LCD monitors placed side by side with equal viewing distance to the observer. The database contains 5000 subjective ratings from 35 different observers, and ratings are reported in the form of DMOS. The distortion types include JPEG compression, JPEG2000 compression, global contrast decrements, additive pink Gaussian noise, and Gaussian blurring [32, 35].

The IVC database consists of 185 distorted images created from 10 original reference images using four types of distortions including JPEG compression, JPEG2000 compression, Local adaptive resolution (LAR) coding, and Blurring. The database was created by Ecole Polytechnique de l'Universite de Nantes [36, 37].

The Toyama-MICT database consists of 196 images, including 168 distorted images generated by JPEG and JPEG2000 compression at Toyama University [38].

The Cornell-A57 database consists of 54 distorted images with 6 types of distortions including quantization with selective step size, additive Gaussian white noise, JPEG compression, JPEG2000 compression with dynamic, contrast-based quantization algorithm, which applies greater quantization to the fine spatial scales relative to the coarse scales in an attempt to preserve global precedence, and blurring using a Gaussian filter [39].

Chapter 3

Generic Image Similarity Metric based on Kolmogorov Complexity

3.1 Normalized Conditional Compression Distance

When NCD was used to quantify image similarities, it did not achieve the same level of success as in other application fields. For example, it was reported in [14] that NCD works well when parts are added or subtracted from an image, but struggles when image variations involve form, material and structure. We believe that this is mainly due to the poor approximation of $K(xy)$ using $C(xy)$, which is often implemented by applying a regular image compressor to the concatenation of two images. For example, when an image is a ninety-degree rotated copy of another, concatenating two images would not facilitate any efficient compression. To avoid this problem, we propose to approximate the conditional Kolmogorov complexity in (2.9) directly by designing a conditional image compressor denoted by C_T, so that

$$K(y|x) \approx C_T(y|x) \quad \text{and} \quad K(x|y) \approx C_T(x|y)\,. \tag{3.1}$$

This leads to a normalized conditional compression distance (NCCD) measure given by

$$\text{NCCD}(x,y) = \frac{\max\{C_T(x|y), C_T(y|x)\}}{\max\{C(x), C(y)\}}\,. \tag{3.2}$$

The critical issue is how to define the conditional compressor C_T. Here we propose

a practical solution by making use of a set of transformations that convert one image to another. Let $\{T_i | i = 1, \cdots, N\}$ be the set of transformations, let $T_i(x)$ represent the transformed image when applying the i-th transform to image x, and let $p(T_i, x)$ denote the parameters used in the transformation. Each type of transformation is also associated with a parameter compressor, and C_i^p denotes the parameter compressor of the i-th transformation. We can then define our conditional compressor as

$$C_T(y|x) = \min_i \{C[y - T_i(x)] + C_i^p[p(T_i, x)] + \log_2(N)\}, \tag{3.3}$$

where C remains to be a practical image compressor which encodes the difference between y and the transformed image $T_i(x)$, and the $\log_2(N)$ term computes the number of bits required to encode the selection of one out of N potential transformations.

The idea of finding the simplest transformation between two images is sensible from the viewpoint of human visual perception, for which it has long been hypothesis that the biological visual system is an efficient coder of the visual world [12]. For example, given two images that are rotated copy of each other, our visual system would not interpret the difference between them by directly differencing their intensity values (which requires a large number of bits to encode the residual), but by estimating the amount of rotation (which can be coded very efficiently).

3.2 Implementation based on Image Compression

An advantage of NCCD (as opposed to NCD) is that it provides a more flexible framework so that different types of transformations can be included. The list of transformations can also be incremental, in the sense that new transformations, when available, can be easily added into the existing list, and expanding the list always improves the approximation of NCCD to NID. Of course, exhausting all possible transformations is practically impossible. However, by going through a handful of transformations, it may be sufficient to appropriately cover most image distortions encountered in real-world applications.

Our current implementation of NCCD are as follows. First, we adopt the content adaptive lossless image compression algorithm (CALIC) [40] as the base image compressor, which achieves superior performance when compared with state-of-the-art lossless image compression algorithms. CALIC is employed in computing the denominator of Eq. (3.2) as well as the first term in Eq. (3.3). Since $y - T_i(x)$ in Eq. (3.3) can generate negative values and CALIC applies to grayscale images with positive intensity values only, the mean

NCCD = 0.8121, SSIM = 0.4649, MSE = 73.4

NCCD = 0.2189, SSIM = 0.7059, MSE = 49.8

NCCD = 0.3901, SSIM=0.8425, MSE = 78.4

NCCD =0.0091, SSIM = 0.9094, MSE = 105.5

Figure 3.1: Comparison of MSE, SSIM and NCCD measures using images distorted by JPEG compression, blur, JPEG2000 compression and contrast reduction.

intensity value of $y - T_i(x)$ is shifted to mid-gray level before the application of CALIC. Second, the types of transformations involved in the computation of C_T include

- *Global contrast and luminance change.* This is computed by a pointwise intensity transformation defined as $s = \alpha(r - \bar{r}) + \bar{r} + \beta$, where r and s are the intensity values before and after the transformation, respectively, $\bar{r}$ is the average value of r, and α and β are the parameters that determine the degrees of contrast and mean luminance changes, respectively. In a special case when $\alpha = 1$ and $\beta = 0$, it reduces to an identity transform, i.e., $T(x) = x$.

- *Global Fourier power spectrum scaling.* This transformation attempts to match two images by scaling the power spectrum of one image in the Fourier transform domain. Let $X(\omega)$ and $Y(\omega)$ be the Fourier transforms of x and y, respectively. We first find the best linear transform parameters p_1 and p_2, such that $\| |Y(\omega)| - (p_1|X(\omega)| + p_2) \|^2$ is minimized. We then define the transform $T(x)$ as the inverse Fourier transform of $p_1 X(\omega) + p_2$.

- *Global affine transform.* This transformation tries to matching one image by applying a global affine transform to another. The transformation can be encoded using six parameters and covers a variety of image changes including translation, scaling (zooming in or zooming out), rotation, and shearing.

- *Local registration transformation.* This is implemented by aligning two images using the coherent point drift registration algorithm [41] that allows for both rigid and affine non-rigid transformations, or by aligning two images using the local affine and global smooth registration [42].

Given a pair of images x and y for comparison, we attempt all the above transformations from both x to y and y to x (multiple transformations are also allowed). This is important because the values of $K(x|y)$ and $K(y|x)$ can be drastically different (and so do the values of $C_T(x|y)$ and $C_T(y|x)$). For example, converting the "Lena" image x to a blank image y is easy (as y can be created by a very short program), but the opposite is not.

3.2.1 Experiment

The goal of our preliminary experimental work is to test the applicability of the proposed NCCD implementation for various distortion types and compare it with existing measures such as the mean squared error (MSE) and SSIM. Figure 3.1 shows four original images

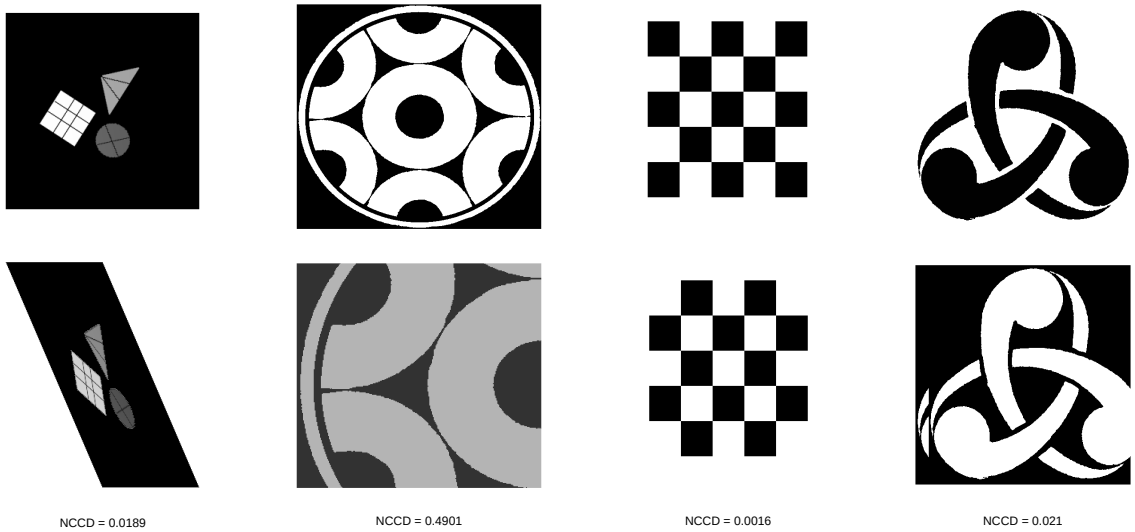


Figure 3.2: Tests using images with geometric and compound distortions.

distorted with JPEG compression, blur, JPEG2000 compression and contrast reduction. MSE appears to be a poor measure in this test, because the best quality image (contrast reduced) that does not exhibit any structural distortion, results in the worst MSE value. Both SSIM and NCCD works reasonably well, which give the best quality values (highest SSIM and lowest NCCD) to the contrast reduced image.

In Fig. 3.2, the test images underwent certain geometric distortions, for which the MSE and SSIM measures do not apply or cannot work appropriately. By contrast, NCCD produces meaningful similarity evaluations. In particular, it works when the two images are of different size and shape; it works if parts of an image are missing; it also works when images are shifted and/or the black/white pixels are reverted. All of these demonstrate the wider applicability of NCCD.

3.2.2 Divisive Normalization Transform

Due to the fact that the notion of Kolmogorov complexity treats all the bits in a program equally, existing NID based approaches do not take into account the degree of perceptual relevance of the information contained in the image. The HVS on the other hand is adapted to match statistical properties of natural stimuli [12]. It is hypothesized that the early sensory systems remove redundancy in the stimuli which results in a set of statistically independent neural responses [43]. In this sense, it is necessary for the HVS to filter

visual stimuli and it cannot treat the information in the stimuli signal equally. In order to compensate for this problem, a theoretical decomposition of NID to perceptually relevant information and residue has been proposed in the past [44]. The model is based on deriving a vector of relevant information which can be used to reconstruct a lossy representation of the image data. However, a practical framework for this decomposition is missing, leaving the feature extraction and selection process to ad-hoc, application specific algorithms.

A practically more useful technique in removing redundancies in the stimuli signal is by using efficient coding transforms. The advantage of using such models is having a theoretical framework which ensures reducing perceptual and statistical dependence of stimuli and represents perceptually relevant information of the signal [43]. Among many nonlinear efficient coding transforms, Divisive Normalization Transform (DNT) has been extensively studied in the past. It has been observed that this simple normalization of elements by a weighted Minkowski combination of its neighboring elements can significantly reduce the dependencies among elements of natural image [45].

Biological sensory systems are believed to have evolved to match the statistical properties of natural stimuli, and efficient coding principle provides a powerful explanation for such an evolutionary optimization by asserting that sensory systems represent information content of the stimuli subject to their inherent limitations [43, 46]. In order to model this mechanism in the mammalian visual cortex, a set of sensory transforms have been proposed in the past, all of which attempt to reduce statistical dependencies of stimuli signals for more efficient signal transmission, representation, and processing [43]. The most simple form of these transforms which was originally proposed to model non-linearities in neurons of visual cortex [47] is Divisive Normalization (DN). Divisive Normalization is often modeled as dividing the value of an image element divided by a weighted average of the amplitudes of adjacent elements in spatial domain which resembles the neural response model in the HVS. Previous studies have shown that DN can reduce statistical dependencies among sensory signals [43, 48], act as a maximum likelihood estimator in noiseless data estimation [49], and play an important role in general decision making mechanism in context-dependent scenarios [50]. Details on how DNT models are applied in our approach will be described later.

3.3 Implementation based on Image Compression

In finding the shortest program which converts one image to another, all combinations of transformations in the list must be accounted for. Since our list includes four transfor-

mations, a total of sixteen combinations are tested. Transformations are applied in the following order:

First global affine transform is used to globally align the two image. In order to achieve this goal, a six-parameter affine matrix is found such that the SSIM value between the transformed source image $T_0(X)$ and the target image Y is maximized. In finding this transform, Matlab's Genetic Algorithm Toolbox is used to find the global optimum to the objective function. Algorithm 1 describes this procedure. The optimum matrix is then applied to the source image and its length is added to the required transform parameters in Eq. 3.3.

Once the images are globally aligned, a locally affine, globally smooth transform is applied to the source image. Assuming f_a and f_b are local regions of the luminance channel of the source and target images we have [42]:

$$cf_a(x,y) + b = f_b(m_1x + m_2y + t_x, m_3x + m_4y + t_y) \tag{3.4}$$

where m_i terms are affine parameters, and c and b are contrast and luminance change parameters. Using this registration, the length of a two- dimensional vector field of local geometric transformations must be included in transform parameters:

$$\overrightarrow{v}(x,y) = \begin{pmatrix} m_1x + m_2y + t_x - x \\ m_3x + m_4y + t_y - y \end{pmatrix}. \tag{3.5}$$

Global contrast and luminance change transform can be modeled as a linear regression problem, where α and β are selected to minimize mean squared error $\|Y - T_0(x)\|^2$. Assuming that the regression problem is formalized by $T_3(x) = L\Gamma$, we have $\Gamma = (L^TL)^{-1}L^TY$, where:

$$\Gamma = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}, L = \begin{bmatrix} (r_1 - \bar{r}) & 1 + \frac{\bar{r}}{\beta} \\ \vdots & \vdots \\ (r_N - \bar{r}) & 1 + \frac{\bar{r}}{\beta} \end{bmatrix}, Y = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}.$$

Global Fourier power spectrum scaling is modeled in a similar way. Assuming $T_4(x) = F^{-1}\{p_1X(\omega) + p_2\}$, and the objective is to minimize $\||Y(\omega)| - (p_1|X(\omega)| + p_2)\|^2$, we have $P = (X_\omega^TX_\omega)^{-1}X_\omega^TY_\omega$, where:

$$P = \begin{bmatrix} p_1 \\ p_2 \end{bmatrix}, Xw = \begin{bmatrix} |x_{\omega_1}| & 1 \\ \vdots & \vdots \\ |x_{\omega_N}| & 1 \end{bmatrix}, Y_w = \begin{bmatrix} |y_{\omega_1}| \\ \vdots \\ |y_{\omega_N}| \end{bmatrix},$$

and $|x_{\omega_i}|$ and $|y_{\omega_i}|$ are magnitude of fourier coefficients of source and target image respectively.

Inspired by the redundancy reduction properties of divisive normalization, we use an energy based form of DNT to create a perceptually uniform conditional image in the spatial domain. Assuming that the target image X is given in the approximation of an upper bound for conditional Kolmogorov complexity of the source image Y, we propose normalizing the conditional image by the following locally adaptive factor:

$$T_0\sqrt{1+\frac{\sigma_x^2}{C_0}} \tag{3.6}$$

where T_0 and C_0 are constants and σ_x^2 is the local energy of the given image computed using an 11×11 sliding Gaussian window with variance of 1.5. This normalization follows the same logic as DNT, and helps reduce the perceptual redundancies in the conditional image. Since the image X is presumed to be available to the decoding machine, the size of this divisive normalization map is not required to be encoded in the approximation of conditional Kolmogorov complexity. Furthermore, the proposed transform can be considered as a lossy compression scheme, and with the proper choice for C_0 and T_0, the resulting normalized image can be used to reconstruct the original image with high structural similarity with the original image.

In order to approximate the conditional Kolmogorov complexity $K(Y|X)$, all sixteen combinations of the four transformations in the list are tested. For each combination of the transformations, a transformed image $T_i(Y)$ is created. The target image X is then subtracted from the transformed source image $T_i(Y)$, resulting in a difference image with a dynamic range of $[-127:127]$ which contains the conditional information required to losslessly recover image Y if image X is available to the decoder. The difference image is then divided by the proposed map in Eq. 3.6 to remove the redundancies among neighboring pixels and transform it into a perceptually uniform space. Finally the result is shifted by a constant in gray level and quantized into integer numbers. Figure 3.3, and algorithm 2 show the step by step process of deriving the final image and approximation of conditional complexity. The new image which we call "Uniform image", is a perceptually compressed form of the original image Y on the condition that image X is available to the decoder.

The quality of the reconstructed image depends on the practical choice of parameters T_0 and C_0. Figure 3.4 shows reconstruction examples of the sample image Y in Figure 3.3 for $T_0 = 2$. It is evident that the choice of the parameter C_0 can greatly affect the visual quality of the reconstructed image, and as C_0 increases the SSIM between the reconstructed image and the original image increases. The optimum values for these parameters are tuned

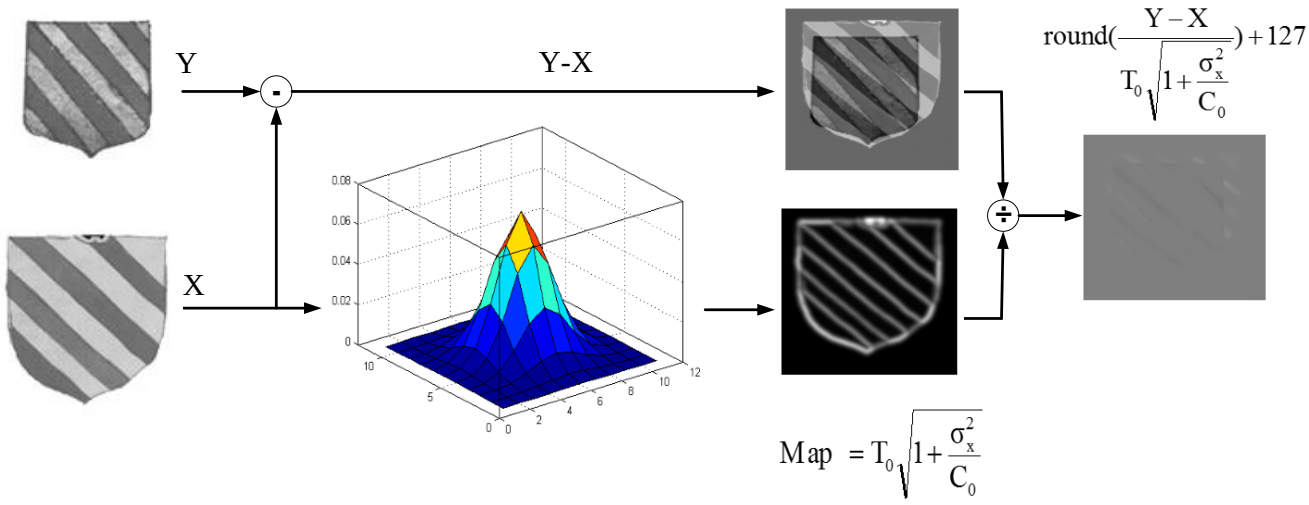


Figure 3.3: Deriving the uniform image

using small datasets. In this work, we select $T_0 = 2$ and $C_0 = 0.1$ in our simulations.

Algorithm 1: Initial Registration of Source (Y) to Target (X) image

Data: Source (Y) and Target (X) Images
Result: Registered Source Image $T(Y)$
Step 1: Apply initial affine transform i = 0 to find $T_0(Y)$;
Step 2: Calculate Objective function = $1 - \text{SSIM}(T_i(Y), X)$;
Step 3: Find affine matrix (M) which minimizes Objective function of Step 2 using Genetic Algorithm;

Algorithm 2: Approximation of $K(T(Y)|X)$ using perceptually normalized difference image

Data: Registered Source $(T(Y))$ and Target (X) Images
Result: Approximation of $K(T(Y)|X)$
Step 1: Difference Image = $T(Y)$ - X ;
Step 2: Perceptually Normalized Difference Image = $\frac{\text{Difference Image}}{T_0\sqrt{\frac{\sigma_x^2}{C_0}+1}} + 127$;
Step 3: $K(T(Y)|X) \leq$ Size of Compressed Perceptually Normalized Difference Image;

Figure 3.4: Reconstruction of the encoded source image for various C_0 values

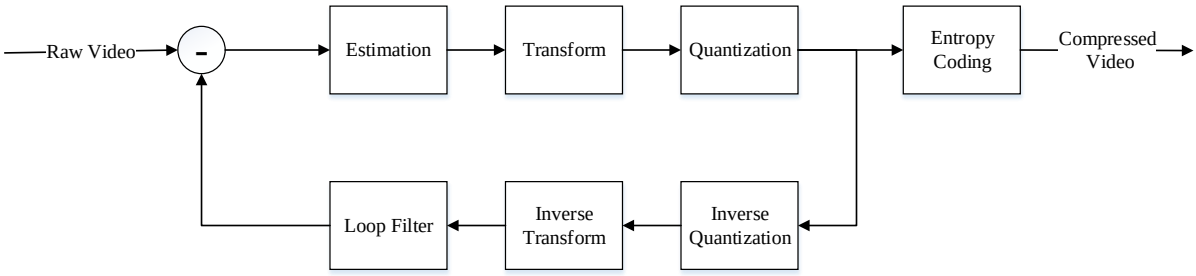


Figure 3.5: H.264 Coding Block Diagram [51]

3.4 Implementation based on Video Compression

Another possible implementation of NCCD is based on video compression. In this scheme, the reference and test images are treated as subsequent frames of a video stream, and a video encoder compresses the two frames by removing the spatial and temporal redundancies among them, and therefore effectively expressing one image based on the other.

Currently the most widely used standard in video coding industry is H.264. Compared with previous standards, H.264 achieves 50% improvement in compression, and substantially improves perceptual quality [51,52]. Figure 3.5 shows the block diagram of the H.264 video encoder. The standard comprises several processing stages, which are fundamental to a conditional compression scheme such as NCCD. Assuming that the input uncompressed video is a two frame video composed of the reference and the test image, the estimation stage attempts to identify and remove spatial redundancies in the reference image and the temporal redundancies among the reference image and the test image in the subsequent frame. This is done by dividing the image into macroblocks of size 16×16, and itera-

tively finding a reference macroblock which is most similar to the one being processed. The estimation is carried out on intra-frame scale to remove the spatial redundancies of the reference image, and inter-frame scale to remove the temporal redundancies between the reference image and the test image. The difference between the test image and its prediction from the reference image is called residual error, and is transformed from the spatial domain to frequency domain using an integer DCT transform. The residual coefficients along with motion vectors and other parameters required to reconstruct the image in the decoder are then sent to an entropy encoder, which ensures that shorter length codes are assigned to more frequent symbols [53]. In effect, this framework finds the most efficient description of the test image given the reference image and acts as a conditional compressor. The encoder also allows to choose between lossless compression and lossy compression with user controlled quantization parameters, which can be used to remove native psychovisual redundancies of the images. In the following sections, we use the H.264 based implementation of NCCD on texture classification and face recognition and compare the results with those of Image compression based implementation and other compression based distances.

3.5 Applications

3.5.1 Digit Recognition

The effectiveness of the proposed framework is demonstrated by a pattern matching test used in [54], and comparing the results to those of MSE, SSIM and CW-SSIM. The dataset used in the test is created from ten standard digit templates of size 32×32, as shown in Figure 3.6. A total of 2430 distorted images (243 for each digit) is then created by shifting, scaling, rotating, and blurring the standard templates. The images are then recognized by comparing each distorted image with the ten standard templates. Table 3.1 shows the recognition rate for image compression based NCCD as compared to MSE, SSIM, and CW-SSIM. It can be observed that the recognition rate for Image compression based NCCD is higher than those of MSE and SSIM. This is due to the fact that both MSE and SSIM are very sensitive to small translations and geometric distortions, while NCCD compensates for that in the preprocessing stage. On the other hand CW-SSIM slightly outperforms NCCD because it is insensitive to small scaling and rotation of images, and it measures structural similarity among images [54].

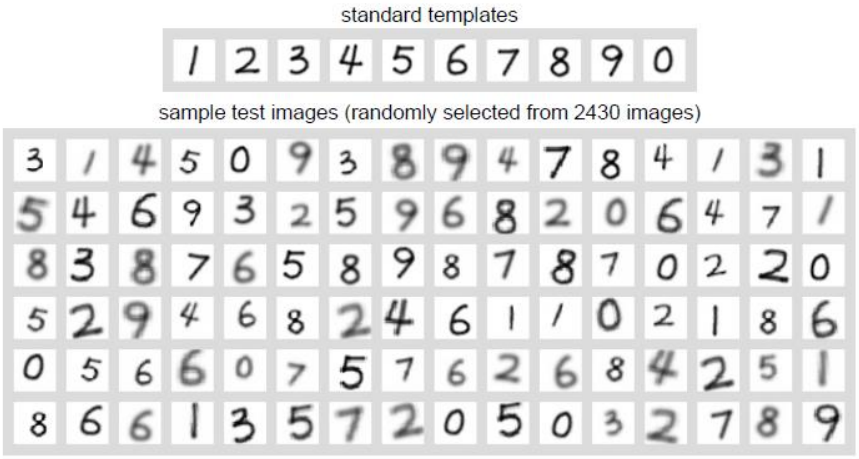


Figure 3.6: Pattern matching. 2430 images are matched to the ten standard templates using MSE, SSIM, CW-SSIM, and NCCD, and each image is recognized as belonging to the category corresponding to the best similarity score. [54]

Table 3.1: Correct recognition rate [54]

Digit	NCCD (%)	MSE (%)	SSIM (%)	CW-SSIM (%)
1	100	84.0	76.1	100
2	96.3	65.4	45.3	98.4
3	90.9	49.4	47.7	97.1
4	93.5	63.8	41.6	100
5	91.2	47.7	18.5	96.3
6	96.6	56.4	42.0	97.9
7	92.9	68.3	60.9	94.2
8	97.4	49.8	39.1	99.6
9	91.9	59.3	51.4	100
0	90.2	51.4	46.5	93.0
All	94.1	59.6	46.9	97.7

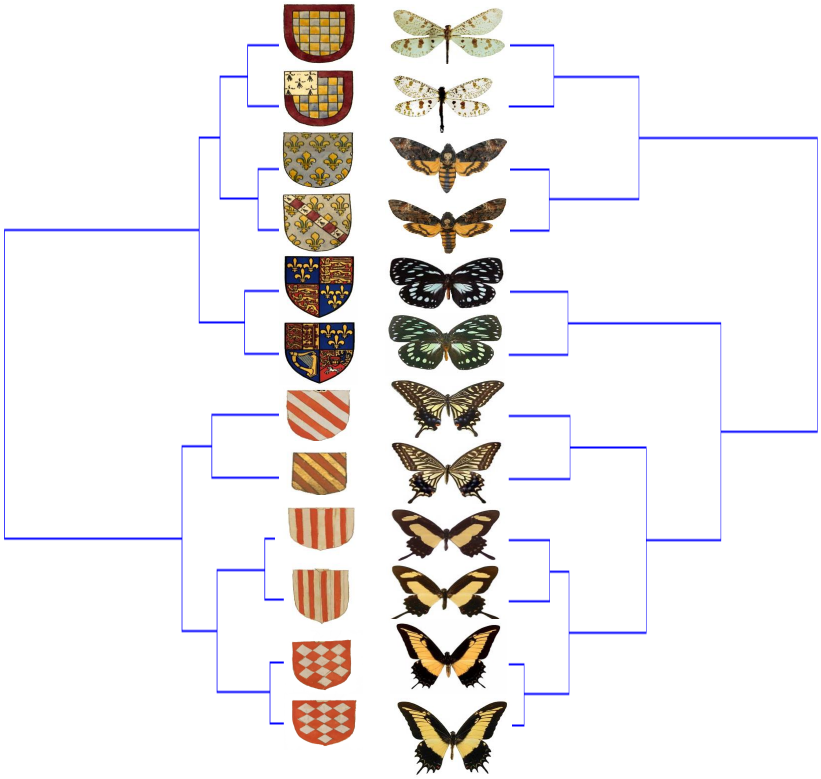


Figure 3.7: Classification of Heraldic Shields and Butterflies datasets by Image compression based NCCD framework

3.5.2 Texture Classification

In order to demonstrate the performance of the proposed framework in texture classification, we apply NCCD to a variety of texture datasets commonly used in the literature and compare our results to two compression based distance methods. In the first step, we classify two small dataset of images of various size and color, which are easy to verify by observation. Figure 3.7 shows the result of clustering the *Heraldic Shields* and the *Butterflies* [55] dataset. Both datasets contain 12 images of different sizes, which are hierarchically clustered according to their similarity in texture using the linkage method. It can be observed that the clustering is consistent with human intuition in both scenarios.

In order to compare the performance of NCCD with other compression based distance methods, we perform leave one out classification experiments on a number of datasets used in [55]. Classification reuslts on these datasets are then compared to those of CK-1 distance

[55], and a sparsity based compression distance proposed in [56]. A brief introduction of these datasets is presented in the following.

Tire Treads: This dataset is a collection of tire tracks, and contains 48 images of 3 tires rolled in 16 different directions [55].

Brodatz Texture: This dataset is a collection of $1,792$ man-made and natural texture images digitalized from a reference photographic album for designers [57].

Camouflage: This dataset is a collection of 80 random orientations of 9 modern US military camouflage [55].

VTT Wood: This dataset is a collection of 200 images of wood which are classified into two subset of healthy and defective woods with 40 types of wood defects [58].

VisTex: This dataset is a collection of homogeneous texture images and texture scenes created by MIT Vision and Texture Group, which do not conform to rigid frontal plane perspectives and studio lighting conditions [59].

In leave-one-out cross validation scheme, a query image is selected from the dataset and all images in the dataset are ranked based on their distance to the query image in ascending order. The first K images are used to develop a hypothesis about the type of the query image. The hypothesis is then checked, and the performance of the classification method is defined as the ratio of correctly classified images to the total number of the images in the dataset. Figure 3.8 shows examples of retrieval of images from Brodatz dataset. In each case the first image in the results is the same as the query image, and the following images are closest images in distance to the query image. Figure 3.9 shows examples of retrieval results of images from Camoflauge dataset. It is notable that although the dataset is in color, the NCCD is capable of finding underlying similarities in both texture and color of the images, and slight variations in color do not cause a significant distance among images belonging to the same class. Results for classification of the above datasets using NCCD and CK-1 as well as Sparsity Based Compression (SBC) distance method proposed in [56], are provided in table 3.2. In this table we use leave-one-out scheme and 1-Nearest Neighbor framework in order to have comparable results to those provided in [56]. It can be observed that the performance of NCCD is comparable to, or better than state-of-the-art compression based classification methods.

The video compression based implementation of NCCD requires approximately 5 seconds to find the distance among a pair of 128×128 images on a core-i3 Intel processor. The Image compression based implementation of NCCD requires 12 seconds to find the distance between the same pair, which is longer due to the computationally expensive preprocessing stage. Both implementations of NCCD in this thesis are relatively more

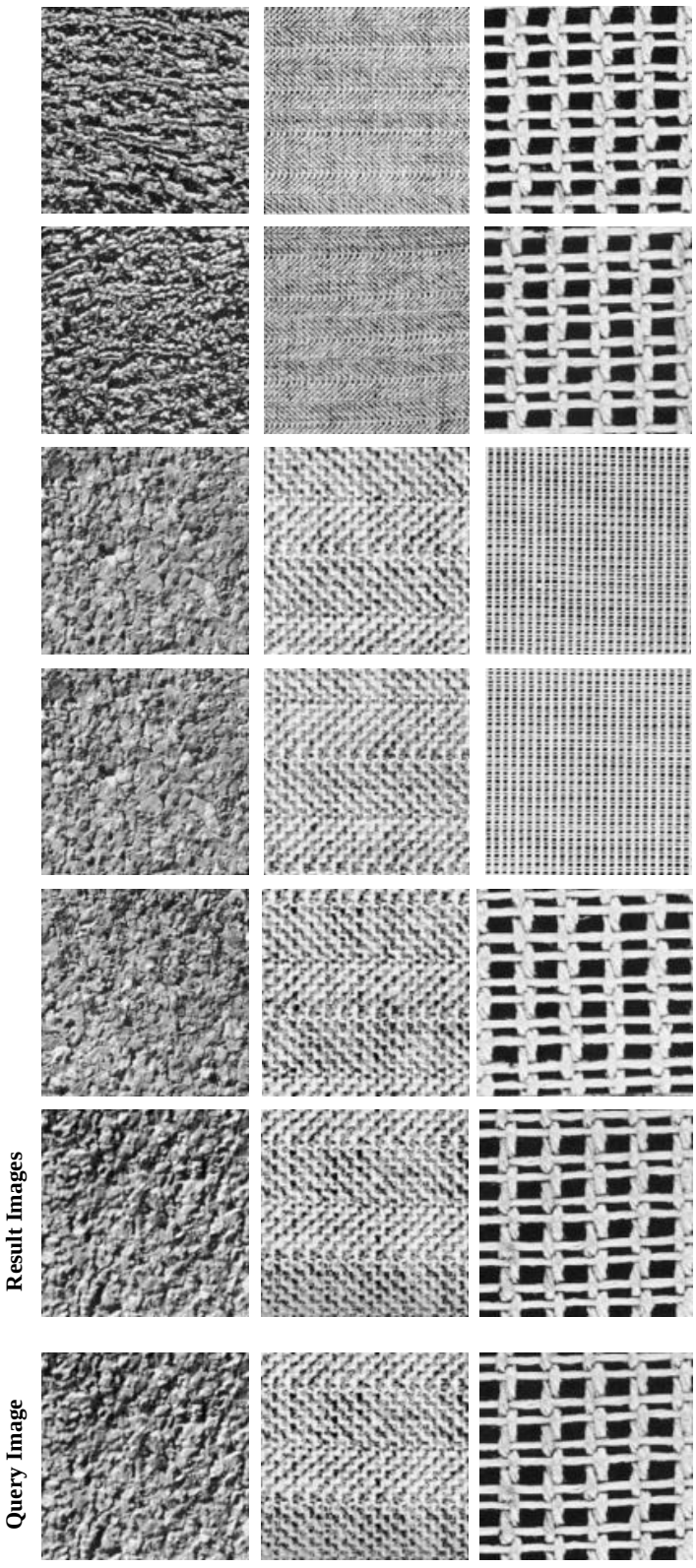


Figure 3.8: Image retrieval in Bordatz [57] Dataset by Image compression based NCCD

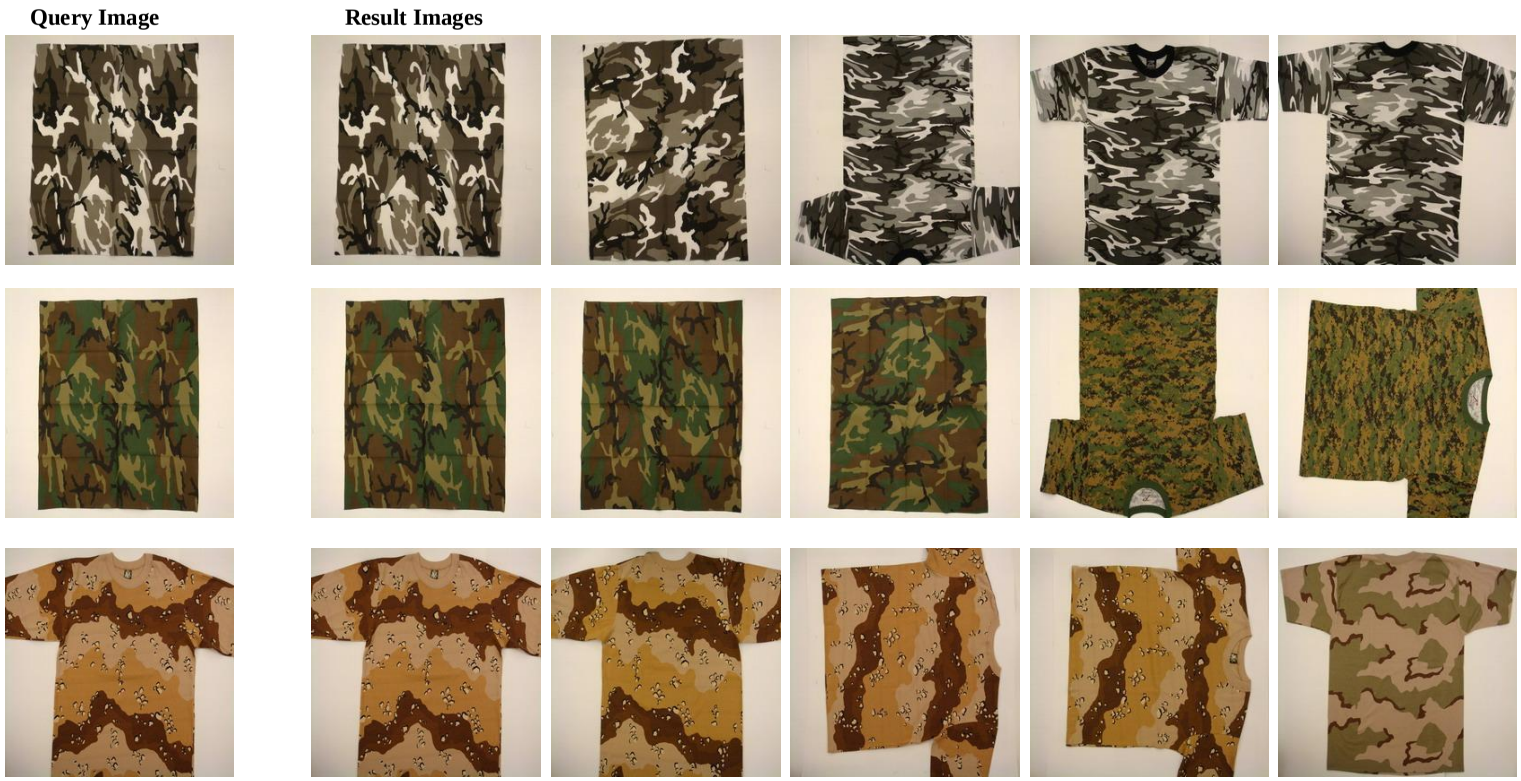


Figure 3.9: Image retrieval in Camoflauge [55] Dataset by Image compression based NCCD

Table 3.2: Classification performance of various datasets using NCCD, CK-1 [55], and SBC [56]

Dataset	NCCD (%)	H.264 NCCD (%)	CK-1 (%) [55]	SBC(%) [56]
Brodatz [57]	73.2	70.1	54.0	**76.2**
Camouflage [55]	81.9	73.0	**87.5**	87.0
Tire tracks [55]	**93.8**	89.5	79.2	79.2
VTT Wood [58]	**91.9**	88.0	80.5	85.2
VisTex [59]	39.3	**42.0**	32.9	N/A

computationally expensive than rival methods such as CK-I, and SBC which use MPEG-I video encoder and dictionary based sparse coding as measures for complexity. However, in general the proposed implementations have wider applicability in scenarios such as global and local geometric distortions, where the rival methods cannot be applied. For practical applications, the implementation could be sped up by the use of Graphical Processing Unit (GPU) programming techniques, which can significantly reduce the processing time.

3.5.3 Face Recognition

In this section, we test the two implementations of NCCD for face recognition on two of the widely cited databases in the literature. Face recognition is essential in many applications such as law enforcement, surveillance and security, and image retrieval. Many algorithms and databases have been developed which report success in recognizing faces under different illumination, pose and occlusion conditions.

In order to demonstrate performance of NCCD in face recognition applications, we apply the distance to AT&T [60] and Yale [61] face datasets, and compare our results to those of CK-1 [55] and SBC [56] compression distances. AT&T [60] dataset consists of images of 40 individuals in 10 different poses, taken under different illumination conditions and facial expressions and details. Yale face dataset [61] contains 165 grayscale images in GIF format of 15 individuals. There are 11 images per subject, one per different facial expression or configuration: center-light, w/glasses, happy, left-light, w/no glasses, normal, right-light, sad, sleepy, surprised, and wink. Figure 3.10 shows examples of image retrieval in AT&T dataset. To create this figure, the NCCD between the query image and every image in the dataset is computed, and images in the dataset are ranked according to their mutual NCCD score with the query image in ascending order. The first image in the result images is the same as the query image, since NCCD between the image and itself is, as

Table 3.3: Clustering performance of AT&T and Yale face datasets using NCCD, CK-1 [55], and SBC [56]

Dataset	NCCD (%)	H.264 NCCD (%)	CK-1 (%) [55]	SBC(%) [56]
AT&T [60]	**82.8**	70.1	76.5	81.6
Yale [61]	**73.1**	73.9	64.1	65.9

expected, close to zero. It is also notable that in cases where NCCD fails to find a match of the same face in a different position, the retrieved image has a visual resemblance to the query image. For instance, if the person in the query image is wearing glasses, the first ten faces retrieved from the dataset are wearing glasses. Figure 3.11 shows examples of image retrieval in Yale face dataset. The images are ranked similar to the AT&T example. It is evident that NCCD is capable of distinguishing individual faces with different facial expressions in presence of occlusions such as glasses.

Similar to [56], we test NCCD framework by measuring its clustering performance on AT&T and Yale face datasets. For each image in the dataset, a distance vector is created based on NCCD among the image and all other images in the dataset. These vectors are then inserted into the rows of a $N \times N$ NCCD matrix, and the matrix is used by a standard spectral clustering algorithm [62] to create a vector of cluster labels. The labels are then used to predict the accuracy of the framework using Hungarian algorithm [63]. Table 3.3 shows the results of clustering AT&T and Yale face datasets using NCCD compared to CK-1 and SBC as reported in [56].

3.6 Summary

In this chapter, we aim to develop a generic image similarity measure based upon the theoretic groundwork of Kolmogorov complexity and the NID metric. The most important contribution of this chapter is to propose a practical framework of NCCD for the approximation of NID. The framework is flexible and expandable to include any image transformations that may help find the shortest description that converts one image to another and vice versa. Two different implementations of the framework are also proposed and tested. Although the first implementation and experimental work is only preliminary, the resulting similarity measure works properly in a wide variety of scenarios. To the best of our knowledge, no existing image similarity measure has achieved the same level of wide applicability. One limitation of this framework is the fact that according to Kolmogorov complexity, all the bits in the definition of images are equally important. However, this

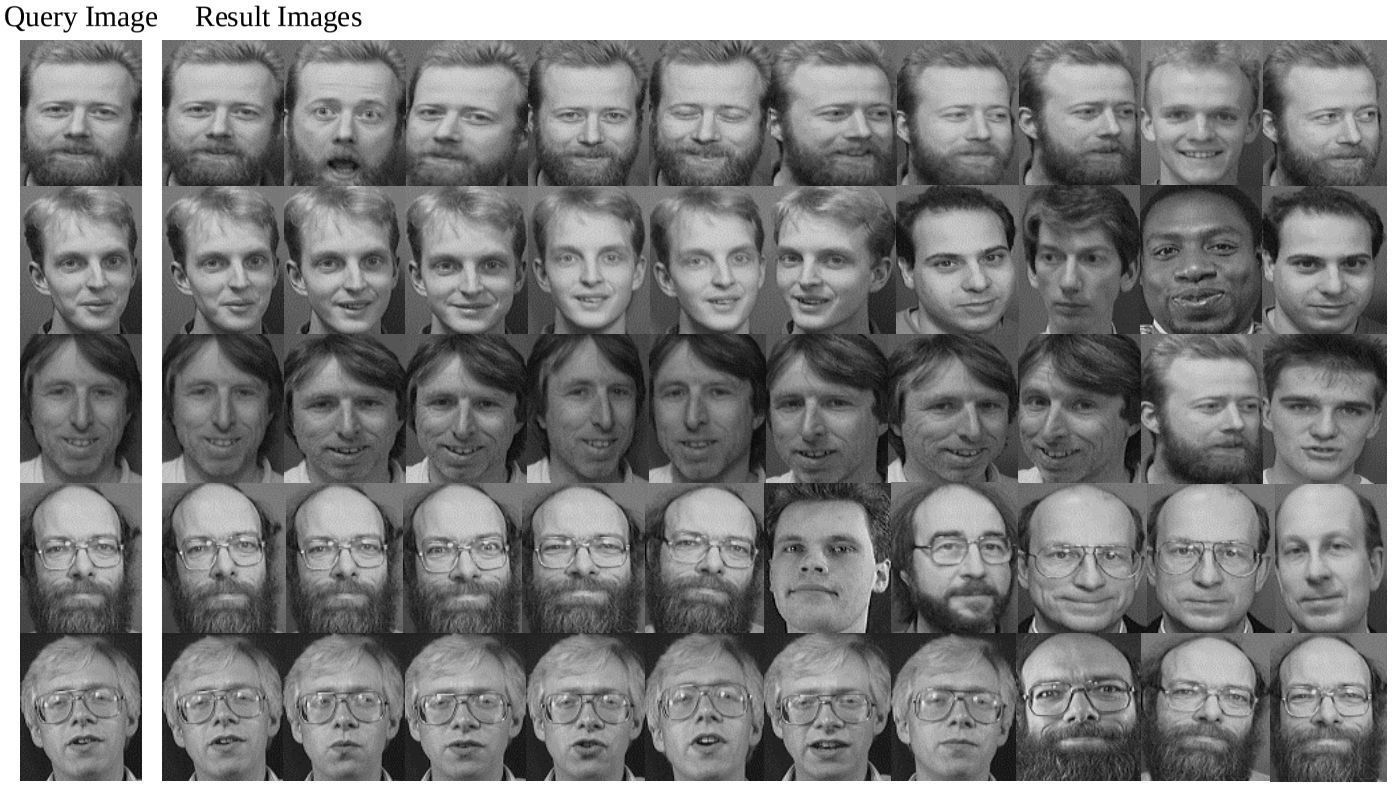


Figure 3.10: Image retrieval in AT&T Face Database [60] by Image compression based NCCD

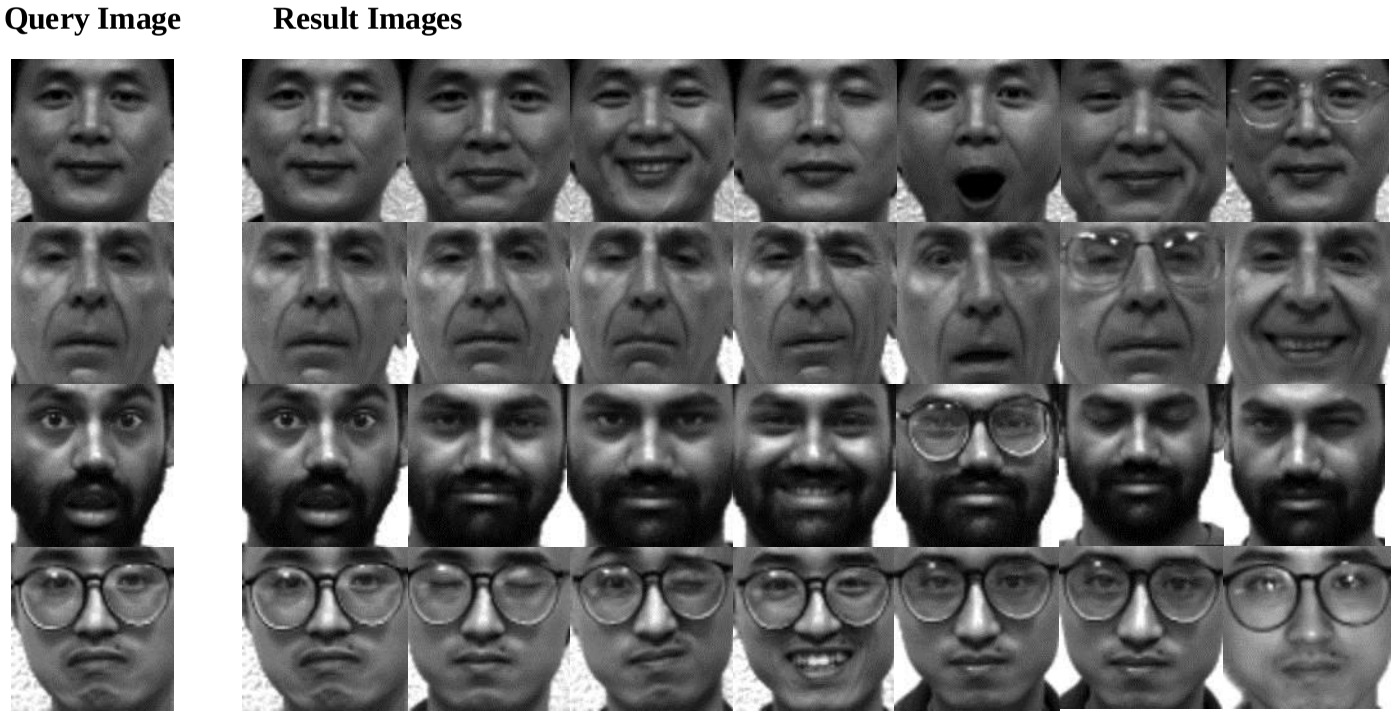


Figure 3.11: Image retrieval in Yale Face Database [61] by Image compression based NCCD

is not the case in HVS, which does not perceive the stimuli equally. Another important contribution of this chapter is the incorporation of the Divisive Normalization Transform (DNT) into the NCCD framework as a nonlinear model to remove statistical redundancies among natural images in finding perceptually relevant information content of the image. We have also made use of the power of H.264 video encoder to approximate conditional compression terms in NCCD. Advantages of using a modern video encoder such as H.264 as conditional compressor in the NCCD framework include a wider range of transformations for NCCD framework, efficient compression of transform parameters and flexible control over quantization parameters of both images. NCCD is a generic framework which allows for applications in a wide range of scenarios where many other rival methods cannot be applied. This includes large global and local geometric distortions among images, contrast and luminance change, blur, and over-sharpening. The framework can also be extended to include more distortions, and the specific algorithm for practical applications can be designed with computational complexity in mind. The proposed algorithms are tested in digit recognition, texture classification, and face recognition applications.It is shown that NCCD-based methods are competitive when compared with state-of-the-art compression based image similarity measures.

Chapter 4

Entropy Approximation of Kolmogorov Complexity with Applications

In this chapter we use Shannon entropy to approximate the non-computable Kolmogorov complexity of images and compute NID. The Shannon entropy approximations of NID are applied in two applications. The first employs wavelet domain Gaussian scale mixture model to present a perceptual image quality measure. The second allows us to choose parameters in tone-mapping operators that minimize the information loss in High Dynamic Range (HDR) to Low Dynamic Range (LDR) conversion.

4.1 Shannon Entropy and Kolmogorov Complexity

Shannon entropy provides a practically useful approximation of Kolmogorov complexity and largely simplifies the implementation of algorithms motivated from Kolmogorov complexity theory. In spite of the fact that Kolmogorov complexity and Shannon entropy are inherently different, it has been shown that they are equivalent for a wide class of information sources.

Assuming stationarity and ergodicity for a source with probability measure μ, the

following equivalence theorem applies [64]:

$$\lim_{n\to\infty} \frac{K(x^n)}{n} = H(\mu) \tag{4.1}$$

Where x^n is an object of the source X and n shows the length of the object, with $K(x^n)$ being Kolmogorov complexity of the object. It can be shown that in the cases that the source is not ergodic, the theorem can hold true if an expectation term is introduced on the Kolmogorov complexity [65]:

$$\lim_{n\to\infty} \frac{E(K(x^n))}{n} = H(\mu) \tag{4.2}$$

The above result has been generalized for any computable probability mass function $f(x) = P(X = x)$ on sample space $\chi = \{0, 1\}^*$ with entropy $H(X) = \sum_x f(x) log 1/f(x)$ [66]:

$$0 \leq \left(\sum_x f(x)K(x) - H(X) \right) \leq K(f) + O(1). \tag{4.3}$$

Therefore, for any computable distribution f, the shortest program x^* which produces the string x of the distribution and its length is equal to the Kolmogorov complexity of the string, compresses on average at least as much as the Shannon-Fano code for f, and guarantees that the expected Kolmogorov complexity of the source is close to the entropy of the source.

In this chapter, we use (4.1) to estimate of Kolmogorov complexity of in transform domains, and (4.3) to estimate Kolmogorov complexity of images in the spatial domain.

4.2 Normalized Perceptual Information Distance for Quality Assessment of Images

Image distortion analysis is a fundamental issue in many image processing problems, including compression, restoration, recognition, classification, and retrieval. Traditional image distortion evaluation approaches tend to be heuristic and are often limited to specific application environment.

In this work, we investigate the problem of image distortion measurement based on the theory of Kolmogorov complexity. To assess the distortions between two images, we

first transform them into the wavelet transform domain. Assuming stationarity and good decorrelation of wavelet coefficients beyond local regions and across wavelet subbands, the Kolmogorov complexity may be approximated using Shannon entropy [67]. Inspired by [4], we adopt a Gaussian scale mixture (GSM) model for clusters of neighboring wavelet coefficients and a Gaussian channel model for the noise distortions in the HVS. Combining these assumptions with the NID framework, we derive a novel Normalized Perceptual Information Distance (NPID) measure, where maximal likelihood estimation and least square regression are employed for parameter fitting. We validate the proposed distortion measure using three large-scale, publicly-available, and subject-rated image databases, which include a wide range of practical image distortion types and levels. Our results demonstrate the good prediction power of the proposed method for perceptual image distortions.

4.2.1 Proposed Framework

Images captured using devices which operate in visual spectrum are classified as natural scenes. Natural scenes are part of an extremely small subset of all possible images, and their statistics is often the subject of interest in many image processing applications such as compression, denoising and texture analysis [4]. Natural Scene Statistics (NSS) is often modeled as output of a stochastic source [68] and it has been used in designing NR [69] and FR information-theoretic IQA methods [4]. In quality assessment applications, natural images of perfect quality are modeled as outputs of a random process, the perceptual quality is quantified by a measuring a shared quantity such as mutual information of the distorted image and reference image.

Our proposed framework uses NID to quantify the perceptual quality of an image as compared to its reference image. In order to estimate the non-computable Kolmogorov complexity terms in NID, we need to use Shannon entropy as an upper bound on the Kolmogorov complexity of the image. The equivalence theorem introduced in section 4.1 states that for Kolmogorov complexity of an infinitely long string to converge to Shannon entropy, the source distribution must be stationary and ergodic. In general, stationarity and ergodicity are not good assumptions of image signals in spatial domain, thus we are interested in transform domain representations of images, where more reasonable assumptions may be made. Specifically, we transform the reference and distorted images into the wavelet domain, and assume stationarity, decorrelation and local independence among the image subbands. Based on these assumptions we have:

$$K(x|y) = \sum_{i=1}^{n} K(x_n|y_n) \tag{4.4}$$

where K stands for Kolmogorov complexity of the images and x_n and y_n are the corresponding wavelet subbands of the reference and distorted image, respectively. Based on (4.4) the NID can be reformulated into:

$$\mathrm{NID}(x,y) = \frac{\max\{\sum_{i=1}^{n} K(x_n|y_n), \sum_{i=1}^{n} K(y_n|x_n)\}}{\max\{\sum_{i=1}^{n} K(x_n), \sum_{i=1}^{n} K(y_n)\}} \tag{4.5}$$

Based on the stationarity and independence assumptions for wavelet coefficients, we may re-write (4.5) in Shannon entropy framework based on (4.1):

$$\mathrm{NID_s}(x,y) = \frac{\max\{\sum_{i=1}^{n} H(x_n|y_n), \sum_{i=1}^{n} H(y_n|x_n)\}}{\max\{\sum_{i=1}^{n} H(x_n), \sum_{i=1}^{n} H(y_n)\}} \tag{4.6}$$

to further simplify the results and inspired by the VIF method introduced in [4], we adopt a wavelet domain GSM model for natural scene statistics. We also use the same distortion and HVS models used in [4].

A Gaussian Scale Mixture (GSM) is a random field which is represented as a product of a zero-mean Gaussian random vector U, and an independent scalar random field, S. In this sense, the GSM, C is defined to be $C = \{\overrightarrow{C}_i : i \in I\}$, where I is the set of spatial indices for the random field [4], and:

$$C = SU = \{S_i \overrightarrow{U}_i : i \in I\} \tag{4.7}$$

where $S = \{S_i : i \in I\}$ is a field of random scalars and $U = \{\overrightarrow{U}_i : i \in I\}$ is the random field of Gaussian vectors. It is easy to show that with the above assumption, the $\overrightarrow{C}_i$ vectors are normally distributed and independent given the random scalar s, which makes GSM easier to use in modeling clusters of coefficients of image subbands. It has also been shown that GSM is capable of modeling important statistical features of natural images, such as heavy-tailed marginal distributions of the wavelet coefficients of natural images and non-linear dependencies among them [4]. This makes the GSM model more appealing for our application.

In the case that C is a GSM representing a cluster of coefficients in a natural image we have:

$$E = C + \mathcal{N} \tag{4.8}$$

$$F = D + \mathcal{N}' \tag{4.9}$$

where $\mathcal{N}$ and $\mathcal{N}'$ are random fields of uncorrelated multivariate Gaussian noise that rep-

resent the internal neural noise in the HVS. D is a distortion model comprised of a signal gain and additive noise:

$$D = gC + \nu = gsU + \nu \tag{4.10}$$

and E and F are the clusters of coefficients in the reference and the distorted images, respectively [4]. Figure 4.1 shows a graphical representation of this model.

Let $\overrightarrow{C}_j = (\overrightarrow{C_1}, \ldots, \overrightarrow{C_N})_j$ represent N elements of a subband C_j, and $\overrightarrow{D}_j$, $\overrightarrow{E}_j$, $\overrightarrow{F}_j$ be defined correspondingly. Assuming that $S^N = s^N$ and is given for all the variables[1], then NIDS becomes:

$$\mathrm{NID_s}(E,F) = \frac{\max\{\sum_{j=1}^{n} H(\overrightarrow{E}_j|\overrightarrow{F}_j), \sum_{j=1}^{n} H(\overrightarrow{F}_j|\overrightarrow{E}_j)\}}{\max\{\sum_{j=1}^{n} H(\overrightarrow{E}_j), \sum_{j=1}^{n} H(\overrightarrow{F}_j)\}} \tag{4.11}$$

which can be further simplified into:

$$\mathrm{NID_s}(E,F) = 1 - \frac{\sum_j I(\overrightarrow{E}_j; \overrightarrow{F}_j)}{\max\{\sum_j H(\overrightarrow{E}_j), \sum_j H(\overrightarrow{F}_j)\}} \tag{4.12}$$

Since both $\overrightarrow{E}_j$ and $\overrightarrow{F}_j$ are continuous random variables, direct calculation of their entropies is difficult while their differential entropies do not provide adequate measures of their information content. To overcome this problem, we use the information content contained in the source $\overrightarrow{C}_j$ as the baseline and replace $H(\overrightarrow{E}_j)$ and $H(\overrightarrow{F}_j)$ with $I(\overrightarrow{C}_j; \overrightarrow{E}_j)$ and $I(\overrightarrow{C}_j; \overrightarrow{F}_j)$, respectively, which quantify the information content that is perceived by the HVS in the original and distorted images. We then define a Normalized Perceptual Information Distance (NPID) as:

$$\mathrm{NPID}(E,F) = 1 - \frac{\sum_j I(\overrightarrow{E}_j; \overrightarrow{F}_j)}{\max\{\sum_j I(\overrightarrow{E}_j; \overrightarrow{C}_j), \sum_j I(\overrightarrow{F}_j; \overrightarrow{C}_j)\}} \tag{4.13}$$

A Normalized Perceptual Information Similarity (NPIS) can be defined as:

$$\mathrm{NPIS}(E,F) = 1 - \mathrm{NPID}(E,F) = \frac{\sum_j I(\overrightarrow{E}_j; \overrightarrow{F}_j)}{\max\{\sum_j I(\overrightarrow{E}_j; \overrightarrow{C}_j), \sum_j I(\overrightarrow{F}_j; \overrightarrow{C}_j)\}} \tag{4.14}$$

[1] hence $.|s^N$ is dropped in all notations

(4.14) can be further simplified by using the fact that $\overrightarrow{E}_j$ and $\overrightarrow{F}_j$ are Gaussian for given s and the mutual information of correlated Gaussians can be calculated based on the determinants of the covariances [32]:

$$I(\overrightarrow{E}_j;\overrightarrow{F}_j) = \frac{1}{2}\log_2[\frac{|\mathbf{C}_E||\mathbf{C}_F|}{|\mathbf{C}_{(F,E)}|}] \tag{4.15}$$

$$I(\overrightarrow{E}_j;\overrightarrow{C}_j) = \frac{1}{2}\log_2[\frac{|\mathbf{C}_C||\mathbf{C}_E|}{|\mathbf{C}_{(C,E)}|}] \tag{4.16}$$

$$I(\overrightarrow{F}_j;\overrightarrow{C}_j) = \frac{1}{2}\log_2[\frac{|\mathbf{C}_C||\mathbf{C}_F|}{|\mathbf{C}_{(C,F)}|}] \tag{4.17}$$

Covariance matrices of C, D, E and F are respectively computed to be [32]:

$$\mathbf{C}_C = s^2\mathbf{C}_\mathbf{U} \tag{4.18}$$

$$\mathbf{C}_D = g^2s^2\mathbf{C}_\mathbf{U} + \sigma_\nu^2\mathbf{I} \tag{4.19}$$

$$\mathbf{C}_E = s^2\mathbf{C}_\mathbf{U} + \sigma_n^2\mathbf{I} \tag{4.20}$$

$$\mathbf{C}_F = g^2s^2\mathbf{C}_\mathbf{U} + (\sigma_\nu^2+\sigma_n^2)\mathbf{I} \tag{4.21}$$

and we also have:

$$|\mathbf{C}_{(E,F)}| = \begin{vmatrix} \mathbf{C}_E & \mathbf{C}_{EF} \\ \mathbf{C}_{FE} & \mathbf{C}_F \end{vmatrix}, \tag{4.22}$$

$$|\mathbf{C}_{(C,E)}| = \begin{vmatrix} \mathbf{C}_C & \mathbf{C}_{CE} \\ \mathbf{C}_{EC} & \mathbf{C}_E \end{vmatrix}, \tag{4.23}$$

$$|\mathbf{C}_{(C,F)}| = \begin{vmatrix} \mathbf{C}_C & \mathbf{C}_{CF} \\ \mathbf{C}_{FC} & \mathbf{C}_F \end{vmatrix}. \tag{4.24}$$

It can be easily shown that: $\mathbf{C}_C = s^2\mathbf{C}_U$, $\mathbf{C}_{EF} = \mathbf{C}_{FE} = gs^2\mathbf{C}_\mathbf{U}$, $\mathbf{C}_{CE} = \mathbf{C}_{EC} = s^2\mathbf{C}_\mathbf{U}$ and $\mathbf{C}_{CF} = \mathbf{C}_{FC} = gs^2\mathbf{C}_\mathbf{U}$. Thus (4.22) - (4.24) can be written as:

$$|\mathbf{C}_{(E,F)}| = |[(\sigma_\nu^2+\sigma_n^2)s^2 + \sigma_n^2g^2s^2]\mathbf{C}_\mathbf{U} + \sigma_n^2(\sigma_\nu^2+\sigma_n^2)\mathbf{I}| \tag{4.25}$$

$$|\mathbf{C}_{(C,E)}| = |\sigma_n^2s^2\mathbf{C}_\mathbf{U}| \tag{4.26}$$

$$|\mathbf{C}_{(C,F)}| = |(\sigma_\nu^2+\sigma_n^2)s^2\mathbf{C}_\mathbf{U}| \tag{4.27}$$

Since $\mathbf{C}_\mathbf{U}$ is a symmetric matrix, a further eigenvalue decomposition can be applied, which gives $\mathbf{C}_\mathbf{U} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T$, where $\mathbf{Q}$ is an orthogonal matrix, and $\mathbf{\Lambda}$ is a diagonal matrix with

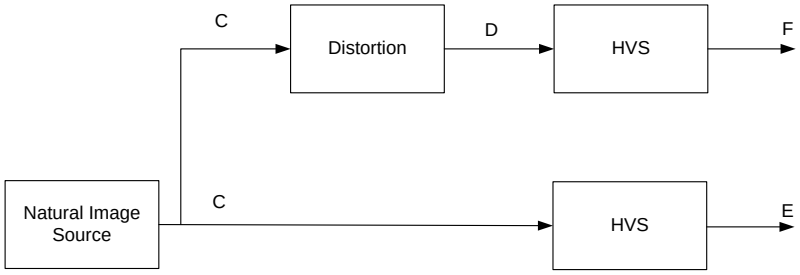


Figure 4.1: Natural Image Source is corrupted by a Distortion channel (D) and then passes through the HVS. Mutual information between C and E quantifies the amount of information extracted by the HVS from an original image and mutual information between C and F quantifies the amount of information extracted from a distorted image [4].

eigenvalues λ_k for $k = 1, 2, \cdots, K$ along the diagonal entries. (4.18), (4.20), (4.21) can then be expressed as:

$$\mathbf{C}_C = \mathbf{Q}\{s^2\mathbf{\Lambda}\}\mathbf{Q}^T \tag{4.28}$$

$$\mathbf{C}_E = \mathbf{Q}\{s^2\mathbf{\Lambda} + \sigma_n^2\mathbf{I}\}\mathbf{Q}^T \tag{4.29}$$

$$\mathbf{C}_F = \mathbf{Q}\{g^2s^2\mathbf{\Lambda} + (\sigma_\nu^2 + \sigma_n^2)\mathbf{I}\}\mathbf{Q}^T \tag{4.30}$$

(4.25) - (4.27) become:

$$|\mathbf{C}_{(E,F)}| = |\mathbf{Q}\{[\sigma_\nu^2 + (1+g^2)\sigma_n^2]s^2\mathbf{\Lambda} + \\ +\sigma_n^2(\sigma_\nu^2 + \sigma_n^2)\mathbf{I}\}\mathbf{Q}^T| \tag{4.31}$$

$$|\mathbf{C}_{(C,E)}| = |\mathbf{Q}\{\sigma_n^2 s^2\mathbf{\Lambda}\}\mathbf{Q}^T| \tag{4.32}$$

$$|\mathbf{C}_{(C,F)}| = |\mathbf{Q}\{(\sigma_\nu^2 + \sigma_n^2)s^2\mathbf{\Lambda}\}\mathbf{Q}^T| \tag{4.33}$$

Since $\mathbf{Q}$ is orthogonal and the middle matrices between $\mathbf{Q}$'s in (4.31),(4.32) and (4.33) are

diagonal, calculation of the desired determinants can be simplified to a great extent [32]:

$$|\mathbf{C}_{E,F}| = \prod_{k=1}^{K}\{[\sigma_\nu^2 + (1+g^2)\sigma_\nu^2]s^2\lambda_k + \sigma_n^2(\sigma_\nu^2 + \sigma_n^2)\} \tag{4.34}$$

$$|\mathbf{C}_{C,E}| = \prod_{k=1}^{K}\{\sigma_n^2 s^2 \lambda_k\} \tag{4.35}$$

$$|\mathbf{C}_{C,F}| = \prod_{k=1}^{K}\{(\sigma_\nu^2 + \sigma_n^2)s^2\lambda_k\} \tag{4.36}$$

Thus the mutual information relations (4.15), (4.16) and (4.17) are simplified:

$$I(\overrightarrow{C}_j; \overrightarrow{E}_j) = \frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{K}\log_2(1 + \frac{s_i^2\lambda_k}{\sigma_n^2}) \tag{4.37}$$

$$I(\overrightarrow{C}_j; \overrightarrow{F}_j) = \frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{K}\log_2(1 + \frac{g^2 s_i^2\lambda_k}{\sigma_n^2 + \sigma_\nu^2}) \tag{4.38}$$

$$I(\overrightarrow{E}_j; \overrightarrow{F}_j) = \frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{K}\log_2(\frac{[g^2 s_i^2\lambda_k + (\sigma_\nu^2 + \sigma_n^2)](s_i^2\lambda_k + \sigma_n^2)}{[(\sigma_\nu^2 + \sigma_n^2)s_i^2 + \sigma_n^2 g^2 s_i^2]\lambda_k + \sigma_n^2(\sigma_n^2 + \sigma_\nu^2)}) \tag{4.39}$$

To finish the computation of NPIS, we conclude by estimating a set of parameters involved in the calculations, including $\mathbf{C_U}$, s^2, g and σ_v^2. We follow the same path taken by [4, 32]:

$$\hat{\mathbf{C}}_\mathbf{U} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{C}_i\mathbf{C}_i^T \tag{4.40}$$

where N is the number of evaluation windows in the corresponding subband and C_i is the *i-th* neighborhood coefficient. The multiplier s is estimated using a maximum-likelihood estimator [32]:

$$\hat{s}^2 = \frac{1}{K}\mathbf{C}^T\mathbf{C}_\mathbf{U}^{-1}\mathbf{C} \tag{4.41}$$

Least square optimization may be used to estimate the parameters g and σ_v^2:

$$\hat{g} = \arg\min_{g} \|\mathbf{D} - g\mathbf{C}\|_2^2 \tag{4.42}$$

$$\tag{4.43}$$

Take derivative of the squared error cost function and let it be zero, we have:

$$\hat{g} = \frac{C^T D}{C^T C} \tag{4.44}$$

by substituting (4.44) into (4.10), we can estimate ${\sigma_\nu}^2$ by:

$$\hat{\sigma}_\nu^2 = \frac{1}{K}(\mathbf{D}^T\mathbf{D} - \hat{g}\mathbf{C}^T\mathbf{D}) \tag{4.45}$$

In practice, when calculating the NPIS, we apply a five-scale Laplacian pyramid decomposition [70] to the original and distorted images, and compute the respective mutual information according to (4.37), (4.38) and (4.39) using a sliding 3×3 window that runs across each subband. At each location, the window contains a spatial neighborhood of ten coefficients (3×3 neighboring coefficients and one parent coefficient, thus $K = 10$).

In [32] the authors propose an optimal pooling strategy for IQA algorithms using a multi-scale information content weighting approach based on a GSM model of natural images. The information weighting scheme is based on estimating the total perceptual information content for the reference and distorted images from evaluation of local information content of the images. It is shown that information content weighting often leads to significant improvements in performance of IQA methods. To incorporate this scheme, we first define a local-NPIS (L-NPIS) measure by:

$$\text{L-NPIS}_i = \frac{I(\overrightarrow{E}_i; \overrightarrow{F}_i)}{\max\{I(\overrightarrow{E}_i; \overrightarrow{C}_i), I(\overrightarrow{F}_i; \overrightarrow{C}_i)\}} \tag{4.46}$$

We can then compute an information content-weighted NPIS (IW-NPIS) measure using:

$$\text{IW-NPIS}_j = \frac{\sum_i \omega_{j,i}\text{L-NPIS}_i}{\sum_i \omega_{j,i}} \tag{4.47}$$

where $w_{j,i}$ is the weight assigned to the L-NPIS value calculated at the *i-th* location at *j-th* scale, and the value of $w_{j,i}$ is calculated based on the information content model given

in [32]. Using the same fine-to-coarse scale weights from [28], we have:

$$\text{IW-NPIS} = \prod_{j=1}^{M} (\text{IW-NPIS}_j)^{\beta_j} \tag{4.48}$$

4.2.2 Tests with Image Quality Databases

To evaluate the performance of the proposed method, we test it using Laboratory for Image and Video Engineering (LIVE) [33], Tampere Image Database 2008 (TID2008) [34], Categorical Image Quality (CSIQ) [35], IVC [36, 37], Cornell-A57 [39] and Toyama-MICT [38] databases and compare the results with a series of widely known and state-of-the-art IQA algorithms, including Peak Signal to Noise Ratio (PSNR) [32], Structural Similarity Index Measure (SSIM) [3], Information Weighted Structural Similarity Index Measure (IW-SSIM) [32], Visual Information Fidelity (VIF) [4], Visual Signal to Noise Ratio (VSNR) [71], HVS-based PSNR [72], Information Weighted PSNR (IW-PSNR) [32], and Most Apparent Distortion (MAD) [73].

Figures 4.2 and 4.4 show a sample scatter plot of NPIS vs. six different subjective quality evaluation databases, where the subjective scores are given by Mean Opinion Score (MOS) or Difference of Mean Opinion Score (DMOS) between a distorted image and its corresponding original reference image. Each point in the scatter plot represents one image in the corresponding database. Figures 4.3 and 4.5 show a similar scatter plot for IW-NPIS. Tables 4.1, 4.2, 4.3, 4.4, 4.5, and 4.6 provide a comparison of the proposed methods with PSNR and state-of-the-art methods. The results of the proposed methods are in bold face. It can be observed that the proposed methods perform significantly and consistently better than PSNR and are in general comparable to many state-of-the-art algorithms. Specifically, IW-NPIS achieves the second best result for the TID2008 database, which is the largest database currently available in terms of the number of test images, and has the most diverse types of image distortions. This is impressive as an early attempt to use Kolmogorov complexity and NID theories for image quality assessment, where many existing methods are in their mature stages.

Table 4.1: Performance comparison based on LIVE [33] database

Model	PLCC	MAE	RMS	SRCC	KRCC
PSNR	0.8723	10.51	13.36	0.8756	0.6865
SSIM [3]	0.9449	6.933	8.946	0.9479	0.7963
IW-SSIM [32]	0.9556	6.212	8.047	0.9570	0.8197
VIF [4]	0.9598	6.148	7.667	0.9632	0.8270
VSNR [71]	0.9229	8.089	10.52	0.9271	0.7610
PSNR-HVS-M [72]	0.9251	7.966	10.37	0.9295	0.7659
MAD [73]	0.9394	7.293	9.368	0.9438	0.7920
NPIS	**0.9211**	**7.458**	**8.491**	**0.9093**	**0.7514**
IW-NPIS	**0.9339**	**7.013**	**8.011**	**0.9376**	**0.7891**

Table 4.2: Performance comparison based on TID2008 [34] database

Model	PLCC	MAE	RMS	SRCC	KRCC
PSNR	0.5223	0.8683	1.1435	0.5531	0.4027
SSIM [3]	0.7732	0.6546	0.8511	0.7749	0.5768
IW-SSIM [32]	0.8579	0.5276	0.6895	0.8559	0.6636
VIF [4]	0.8090	0.5990	0.7888	0.7496	0.5863
VSNR [71]	0.6820	0.6908	0.9815	0.7046	0.5340
PSNR-HVS-M [72]	0.5519	0.8036	1.1190	0.5612	0.4509
MAD [73]	0.7480	0.6641	0.8907	0.7708	0.5734
NPIS	**0.7855**	**0.6239**	**0.8111**	**0.7682**	**0.5013**
IW-NPIS	**0.8244**	**0.5846**	**0.7637**	**0.8167**	**0.5819**

Table 4.3: Performance comparison based on CSIQ [35] database

Model	PLCC	MAE	RMS	SRCC	KRCC
PSNR	0.7512	0.1366	0.1733	0.8058	0.6084
SSIM [3]	0.8612	0.0992	0.1334	0.8756	0.6907
IW-SSIM [32]	0.9144	0.0801	0.1063	0.9213	0.7529
VIF [4]	0.9277	0.0743	0.0980	0.9195	0.7537
VSNR [71]	0.7355	0.1335	0.1779	0.8109	0.6248
PSNR-HVS-M [72]	0.7725	0.1290	0.1667	0.8222	0.6529
MAD [73]	0.8202	0.1258	0.1502	0.8988	0.7272
NPIS	**0.8999**	**0.1024**	**0.1188**	**0.8643**	**0.5920**
IW-NPIS	**0.9023**	**0.0923**	**0.1070**	**0.8985**	**0.6413**

Table 4.4: Performance comparison based on IVC [36, 37] database

Model	PLCC	MAE	RMS	SRCC	KRCC
PSNR	0.6719	0.7191	0.9023	0.6884	0.5218
SSIM [3]	0.9119	0.3777	0.4999	0.9018	0.7223
IW-SSIM [32]	0.9231	0.3694	0.4686	0.9125	0.7339
VIF [4]	0.9028	0.4104	0.5239	0.8964	0.7158
VSNR [71]	0.7904	0.5860	0.7463	0.7993	0.6053
PSNR-HVS-M [72]	0.8788	0.4614	0.5815	0.8832	0.6935
MAD [73]	0.8741	0.4728	0.5918	0.9150	0.7406
NPIS	**0.8865**	**0.4321**	**0.5689**	**0.8819**	**0.6920**
IW-NPIS	**0.9069**	**0.3978**	**0.5149**	**0.9029**	**0.7174**

Table 4.5: Performance comparison based on Cornell A57 [39] database

Model	PLCC	MAE	RMS	SRCC	KRCC
PSNR	0.6347	0.1607	0.1899	0.6189	0.4309
SSIM [3]	0.8017	0.1209	0.1469	0.8066	0.6058
IW-SSIM [32]	0.9034	0.0892	0.1054	0.8709	0.6842
VIF [4]	0.6157	0.1397	0.1937	0.6223	0.4589
VSNR [71]	0.9146	0.0.0809	0.0.0994	0.9355	0.8031
PSNR-HVS-M [72]	0.8748	0.0923	0.1190	0.8962	0.7261
MAD [73]	0.8816	0.0942	0.1160	0.8645	0.6702
NPIS	**0.8850**	**0.0987**	**0.1164**	**0.8662**	**0.6789**
IW-NPIS	**0.8894**	**0.0912**	**0.1018**	**0.8853**	**0.6951**

Table 4.6: Performance comparison based on Toyama-MICT [38] database

Model	PLCC	MAE	RMS	SRCC	KRCC
PSNR	0.6329	0.7817	0.9689	0.6132	0.4443
SSIM [3]	0.8887	0.4386	0.5738	0.8794	0.6939
IW-SSIM [32]	0.9248	0.3677	0.4761	0.9202	0.7537
VIF [4]	0.9138	0.4038	0.5084	0.9077	0.7315
VSNR [71]	0.8705	0.4654	0.6159	0.8608	0.6745
PSNR-HVS-M [72]	0.8406	0.5541	0.6777	0.8480	0.6568
MAD [73]	0.9116	0.3951	0.5145	0.9086	0.7354
NPIS	**0.9063**	**0.4087**	**0.5242**	**0.8765**	**0.7049**
IW-NPIS	**0.9123**	**0.3726**	**0.4890**	**0.9096**	**0.7432**

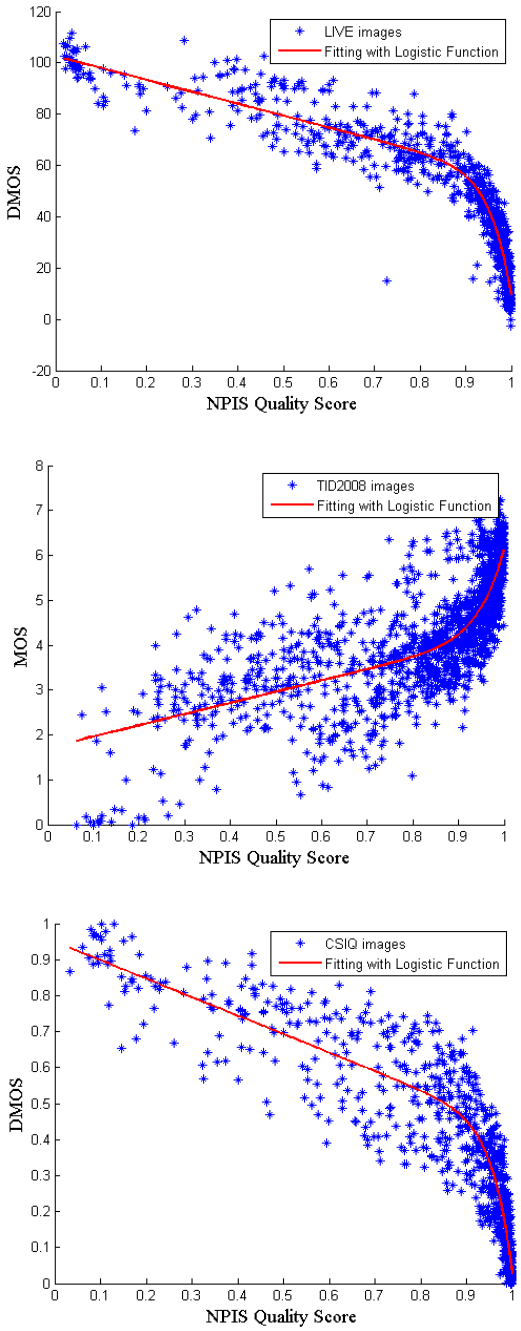


Figure 4.2: Scatter plots of NPIS vs. subjective scores for LIVE [33] TID2008 [34] and CSIQ [35] databases by NPIS

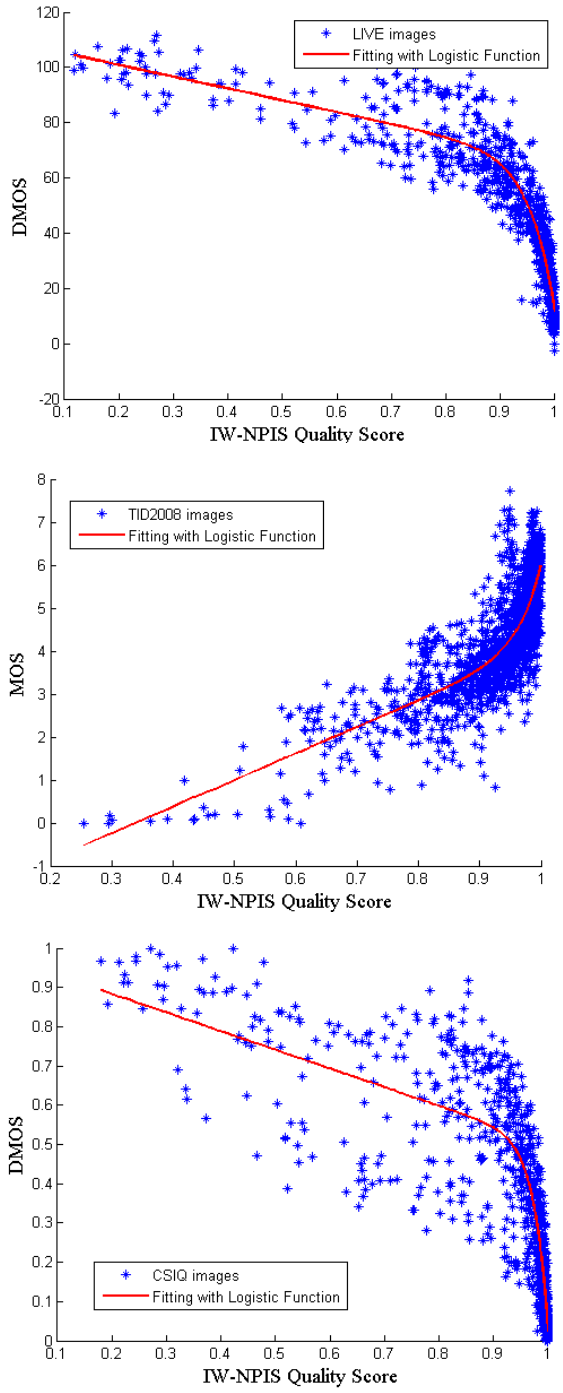


Figure 4.3: Scatter plots of IW-NPIS vs. subjective scores for LIVE [33] TID2008 [34] and CSIQ [35] databases by IW-NPIS

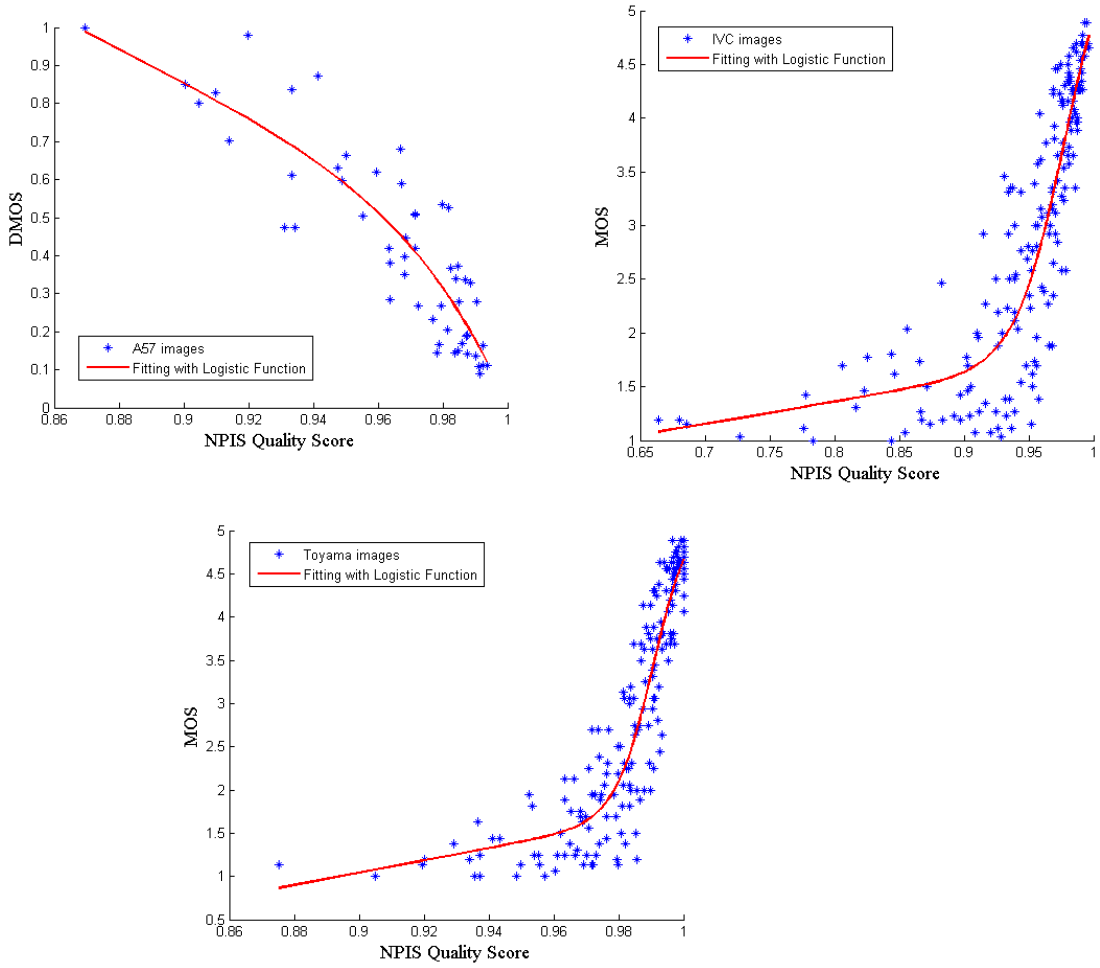


Figure 4.4: Scatter plots of NPIS vs. subjective scores for Cornell-A57 [39] IVC [36, 37] and Toyama [38] databases by NPIS

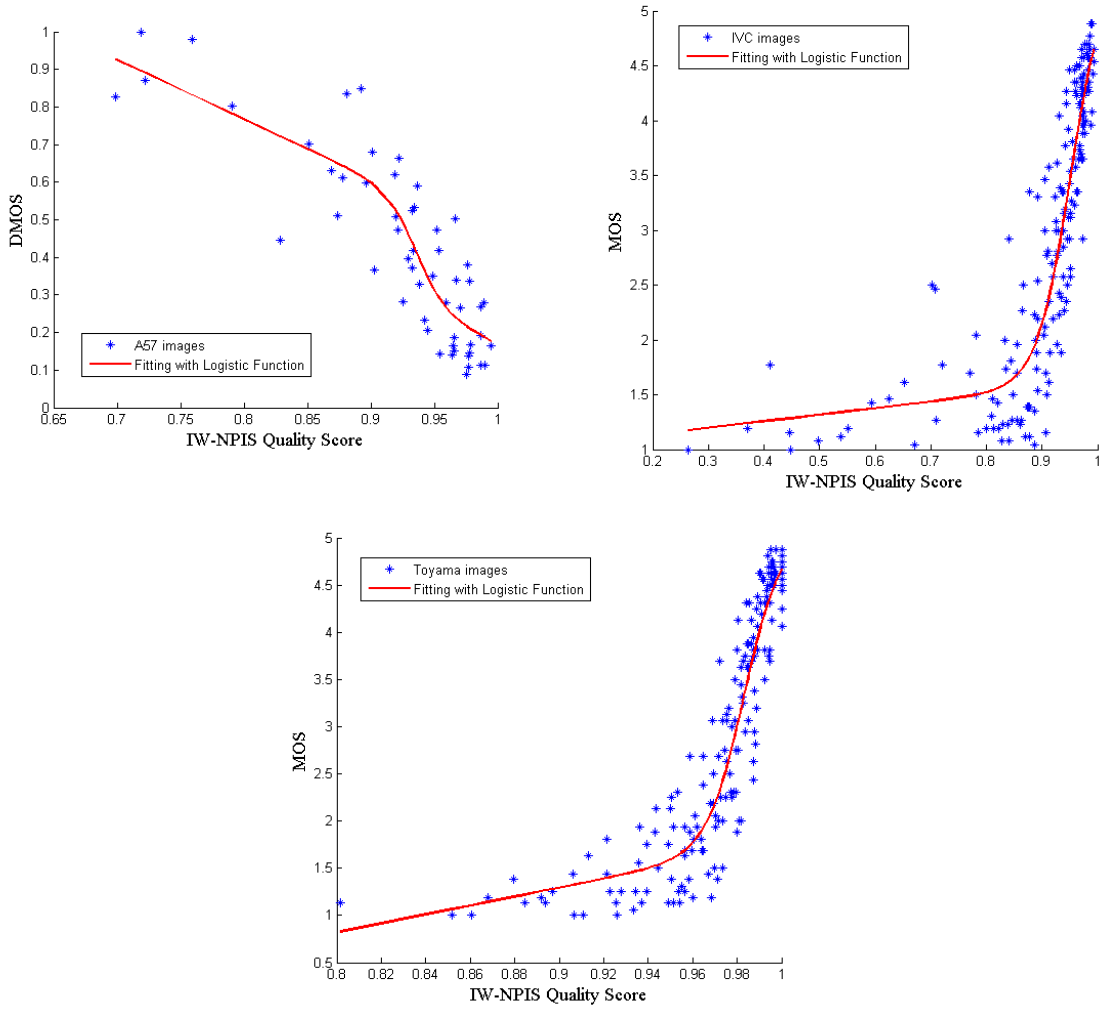


Figure 4.5: Scatter plots of IW-NPIS vs. subjective scores for Cornell-A57 [39] IVC [36,37] and Toyama [38] databases by IW-NPIS

4.3 NID Based Parameter Tuning For HDR Tone-Mapping

High Dynamic Range (HDR) images can capture a wider range of lighting variations and allow for a more accurate representation of luminance changes in over-exposed and under-exposed areas of a photograph. This can provide a better overall quality for images which have both very dark and bright areas in the same scene. Most medical imaging devices capture and store raw data in high precision format in order to preserve as much critical detail as possible for the medical experts. Since the advent of HDR imaging technology, visualization of such images on traditional displays has been a challenge. In order to visualize a HDR image on a standard display, the contrast ratio of the entire photograph must be reduced, i.e. the high dynamic range of the photograph must be mapped to a lower range. This operation is called Tone mapping and inevitably results in loss of information from the original image. Several tone mapping algorithms have been proposed [74–77], many of which have one or more image dependent parameters, and proper selection of these parameters can greatly affect the perceptual quality, structural fidelity and information loss in the Low Dynamic Range (LDR) image.

Medical images are typically captured with higher precisions or higher dynamic ranges of intensity values than what can be directly shown on standard displays with 8-bit depth. Standard medial image formats such as DICOM allow to store such HDR images with more bit depths, but to visualize them on regular displays becomes a challenge. In practice, a so-called "windowing approach is often employed, which linearly maps an intensity interval of interest to the dynamic range of the display. These intervals are defined using two parameters: (i) window width, or the range of the interval, W (which is typically larger than 255); and (ii) window center, or the center of this interval, C. Thus a windowing operator maps the range of intensity values $[C - \frac{1}{2}W, C + \frac{1}{2}W]$ to a low dynamic range [0,255]. The default values for W and C may be embedded in the headers of DICOM image files, or determined manually by the end users (radiologists so that the structural details for specific body region become more visible.

In this section, we aim to develop a new parameter selection algorithm for tone-mapping operators and windowing operators for optimal visualization of medical and non-medical HDR images, where the optimality is defined as maximization of the information similarity between the HDR image and the mapped LDR image. A key step in our approach is to approximate NID based information similarity using a Shannon entropy approach. Our experiments show that when the new similarity measure is employed in optimizing non-medical operators and two types parametric windowing medical operators, perceptually

appealing images with higher contrast and more visible structural details are obtained.

4.3.1 Tone mapping operators

Tone mapping operators compress the dynamic range of HDR content to a lower dynamic range and makes it adaptable to traditional displays. The process often starts from an image with floating point accuracy and results in an integer image with byte accuracy in the range of [0-255]. More formally this is defined as [78]:

$$f(I) : \mathbb{R}_i^{w \times h \times 3} \rightarrow \mathbb{D}_o^{w \times h \times 3} \tag{4.49}$$

where I represents the image, and f is the Tone Mapping Operator (TMO) defined from a subset of real numbers $\mathbb{R}_i$ to a subset of integer numbers $\mathbb{D}_0$. Tone mapping algorithms have been researched in the past two decades and several successful techniques have been developed. These operators are categorized into four groups, namely global operators, local operators, frequency operators, and gradient operators [79]. Global operators apply a standard function on the whole image, and in most cases, optimal parameters for this function vary from image to image. In the absence of human interaction, choosing these parameters can be quite challenging. Our goal is to select these parameters such that loss of information from the HDR to LDR domain is minimized. In this work, we use two global tone mapping operators; Piecewise linear basis windowing, and Sine basis windowing proposed in [80], as showcase examples of parameter tuning based on NID.

Piecewise linear function is defined by an equipartition of HDR dynamic range $[l_l, l_u]$ into n subintervals $I_k = [l_{k-1}, l_k]$ for $1 \leq k \leq K$ of length $\Delta l = (l_u - l_l)/K$. Partition points are then found by $l_k = l_l + k\Delta l$ with $0 \leq k \leq n$, and Window width and center are:

$$W = l_n - l_0 = n\Delta l, \quad C = \frac{1}{2}(l_0 + l_n) = l_0 + \frac{n\Delta l}{2} \tag{4.50}$$

It is shown that the equipartitions of piecewise linear function can be represented as linear combination of n basis ramp functions [80]:

$$f(t) = \sum_{k=0}^{n-1} c_k \phi_k(l) = \phi_0(l) + \sum_{k=1}^{n-1} c_k \phi_k(l) \tag{4.51}$$

where:

$$\phi_0(l) = \begin{cases} (l - l_0)/W, & l_0 \leq l \leq l_n \\ 0, & \text{otherwise} \end{cases}$$

and

$$\phi_k(l) = t(\frac{l - l_k}{\Delta l}) \qquad \text{for } k = 1, \cdots, n-1 \tag{4.52}$$

with function $t(l)$ defined as:

$$t(l) = \begin{cases} 1 - |l|, & l_0 \leq l \leq l_n \\ 0, & \text{otherwise} \end{cases} \tag{4.53}$$

In order to use this function as a tone mapping operator, the function must be defined such that it is monotonically increasing. To ensure this, the coefficients are selected to meet the following condition:

$$0 \leq \cdots \leq c_{k-1} + \frac{k-1}{n} \leq c_k + \frac{k}{n} \leq \cdots \leq 1. \tag{4.54}$$

It is shown that choosing piecewise linear function as tone mapping operator results in higher structural fidelity LDR images compared to the linear function that is usually used in medical imaging industry [80].

It is also possible to define the windowing function as a linear combination of a family of sine basis functions [80]:

$$\phi_k(l) = \sin(\frac{k\pi(l - l_l)}{W}) \qquad \text{for } l_l \leq l \leq l_u \text{ and } k = 1, 2, \cdots \tag{4.55}$$

In the special case of $n = 3$ we have:

$$f(l) = \frac{l - l_l}{W} + c_1 \sin(\frac{\pi(l - l_l)}{W}) + c_2 \sin(\frac{2\pi(l - l_l)}{W}). \tag{4.56}$$

4.3.2 Parameter Selection Scheme

Assuming that f is an intensity mapping function, and $T_f(.)$ is the tone mapping operator which uses f to compress the dynamic range of the HDR image and quantizes the compressed range into lower dynamic range, we intend to find the parameters for the mapping function such that the loss of information in tone mapping operation is minimized. The windowing process in medical imaging may also be understood as a special case of the tone-mapping operation (TMO) that converts HDR images to LDR images. In the context of medical imaging, global operators implemented using monotonic intensity transformations are preferred because it is the only category that maintains one-to-one mapping of

intensity values and preserves the ranks of pixel intensity values. By contrast, other TMOs may map the same intensity value in the HDR image to different values in the LDR image, which may confuse the understanding of the physical meanings behind the intensity values.

Standard windowing operation in medical imaging linearly maps the intensity interval of interest $[l_l, l_u]$ to the dynamic range of the LDR image, typically $[0, 255]$. This has often been shown to be far from optimal in terms of perceived image quality [80]. To develop a better windowing method, we relax the mapping operation to be a continuous and monotonically increasing function f lives in the function space of

$$\mathcal{F}_{[l_l,l_u]} = \{f : [l_l, l_u] \to [0, 255] | f \text{ monotonically increasing}\} \tag{4.57}$$

For any given f, we can then define a windowing operator T_f over an input HDR image x by

$$y = T_f(x) = \text{round}\{f(x)\}\,, \tag{4.58}$$

where since both images can take only integer intensity values, a rounding operator is necessary. The key question now is to obtain an LDR image y that is optimal in certain criterion. Motivated by the ideas behind NID, we would want to find an image y such that the normalized information similarity between x and y is maximized. Therefore, the problem of finding the optimal windowing operator can be expressed as

$$f_{opt-\text{NID}} = \underset{f \in \mathcal{F}_{[l_l,l_u]}}{\arg\min}\, \text{NID}(x, T_f(x))\,. \tag{4.59}$$

To provide a practical algorithm to compute NID, we resort to a Shannon entropy approximation of the Kolmogorov complexity, leading to a normalized Shannon information distance

$$\text{NID}(x, y) \approx \frac{\max\{H(x|y), H(y|x)\}}{\max\{H(x), H(y)\}} \tag{4.60}$$

Since the conversion from x to y is unique, there is no uncertainty in y given x, thus $H(y|x) = 0$. To compute $H(x|y)$, we first need to apply a reconstruction operator that "invert" the windowing function f:

$$\hat{x} = R_{f^{-1}}(y) = \text{round}\{f^{-1}(y)\}\,. \tag{4.61}$$

Note that such an "inversion" will not fully reconstruct x because there is information loss in the forward conversion and all values are integers that create rounding errors. Therefore, the actual uncertainty of $H(x|y)$ roughly lies in the prediction residual between x and $\hat{x}$. Also note that x as an HDR image contains more information (and uncertainty) than

the LDR image y, thus $H(x) > H(y)$. Considering all the above factors, the actually computation simplifies to

$$\mathrm{NID}(x, y) \approx \frac{H(x - R_{f^{-1}}(y))}{H(x)} \,. \tag{4.62}$$

Combining this with Eq. (4.59), the actual optimization problem we would need to solve reduces to

$$f_{opt-\mathrm{NID}} = \arg\min_{f \in \mathcal{F}_{[l_l, l_u]}} \frac{H(x - R_{f^{-1}}(T_f(x)))}{H(x)} \,. \tag{4.63}$$

To fully solve Eq. (4.63) requires finding the best function in the function space $\mathcal{F}_{[l_l, l_u]}$, and is in general a difficult problem. Here we constrain the solutions to live in two families of parametric functions. In both cases, we express f as a linear combination of basis functions by

$$f(l) = \sum_{k=0}^{n-1} c_k \phi_k(l) = \phi_0(l) + \sum_{k=1}^{n-1} c_k \phi_k(l) \,, \tag{4.64}$$

where $c_0 = 1$ and $\phi_0(l)$ is a "ramp" function that corresponds to direct linear mapping given by

$$\phi_0(l) = \begin{cases} (l - l_l)/(l_u - l_l), & l_0 \le l \le l_u \\ 0, & \text{otherwise} \end{cases} \,. \tag{4.65}$$

The other basis functions are different for the two cases.

In the first case, we consider equipartition piecewise linear approximation, where we divide the full intensity interval $[l_l, l_u]$ into n subintervals $I_k = [l_{k-1}, l_k]$ for $1 \le k \le K$ of length $\Delta l = (l_u - l_l)/K$. The partition points are given by $l_k = l_l + k\Delta l$, $0 \le k \le n$, as such $l_l = l_0$ and $l_u = l_n$. The basis functions for piecewise linear approximation are "hat" function given by

$$\phi_k(l) = t\left(\frac{l - l_k}{\Delta l}\right) \,, \text{ for } k = 1, \cdots, n-1 \,, \tag{4.66}$$

where

$$t(l) = \begin{cases} 1 - |l|, & -1 \le l \le 1 \\ 0, & \text{otherwise} \end{cases} \,. \tag{4.67}$$

For the function $f(l)$ to be monotonically increasing, we need $0 \le \cdots \le f(l_{k-1}) \le f(l_k) \le \cdots \le 1$, which yields

$$0 \le \cdots \le c_{k-1} + \frac{k-1}{n} \le c_k + \frac{k}{n} \le \cdots \le 1 \,. \tag{4.68}$$

For example, in the case that $n = 3$, we can derive the following constraints on the solutions of the coefficients:

$$\begin{cases} c_1 \geq -\frac{1}{3}; \\ c_2 - c_1 \geq -\frac{1}{3}; \\ c_2 \leq \frac{1}{3}. \end{cases} \tag{4.69}$$

In the second case, we approximate the mapping function using the family of sine functions by

$$\phi_k(l) = \sin\left(\frac{k\pi(l - l_l)}{l_u - l_l}\right) \text{ for } l_l \leq l \leq l_u \text{ and } k = 1, 2, \cdots, n \tag{4.70}$$

To ensure that the mapping function $f(l)$ to be monotonically increasing, we would need $f'(l) \geq 0$. Plug Eq. (4.70) into Eq. (4.64) and take derivatives with respect to l and let it be no less than 0, we can obtain a set of constraints on the solutions of the coefficients. For example, in the case of $n = 3$, the constraints are given by

$$\begin{cases} c_1 + 2c_2 \geq -\frac{1}{\pi} \\ -c_1 + 2c_2 \geq -\frac{1}{\pi} \\ \frac{c_1^2}{16c_2} + 2c_2 \leq \frac{1}{\pi} \end{cases} \tag{4.71}$$

Having the aforementioned two types of parametric windowing functions, we can then search in the coefficient space $(c_1, c_2, \cdots, c_n)$ to solve for the optimization problem defined in Eq. (4.63 under the constraints on the coefficients (e.g., for the case of $n = 3$, the constraints are (4.69) for piecewise linear functions or (4.71) for sine basis functions). The search space is typically complex and to solve the problem, we would need to employ numerical optimization methods or resort to software optimization tools (e.g., Matlab *fmincon* function). Examples and detailed experimental results are presented in Section 4.3.3.

4.3.3 Results and Discussion

. To demonstrate the performance of the proposed scheme, we implement the parameter selection framework for natural images using the piecewise linear tone mapping, and continue by implementing sine basis windowing for medical images, using $n = 3$ for both mappings.

To demonstrate performance of the proposed scheme on medical images, sine basis windowing is used on a set of images in Digital Imaging and Communication in Medicine

(DICOM) format provided to our group by AGFA Healthcare Inc. DICOM images have window width and window center parameters which are preset values for optimal viewing of the intensity interval of interest embedded in their headers and traditional approaches in displaying such images use linear tone mapping operators which map the range of interest to the dynamic range of the display linearly [80].

We use real-world medical images in DICOM format to test the proposed method. In addition, we compare it with the most widely used image similarity/distortion measures in the literature, i.e., MSE and SSIM [3]. Note that the images before and after windowing have different dynamic ranges, and thus direct computation of MSE and SSIM is not feasible. Therefore, we search for the best windowing methods by optimizing MSE or SSIM between the original and reconstructed HDR images. These can be expressed as

$$f_{opt-\mathrm{MSE}} = \arg\min_{f \in \mathcal{F}_{[l_l, l_u]}} \mathrm{MSE}(x, R_{f^{-1}}(T_f(x))) \,, \tag{4.72}$$

$$f_{opt-\mathrm{SSIM}} = \arg\max_{f \in \mathcal{F}_{[l_l, l_u]}} \mathrm{SSIM}(x, R_{f^{-1}}(T_f(x))) \,. \tag{4.73}$$

In DICOM images, the window width and window center parameters are embedded in the image header, and thus the values of l_l and l_u are fixed. All windowing methods under test do not change these values, but attempt to find the best mapping functions with different optimization criteria.

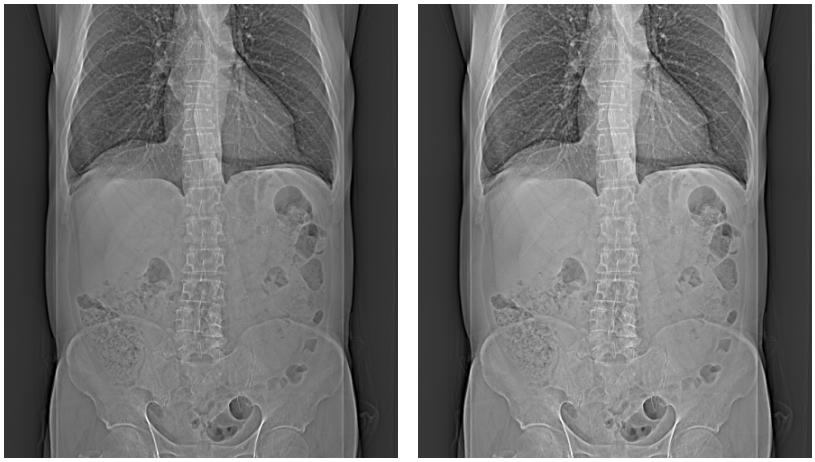
Figure 4.6 compares the images created using default DICOM direct linear windowing, and optimal MSE windowing, optimal SSIM windowing, and optimal NID windowing, all using sine basis. Their corresponding windowing functions are also given. It can be observed that the structural details are best preserved in NID optimal windowing image, which also appears to have higher contrast and better perceptual quality. To better illustrate how NID behaves in the parameter space, Figure 4.7 shows NID as a function of the c_1 and c_2 parameters in piecewise linear windowing, where brighter pixels indicate larger NID values. Sample images corresponding to different choices of c_1 and c_2 values are also given. It can be seen that the quality of the windowing results is quite sensitive to the selection of the parameters, and NID provides a useful tool to automatically select the best parameters that produces the best quality image.

Figure 4.8 shows the NID and MSE surfaces created using an exhaustive search over the domain of possible c_1 and c_2 parameters ($c_1 \leq c_2$) for piecewise linear tone mapping with the LDR tone mapped result image for each case. It can be observed that the image which is tone-mapped using parameters selected by the proposed scheme has enhanced quality. Figure 4.9 shows a top view of NID surface in the domain of possible c_1 and c_2

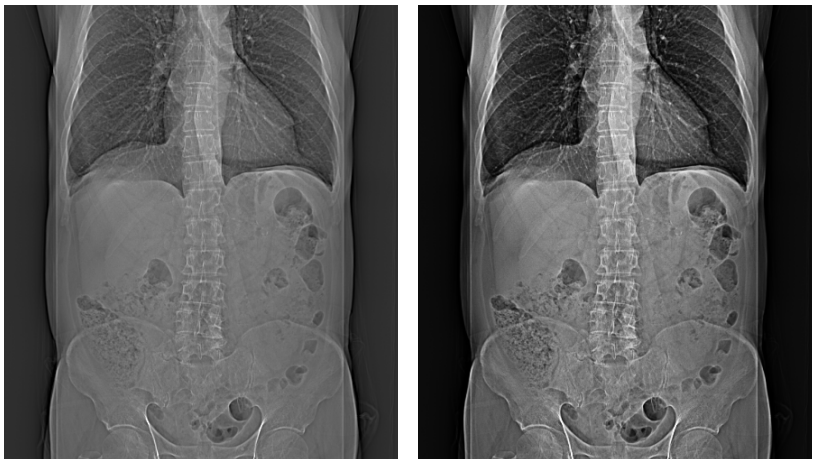
parameters. It can be observed that as NID decreases, the quality of the tone mapped image increases, and the best quality is achieved at the global minimum of this surface. Figure 4.10 (b) shows histogram of the high dynamic range desk image with the tone mapping operators selected by NID and MSE schemes. The tone mapped images using the selected parameters by each scheme are shown in (c) and (d) respectively. The quality of the image tone mapped using the NID parameters is notably enhanced compared to that of MSE.

Figure 4.11 (a) shows the result of sine basis windowing using parameters selected by the NID scheme in the range of $[-125 : 225]$. It is notable that the image has enhanced quality compared to the sine basis windowing using MSE scheme (b) and linear tone-mapping (c), and the details in the chest, spinal cord and intestine are more discernable. All three operators are shown in Figure 4.11 (d). Figure 4.12 (c) shows the enhanced quality of the tone mapped image when tone mapping parameters are selected according to the proposed scheme.

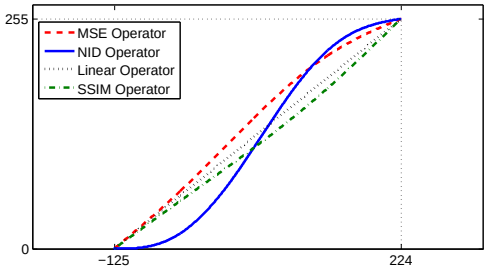
The major computational cost of the proposed method lies in the search procedure in the parameter space. In our experiment using a computer with a Core-i5 CPU running at 2.27Ghz, it takes about 250 seconds for our unoptimized program to find the optimal NID windowing operator for an 512×512 image using an exhaustive search method on a grid of 0.02×0.02 precision. The time can be largely shortened by using advanced optimization method. For example, an MATLAB *fmincon* function that employs gradient optimization and trust-region-reflective algorithm reduces the search time to about 13 seconds.



(a) (b)



(c) (d)



(e)

Figure 4.6: Adaptive windowing using (a) Linear, (b) MSE, (c) SSIM, and (d) NID optimization of sine basis operators. (e) Corresponding optimal windowing function.

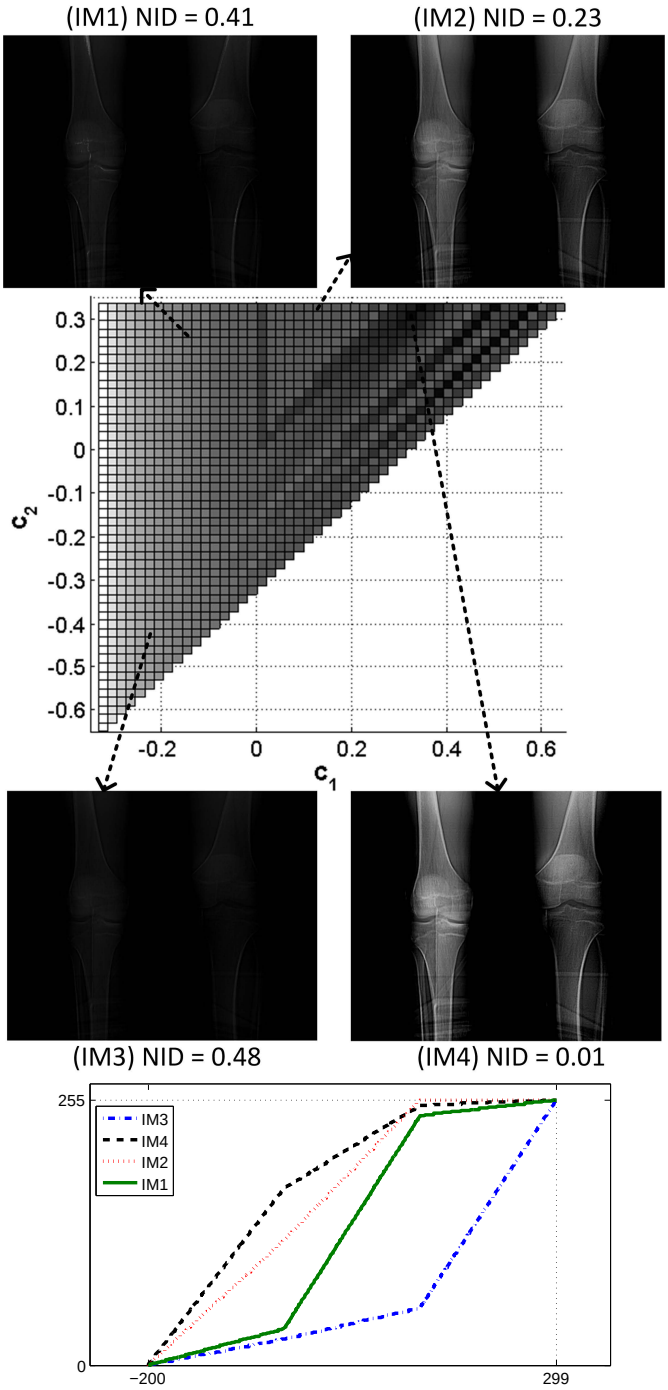
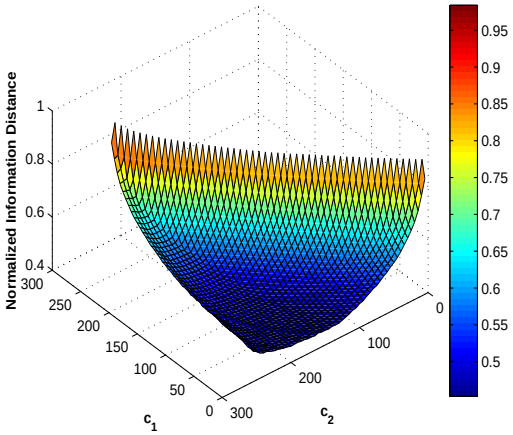
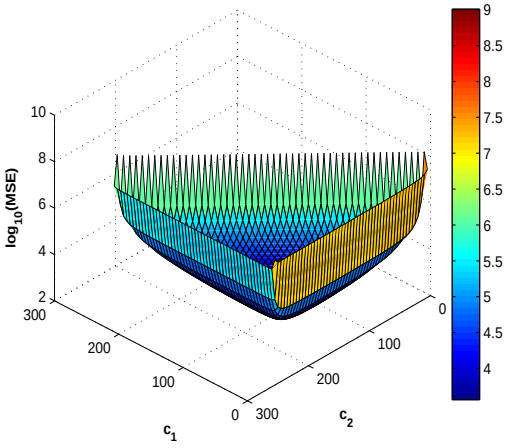


Figure 4.7: NID as a function of the parameters in piecewise linear windowing operator. (IM1)-(IM4): images correspond to 4 different options of c_1 and c_2 parameters, which result in different image quality and NID values.



(a)



(b)

(c)

(d)

Figure 4.8: (a) Normalized Information Distance (NID) surface over the domain of possible c_1 and c_2 with $c_1^* = 10$, $c_2^* = 175$, (b) Mean Square Error (MSE) in logarithmic scale with $c_1^* = 90$, $c_2^* = 215$, (c) Tone mapped image using parameters selected by NID scheme (d) Tone mapped image using parameters selected by MSE scheme

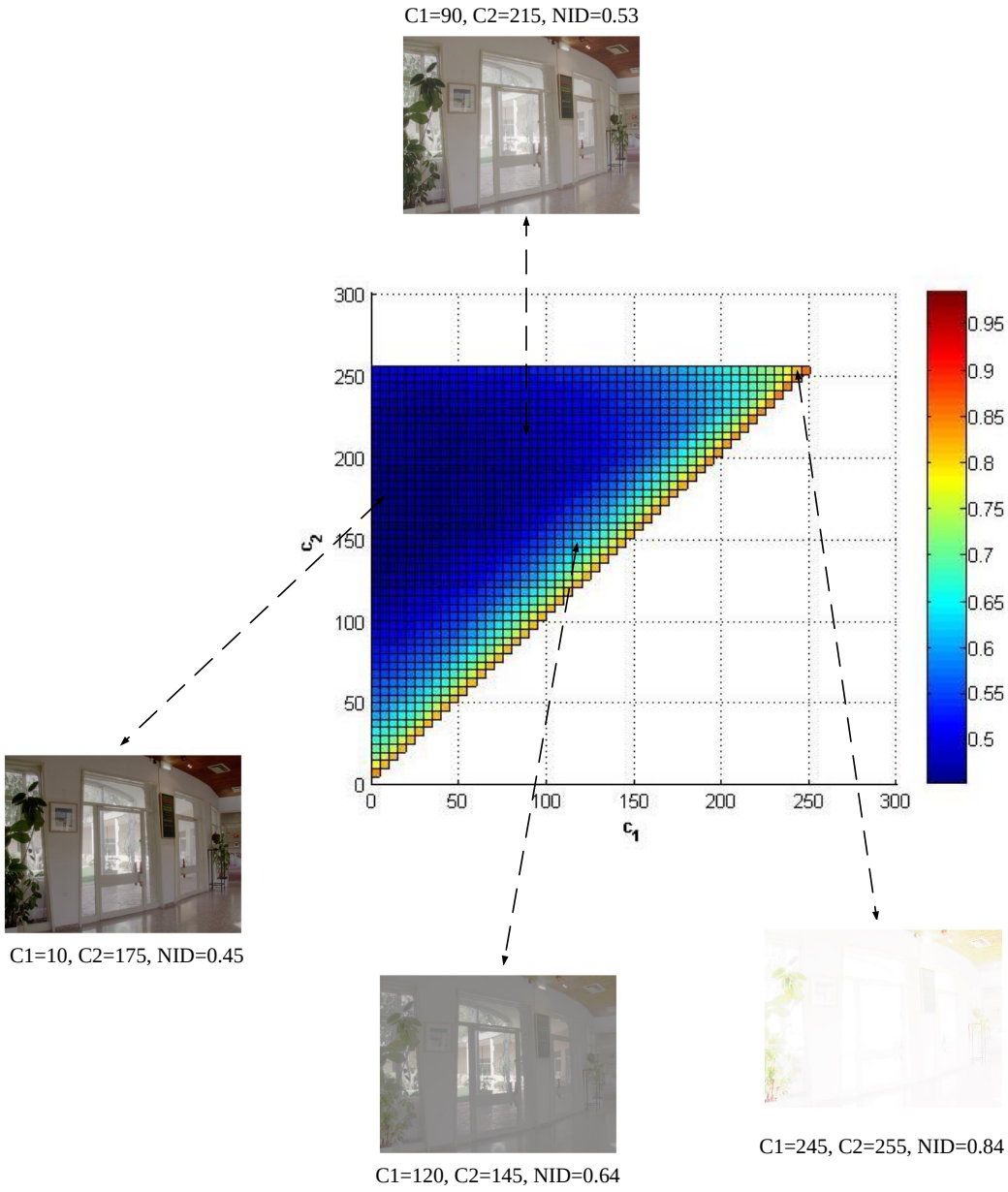
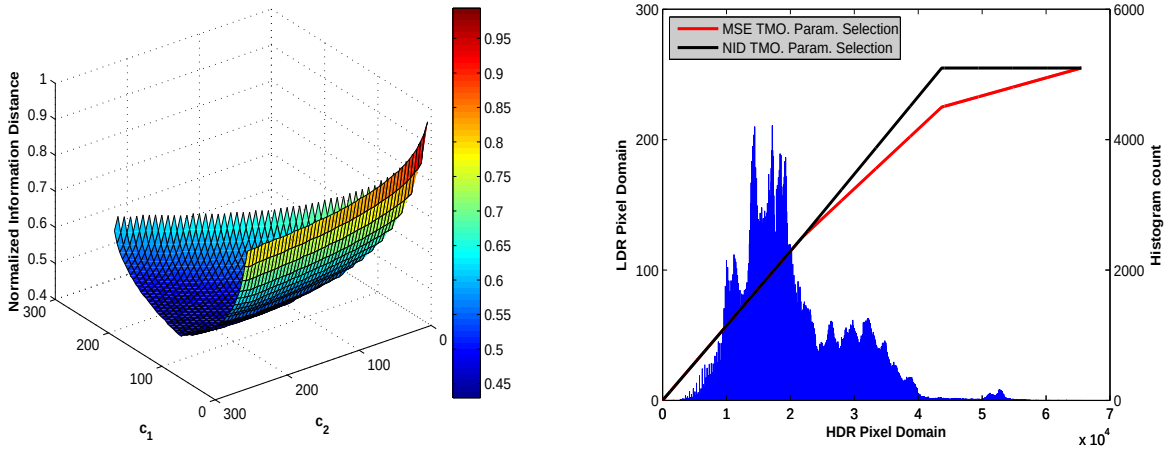


Figure 4.9: NID as a function of the parameters in piecewise linear windowing operator, with piecewise tone-mapping carried out at four different locations



(a) (b)

(c) (d)

Figure 4.10: (a) Surface of NID with $c_1^* = 125$ and $c_2^* = 255$, (b) Histogram of HDR image with tone mapping operators using NID and MSE parameters (c) Tone mapped image using parameters selected by NID scheme, (d) Tone mapped image using parameters selected by MSE scheme. Image courtesy of AGFA Healthcare Inc.

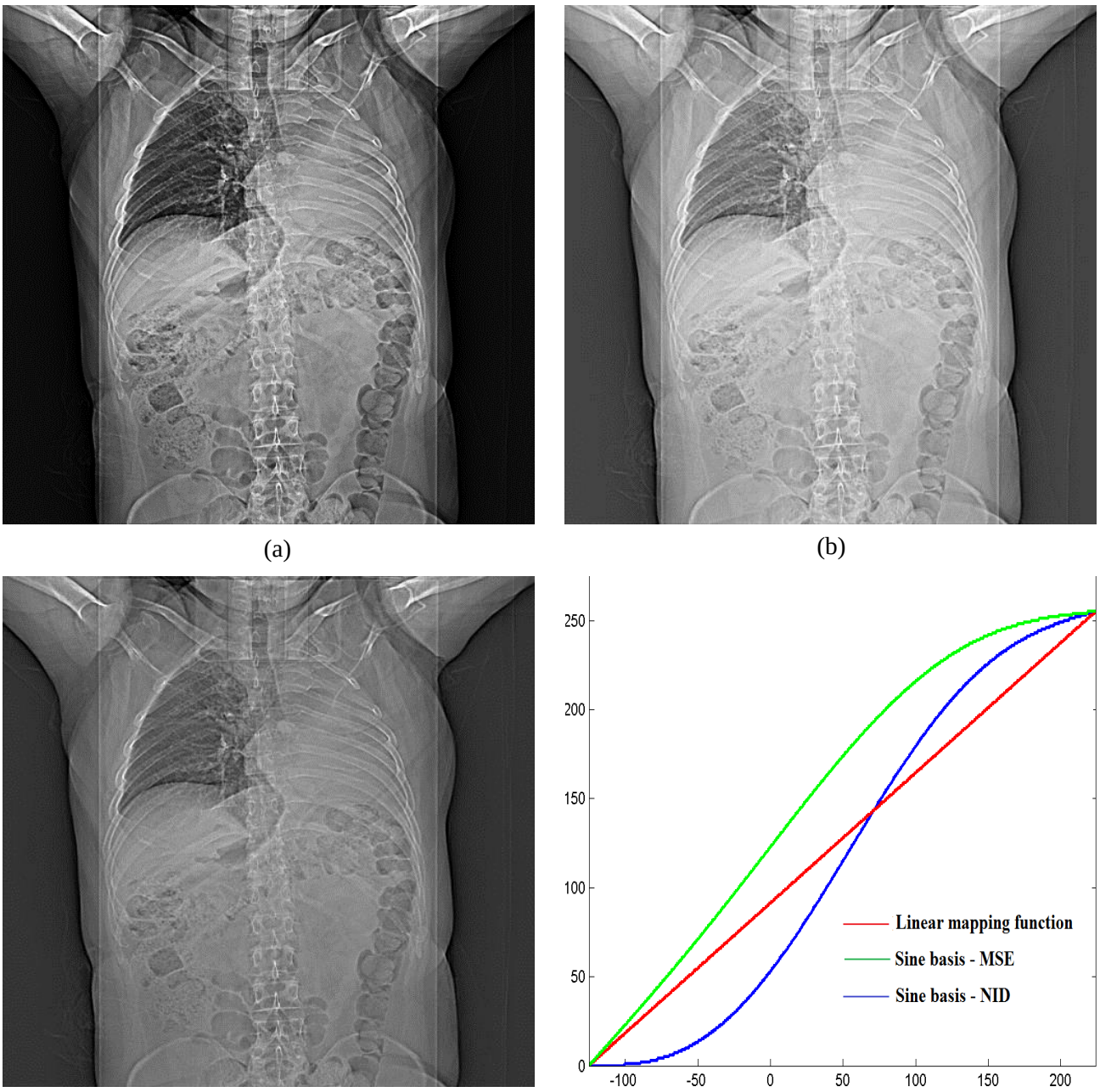
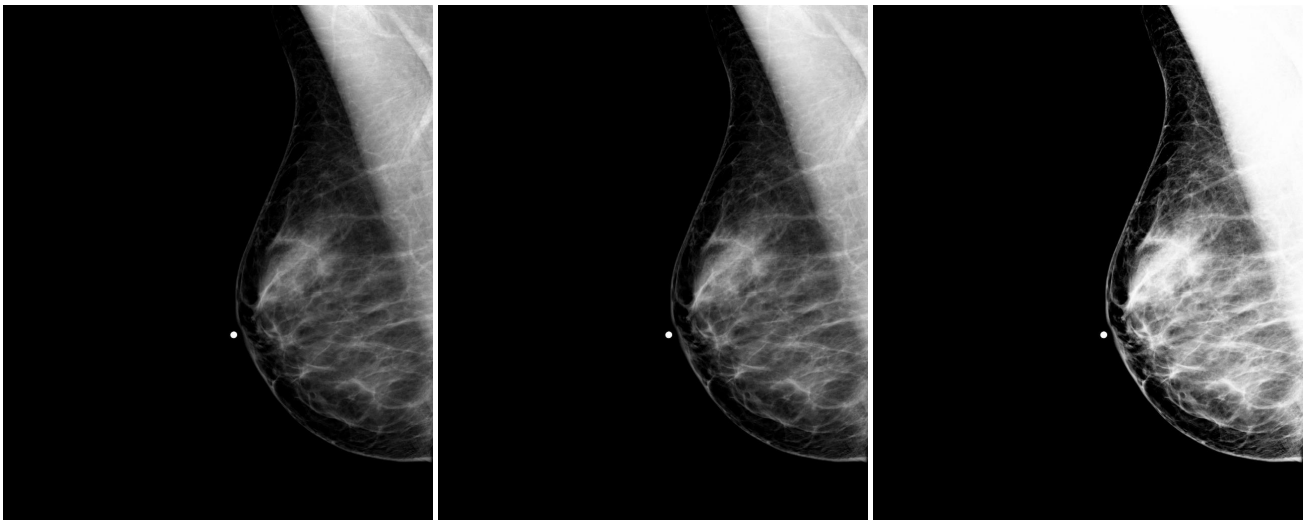


Figure 4.11: Chest image : (a) Adaptive Sine basis windowing with NID parameter selection, (b) Adaptive Sine basis windowing with MSE parameter selection, (c) Linear tone mapping. Image courtesy of AGFA Healthcare Inc.



(a) (b) (c)

Figure 4.12: Breast mammogram image: (a) Linear tone mapping, (b) Sine basis windowing with MSE parameter selection, (c) Sine basis windowing with NID parameter selection. Image courtesy of AGFA Healthcare Inc.

(a) (b) (c)

Figure 4.13: Bone image: (a) Linear tone mapping, (b) Sine basis windowing with MSE parameter selection, (c) Sine basis windowing with NID parameter selection. Image courtesy of AGFA Healthcare Inc.

Figure 4.14: Torso image: (a) Linear tone mapping, (b) Sine basis windowing with MSE parameter selection, (c) Sine basis windowing with NID parameter selection. Image courtesy of AGFA Healthcare Inc.

4.4 Summary

In this chapter, we used Shannon entropy to approximate Kolmogorov complexity of the images and compute NID. The Shannon entropy based approximation of NID was then used to develop two frameworks for image distortion analysis, and parameter selection for tone-mapping of HDR images.

In the first framework, we extended the application of Kolmogorov complexity and NID to image distortion and quality assessment by employing a wavelet domain Gaussian scale mixture model of images and by estimating Kolmogorov complexity based on Shannon entropy approach. We showed that the resulting image similarity measure is competitive with respect to state-of-the-art image quality assessment algorithms when tested using publicly-available subject-rated large-scale image databases. The proposed method draws some connections between the theories of Kolmogorov Complexity, NID, Shannon entropy, statistical image modeling, perceptual modeling, and real-world image processing applications.

In the second framework, an NID-motivated criterion in the optimal design of windowing operators for the visualization of HDR medical and natural images on standard displays. A Shannon entropy based approximation was made that converts the uncomputable NID minimization problem into a practical algorithm that optimizes parametric windowing operators. Experiments using both medical and natural images demonstrate that the proposed method provides a powerful tool in finding the best parametric windowing functions and tuning tone-mapping operators, which create images with higher contrast and more visible structural details. In the future, the proposed method may be extended to higher order parametric windowing functions. The promising results obtained in this work also inspires us to explore more applications of Kolmogorov complexity in the field of medical image processing.

Chapter 5

Information Distance Based Feature Extraction with Applications

In this chapter we introduce a set of information distance based features which can be used to predict perceived modification to images. A Support Vector Regression (SVR) algorithm is trained using a human observer rated database of original and edited images, and it is shown that the predicted scores are highly consistent with mean observer ratings. The algorithm can be used to predict a perceptual score for new images added to the database.

5.1 Introduction

Editors of fashion magazines have long been criticized for publishing altered images of celebrities and fashion models, often making them look flawlessly attractive and natural. Such modifications, while appealing to the average reader, have been linked to mental and physical disorders in younger generations, and have given rise to calls by politicians in many countries such as UK, France, and Israel to provide labels for altered images of this kind [81–83]. In similar cases, photoshop scandals of celebrities on the Internet have shocked Hollywood community, and have raised the question of how much editing is considered too much in this business.

Inspired by the recurring photoshop scandals in fashion industry, and a dire need for an automatic rating system of photoshopped images for fashion industry, Farid and Kee

proposed an algorithm which represents human perception of photo manipulations [84]. In this chapter, we intend to address this problem by introducing a set of information theoretic features which are capable of predicting a perceptual score for a pair of original and modified images using a SVR tool, and show that the predicted score is correlated with the human perception of modifications.

Information distance features describe average number of bits required to transform certain features of original image to the edited image and vice versa. While these features are capable of describing the modifications in detail, in most cases, HVS does not distinguish subtle modifications that might translate into large number of bits when quantified using the features. A machine learning algorithm built upon such features an trained by a human observer rated database can be used to predict human observer ratings. In effect, the machine learning algorithm gives the information distance features the flexibility they require to predict scores that are not only consistent with the amount of modifications, but also consider limits of the HVS in distinguishing these modifications.

5.2 Perceptual metric for photo retouching by Eric Kee and Hany Farid

Kee and Farid proposed a measure that quantifies the modifications perceptually [84]. The proposed algorithm is based on using eight summary statistics features which are extracted from 468 pair of original and altered images rated by human observers to train a Support Vector Regression (SVR) and predict scores with minimum distortion and maximum correlation compared to Mean Opinion Scores (MOS). The features are divided into two groups, and are intended to measure the impact of geometric modifications and photometric modifications on the modified picture.

In order to train the SVR, a diverse set of 468 original and retouched images was collected from online sources and 390 Users were paid to rate these images on the scale of 1 to 5 on Amazon's Mechanical Turk website. Each user was shown a total of 50 images, including a random set of five images three times to measure consistency of responses.

In this section, we briefly review this algorithm, and its implementation by two groups of independent researchers [85, 86] who report moderate success in replicating the original paper's result using the same training database as of original paper. Due to the fact that the original paper's code has not been published by the authors, we use our own implementation of the algorithm as a baseline for the purpose of comparison in future sections.

The first set of features used in the algorithm are chosen to measure the impact of geometric modifications on the images. Geometric distortions between the original and edited images are modeled using a 6-parameter affine model along with two extra parameters to model the brightness and contrast changes of the images [42]. Assuming that f_a and f_b are local regions of the luminance channel of the original and edited images we have:

$$cf_a(x,y) + b = f_b(m_1x + m_2y + t_x, m_3x + m_4y + t_y) \tag{5.1}$$

where m_i terms are affine parameters, and c and b are contrast and luminance change parameters. Using this registration, a two dimensional vector field of geometric transformations can be built as follows [42]:

$$\overrightarrow{v}(x,y) = \begin{pmatrix} m_1x + m_2y + t_x - x \\ m_3x + m_4y + t_y - y \end{pmatrix}. \tag{5.2}$$

The mean and standard deviation of magnitude of the vector field in 5.2 projected onto the gradient vector field of the image, and computed over the face and body region are taken to be the geometric features.

The second set of features used in the algorithm are called photometric features, and are designed to capture the effect of sharpening or blurring, and other structural distortions that might occur during modification of images. After initial alignment of the images using the vector field in 5.2, the aligned before image ($\tilde{f}_b$) and the after image are used to estimate a 9×9 linear filter h:

$$f_a(x,y) = h(x,y) * \tilde{f}_b(x,y), \tag{5.3}$$

where $*$ is the convolution operator, and the filter h is computed by applying a conjugate gradient descent optimization with Tikhonov regularization [84]. Frequency response of this filter is then used to compute a measure, D, which quantifies the measure of sharpening or blurring in images as follows:

$$D = \sum_{\omega} |\tilde{F}_b(\omega)|\omega - \sum_{\omega} |H(\omega)\tilde{F}_b(\omega)|\omega, \tag{5.4}$$

where $H(\omega)$ and $\tilde{F}_b(\omega)$ are unit sum normalized frequency responses of the filter h and warped before image [84]. Mean and standard deviation of D are taken to be blurring features in this algorithm.

In order to quantify the effect of photometric modifications that are not captured by

previous steps, SSIM is calculated on the luminance channel of the warped before image and the after image, and mean and standard deviation of SSIM are used as two features which embody basic blurring, sharpening and special effects by various photoshop filters [84].

These features are then scaled into the range [-1,1] and are fed into a nu-SVR tool with Gaussian radial basis kernel [87] along with human observer ratings. The SVR parameters γ and c are then fine tuned using a 2D grid search to maximize correlation coefficient of each training set. Performance of the algorithm is tested using leave-one-out cross-validation by training the SVR using 467 images and predicting the score for the remaining image. This process is repeated 468 times and the score for each image is predicted. The authors report 80% correlation between the predicted scores and the mean observer scores with a mean/median absolute error of 0.3/0.24, maximum absolute error of 1.19 and standard deviation of 0.249.

5.2.1 Implementations of the Algorithm

Due to the fact that the authors do not publish or share the code of their implementation, we use two reimplementations by two groups of Stanford University researchers and create our own implementation for the purpose of having a benchmark for comparison.

The first group report little success in reimplementing the algorithm, and apply the method to a subset of images in the database due to computational limitations. Their result has less than six percent correlation between the predicted scores and mean observer ratings using 5-fold cross validation [85].

The second group develop a dataset of 137 images and apply a modified version of the original algorithm to their dataset. They report 65.2 percent correlation between the predicted scores and mean observer ratings for their dataset [86].

Both groups have published the Matlab code of their implementations online, and neither has been able to fully duplicate the original results. Applying their published code to the original dataset using leave-one-out cross-validation results in predicted scores that have a correlation of less than 30% with the mean observer rating in both cases. Using some parts of the published codes from both groups, we implemented our version of the algorithm, and applied it to the original 468 dataset image. The main modification carried out to both codes is to enforce symmetry in estimating the filter h using gradient descent method as proposed in the original algorithm, and some other details such as projecting the vector fields onto a gradient of luminance channel before computing the related features. Our implementation results' predicted scores have 57.4% correlation with the mean observer ratings, with mean/median absolute error of 0.42/0.37, maximum error

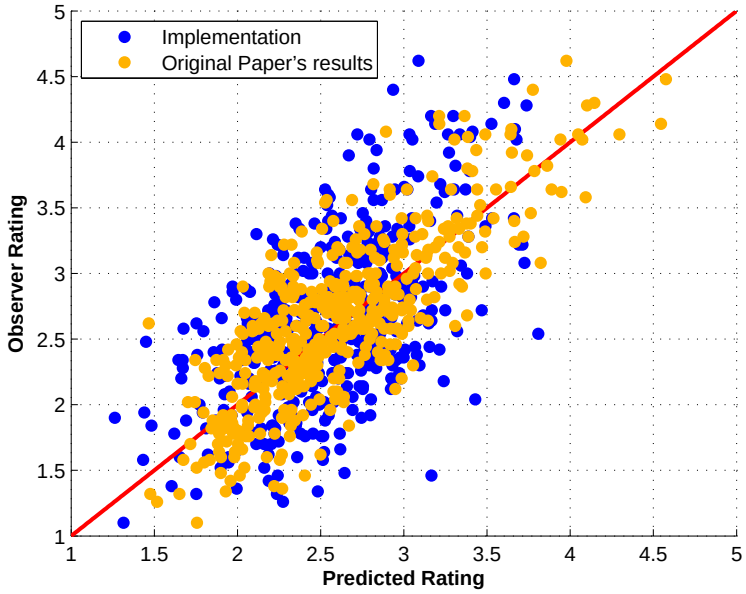


Figure 5.1: Our Implementation results as opposed to the results in [84]

of 1.7063 and standard deviation of 0.3. Figure 5.1 shows the results of our implementation of the algorithm superimposed on the reported results in the original paper [84].

5.3 Normalized Information Distance for Photo retouching

In our first attempt to apply information theoretic approaches to the problem of perceptual retouching metric, we use the two implementations of Normalized Conditional Compression Distance (NCCD) presented in sections 3.2 and 3.4 of chapter 3. The first implementation uses CALIC compressor to estimate non-computable Kolmogorov complexity terms in Normalized Information Distance (NID). The second uses H.264 video encoder to approximate conditional Kolmogorov complexity terms in NID effectively. Table 5.1 shows the Pearson linear correlation coefficient and Spearman ranking correlation coefficients of the two implementations of NCCD of a subset of 48 images of the dataset with the mean observer ratings. While the results are promising for a non-supervised method, they are

Table 5.1: Normalized Information Distance (NID) Predictions

Model	PLCC	SRCC
NCCD (H.264)	0.3518	0.3621
NCCD (CALIC)	0.1811	0.1901

not competitive with the reported results in the original paper, and also require significant computation time compared to supervised methods.

5.4 Information Distance based features

In this section, we introduce a set of information distance based features which will be used for training the SVR tool using human observer data. We start by aligning the before and after images using the registration technique in [42]. The aligned images are then transformed into the La^*b^* color space.

The first set of features are La^*b^* differential entropy features which are designed to quantify modifications to La^*b^* pixel domain such as color change, noise and structural changes not captured by registration. These features are defined for face, body and hair separately by subtracting each channel of before and after images directly and inserting it into column vectors ΔL, Δa^*, and Δb^*. The vectors are then inserted in the columns of a matrix Δ:

$$\Delta = \begin{pmatrix} \Delta L^T \\ \Delta a^{*T} \\ \Delta b^{*T} \end{pmatrix}^{\mathrm{T}}. \tag{5.5}$$

Since the differential entropy of joint Gaussian distribution is a higher bound for differential entropy of all distributions [67], we can assume that the observation vectors follow a joint Gaussian distribution without loss of generality. In this case the differential entropy of joint Gaussian distribution $\Delta L, \Delta a^*, \Delta b^* \sim N(\mu_\Delta, \Sigma_\Delta)$ is:

$$h(\Delta L, \Delta a^*, \Delta b^*) = \frac{1}{2}\log(2\pi e)^3|\Sigma_\Delta|, \tag{5.6}$$

where Σ represents covariance matrix of Δ, and $|\Sigma_\Delta|$ represents its determinant. We take $\log(|\Sigma_\Delta|)$ computed over regions of face, hair and body as our La^*b^* entropy features.

The second set of features are selected to measure the effect of geometric modifications. Similar to the entropy features, these features are defined for face, hair, and body regions,

and are computed using motion vectors in the direction of X and Y extracted from vector field 5.2. The motion vectors are inserted into column vectors of a matrix M such that:

$$\mathrm{M} = \begin{pmatrix} M_x^T \\ M_y^T \end{pmatrix}^{\mathrm{T}}. \tag{5.7}$$

Assuming that motion vectors in X and Y directions are following a joint Gaussian distribution, $M_x, M_y \sim N(\mu_M, \Sigma_M)$ we have:

$$h(M_x, M_y) = \frac{1}{2}\log(2\pi e)^2|\Sigma_M|, \tag{5.8}$$

where $|\Sigma_M|$ is the determinant of covariance of M, and we take $\log(|\Sigma_M|)$ computed over face, body, and hair regions as our geometric features.

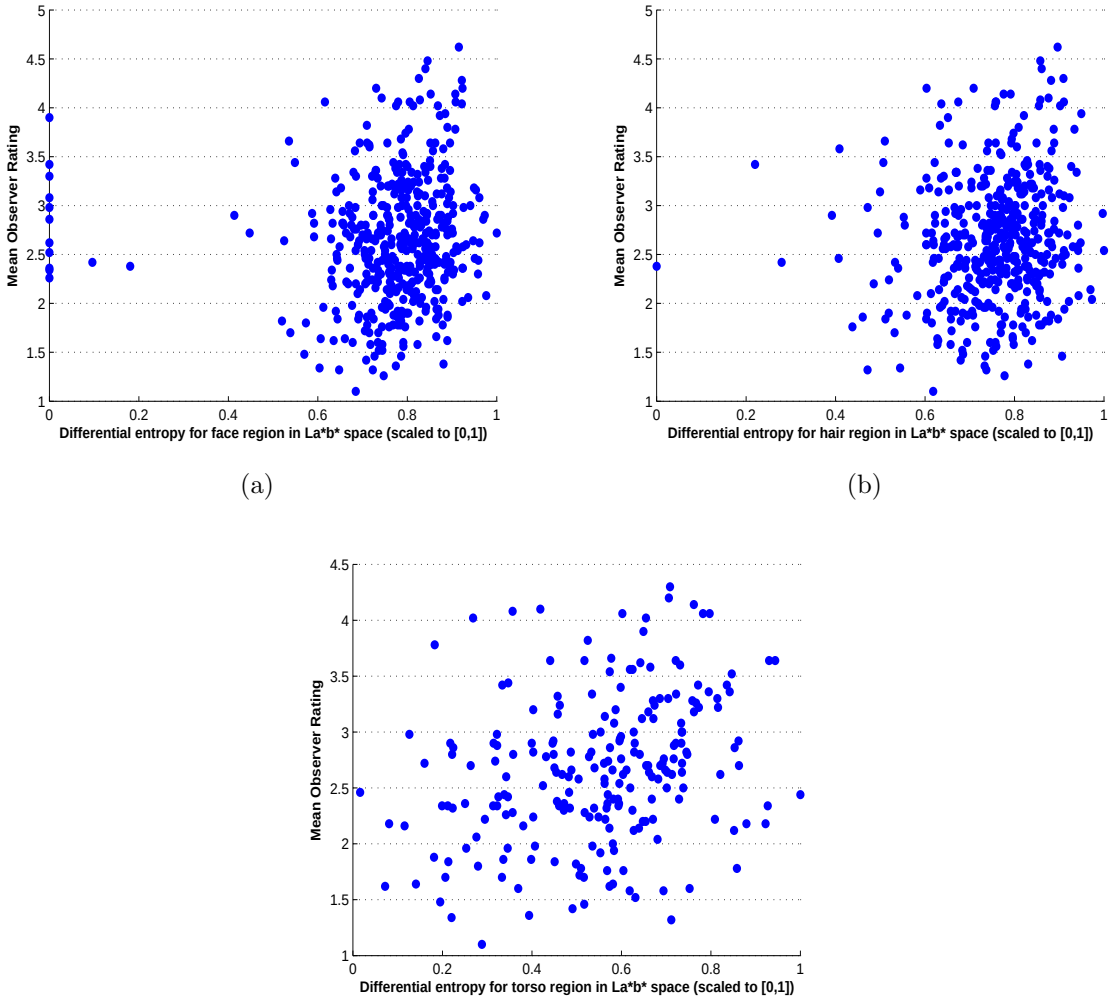
5.5 Results and Discussion

Both sets of features are carefully selected to quantify a higher bound on the average number of bits required to describe the modifications carried out on the original image to reach the edited image. Figures 5.2 and 5.3 shows these features extracted for 468 images of the original paper's dataset and plotted against mean observer ratings and table 5.2 shows the correlation of these features with each other as well as with the mean observer ratings. While these features show some correlation with the mean observer ratings, it must be taken into consideration that HVS might not notice some of the subtle changes in the bit-plane of images. In fact, as far as HVS is concerned, some of these bits are not noticed at all.

A pre-trained SVR tool is then used to predict scores that are consistent with the modifications that the features quantify, taking into consideration the limits of HVS. Similar to the original paper, we used a nu-SVR with Gaussian kernel implemented in LIBSVM [87]. The parameters, γ and c were selected to maximize correlation of predicted scores with mean observer ratings in both leave-one-out cross-validation and 1-fold cross-validation (train-all/test-all) schemes by an exhaustive search over a range of possible values. Figure 5.4 shows a top view of correlation values over the domain of parameters for both leave-one-out scheme and train-all/test all. Figure 5.5 shows the result of training the SVR with all features and labels and then testing the same features to predict perceptual scores. The predicted scores have 76 percent correlation with mean observer ratings in this case. Figure 5.6 shows the result of leave-one-out cross validation of the method on (a)

Table 5.2: Correlation of features with MOS

Features	Feat. 1	Feat. 2	Feat. 3	Feat. 4	Feat. 5	Feat. 6	MOS
Differential Entropy Face	1.0000	-0.4303	0.7351	0.7551	0.4814	-0.4367	0.0567
Differential Entropy Torso	-0.4304	1.0000	-0.2243	-0.5511	-0.2299	0.9549	0.0487
Differential Entropy Hair	0.7351	-0.2243	1.0000	0.5041	0.6950	-0.2210	0.0789
Correlation XY vector Face	0.7551	-0.5511	0.5041	1.0000	0.5811	-0.5198	0.2711
Correlation XY vector Hair	0.4814	-0.2299	0.6950	0.5811	1.0000	-0.1755	0.2884
Correlation XY vector Torso	-0.4367	0.9549	-0.2210	-0.5198	-0.1755	1.0000	0.1203
Mean Opinion Score (MOS)	0.0567	0.0487	0.0789	0.2711	0.2884	0.1203	1.0000

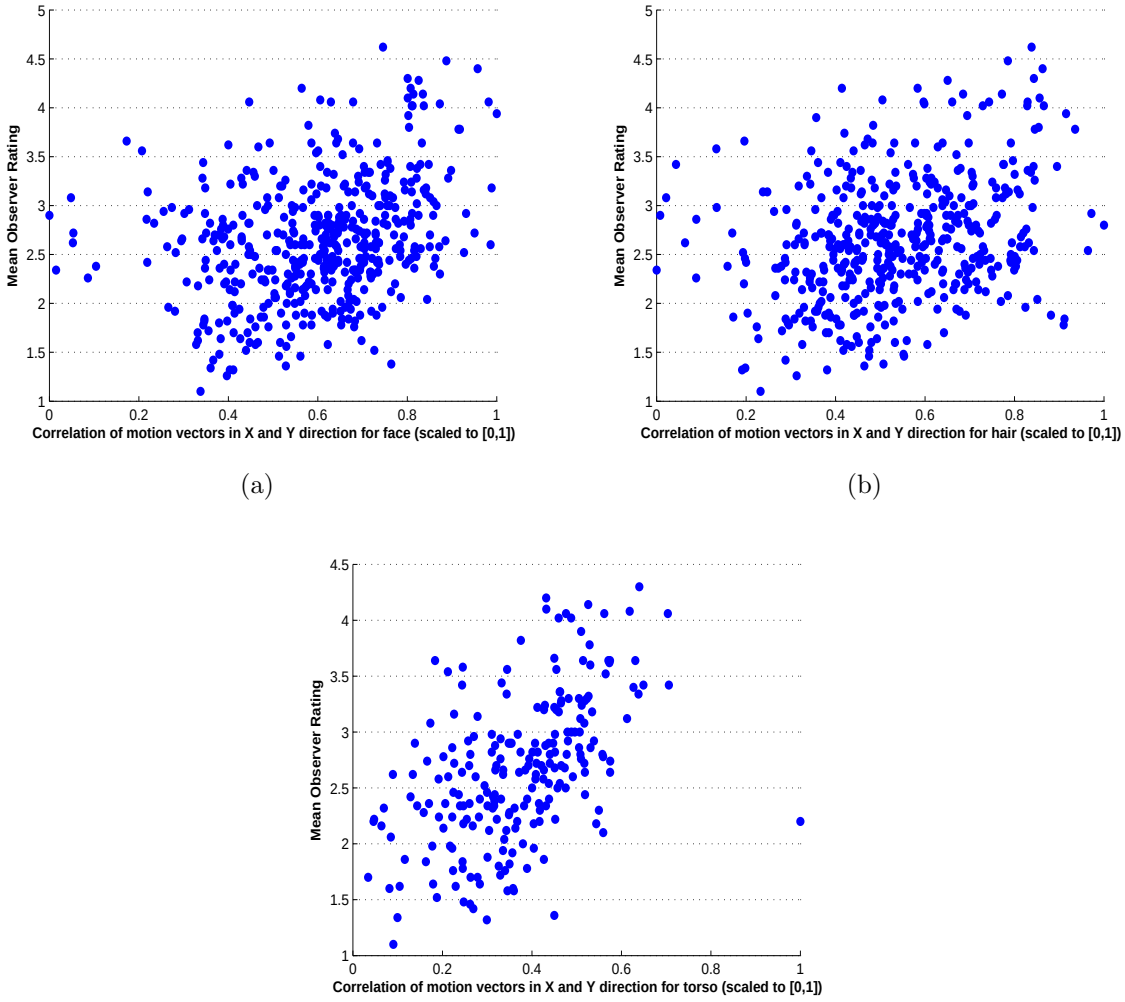


(a) (b)

(c)

Figure 5.2: Entropy Features Vs. MOS: (a) Differential entropy of face, (b) Differential entropy of hair, (c) Differential entropy of torso

234 and (b) 100 randomly selected images. The correlation between predicted scores and mean observer ratings in this case is 72% and 74% respectively. The algorithm was also verified using repeated random sub-sampling validation, by randomly dividing the dataset into training and testing data, and averaging correlation of the predicted scores with mean observer scores for 1000 times. The average correlation of the predicted scores with MOS



(a) (b)

(c)

Figure 5.3: Geometric Features Vs. MOS: (a) entropy of motion vectors for face, (b) entropy of motion vectors for hair, (c) entropy of motion vectors for torso

in this case is 65% with standard deviation of 13%.

In comparison, the prediction of our method based on information distance features performs significantly better than our reimplementation of the algorithm described in [84], and the predicted scores are similar to the results reported in [84]. Figure 5.7 shows

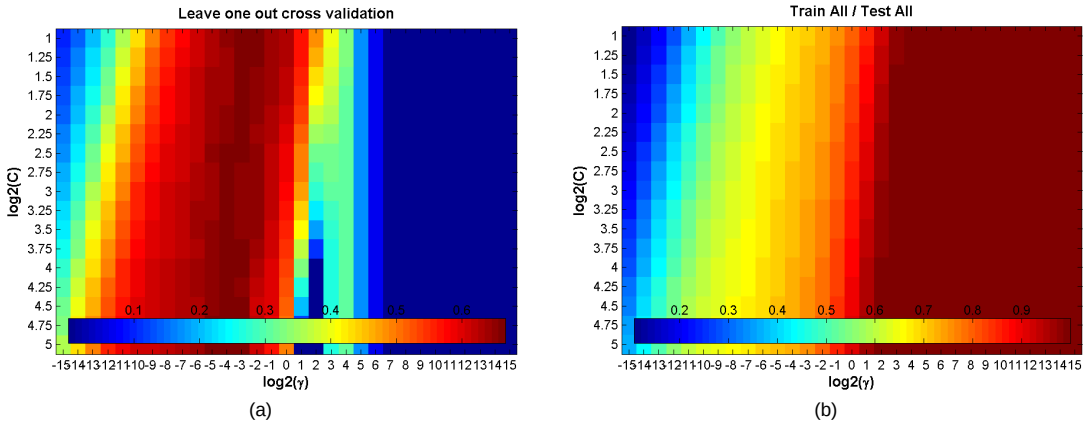


Figure 5.4: Parameter selection: (a) Leave-one-out scheme, (b) Train all / test all

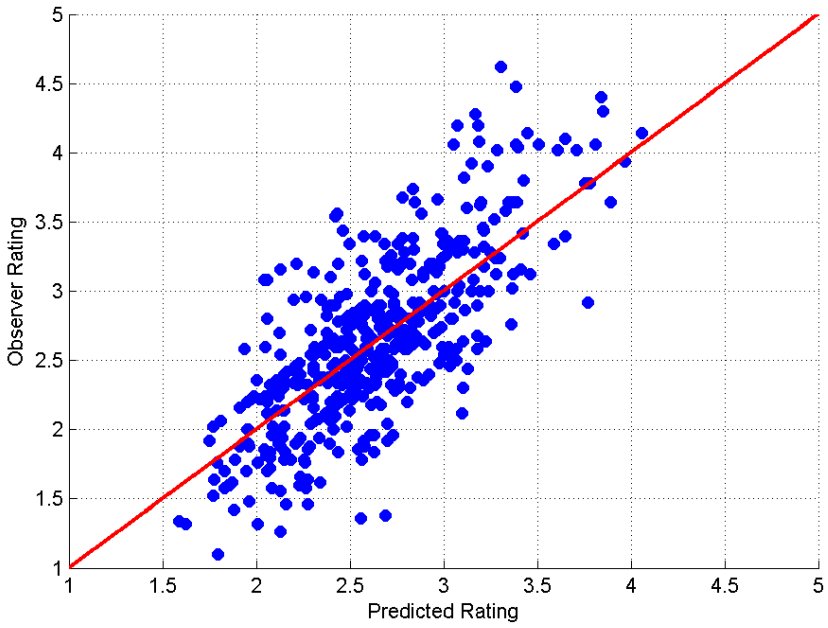
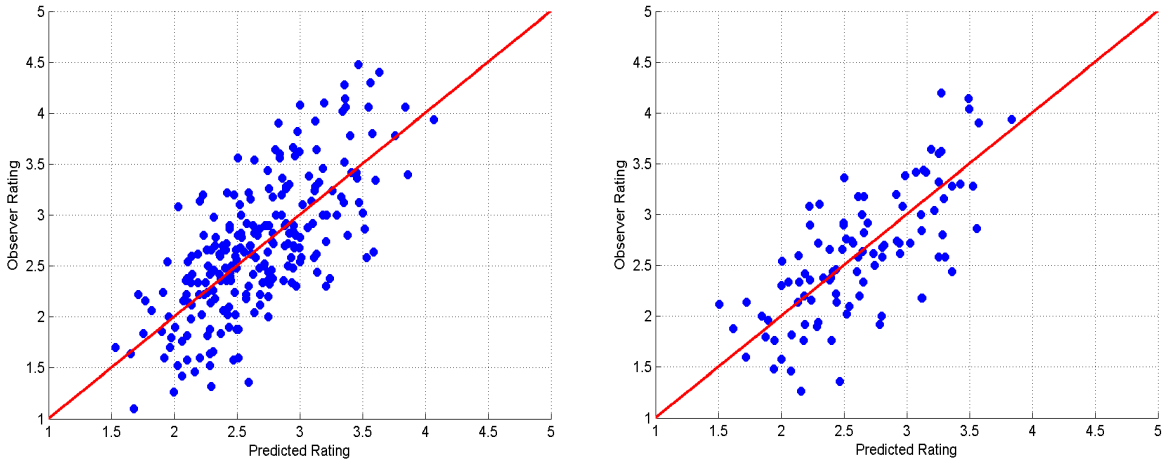


Figure 5.5: Results by training SVR on the whole data and testing it on the same set

predicted scores using information distance features as well as the original paper's scores for 18 images of the dataset. The images are ranked according to their mean observer score



(a) (b)

Figure 5.6: Results of leave-one-out cross-validation on a set of randomly selected images: (a) 234 images with 72% correlation (b) 100 images with 74% correlation

in the ascending order. The error bars represent standard deviation of the observer scores. It is evident that the proposed method and the method in [84] are both overcompensating when observer scores are in the lower range and undercompensating when the observer scores are in the higher range. The result of both methods stay in the range of standard deviation of the observer ratings. The distribution of mean observer scores and predicted scores by individual observer ratings are shown in Figure 5.8. As expected, the density of individual observer ratings are higher in the intervals where the mean observer ratings and predicted ratings are consistent, which shows the soundness of both the subjective test and the predicted results. Figure 5.9 shows examples of the original and edited images with their corresponding mean observer scores and predicted scores. It can be seen that the proposed method is capable of predicting the scores with higher accuracy when the nature of modifications are geometric, and the performance degrades when more complicated editing filters have been used.

5.6 Summary

In this chapter, we presented a set of information theoretic distance based which quantify a higher bound for the average number of bits required to describe modifications carried

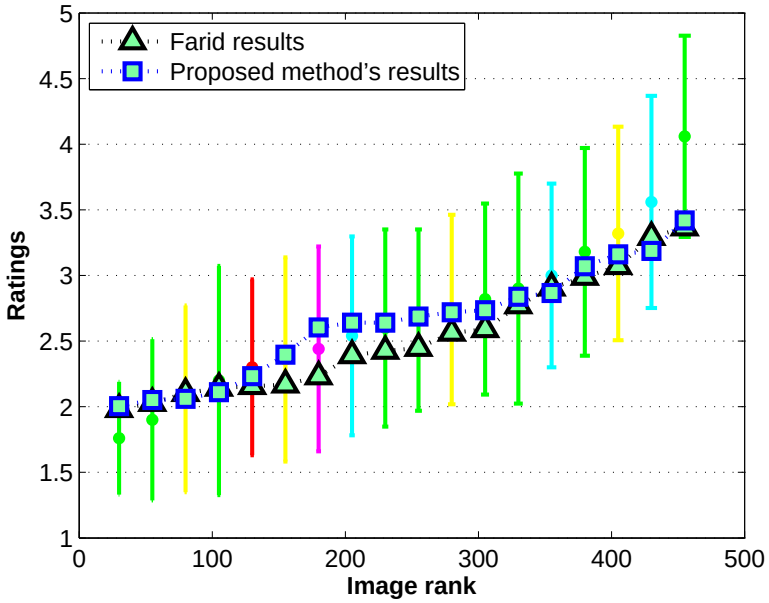


Figure 5.7: Farid's results Vs. proposed method's results

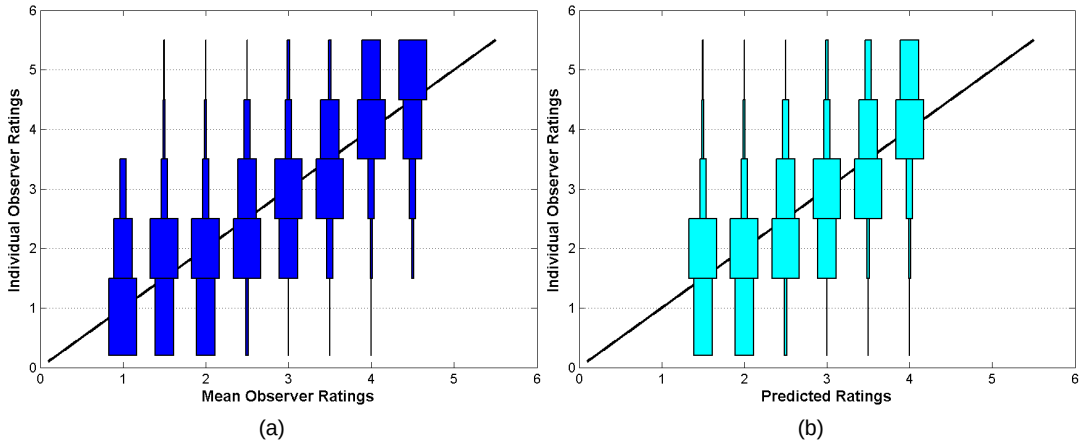


Figure 5.8: Distributions: (a) MOS by individual observer ratings, (b) Predicted Ratings by individual observer ratings

Figure 5.9: Examples of Before and After images with mean observer ratings and predicted ratings

out on images. The features were tested on an observer rated dataset of original and modified images, and it was shown that the features can be used to predict perceptual scores for retouched images. Compared with the existing method, the proposed method has significantly lower complexity in feature extraction and competitive performance in predicting perceptual scores.

Chapter 6

Concluding Remarks and Future Works

This chapter summarizes the results and contributions of the thesis and outlines possible directions for future research in using information distances in perceptual image similarity and image quality assessment.

6.1 Concluding Remarks

In this thesis we introduced information distance framework into the field of image similarity and image quality assessment, The major contributions of the thesis are as follows:

- A practical framework for the approximation of NID is developed. Normalized Conditional Compression Distance (NCCD), is a flexible and expandable framework that uses a list of possible transformations that convert one image to another along with an image compressor to approximate the non-computable Kolmogorov complexity terms in NID, and in effect finds the simplest transformation of one image to another in accordance with visual sensory coding principle. Two different implementations of NCCD based on image compression and video compression are proposed and it is shown that the NID based similarity framework has a wide range of applicability in problems where most other image similarity measures fail.
- A NID based perceptual distortion analysis method is developed based on approximating Kolmogorov complexity using Shannon's entropy, and it is shown that the

method is competitive with state-of-the-art image quality assessment algorithms when tested using subject-rated image databases.

- A parameter selection scheme based on entropy approximation of NID is proposed for tone-mapping of High Dynamic Range (HDR) images, and it is shown that the Low Dynamic Range (LDR) images tone-mapped using this scheme have enhanced quality.

- A set of information distance based features which quantify a higher bound for the average number of bits required to describe modifications carried out on a pair of original and edited images is proposed, and it is shown that the features can be used to predict perceptual scores for quantifying the impact of modifications carried out on images when used along with a Support Vector Regression (SVR) tool that is trained by human observer data.

6.2 Future Work

6.2.1 Refinement of The Perceptual Quality Metric

In the motivation section of this thesis we discussed the efficient coding principle [12], and redundancy in statistics of natural images [11], and we proposed to use Kolmogorov complexity as a framework to model the compression that occurs in the HVS [13]. The theories of Kolmogorov complexity and NID provide a solid groundwork to build a universal and generic information distance between virtually any two objects [15]. However, Kolmogorov complexity is non-computable, and it is usually approximated using compressors [15]. Furthermore, the concept ignores the time, energy, and hardware constraints that the HVS faces in processing the visual information. Another more important problem is that Kolmogorov complexity treats every bit of visual information with equal importance, and all lossless compression algorithms that are commonly used in the literature give the same weight to all the information bits present in the data. On the other hand, it is a well documented fact that HVS is lossy, and requires selective filtering of visual information in order to transmit the visual information through a band limited neuron channel [88, 89].

To effectively model such constraints, we must revisit the notion of perceptual information of images. Similar to VIF [4], we conjecture that the perceptual information content of an image can be defined as the amount of inherent physical information of an image that can be extracted by HVS. Based on this conjecture, we may incorporate a perceptual

component into the Kolmogorov complexity framework. The class of visually lossless images can be defined as a set of images which are not exactly the same in the description but are indistinguishable by the HVS. In order to bring the concept of visually lossless equivalency into the regular definition of Kolmogorov complexity, we need a perceptual distance, preferably a metric, that defines boundaries of the equivalent set. Assuming that such a distance exists, and we have $x \in X$ being a typical output of random source X with Kolmogorov complexity of $C(x)$, then y, represented with $r < C(x)$ bits, is a lossy compression of x, and a distortion ball is defined as follows:

$$B_D^n = \{y^n \in A^n : d_n(y^n; x^n) \leq D\} \tag{6.1}$$

where x^n and y^n are strings of n outputs of sources X and Y, and B_D^n is a distortion ball centered around x^n for equivalent images over the alphabet A^n, given a distance measure $d_n(x^n, y^n)$ and a distortion threshold D. Visual complexity of the class of images in that distortion ball is defined based on the notion of a D-ball centered around x^n:

$$Q_D(x^n) = \arg \min_{y^n \in A^n : d(x^n, y^n) \leq D} C(y^n) \tag{6.2}$$

and $C_D(x^n)$ is defined to be:

$$C_D(x^n) = C(Q_D(x^n)) \tag{6.3}$$

If there exists many such sequences in D-ball with smallest complexity, the one with minimum $d_n(x^n, y^n)$ is used. It is shown in [64] that for a stationary ergodic source with probability measure μ, we have:

$$\lim_{n \to \infty} \frac{C_D(x^n)}{n} = R(D) \tag{6.4}$$

where $R(D)$ is the rate distortion function.

The idea of iso-visual complexity spheres can be incorporated into the NCCD framework by replacing $C_T(x|y)$ and $C(x)$ with $C_D T(x|y)$ and $C_D(x)$. This definition will formally extend the NCCD, and any other NID based framework to include lossy descriptions of image information using the concept of complexity distortion function. Furthermore, since complexity of the distortion balls can be replaced by rate distortion function as shown in 6.4, the Kolmogorov complexity terms in NID based frameworks can be replaced by Shannon's rate distortion function, and optimization algorithms can be developed to control the trade-off between complexity and perceptual quality of the images belonging to the same class. A natural extension of complexity distortion function has also been proposed in [65] which can be used to model limited computational resources of the HVS.

In chapter 4, we presented a NID based perceptual quality assessment method and compared its performance with existing state-of-the-art perceptual quality assessment algorithms. We believe that future research on incorporating the notion of visually lossless complexity into NPIS algorithm can lead to formulation and development of an elegant information theoretic framework for perceptual image and video quality assessment based on Kolmogorov complexity and NID, which will resolve many of the shortcomings of the existing methods.

6.2.2 Refinement of The Image Similarity Framework For Texture Classification and Face Recognition

Normalized Information Distance is a universal metric, in the sense that it minorizes every other distance among the two objects being compared [6]. This makes NID a powerful tool in detecting underlying similarities among images. However, as presented in this thesis, non-computability of Kolmogorov complexity terms in NID makes it difficult to apply the metric to applications such as texture and face recognition. In chapter 3, we presented a novel framework to estimate the non-computable NID with Normalized Conditional Compression Distance (NCCD), and demonstrated its performance in texture classification and face recognition. During implementation of NCCD using H.264 video encoder in section 3.4, it was observed that high quantization parameters for both I and P frames of the video resulted in higher accuracy rates for classification of texture and recognition of faces in all tested databases. This phenomena can be attributed to the fact that the visual data is highly redundant, and underlying similarities among images, such as structural properties, can be better detected using relevant data to the features of interest. It is therefore, crucial to the performance of NID based texture classification and face recognition frameworks to minimize redundancy of the images under test, while keeping important visual information of the images for proper detection.

The problem of building a compact representation with relevant information of objects has been addressed in Kolmogorov complexity literature [90] and preliminary application of the method to image similarity has been explored [44]. The method is based on a theoretical decomposition of NID to distance among model and residue components of the images. Another solution to this problem has been taking sparse representation-based approach in encoding visual information of the images as proposed in [56]. The proposed solutions provide promising preliminary results, but remain incapable of providing a strong theoretical framework in model selection for general applications. Proper formulation of compactness for NID has yet to be developed as well. Future research in this direction

can formalize optimum model selection in applying NID to texture classification, face recognition and other detection and retrieval applications applied to images.

6.2.3 Further study on Information Distances as Features

In this thesis, we presented applications of Normalized Information Distance to the field of perceptual image quality assessment, image classification and recognition. Although the results have been promising, there are still many challenges ahead before NID or its variants such as NCCD can outperform traditional image similarity and quality assessment methods such as SSIM in assessment of perceptual quality and similarity of images. As discussed in section 6.2.1, we believe that in order to make NID framework competitive, a perceptual component must be incorporated in the framework. However, modification of the framework results in higher computational complexity, and further complicates the simple and elegant NID.

An alternative to direct modification of the NID framework can be using information theoretic distances as input features to a machine learning algorithm which has been trained by human observer data. Using this technique, a properly trained algorithm can predict a perceptual quality or similarity score based on the previously trained perceptual scores from human observers and a vector of information theoretic distance features, which usually represent the number of bits required to modify certain features of one of the images and reach the modified image.

In chapter 5, we presented a set of information theoretic features which were used by Support Vector Regression tool to quantify a perceptual metric for modification of images, and showed that the predicted scores highly correlate with mean observer ratings. We believe that further research in this direction can lead to developing competitive information theoretic based quality metrics with higher flexibility and significantly lower complexity.

6.2.4 Information Distance as a Parameter Selection Tool

Normalized Information Distance provides a universal tool which can be used to detect underlying similarities of any two objects that can be created on a Universal Turing Machine (UTM), regardless of their type, shape, size and format [6]. Thanks to these unique properties, the NID framework is applicable in scenarios where no other similarity measurement is.

There are many algorithms in image processing in which the input image are of different size, shape or nature of the output image. In such scenarios, a direct comparison of the

input and output image using traditional algorithms such as MSE or SSIM is impossible. Furthermore if there are any parameters involved in these algorithms, the only way to fine tune these parameters is to use human observer data. Normalized Information Distance can be used to compare the similarity of the input and output images in such scenarios, and also find the parameters such that information of the input image is maximally preserved in the processing and the output image has minimum information distance with the input image. Examples of such applications include 3D to 2D conversion, image fusion, HDR to LDR tone mapping, and image scaling.

In chapter 4.3, we presented a parameter selection algorithm for tone mapping of HDR images, and showed that the scheme results in enhanced quality LDR images compared to those created using MSE parameter selection or tone mapped linearly. We believe that in the future, applications of NID can be extended to address similar problems.

APPENDIX

List of related publications:

Nima Nikvand and Zhou Wang. "Generic image similarity based on Kolmogorov complexity," 17th IEEE International Conference on Image Processing (ICIP), Sept. 2010.

Nima Nikvand and Zhou Wang. "Perceptual normalized information distance for image distortion analysis based on Kolmogorov complexity", International Conference on Applied Mathematics, Modeling and Computational Science (AMMCS), July 2011.

Nima Nikvand and Zhou Wang. "Image Distortion Analysis Based on Normalized Perceptual Information Distance", Springer Journal of Signal, Image and Video Processing, vol. no. 7, issue 3, pp 403-410 May 2013.

Nima Nikvand, Hojat Yeganeh and Zhou Wang. "Adaptive Windowing For Optimal Visualization of Medical Images Based on Normalized Information Distance", accepted in proceedings of ICASSP 2014.

Nima Nikvand and Zhou Wang. "Image Similarity Based on Perceptually Inspired Normalized Conditional Compression Distance", submitted to IEEE Transactions on Image Processing.

References

[1] Z. Wang and E.P. Simoncelli. Translation insensitive image similarity in complex wavelet domain. In *in Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, pages 573–576, Mar. 2005.

[2] Zhou Wang and Eero P. Simoncelli. Reduced-reference image quality assessment using a wavelet-domain natural image statistic model. In *Proc. of SPIE Human Vision and Electronic Imaging*, pages 149–159, 2005.

[3] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Processing*, 13(4):600–612, April 2004.

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

[5] Zhou 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(1):98 –117, jan. 2009.

[6] M. Li, X. Chen, X. Li, B. Ma, and P. M. B. Vitányi. The similarity metric. *IEEE Trans. Info. Theory*, 50:3250–3264, December 2004.

[7] R. Cilibrasi and P. M.B. Vitányi. Clustering by compression. *IEEE Trans. Info. Theory*, 51:1523–1545, April 2005.

[8] I. Gondra and D. R. Heisterkamp. Content-based image retrieval with the normalized information distance. *Computer Vision and Image Understanding archive*, 111:219–228, August 2008.

[9] B. Hescott and D. Koulomzin. On clustering images using compression. Technical report, CS Department, Boston University, 2007.

[10] K. Kaabneh, A. Abdullah, and Z. Al-Halalemah. Video classification using normalized information distance. In *Geometric Modeling and Imaging: New Trends*, pages 34–40, August 2006.

[11] Mariano Sigman. Bridging psychology and mathematics: Can the brain understand the brain? *PLoS Biol*, 2(9):e297, 09 2004.

[12] H. Barlow. *Possible principles underlying the transformation of sensory messages*. MIT Press, Cambridge, MA, 1961.

[13] Fred Attneave. Some informational aspects of visual perception. *Psychological Review*, 61(3), 05 1954.

[14] N. Tran. The normalized compression distance and image distinguishability. In *The 19th IS&T/SPIE Symposium on Electronic Imaging Science and Technology*, San Jose, January 2007.

[15] M. Li and P. Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer, Berlin, 2nd edition, 1997.

[16] P. Gács. On the symmetry of algorithmic information. *Sov. Math-Dokl*, 15:1447–1480, 1974.

[17] C. H. Bennett, P. Gács, M. Li, P.M.B Vitány, and W. Zurek. Information distance. *IEEE Trans. Inform. Theory.*, 44:1407–1423, July 1998.

[18] M. Li and R. Sleep. Melody classification using a similarity metric based on kolmogorov complexity. In *Proceeding of Conference on Sound and Music Computing*, October 2004.

[19] D. Cerra, A. Mallet, L. Gueguen, and M. Datcu. Complexity based analysis of earth observation imagery: an assessment. In *ESA-EUSC*, March 2008.

[20] A. Bardera, M. Feixas, I. Boada, and M. Sbert. Compression-based image registration. In *IEEE International Symposium on Information Theory*, Seattle, July 2006.

[21] B. Campana and E. J. Keogh. A compression based distance measure for texture. In *Proceeding of SIAM International Conference on Data Mining*, April 2010.

[22] Allexandre Mallet, Lionel Gueguen, and Mihai. Datcu. Complexity based image artifact detection. In *Data Compression Conference*, page 534, 2008.

[23] A. Kaltchenko. Algorithms for estimating information distance with application to bioinformatics and linguistics. In *Canadian Conf. Electrical and Computer Engineering*, volume 4, pages 2255–2258, May 2004.

[24] Z. Wang and A. C. Bovik. A universal image quality index. *IEEE Signal Processing Letters*, 9:81–84, March 2002.

[25] W. S. Geisler and M. S. Banks. *Visual performance*. Handbook of Optics, M. Bass, Ed. New York: McGraw Hill, 1995.

[26] Z. Wang and A. C. Bovik. Embedded foveation image coding. *IEEE Trans. Image Processing*, 10:1397–1410, October 2001.

[27] Z. Wang. *Rate Scalabale Foveated image and video communications*. PhD thesis, Dept. Elect. Comput. Eng., Univ. Texas at Austin, Austin, TX, December 2001.

[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*, pages 1398–1402, Nov. 2003.

[29] J Portilla, V Strela, M J Wainwright, and E P Simoncelli. Image denoising using scale mixtures of Gaussians in the wavelet domain. *IEEE Trans Image Processing*, 12(11):1338–1351, November 2003. Recipient, IEEE Signal Processing Society Best Paper Award, 2008.

[30] H.R. Sheikh, M.F. Sabir, and A.C. Bovik. A statistical evaluation of recent full reference image quality assessment algorithms in the wavelet domain. *Image Processing, IEEE Transactions on*, 15(11):3449–3451, November 2006.

[31] Hamid 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, Scottsdale, AZ*, pages 23–25, January 2005.

[32] Z. Wang and Q. Li. Information content weighting for perceptual image quality assessment. *IEEE Trans. on Image Processing*, 20(5):1185–1198, May 2011.

[33] H. R. Sheikh, K. Seshadrinathan, K. Moorthy, Z. Wang, A. C. Bovik, and L. K. Cormack. Image and video quality assessment research at LIVE. http://live.ece.utexas.edu/research/quality.

[34] N. Ponomarenko and K. Egiazarrian. Tampere Image Database 2008 TID2008. http://www.ponomarenko.info/tid2008.htm.

[35] E. C. Larson and D. M. Chandler. Categorial Image Quality (CSIQ) Database. http://vision.okstate.edu/csiq.

[36] A. Ninassi, P. Le Callet, and F. Autrusseau. Pseudo no reference image quality metric using perceptual data hiding. In *Proc. SPIE: Human Vision and Electronic Imaging (San Jose, CA, USA)*, volume 6057, pages 146–157, Jan. 2006.

[37] A. Ninassi, P. Le Callet, and F. Autrusseau. Subjective quality assessment-ivc database, 2006. http://www2.irccyn.ec-nantes.fr/ivcdb.

[38] Y. Horita, K. shibata, Y. Kawayoke, and Z. M. Parvez Sazzad. MICT Image Quality Evaluation Database. http://mict.eng.u-toyama.ac.jp/mict/index2.html.

[39] D. M. Chandler and S. S. Hemami. VSNR: A wavelet-based visual signal-to-noise ratio for natural images. http://foulard.ece.cornell.edu/dmc27/vsnr/vsnr.html.

[40] X. Wu and N. Memon. Context-based, adaptive, lossless image codec. *IEEE Trans. Comm*, 45(4):437–444, April 1997.

[41] A. Myronenko and X. Song. Point-set registration: Coherent point drift, 2009. http://www.citebase.org/abstract?id=oai:arXiv.org:0905.2635.

[42] S. Periaswamy and H. Farid. Elastic registration in the presence of intensity variations. *IEEE Trans. Medical Imaging*, 22(7):865–874, July 2003.

[43] Siwei Lyu. Divisive normalization: Justification and effectiveness as efficient coding transform. In *In Advances in Neural Information Processing Systems 23*, pages 1522–1530, 2010.

[44] J.J. Chua and P.E. Tischer. Focusing the normalised information distance on the relevant information content for image similarity. In *Digital Image Computing: Techniques and Applications (DICTA), 2010 International Conference on*, pages 1–7, 2010.

[45] Eero P Simoncelli and Bruno A Olshausen. Natural image statistics and neural representation. *Annual review of neuroscience*, 24(1):1193–1216, 2001.

[46] Yan Karklin and Eero P. Simoncelli. Efficient coding of natural images with a population of noisy linear-nonlinear neurons. In *In Adv. Neural Information Processing Systems (NIPS*11*, pages 13–15. MIT Press, 2012.

[47] Matteo Carandini and David J. Heeger. Normalization as a canonical neural computation. *Nat. Rev. Neurosci.*, 13(1):51–62, 2012.

[48] Siwei Lyu and E.P. Simoncelli. Nonlinear image representation using divisive normalization. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8, 2008.

[49] Sophie Deneve, Alexandre Pouget, and P.E. Latham. Divisive normalization, line attractor networks and ideal observers. In *Advances in Neural Processing Systems*, pages 104–110. The MIT Press, 1999.

[50] K Louie, MW Khaw, and PW Glimcher. Normalization is a general neural mechanism for context-dependent decision making. *Proceedings of the National Academy of Sciences*, 110(15):6139–6144, 2013.

[51] LSI Logic Corporation. H.264/MPEG-4 AVC Video Compression Tutorial, 2003.

[52] Jian-Wen Chen, Chao-Yang Kao, and Youn-Long Lin. Introduction to H.264 advanced video coding. In *Design Automation, 2006. Asia and South Pacific Conference on*, 2006.

[53] T. Wiegand, G.J. Sullivan, G. Bjontegaard, and A. Luthra. Overview of the H.264/AVC video coding standard. *Circuits and Systems for Video Technology, IEEE Transactions on*, 13(7):560–576, 2003.

[54] Zhou Wang and E.P. Simoncelli. Translation insensitive image similarity in complex wavelet domain. In *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, volume 2, pages 573–576, March 2005.

[55] Bilson J. L. Campana and Eamonn J. Keogh. A compression based distance measure for texture. *Stat. Anal. Data Min.*, 3(6):381–398, December 2010.

[56] Tanaya Guha and Rabab K. Ward. On image similarity, sparse representation and kolmogorov complexity. *CoRR*, abs/1206.2627, 2012.

[57] T. Randen. Brodatz texture image database. http://www.ux.uis.no/~tranden/brodatz.html.

[58] Olli Silvén, Matti Niskanen, and Hannu Kauppinen. Wood inspection with non-supervised clustering. *Mach. Vision Appl.*, 13(5-6):275–285, March 2003.

[59] MIT Vision and Modeling Group. Vision Texture Database. http://vismod.media.mit.edu/vismod.

[60] F. S. Samaria and A. C. Harter. Parameterisation of a stochastic model for human face identification. In *Applications of Computer Vision, 1994., Proceedings of the Second IEEE Workshop on*, pages 138–142, 1994.

[61] P.N. Belhumeur, J.P. Hespanha, and D. Kriegman. Eigenfaces vs. fisherfaces: recognition using class specific linear projection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(7):711–720, 1997.

[62] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Proc. NIPS.*, pages 849–856. MIT Press, 2001.

[63] C.H. Papadimitriou and K. Steiglitz. *Combinatiorial Optimization: Algorithms and Complexity*. Dover Publications, 1998.

[64] D. M. Sow and A. Eleftheriadis. Complexity distortion theory. *IEEE Trans. Information Theory*, 49(3):604–608, March 2003.

[65] Daby M. Sow. *Algorithmic Representation of Visual Information*. PhD thesis, Graduate School of Arts and Sciences, Columbia University, 2000.

[66] Peter Grünwald and Paul M. B. Vitányi. Shannon information and kolmogorov complexity. *CoRR*, cs.IT/0410002, 2004.

[67] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, New York, 1991.

[68] A. Srivastava, A. B. Lee, E. P. Simoncelli, and S c. Zhu. On advances in statistical modeling of natural images. *Journal of Mathematical Imaging and Vision*, 18:17–33, 2003.

[69] H.R. Sheikh, A.C. Bovik, and L. Cormack. No-reference quality assessment using natural scene statistics: Jpeg2000. *Image Processing, IEEE Transactions on*, 14(11):1918–1927, 2005.

[70] P.J. Burt and E.H. Adelson. The laplacian pyramid as a compact image code. *IEEE Trans. Communications*, 31:532–540, April 1983.

[71] D. M. Chandler and S. S, Hemami. Vsnr: A wavelet-based visual signal-to-noise-ratio for natural images. *IEEE Trans. Image Processing*, 16:2284–2298, Sept. 2007.

[72] N. Ponomarenko, F. Silvestri, K. Egiazarian, M. Carli, J. Astola, and V. Lukin. On between-coefficient contrast masking of dct basis functions. In *Third International Workshop on Video Processing and functions, (Scottsdale, Arizona,USA)*, Jan. 2007.

[73] E. C. Larson and D. M. Chandler. Most apparent distortion: full reference image quality assessment and the role of strategy. *Journal of Electronic Imaging*, 19:011006:1–21, Jan.-Mar. 2010.

[74] Erik Reinhard, Michael Stark, Peter Shirley, and James Ferwerda. Photographic tone reproduction for digital images. In *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, SIGGRAPH '02, pages 267–276, New York, NY, USA, 2002. ACM.

[75] Gregory Ward Larson, Gregory Ward Larson, Holly Rushmeier, Holly Rushmeier, Christine Piatko, and Christine Piatko. A visibility matching tone reproduction operator for high dynamic range scenes. *IEEE Transactions on Visualization and Computer Graphics*, 3:291–306, 1997.

[76] F. Drago, K. Myszkowski, T. Annen, and N. Chiba. Adaptive logarithmic mapping for displaying high contrast scenes. *Computer Graphics Forum*, 22:419–426, 2003.

[77] Raanan Fattal, Dani Lischinski, and Michael Werman. Gradient domain high dynamic range compression. In *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, SIGGRAPH '02, pages 249–256, New York, NY, USA, 2002. ACM.

[78] Francesco Banterle, Alessandro Artusi, Kurt Debattista, and Alan Chalmers. *Advanced High Dynamic Range Imaging*. A K Peters, Ltd., Natick, Massachusetts, 2011.

[79] Y. Salih, W. bt Md-Esa, A.S. Malik, and N. Saad. Tone mapping of hdr images: A review. In *Intelligent and Advanced Systems (ICIAS), 2012 4th International Conference on*, volume 1, pages 368–373, 2012.

[80] Hojatollah Yeganeh, Zhou Wang, and Edward R. Vrscay. Adaptive windowing for optimal visualization of medical images based on a structural fidelity measure. In Aurlio J. C. Campilho and Mohamed S. Kamel, editors, *ICIAR (2)*, volume 7325 of *Lecture Notes in Computer Science*, pages 321–330. Springer, 2012.

[81] J.K. Thmpson and L.J. Heinberg. The media's influence on body image disturbance and eating disorders: We've reviled them, now can we rehabilitate them? *Journal of Social Issues*, 55:339–353, Summer 1999.

[82] P.N. Myers and F.A. Biocca. The elastic body image: The effect of television advertising and programming on body image distortions in young women. *Journal of Communication*, 42:108–133, 1992.

[83] D. Smeesters, T. Mussweiler, and N. Mandel. The effects of thin and heavy media images on overweight and underweight consumers: social comparison processes and behavioral implications. *Journal of Consumer Research*, 36:930–949, Apr. 2010.

[84] Erik Kee and Hany Farid. A perceptual metric for photo retouching. *Proceedings of the National Academy of Sciences*, 108(47), November 2011. http://www.pnas.org/content/early/2011/11/21/1110747108.abstract.

[85] B. Collins, A. Danowitz, and A. Zvinakis. Photo retouching metric, 2012. http://white.stanford.edu/teach/index.php/CollinsZvinakisDanowitz.

[86] A. Gellineau, J. Huang, and C. Lee. Photo retouching metric, 2012. http://white.stanford.edu/teach/index.php/GellineauHuangLee.

[87] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[88] A. Yuille and D. Kersten. Vision as Bayesian inference: analysis by synthesis? *Trends in Cognitive Sciences*, 10:301–308, 2006.

[89] S. Edlman. *Representation and Recognition in Vision*. The MIT Press, Cambridge, Massachusetts, 1998.

[90] L. Gueguen and M. Datcu. The model based similarity metric. In *Data Compression Conference, 2007. DCC '07*, pages 382–382, 2007.