THE STRATEGY OF IMAGE QUALITY ASSESSMENT

A New Fidelity Metric Based upon Distortion Contrast Decoupling

By

ERIC LARSON

Bachelor of Science in Electrical Engineering
Oklahoma State University
Stillwater, OK, U.S.A
2006

Submitted to the Faculty of the
Graduate College of
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
MASTER OF SCIENCE
July 2008

THE STRATEGY OF IMAGE QUALITY ASSESSMENT

A New Fidelity Metric Based upon Distortion Contrast Decoupling

Thesis Approved:

Dr. Damon Chandler, Thesis Advisor

Dr. Guoliang Fan

Dr. Keith Teague

Dr. Gary Yen

Dr. A. Gordon Emslie

ACKNOWLEDGMENTS

I would like to thank my Adviser, Dr. Damon Chandler for his help and support in the work presented in this thesis. I would also like to thank my lab mates for their help in reviewing and many late nights and discussions that helped to shape the direction of this thesis work.

I would also like to thank Dr. Keith Teague for giving me my first research position and for setting me on the road to complete my Masters Degree. Without his assistance, I would surely not be in the position I am today.

I would like to thank Dr. Gouliang Fan and Dr. Gary Yen for their guidance in the fields of computer vision and global optimization, which helped to largely shape the direction this thesis explored.

A special thanks to the researchers at LIVE for generously providing a subjective fidelity database, without which this thesis would not have been possible.

Lastly, I would like to thank my wife, Chelsea, for putting up with my work habits and for all her help and support.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER 1

INTRODUCTION

This chapter introduces several aspects of this thesis work, namely the background, motivation, current problems, and objectives. Also included are two clerical items to the thesis: the outline and definition of terms. In this chapter we will make a case for using different strategies for assessing image quality based upon the level of visual distortions in the image. Ultimately, we will present a new measure of image fidelity coined *Most Apparent Distortion* (MAD). Section 1.1 presents an annotated history of image fidelity assessment algorithms. Then, section 1.2 illustrates the need for this evaluation study, in terms of quantifying the direction image fidelity metrics should pursue. Section 1.3 gives some examples that motivate the decoupling of quality assessment strategies for high and low quality images. Finally, Section 1.4 details the key questions this thesis addresses.

1.1 Background

Various image processing disciplines require subjective ratings of fidelity based upon a reference image. Image compression, printing calibration, image enhancement, watermarking, image transmission, computer graphics, etc. all rely on producing a visually pleasing result. Many times, this means quantifying a user's ability to detect artifacts in images and the degree to which visible artifacts degrade or enhance its perceived fidelity. As it turns out, the fidelity of an image can be judged quite easily and consistently by a human. That is to say, given a reference image and distorted version of the image, a group of individuals will give roughly the same scoring for the image [1].

In image compression, for example, given subjective feedback and a target bit/pixel ratio, a system can be optimized to create the most visually pleasing result. Therein lies the problem. Using human feedback inside an iterative process is impractically slow. It predicates the need for a machine evaluation of image fidelity.

The first image quality metrics used computationally convenient measures such as *mean-squared error*(MSE) and *peak signal-to-noise ratio*(PSNR). Though still widely used at present, these metrics do not correspond well with subjective ratings of fidelity [2] for a wide range of images. For example, Figure 1.1 shows three images of the exact same MSE. However, observers would rank the quality of these images radically different.

(a) (b) (c)

Figure 1.1: Three pictures are shown with the same mean squared error. The distortions present are (a) additive white Gaussian noise, (b) JPEG compression artifacts, and (c) JPEG2000 compression artifacts. It can be easily seen that the images shown do not illustrate what the average observer would consider equal quality images.

It seems obvious from the example in Figure 1.1 that to improve our estimation of fidelity we need to use properties of the *human visual system* (HVS). However, which properties of the HVS are most important to model is not exactly clear. During the evolution of fidelity prediction, metrics began to follow one of two distinct lines. The first line of fidelity prediction algorithms use only bottom-up properties of the HVS (such as masking and contrast sensitivity). Another line of fidelity metrics define quality based upon what they believe the HVS is ultimately trying to achieve, namely structural and information content. The first camp is rooted in near thresh-

old psychophysics and the second in biologically motivated aspects of visual signal processing.

1.1.1 Metrics based upon near threshold psychophysics

The first improvement to MSE and PSNR built upon them using contrast properties of low level vision. For example, *weighted signal-to-noise ratio*(WSNR) takes contrast sensitivity into account when assessing error. Visual contrast sensitivity experiments show that the HVS peaks in sensitivity at about 1-6 cycles per degree of visual angle, commonly referred to as the *contrast sensitivity function*(CSF)[3][4][5][6]. WSNR exploits the CSF to weight errors that occurr in high frequency areas less when predicting fidelity. Although only a slight improvement to PSNR, WSNR opened the door to using bottom-up properties of vision for fidelity assessment.

In addition to contrast sensitivity, it is well known that masking is effected by properties of the image onto which the distortion appears. Typically, regions of high spatial frequency can mask distortion better than smooth areas[7][8]. Spatial frequency masking is commonly captured using statistics of oriented sub-band filters.

Fidelity algorithms soon evolved to account for spatial frequency masking using multi-channel filtering and different models of contrast sensitivity thresholds [9] [10]. Daly's *Visual Difference Predictor* (VDP) [11] was one of the first and most comprehensive algorithms to use contrast sensitivity and detection mechanisms to define the perceived fidelity of an image. The concept of elevated masking thresholds versus spatial frequency is key to its performance. However, when the VDP prediction maps are collapsed into a single measure of image quality, the overall performance is only marginally better than PSNR.

When the wavelet was shown to correlate highly with human cortical responses[12], fidelity algorithms were quick to exploit the use of wavelet based decompositions. Algorithms soon expanded to efficiently define contrast sensitivity and distortion mask-

ing through wavelet sub-bands [13] [14]. Although these metrics are better models of subjective fidelity, they are still far from being exact predictions.

The incorporation of masking models and image fidelity prediction has been limited mainly from the fact that most models of masking are motivated by *detection thresholds*. It is still largely unclear if the same principles can be applied when distortions are clearly supra-threshold. This is one of the key questions that this thesis addresses.

1.1.2 Metrics based upon structural and information content

A second camp of image quality metrics evolved based upon measures of image structure and information extraction. Namely, *Mean Structural SIMilarity* (MSSIM or SSIM) [15] and multi-scale MSSIM [16] bases measures upon statistical differences between the reference and distorted image using a Gaussian weighted sliding window approach. This method has been shown to correlate highly with subjective ratings of distortion. However, there is no explicit use of a masking model or contrast thresholds in SSIM.

Another approach, *Information Fidelity Criterion*(IFC), uses statistics of wavelet sub-band coefficients to quantify the amount of information in a reference and distorted image [17]. The information content is judged based upon a Gaussian mixture model learned from natural scene statistics. An extension of this metric, *Visual Information Fidelity*(VIF), builds upon IFC using a wavelet based model of natural scene statistics *and* human vision [18]. Even so, the HVS model used in VIF is crude and does not contain an explicit masking threshold to speak of. The performance of VIF has been shown to correlate highly with subjective ratings of image fidelity, even without a sophisticated masking model.

More recently, a metric, *Visual Signal-to-Noise Ratio*(VSNR), took both visual detection thresholds and perception of distortion in natural images into account [19].

This metric uses the detectability of distortion, perceived contrast, and degree of global precedence disruption to predict fidelity (defined by wavelet subband statistics). More information on SSIM, VIF, and VSNR can be found in Chapter 2.

1.1.3 Top-down approaches to image fidelity prediction

Other metrics have evolved for quality assessment based upon top-down properties of human vision, but have had limited success. Preliminary work was performed by Osberger *et al.*, that attempted to quantify importance based upon low level properties of an object such as size, location, and shape [20]. The importances are then used to weight errors differently in the distorted image. The results, however, only show slight improvement over PSNR. Still others believe that weighting more salient regions in images helps in studying fidelity. In [21], [22], and [23] saliency maps and eye-tracking data were incorporated into existing metrics of visual quality. The improvements, however, are only marginally better than the baseline metrics. In [24], the use of content based region-of-interest maps was applied to existing quality metrics. The improvements, however, are also insignificant. In light of these results we have chosen not to employ saliency or region of interest information into our measure of fidelity.

1.2 Motivation

Before moving on, it is interesting to clarify what benchmarks for defining quality might be possible. Subjective scores for fidelity assessment are not identical from observer to observer. There is a degree of variance to the data. Sheikh, *et al.* devised a measure known as the outlier ratio that defines a false fidelity prediction score as one that lies outside two standard deviations above or below the average subjective fidelity score [1]. If we divide the false predictions by the total predictions, we get a percentage of correctness, $R_{out} = \frac{N_{false}}{N_{total}}$. In this way, we can evaluate different metrics

without ignoring the variations that exist between observers. However, even the best fidelity assessment predictors have false predictions over 50% of the time. The current state of the art is not wholly acceptable. This leads us to ask what factors are we not addressing in image fidelity assessment? What else could be used to augment fidelity predictors, such that they can achieve near perfect correlations?

Clearly, current metrics are capable of defining different aspects of fidelity assessment. However, they fail to capture *all* features that define fidelity, as denoted by the high outlier ratios. This motivates the need for a different strategy of quality assessment prediction. In this thesis we argue that the most informative features for image quality assessment are (1) the perception of structural appearance degradations when the distortions are highly supra-threshold, (2) the perception of masking and local error intensity when the distortions are mild, and (3) effectively modeling the interaction of the two strategies as the distortions become increasingly visible.

(a) Reference Image (b) Mild Dist. (c) Heavy Dist.

Figure 1.2: A reference image and two distorted images are shown. The distortions present are additive pink Gaussian noise. Observe that in the high quality image we rate quality largely on how well we can see the distortions, but we rate the low quality image based on how much the distortion disrupt appearance.

1.3 Problem Description

This section describes image quality assessment as a task with different objectives depending on the overall visibility of distortions. It may not be obvious why we

have chosen to decouple the estimation of quality based upon the visibility of the distortions. We present the problem and argument using some motivating examples.

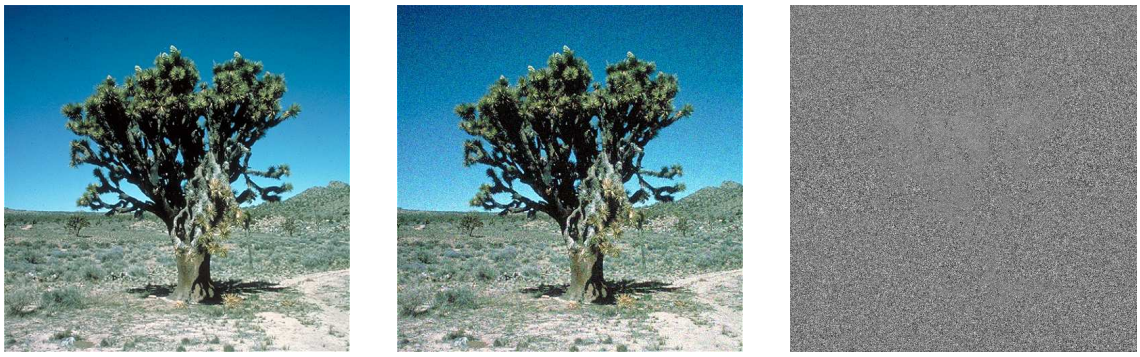
Figure 1.2 shows a reference image and two distorted versions of the image, one with mild distortion and another with heavy distortion. For the high quality image, imagine trying to rank the quality of many similarly distorted images. In this regime, it is easy to see how we might interpret quality assessment as a detection task. We begin carefully looking for any distortions in the images and determine quality based upon how many distortions we can detect and how intense the distortions appear.

Now imagine performing a ranking for images similar to the heavily distorted image. Would we again judge quality using the same objectives? We believe not. In the low quality regime, the task of judging quality shifts from detection to one of appearance recognition. We no longer care about distortion intensity; we want to know how different distorted regions look compared to their original appearance. Note that when we say appearance here, we are referring to how similar two regions of an image look. We will further define our exact definition of appearance in Chapter 3. In a nutshell, we would like to capture how the distortion alters the visual appearance of textures and edges in the image.

Let us return to the high quality regime for a moment. To further motivate the importance of masking and detection thresholds in high quality images, look at Figure 1.3. A reference image and distorted version are shown with the difference image (additive noise source) between the two. Notice from the difference map that the distortion is nearly uniform across the image (except for some clipping in overly dark regions), but that the distortion is only visible in smooth areas of the image. It is easy to see from this example that we need to account for some degree detection thresholds when assessing the quality of an image.

We would now like to present an example of how frequency effects the appearance of images differently. Figure 1.4 shows an example image and two distorted

versions. The first image is distorted by eliminating *discrete cosine transform* (DCT) coefficients at a low frequency band. The second is distorted by eliminating the same number of coefficients at a slightly higher frequency band. Both distortions are highly supra-threshold but the lower frequency loss is more apparent. This example is meant to motivate our creation of an appearance measure for low quality images. Clearly, the appearance model will need to rank high frequency distortion as less important than low frequency distortion. In this way the metric for low quality images will still need to use elements of spatial frequency masking, albeit in a manner different from that used when distortions are near threshold.

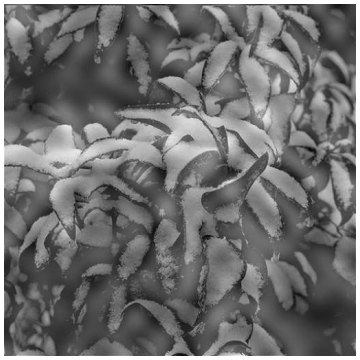


(a) Reference Image (b) Dist. Image (c) Error Image

Figure 1.3: A reference image and distorted image is shown. The distortion present is additive white Gaussian noise. The absolute difference image is also shown where gray represents no difference and white or black represent differences between the reference and distorted image. The difference image has been contrast enhanced to accentuate visibility. Observe that The distortion energy is largely uniform but only noticeable in the smoother area of the image. Clearly a masking model is needed to correctly predict the quality of this image.

We have presented exemplary evidence for decoupling the estimation of fidelity based upon the degree to which distortions are visible. A key question in this type of analysis is what occurs in the transition region. Some distortions may be near threshold and others may degrade the appearance of the image. We believe that in this regime observers use a mixture of the strategies presented. In the end, we decided to use an α-blending of the two metrics to approximate this interaction. The exact

(a) Reference Image



(b) Low Freq. Loss

(c) Mid Freq. Loss

Figure 1.4: A reference image and two distorted images are shown. The distortions present are artifacts from information loss of discrete cosine transform (DCT) coefficients. (b) shows the result of throwing away a subset of low frequency DCT coefficients. (c) shows the results of throwing away the same number of DCT coefficients, only shifted to a slightly higher frequency.

implementation can be found in Chapter 3.

In this thesis we will present a metric that approximates both of the aforementioned strategies. Because the metric attempts to quantify which distortions are most apparent to the observer, we have coined it *Most Apparent Distortion* (MAD).

1.4 Research Objectives

The questions this research strives to find the answer to are listed below. The ramifications of each objective are briefly discussed.

1.4.1 Can a fidelity metric with visual masking compensation perform well when the images are of high quality?

We wish to know how well a metric can comparatively perform when masking is implemented on high quality images. We expect that at high quality, visual masking will greatly help the prediction accuracy while it will be a hindrance at low quality.

1.4.2 Can a fidelity metric be tuned to recognize the extent of visual appearance disruption for low quality images?

We wish to know how well a metric can identify the extent to which supra-threshold distortions disrupt appearance. We expect such a model to approximate fidelity of low quality images well, but perform poorly on high quality images.

1.4.3 Can predictions of perceived fidelity be significantly improved using different strategies for low and high quality images?

This is the key objective of this thesis study. We expect strategic interaction modeling to improve the accuracy of fidelity assessment. However, statistical significance is the key issue here.

1.5 Outline

The outline of the Thesis is as follows:

- Chapter 2 discusses the current state of the art in fidelity assessment prediction and introduces the intricacies of different metrics.

- Chapter 3 explains our methodology, namely how we define a measure of fidelity for low and high quality images and train their interaction. We also explain the methods used to create a sufficient database for use in their evaluation.
- Chapter 4 summarizes our findings.
- Chapter 5 concludes this thesis and presents ideas for future work.

1.6 Definition of Terms

ς	Skewness
κ	Kurtosis
μ	Mean
σ	Standard Deviation
R_{out}	Outlier Ratio
R_{SOD}	Sum Outlier Distance
CC	Correlation Coefficient
CIELAB	International Commision on Illumination Lightness and Color Opponent
CSF	Contrast Sensitivity Function
DCT	Discrete Cosine Transform
DMOS	Differential Mean Opinion Score
HSV	Human Visual System
IFC	Information Fidelity Criterion
JPEG	Joint Photographer Experts Group
LMSE	Local Mean Squared Error
MAD	Most Apparent Distortion
MSE	Mean Squared Error
MSSIM	Mean Structural Similarity
PSNR	Peak Signal to Noise Ratio
RMSE	Root Mean Squared Error
ROCC	Rank Order Correlation Coefficient
VDP	Visual Difference Predictor
VIF	Visual Information Fidelity
VQEG	Visual Quality Experts Group
VSNR	Visual Signal to Noise Ratio

CHAPTER 2

STATE OF THE ART

This section expands on the annotated history from Chapter 1. In particular, we explain several metrics of fidelity that are widely used by the research community and that are used to evaluate our results against. Section 2.1 explains metrics used for their computational efficiency. Section 2.2 explains *Visual Signal to Noise Ratio* (VSNR)[19], a wavelet based measure of quality. Lastly Section 2.3 explains SSIM, a measure devised by Wang *et al.*[25], and VIF, a measure introduced by Sheikh *et al.*[18].

2.1 Pixel and luminance differences

Certainly PSNR is one the most used metrics of fidelity. It is defined as the decibel measure of the maximum MSE over the actual MSE, as shown in equations (2.1) and (2.2). For 8-bit grayscale images,

$$PSNR = 10 \log_{10} \left(\frac{255^2}{MSE} \right) \tag{2.1}$$

where

$$MSE = \frac{1}{MN} \sum_{i \in M, j \in N} \left(\widehat{I(i,j)} - I(i,j) \right)^2 \tag{2.2}$$

where M and N are the dimensions of the image, I is the original image, and $\widehat{I}$ is the distorted image. The metric is based explicitly on the MSE and is a widely used measure of quality in image processing. It contains no alterations for masking, contrast sensitivity, or other factors that attempt to describe the HVS. The advantage of PSNR is that it is (comparatively) efficient in terms of computational complexity.

Many researchers will opt to use PSNR inside iterative processes as other metrics can have substantially longer run times when used in this fashion. However, except for the special case of white noise, PSNR cannot describe subjective ratings of image fidelity with the accuracy of other metrics.

2.2 Wavelet Decomposition

VSNR attempts to define quality using concepts of visual contrast masking and global precedence. The metric uses contrast detection thresholds for the reference image and distortions to define if the distortions in the image are visible. VSNR, firstly, computes the perceived contrast of the distortions in the image, denoted by d_{pc}. It then computes the degree to which these disrupt the global-precedence-preserving contrast, defined by wavelet sub-band statistics in the reference image and distortion channel. The measure of disruption is collapsed into a single distance measure, d_{gp}. The *visual distortion* is then defined as a linear combination of d_{pc} and d_{gp}.

$$VD = \alpha d_{pc} + (1 - \alpha)\frac{d_{gp}}{\sqrt{2}} \tag{2.3}$$

where α is a tunable parameter for defining the relative contribution of each distance measure. If $C(I)$ is the RMS contrast of the reference image, then VSNR can be defined according to:

$$VSNR = 10 \log_{10}\left(\frac{C^2(I)}{VD^2}\right) \tag{2.4}$$

The result is a decibel measure of image quality. Note that VSNR is a measure of image fidelity. If the distortions in the image are sub-threshold, VSNR returns a value of ∞. VSNR also has the advantage of being more computationally relaxed than other metrics of fidelity because it uses the separable 9/7 discrete wavelet transform for image analysis.

2.3 Information Content and Structure of Natural Scenes

SSIM attempts to account for the "structure" differences between a reference image and a distorted image. It is based on the assertion that, at a high level, the HVS scans a scene (from the fovea outwards) and makes judgments about the luminance and contrast in the scene. It attempts to mimic the fovea input by using a 13x13 pixel Gaussian window. Running this window pixel by pixel across the reference image and distorted image, the metric calculates a weighted approximation of the difference between kernel luminance and kernel contrast in the images at each pixel. The kernel difference is calculated using equation (2.5).

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{2.5}$$

where x and y are kernels from the reference and distorted images, σ is the variance of each kernel or covariance between the kernels, and μ is the mean of each kernel. Constants are added to prevent the equation from approaching infinity. This results in a "structural" map of the differences in the images as the kernel is passed across the images. An example map appears in Figure 2.1. SSIM collapses the map into a single measure of quality using the mean value.

The VIF criterion is by far the most computationally intensive algorithm looked at in this thesis. The model uses the steerable pyramid and Gaussian scale mixtures to characterize random fields in the wavelet domain. The steerable pyramid is a set of multiple scale and orientation filters in quadrature. The use of the pyramid is not essential for VIF. Without loss of generality, any multi-scale frequency decomposed method of image analysis may be used, such as wavelet sub-bands or Gabor filter banks. The general idea is to transform the image into a domain where the coefficients are less redundant than the spatial domain, making the calculation of mutual information more accurate.

At the heart of the algorithm is the use of a covariance matrix of the steerable

(a) Reference Image

(b) Dist. Image

(c) SSIM Map

Figure 2.1: A reference image and Gaussian blurred distorted version are shown. The SSIM map attempts to rate portions of high distortion less than regions of the image that are not structurally distorted. Black corresponds to regions that SSIM predicts to be of low quality.

pyramid coefficients to estimate a visual distortion model. The model is used to estimate the amount of information (a single number) that the HVS could discern from the reference image and the information actually discerned from the distorted image. Once the "information values" are found the two numbers are divided to measure the amount of mutual information that they share, and subsequently, image quality. Figure 2.2 shows the flowchart for VIF. To the best of our knowledge, VIF has been shown to be the best performing algorithm for prediction of subjective image quality.

Note that some strategies used in VIF are similar to those used in visual masking models. However, VIF ignores most aspects of visual masking such as contrast sensitivity and conversion to luminance using the point spread function. VIF also

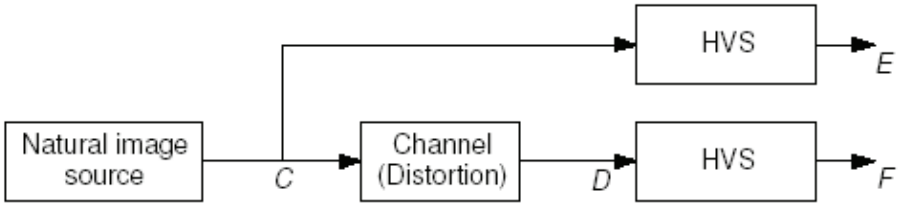


Figure 2.2: The example model of the reference and distorted image information used by VIF.

has the capability of detecting enhancements in images. For example, VIF predicts that contrast enhancements will results in ratings of quality greater than the original image.

CHAPTER 3

METHODOLOGY

This chapter describes the various implementations used in *MAD*. In section 3.1, for high quality images, the masking model and distortion energy model are discussed. Then, for low quality images, section 3.2 introduces the appearance relation model is discussed. Section 3.3 explains the implementation of blending the two strategies. Finally, section 3.4 presents the methods used to create a new subjective image quality database.

3.1 A Strategy for High Quality Images

We have argued that observers use a different strategy for assessing high quality images. Namely, that assessing high quality images is a distortion detection task. In a nutshell, we argue that observers discriminate between high quality images by asking two questions: Can we see the distortion? And, how much does the distortion degrade the image?

In order to capture these aspects we first need a masking model that will tell us *if* the distortion is visible and then we need a measure of *how* visible it appears. However, current masking models are largely created from the point of view of compression. Their main purpose is to evaluate where distortions appear in an image first, not *if* an already present distortion is visible. In light of this, we decided to create a new model of masking that is tuned to detect visibility given a distorted and reference image. Once the masking map is created, we use the local energy of the distortion as a measure of how much the distortion degrades quality.

3.1.1 A New Masking Model Tuned for Quality Assessment

When viewing many images of high quality, the only way to discriminate is to go looking for distortions. In order to approximate that task we have chosen to combine a masking map of visible distortions in the image with the local mean squared error($LMSE$). In this way, we model observers detecting distortions and use distortion energy as a means of quantifying how much each degrades quality. Previously, we have stated that MSE is not a good model of quality. However, we will show that when done locally and combined with the proper masking model, it is a good indication of fidelity in high quality images.

(a) (b) (c)

Figure 3.1: Three pictures are shown with the same type an intensity of distortions. The distortions present are JPEG blocking artifacts and DCT edge effects. (a) Shows the visibility of artifacts in a smooth area of average brightness, (b) shows the visibility of the distortions in busy areas, and (c) shows the visibility in bright areas. It can be easily seen that the different properties of the image illustrate how well the patch can mask distortions.

To effectively model the visibility of distortions in images we need to model various elements of low level vision. Firstly, pixel values must be converted to their approximate luminance and then to their perceived level of brightness. Secondly, we must account for contrast sensitivity using the CSF. If applied to the reference and distorted version of the image, these modifications will bring the images into their approximate perceived level of brightness and contrast. After this, we must account for the fact that *busy* areas in an image can mask distortion better than smooth areas. Also, we must account for luminance masking; that is to say we must model that it

is more difficult to see errors in very bright regions.

Figure 3.2: Five pictures are shown with the increasing sinusoidal contrast intensities from left to right. Notice that the smooth area shows the distortion immediately and the turtle's skin masks for quite several images afterward. Also notice that the bright patch in the bottom right of the turtle's skin completely masks the distortion even after it is visible in other portions of the skin. This illustrates both contrast and luminance masking.

Figures 3.1 and 3.2 show some examples of these masking elements. Notice that smooth areas in the images are easy indicators of the added distortions. Also notice that bright areas do well at hiding many of the distortions that are visible in other areas.

We begin by converting pixel values to luminance according to

$$L = k \times I^{\gamma} + b \tag{3.1}$$

where I is the integer pixel values (0-255) in the reference or distorted images, and the parameters are constants defined as $\gamma = 2.2$, $k = 0.02874$, and $b = 0$. Gamma has a special meaning. In the days of the CRT, it was well known that the monitor applied a nonlinear gamma curve to pixels. In response, image acquisition has long adapted to the common CRT nonlinearity using an *inverse* gamma. The idea here is that when displayed on a CRT, the images appear about how they looked originally. For us, this means that before processing the pixels, we must undo the nonlinearity imposed during image acquisition using $\gamma = 2.2$.

To convert them to luminance we use the k and b variables. This operation scales the values into cd/m^2 rather than their integer values. So L is the absolute luminance

of the pixels in the image.

Once we have the approximate cd/m^2, we then convert the absolute luminance values into $L*$ space using

$$L* = \sqrt[3]{L} \tag{3.2}$$

The idea here is that absolute luminance is not linearly associated with our *perception* of brightness. $L*$ is the luminance portion of the CIELAB color conversion and has been shown to linearly approximate the average *perception* of greyness. We perform this operation for the reference and distorted images, resulting in the respective perceived luminances $L*_{ref}$ and $L*_{dst}$. Of course, if one knows more about how the image was acquired, the conversion parameters, γ, k, and b, can be tuned even further. We just use the most common conversion values.

We then define the difference image according to $L_{diff} = L*_{ref} - L*_{dst}$. We account for contrast and spatial frequency by filtering the reference image and difference image by the *CSF*. The *CSF* filtering is performed in the Fourier domain assuming a maximum of 32 cycles per degree, and a rolloff of 3dB along the diagonals of the *CSF*. We use a peak sensitivity of 6 cycles per degree as studies show that our sensitivity peaks somewhere between 4 and 8 cycles. The filtered image is then converted back to the spatial domain.

$$I'_x = F^{-1}\left[CSF \times F\left[L*_x\right]\right] \tag{3.3}$$

where $F[\cdot]$ is the Fourier transform operation and $F^{-1}[\cdot]$ is the inverse Fourier transform operation. $L*_x$ denotes the reference or difference image, with $_x$ replaced by $_{ref}$ or $_{diff}$. So at this point, I'_{ref} and I'_{diff} are the reference and distorted images that are linearly associated with our perception of brightness *and* contrast. You can think of I'_{diff} as the distortions in the image that our eye could see if they were on a completely smooth, medium brightness background. We now need to see if the reference image is masking those distortions.

A good measure of the masking potential in an image is the local standard devia-

tion of I'_{ref}. Other researchers have used local Fourier amplitude slopes and wavelet bases to measure this. Each method has it advantages and disadvantages, however, the standard deviation is a simple and straight forward measure that can be performed without losing spatial resolution.

We calculate the local standard deviation of the reference image in the pth patch as $\sigma_{ref}(p)$, using a sliding window of 16x16 pixels, where each window overlaps three quarters of the immediately surrounding windows. The idea is that large values of σ_{ref} indicate a busy texture (i.e.-high spatial frequency) and can mask well. The caveat, however, is that edges will also have high σ_{ref} values, but cannot hide distortion. To overcome this limitation we use a modified measure of standard deviation. We break the window into four sub-windows and set $\sigma_{ref}(p)$ to be the minimum standard deviation of the sub-windows. So the standard deviation of the reference image becomes

$$\sigma_{ref}(p) = \min\left[\sigma(p_{11}), \sigma(p_{12}), \sigma(p_{21}), \sigma(p_{22})\right] \tag{3.4}$$

where the four standard deviations of the sub-windows are denoted by $\sigma(p_{xy})$ (x, y represent the placement of the sub-window as seen in Figure 3.13).

The beauty of this is that when we encounter an edge (middle figure), it likely does not intersect all four sub-windows. So the minimum standard deviation is in sub-patch p_{21}, and we know that the edge cannot mask. But when we get out to a busy patch (assuming stationarity) the standard deviation will be about the same as the block as a whole - and, thus, we get a measure of masking potential.

We then define the local contrast of the reference image using the modified standard deviation normalized by the mean, $C_{ref}(p) = \sigma_{ref}(p)/\mu_{ref}(p)$, where $\mu_{ref}(p)$ is the local mean of each block p in the reference image, I'_{ref}. This normalization accounts for luminance masking in the image. Large values of $C_{ref}(p)$ indicate that the block can mask distortions well.

We now want to know if the distortions in the difference image are visible. To do

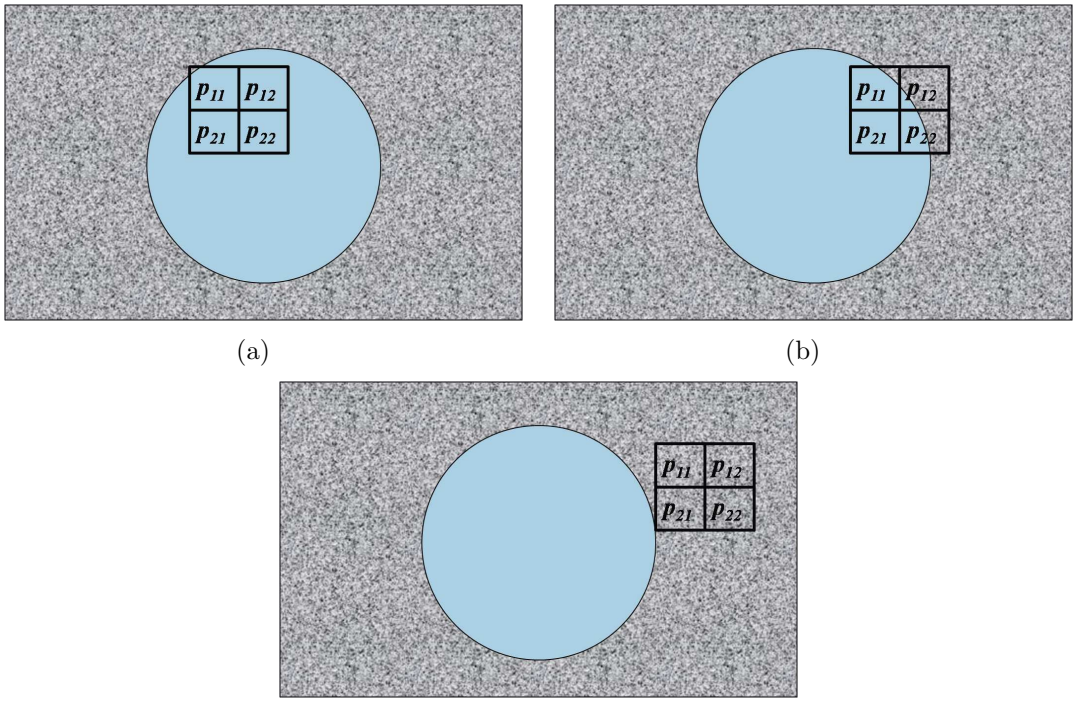


Figure 3.3: The sliding window divided into subsections can be seen at two points in the example image. As the window begins to overlap the edge of the circle, you can see how the sub window method avoids large variances. Because two of the sub-windows do not contain the edge, the minimum standard deviation is still small and we do nor falsely assume that the patch can mask well.

this we compare the modified contrast of the reference image to the contrast of the distortions in the difference image. The local contrast of the distortions is defined as:

$$C_{diff}(p) = \begin{cases} \sigma_{diff}(p)/\mu_{ref}(p), & \text{if } \mu_{ref}(p) > 0.9 \ ; \\ 0, & \text{otherwise} \end{cases} \tag{3.5}$$

Here we use the standard deviation, σ_{diff}, of the difference image, I'_{diff}. Intense distortions will have a large standard deviation. Notice that this has a luminance threshold. The brightness threshold of 0.9 is meant to account for the fact the *HVS* is insensitive to changes in extremely dark regions. We adjust our eyes to the surrounding light so even on an LCD monitor, the blackest black can be outside of our level of sensitivity, especially when in a well lit room. Low luminance values are thus discarded.

From here we can apply a threshold of visibility by comparing the local contrast of each block. If $C_{diff}(p) > 0.75 \times C_{ref}(p)$, then the distortion is considered visible. Or more formally,

$$Visib(p) = \begin{cases} 1, & \text{if } C_{diff}(p) > \frac{3}{4} \times C_{ref}(p) \ ; \\ 0, & \text{otherwise} \end{cases} \tag{3.6}$$

The value of 0.75 was chosen experimentally by the author by comparing a set of masking maps and distorted images. Figure 3.4 shows an example masking map with the reference and distorted image. Notice that the masking inside the skin of the turtle is captured well by the model. Note also that the masking map has different levels of gray, but that any gray level other than black means the distortion is visible.

(a) (b) (c)

Figure 3.4: Three pictures are shown. (a) shows the reference image, (b) shows the image with mild additive white Gaussian noise distortion, and (c) shows the masking map. Bright areas do not show the noise, which the model captures. In addition, the busy area on the turtle's head masks well, which is also captured by the model.

3.1.2 Combining the Masking Map and Local Errors

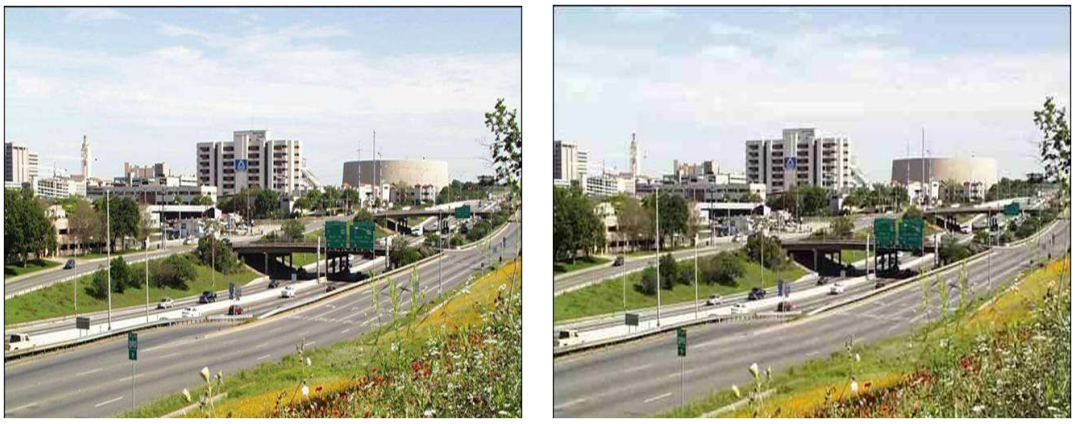
Once the map of visible distortions is created, we use the local distortion energy to determine how much it degrades quality. We define the distortion energy using the local mean squared error ($LMSE$) of a 16x16 pixel block.

$$LMSE(p) = \frac{1}{16^2} \sum_{i,j \in N_p} \left(I'_{ref}(i,j) - I'_{dst}(i,j) \right)^2 \tag{3.7}$$

where N_p is the set of pixels inside the pth 16x16 block. At this point we have a map of the local errors and the local visibility of those erros. To define quality we use the point wise multiplication of the two maps. In essence, this multiplication creates a map of the *LMSE* of *visible* blocks in the image. Eventually, we plan to multiply the *LMSE* map by an *absolute* measure of visibility. However, preliminary results show that quality is well captured using a hard binary threshold for the visibility map (i.e.- only 0's or 1's in the map). We define the quality of the image as:

$$Q_{high} = \frac{1}{N}\sqrt{\sum_{p\in\Lambda_p} Visib(p) \times LMSE(p)^2} \tag{3.8}$$

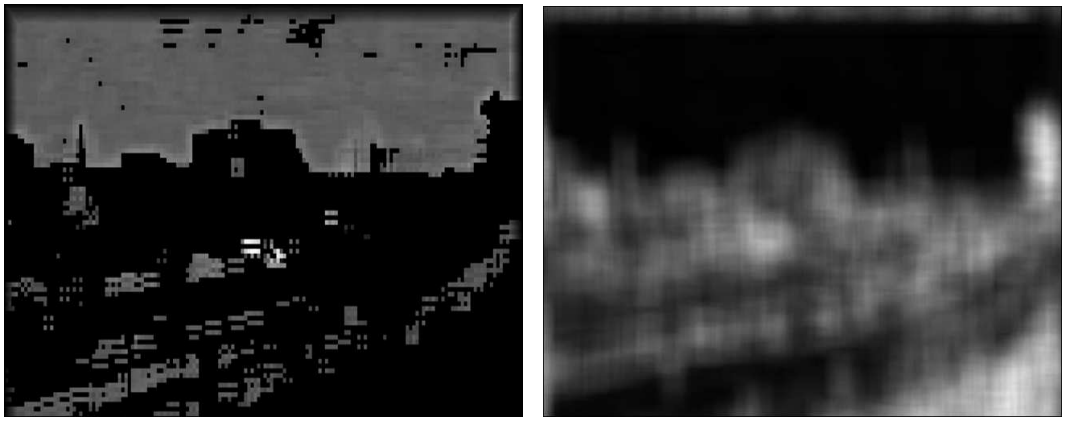
where Λ_p is the set of all blocks where the masking model detects visible distortion, and N is the number of blocks in the entire image. In this way, equation (3.8) collapses the visible *LMSE* into a single quantity using the vector two norm. The two norm gives more weight to larger local distortions than just taking the mean would. This makes sense. The visibility of a distortion and the way it degrades an image do not need to be linearly related. Our research indicates that the conversion from visibility to quality degradation is well modeled by the two norm.



(a) (b)

Figure 3.5: This shows an overview of the entire process for high quality using an example distorted image and reference -(a) and (b). The images are first loaded into memory.

The final measure of distortion, Q_{high}, has the following properties: A value of



(a) (b)

Figure 3.6: The masking map is created from the reference and distorted images (a), using the contrasts, C_{ref} and C_{diff}. Separately, the map of the *LMSE* is calculated - (b). Then the visibility map is thresholded to 0 or 1.

Figure 3.7: The two maps are multiplied together. The resultant map shows the *LMSE* of only visible patches in the image. This map is collapsed using the vector two norm of equation (3.8).

zero denotes the distortions in the image are not present or not visible. Increasing values denote decreasing quality.

Figure 3.5 shows an overview of the entire process for high quality using an example distorted image. The images are from the *LIVE* database[26]. The distorted version shown contains *JPEG* blocking artifacts. Notice that the only real visible distortions in the image are in the sky and the highway lines.

The images are first loaded into memory. Then, the masking map is created from the reference and distorted images, using the contrasts, C_{ref} and C_{diff} from equation

(3.6). Separately, the map of the *LMSE* is calculated. Figure 3.6 shows each of these maps.

The visibility map is thresholded to 0 or 1 and the two maps are multiplied together. The resultant map is shown at the bottom in Figure 3.7 and only displays the *LMSE* of visibly distorted patches in the image. Notice that the map captures well that some artifacts are visible in the sky but do not really degrade the image quality, and that the most annoying artifacts occur in the street.

This map is collapsed using the vector two norm of equation (3.8). Thus, we are left with a single measure of a high quality image. This is the complete methodology for the high quality metric portion of *MAD*. The next section discusses how *MAD* assesses low quality images.

3.2 A Strategy for Low Quality Images

At low quality we have argued that visual masking is of less importance to our perception of quality. The distortions in the image are highly supra-threshold and we are interested in the extent to which they degrade the edge and texture structure of the image. Basically, the task of assessing quality changes from detecting distortion to an appearance relation task - are edges and textures preserved?

We further argue that the degree to which supra-threshold distortions degrade quality is based upon the degree to which they change what our vision system expects to see. We propose using a biologically motivated model of appearance: local statistics of multi-scale log-Gabor filters. This method has long been used by the computer vision community to classify the appearance of textures[27], and we showed in [28] that these statistics are good indicators of how camouflaged an animal appears in its natural environment. In this respect, it shows great promise for approximating the extent to which distortions disrupt appearance.

In [12] and [29] it was shown that the visual cortex may be performing a similar

analysis strategy as log-Gabor filter banks. With the proper values for bandwidths and frequency overlaps of each filter, one can theoretically approximate the function of the first visual cortex. Furthermore, in [30] it was shown that synthetic textures can be shifted visually using statistics of log-Gabors. It was also shown that changes in these statistics were more discernible than changes in spatial statistics. All of these studies indicate that the log-Gabor filter bank can be powerful tools for measuring how distortions change the edges and textures in an image, approximating our needed appearance relation task.

3.2.1 Assessing Quality Using the log-Gabor Filter Bank

The log-Gabor filter has been shown to approximate cortical responses in the visual cortex [29]. In the frequency domain the log-Gabor filter is defined as:

$$S_{log}(f) = \exp\left(-\frac{(\log f/f_o)^2}{2\,(\log \sigma_s/f_o)^2}\right) \tag{3.9}$$

where f_o is the center frequency and σ_s/f_o is a measure of the bandwidth. Notice that S_{log} is only a function of the frequency, f, so it will be concentrically symmetric about the origin in the 2D frequency domain. The parameters of S_{log}, f_o and σ_s, must be tuned to values of sensitivity and bandwidth for the mammalian visual system. In addition the overlap between filters (ratio between increasing values of σ_s/f_o in each scale) must be calculated to be something biologically motivated. We use values defined from the previous research of [31]. The Fourier spectrum of an example filter can be seen in Figure 3.8.

In addition to frequency scaling, each filter has an orientation. In frequency, this translates to a Gaussian defined in the angular component of the frequency domain. or mathematically as,

$$O_{log}(\theta) = \exp\left(-\frac{(\theta - \mu_\theta)^2}{2\sigma_\theta^2}\right) \tag{3.10}$$

where μ_θ is the orientation of the filter and σ_θ is the angular spread. These values

define the orientation of the filter and the overlap between adjacent filters in frequency. These values are also chosen from [31] and are motivated biologically from the mammalian vision system.

Each filter, then, is defined by the multiplication of the scaling and orientation functions,

$$G_{so} = O_{log} \times S_{log} \tag{3.11}$$

where G_{so} is the filter defined at orientation o and scale s.

The final implementation of each filter bank contains four parameters for each filter, f_o, σ_s, μ_θ, and σ_θ. An example of this graphically can be seen in Figure 3.8.

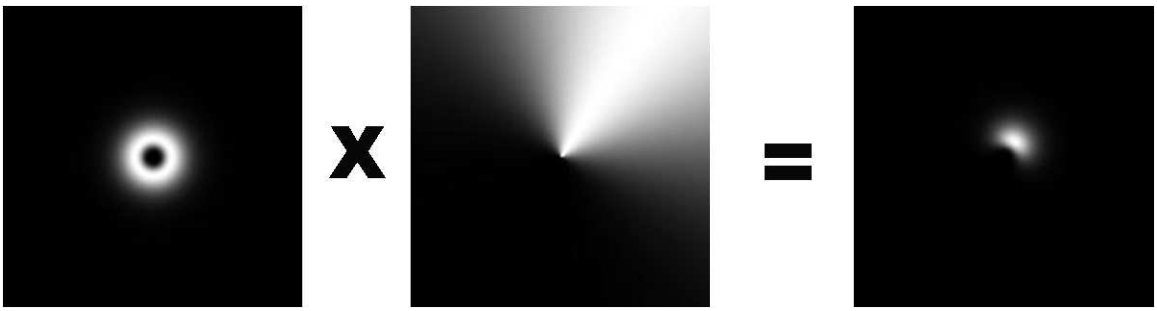


Figure 3.8: Three images are shown of a single log-Gabor filter. The first is the frequency scaling operation defined by equation (3.9) and the second is the orientation defined by equation (3.10). The last is the multiplication of the filters defined by equation (3.11). All pictures are magnitudes in the frequency domain.

Sadly, the singularity of the log function in the Fourier domain prohibits the definition of a closed form spatial solution. Additionally, we must talk about the bandwidth of each filter in terms of octaves due to the log scale.

We begin our log-Gabor decomposition by filtering the image with filters at five scales and four orientations. The value of σ_s/f_o is set to 0.65, which translates to each filter spanning between one and two frequency octaves. The bandwidth of each frequency scale is chosen to minimize overlap between scales, but still uniformly cover most of the frequency spectrum (as in the mammalian visual system). In the spatial domain this translates numerically to convolving the image with log-Gabors of pixel

size 3, 6, 13, 27, and 61, each at four different orientations. Figure 3.9 shows the frequency domain equivalent of all the log-Gabor filters used in the analysis.

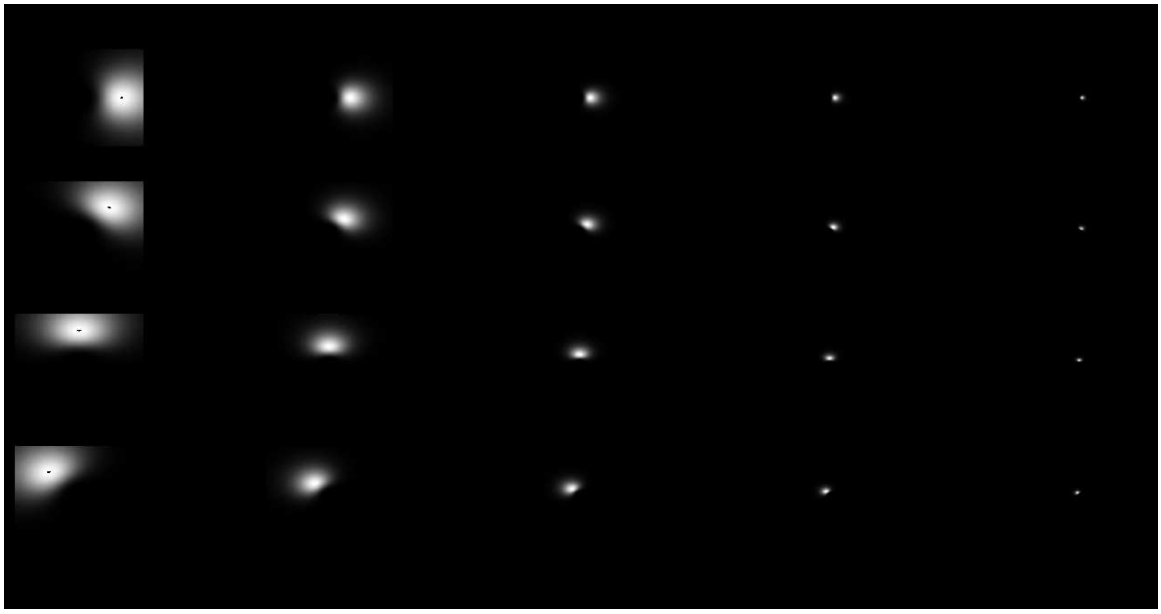


Figure 3.9: All scales and orientations of the log-Gabor filters are shown. Each is a combination of the scale and orientations as seen in the example from Figure 3.8

We can add the filter bank outputs to get the entire frequency domain coverage. This is shown in Figure 3.10. Notice that it only covers one half of the frequency spectrum. This may seem odd at first glance, but in actuality the entire domain *is* covered. This can be explained by looking at the individual filters.

The individual filters are shown in Figure 3.9 and are only one sided, not symmetric. By using non-symmetric filters in frequency we actually save some computation time but cover the entire spectrum - it performs two filtering operations with only one Fourier transform. We can achieve the one sided filters shown by adding two *symmetric* filters of the same bandwidth, as shown in Figure 3.11. One filter is even symmetric and the other is odd symmetric in frequency.

In the frequency domain, an even filter (with no orientation) is the equivalent of a completely real spatial domain function. By the same respect, an odd frequency filter (with orientation on the imaginary line) is the equivalent of a completely imag-

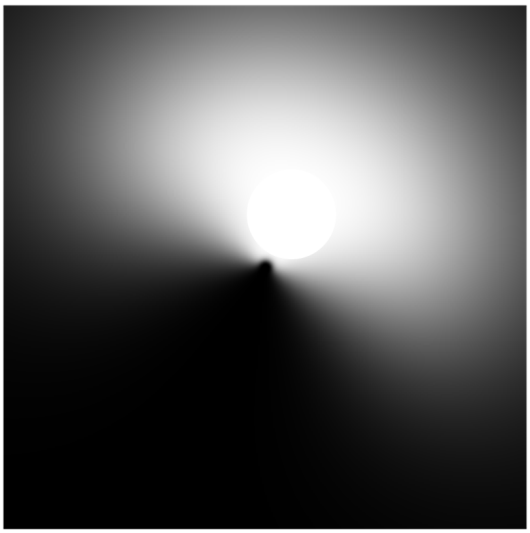


Figure 3.10: All scales and orientations of the log-Gabor filters are added together to form the complete coverage of the log-Gabors.

inary spatial domain filter. If we multiply the odd filter by j and add them, we can get a non-symmetric filter. This means that when we filter and perform an inverse transform using the non-symmetric filter, the spatial domain output will have a real and imaginary component. The real part is the result of one filter; the imaginary is the result of another filter (because it was multiplied by j before being added).

Since each non-symmetric filter result has same orientation and bandwidth, you may also be wondering what the difference between the real and imaginary result is. The real component is a log-Gabor that is even symmetric (in the spatial domain), G_{even}, and the imaginary component is odd symmetric (in the spatial domain),G_{odd}. In this way, the one sided filter gives us results from even and odd log-Gabor convolutions. Odd log-Gabors tend to capture edges well and even log-Gabors tend to capture bars well. This can be seen in Figure 3.12. The even convolution will be maximized when it overlaps a bar of about the same size and the odd convolution will be maximized when it overlaps an edge.

This is biologically motivated as well, as many believe that the *HVS* is performing

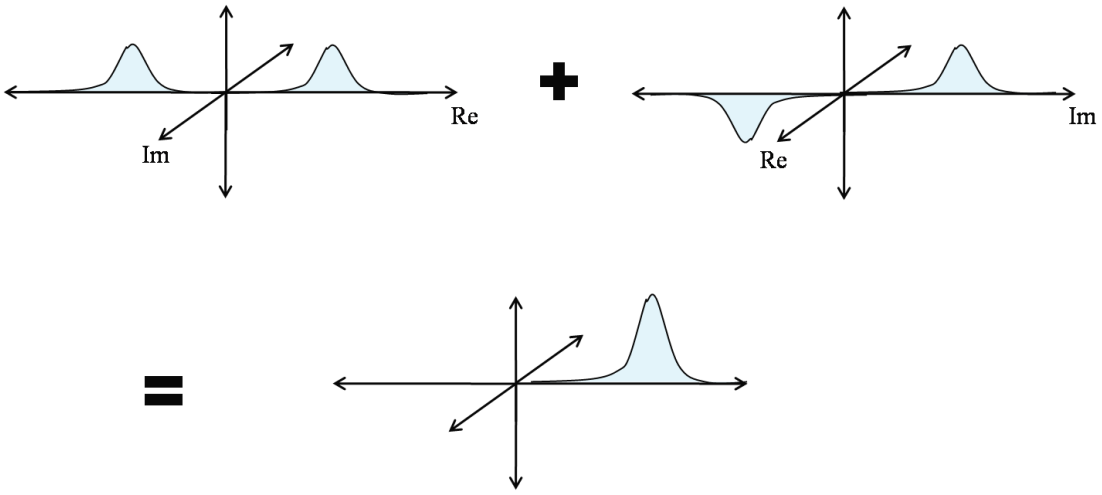


Figure 3.11: All scales and orientations of the log-Gabor filters are added together to form the complete coverage of the log-Gabors.

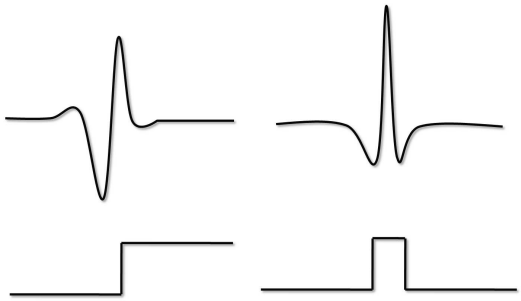


Figure 3.12: The bars and edge equivalents for the log-Gabor filters are shown.

a bar and edge detection operation and that the visual cortex combines them to form everything we see in between. We approximate this by collapsing the even and odd output amplitudes into a single magnitude, $G_{out} = \sqrt{G_{even}^2 + G_{odd}^2}$. This creates a log-Gabor filter bank with 20 magnitude outputs (5 scales x 4 orientations).

At this point we have 20 analysis images (the filter outputs) for the reference and distorted image. Each analysis image is the output of a filter at one scale and orientation. We break up the analysis images further into 16x16 overlapping patches, p. For each patch in the reference and distorted images, this results in 20 filter analysis patches. We then compute the statistics in each of the reference and distorted image patches. Namely, we compute the standard deviation, skewness, and kurtosis of each

analysis filter patch (that is 20 values of standard deviation, skewness and kurtosis per patch). We denote the standard deviation of the pth reference image patch as $\sigma_{so}^{ref}(p)$, where $_{so}$ is the scale and orientation of the log-Gabor filter where the statistic was calculated. Similarly, the skewness can be represented as $\xi_{so}^{ref}(p)$ for the reference image and the kurtosis as $\kappa_{so}^{ref}(p)$. The distorted image statistics are signified by $\sigma_{so}^{dst}(p)$, $\xi_{so}^{dst}(p)$, and $\kappa_{so}^{dst}(p)$.

(a) (b)

Figure 3.13: An example of a reference image and highly distorted version from the LIVE database are shown. The distortion image contains *JPEG2000* artifacts.

The statistics of the log-Gabor filter outputs have been widely used to define visual appearance and texture. Specifically, the change in standard deviation, skewness, and kurtosis have been shown to be good indications of discriminable texture statistics[30]. To capture this characteristic, we compute absolute differences in patch statistics of overlapping portions for the reference and distorted images according to equation (3.12). This results in *difference* measures of local standard deviation, skewness, and kurtosis for each scale and orientation from the filter bank outputs. We take a weighted sum of the statistical differences between the reference and distorted image. For the pth block in the image,

$$\eta(p) = \sum_{s \in S_p, o \in O_p} w_s \left[|\sigma_{so}^{ref} - \sigma_{so}^{dst}| + 2|\xi_{so}^{ref} - \xi_{so}^{dst}| + |\kappa_{so}^{ref} - \kappa_{so}^{dst}| \right] \tag{3.12}$$

where S_p and O_p are the set of all scales and orientations in patch p. σ_{so}, ξ_{so}, and κ_{so}

are the standard deviation, skewness, and kurtosis of the patch at scale s and orientation o defined previously. Note that the skewness difference is multiplied by a factor of 2 to bring it on approximately the same scale as the σ and κ differences. Notice also that each scale is multiplied by a different weight, w_s. This is to account for the differences in perception of statistical distortions at different frequency octaves. This weighting takes advantage of the observation that low frequency distortions are perceptually more annoying than high frequency distortions. See Figure 1.4 for an example of this. The values chosen for w_s from the finest to coarsest scale are 1, 2, 6, 10, and 12. These parameters were chosen for best performance. They are not optimized, however, but simply selected as the best performing integer weights. We can also represent the integer weights as percentages. The finest scale contributes 3.2%, and increasing scales contribute 6.4%, 19.4%, 32.2%, and 38.7%. In this format, one can immediately notice the similarity between the chosen weights and the weights recommended for lossy *JPEG2000* encoded images. Even so, the weights presented here are an area of further research. Adjustment of w_s can greatly effect the performance of the low quality metric.

At this point, the statistical differences accumulated in each patch are summed together to form a stuctural difference map, $\eta(p)$. To combine the accumulated differences of each patch we use the vector two norm,

$$Q_{low} = \frac{1}{N}\sqrt{\sum_{p \in P_N} \eta(p)^2} \tag{3.13}$$

where P_N is the set of all patches in the image and N is the total number of image patches. This results in a single value of quality for highly distorted images where zero denotes the highest quality and increasing values correspond to increasingly worse image quality. Again, using the two norm weights larger statistical variations more when collapsing the map (because they are squared; large values become larger).

Figure 3.13 shows an overview of the entire process for low quality using an ex-

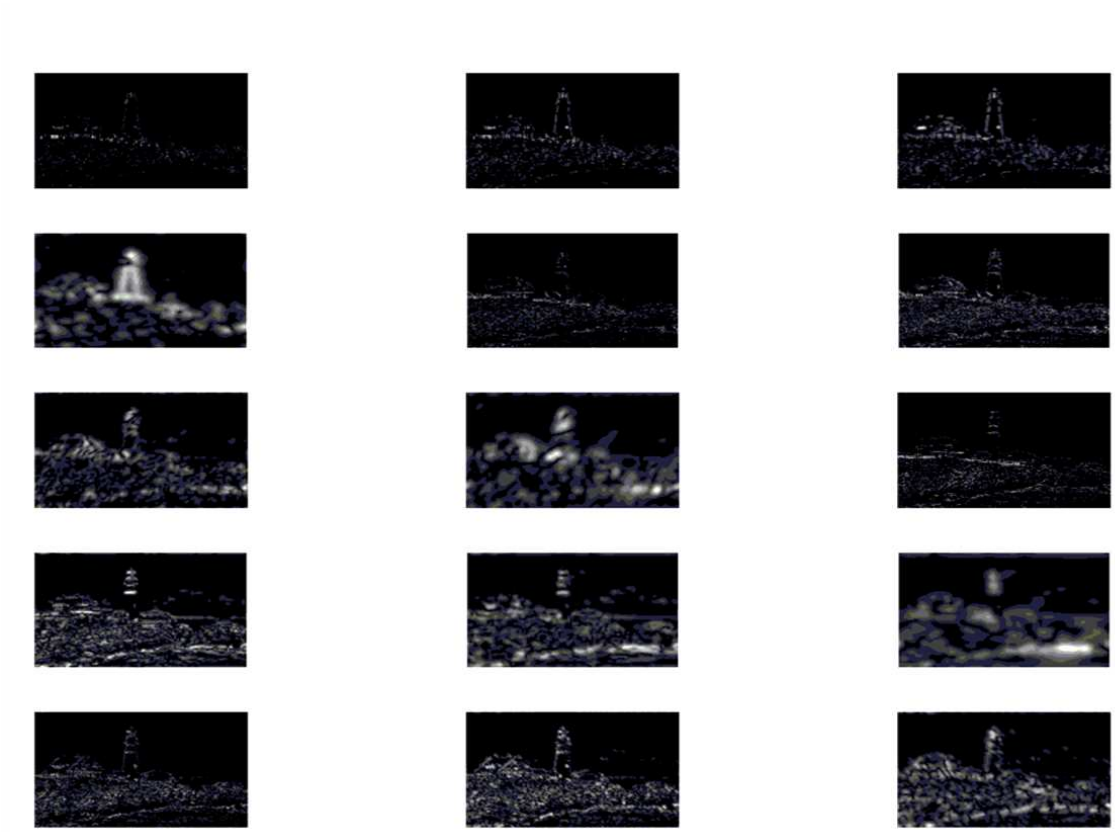


Figure 3.14: Various outputs from the log-Gabor filter bank outputs of the light house image are shown. Statistics of these analysis filters are used to create a difference map.

ample distorted image. The images are from the *LIVE* database[26]. The distorted version contains *JPEG2000* artifacts, namely ringing and blurring. Notice that the most disturbing artifacts occur in the sky of the lighthouse image.

The images are first loaded into memory. Then, images are analyzed using the log-Gabor filter bank. Some of the resultant analysis images are shown in Figure 3.14. Then the statistical differences between the patches of the filter outputs are calculated and summed together to form the statistical difference map, shown in Figure 3.15. Notice that the map captures the most annoying artifacts, i.e.- ringing around the hard edges in the image and blurring in the sky.

This map is collapsed using the vector two norm of equation (3.13), resulting in a

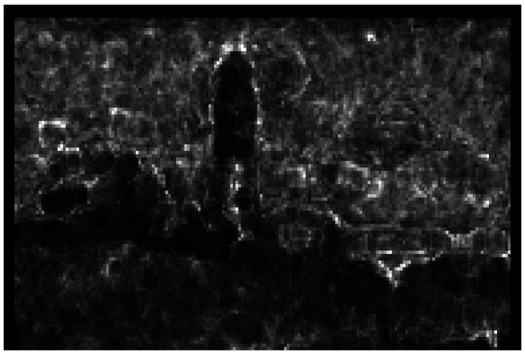


Figure 3.15: An example map of the statistical differences in an image are shown. The staistica are calculated in the log-Gabor analysis domain.

measure of fidelity for the low quality image. Thus, we have a single numeric value of fidelity for a low quality image. This is the complete methodology for the low quality metric portion of *MAD*. At this point, we still need to combine both metrics into a single measure. The next section discusses this aspect.

3.3 Combining the Two Strategies

At this point we have two metrics of quality that must be combined into a single measure based upon how apparently distorted the image is. High quality images should attain their value mostly from Q_{high} and low quality images from Q_{low}. We argue that in the transition between high and low quality assessment, observers use a mixture of strategies. This makes sense in images like those compressed with *JPEG* or *JPEG2000*, for example. Some regions are of very high quality. Other regions may contain highly visible blocking or ringing artifacts that disrupt appearance. An example transition image can be seen in Figure 3.16. Clearly, the bridge has statistical appearance changes (the wall is completely changed), but the trees and sky are largely masking the distortions. In addition, the middle portion of the bridge looks mostly normal, with some statistical changes and some masking.

To capture the interacting strategies, we propose using an α-blending of the two metrics according to the output of the Q_{high} quality measure. This makes sense as

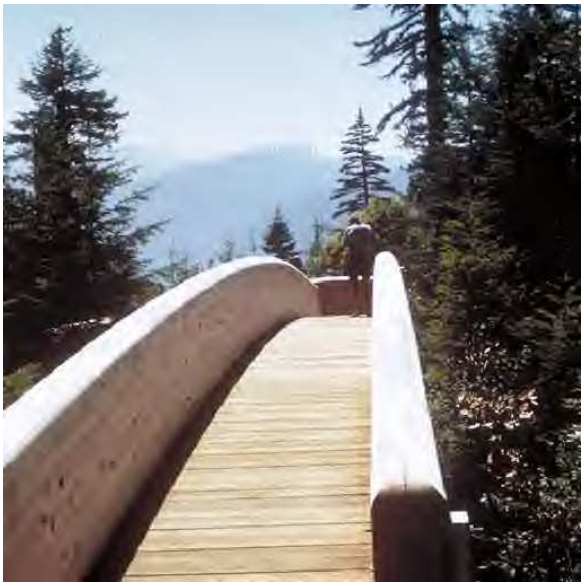


Figure 3.16: An example transition image with *JPEG2000* compression artifacts is shown. Clearly, the bridge has statistical appearance changes (the wall is completely changed), but the trees and sky are largely masking the distortions. And middle part of the bridge looks mostly normal.

Q_{high} will be a rough indication of how high the quality of the image is. When Q_{high} is low, MAD should mostly attain its value from Q_{high}. Otherwise, it should attain most of its value from Q_{low}.

Before blending, the two metrics must be brought to approximately the same scale. To do this, we take the $\log_{10}(\cdot)$ of each metric. This operation makes the two datasets appear as if they came from the same dataset when plotted versus the actual image quality. If both dataset were plotted without the log (i.e. - a linear scale), then there exists too much of a gap between the metrics. See Figure 3.17 for a graphical interpretation of the log scaling operation. Therefore, the metrics become,

$$LQ_{low} = \log_{10} Q_{low} \text{ and } LQ_{high} = \log_{10} Q_{high} \tag{3.14}$$

Once on the same scale, we argue that we can capture the transition aspects by blending the metrics together with a dual parameter sigmoid. We define the sigmoidal

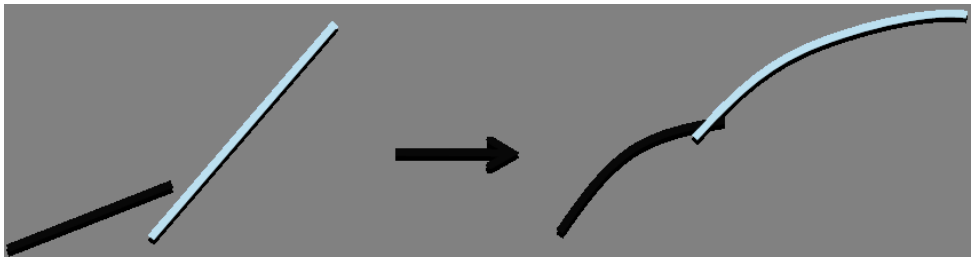


Figure 3.17: An example of how the log operation makes the two datasets appear as if they are from the same dataset. If each is plotted verse the perceived quality, the datasets might look like the leftmost lines. The log operation turns thos lines into the rightmost intersecting lines, which are more like they form one dataset when plotted together.

α-blending scheme according to LQ_{high}:

$$\Lambda(LQ_{high}) = \frac{1}{1 + \exp \frac{LQ_{high} - \tau_1}{\tau_2}} \tag{3.15}$$

where τ_1 and τ_2 are free parameters that may be tuned to a specific database. We will use values of -1.2 and 0.6 for τ_1 and τ_2, respectively. Later on, we will show that the specific values are not crucial, as long as they define a gradual blend of the strategies. Figure 3.18 shows an example of Λ plotted versus LQ_{high}. As shown, τ_1 defines the threshold value for blending the metrics. τ_2 defines how soft the threshold is. Too hard of a threshold will not capture the blending of strategies in transition. Too soft will use an inappropriate strategy in the high or low quality range.

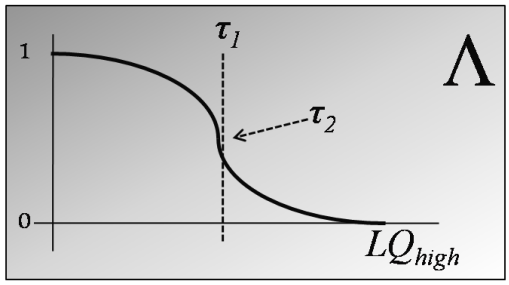


Figure 3.18: An example of the sigmoidal blending function as a function of LQ_{high} is shown. τ_1 and τ_2 control the threshold value and threshold softness for blending the metrics.

We then present the final output of the MAD fidelity metric as:

$$MAD = 10^{\left(\Lambda \times LQ_{high} + [1-\Lambda] \times LQ_{low}\right)} \tag{3.16}$$

where the metric is raised to a power of 10 to get rid of the non-linearity induced by the log scaling operation. Note that MAD is a metric of fidelity, not quality. If the distortions are not visible, it returns a value of one. This is an important property in many applications. Increasing values indicate decreasing quality.

3.4 Building a New Subjective Quality Database

This section describes the methods used in creating a new quality database for various images. The database is currently a preliminary model with ten images chosen to fit into five categories. We are currently working to increase the number of observers involved in the database and the total number of images. Eventually we plan to have many images assigned by category, thus the name Categorized Subjective Image Quality (CSIQ, pronounced sea-sick).

3.4.1 Subjective Ratings of Perceived Distortion

This section provides a summary of the psychophysical scaling experiment which was performed to obtain subjective ratings of visual fidelity for images containing a variety of distortion types.

Stimuli: Ten natural images, obtained from the National Park Image Archives [32]served as original images in this study. The digital images were of size 512×512 pixels and were 24-bit RGB with each color channel ranging from 0-255. These images were distorted with six types of distortions:

1. Additive Gaussian white noise (group NOZ).
2. Baseline JPEG compression of the image using the standard quantization matrix (group JPG).

3. JPEG-2000 compression of the image (group JP2).

4. Blurring by using a Gaussian filter (group BLR).

5. Additive Gaussian pink noise, $1/f$ noise (group FNZ).

6. Global Contrast reduction (group CST).

The distortions were generated such that five different PSNR levels were achieved for each distortion type and each image. PSNR was calculated in the RGB color space. For the compression-type distortions, these criterion PSNR levels were met by adjusting the granularity of the quantizer; for the blurring, pink noise, and white noise, the levels were met by adjusting the standard deviation of the underlying Gaussian. The distortion was applied to each color channel separately and with the same magnitude. Thus, a total of 310 images were tested: 10 original images and 300 distorted images (10 images $\times$ 6 distortion types $\times$ 5 PSNR levels). Groups NOZ, FNZ, BLR, and CST had target PSNR levels of 20dB to 38dB spaced logarithmically. Groups JPG and JP2 had target PSNR levels of 20dB to 33dB spaced logarithmically. The different spacing for compression type images was chosen due to the high quality of many images even at a PSNR of 33dB. We did not want the visually lossless images to overpower the database.

Subjective Ratings: Subjective ratings of visual fidelity were obtained from four adult imaging-expert observers by using a continuous rating system in which each original image was tested against the distorted versions of that image. The images were placed on a four monitor calibrated display array, with solid gray background; the original was fixed in position at one end of the display array, and subjects were instructed to position the distorted images such that the pixel displacement between each distorted image and the original was linearly proportional to their subjective assessment of distortion. Thus, images which were placed further away from the original were judged to contain greater perceived distortion (lower visual fidelity) relative

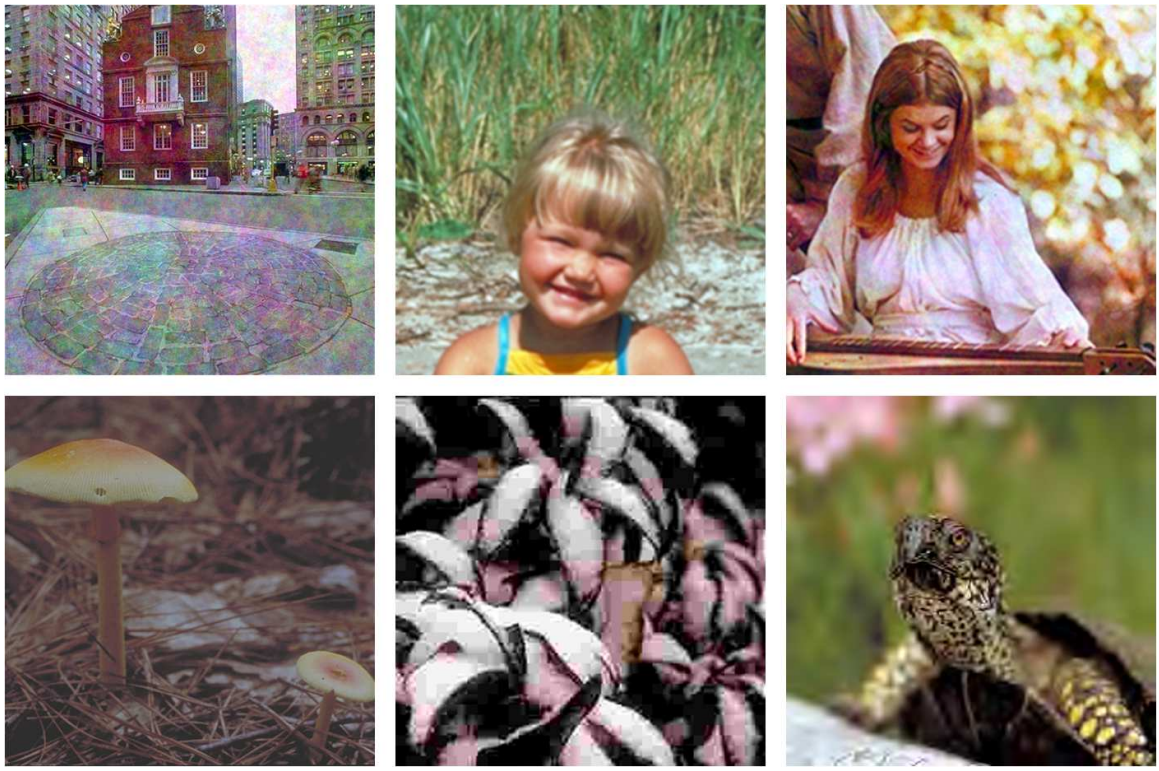


Figure 3.19: Some representative stimuli for the database are shown. Notice that many distortions are included, including unconventional distortions such as additive pink noise and contrast degradations.

to the original. Unlike pairwise comparison procedures, this technique allowed unlimited, simultaneous comparisons of multiple images, and therefore allowed subjects to account for any bias induced by previous judgments; subjects frequently made adjustments to previous placements, and all subjects performed a final pass to ensure their satisfaction in the results.

The four monitors were carefully calibrated to ensure the same appearance of the images across the entire array. Even so, observers were asked to make comparisons between images on the same display whenever possible. Stimuli were displayed on four high-resolution, Sceptre X 22WG 22-inch LCD monitors with resolution of 1680x1050. The displays yielded minimum and maximum luminance of respectively, 0.9 and 190 cd/m^2, and an overall gamma of approximately 2.2; luminance measurements were made by using a Minolta CS-100A photometer (Minolta Corporation, Tokyo, Japan).

Stimuli were viewed binocularly through natural pupils in a darkened room at a distance of approximately 46 cm (about one arm length) under D65 lighting.

Upon completion of the experiment, all subjects performed an additional realignment using the same procedures, but with simultaneous presentation of all 310 images. The original images were aligned and fixed in position at one end of the display array, and subjects were instructed to position the distorted images such that the physical displacement between each distorted image and its original was linearly proportional to perceived distortion. This step provided the necessary factor by which to scale the within-image results to obtain between-image scores. The raw scores for each observer on each image were converted to z-scores (zero-mean and unit variance scores) and the average z-scores over all subjects were scaled to span the range [0, 1.0], where a score of 1.0 corresponded to the image containing the greatest perceived distortion (i.e., the image placed furthest from the original). These within-image scores were then scaled to between-image scores, and then the average scores over all observers and over all images were rescaled to span the range [0, 100]. Some representative stimuli for the *CSIQ* database can be seen in Figure 3.19.

CHAPTER 4

RESULTS

The results obtained by MAD are reported on two databases of subjective image quality. The LIVE database[33] and CSIQ database. Further information on the CSIQ database can be found in section 3.4.

4.1 Results on LIVE Fidelity Database

The LIVE database contains 29 original images, 26 to 29 distorted versions of each original image, and subjective ratings of fidelity (differential mean opinion score, DMOS values) for each distorted image. The distortions present in the database were: Gaussian blurring, additive white noise, JPEG compression (DCT based), JPEG2000 compression (wavelet based), and simulated data packet loss of transmitted JPEG2000 compressed images. For the remainder of this section we will refer to the following groups:

1. Group ALL, the set of all 779 distorted images in the LIVE database.
2. Group JP2, the set of 169 JPEG2000 compressed images.
3. Group JPG, the set of 175 JPEG compressed images.
4. Group NOZ, the set of 145 addititive white noise compressed images.
5. Group BLR, the set of 145 Gaussian blurred images.
6. Group RAY, the set of 145 simulated fast fading packet loss JPEG2000 compressed images.

The DMOS values were computed by averaging z-scores obtained from subjective ratings of fidelity on a continuous linear scale that was divided into five equal regions

labeled "Bad," "Poor," "Fair," "Good," and "Excellent." Approximately 20 to 29 human observers rated each distorted image. Note that the DMOS values were provided as part of the LIVE image database; they were not experimentally determined nor verified in the current study.

The images ranged in size from 408×704 pixels to 768×512 pixels: Six of the images were of size 408×704, three of the images were of size 640×512, and 14 of the images were of size 768×512; the remaining six images ranged in size from 608×416 to 608×480. Grayscale versions of the original and distorted images were obtained via a pixel-wise transformation of I = 0.2989 R + 0.5870 G + 0.1140 B, where I, R, G, and B denote the 8-bit grayscale, red, green, and blue intensities, respectively.

We will compare our results to those obtained by PSNR, MSSIM, VSNR, and VIF on LIVE. Before comparing performance on the database it is customary to apply a logistic transform to each metric that brings it on the same scale and linearity as the DMOS values. The logistic transform recommended by the Video Quality Experts Group (VQEG) is a four parameter sigmoid [34]:

$$f(x) = \frac{\tau_1 - \tau_2}{1 + \exp \frac{x - \tau_3}{\tau_4}} + \tau_2 \tag{4.1}$$

The parameters τ_1, τ_2, τ_3, and τ_4 are chosen such that they minimize the MSE between the DMOS values and metric output. The parameters are found using the Nelder-Mead downhill simplex method at many random starting parameters [35]. The logistic transform is completely monotonic and chosen mainly for its ability to bring the datasets to the same scale. Once applied, it enables the use of various performance measures.

The first measures of performance reported are the Pearson correlation coefficient (CC) and Spearman rank order correlation coefficient(ROCC). The Pearson CC provides a measure of how similar two sets of data are. The Spearman ROCC assigns a rank to increasing values in each dataset and then calculates the CC. In this way, it provides a measure of how similar each metric ranks the ordering of image qualities. It

is important to note here that the CC and Spearman ROCC are not perfect measures of similarity in regression problems. The CC is only a good comparison measure for regression when the compared fits satisfy the property of residual homoskedasticity (also known as homogeneity of variance). Basically, we assume that the variance of the residuals is independent of the x value; no portion of the data is better fit than any other portion. Because of this we will report on various measures of performance, not just CC. Also note that the CC does not provide a linear measure of similarity. For instance, a CC of 0.8 does not mean that the fit is twice as good as a fit with CC 0.4.

Table 4.1: This table contains a comprehensive review of the performance of each metric for group ALL in the LIVE database. The group contains 779 ratings of quality. The Correlation Coefficient (CC), Spearman rank-order correlation coefficient(SROCC), root mean-squared error(RMSE), outlier ratio, R_{out}, and sum outlier distance, R_{SOD}, are given.

ALL	*PSNR*	*SSIM*	*VSNR*	*VIF*	*MAD*
CC	0.8707	0.9378	0.9233	0.9595	0.9700
SROCC	0.8763	0.9473	0.9278	0.9633	0.9705
RMSE	375.03	264.84	292.90	214.92	185.43
R_{out}	0.682	0.592	0.588	0.546	0.399
R_{SOD}	4943.3	2814.1	3246.8	1890.4	1270.3

Table 4.1 shows the CC, Spearman ROCC, and root mean squared error (RMSE) for different metrics on group ALL from the LIVE database. The RMSE is defined from squared differences between the DMOS and logistically transformed metric output. Also shown in the table are two measures, the outlier ratio, R_{out} and the sum outlier distance, R_{SOD}. These two measures attempt to account for the inherent variation in human subjective ratings of quality. If the perceived quality of a particular image has large variation between observers, then the *average* DMOS rating is not necessarily a good measure of what the metric should predict. Instead, some leeway should be given around certain ratings. The outlier ratio is defined as [1]:

$$R_{out} = \frac{N_{false}}{N_{total}} \tag{4.2}$$

where N_{false} is the number of predictions outside two standard deviations of the average observer and N_{total} is the total number of predicted qualities. In this way, we have a percentage for the metric predicting quality inside a range of inherent error bars. Two standard deviations was chosen because it ensures inclusion of 95% of the observer's scores.

The R_{SOD} is a new performance measure that we have devised. In addition to knowing if the metric is inside the error bars, we want to know how close the metric is from the closest error bar when it fails. Mathematically,

$$R_{SOD} = \sum_{x \in X_{false}} \min |f(x) - [DMOS(x) \pm 2\sigma_{obs}(x)]| \tag{4.3}$$

where σ_{obs} is the standard deviation of ratings from different observers and X_{false} is the set of all quality ratings outside $2\sigma_{obs}$.

As seen from table 4.1, MAD performs superiorly to all other metrics with respect to every performance measure. Note that these figures do not report if MAD has statistically significant performance. Significance involving regression is somewhat tricky and is given a separate subsection. To the best of our knowledge, these are the best performance measures achieved on the LIVE database to date.

It is also interesting to look at the performance on a per distortion rate. Some metrics can capture specific distortions better than others. To measure this, each metric was fit logistically to the DMOS ratings of a particular distortion category. Tables 4.2 through 4.6 show performance for each distortion group. Observe that for group JP2, MAD also has the best performance across all measures. For group JPG, MAD performs slightly worse than VIF, but better than all other metrics.

For group NOZ, the best performing metric is PSNR, followed by MAD for all performance measures except SROCC. In terms of rank order, VIF performs superiorly, followed by PSNR, followed by MAD.

For group BLR, VIF outperforms every metric, followed by MAD. It is of note that the performance of VIF in this category is quite substantial. MAD is a (comparatively) distant second when predicting quality of Gaussian blurred images.

For group RAY, VIF performs the best in all categories except outlier ratio, where MAD performs best. It seems that MAD predicts subjective quality well in this region, but when it guesses incorrectly, the answer is comparatively far from the DMOS. The statistical significance of these performances can be seen in the next section.

Table 4.2: This table contains a comprehensive review of the performance of each metric for group JP2 in the LIVE database. The group contains 169 ratings of quality. The measures are the same as those given in table 4.1.

JP2	*PSNR*	*SSIM*	*VSNR*	*VIF*	*MAD*
CC	0.8993	0.9662	0.9629	0.9771	0.9785
SROCC	0.8947	0.9607	0.9560	0.9694	0.9704
RMSE	143.44	84.60	88.52	69.81	67.65
R_{out}	0.604	0.467	0.485	0.385	0.325
R_{SOD}	780.1	304.4	312.3	183.5	174.8

Table 4.3: This table contains a comprehensive review of the performance of each metric for group JPG in the LIVE database. The group contains 175 ratings of quality. The measures are the same as those given in table 4.1.

JPG	*PSNR*	*SSIM*	*VSNR*	*VIF*	*MAD*
CC	0.8883	0.9784	0.9723	0.9863	0.9836
SROCC	0.8814	0.9755	0.9657	0.9844	0.9809
RMSE	193.51	87.14	98.43	69.56	76.04
R_{out}	0.634	0.394	0.457	0.257	0.303
R_{SOD}	1173.0	249.0	330.5	137.5	182.8

To further our comparison, the graph of logistic MAD and the next best performing metric versus DMOS are shown in Figure 4.1, Figure 4.2, and Figure 4.3 for the overall and per distortion groupings. Notice from the overall fit (Figure 4.1) that MAD fails on a subset of about seven high quality images. Also notice from the compression distortion fits (Figure 4.2) that all residuals appear to be homoskedastic except for group RAY. MAD is fitting low quality images in group RAY better than

Table 4.4: This table contains a comprehensive review of the performance of each metric for group NOZ in the LIVE database. The group contains 145 ratings of quality. The measures are the same as those given in table 4.1.

NOZ	*PSNR*	*SSIM*	*VSNR*	*VIF*	*MAD*
CC	0.9858	0.9700	0.9777	0.9839	0.9841
SROCC	0.9852	0.9691	0.9785	0.9854	0.9807
RMSE	56.67	81.91	70.78	60.28	59.75
R_{out}	0.262	0.531	0.414	0.352	0.317
R_{SOD}	103.2	282.7	200.3	136.0	118.6

Table 4.5: This table contains a comprehensive review of the performance of each metric for group BLR in the LIVE database. The group contains 145 ratings of quality. The measures are the same as those given in table 4.1.

BLR	*PSNR*	*SSIM*	*VSNR*	*VIF*	*MAD*
CC	0.7836	0.9450	0.9329	0.9744	0.9524
SROCC	0.7813	0.9512	0.9418	0.9731	0.9517
RMSE	138.18	72.75	80.08	49.97	67.83
R_{out}	0.717	0.545	0.476	0.372	0.469
R_{SOD}	830.1	268.5	321.8	112.3	217.5

high quality. This could be an indication that the masking model is failing for this particular distortion type.

Looking at the photographic distortions (Figure 4.3), MAD tends to have a non-linear relationship with Gaussian blurring. Again high quality images are not fit as well as low quality images. This may be partly the fault of the masking model used. However, notice that MAD predicts the fidelity of high quality white noise distorted images extremely well. In light of this, we may be able to further improve MAD using a distortion adaptive masking model.

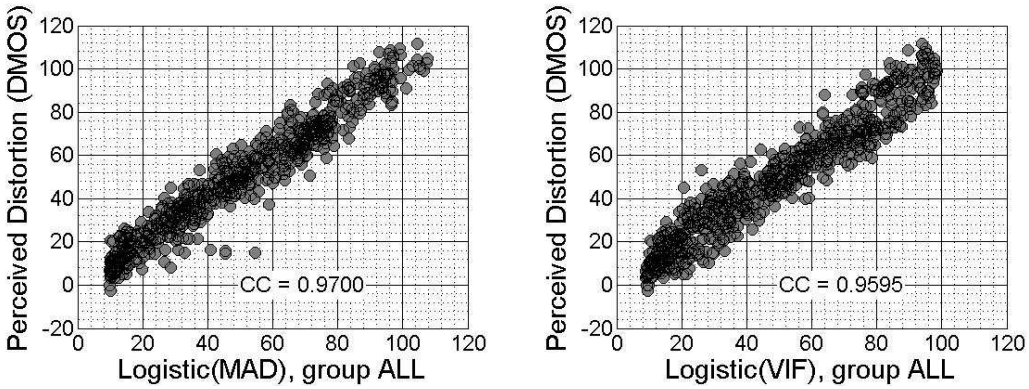
4.1.1 Statistical Significance on LIVE

When we talk about statistical significance of two metrics fitting to DMOS we usually approach it like a regression analysis: Which set of data better fits the observations and is the fit significant?

To establish significance we compare the residuals of each metric. Ideally, all the

Table 4.6: This table contains a comprehensive review of the performance of each metric for group RAY in the LIVE database. The group contains 145 ratings of quality. The measures are the same as those given in table 4.1.

RAY	*PSNR*	*SSIM*	*VSNR*	*VIF*	*MAD*
CC	0.8921	0.9481	0.9023	0.9613	0.9497
SROCC	0.8928	0.9554	0.9054	0.9648	0.9488
RMSE	155.02	109.07	147.90	94.51	107.43
R_{out}	0.566	0.579	0.648	0.441	0.386
R_{SOD}	763.3	399.0	699.7	287.6	338.2



(a) (b)

Figure 4.1: The DMOS ratings are plotted verse the logistic transformed MAD prediction for group ALL in the LIVE database. VIF, the next best performing metric, is also shown.

residuals would equal zero and have no variance. Smaller residual variance denotes a better fit. In addition to this, if the residuals are Gaussian, we can establish the probability that the samples are drawn from two different distributions or from the same Gaussian distribution. Given the number of parameters in each dataset and the variance of each set, we can use the incomplete Beta function, $I_x(z, w)$, to determine the probability that the residuals are drawn from the same underlying distribution. This type of analysis is commonly known as the F-statistic or F-test. We assume that the datasets are part of the same group (the NULL hypothesis) and see with what confidence we can reject this hypothesis.

If the variance of the residuals of one dataset are smaller than another and we can say with 99% that the two datasets are not drawn from the same underlying

distribution, then we can say that one fit is statistically better than the other. Note that if the residuals are not Gaussian, the test for significance is considerably more difficult and many times inconclusive.

If the skewness is between -1 and 1, and kurtosis is between 2 and 4, we assume that the residuals can be deemed Gaussian. However, a more formal test of Gaussianity is the Jarque-Bera Statistic or J.B Test[36]. This test assumes the NULL hypothesis that the dataset is from a Gaussian distribution. If it cannot say with 95% confidence that the dataset is not Gaussian, then we assume Gaussianity. If the JB statistic is sufficiently small enough, then we are safe to assume it is Gaussian. Larger values of the JB statistic denote how far from Gaussian the distribution lies.

Table 4.7 shows the summary for overall statistical performance of each metric. An F-test is performed between each pair of metrics. A "-1" denotes that the metric in the row has statistically larger residuals with confidence greater then 99%. A "0" denotes that there is statistically no difference between the residuals of the two metrics. Table (b) contains the measures of Gaussianity. A zero in the Gaussian row denotes that the NULL hypothesis holds, according to the JB test. Notice that the F-tests conclude that MAD has statistically smaller residuals than every other metric. However, MAD is also deemed as the most non-Gaussian of all the metrics as denoted by the JB statistic and abnormally high kurtosis. It is interesting to look at why MAD has such high kurtosis.

Figure 4.5 shows the histogram of residuals for MAD. Notice that the distribution is mostly Gaussian except for about seven outlying images where MAD fails by a large amount. If these images are removed, the residuals can be deemed Gaussian with p value of 0.42 to reject the NULL hypothesis and JB statistic = 1.2. MAD rates these images as being low quality, even though they have fairly low DMOS values. All of the DMOS values are around 15, while MAD predicts that they have a rating of around 50. Figure 4.4 shows the two images where MAD performs the worst. Both

Table 4.7: Table (a) shows the statistical relationships between the metrics in the rows and columns on group ALL of the LIVE database. A "1" denotes that the metric in the row has statistically smaller residuals with confidence greater then 99%. A "-1" denotes that the metric in the row has statistically larger residuals with confidence greater then 99%. A "0" denotes that there is statistically no difference between the residuals of the two metrics. Table (b) contains the measures of Gaussianity. A zero in the Gaussian row denotes that the NULL hypothesis holds, according to the JB test. A one denotes that with 95% confidence the residuals are not Gaussian. Also reported are the p-value confidences and residual skewness and kurtosis.

(a) Significance Table

ALL	*PSNR*	*SSIM*	*VSNR*	*VIF*	*MAD*
PSNR	0	-1	-1	-1	-1
SSIM	1	0	1	-1	-1
VSNR	1	-1	0	-1	-1
VIF	1	1	1	0	-1
MAD	1	1	1	1	0

(b) Measures of Gaussianity

ALL	*PSNR*	*SSIM*	*VSNR*	*VIF*	*MAD*
Gaussian	1	0	1	0	1
Conf.	0.007	0.257	0.001	0.081	0.001
JB Stat	11.768	2.583	20.011	4.843	246.610
Skew	0.292	-0.139	0.091	0.170	-0.518
Kurt	3.143	2.957	3.764	2.818	5.554

are from group RAY. The fundamental failure of MAD for these images occurs in the masking model used. Each image contains heavy ghosting, but it is largely masked. Our model predicts that the the distortions are visible, but the LMSE gives them such weight that MAD assesses quality using the log-Gabor structural content, which is a large value due to the change that ghosting induces. Perhaps a better model of masking here would help to correctly classify the images as being below the detection threshold.

Tables 4.8 through 4.12 show the statistical significance of all metrics per distortion group. For group JP2 and JPG, MAD, SSIM, and VIF are statistically the same and have the best performance. For group NOZ, MAD, VIF, PSNR, and VSNR are statistically the same. Only VSNR does not perform better than SSIM. For group

BLR, VIF performs significantly better than all other metrics. For group RAY, MAD, SSIM, and VIF perform the best and are statistically no different from one another.

Table 4.8: Residual significances and statistics are shown for group JP2 in LIVE database, containing 169 images. The measures are the same as those explained in table 4.7

(a) Significance Table

JP2	*PSNR*	*SSIM*	*VSNR*	*VIF*	*MAD*
PSNR	0	-1	-1	-1	-1
SSIM	1	0	0	0	-1
VSNR	1	0	0	-1	-1
VIF	1	0	1	0	0
MAD	1	1	1	0	0

(b) Measures of Gaussianity

JP2	*PSNR*	*SSIM*	*VSNR*	*VIF*	*MAD*
Gaussian	0	0	0	0	0
Conf.	0.406	0.500	0.063	0.500	0.185
JB Stat	1.559	0.908	5.019	0.065	2.832
Skew	-0.207	0.085	-0.007	-0.046	0.005
Kurt	3.224	2.683	3.844	3.026	3.634

Tables 4.8 through 4.12 show significance for each distortion group. Note that because the groups are smaller, it is easier to make a type II error (that there is no significant difference when one actually exists). The only way to be certain is to collect more subjective ratings for more images of the particular distortion type.

A possible explanation for the performance of MAD on LIVE is that it has two additional free parameters that help to fit it to the database from the α-blend sigmoid. This certainly seems like a plausible explanation at first glance. For fairness, we adjusted the parameters of the α-blend sigmoid and noted when the results were significantly better than VIF, significantly worse than VIF, or the same. Figure 4.6 shows the sensitivity of each parameter graphically using three plots. Each plot has a constant bias parameter, from equation (3.15) τ_1 is held constant and τ_2 is adjusted from 0.2 to 1.4 in steps of 0.2. τ_1 takes values -1.5, -1.2, and -0.8 in each plot. Red sigmoids denote that MAD is significantly inferior to MAD with 95% confidence. Blue

Table 4.9: Residual significances and statistics are shown for group JPG in LIVE database, containing 175 images. The measures are the same as those explained in table 4.7

(a) Significance Table

JPG	*PSNR*	*SSIM*	*VSNR*	*VIF*	*MAD*
PSNR	0	-1	-1	-1	-1
SSIM	1	0	0	-1	0
VSNR	1	0	0	-1	-1
VIF	1	1	1	0	0
MAD	1	0	1	0	0

(b) Measures of Gaussianity

JPG	*PSNR*	*SSIM*	*VSNR*	*VIF*	*MAD*
Gaussian	0	0	1	1	0
Conf.	0.500	0.334	0.021	0.003	0.500
JB Stat	1.161	1.879	8.587	19.728	0.052
Skew	-0.199	0.249	0.521	0.678	0.042
Kurt	2.989	3.099	3.305	3.930	3.001

sigmoids denote that MAD is statistically the same as VIF. Green sigmoids denote that MAD is significantly superior to VIF with 95% confidence. No red sigmoids are present in the graphs. Also notice that insignificant blends can be explained because they do not capture enough of each metric in each quality region or apply too hard of a threshold.

Note that it is possible to make MAD inferior to VIF using parameters of the blend. Again, this usually occurs such that the α-blend does not mix the strategies, but rather applies a hard threshold or uses too much of one metric in a quality range that it is not suited for.

4.2 Results on CSIQ Fidelity Database

The CSIQ database consisted of six different types of distortions at five different levels. Details can be seen in section 3.4. Currently, the database includes 10 original images and 300 distorted versions of the images. We are currently working to expand the database, but we wish to report our preliminary findings from four observers. We

Table 4.10: Residual significances and statistics are shown for group NOZ in LIVE database, containing 145 images. The measures are the same as those explained in table 4.7

(a) Significance Table

NOZ	*PSNR*	*SSIM*	*VSNR*	*VIF*	*MAD*
PSNR	0	1	1	0	0
SSIM	-1	0	0	-1	-1
VSNR	-1	0	0	0	0
VIF	0	1	0	0	0
MAD	0	1	0	0	0

(b) Measures of Gaussianity

NOZ	*PSNR*	*SSIM*	*VSNR*	*VIF*	*MAD*
Gaussian	0	0	0	0	0
Conf.	0.500	0.073	0.343	0.500	0.089
JB Stat	0.294	4.563	1.791	1.069	4.131
Skew	0.108	0.059	0.238	0.113	0.003
Kurt	2.952	2.139	3.266	2.645	3.827

will not talk about the individual groups of distortions in the database as the limited number of ratings prohibits a valid statistical analysis. However, we can talk above overall performance on the database. When the database is completed, it will be freely available to others for use. We believe that this is the only way for the research community to work together in solving the problem of image quality. We further push for other researchers to make their databases freely available so that they may be peer reviewed and be open for other researchers to compare results.

Table 4.13 shows the performance of PSNR, SSIM, VSNR, VIF, and MAD on the CSIQ database. Observe that MAD is the superior measure of quality. Also notice that the results are similar to those from the LIVE database, with MAD having the best performance, followed by SSIM and VIF.

The logistic transform of each metric can be seen in figure 4.7. Observe that MAD is obviously the tightest fit. Also note that VSNR has a band of images at high quality that it does not predict accurately. Also note that the residuals of many of the metrics appear to be heteroskedastic, indicating that CC should not be the

Table 4.11: Residual significances and statistics are shown for group BLR in LIVE database, containing 145 images. The measures are the same as those explained in table 4.7

(a) Significance Table

BLR	*PSNR*	*SSIM*	*VSNR*	*VIF*	*MAD*
PSNR	0	-1	-1	-1	-1
SSIM	1	0	0	-1	0
VSNR	1	0	0	-1	0
VIF	1	1	1	0	1
MAD	1	0	0	-1	0

(b) Measures of Gaussianity

BLR	*PSNR*	*SSIM*	*VSNR*	*VIF*	*MAD*
Gaussian	0	0	0	0	1
Conf.	0.243	0.500	0.494	0.401	0.003
JB Stat	2.328	0.336	1.222	1.550	20.758
Skew	0.298	-0.113	0.177	-0.173	-0.613
Kurt	2.826	2.933	3.277	2.629	4.390

primary means of comparison. In that respect, the considerably higher CC achieved by MAD could be in part be attributed to the homoskedasticity of its residuals.

The statistical significance of each metric comparatively can be seen in table 4.14 along with the measures of Gaussianity. In CSIQ as with LIVE, MAD performs significantly better than any other metric. Also notice that without the packet-loss distortions present in the LIVE database, MAD also has highly Gaussian residuals according to the JB statistic. The results from CSIQ are encouraging. However, we caution readers from placing extensive trust in the exact degree of performance. The database is a work in progress and needs more observers and more images before its reliability can be without question.

It is of note that the correlations on the database are markedly lower than on LIVE. We believe this is due to the contrast distortions in CSIQ. None of the metrics were constructed to handle these types of distortions. Even still, MAD performs well, even with the contrast distortions present. To make the measures more valid, we also look at the performance and significance of each metric on CSIQ without the contrast

Table 4.12: Residual significances and statistics are shown for group RAY in LIVE database, containing 145 images. The measures are the same as those explained in table 4.7

(a) Significance Table

RAY	*PSNR*	*SSIM*	*VSNR*	*VIF*	*MAD*
PSNR	0	-1	0	-1	-1
SSIM	1	0	1	0	0
VSNR	0	-1	0	-1	-1
VIF	1	0	1	0	0
MAD	1	0	1	0	0

(b) Measures of Gaussianity

RAY	*PSNR*	*SSIM*	*VSNR*	*VIF*	*MAD*
Gaussian	0	0	0	0	1
Conf.	0.500	0.500	0.239	0.056	0.001
JB Stat	0.079	0.054	2.359	5.258	102.025
Skew	0.044	-0.006	-0.222	0.448	-1.139
Kurt	3.072	2.906	3.439	3.262	6.420

Table 4.13: This table contains a comprehensive review of the performance of each metric for group ALL in the CSIQ database. The group contains 300 ratings of quality. The Correlation Coefficient (CC), Spearman Rank-Order Correlation Coefficient (SROCC), root mean-squared error(RMSE), outlier ratio, R_{out}, and sum outlier distance, R_{SOD}, are given.

ALL	*PSNR*	*SSIM*	*VSNR*	*VIF*	*MAD*
CC	0.8455	0.8893	0.8472	0.9079	0.9507
SROCC	0.8428	0.9019	0.8577	0.9063	0.9484
RMSE	235.65	201.82	234.45	185.01	136.85
R_{out}	0.356	0.305	0.282	0.339	0.228
R_{SOD}	782.3	497.1	737.9	577.8	300.4

images. Table 4.15 shows the performance of each metric and table 4.16 shows the corresponding statistical significances. After this adjustment, MAD performs even better. However, VIF only increases marginally in performance and VSNR becomes statistically indistinguishable from MAD. This development has several repercussions. First, VIF is a decent measure of contrast distortion. Secondly, the performance of VSNR was greatly dampened by contrast distortions. And lastly, MAD is the only metric that has approximately the same performance on both databases. In this way, the reliability of the MAD is further verified.

Table 4.14: Table (a) shows the statistical relationships between the metrics in the rows and columns for the CSIQ database group ALL. A "1" denotes that the metric in the row has statistically smaller residuals with confidence greater then 99%. A "-1" denotes that the metric in the row has statistically larger residuals with confidence greater then 99%. A "0" denotes that there is statistically no difference between the residuals of the two metrics. Table (b) contains the measures of Gaussianity. A zero in the "Gaussian" row denotes that the NULL hypothesis holds, according to the JB test. A one denotes that with 95% confidence the residuals are not Gaussian. Also reported are the p-value confidences and residual skewness and kurtosis.

(a) Significance Table

ALL	*PSNR*	*SSIM*	*VSNR*	*VIF*	*MAD*
PSNR	0	-1	0	-1	-1
SSIM	1	0	1	0	-1
VSNR	0	-1	0	-1	-1
VIF	1	0	1	0	-1
MAD	1	1	1	1	0

(b) Measures of Gaussianity

ALL	*PSNR*	*SSIM*	*VSNR*	*VIF*	*MAD*
Gaussian	0	1	1	0	0
Conf.	0.500	0.003	0.001	0.500	0.425
JB Stat	0.414	17.619	24.255	0.652	1.560
Skew	-0.053	-0.303	-0.569	0.084	0.108
Kurt	3.149	4.026	3.812	3.156	3.281

Figure 4.8 shows the logistic transformed metrics with contrast removed from the database. Notice that MAD and VSNR get noticeably tighter fits. PSNR improves greatly as well, but the overall fit is lackluster.

Also of note is that the α-blend parameters of MAD were held at the same value as on the LIVE database. Tuning these parameters to the CSIQ database results in a CC of 0.9521, R_{out} of 0.2215, and R_{SOD} of 264.76, including the contrast images. Without contrast, tuned MAD can achieve a CC of 0.9659, R_{out} of 0.2200, and R_{SOD} of 215.30. These performance measures clearly delineate MAD as a superior measure of image fidelity prediction.

Table 4.15: This table contains a comprehensive review of the performance of each metric for the CSIQ database group ALL, without group CST. The group contains 250 ratings of quality. The Correlation Coefficient (CC), Spearman Rank-Order Correlation Coefficient (SROCC), root mean-squared error(RMSE), outlier ratio, R_{out}, and sum outlier distance, R_{SOD}, are given.

ALL-CST	*PSNR*	*SSIM*	*VSNR*	*VIF*	*MAD*
CC	0.9178	0.9313	0.9499	0.9263	0.9637
SROCC	0.9185	0.9364	0.9478	0.9294	0.9594
RMSE	166.55	152.80	131.10	158.11	111.94
R_{out}	0.344	0.316	0.240	0.312	0.236
R_{SOD}	447.9	383.3	256.6	387.0	237.1

Table 4.16: The measures submitted here are identical to those used in table 4.7 except performed on the CSIQ database group ALL without group CST.

(a) Significance Table

ALL-CST	*PSNR*	*SSIM*	*VSNR*	*VIF*	*MAD*
PSNR	0	0	-1	0	-1
SSIM	0	0	0	0	-1
VSNR	1	0	0	1	0
VIF	0	0	-1	0	-1
MAD	1	1	0	1	0

(b) Measures of Gaussianity

ALL-CST	*PSNR*	*SSIM*	*VSNR*	*VIF*	*MAD*
Gaussian	0	1	1	1	1
Conf.	0.500	0.001	0.003	0.004	0.044
JB Stat	0.699	26.441	17.912	15.964	6.055
Skew	0.127	-0.172	-0.292	0.408	0.022
Kurt	3.054	4.556	4.174	3.932	3.761

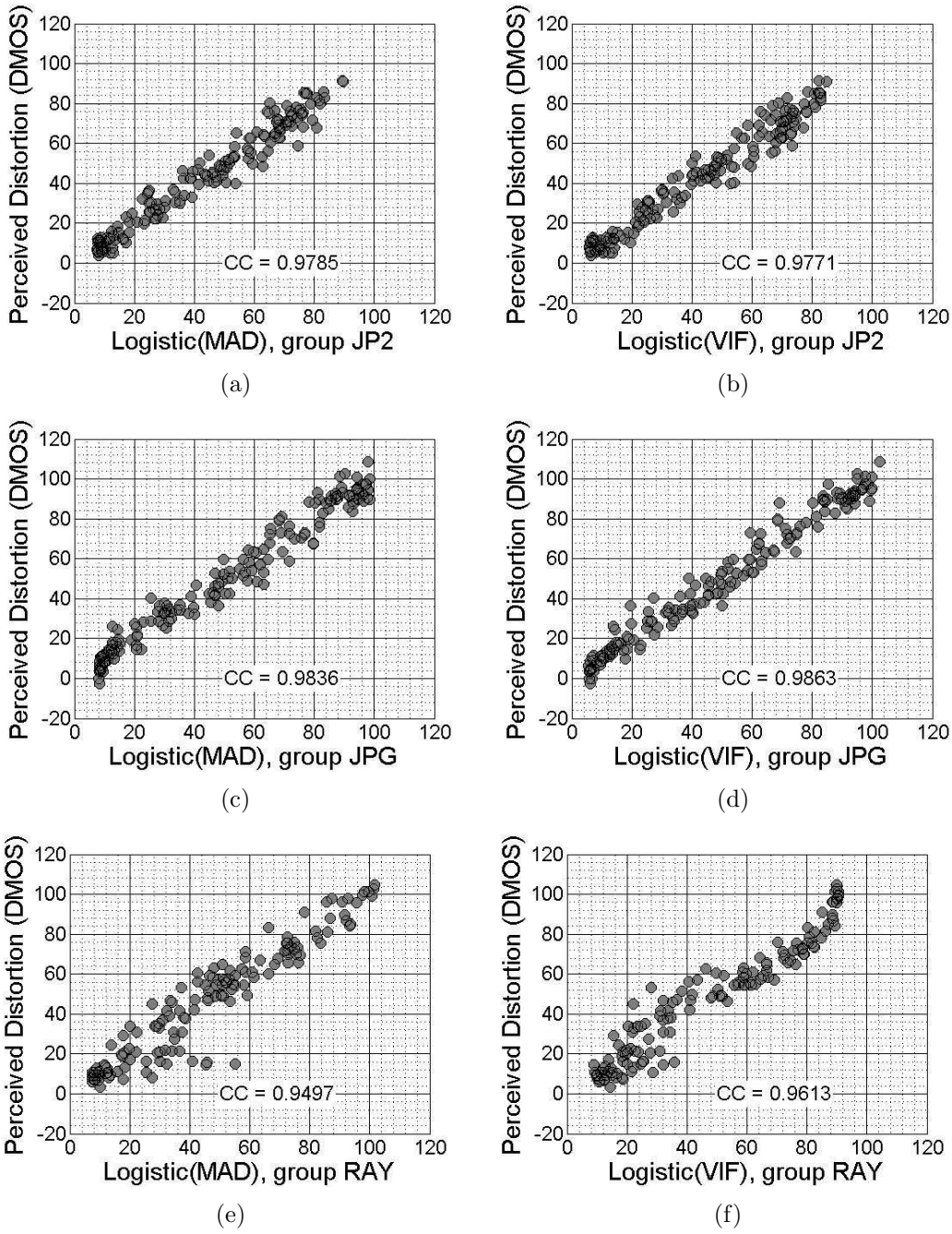


Figure 4.2: The DMOS ratings are plotted verse the logistic transformed MAD prediction for each type of compression distortion in the LIVE database. Also shown are the best or next best (if MAD is best) performing metric.

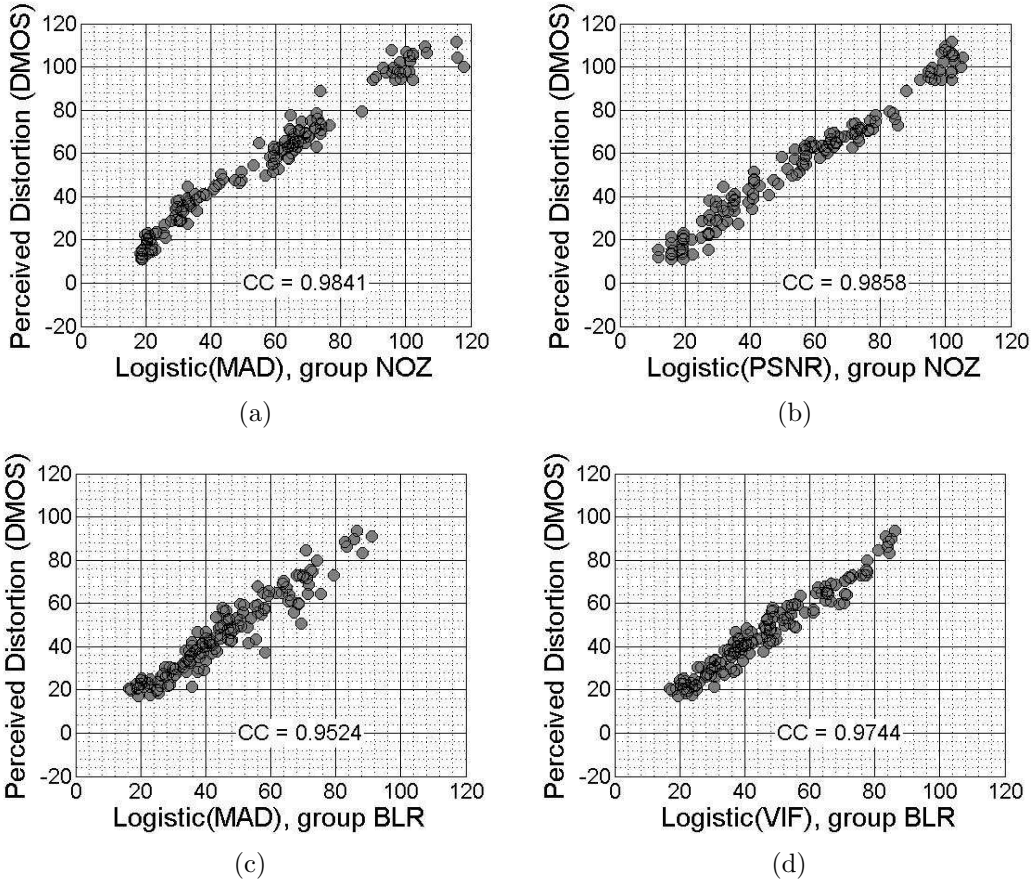
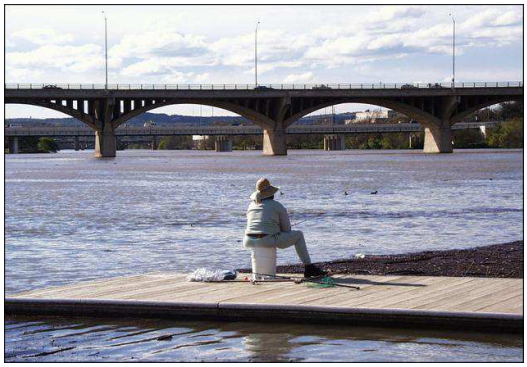
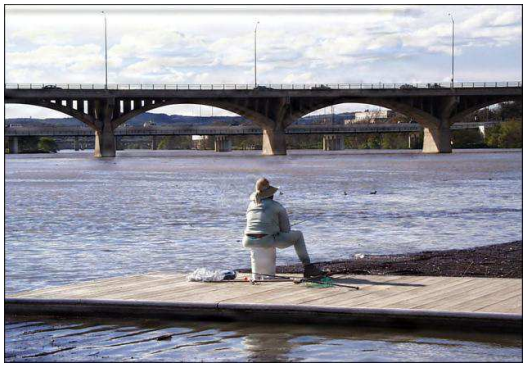


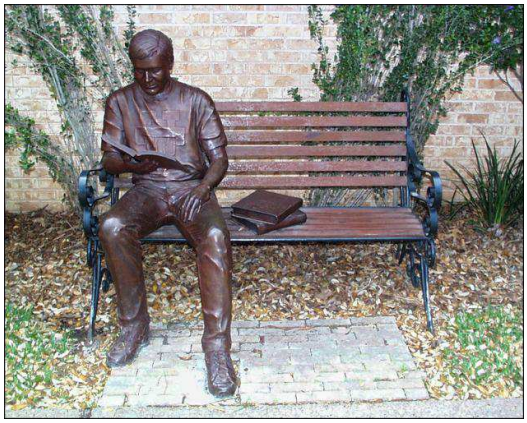
Figure 4.3: The DMOS ratings are plotted verse the logistic transformed MAD prediction for each type of photographic distortion in the LIVE database (Gaussian white noise and Gaussian blurring). These distortions are common in photography. Also shown are the best or next best (if MAD is best) performing metric.



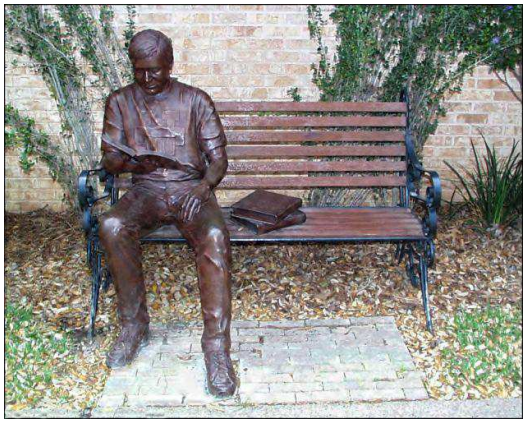
(a) Reference Image 1



(b) Distorted Image 1



(c) Reference Image 2



(d) Distorted Image 2

Figure 4.4: Two reference and distorted images are shown. The distortions present are fast fading Rayleigh simulated packet loss distortion.

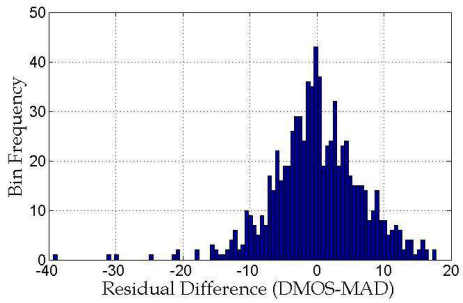


Figure 4.5: The residual histogram of metric MAD. Notice that the distribution is quite Gaussian except for some large outlier images which MAD believes are of higher quality than rated by observers.

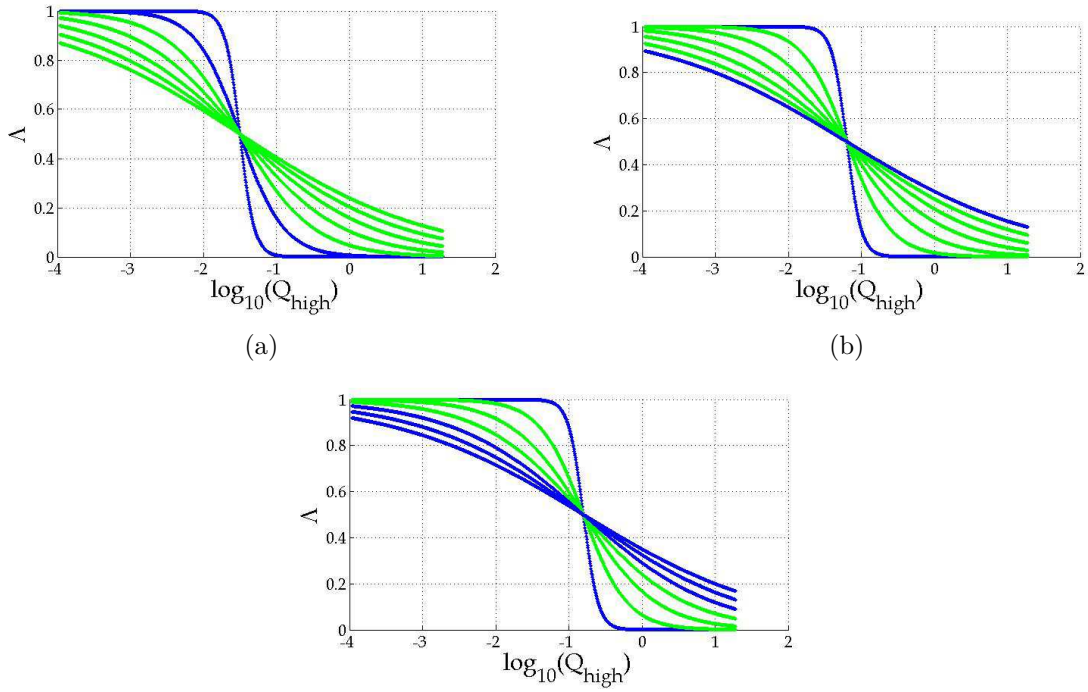


Figure 4.6: Three plot are shown with varying parameters for the α-blend of the two metrics. Each plot has a constant bias parameter, τ_1 from equation (3.15) and τ_2 is adjusted from 0.2 to 1.4 in steps of 0.2. τ_1 takes values -1.5, -1.2, and -0.8 in plots (a), (b), and (c), respectively. Red sigmoids denote that MAD is significantly inferior to MAD with 95% confidence. Blue sigmoids denote that MAD is statistically the same as VIF. Green sigmoids denote that MAD is significantly superior to VIF with 95% confidence.

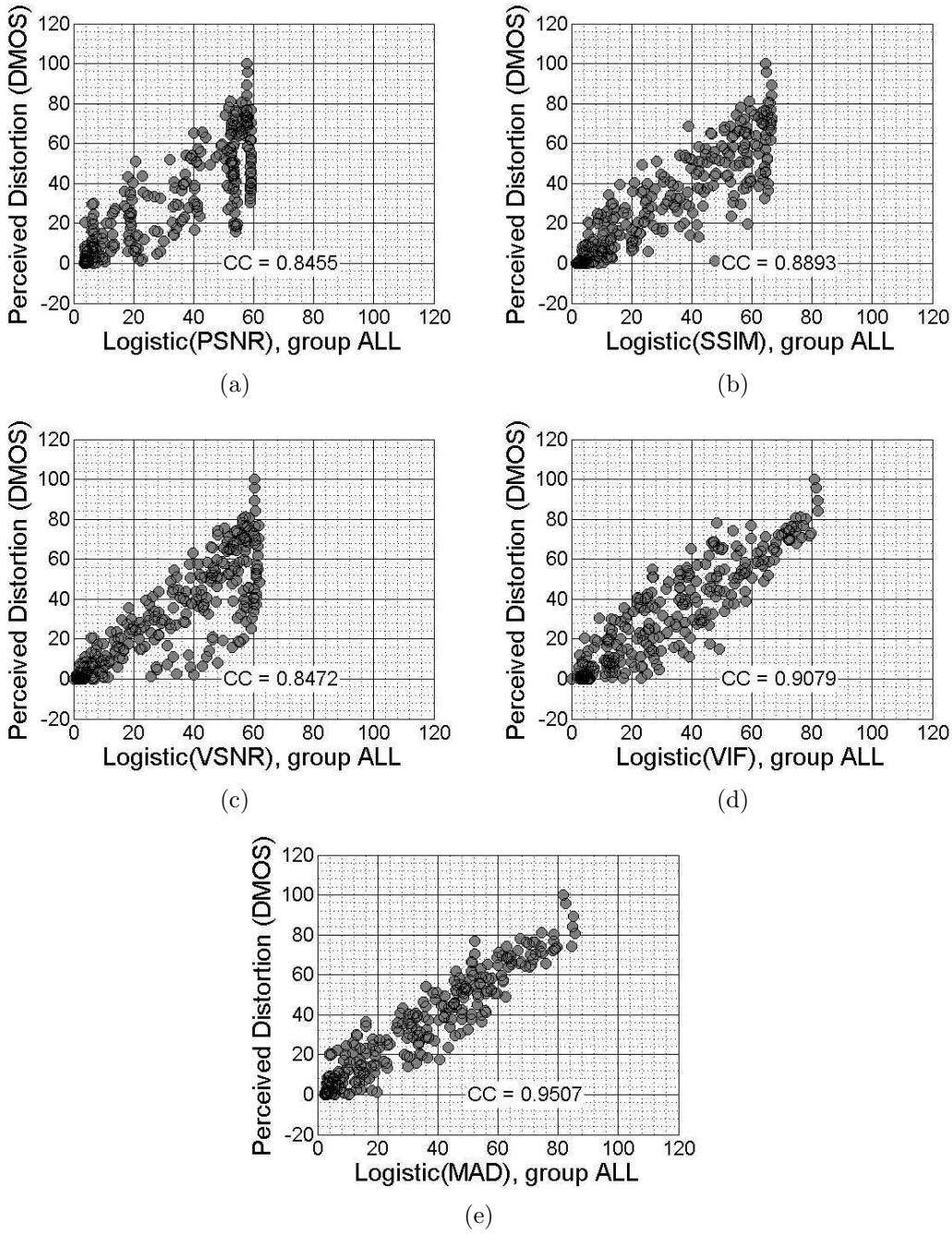


Figure 4.7: The DMOS ratings are plotted verse the logistic transformed prediction for each metric on the CSIQ database for group ALL.

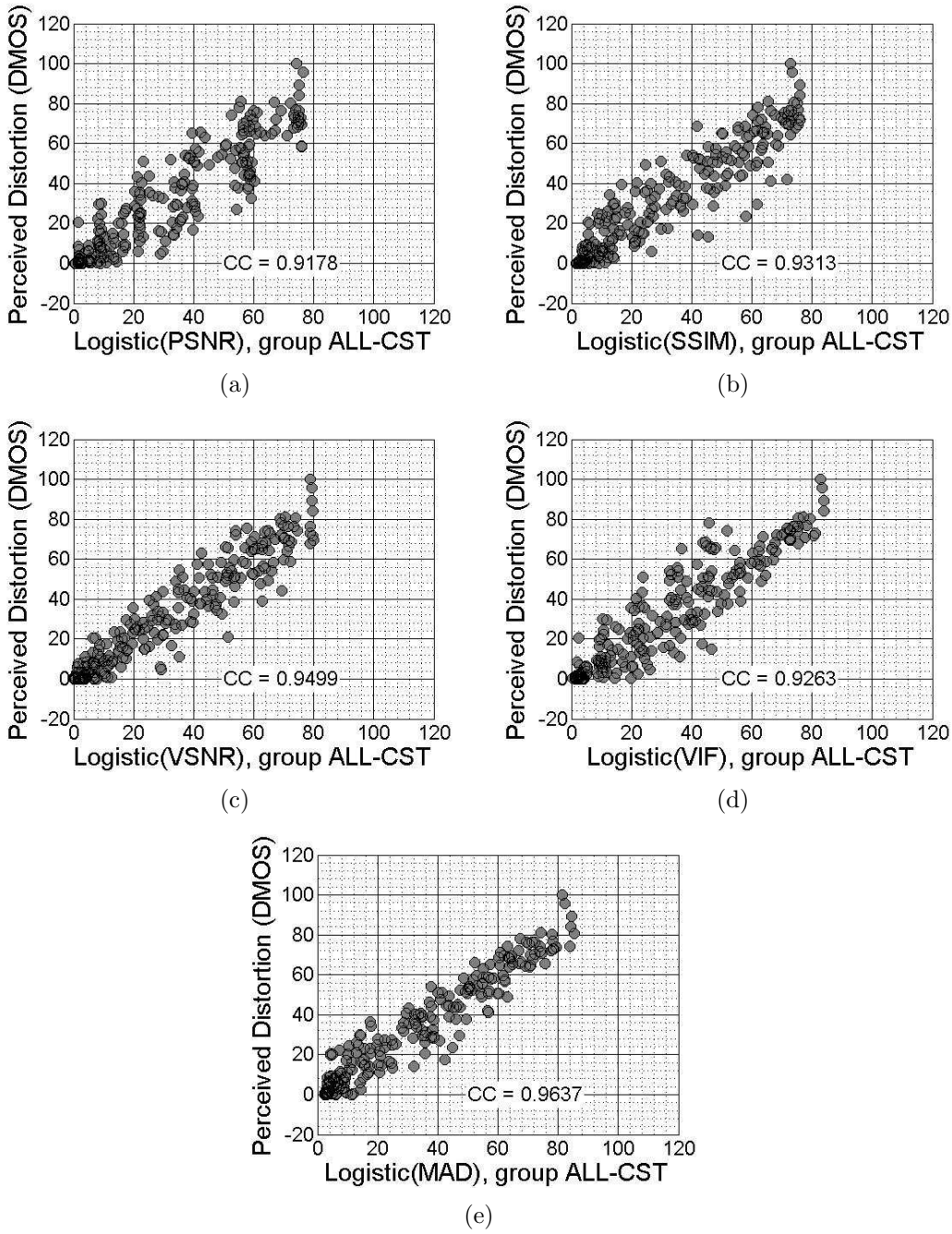


Figure 4.8: The DMOS ratings are plotted verse the logistic transformed prediction for each metric. Contrast distortions are removed from the database.

CHAPTER 5

CONCLUSIONS

We have presented a metric that uses properties of both low level vision and visual appearance to define a measure of fidelity. As far as we know, it achieves the highest level of performance thus far at predicting subjective image fidelity.

Because of the excellent performance of MAD, we foresee its usefulness in compression optimization, dynamic image bandwidth allocation, calibration for denoising, image resizing, or any algorithm were the idea is to preserve quality. MAD could also be used in assessing print quality and inside consumer electronics (camera phones, etc.).

In addition, because MAD is also modular for high and low quality, it can be used in a highly efficiently fashion if the relative quality range of images is known beforehand. Also, the low quality range of the metric may be useful in predicting human recognition tasks or restoring highly degraded images.

We are currently working to integrate a more sophisticated model of masking into MAD and decrease the complexity of the algorithm (MAD++) so that it can be more easily used inside an optimization loop. We are also working to expand MAD to video, high-dynamic range images, and photography.

BIBLIOGRAPHY

[1] H. Sheikh, M. Sabir, and A. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Transactions on Image Processing*, vol. 15, pp. 1349–1364, Nov. 2006.

[2] B. Girod, *What's wrong with mean squared error?*, pp. 207–240. MIT Press, 2nd ed., 1993.

[3] D. M. Chandler and S. S. Hemami, "Effects of natural images on the detectability of simple and compound wavelet subband quantization distortions," *Journal of the Optical Society of America*, vol. 20, pp. 1164–1180, 2003.

[4] R. L. DeValois and K. K. DeValois, *Spatial Vision*. Oxford University Press, 1990.

[5] N. Graham, *Visual Pattern Analyzers*. New York: Oxford University Press, 1989.

[6] E. Peli, L. E. Arend, G. M. Young, and R. B. Goldstein, "Contrast sensitivity to patch stimuli: Effects of spatial bandwidth and temporal presentation," *Spatial Vision*, vol. 7, pp. 1–14, 1993.

[7] M. G. Ramos and S. S. Hemami, "Suprathreshold wavelet coefficient quantization in complex stimuli: psychophysical evaluation and analysis," *J. of Opt. Soc. Am. A*, vol. 18, pp. 2385–2397, 2001.

[8] G. E. Legge and J. M. Foley, "Contrast masking in human vision," *J. of Opt. Soc. Am.*, vol. 70, pp. 1458–1470, 1980.

[9] D. J. Heeger and P. Teo, "A model of perceptual image fidelity," *IEEE International Conference on Image Processing*, vol. 2, pp. 343–345, Oct. 1995.

[10] S. Winkler, "Visual quality assesment using a contrast gain control model," in *Proc. IEEE Signal Processing Society Workshop on Multimedia Signal Processing*, pp. 527–532, Sept. 1999.

[11] S. Daly, "Visible differences predictor: an algorithm for the assessment of image fidelity," in *Proc. SPIE Vol. 1666, p. 2-15, Human Vision, Visual Processing, and Digital Display III, Bernice E. Rogowitz; Ed.* (B. E. Rogowitz, ed.), vol. 1666 of *Presented at the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference*, pp. 2–15, Aug. 1992.

[12] D. J. Field, "Relations between the statistics of natural images and the response properties of cortical cells," *J. Opt. Soc. Am. A*, vol. 4, pp. 2379–2394, 1987.

[13] A. Beghadi and B. Pesquet-Popescn, "A new image distortion measure based on wavelet decomposition," *Seventh International Symposium on Signal Processing and its Applications*, vol. 1, pp. 485–488, July 2003.

[14] Y. Lao and C. C. J. Kuo, "A haar wavelet approach to compressed image quality measurement," *Journal of Electronic Imaging*, vol. 10, no. 1, pp. 20–29, 2001.

[15] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, pp. 600–612, April 2004.

[16] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multi-scale structural similarity for image quality assessment," in *37th IEEE Asilomar Conference on Signals, Systems and Computers*, Nov. 2003.

[17] H. R. Sheikh, A. C. Bovik, and G. de Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Transactions on Image Processing*, vol. 14, pp. 2117–2128, Dec. 2005.

[18] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Transactions on Image Processing*, vol. 15, pp. 430–444, Feb. 2006.

[19] D. Chandler and S. Hemami, "Vsnr: A wavelet based visual signal to noise ratio for natural images," *IEEE Transactions on Image Processing*, vol. 16, pp. 2284–2298, Sept. 2007.

[20] W. Osberger, N. Bergmen, and A. Maeder, "An automatic image quality assessment technique incorporating higher level perceptual factors," *IEEE International Conference on Image Processing*, vol. 1, pp. 414–418, 1998.

[21] M. Pederson, "Importance of region of interest on image difference metrics," Master's thesis, Gjovik University, 2007.

[22] A. Ninassi, O. L. Meur, P. Callet, and D. Barba, "Does where you gaze on an image affect your perception of quality? applying visual attention to image quality," *IEEE International Conference on Image Processing*, 2007.

[23] J. Nachmias, "Masked detection of gratings: The standard model revisited," *Vis. Res.*, vol. 33, pp. 1359–1365, 1993.

[24] E. C. Larson and D. M. Chandler, "Unveiling relationships between regions of interest and image fidelity metrics," *Visual Communications and Image Processing*, 2007.

[25] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multi-scale structural similarity for image quality assessment," in *Proc. Asilomar Conf. on Signals, Systems, and Computers*, November 2003.

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

[27] A. K. Jain and F. Farrokhnia, "Unsupervised texture segmentation using gabor filters," *Pattern Recognition*, vol. 24, pp. 1167–1186, May 1991.

[28] E. C. Larson and D. M. Chandler, "Explaining crypsis and information content in the visual pathway using statistical properties of animal camouflage and natural scenes," *OSA Fall Vision Meeting*, 2007.

[29] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by v1?," *Vision Research*, vol. 37, pp. 3311–3325, Dec. 1997.

[30] F. A. A. Kingdom, A. Hayes, and D. J. Field, "Sensitivity to contrast histogram differences in synthetic wavelet-textures," *Vis. Res.*, vol. 41, pp. 585–598, 1995.

[31] P. Kosevi, *Invariant Measures of Image Features from Phase Information*. PhD thesis, University of Western Australia, 1996.

[32] N. P. S. D. I. A. [Online].

[33] Z. W. H. R. Sheikh, A. C. Bovik, and L. K. Cormack. Image and Video Quality Assessment Research at LIVE [Online]. Available: http://live.ece.utexas.edu/research/quality/.

[34] "Vqeg, final report from the video quality experts group on the validation of objective models of video quality assessment, phase ii," August 2003 [Online]. Available: http://www.vqeg.org.

[35] J. A. Nelder and R. Mead, "A simplex method for function minimization," *J. Comput.*, vol. 7, pp. 308–313, 1965.

[36] A. K. Bera and C. M. Jarque, "Efficient tests for normality, homoscedasticity and serial independence of regression residuals," *Econ.Lett.*, vol. 6, p. 25525, 1980.

VITA

Eric C. Larson

Candidate for the Degree of

Master of Science

Thesis: THE STRATEGY OF IMAGE QUALITY ASSESSMENT
A New Fidelity Metric Based upon Distortion Contrast Decoupling

Major Field: Electrical Engineering

Biographical:

Personal Data: Born in Spokane, Washington, U.S.A. on August 6, 1982.

Education:

Received the B.S. degree from Oklahoma State University, Stillwater, OK, U.S.A., 2006, in Electrical Engineering
Completed the requirements for the degree of Master of Science with a major in Electrical Engineering from Oklahoma State University in July, 2008.

Experience:

Worked as teaching assistant for two and one half years. Three years of research experience during undergraduate and graduate semesters. Worked as design Engineering Intern for Garmin International in Olathe, Kansas.

Name: Eric Larson Date of Degree: July 2008

Institution: Oklahoma State University Location: Stillwater, Oklahoma

Title of Study: THE STRATEGY OF IMAGE QUALITY ASSESSMENT
A New Fidelity Metric Based upon Distortion Contrast Decoupling

Pages in Study: 70 Candidate for the Degree of Master of Science

Major Field: Electrical Engineering

This thesis presents a new image fidelity metric, Most Apparent Distortion (MAD), that uses a visual masking model and a measure of appearance distortion strategically to define the fidelity of a distorted image. Subjective image fidelity has been shown to be largely influenced by visual contrast masking of distortions and distortion energy. However, recent image fidelity metrics without an explicit visual masking model have been shown to correlate highly with subjective ratings. We argue that, at high quality, viewers use a different strategy for the task of rating distorted images than at low quality. We then evaluate the performance of MAD on two fidelity databases. In particular, we compare the performance of MAD to Peak Signal to Noise Ratio, Visual Signal to Noise Ratio, Structural Similarity, and Visual Information Fidelity. The results show that MAD performs statistically better than all other fidelity algorithms using various evaluation criteria for both databases.

ADVISOR'S APPROVAL: _____