Evaluating and Improving Image Quality of Tiled Displays

by

Steven McFadden

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

Waterloo, Ontario, Canada, 2015

Author's Declaration

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

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

Abstract

Tiled displays are created by grouping multiple displays together to form one very large display. These tiled displays are often the only suitable option for displaying very large images but suffer from a grid distortion caused by gaps between each sub-display's active region. This grid distortion is fundamentally different from other, well-studied, image distortions (e.g., blur, noise, compression) and the impact of these grid distortions has thus far not been studied. This research addresses this lack of attention by investigating the grid distortion's quality impact and creating perceptual algorithms to reduce this impact.

We measure the quality impact of the grid distortion by creating two new image quality assessment (IQA) databases for tiled images. These databases provide significant insight into the unique characteristics of the grid distortion and provide a baseline against which to measure the performance of current IQA metrics. We use these databases to show that current metrics do not adequately reflect the quality impact of the grid distortions, and we create a new metric specifically for tiled images that statistically (with 95% confidence) outperforms current metrics.

We improve perceived tiled display image quality by creating new image-correction algorithms based on elements of the human visual system (HVS). These correction techniques modify the perceived quality of the displayed images without directly modifying the static grid distortion. These algorithms are shown, through the use of a third subjective user study, to clearly and consistently improve the perceived quality of tiled images.

Acknowledgements

I wish to thank Christie Digital Systems and the Natural Sciences and Engineering Research Council of Canada for their generous funding support, without which this dissertation would not have been possible. I am also grateful to my supervisor Paul Ward for his support, guidance, and encouragement throughout this endeavour. Many thanks also to my thesis committee members – Christopher Nielsen, Thrasyvoulos Pappas, Justin Wan, and Zhou Wang – for their comments and advice. I wish to also express my ongoing appreciation to Maher Sid-Ahmed, who got me hooked on research for the first time many years ago. Last but certainly not least, I am extremely grateful to my family for their ongoing patience and support.

Dedication

This dissertation is dedicated to my family:

To my wonderful son, I hope you grow up with my love of lifelong learning (but please don't drag out the formal education component like I did).

In loving memory of my mother, who inspired me with her endless perseverance.

To my father, who taught me the value of hard work.

To my brother, who has been everything a big brother should be.

To my aunts, uncles, and cousins, who have always been supportive and encouraging.

To my nieces and nephews, I am very proud and consider myself fortunate to be your uncle.

Table of Contents

List of Tables

List of Figures

Acronyms

ACR	Absolute Category Rating
CSF	Contrast Sensitivity Function
DCT	Discrete Cosine Transform
DLP	Digital Light Projection
DMOS	Differential Mean Opinion Score
DSCQS	Double Stimulus Continuous Quality Scale
DSIS	Double Stimulus Impairment Scale
HVS	Human Visual System
IW-SSIM	Information content Weighted Structural SIMilarity
IQA	Image Quality Assessment
IPS	In-Plane Switching
ITU	International Telecommunication Union
JND	Just Noticeable Difference
LCD	Liquid Crystal Display
KRCC	Kendall's Rank order Correlation Coefficient
LIVE	Laboratory for Image and Video Engineering
MAD	Most Apparent Distortion
MAE	Mean Absolute Error
MOS	Mean Opinion Score
MS-SSIM	Multi-Scale Structural SIMilarity
OR	Outlier Ratio

PLCC	Pearson's Linear Correlation Coefficient
PSF	Point Spread Function
PSNR	Peak Signal-to-Noise Ratio
QA	Quality Assessment
RGB	Red Green Blue
RMSE	Root Mean Squared Error
sRGB	standardized Red Green Blue (colour space)
SSCQE	Single Stimulus Continuous Quality Evaluation
SNR	Signal-to-Noise Ratio
SRCC	Spearman's Rank order Correlation Coefficient
SSIM	Structural SIMilarity
TDQM	Tiled Display Quality Metric
TID	Tampere Image Database
VDP	Visible Difference Predictor
VIF	Visual Information Fidelity
VQEG	Video Quality Experts Group
VSNR	Visual Signal-to-Noise Ratio

Chapter 1

Introduction

Tiled displays allow for visualization of images that cannot be practically viewed on individual displays. They support sizes that are orders of magnitude greater than the largest individual display, with equivalent or superior pixel densities, and they offer this support with the option of different shapes and configurations that are infeasible using individual displays.

Large tiled displays are commonly used for multiple purposes including analytics [45, 34], command and control [28, 18], and information display [57]. These displays aid in visualization required to gain important insights into large and/or complex data sets [25].

For very large displays, tiled displays are more economical than individual displays. As the size, and pixel count, of an individual display increases, the cost rises quickly due to decreased yield. Using tiled displays mitigates the yield issue because a handful of defective pixels no longer wastes an entire high-definition panel; it instead leads to the discard of a smaller lower-resolution (and lower cost) panel[1]. This cost benefit extends to maintenance of the large display. If a large individual display fails, the entire unit must be replaced. For a tiled array, only the defective sub-unit requires replacement.

In addition to cost, tiled displays can be constructed orders of magnitude larger than what is possible for individual displays [57] while maintaining pixel densities equivalent to those of individual displays. It is important to note how this is different from creating very large images using a single projection display. A single projector can scale an image to large physical dimensions but it does so by stretching the image and sacrificing the pixel

[1]We refer to LCD "panels", but this concept also applies to other display technologies (e.g., optical projection DLP "chips" used in projectors).

density. Tiled displays can achieve these large physical dimensions while maintaining pixel density.

Tiled displays also support custom aspect ratios and even novel screen shapes [58]. Individual displays are mass produced in standard aspect ratios (e.g., 16:9, 4:3, etc.) but an array of individual displays can be shaped with a great deal of flexibility, as shown in Figure 1.1.

1.1 Motivation

These advantages come at the cost of certain distortions that are unique to tiled displays such as non-uniformity, brightness and/or colour mismatch between tiles, and misaligned tiles [5, 16]. These distortions can generally be managed through careful design and manufacturing decisions. This dissertation focuses on another distortion inherent to tiled displays, caused by the gaps between each active region, that creates the appearance of a grid overtop of any image displayed. This grid distortion is not correctable with current manufacturing techniques, making it an objectionable artifact on every tiled display.

Since this grid distortion is currently uncorrectable, the ability to measure, and potentially affect, its quality impact is of great significance. Accurate quality measurements allow for better design and manufacturing decisions and creates opportunity for quality improvements through real-time image processing. General-purpose image quality assessment (IQA) metrics have existed for some time and work well for many common image distortions (e.g., compression, noise, and blur), but the grid distortion of tiled displays had never been studied. Image quality databases used to develop and test objective IQA metrics have never included images with any kind of grid distortion, making it unknown whether current metrics would be effective on this unique distortion.[2]

In addition to *measuring the quality* of tiled displays, there are potential means of *improving the quality* by minimizing the visual impact of the grid distortion. A "typical" image enhancement problem involves determining the best pixel values to display in a given physical location. Enhancement of grid-distorted images presented a unique challenge because the "best" pixel values are already known for a given area (i.e., the grid) but there are no physical pixels in that area to display those values.

[2]The grid distortion was considered in [13] but only in a narrow sense (pertaining to vernier acuity).

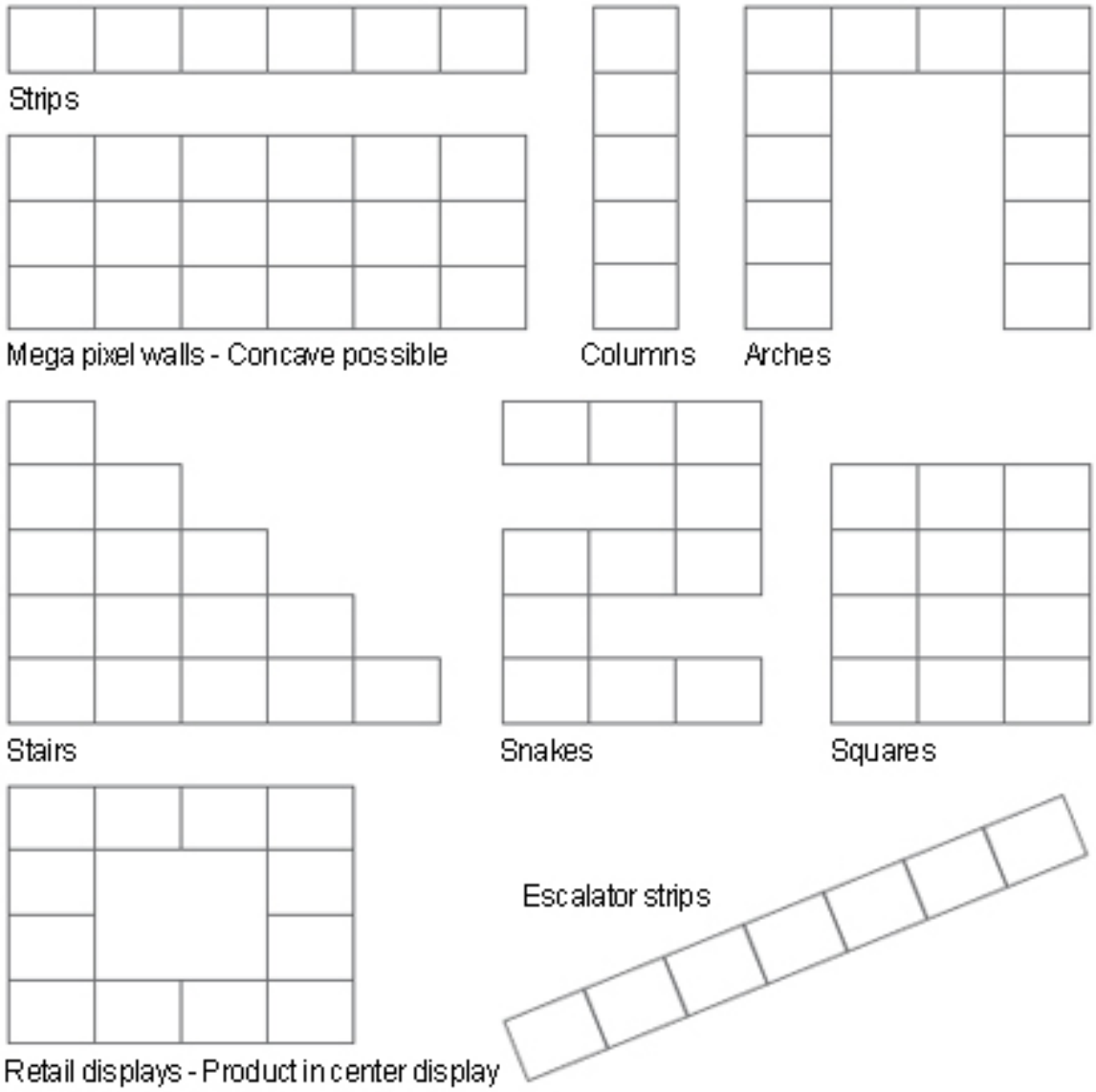


Figure 1.1: Some potential shapes of tiled displays (©Christie Digital).

1.2 Contributions

This dissertation provides the following significant contributions to the field of image processing:

1. Two IQA databases created through formal subjective user studies. These databases contain quality scores for 248 grid-distorted images evaluated by a total of 71 subjects.
2. Evidence that current objective IQA metrics perform poorly when applied to tiled images. The best traditional metric only accounted for roughly 36% of the variance in quality scores.
3. A new objective IQA metric that significantly outperforms (with a 95% confidence interval) current metrics when measuring tiled image quality. Our new metric accounts for 62% more variance than the leading traditional metric (60% vs. 36%).
4. Four new image-correction algorithms that improve perceptual quality of tiled images and mitigate the visual effect of the grid distortion. Our top-performing algorithm was preferred over the unmodified image more than 90% of the time.

To the best of our knowledge, ours is the first research performed on the grid distortion of a tiled display and its impact on image quality. As a result, there previously existed no IQA databases containing subjective quality scores for grid-distorted images. IQA databases contain the "ground truth" data, in the form of average subjective quality scores, necessary for understanding and objectively measuring image quality. The creation of two such databases for tiled displays was our first contribution.

With the new tiled IQA databases available, it was then possible to evaluate a selection of objective IQA metrics to determine their performance (i.e., how well they matched the subjective results stored in the databases). This evaluation showed clear evidence that current objective metrics perform poorly when applied to grid-distorted images.

With current IQA metrics performing poorly for tiled images, we used the new image databases to develop and test a new quality metric: the Tiled Display Quality Metric (TDQM). This metric proved to be be statistically better (with $p < 0.05$) at correlating with subjective quality scores than any other metrics tested.

We also developed four new image-correction algorithms designed to perceptually improve image quality by minimizing the effects of the grid distortion. A subjective user study showed statistically significant (with $p < 0.05$) improvements in quality between the

corrected and reference images. In addition to verifying our correction algorithms, this image-correction study contributed to the knowledge of correcting tiled images, opening avenues for further improvements.

1.3 Organization

This dissertation is organized into the following two parts:

1.3.1 Fundamentals

We begin by reviewing relevant background and fundamentals used throughout this dissertation. This background includes an overview of tiled display technologies and their inherent distortion types (Chapter 2), a review of current techniques for evaluating image quality (Chapter 3), and a review of methods for testing the performance of objective IQA metrics (Chapter 4).

1.3.2 Solutions and Contributions

The chapters in this part detail the contributions listed in Section 1.2: the subjective user studies used to develop the new IQA databases (Chapters 5 and 6), evaluation and analysis of current metrics (Chapter 6), development of our new TDQM metric (Chapter 7), development of our image-correction algorithms (Chapter 8), and the design of the subjective study to test these algorithms (Chapter 9).

Part I

Fundamentals

Chapter 2

Tiled Displays

Tiled displays are commonly used for the display and visualization of large images. By extending an image across multiple sub-displays, or "tiles", display walls can be created that are orders of magnitude larger than what is possible using a single display, while still maintaining the same pixel density. In addition to their size flexibility, these displays also have superior shape flexibility; non-standard aspect ratios and even non-rectangular shapes can be obtained with relative ease (refer to Figure 1.1 for some examples).

This flexibility is not without costs as tiled displays are subject to unique distortions that are rarely (or never) an issue with individual displays. We introduce different tiled display technologies in Section 2.1 and discuss their inherent distortions in Section 2.2.

2.1 Types of Tiled Displays

There are four common types of tiled displays [36]: front-projection, rear-projection with single screen, rear-projection cubes, and tiled LCD panels.

2.1.1 Front Projection Tiled Displays

Front projection tiled displays use an array of projectors displaying to a single (reflective) screen. The projectors are mounted in a grid array and aligned to allow for some overlap between displayed images. This overlap is used for edge-blending as determined through image processing methods.

The seamless images created through edge blending are the primary advantage of front projection tiled displays. Their disadvantages include high manufacturing costs (i.e., no economy of scale), high maintenance costs (i.e., maintenance of strict alignment), special environment requirements (i.e., reduced location lighting and space requirements), and the potential for obstructed viewing (i.e., when viewers or objects come between the projector light source and the screen). In addition to these disadvantages, front projection tiled displays are generally not portable and require a fixed installation. This is due to the rigid mounting structure that is generally required to ensure projector alignment.

2.1.2 Single-Screen Rear Tiled Projection Displays

Single-screen rear projection tiled displays use an array of projectors mounted behind a rear projection (transmissive) screen. These projectors are tiled in a grid array and aligned to allow overlap between displayed images, similar to front projection tiled arrays. As with front-projection arrays, image processing is used to blend the edges of each individual image. This allows single-screen rear projection displays to share the primary advantage of front projection tiled arrays: a seamless image. Single-screen rear projection arrays also have the advantage of avoiding image occlusion because the projectors are behind the screen.

Aside from the lack of image occlusion, these tiled displays share the main disadvantages of front projection tiled displays: high maintenance costs, lack of portability, and special environment requirements (though these are more flexible since the environment behind the screen need not be the same as in front where the viewers are positioned). These displays can not be made in narrow profile form factors because the Fresnel lenses required for shorter throw lengths would interfere with edge-blending capabilities. These displays require large seamless sheets of rear projection screen material and are not reconfigurable after the initial aspect ratio and screen shapes are selected.

2.1.3 Rear Projection Tiled Cube Displays

Rear-projection tiled cube displays consist of an array of individual, self-contained rear-projection display units, each consisting of a frame, a projection unit, and a screen. These displays are stacked edge to edge in a manner where the distance between each unit is minimized.

Rear projection tiled cube arrays do not require the same special environments needed by front or rear single screen projection tiled displays. Each display is a self-contained unit

and is therefore tolerant to different ambient lighting conditions. This self-containment also ensures there are no issues with image occlusion. These arrays also require much less space because a narrow display depth is attainable by including a Fresnel as part of each screen; this redirects the light from the projection unit and allows for a shorter throw distance. The modular nature of these arrays provides for simple maintenance because the cubes can be designed to allow for front access, and any damaged screens can be easily replaced without replacing the entire screen or accessing the rear of the display [58].

The primary disadvantage of rear projection display cubes is the gap present between the individual screens of each display unit. These gaps create a grid-like seam and cannot be removed because they are required to allow for changes in temperature and humidity. Through careful design and selection of screen materials, these seams can be (at the time of this writing) as small as 0.2 mm [10].

2.1.4 LCD Tiled Displays

LCD tiled displays are created by tiling multiple LCD panels together edge-to-edge, usually mounted to an external rear frame or structure. These displays are the cheapest to build [25] because they use mass produced commodity LCD panels and require minimal maintenance (e.g., lack of alignment issues, colour shift, etc.). They are also the thinnest displays available, with most of their thickness taken up by the support structure and electronics.

The primary disadvantage of LCD tiled arrays is the introduction of image seams as a result of the individual display bezels. These bezels provide structural integrity to each panel and cannot be entirely removed. Custom thin-bezel panels are available with bezels as small as 2 mm. Another disadvantage, a result of the thinness, is the requirement for rear access maintenance. There is no capacity for front access replacement of components.

2.1.5 Common Use

Most tiled displays in use today are based on LCD or rear projection cube technology [35]. Single-screen projection technologies are used primarily in custom environments such as simulators. This dissertation focuses on LCD and rear projection cube display technologies.

2.2 Distortions of Tiled Displays

Tiled displays are subject to unique distortions that rarely, or never, impact the visual quality of single displays in isolation. Examples of these distortion are shown in Figure 2.1.

2.2.1 Colour Mismatch Between Tiles

Colour mismatch distortion can be found in all types of tiled displays. Every display has a particular colour gamut; a range of colours it is capable of displaying. A deficient colour gamut that may be imperceptible on a single display becomes very noticeable when multiple displays are tiled together. As a result, the colour gamuts must be matched between individual displays to ensure consistency across the array (mismatches manifest as "hotspots" or "darkspots" in the array). This distortion can be managed through real time monitoring and adjustment. Gamut matching at the time of manufacture is often not sufficient because the colour range can shift as the age and/or temperature of the light source changes. The colour gamut of the entire array is dictated by that of the individual display with the smallest range of colour support.

2.2.2 Brightness Mismatch Between Tiles

Brightness mismatch distortion is very similar to colour mismatch distortion and is applicable to all tiled displays. When viewing an individual display, brightness can vary considerably from its default setting with no objectionable effect. When displays are tiled together, even small differences in brightness between tiles become very obvious and objectionable (mismatches manifest as "hotspots" or "darkspots" in the array). As with colour management, the brightness of each tile can be managed through real time monitoring and adjustment (brightness can shift with age and/or temperature of the light source). The peak brightness of the array is determined by the darkest individual display.

2.2.3 Misaligned Tiles

Tile misalignments are applicable primarily to projection displays. Misalignments in the projection units, often caused by vibration over time, can cause an image to be slightly misplaced on the screen. Slight alignment issues that may be acceptable in single displays become very noticeable when displays are tiled together (e.g., consider a single-pixel line

displayed across multiple individual displays). This distortion can be corrected through mechanical means (e.g., using a rigid structure to ensure no variance in the alignment), optical means (e.g., many higher-end projectors support some level of fine lens adjustment), electronic means (e.g., transforming/shifting the image through image processing), or some combination of the three methods.

2.2.4 Non-Uniform Brightness Within Individual Tiles

Non-uniform brightness within tiles is applicable mostly to projection displays. Minor non-uniformity distortions are not typically noticeable on individual display tiles but create an objectionable "dimpling" effect when part of an array. This distortion is a result of the geometry involved with projecting an image from a (roughly) point source onto a two-dimensional screen; the lens is not equidistant to all parts of the screen. For a display where the lens is centred, care must be taken to ensure the centre of the image is not brighter than the edges (both because the centre is closer to light source than the screen edges, and because the light is striking the edges at a different angle). This distortion can be corrected through use of a Fresnel as part of the screen to focus/direct the light and through image processing means (the peak brightness of the individual display is limited to the darkest portion of the screen).

2.2.5 Non-Uniform Scaling Within Individual Tiles

Non-uniform scaling within an individual tile is similar to the misaligned tiled distortion and is applicable to projection displays. If a projector is not properly aligned, the image will not display properly on the screen. An example of this is the keystone effect where a square image takes the shape of a trapezoid. These distortions are often very minor on individual displays but are much more noticeable when displays are tiled together and differences become apparent, similar to the distortions caused by misaligned tiles. This distortion can be corrected through optical or electronic means, as described for the misaligned tile distortion.

2.2.6 Grid Distortion

Grid distortion (a.k.a., display seam distortion) is found in all rear projection cube arrays and LCD arrays. Unlike the other distortions described in this section, the grid distortion can not be completely eliminated with current manufacturing methods. For LCD arrays, a

bezel is required for each individual display to provide structural support. Thin-bezel LCD panels are becoming more common at lower prices but the smallest bezel available in LCD tiled arrays, at the time of this writing, is 2 mm. Rear projection cube arrays require a small expansion gap between each individual screen to allow for changes in temperature and humidity. These expansion gaps can be minimized through careful alignment and selection of screen materials but the smallest gap in rear projection cube arrays, at the time of this writing, is 0.2 mm (nominal). These grid distortions are the primary disadvantage of rear projection cube and LCD technologies compared with their single-screen projection alternatives. This dissertation focuses primarily on the grid distortion because it is always present on any tiled display where array depth is a constraint.

Current Techniques

Industry players currently use optical and mechanical means to minimize the gaps in tiled displays. LCD manufacturers continue to shrink the bezel width while screen manufacturers use better materials to minimize the expansion of projection screens. Nobody has approached the grid distortion problem from an image processing point of view. Minimum gaps were roughly 5 mm as recently as 2007 and there was little that could be done to perceptually reduce their appearance. With gaps now less than 2 mm, there is potential to use elements of the human visual system (HVS) to reduce the quality impact of the gap through modification of the active pixels in an image. This potential to improve the appearance of images with narrow gaps is a recent occurrence that has not yet been investigated in the literature.

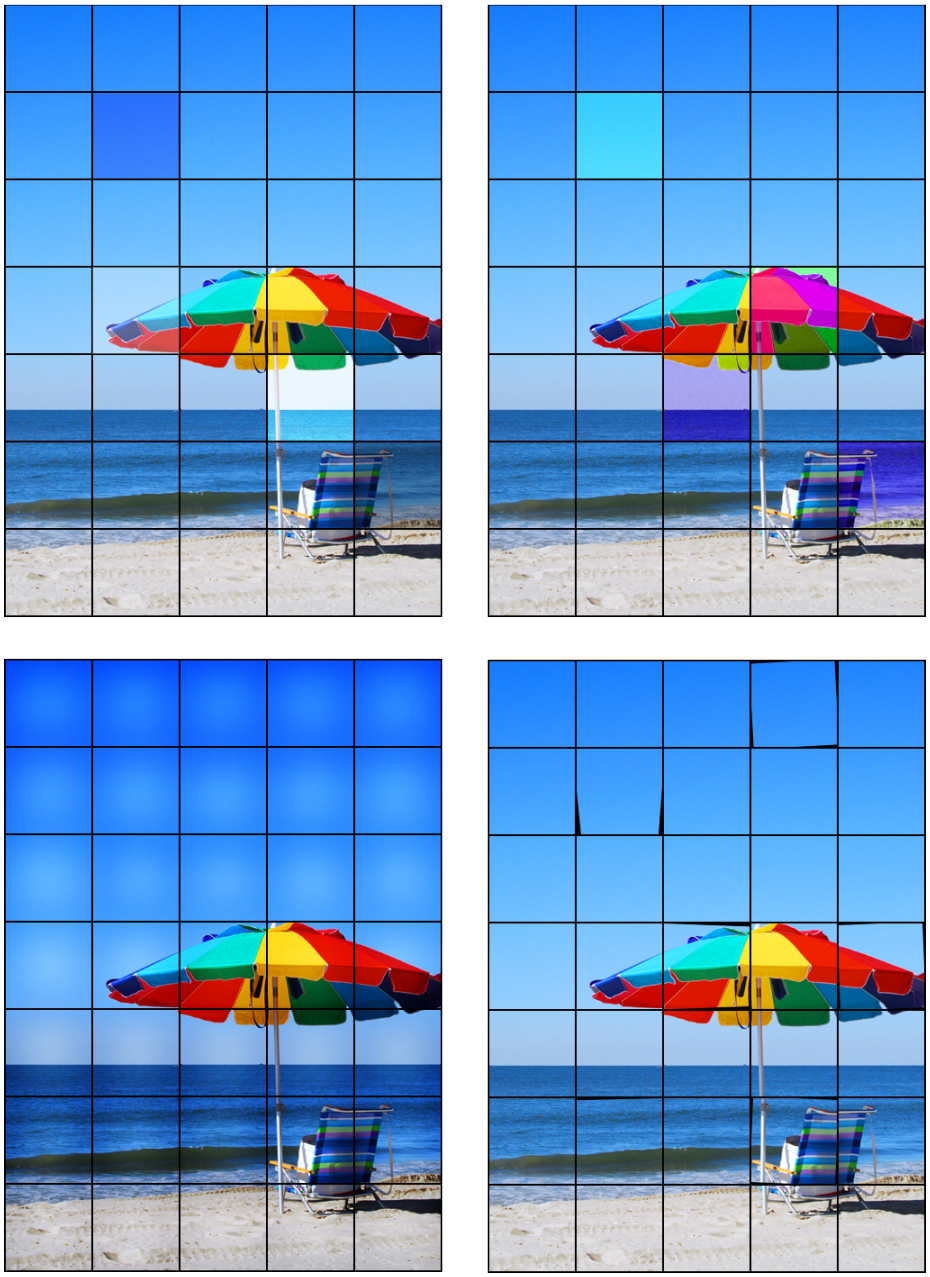


Figure 2.1: Examples of different tiled display distortions. Top left: brightness mismatch. Top right: colour mismatch. Bottom left: brightness nonuniformity. Bottom right: tile misalignment and/or non-uniform scaling. All: Grid distortion.

Chapter 3

Image Quality Assessment (IQA)

Image Quality Assessment (IQA) is an enormous area of research and this dissertation only touches on a relatively small portion. This chapter gives a brief overview of the subjective IQA methods used to obtain "ground truth" image quality data and the objective methods that attempt to achieve high correlation with this data.

3.1 Subjective Image Quality Evaluation

Subjective image quality evaluation is at the heart of any IQA metric. The worth of any objective quality metric for a group of images is determined by its correlation to the corresponding mean opinion scores (MOS) or differential mean opinion scores (DMOS). These scores are obtained through subjective image quality testing.

3.1.1 Methods of Subjective Quality Evaluation

There are many methods of evaluating subjective image quality and some of the most common standardized methods are listed below. These methods have multiple variations and only the main differentiators are described here. Note these methods were developed for video quality evaluation and modifications are made when applying them to image quality evaluation.

Double Stimulus Impairment Scale (DSIS) [4] In this method, each viewer is shown a series of image sequence pairs (reference sequence followed the impaired sequence). The

viewer, after viewing each pair, provides a rating for the difference between the two sequences in terms of impairment.

Double Stimulus Continuous Quality Scale (DSCQS) [4] Viewers are shown a series of reference and impaired image sequence pairs in random order. They provide an absolute quality rating for each sequence after viewing each pair (independent of other image pairs).

Single Stimulus Continuous Quality Evaluation (SSCQE) [4] In this method, viewers are shown a single continuous image sequence and provide an absolute quality rating using a slider in real time.

Absolute Category Rating (ACR) [17] Viewers are shown a number of individual image sequences and provide a rating for each on a discrete scale after viewing. When the reference sequence is included for viewing (without any indication of such), this is known as the 'hidden reference' variation.

3.1.2 Publicly Available IQA Databases

Results from large subjective quality studies are often made available in the form of IQA databases for use by other image quality researchers. These databases typically consist of a large number of distorted images along with their corresponding 'perfect quality' reference images. Each database contains a subjective quality score (MOS or DMOS) for each distorted image, obtained through subjective testing (often, but not always, a variation of one of the methods listed in Section 3.1.1). Six of the most commonly used (and publicly available) IQA databases are listed below along with the distortion types they contain:

- The LIVE IQA Database [43, 44, 53] was developed at the University of Texas at Austin and consists of 779 images distorted by the following means:
  - JPEG compression
  - JPEG2000 compression
  - Gaussian blur
  - White noise
  - Bit errors in a JPEG2000 transmission
- The A57 Database [8] was developed at Cornell University and consists of 54 images distorted by the following means:
  - LH subband quantization

  - Gaussian noise
  - JPEG compression
  - JPEG2000 compression
  - JPEG2000 compression with the dynamic contrast-based quantization (DCQ) algorithm
  - Gaussian blur

- The Toyama Database [41] was developed at the University of Toyama and consists of 168 images distorted by the following means:
  - JPEG compression
  - JPEG2000 compression

- The IVC Database [24] was developed at L'Université de Nantes and contains 185 images distorted by the following means:
  - JPEG compression
  - JPEG2000 compression
  - LAR coding
  - Blurring

- The CSIQ Database [22] was developed at Oklahoma State University and consists of 866 images distorted by the following means:
  - JPEG compression
  - JPEG2000 compression
  - Global contrast decrements
  - Pink Gaussian noise
  - Gaussian blurring

- The TID2008 Database [38] was jointly developed in Finland, Italy, and Ukraine and consists of 1700 images distorted by the following means:
  - Additive Gaussian noise
  - Additive noise in colour components more intensive than noise in luminance components

- Spatially correlated noise
- Masked noise
- High frequency noise
- Impulse noise
- Quantization noise
- Gaussian blur
- Image denoising
- JPEG compression
- JPEG2000 compression
- JPEG transmission errors
- JPEG2000 transmission errors
- Non-eccentricity pattern noise
- Local blockwise distortions
- Mean (intensity) shift
- Contrast change

3.2 Objective Image Quality Evaluation

While subjective image quality evaluation is the most reliable measure of image quality, it is expensive and time consuming. For repeatable results in real time, objective image quality metrics must be used. Objective metrics additionally allow for dynamic quality adjustment and image optimization, potentially in real time. Full reference metrics are the simplest class of image quality metrics and can be used whenever a reference source is available. Full-reference image quality algorithms are commonly divided into two categories: "bottom-up" algorithms using direct modelling of the human visual system (HVS), and "top-down" algorithms that treat the HVS as a "black-box".

3.2.1 Bottom-up Algorithms Using Direct HVS Modeling

In this section, we briefly discuss some fundamental characteristics of the human visual system ("HVS Fundamentals") and list some of the "top-down" algorithms that use these characteristics directly ("HVS Models").

HVS Fundamentals [56, 37]

Preprocessing Most QA algorithms have some type of preprocessing stage which commonly includes image calibration and registration. Calibration takes into account factors such as viewing distance and pixel spacing to map an image to cycles per degree of visual angle. Registration aligns the two images to ensure pixels and local regions are compared against their correct counterparts in the other image. Other preprocessing may include colour space transformations and low-pass filtering to simulate the point spread function (PSF) of the human eye.

Frequency analysis The HVS is sensitive to various 'bands' of frequency and orientation. Therefore, a decomposition is often performed to separate an image into different bands for analysis. Various decomposition methods include Fourier, wavelet, Discrete Cosine Transfer (DCT), and Gabor.

Contrast sensitivity function (CSF) The CSF models the sensitivity of the HVS as a function of spatial frequency. In general, the CSF has a band-pass nature for luminance [12] and a low-pass nature for chrominance [33].

Light adaptation Also known as 'luminance masking', light adaptation models the just noticeable luminance difference over the background as a function of the background luminance itself. This relationship is described by Weber's law which states that the ratio of the just noticeable difference to the background is a constant.

Contrast masking Sometimes referred to as 'texture masking', contrast masking refers to the reduction of visibility of one image component (or signal) caused by the presence of another image component (the 'masker'). The masking effect is generally strongest when the two image components possess similar spatial, frequency, and orientation properties. The effect also depends on the intensity of the mask component.

Foveated vision Of particular interest in large displays, foveated vision refers to the higher sampling resolution associated with a viewer's fixation point. Due to the distribution of cone receptors in the retina, this resolution drops off sharply as the distance from this point increases. The high resolution near the fixation point is referred to as 'foveal vision', while the lower resolution away from the fixation point is referred to as 'peripheral vision'. Conversely, temporal resolution is higher in the 'peripheral vision' than in the 'foveal vision'.

Error pooling The final step in the QA algorithm, error pooling combines the results from each of the preceding stages. This pooling can result in either a quality/error

map (with values for each pixel or group of pixels) or a single value for the entire image. Pooling to a single value is used when measuring or comparing performance of IQA metrics.

HVS Models There are a number of existing QA metrics based on the HVS. The well known models include

- Daly Model (also known as the Visible Difference Predictor, or VDP, model) [12]
- Lubin Model (also known as the Sarnoff JND model) [26]
- Safranek-Johnson Model [40]
- Teo-Heeger Model [49]
- Visual Signal to Noise Ratio (VSNR) [9]
- PSNR-HVS and PSNR-HVS-M [15, 39]
- Most Apparent Distortion (MAD) [23]

Discussion The older methods (Daly, Lubin, Safranek-Johnson, and Teo-Heeger) rely strongly on models that are based on overly simplistic images that are not as accurate as natural images. They are also based on the just noticeable difference (JND) instead of general image quality. As a result, these models often break down in the supra-threshold region of visibility.

The newer models (VSNR, PSNR-HVS, PSNR-HVS-M, and MAD) have largely overcome these supra-threshold limitations.

3.2.2 Top-Down Metrics

The following approaches are based on mathematical measures developed by treating the HVS as a "black box" system under test.

Peak Signal-to-Noise Ratio (PSNR) The PSNR is a simple metric based on measuring the energy of the distortion. It uses point-wise differences between pixel values in the reference and distorted images, and has been shown to correlate poorly with ground-truth (subjective study) results. [51, 56] In spite of this, it is still in common use due to it's simplicity.

Structural Similarity Index (SSIM) [53] The SSIM is based on the fundamental assumption that the HVS is highly adapted to extract structural information from a visual scene. It consists of three independent component measures – luminance comparison, contrast comparison, and structural comparison – of which the the structural comparison is the most significant. [7] By designating one image as a perfect reference, the quality of the second image can be measured by computing the similarity between the two images.

Multi-Scale SSIM (MS-SSIM) [55] The MS-SSIM is an extension of the SSIM where the luminance difference is calculated as in SSIM but the contrast and structural difference terms are calculated through successive downscaling steps of the reference and test images. Each scale is weighted based on empirical testing against IQA databases before all components are combined, providing a quality measure incorporating variations of viewing distance.

Visual Information Fidelity Index (VIF) [42] The VIF is an information theoretic approach that treats QA as an information fidelity problem (as opposed to a signal fidelity problem). It makes heavy use of statistical characteristics of "natural images". It assumes the test image and original reference image both pass through an HVS "distortion channel", while the test image passes through an additional "distortion channel" (e.g., blur, compression, etc.).

Information Content Weighted SSIM (IW-SSIM) [54] The IW-SSIM combines an information theoretic analysis of visual information content (similar to VIF), structural similarity based local quality measurement (as in SSIM), and multi-scale image decomposition followed by scale variant weighting (similar to MS-SSIM). The IW-SSIM provides the best overall performance reported in the literature when tested against six independent publicly available IQA databases.

Discussion These methods avoid many of the disadvantages of HVS-based methods by using real "natural" images instead of artificial test patterns. Their main disadvantage lies in the enormous space of possible images to model against. The overall effectiveness of these metrics is limited by the relatively small number of available test images.

Chapter 4

Evaluating Performance of IQA Metrics

This chapter describes the methods for evaluating performance of objective image quality metrics. These methods are categorized by three broad characteristics: prediction accuracy, prediction monotonicity, and prediction consistency. A non-linear mapping is applied to the objective quality scores before calculating prediction accuracy and prediction consistency scores. This mapping serves the dual purposes of accounting for nonlinearities in the subjective testing, and providing a common analysis space for multiple IQA metrics. [2, 3] No mapping is required for prediction monotonicity scores because they are non-parametric rank correlations.

4.1 Nonlinear Mapping

Given image i of N images, subjective opinion score o_i, and raw objective score r_i, a function q is generated (Equation 4.1) where the coefficients a_1 to a_5 are calculated through nonlinear regression to maximize the correlation between the subjective and objective scores.

$$q(r) = a_1 \left\{ \frac{1}{2} - \frac{1}{1 + exp[a_2(r - a_3)]} \right\} + a_4 r + a_5 \tag{4.1}$$

where r refers to raw objective quality scores and a_1 to a_5 are the fitted model parameters. No mapping is required for the prediction monotonicity measures because they use ranked values instead of objective scores.

4.2 Prediction Accuracy

The prediction accuracy measures indicate a model's ability to predict the subjective quality scores with minimal "average error". We use three different measures of prediction accuracy: Pearson's linear correlation coefficient (PLCC), mean absolute error (MAE), and root mean square error (RMSE).

4.2.1 Pearson Linear Correlation Coefficient (PLCC)

The PLCC is a parametric statistical measure of dependence between two variables (defined in Equation 4.2):

$$PLCC = \frac{\sum_i (q_i - \bar{q})(o_i - \bar{o})}{\sqrt{\sum_i (q_i - \bar{q})^2 \cdot \sum_i (o_i - \bar{o})^2}} \tag{4.2}$$

where o_i and q_i are the subjective and mapped objective scores, respectively. Values may range from -1 to 1, with 1 indicating perfect correlation and a value of 0 indicating no correlation. The sign of the result indicates the direction of correlation; we ignore the sign in our results (i.e., for our purposes, both 1 and -1 indicate perfect correlation).

4.2.2 Mean Absolute Error (MAE)

The MAE provides a more intuitive measure of error than the correlation coefficients because its units match those of the subjective scores being predicted. It is calculated according to Equation 4.3:

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |q_i - o_i| \tag{4.3}$$

where o_i and q_i are the subjective and mapped objective scores (respectively) and N is the number of scores.

4.2.3 Root Mean Squared Error (RMSE)

RMSE is similar to the MAE because it also provides an intuitive measure of error in subjective quality units. It differs in how it removes negative values; squaring the errors

and then taking the square root. This results in a measure that gives more weight to outliers than MAE:

$$RMSE = \sqrt{\frac{1}{N}\sum_i (q_i - o_i)^2} \tag{4.4}$$

where o_i and q_i are the subjective and mapped objective scores (respectively) and N is the number of scores.

4.3 Prediction Monotonicity

Measures of prediction monotonicity describe how well a model predicts changes in subjective scores; the model should predict a change with the same sign as any subjective score change. Large values indicate the two parameters tend to increase or decrease together. We use Spearman's rank order correlation coefficient (SRCC), as recommended in the VQEG final reports, and Kendall's rank order correlation coefficient (KRCC) as used in recent IQA research. [38, 54]

4.3.1 Spearman's Rank Order Correlation Coefficient (SRCC)

SRCC is similar to PLCC but only uses the ranks of the values. This makes it less susceptible to outliers than PLCC but also less sensitive to the distances between different values.

$$SRCC = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \cdot \sum_i (y_i - \bar{y})^2}} \tag{4.5}$$

where x_i and y_i represent the ranks of the subjective and objective scores.

4.3.2 Kendall's Rank Order Correlation Coefficient (KRCC)

KRCC is similar to SRCC but represents a probability (i.e., probability data is in same order vs. probability data is not in same order) where SRCC represents the proportion of variability accounted for.

$$KRCC = \frac{N_c - N_d}{\frac{1}{2}N(N-1)} \tag{4.6}$$

where N_c and N_d are the number of concordant and discordant pairs, respectively.

Table 4.1: Rough categorizations of correlation coefficient r.

$r \leq 0.35$	Weak correlation
$0.35 \leq r < 0.68$	Moderate correlation
$0.68 \leq r < 0.90$	Strong correlation
$r \geq 0.90$	Very strong correlation

4.4 Prediction Consistency

The outlier ratio (OR) gives a measure of how consistently the model predicts the subjective scores. It is a unitless value calculated by dividing the number of outliers (defined as values greater than two standard deviations from the mean) by the total number of values.

$$OutlierRatio = \frac{(number\ of\ outliers)}{N} \tag{4.7}$$

with an outlier defined as any value for which

$$|e_i| > 2 \times (DMOS\ standard\ deviation)_i \tag{4.8}$$

where e_i is the i_{th} residual between subjective and mapped objective scores.

4.5 Interpreting Correlation Coefficients

The correlation coefficients described above are abstract measures and cannot be precisely interpreted, but rough categorizations of correlation do exist, such as that shown in Table 4.1[1]. [47]

A fuller interpretation can be obtained through the coefficient of determination which is simply the squared value of the correlation coefficient (i.e., r^2). When applied to IQA, the r^2 value represents the percent of subjective quality variation (i.e., MOS or DMOS) that can be "explained" by variations in the objective model scores.

[1]Here and throughout the rest of this thesis, r refers to the correlation coefficient; not to be confused with the raw objective score used in Equation 4.1.

Part II

Solutions and Contributions

Chapter 5

Informal Evaluation of Existing Metric Performance

Our first task in evaluating tiled display image quality was to evaluate the performance of existing IQA metrics. As described in Chapter 3, the performance of these metrics is judged based on how well they correlate with subjective data stored in publicly available IQA databases. This presented our first problem because tiled distortion is a new distortion type and there had never been any subjective user studies conducted to obtain results for comparison. We therefore prepared a small, and informal, user study (roughly inspired by the method used in the CSIQ database, described in Section 3.1.2) to provide some insight into the performance of a well respected general-purpose IQA metric: the structural similarity (SSIM) index (described in Section 3.2.2).

5.1 Initial User Study (Informal)

We began by selecting one reference image from the LIVE database ("womanhat.bmp"). We generated a variety of grid-distorted images by applying a set of grids to this image that varied in width (from 1 to 3 pixels), frequency (4×4, 5×5, and 6×6 grid arrays), and intensity (black, gray, and white). The SSIM score was calculated (using the publicly available Matlab implementation [52]) for each grid-distorted image and a subset of 7 distorted images were selected to represent a broad distribution of SSIM scores. We then selected a subset of blur-distorted images from the LIVE database with a distribution of SSIM scores roughly equivalent to that of the grid-distorted images. These images

Figure 5.1: The photographs used in the informal user study interface, prior to sorting.

(1 reference image, 7 grid-distorted images, and 5 blurred images) were printed out on photo-quality paper for use in the user study. Figure 5.1 shows an example of the image photographs before sorting by the user.

The user study consisted of 2 stages: a training phase and an ordering stage. We used a different reference image for the training images to avoid influencing the user selections; the training images were meant only to acquaint the user with the procedure and provide a rough introduction to the ranges of quality he/she would encounter during the ordering stage. Aside from use of a difference reference, the training phase images were generated using the same procedure as those used in the ordering stage.

5.1.1 Training Stage

With the reference image in the middle, users were instructed to place one random image to the left and one different random image to the right. They were asked to look closely at the reference image (with instruction to consider that as "perfect" quality by definition) and then look at the others and decide which looks "better" with respect to the reference. This procedure was repeated for 6 image pairs (3 blur-distorted pairs and 3 grid-distorted pairs).

5.1.2 Ordering Stage

All photos were placed in random order on a large table. Users were provided with the reference image and instructed to place it at one end of the table (either left or right, as preferred by the user). Each user then arranged the other images in order of quality, with the "best" images on one end near the reference image and the "worst" images farther away from the reference. Extra care was taken to avoid effects of glare from lighting sources when comparing images.

5.1.3 Informal User Study Results

The study was performed using 10 people. One user's results were discarded as outliers based on discussion which indicated a misunderstanding of the instructions. Opinion scores were assigned to each image based on its placement relative to the reference image. Results were separated based on the distortion types (i.e., blur and grid) of the images. Blur images showed perfect (non-parametric) correlation (Figure 5.2) for every user while grid-distorted images fared relatively much worse (Figure 5.3).

The purpose of this study was not to provide statistically valid results upon which new metrics could be designed. The small sample size (10 subjects), imperfect image reproduction (printed photographs and their associated limited colour gamut instead of computer monitors), and lack of environment control (subjects were asked to evaluate images in various locations with varying lighting and other environmental factors) made this study a poor vehicle for evaluating firm results. Instead, this study served two purposes: it indicated the need for a larger and more formal user study (based on the poor performance of SSIM on the grid-distorted images) and it provided a "test run" to identify potential user study errors before performing a larger and more expensive formal user study (e.g., the

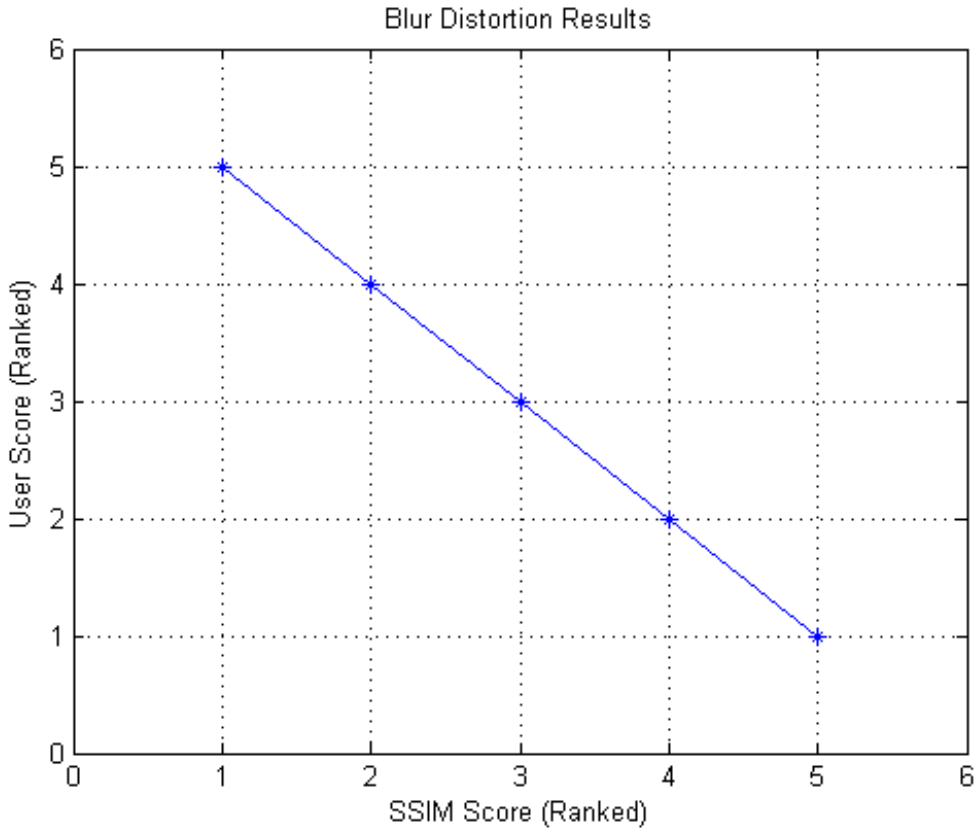


Figure 5.2: Results showing correlation between a typical user ranking of blurred image quality and corresponding (ranked) SSIM scores. Perfect correlation for all users.

importance of clear and specific instruction for the subjects was identified in the informal user study). Further details of this user study can be found in Appendix A.1.

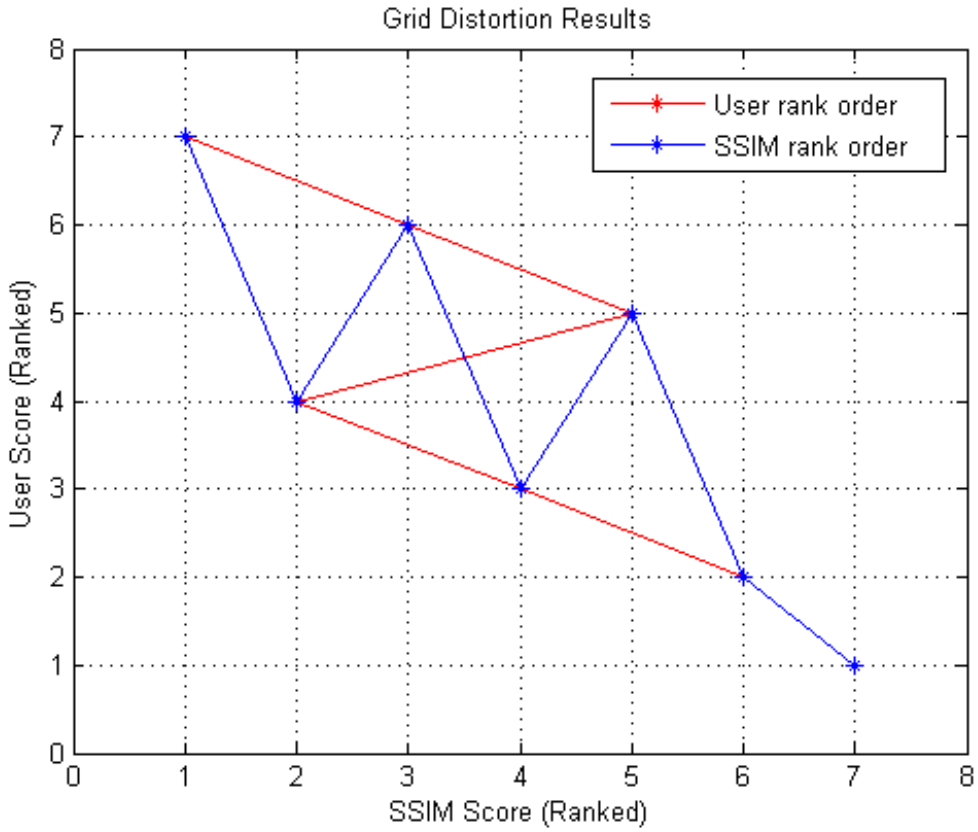


Figure 5.3: Results showing correlation between a typical user ranking of grid-distorted image quality and corresponding (ranked) SSIM scores. Note the correlation is much poorer than in the case of blur (average correlation of 0.8393).

Chapter 6

Formal Evaluation of Existing Metric Performance

Our informal user study (Chapter 5) suggested further research of existing metric performance was warranted but it was not robust enough to facilitate this research. Based on these results we performed two formal subjective quality studies to provide statistically valid data for development and testing.

6.1 Initial Formal User Study [29]

Our first formal user study was modelled after the procedure used to create the LIVE IQA database [44, 43]. The study consisted of 27 subjects; predominantly male undergraduate engineering students in the range of 18 to 22 years of age. Each subject viewed a series of images on a 27" ASUS VG278H IPS LCD monitor and provided a subjective quality score for each image. These image sequences contained a total of 144 images: 78 grid-distorted, 40 blur-distorted [1], and 26 undistorted reference images. Each grid-distorted image was corrupted by a two-pixel-wide grid (simulating a grid of roughly $1mm$ width) consisting of roughly 7×5 tiles (or 5×7 tiles for portrait images) with a pseudo-random intensity from one of three ranges: black [0,85], grey [86,170], or white [171,255]. The blur-distorted images were selected from a set of blur distortions found in the LIVE IQA database that covered a broad range of subjective quality (i.e., DMOS) scores. Further details of this user study can be found in Appendix A.2.

[1] The blur-distorted images were included primarily for cross referencing against the LIVE IQA database and to provide a "sanity test" to monitor the effectiveness of our methodology.

6.2 Expanded Formal User Study

Based on the experimental control provided by the blur-distorted images in our first formal user study, we were able to confirm our methodology was reliable. We then performed another, larger, formal user study where we removed the blur-distorted images to make room for more grid-distorted images. The new user study was performed in a very similar manner to that of Section 6.1 (and its proven methodology) but with the following differences:

The new user study was larger, with 33 subjects compared to 27 for the previous study, and recruited from a broader range of candidates with better gender representation. Where the first study consisted primarily of male engineering students, the second formal user study recruited students from a broad range of university faculties which resulted in a nearly-even division of gender. The modified recruitment also had the effect of lowering the number of subjects who had experience with image quality evaluation.

More images were evaluated by each subject in the second formal user study. In addition to the 26 undistorted reference images from the first formal study (25 images from the Kodak Lossless True Colour Image Suite [14] and one created using OpenStreetMap [1]), we added an additional eight new source images chosen from the Tecnick Testimages archive [48].

As in the first formal user study, each source image was distorted by the addition of a two-pixel-wide grid of roughly 5×7 for landscape images and 7×5 for portrait images. For our new study, we expanded the number of intensity ranges from three to five: "black" [0,50], "dark-grey" [51,101], "grey" [102,152], "light-grey" [153,203], and "white" [204,255]. We removed the blur-distorted images because they were no longer necessary for cross-referencing with an established image database. Our second formal user study contained a total of 204 images (170 grid-distorted and 34 reference) evaluated by each subject; this compares to 144 images per subject (78 grid-distorted, 40 blur-distorted, and 26 reference) in the first formal study. Even with the increased number of images, all sessions were still completed in under 30 minutes as recommended by the ITU BT.500 standard [4].

6.3 Formal User Study Results

Our metric testing results are shown in three separate tables: Table 6.1 provides results for the first formal user study, Table 6.2 shows results for the larger second study, and combined results are given in Table 6.3.

In Table 6.1, the relative rankings of the general purpose metrics roughly correspond to their rankings when tested against other common IQA databases (a reference set listing results of multiple metrics tested against multiple databases can be found in [54]), with the exception of VIF. The VIF metric performs better than SSIM when tested against most databases but our results show it performing below even PSNR in our first user study. Figures 6.1 and 6.2 show scatter plots of each metric against DMOS values for the first formal user study (blur distortions[2] and grid distortions, respectively).

Table 6.2 shows the results for our second (expanded) formal subjective user study. With 20% more subjects (33 vs. 27) and more than twice as many grid-distorted images (170 vs. 78), this study provides a better sample of subjective quality scores for tiled images. We note that VIF continues to perform poorly for tiled images but this time IW-SSIM also performs worse than normal, with performance even below SSIM despite being two "generations" newer. These results suggest two things: 1) information theoretic approaches do not work as well as structural approaches when measuring tiled image quality, and 2) evaluating grid-distorted images at multiple scales provides little advantage (based on the negligible improvements of MS-SSIM and IW-SSIM over basic SSIM). In fact, despite their much higher computational complexity, there is no statistically significant difference between the results for SSIM, MS-SSIM, and IW-SSIM. Figure 6.3 shows a scatter plots of each metric against DMOS values for the first formal user study.

The results in Tables 6.1-6.3 show that every general purpose metric we tested performs poorly for tiled images relative to traditional distortions. Pearson and Spearman correlation values that are typically above 0.85 [54] for traditional distortions are barely above 0.6 for tiled images (for example, IW-SSIM never drops below 0.8579, even on the "difficult" TID2008 database).

[2]We include the blur distortion results to illustrate the poor relative performance of the grid distortion results.

Table 6.1: IQA metric results for first formal user study.

	PLCC	MAE	RMS	SRCC	KRCC	OR
PSNR	0.5052	4.2268	5.2580	0.4914	0.3253	0.2308
SSIM	0.5683	3.8598	5.0134	0.5812	0.4086	0.2308
MS-SSIM	0.5954	3.7776	4.8952	0.6049	0.4279	0.2436
IW-SSIM	0.5968	3.7797	4.8888	0.5957	0.4252	0.2564
VIF	0.4646	4.2574	5.3951	0.4639	0.3134	0.2051
PSNR-HVS-M	0.5147	4.1664	5.2239	0.5017	0.3347	0.2436
MAD	0.5414	3.8751	5.1225	0.5811	0.4036	0.2051

Table 6.2: IQA metric results for second formal user study.

	PLCC	MAE	RMS	SRCC	KRCC	OR
PSNR	0.5552	5.5503	6.8115	0.5260	0.3692	0.3118
SSIM	0.6164	5.1986	6.4487	0.5989	0.4247	0.2882
MS-SSIM	0.6202	5.1975	6.4241	0.5882	0.4187	0.3000
IW-SSIM	0.6072	5.2860	6.5070	0.5739	0.4109	0.2941
VIF	0.5721	5.3599	6.7169	0.5320	0.3728	0.3059
PSNR-HVS-M	0.5603	5.5022	6.782	0.5381	0.3786	0.3177
MAD	0.6072	5.2553	6.5073	0.5895	0.4228	0.2765

Table 6.3: Combined results of first and second formal user studies.

	PLCC	SRCC
PSNR	0.5334	0.5108
SSIM	0.5955	0.5911
MS-SSIM	0.6093	0.5957
IW-SSIM	0.6026	0.5837
VIF	0.5264	0.5025
PSNR-HVS-M	0.5406	0.5221
MAD	0.5788	0.5858

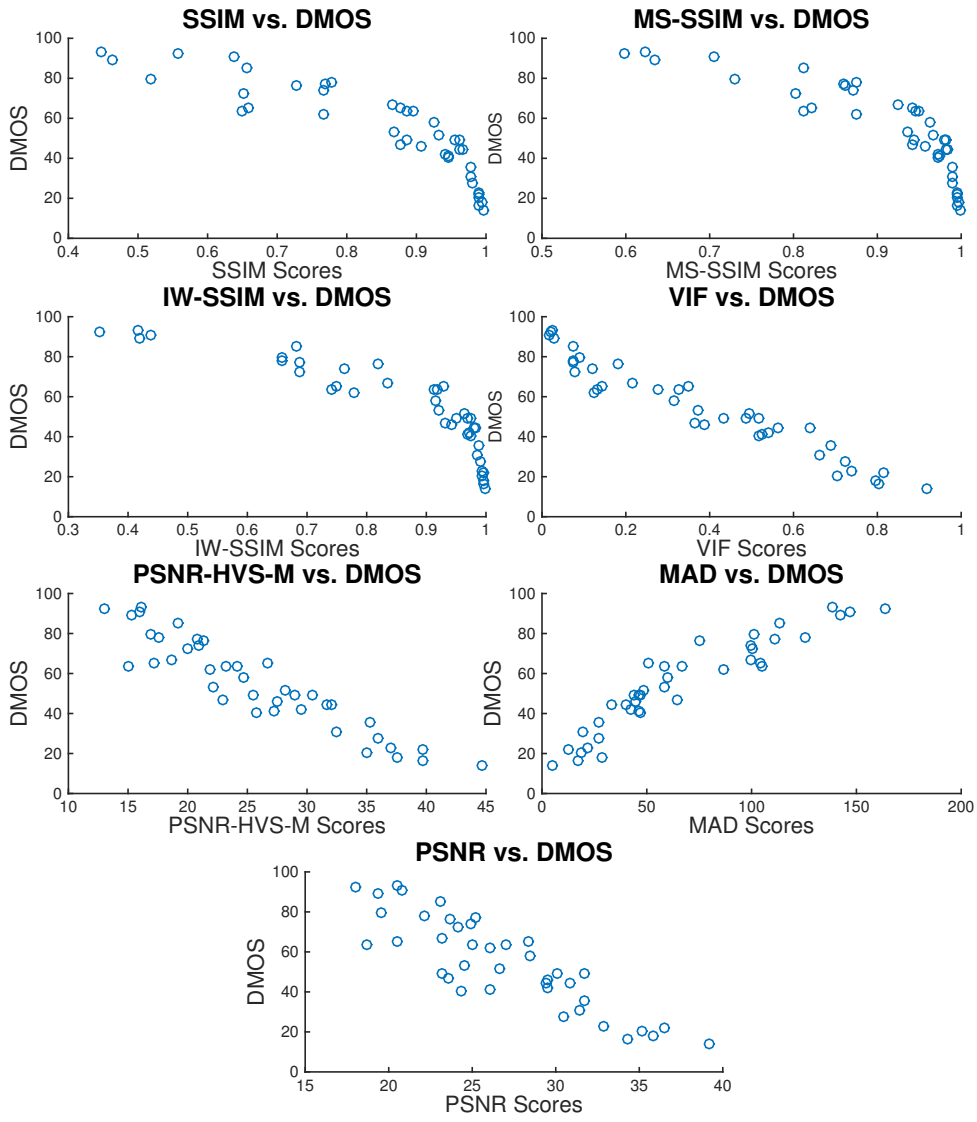


Figure 6.1: Results showing correlations between traditional IQA metrics and DMOS scores for the blur-distorted images in the first formal user study.

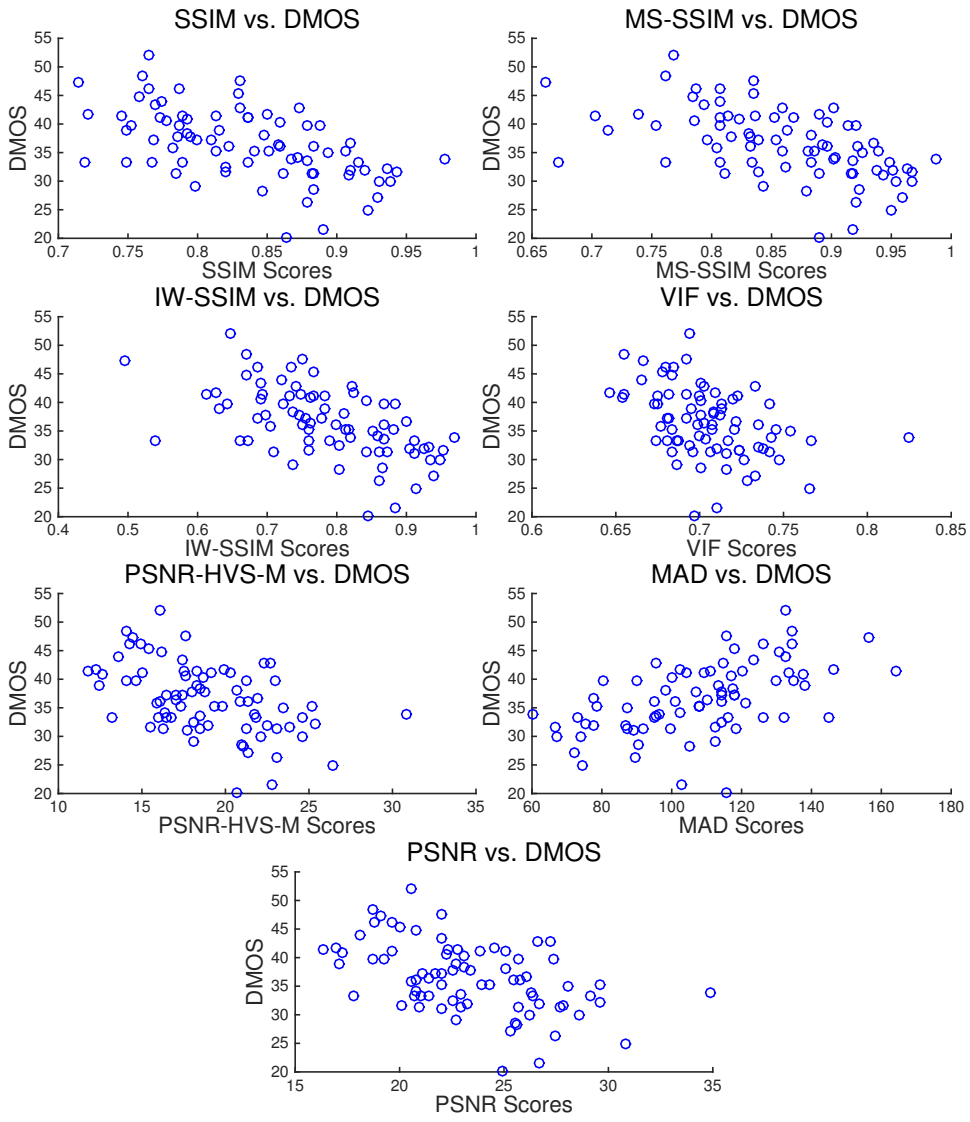


Figure 6.2: Results showing correlations between traditional IQA metrics and DMOS scores for the grid-distorted images in the first formal user study.

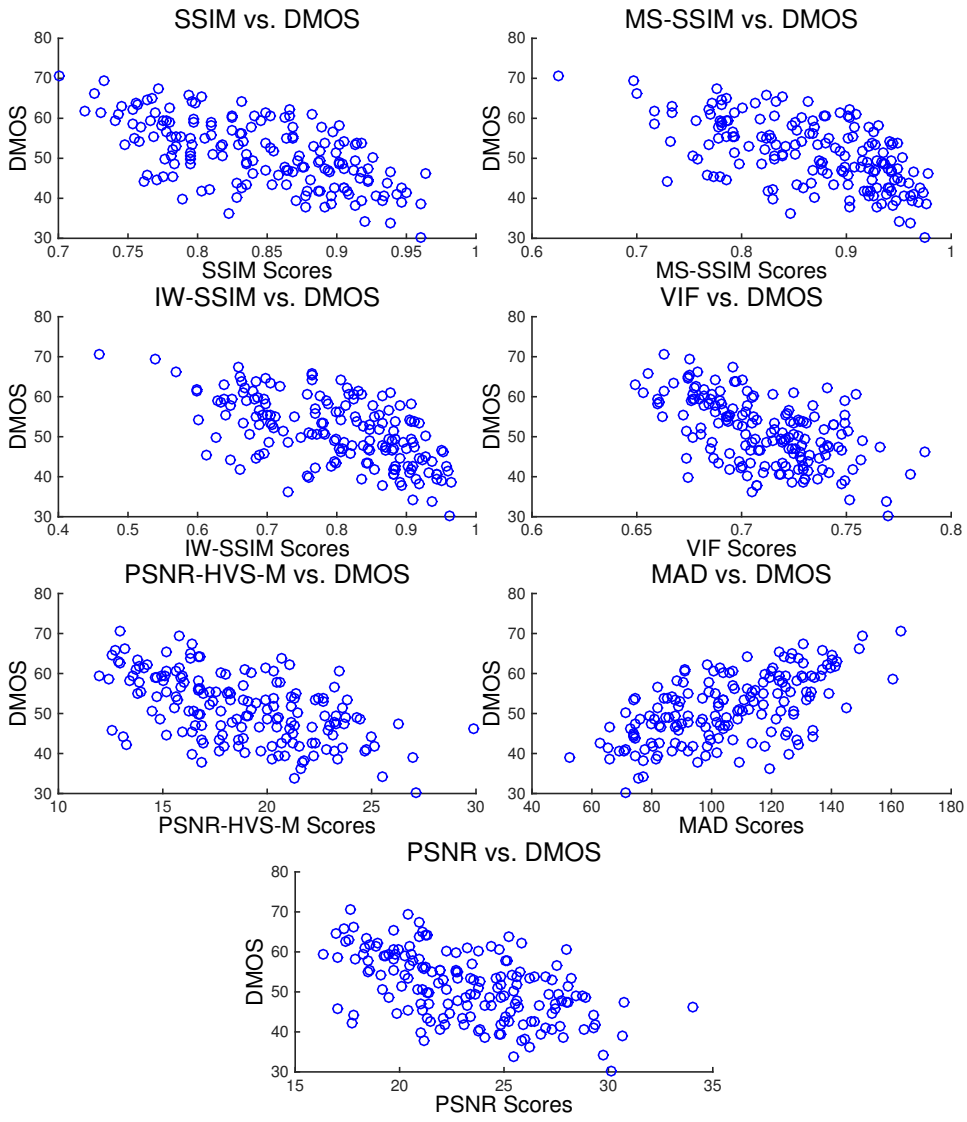


Figure 6.3: Results showing correlations between traditional IQA metrics and DMOS scores for the second formal user study.

Chapter 7

New Model Development

The data from our formal subjective user studies indicated a need for a new IQA metric for measuring quality of tiled images ([29, 31]). This chapter describes our development of a new, improved, metric for grid-distorted images.

7.1 Building Upon an Existing Metric

The metrics described in Chapter 3 and tested in Chapter 6 do not perform well for grid-distorted images but they can still be useful as a starting point when developing a new metric. The top-performing metrics (SSIM, MS-SSIM, and IW-SSIM) all had correlation coefficients of roughly 0.6, placing them near the high end of the "moderate correlation" category in Table 4.1. This moderate correlation represents an r^2 value of roughly 0.36; in other words, these metrics can account for approximately $\frac{1}{3}$ of the variation in subjective quality scores. Based on this pre-existing "moderate correlation", we elected to build on the performance of an existing metric instead of creating a new metric "from the ground up".

7.1.1 Metric Selection

Based on the performance results of Chapter 6, we selected the SSIM metric as a starting point for our new model development. The following factors influenced this decision:

Metric Performance: The structural metrics (SSIM, MS-SSIM, and IW-SSIM) performed best among the metrics tested. Though the differences were not statistically

significant, they were consistent across both user studies. (MAD was close behind on the second user study but performed poorly on the first).

Metric Recognition: The SSIM metric is among the most popular IQA metrics in use today. This metric has been extensively tested and has even been added to the Matlab Image Processing Toolbox as the only "modern" image quality assessment function [27] (i.e., excluding PSNR and the associated MSE).

Potential Gains: SSIM is an "untuned" IQA metric. There are no tuning parameters based on training data sets. This suggested more potential room for optimization than other tuned and optimized metrics.

Computational Simplicity: SSIM performed nearly as well as its derivatives, MS-SSIM and IW-SSIM, but SSIM is a far more computationally simple quality measure compared not only to its derivatives, but to all metrics we tested (with the exception of the PSNR metric). For example, results from [54] indicate MS-SSIM is roughly three times the complexity, IW-SSIM is roughly 13 times more complex, and MAD is more than 300 times the complexity (based on unoptimized computation times).

Metric Familiarity: We were already familiar with SSIM and derivatives from our previous work [30] and could leverage pre-existing software infrastructure we had created. This aided in test setup and experimentation. While not a strong factor on its own (we would have used the "best" metric regardless of our software infrastructure), it added "one more reason" to the other factors mentioned here.

7.1.2 Metric Analysis and Modification

Based on our results from [30] (which showed different behaviour between the luminance and constrast/structure terms of SSIM), one of our first steps in studying SSIM's effectiveness was a separation of its luminance and contrast/structure components. In doing so, we noted (experimentally) the impact of the luminance component was negligible for our tiled image results. Table 7.1 and Figure 7.1 compare SSIM performance on tiled images with and without the luminance component included. This lack of impact conflicted with our intuitive expectations that this relationship would be strong. For example, one would intuitively expect a black grid to look "better" (i.e., less noticeable) on a dark image while a white grid would look "better" on a white image.

To understand the discrepancy between these results and our intuitive expectations, we focused on an analysis of the luminance component. We found that a direct (weighted)

Table 7.1: Comparison of SSIM performance (correlation with subjective scores) with and without the luminance component. Differences are negligible.

Full SSIM	Contrast/Structure
srcc = 0.59893	srcc = 0.59898
krcc = 0.42471	krcc = 0.42513

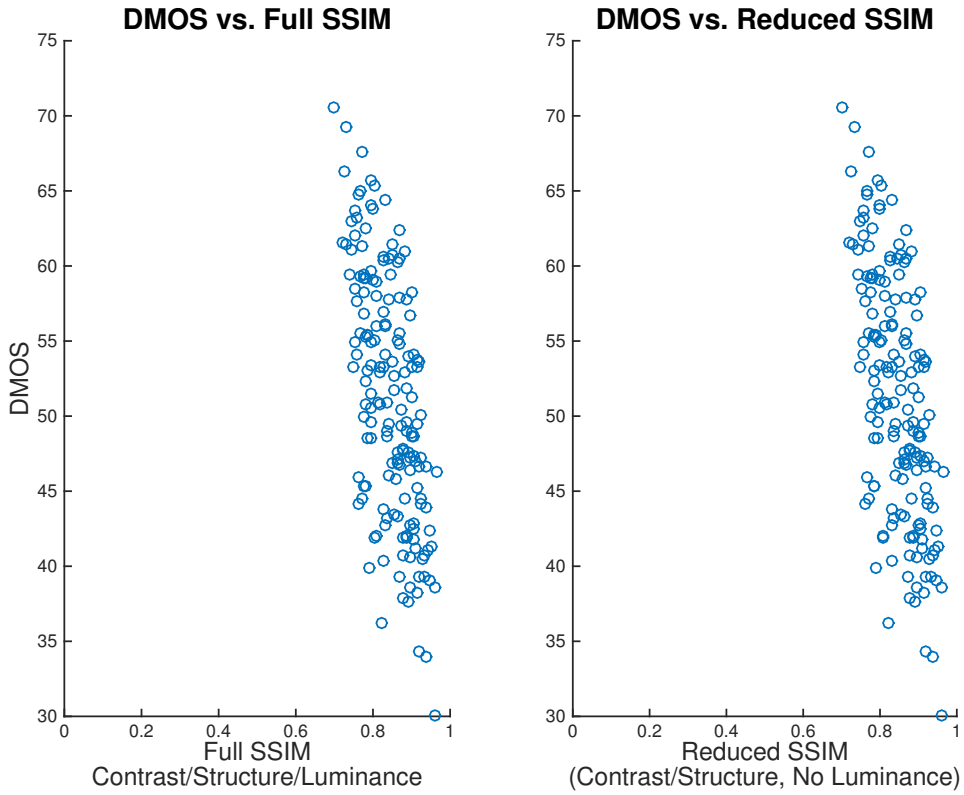


Figure 7.1: Comparison of SSIM performance with (left) and without (right) the luminance component. Scatter plots appear identical.

differencing of the grid intensity and image mean gave a much higher correlation with subjective scores than the SSIM luminance component. As a result, we removed the luminance term of the SSIM algorithm and replaced it with a "grid differential" term:

$$\Delta_g = abs(2\mu_{ref} - I_g) \tag{7.1}$$

where μ_{ref} is the mean intensity of the reference image and I_g is the grid intensity.

The weightings of this term were found through fitting of the data from the first formal user study. The reason for the heavier weighting on the reference image mean ($2\mu_{ref}$) is not immediately apparent until one examines the scatter plot shown in Figure 7.2. This plot shows a cluster analysis of the subjective quality scores as a function of the difference between grid intensity and reference image mean. The analysis generates two clusters, roughly divided by the $x = 0$ line (i.e., where the grid intensity and reference image mean are equal). While the clusters divide almost perfectly across the $x = 0$ line, it is important to note that the plot is not symmetric about this line. The points on the positive x-axis have a noticeably different slope and pattern compared to the points on the negative x-axis. This indicates that grids with intensities lower than the image mean correlate differently with subjective quality than grids that have intensities higher than the image mean. Table 7.2 provides an objective measure of the intuitive observation of Figure 7.2, showing the significantly different correlations as measured by PLCC and SRCC.

Table 7.2: Results of grid differential cluster analysis. Much higher correlation when grid is brighter than mean image intensity.

Grid Intensity > Reference Mean	SRCC	KRCC
TRUE	0.71679	0.51519
FALSE	0.25753	0.16815

These plots also explain the poor correlation of SSIM's luminance component. The calculation for this component is shown in Equation 7.2 for reference:

$$l(x, y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \tag{7.2}$$

where μ_x and μ_y represent the mean intensities of the reference and distorted images. (C_1 is a small constant selected to prevent instability when the denominator may otherwise approach zero).

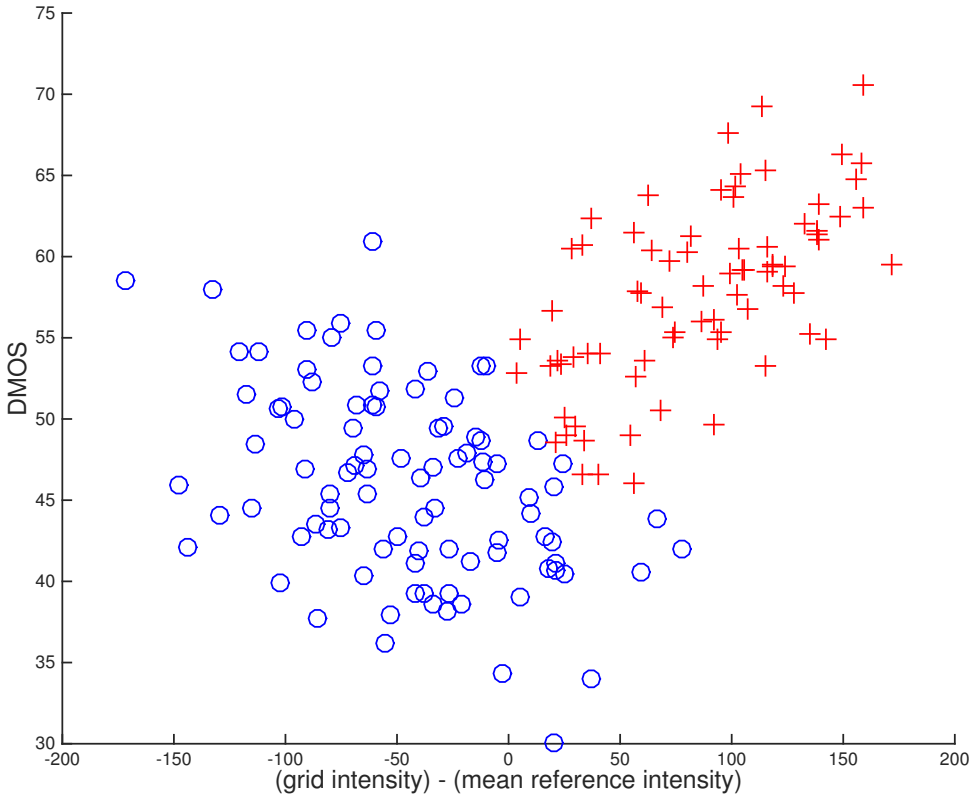


Figure 7.2: Cluster analysis of DMOS values vs. grid differential. Note the different relationships on either side of the $x = 0$ line.

Considering Equation 7.2 in tandem with Figure 7.2 and Table 7.2 illustrates why our grid differential term (Δ_g) outperforms SSIM's luminance term. Our grid differential term and the SSIM luminance term both measure the "difference" between reference and distorted images[1], but the SSIM term does not account for the "direction" (or "sign") of this difference (i.e., whether the grid, and resulting mean distorted image intensity, is higher or lower than the mean reference image intensity). By ignoring this sign, SSIM "assumes" the plot of Figure 7.2 is symmetric about the $x = 0$ line, which is clearly not the case.

Though we did not significantly modify the contrast/structure term (as we did the luminance component), we did apply a downscaling of this term. Unlike MS-SSIM (and its multiple downscaling passes), where performance was no better (or only marginally better) than SSIM, a single downscaling change was significant for the contrast/structure term and its correlation against subjective scores. This downscale operation gives less emphasis to localized distortions computed by the sliding window. The resulting emphasis on global measurements reflects that the quality impact of a grid distortion outweighs its local distortion weights because of its distributed nature.

Our new Tiled Display Quality Metric (TDQM) is formed as a weighted sum of our grid-differential term (Δ_g) and the contrast-structure component of the SSIM after downscaling the image ($CS_{\downarrow}$) [31]:

$$TDQM = W_1 \cdot \Delta_g + W_2 \cdot CS_{\downarrow} \tag{7.3}$$

W_1 and W_2 were experimentally determined (exclusively using the first user study) to be 1 and 4, respectively. The downscaling of the $CS_{\downarrow}$ term was experimentally found to be a factor of eight.

7.2 Results

We tested our new TDQM metric against the subjective quality scores obtained in Chapter 6 and present our results in three tables[2]: Table 7.3 provides results from the first formal user study, Table 7.4 provides results from the larger second study, and combined

[1] The grid differential term does this in a less direct manner, but the result is the same; the grid intensity will be generally be darker or lighter than the mean reference intensity and will cause a corresponding change in the intensity of the distorted image.

[2] These tables are the same as those in Chapter 6 but with the addition of the TDQM results.

results (PLCC and SRCC, combined using the methods described in Appendix B.3) are given in Table 7.5. The new TDQM metric statistically outperforms all other metrics based on PLCC comparison (with $p < 0.05$), and special note should be taken of the results for prediction consistency. The relative outlier ratios for the other metrics drop as their accuracy and monotonicity measures improve (e.g., MS-SSIM outperforms SSIM in every measure except for the outlier ratio). This indicates that only TDQM does not sacrifice prediction consistency for improved prediction accuracy and monotonicity. Scatter plots of the subjective scores vs. MS-SSIM and TDQM are shown in Figure 7.3.

Table 7.3: Expanded IQA metric results for first formal user study.

	PLCC	MAE	RMS	SRCC	KRCC	OR
PSNR	0.5052	4.2268	5.2580	0.4914	0.3253	0.2308
SSIM	0.5683	3.8598	5.0134	0.5812	0.4086	0.2308
MS-SSIM	0.5954	3.7776	4.8952	0.6049	0.4279	0.2436
IW-SSIM	0.5968	3.7797	4.8888	0.5957	0.4252	0.2564
VIF	0.4646	4.2574	5.3951	0.4639	0.3134	0.2051
PSNR-HVS-M	0.5147	4.1664	5.2239	0.5017	0.3347	0.2436
MAD	0.5414	3.8751	5.1225	0.5811	0.4036	0.2051
TDQM	**0.8347**	**2.7207**	**3.3550**	**0.8269**	**0.6330**	**0.0385**

Table 7.4: Expanded IQA metric results for second formal user study.

	PLCC	MAE	RMS	SRCC	KRCC	OR
PSNR	0.5552	5.5503	6.8115	0.5260	0.3692	0.3118
SSIM	0.6164	5.1986	6.4487	0.5989	0.4247	0.2882
MS-SSIM	0.6202	5.1975	6.4241	0.5882	0.4187	0.3000
IW-SSIM	0.6072	5.2860	6.5070	0.5739	0.4109	0.2941
VIF	0.5721	5.3599	6.7169	0.5320	0.3728	0.3059
PSNR-HVS-M	0.5603	5.5022	6.782	0.5381	0.3786	0.3177
MAD	0.6072	5.2553	6.5073	0.5895	0.4228	0.2765
TDQM	**0.7224**	**4.5308**	**5.6628**	**0.6873**	**0.4956**	**0.2059**

Table 7.5: Combined expanded results of first and second formal user studies.

	PLCC	SRCC
PSNR	0.5334	0.5108
SSIM	0.5955	0.5911
MS-SSIM	0.6093	0.5957
IW-SSIM	0.6026	0.5837
VIF	0.5264	0.5025
PSNR-HVS-M	0.5406	0.5221
MAD	0.5788	0.5858
TDQM	**0.7787**	**0.7582**

7.3 Conclusions

We have developed, using SSIM as a reference, a new quality metric (TDQM) specifically targeted towards measuring grid-distorted images. Our new metric shows a statistically significant improvement (with 95% confidence) over the best general-purpose image quality metrics at a computational cost below that of SSIM (one of the least computationally expensive modern metrics). The combined PLCC of 0.7787 indicates roughly 60% of the DMOS can be explained by our new metric; a significant improvement over the roughly 37% explained by the next best metric (MS-SSIM). Based on the categories of Table 4.1, we have moved from "moderate correlation" to "strong correlation" with our new metric.

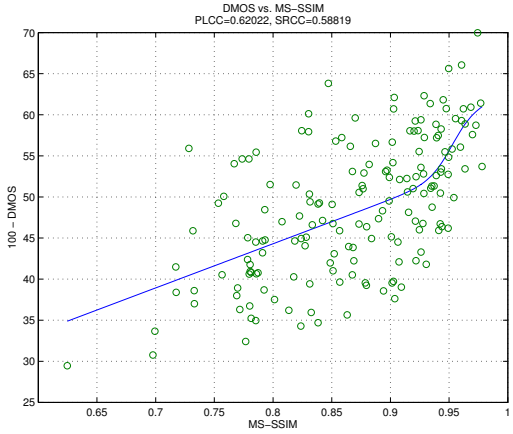
The performance of our new metric is also competitive with that of other modern IQA metrics when they are applied to "traditional" distortions such as blur and compression.[3] At first glance, the performance of the TDQM appears much lower than that of MS-SSIM. For example, when tested against the full TID2008 IQA database, MS-SSIM scores an SSRC of 0.8542 which is significantly higher than the SRCC of 0.7582 for TDQM on our new tiled databases. However, these MS-SSIM results are based on using images with a much wider range of subjective quality than what our tiled database contains (i.e., many image distortions in common IQA databases are sub- or near-threshold but all distortions in our tiled databases are supra-threshold). It was shown in [54] that metrics perform worse for low-quality images than for high-quality images (where a "low-quality image" was roughly defined as having a subjective quality in the bottom half for a given database). To obtain a better comparison, we evaluated the performance of some "traditional" metrics using the TID2008 database, but we did so while restricting the images to those in the approximate subjective quality range of tiled images.[4] The results of this evaluation are shown in Table 7.6. With these new "reduced quality range" performance scores, the performance of TDQM is competitive with (and even slightly better than) that of common general-purpose metrics.

[3] To clarify, we are referring to the performance of TDQM when measuring grid-distorted images compared to the performance of "traditional" metrics when measuring "traditional" distortions. TDQM is not a general-purpose metric and cannot be used for "traditional" distortions.

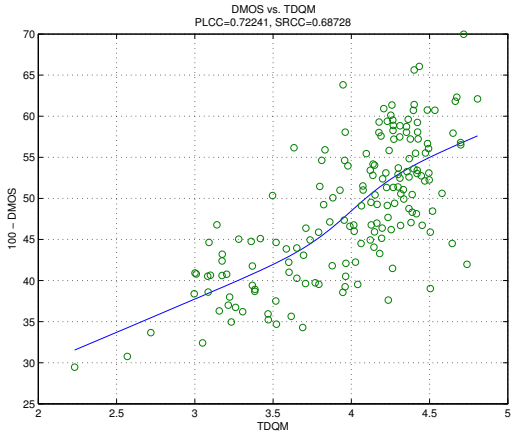
[4] This reduced quality range was estimated from the results of our first formal user study where blur distortions were included in the same sessions as grid distortions.

Table 7.6: Metric performance on TID2008 database when restricted to images within the same approximate quality range of tiled images. The SRCC for the TDQM metric is **0.7582**; roughly equivalent to the performance of MS-SSIM (highlighted) when applied to "traditional" distortions in the same subjective quality range.

	SRCC (full MOS range)	**SRCC** (reduced MOS range)	**Difference**
PSNR	0.5531	0.3845	-0.1686
SSIM	0.7749	0.6664	-0.1085
MS-SSIM	0.8542	**0.7553**	-0.0989
VIF	0.7496	0.6407	-0.1089
PSNR-HVS-M	0.5612	0.3739	-0.1873



(a) DMOS vs. MS-SSIM.



(b) DMOS vs. TDQM.

Figure 7.3: DMOS prediction of MS-SSIM and TDQM.

Chapter 8

New Algorithms for Improving Tiled Display Image Quality

Our work thus far has focussed on the measurement of tiled display image quality, but measurement is only half of the problem we wished to solve. In many image processing applications, the goal is to determine the correct pixels for a given spatial location in an image. The problem in tiled-display image processing is different: we know what pixel values should be in the location of the grid, but we have no means of directly displaying those pixels. Our problem thus becomes a question of perceptual image processing; we wish to modify the image in such a way that the grid (i.e., pixels we cannot modify) appears less objectionable. The following sections explain the algorithms we developed to perceptually improve the image quality of tiled displays.

8.1 Image-Correction Algorithm Theory

This section introduces the fundamental concepts that we used to develop algorithms to improve the perceived quality of grid-distorted images: edge brightening, and its trade-off, global darkening.

8.1.1 Edge Brightening

The primary concept for reducing the grid visibility is edge brightening[1], where the darkened grid "pixels" (which we cannot directly modify) are compensated for by increasing the intensity of adjacent pixels which we can directly modify.

Edge brightening makes use of the Point Spread Function (PSF) of the human eye. The PSF refers to the effect of passing a point source of light through an imperfect lens[6]. The diffraction-limited PSF, where the effects of defocus, aberrations, and scatter are ignored, provides the luminance distribution in the resulting image according to Equation 8.1:

$$L(\zeta) = \frac{[2J_1(\zeta)]^2}{\zeta^2} \tag{8.1}$$

where $L(\zeta)$ represents the relative light level at distance ζ from the center of the PSF, and $J_1(\zeta)$ is a Bessel function. In object space with the object at infinity,

$$\zeta = \frac{\pi\theta D}{\lambda} \tag{8.2}$$

where θ is the angular distance (in radians), D is the pupil diameter, and λ is the light wavelength.

An example of a point-spread function is shown in Figure 8.1.

The application of the PSF to improving tiled image quality relies on the effect shown in Figure 8.1. At sufficient viewing distances, the "spread" of any point source of light (i.e., any pixel) overlaps with one or more adjacent points (Figure 8.2). It is in this way that we can modify the *perceived* values of unmodifiable "grid pixels"; not by directly changing their values, but by changing the values of nearby pixels. A similar procedure has been used to hide individual "dead" display pixels[19, 20, 21, 32, 46] but these procedures aim only to hide a single defective pixel. Hiding a large supra-threshold distortion such as a grid is more difficult because each "grid pixel" has fewer adjacent "compensation pixels", and the grid is a global distortion that spread across the entire image (Figure 8.3).

It is worth noting that corner brightening is a special case of edge brightening. As illustrated in Figure 8.4, corner "grid pixels" have fewer adjacent "correction pixels". Therefore, any correction applied to these pixels must be greater than that of a typical grid correction pixel.

[1] In theory, edge darkening could be used for non-black grids but we focus on black grids based on the results from our prior subjective user studies and their common use in practice.

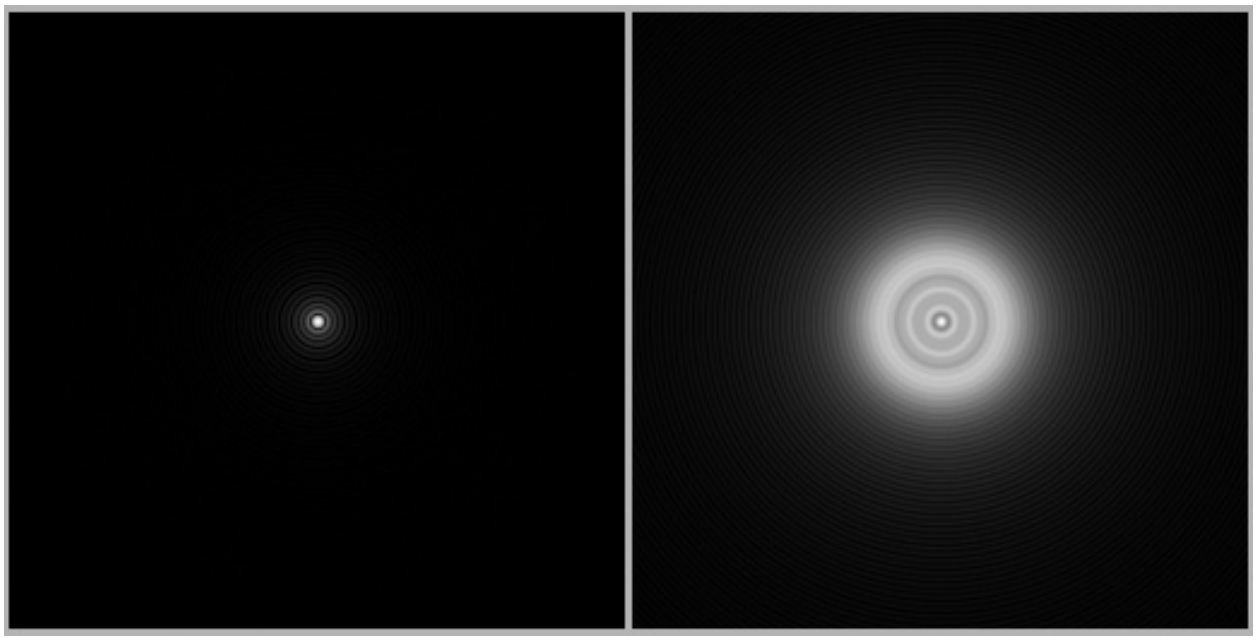


Figure 8.1: PSF example; (Left) Input point source; (Right) Output image.

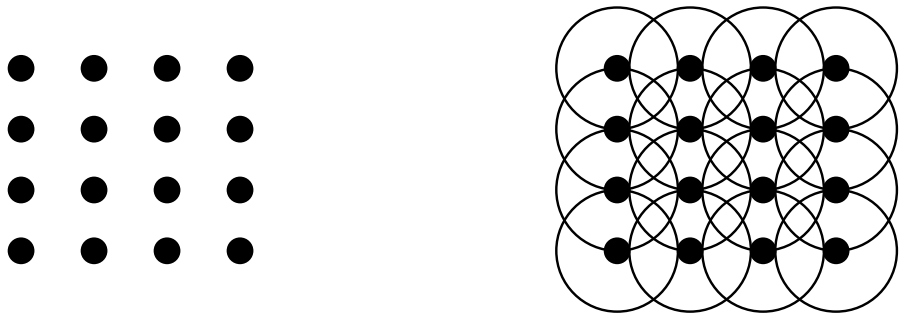


Figure 8.2: PSF illustration; (Left) Input point grid (i.e., pixels); (Right) Perceived image.

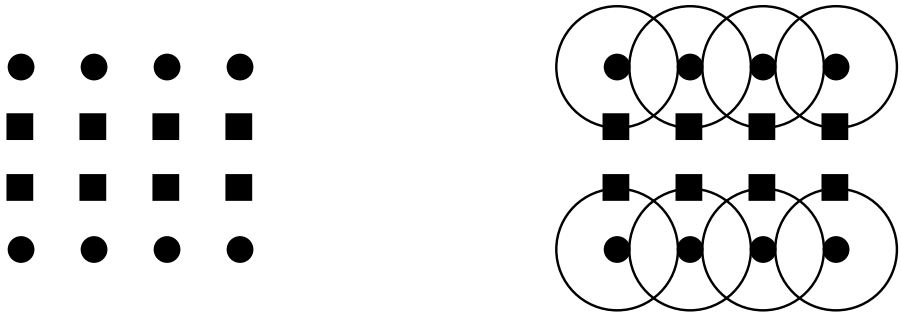


Figure 8.3: PSF illustration with Grid Line; (Left) Input point grid (i.e., pixels); (Right) Perceived image; squares represent "grid pixels". Note that each "grid pixel" has a minimum of three adjacent "correction pixels".

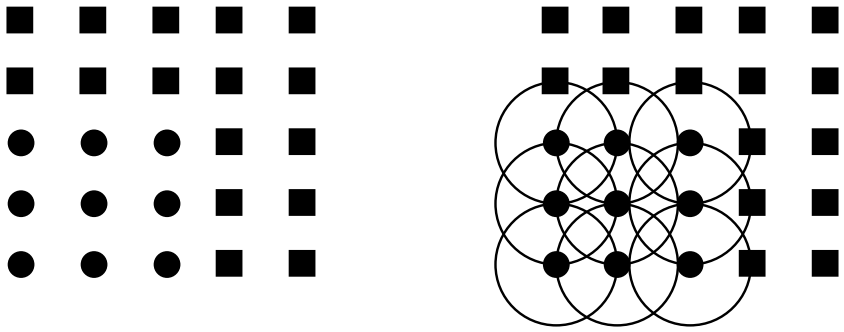


Figure 8.4: PSF illustration with Grid Corner; (Left) Input point grid (i.e., pixels); (Right) Perceived image; squares represent "grid pixels". Note that "grid pixels" have fewer adjacent "correction pixels" as they approach a corner.

8.1.2 Edge-Brightening Scenarios

There are two scenarios for use of the edge-brightening correction of Section 8.1.1:

- The image brightness is below maximum
- The image brightness is at maximum

Image Brightness Below Maximum

This scenario exists when a certain display has the capability to exceed the maximum desired brightness for a given environment (e.g., a darkened room). In such a case, edge brightening correction can be achieved by modifying the light source (i.e., back panel for LCD, lamp or LED brightness for projections) for the lines to be corrected. This may theoretically be done through optical or electronic means; for example, physical modification of screens, or by firmware modifications to internal electronics (e.g., a DLP chip in the case of projection displays). In this scenario, image-correction is entirely a function of determining the best "correction" values for the edge pixels.

Image Brightness At Maximum

This scenario reflects situations where a display is already operating at maximum brightness (e.g., an outdoor or otherwise brightly lit environment). In such cases, direct application of extra brightening to grid-edge pixels is not an option. An alternative is to apply contrast compression and map the pixel intensities of the original image to a smaller range. After such a compression, the image will appear darker, but there will be "room" to increase the brightness of pixels adjacent to the grid. Since brighter images are generally preferred, there exists a trade-off between perceptual thickness of the grid and global brightness (and contrast) of the image. We refer to this contrast compression as "global darkening" throughout this dissertation.

Chapter 9

Formal Evaluation of Image-Correction Algorithms

To test the effectiveness of perceptual grid correction, and develop a basic understanding of the dynamic between edge brightening and global darkening, we developed a user study incorporating six different algorithms[1] for comparison.

This formal user study was based on the methodology used for the TID2008 IQA database [38] but with some significant modifications. We recruited 31 subjects from undergraduate engineering and general graduate programs. Each subject was shown a series of image pairs and asked to provide their preferences between each pair. No formal visual acuity testing was performed on viewers, with verbal assurance of 20/20 vision accepted from each subject.[2]

9.1 Equipment

All images were displayed using a 23" Acer H236HLbid IPS LCD monitor set to its native resolution of 1920×1080 and factory default settings. The dot pitch of the monitor was slightly smaller than that of the display used in our first two formal studies (0.265mm vs. 0.311mm). No explicit calibration of the monitor was performed beyond visual inspection. Subjects were seated at a fixed distance of 1.5 metres from the display in a windowless room

[1] Strictly speaking, we use five algorithms (i.e., images modifications) plus an unmodified reference "algorithm".

[2] An "informal vision check" was performed by ensuring each subject could read the text on screen.

with typical office lighting. This distance is greater than the typical recommended viewing distance of $3-4$ times the image height and was deliberately selected to accommodate the testing of the grid-correction algorithms (recall from Chapter 8 that the PSF is dependent on viewing distance). This setup resulted in a density of roughly 99 pixels per degree.

Unlike the first two formal user studies, we accelerated this user study by running two sessions simultaneously on identical monitors driven by digital outputs (calibrated to sRGB) from two Macbook Pro laptop computers. This parallel testing allowed us to complete our study over a period of two days instead of the four days required for each of the first two formal studies. Aside from some minor inconveniences (sessions were more likely to be delayed if one participant was late, cancelled slots were more difficult to fill, etc.), there were no significant disadvantages created from this modification.

9.2 Images

We used a different methodology from our first two user studies and this change (described in Section 9.4) required us to reduce the number of source images we could use. As a result, we selected 16 reference source images from our second formal study (refer to Table 9.1 and Figure 9.1 for the images used; refer to Appendix B.4 for a detailed explanation of why we changed our methodology). Each source image was corrupted by a single grid distortion with a width of two pixels (as in the first two formal user studies) and a fixed intensity of '0' (based on our findings that this was the "best" quality fixed grid intensity). We then applied five different image-correction algorithms to each of these grid-distorted images to be evaluated alongside the uncorrected grid images (the original reference images, without the grid distortions, were not included in the study). This resulted in a total of 96 distinct images used in the study. The grid width was left at two pixels (to simulate a gap of roughly 1mm on a tiled display) because the difference in dot pitch between the Acer monitors and the ASUS display was not considered significant enough to warrant a modification.

9.2.1 Image-Correction Algorithms

We applied six corrections to our reference images for evaluation (illustrated in Figure 9.2). We used only algorithms with fixed parameters (e.g., no dynamic global darkening based on image brightness) to gain a clearer understanding of the different components (i.e., edge brightening vs. global darkening). Due to the restricted study size, dictated by our choice to use round-robin evaluation, we could not include a dynamic algorithm for comparison.

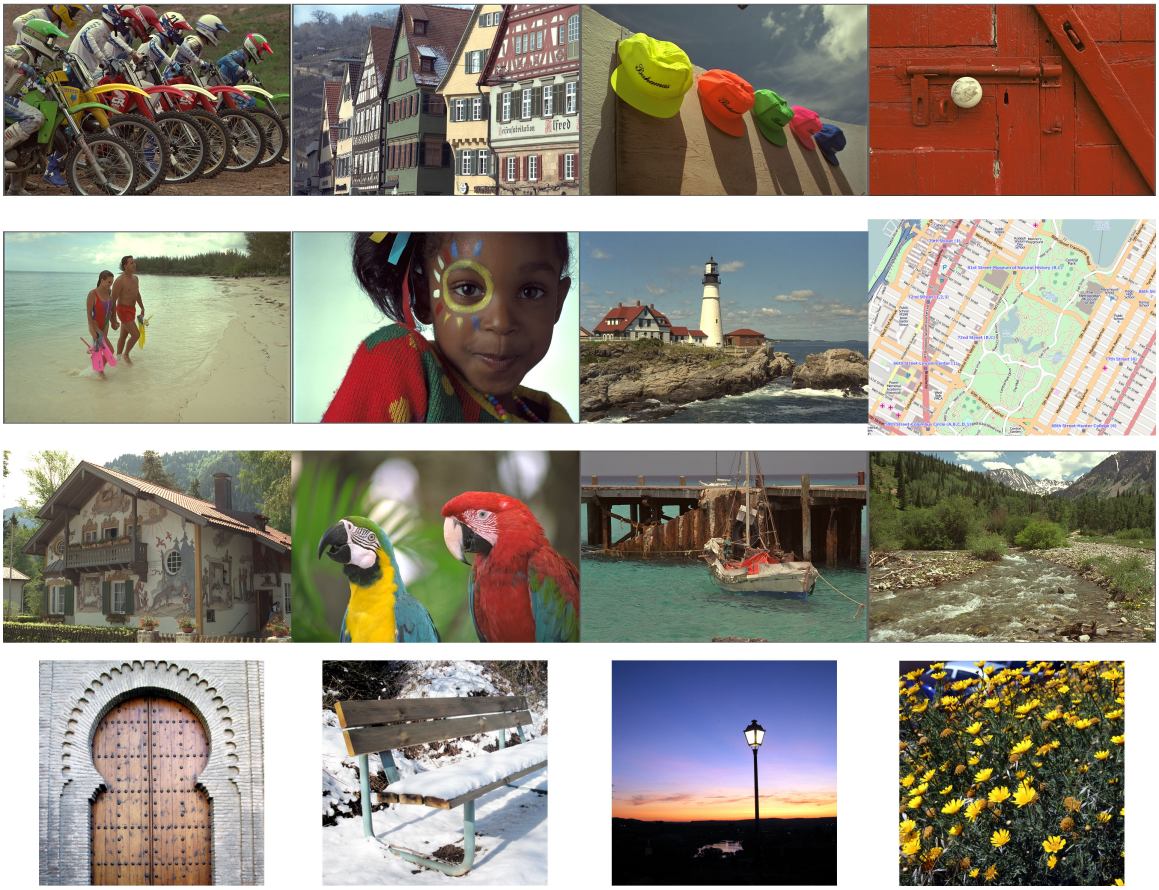


Figure 9.1: Reference images used in the image-correction user study.

Table 9.1: Reference images used in image-correction user study.

Reference	Mean	Median	StdDev	Description
bikes	82.8	77	48.3	Off-road motorcycles in a row
buildings	122.8	107	63.1	Buildings viewed at an angle
caps	102.1	99	39.2	Hats hanging on a board with sky background
kodim02	79.4	79	20.6	Red door and latch with small white handle
kodim12	162.1	168	45.8	Couple walking on beach
kodim15	108.7	74	82	Face with paint around one eye
lighthouse2	115.9	125	42.3	House and lighthouse against sky and water
map	216.8	229	35.8	Road map showing subset of Manhattan
paintedhouse	109	98	53.5	House with murals painted on sides and front
parrots	109.5	98	46.4	Two parrots against a blurred background
sailing4	93.2	106	41.5	Boat in water with dock in background
stream	108	98	52.6	Stream flowing from mountain range
testim008	163.6	175	52.2	Large wooden door inside stone arch
testim027	161.4	179	73.7	Bench covered in snow
testim036	102.5	101	77.5	Lamp post at dusk
testim098	78.6	59	61.5	Yellow daisies

Algorithm 0

This "algorithm" left the grid-distorted reference images unchanged, with no edge brightening and no darkening. These images represent typical uncorrected images shown on tiled displays.

Algorithm 1

This algorithm performs no edge brightening but applies global darkening (i.e., contrast compression) of 40%: the pixel range of 0-255 is scaled to the range 0-182. These images represent a common reference for comparing edge brightening with no clipping concerns.

Algorithm 2

This algorithm performs 40% global darkening of the images, followed by a 40% "step correction" edge brightening. Step correction refers to brightening the single row (or column) adjacent to the grid (on each side). This is the simplest form of edge brightening.

Algorithm 3

Algorithm 3 applies a "sinc correction" brightening to the undarkened grid-distorted reference images. This correction applies 40% brightening to the first row/column, 20% darkening to the second, and 10% brightening to the third. Since there is no global darkening, pixels that are already above the level of 182 will clip at 255.

Algorithm 4

This algorithm applies the same sinc correction as Algorithm 3 (40/-20/10), but does so after applying a 20% darkening to the image ([0,255] scaled to [0,212]). This algorithm represents a trade-off between global darkening and potential clipping of pixel values.

Algorithm 5

Algorithm 5 applies the same sinc correction (40/-20/10) as Algorithm 3 and Algorithm 4, but does so after darkening the image by 40% ([0,255] scaled to [0,182]). These images allow for full effects of edge brightening with no clipping.

Corner Correction

All algorithms that use edge brightening (i.e., Algorithms 2–5) applied an extra corner brightening of 20%.

9.3 Subjects

For this study, we collected (voluntary) information from the user study participants. We recorded each subject's gender, age (or age range if the subject preferred), correction of vision (i.e., uncorrected or glasses/contacts), and naivety in regard to image quality evaluation. We also noted the amount of time each subject required to complete the experiment portion of the session. These results are summarized in Table 9.2.

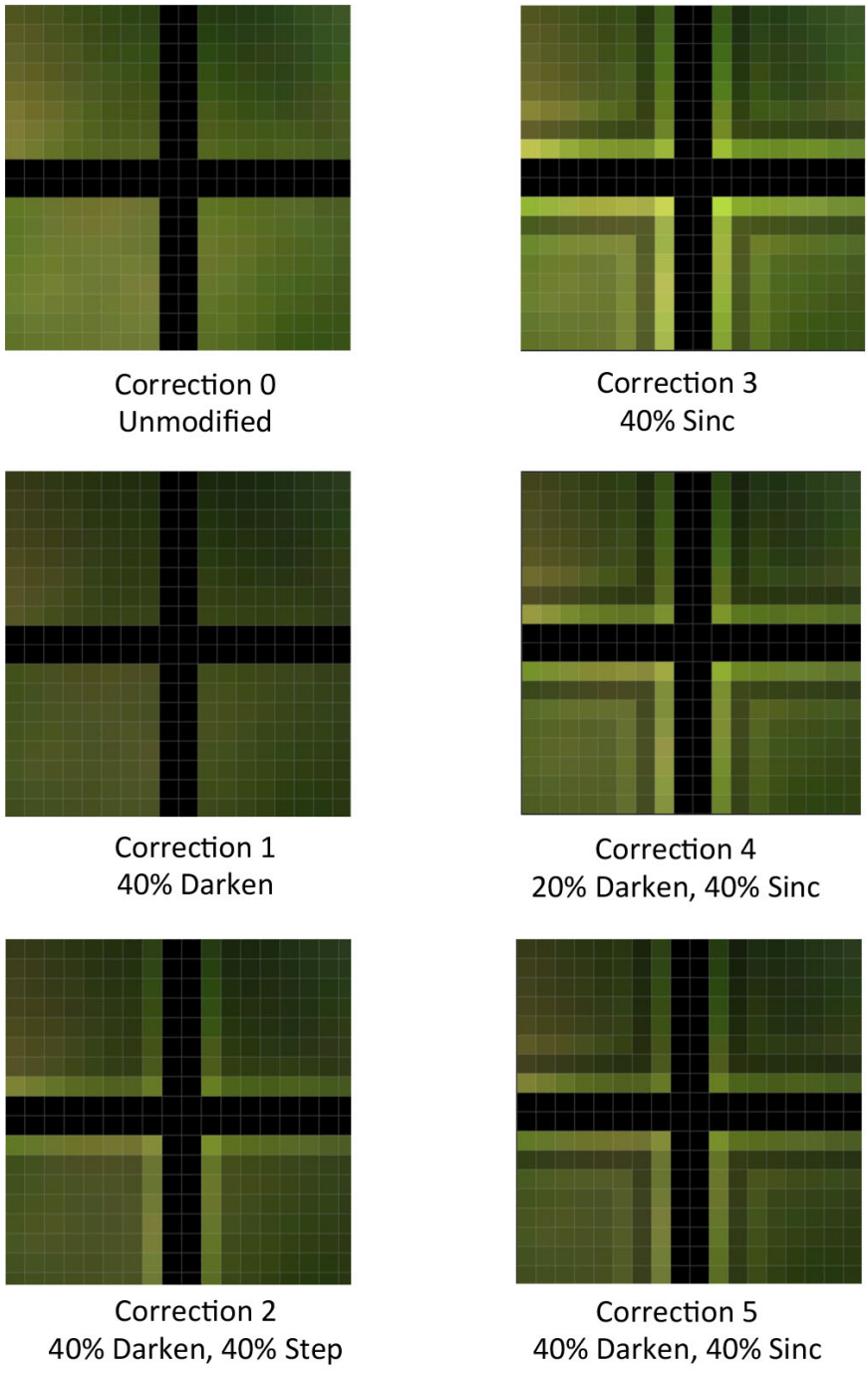


Figure 9.2: Image correction algorithms.

Table 9.2: Participant summary for image-correction user study.

Duration Mean	17:39 min
Duration StdDev	3:45 min
Male Subjects	22
Female Subjects	9
Uncorrected Vision	18
Corrected Vision	13
Age Mean	26.4 years
Age StdDev	5.1 years

9.4 Methodology

Each session of our initial user study was divided into three parts:

1. Instruction:
   - Subjects were provided written and verbal instructions for the session.
2. Training:
   - Identical to "Experiment" (see below) except for a shorter duration (i.e., fewer images) and use of different reference images. Subjects were encouraged to ask any questions during this phase.
3. Experiment:
   - Unlike the first two user experiments, we used a paired image, forced choice methodology similar to that used in creating the TID2008 IQA database [38]. This paired comparison method is commonly used in detection studies (e.g., detecting an object or distortion present in an image), but Ponomarenko et. al. demonstrated its usefulness for computing MOS scores. We modified their methodology in a few significant ways:
     – We did not show the undistorted (i.e., no grid distortion) source image alongside each corrected image pair (Refer to Figure 9.3). Our goal in the study was not to determine fidelity to the original undistorted image, but

instead to determine subject preferences between various distorted images.[3] With a supra-threshold distortion such as a grid, we believed including an undistorted reference image would cause subjects to ignore subtle differences between grid-distorted and corrected-grid-distorted images and view both images as "bad".

  - We provided four choices instead of two for each image pair. Each subject selected their preferred image, as in the TID2008 study, but also indicated how certain they were of their selection (refer to Figure 9.3). This change in methodology served two purposes: 1) users were given a "less severe" option for cases where they believed the quality differences were minimal (or even non-existent), and 2) we were given more data to distinguish between the effectiveness of different algorithms.
  - We used a Round-Robin Tournament scoring method instead of the Swiss Tournament scoring method used in the TID2008 database. This decision provided better granularity in our scoring results at the expense of reducing the number of images we could include for evaluation. We further explain our motivation for this decision, and the resulting trade-offs, in Appendix B.4.

- Each subject selected a quality score for each image pair using one of four radio buttons in a Java application similar to that shown in Fig. 9.3. Two radio buttons were placed under each images with buttons having one of the following labels: "Certainly Better" or "Probably Better". Subjects were required to select a quality score for each image pair before the next pair could be displayed. Image pairs were shown in pseudo-random order with the restriction that no consecutive images could share the same correction algorithms. All sessions were completed in under 30 minutes, as recommended in the ITU standard. To maximize the number of images and corrections in the study, each subject viewed every image only once. To account for potential bias in left/right vs. right/left image pair order, we ensured half of our subjects viewed the sequence with the image placement reversed.

[3]From a full-reference IQA perspective, our "image corrections" can technically be considered "image distortions" since they modify pixels that already perfectly match those in the reference source.

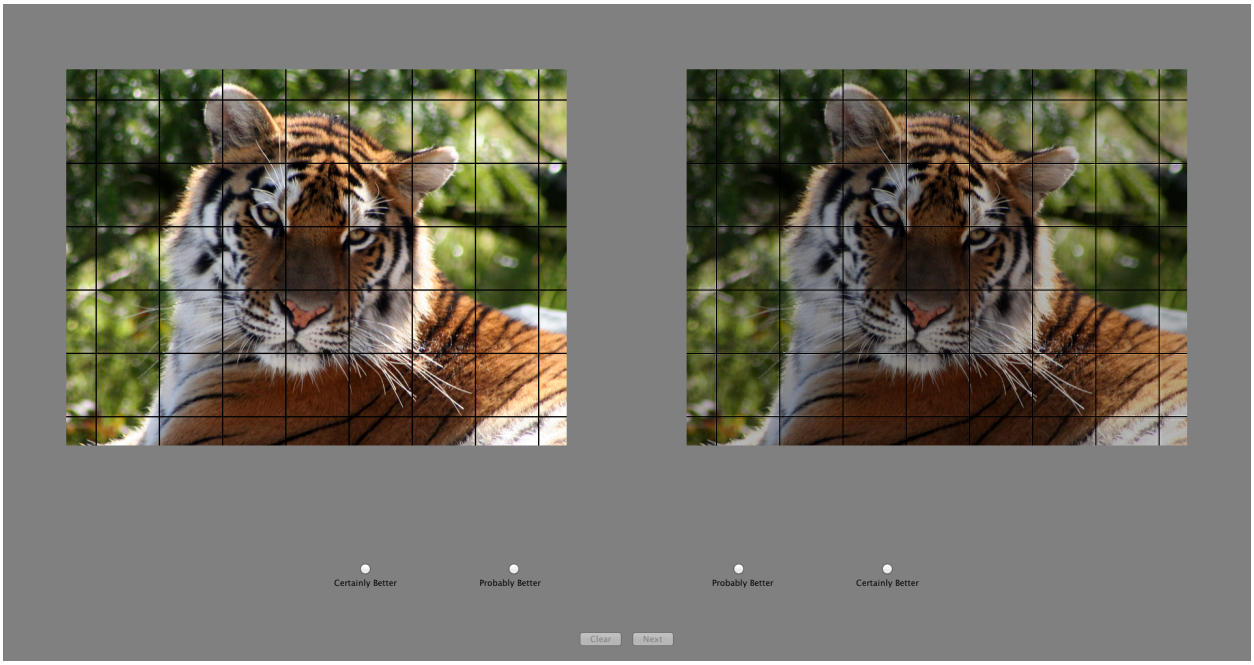


Figure 9.3: The image-correction user study interface. The 'Next' button is shown inactive because the subject must select a score before moving to the next image. Left/right ordering of images is reversed between viewing sessions.

9.4.1 Scoring

We converted user selections to opinion scores by assigning "points" to each image based on each selection according to Table 9.3:

Table 9.3: Scoring of images in the correction user study.

User Selection	Points
"Certainly Better"	2
"Probably Better"	1
"Not Selected"	0

Every image begins a session with a score of zero points, and this score is increased by the amount listed in Table 9.3 every time a selection is made (i.e., once for each time the image is displayed). For the Round-Robin Tournament scoring method we used, each image is compared once against every other image (that shares the same reference). Therefore, with six image-correction algorithms applied to each reference, each image is compared with another image a total of five times. This gives a possible total score of between "0" and "10" for each image (specific details of the Round-Robin Tournament method are described in Appendix B.4). The total scores are then averaged across all users to obtain a mean opinion score for each image. Based on the non-symmetric distributions of image scores among participants, we favour median opinion scores over the more commonly used mean opinion scores, but we include both in our results for comparison.

9.5 Results

Results from our user study are shown in Figures 9.4 – 9.6.

Figure 9.4 shows the distributions of opinion scores, by correction algorithm, across all reference images. We include both mean and median opinion scores because the score distributions for each image were highly non-normal and non-symmetrical (thus justifying the inclusion of median opinion scores in addition to the more common mean opinion scores; all plots are included in the appendices). Notches on each box plot indicate the 95% confidence interval of the median.

Figure 9.5 shows the distribution of median opinion scores across all images for each correction. This plot clearly shows that Correction Algorithm 1 (darkening-only) is unani-

mously the worst "correction" while the others are less agreed upon. (We include a similar plot, using mean scores, in the appendices for completeness).

Figure 9.6 presents the algorithm scores by considering only their rankings (i.e., how many times each algorithm finished first, second, etc.). These distributions closely mirror the results of Figure 9.5 but with a smaller spread. Rankings averaged over all images are shown in Table 9.4.

Table 9.4: Correction algorithm rankings averaged over all images. Lower is better.

Correction Algorithm	**Average Ranking** (out of 6)
Algorithm 3	**1.5**
Algorithm 4	1.6875
Algorithm 5	2.875
Algorithm 2	2.875
Algorithm 0	4.875
Algorithm 1	6.0

Detailed score distributions for each image and correction can be found in Figures A.2 through A.7 in the Appendices.

We do not include results for TDQM or other objective models because our sample sizes are too small to provide consistent and meaningful data. Future user studies will correct this by including an extra realignment component to allow comparisons between different source images.

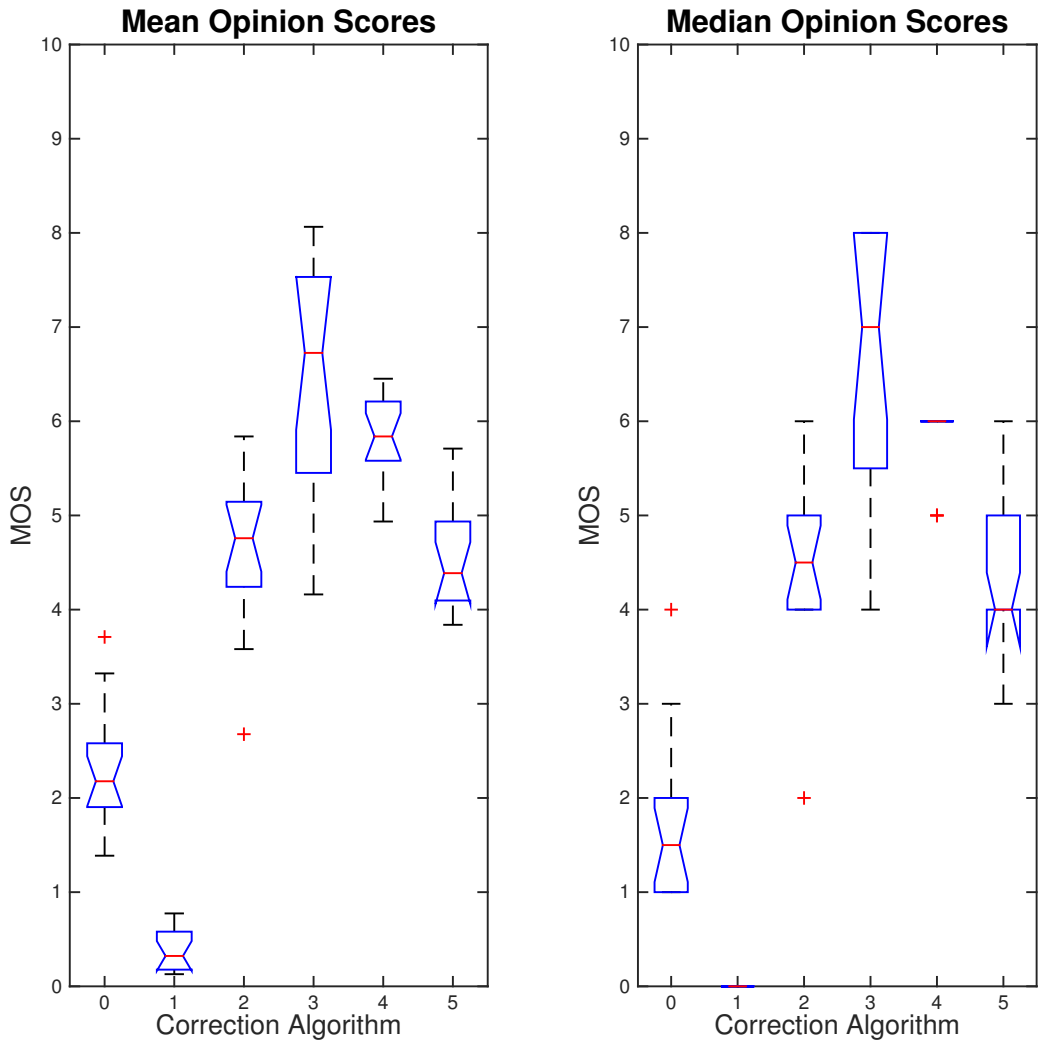


Figure 9.4: Mean and Median Opinion Scores across all images. Median scores are included with the traditional mean scores due to the non-symmetric score distributions of many images (included in appendices).

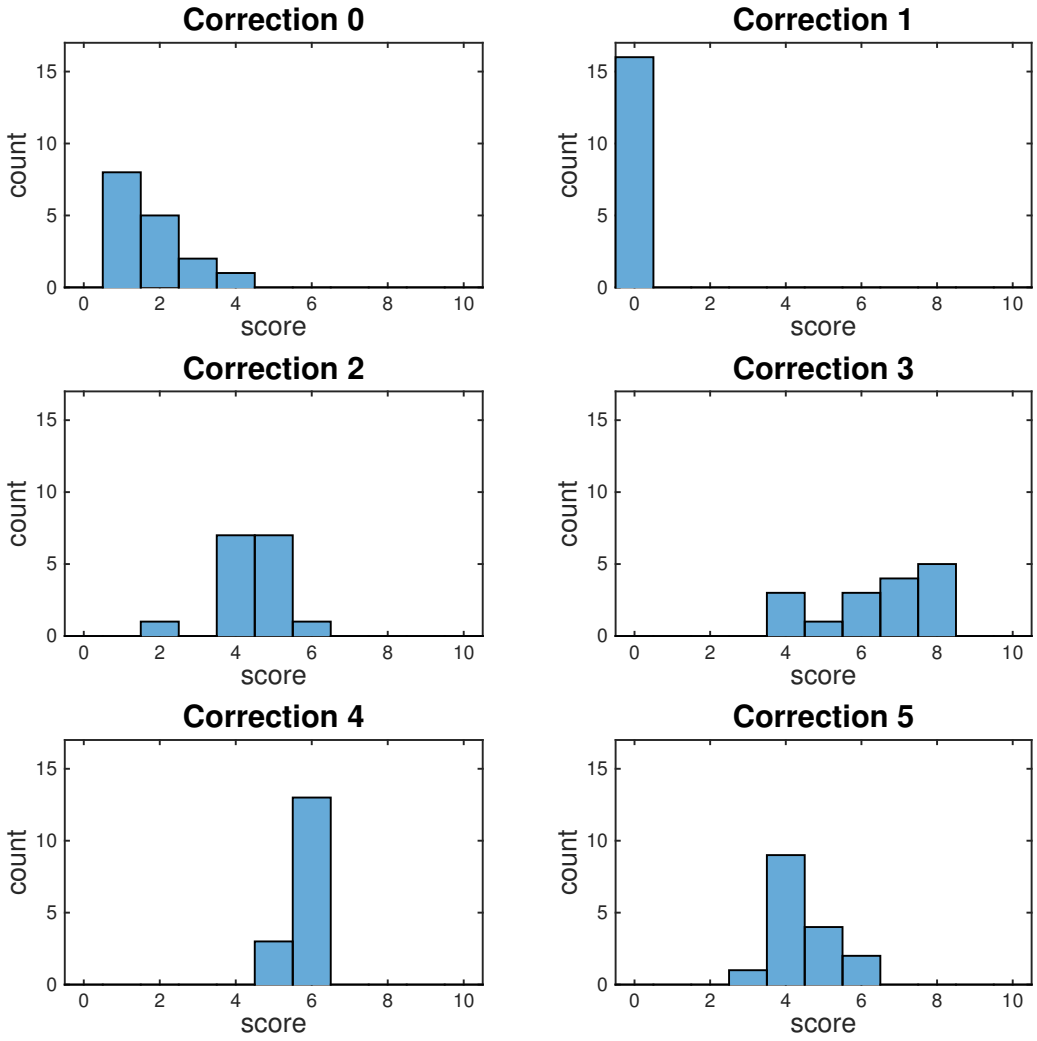


Figure 9.5: Distribution of median opinion scores for each correction across all images. (Mean scores are included for reference purposes in the appendices).

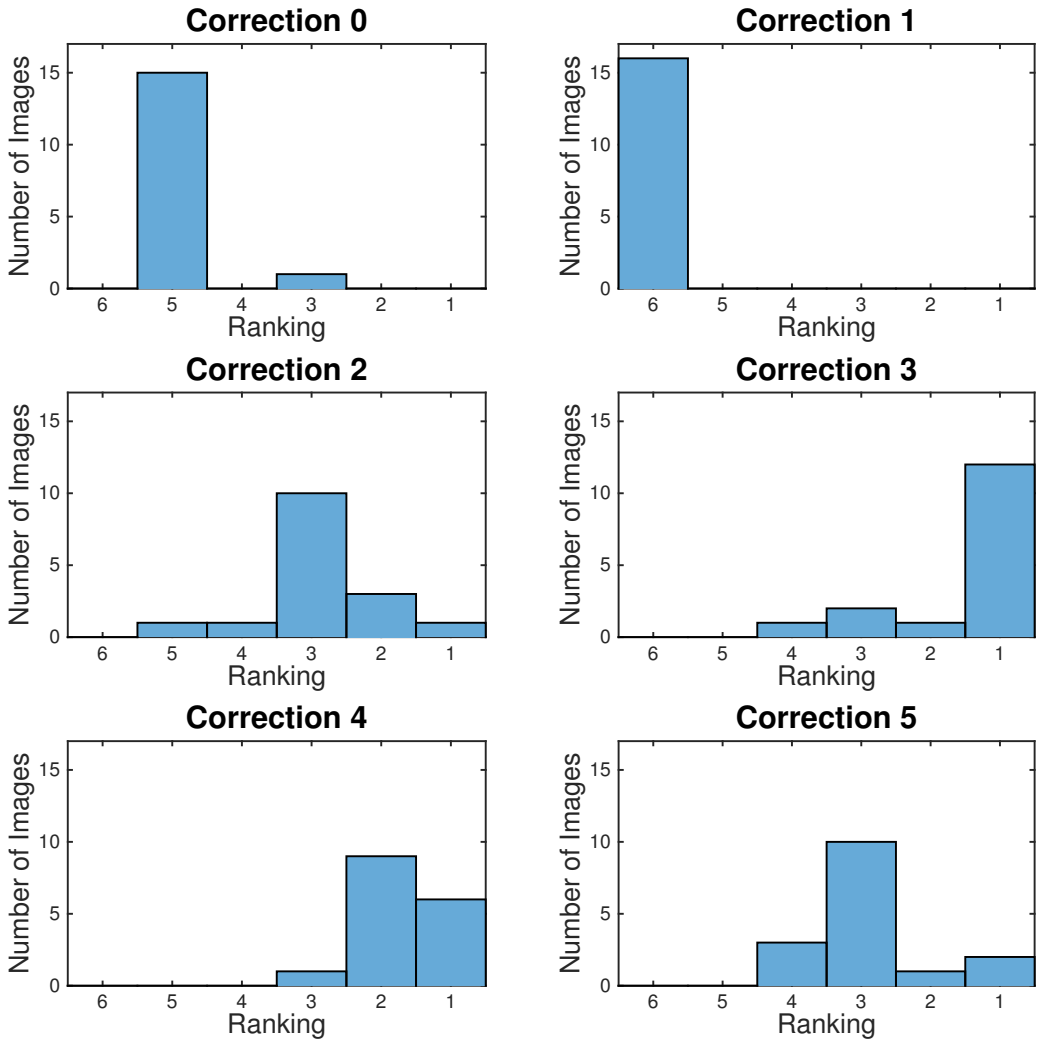


Figure 9.6: Distribution of rankings for each correction across all images.

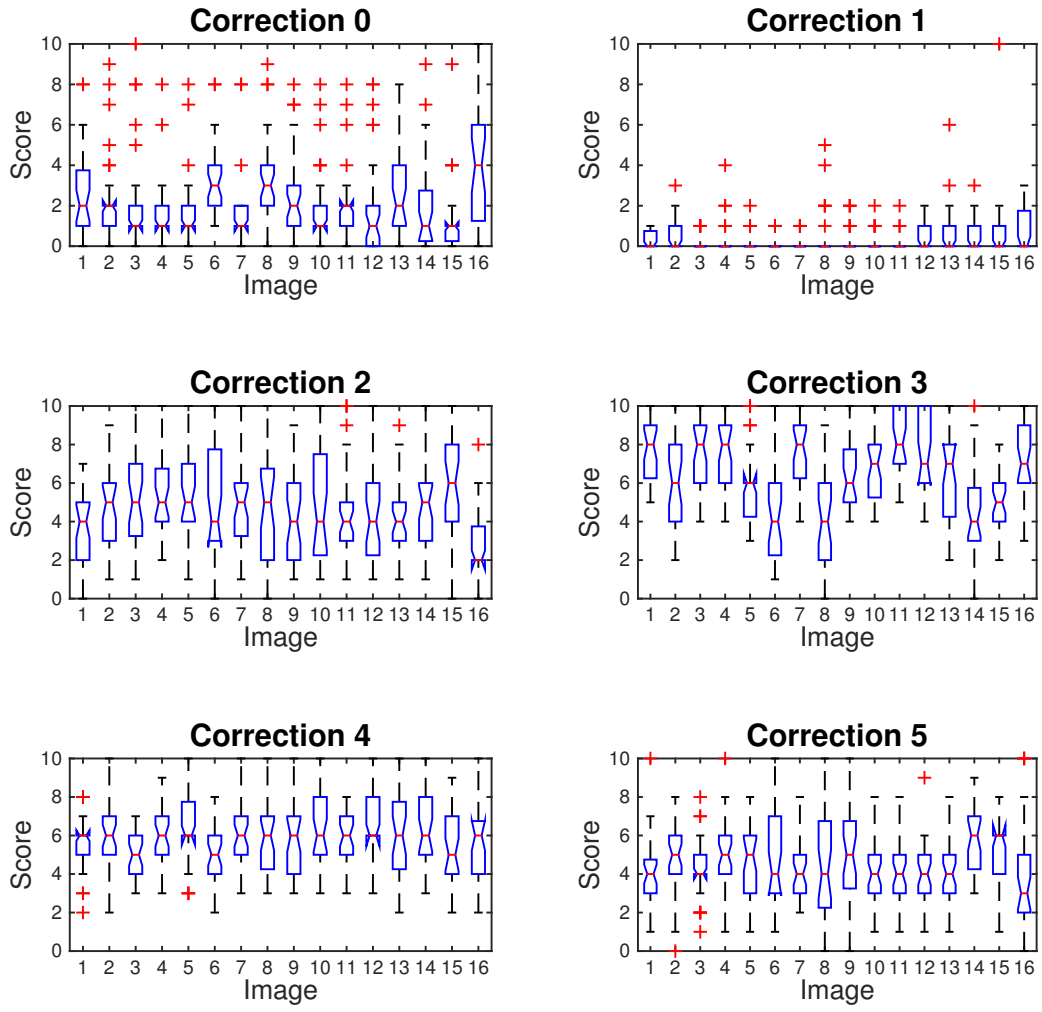


Figure 9.7: Distribution of opinion scores for each image and correction.

9.6 Conclusions

Based on the results of Section 9.5, we note the following key points (all statements of "statistically better" or "statistically worse" refer to a 95% confidence interval):

1. The "darken-only" algorithm ("Correction 1") is statistically worse than the "no-modification" algorithm ("Correction 0"), with 100% of samples (both for mean and median opinion scores) supporting this conclusion.

2. All correction algorithms, with the notable exception of the "darken-only" algorithm ("Correction 1"), result in statistically better quality images than that for the unmodified grid-distorted image ("Correction 0").

3. Algorithms with significant darkening (i.e., "Correction 2" and "Correction 5") produce images that are statistically better than images that are uncorrected (or only darkened), but statistically worse than those produced by "Correction 3" (40/-20/10 sinc, no darkening) and "Correction 4" (40/-20/10 sinc, 20% darkening).

4. "Correction 3" (40/-20/10 sinc, no darkening) and "Correction 4" (40/-20/10 sinc, 20% darkening) are closer and more difficult to compare than the other conclusions drawn thus far. "Correction 3" has a higher median MOS score, but also has a much higher spread of MOS scores. As a result, the statistical significance is also questionable: referring to the *median* opinion scores, "Correction 3" is statistically better than "Correction 4", but if one refers to the *mean* opinion scores, the improvement is not statistically significant[4].

5. "Correction 3" performed poorly for three specific images ("kodim15", "map", and "testim027"), all of which had significant bright areas intersecting the grid (refer to Figure 9.7). These bright areas caused the edge brightening to be ineffective (i.e., clipping occurred) on significant, visible areas of the images. It is worth noting that "Correction 4" performed well on these images, suggesting a dynamic algorithm (based on image brightness) would outperform both.

Based on these observations, we draw the following conclusions about the image-correction algorithms and their effects upon the perceived quality of tiled display images:

[4]Both mean and median opinion scores are very close to the threshold signifying 95% confidence: *median* opinion scores barely satisfy this condition while the confidence intervals for the *mean* opinion scores slightly overlap.

1. Darkening of the image is always[5] undesirable.

2. Edge brightening is always desirable, even if the image needs to be darkened to accommodate this.

   (a) All edge brightening correction algorithms produced images that were statistically better than the unmodified grid-distorted images. In other words, on a given tiled display, any of these correction algorithms provide an improvement to image quality.

   (b) In cases were darkening is required to allow for edge brightening, the amount of darkening should be minimized as much as possible.

3. The "best" image-correction algorithm studied here is either Correction 3 (40/-20/10 sinc with no darkening) or Correction 4 (40/-20/10 sinc, 20% darkening), subject to preference and interpretation:

   (a) If consistency is valued over maximum and average quality, then Correction 4 is the "best" algorithm.

   (b) If average and maximum potential quality are valued over consistency, then Correction 3 is the "best" algorithm.

   (c) If "better quality in the majority of cases" is a priority, Correction 3 is the best algorithm. In direct comparisons, Correction 3 was selected over Correction 4 by a ratio of nearly 2:1.

4. The effectiveness of edge brightening is dependent on the content of a given image. Brightening (and as a result, perceptual correction) is restricted in image areas that already approach the maximum display brightness.

   (a) A dynamic algorithm that determines edge brightening and global darkening based on image parameters will theoretically exceed the performance of all algorithms presented here.

   (b) Further research is required to confirm this and determine details such as optimal edge brightening levels, optimal global darkening amount, maximum allowable pixels clipped, etc.

[5] At least in the case of the environment (i.e., ambient lighting and screen brightness) of our user study.

Conclusions

Chapter 10

Conclusions

Tiled displays are an important, and growing, segment of the display market but one of their largest inherent distortions have been largely un-researched until now.

10.1 Contributions

This dissertation provides four significant contributions to the field of image quality assessment:

1. Creation of two new IQA image databases to provide previously unavailable ground-truth data.
2. Analysis of current objective IQA metrics that demonstrates their poor performance for measuring tiled image quality.
3. Creation of the new tiled display quality metric (TDQM) that significantly outperforms current metrics (when measuring tiled image quality).
4. Creation and verification of four new image-correction algorithms that significantly improve the perceptual quality of tiled images and mitigate the visual effect of the grid distortion. These algorithms are simple and could easily be incorporated into existing tiled display technology.

10.1.1 Future Work

The area of tiled display image quality is a very new one, and there are multiple directions for future research:

- Improvement of TDQM: Though our new metric significantly outperforms current objective metrics, there is still much room for improvement with roughly 40% of subjective score variance unaccounted for.
- Extension of TDQM to a general form: TDQM is currently a single-task objective metric (only for tiled images). There is value in extending the concepts to general-purpose objective metrics to make them more complete.
- Improvement of image-correction algorithms: Our new correction algorithms clearly improve the perceived quality of tiled display images but there are still many steps that can be taken to further improve them:
  - Tuning of edge-brightening and global-darkening tradeoffs, including determination of optimal values for each.
  - Investigation of potential overcorrection at close viewing distances.
  - Development of a dynamic algorithm incorporating the best qualities of the top-performing algorithms.
  - Examine potential improvements offered by independent brightening parameters for red, green, and blue colour components.
- Investigation of other tiled display distortions: The grid distortion was selected because it cannot be removed using current technology, but there is still value in understanding and measuring other distortions inherent to tiled displays. For example, the cost of matching brightness and colour across multiple display tiles could be minimized given an ability to dynamically monitor the quality impact of such mismatches.

APPENDICES

Appendix A

Subjective User Study Details

We performed one informal user study and two formal user studies using similar methodologies (but with a few differences). Our initial user study verified the appropriateness of our methodology by including control data from a widely recognized IQA database (LIVE). The second user study expanded upon the tiled-image results by increasing the study size and replacing the control data with more grid-distorted test cases.

A.1 Informal User Study Details

Our informal user study was loosely modelled after the user study performed for the CSIQ image database. [22] Our goal was not to develop a database with strong statistical reliability. Instead, this database was created to provide a rough sense of the suitability of current metrics for grid distortions. It also served as a platform for us to learn and avoid potential mistakes while running a user study, but it's primary goal was to provide an indication whether further, formal, user studies were of value.

We randomly selected the file "womanhat.bmp" (from the LIVE IQA database) for use as our reference image. There are many potential distortions associated with tiled displays but we chose to focus on the grid distortion (as discussed in Chapter 2. We selected the following grid variations to include in our informal study: grid width, grid frequency, and grid intensity.

Grid Width

Grid width rarely varies in a single display (unless misaligned) but does vary between different displays. Even similar displays can potentially have different widths when deployed in different environments. For example, a rear projection cube array such as MicroTiles can have a typical screen gap of $0.7mm$ or $1.3mm$ depending on the screen selected [11], which in turn is determined by the desired viewing properties for the array. We selected grid widths of 1, 2, and 3 pixels wide for our informal user study. For illustration purposes, these widths would equate to screen gaps of roughly $0.5674mm$, $1.134mm$, and $1.701mm$ (respectively) on a MicroTile array (based on a pixel pitch of $0.567mm$).

Grid Frequency

Like the seam width, the frequency of the seams is also fixed for any given display. This variable is meant to simulate the effects of choosing different tile sizes in an array. For example, a single MicroTile unit has a screen dimension of roughly $16inches \times 12inches$ ($408 \times 306mm$) while a Christie Entero unit can be as large as $63in \times 47in$ ($1600mm \times 1200mm$). There are multiple tradeoffs when determining the size of an arrays individual tiles, and image quality is one of them. We selected arrays of 4×4, 5×5, and 6×6. For illustration, these frequencies would equate to arrays of roughly $5\frac{1/3}{f}t \times 4ft$, $6\frac{2/3}{f}t \times 5ft$, and $8ft \times 6ft$ (respectively) on a MicroTile array.

Grid Intensity

All tiled displays we are aware of use a grid intensity of black, but we do not know of any research to support this decision. For our informal user study we selected three levels of grid intensity: black, grey, and white. This represents the range of (monochromatic) options for grid colour when manufacturing displays.

We used the above variations to distort the reference image and produce 81 (i.e., $3 \times 3 \times 3$) distorted grid images. We computed the SSIM score for each image and selected 7 images that represented a broad distribution of scores to include in our user study.

To provide a baseline (to ensure our testing procedure was valid), we included in our test a set of blur-distorted images (from the LIVE database) with a SSIM score distribution roughly equivalent to that of the grid-distorted images. The SSIM distribution dictated by the grid-distorted images led to a selection of blur-distorted for which non-parametric

correlation (i.e., SRCC) was perfect. The validity of our testing procedure would therefore be determined by the correlation between the blur-distorted images and their SSIM scores (or alternatively, their DMOS scores since these were known).

Figure 5.1 shows an example of the image photographs before sorting by the user.

To allow for portability and to avoid issues such as differing computer displays, we elected to print our images as photographs instead of using the multiple-monitor setup of [22]. This resulted in a tradeoff where the images were less accurate than what could have been displayed on a computer screen, but each user had the advantage of tactile touch in moving images to their desired placement in the sequence. Unlike in [22], we did not consider (or ask the users to consider) distance between images to reflect quality difference. As such, our results were purely non-parametric.

The user study consisted of 2 stages: a training phase and an ordering stage. We used a different reference image for the training images to avoid influencing the user selections; the training images were meant only to acquaint the user with the procedure and provide a rough introduction to the ranges of quality he/she would encounter during the ordering stage. Aside from use of a difference reference, the training phase images were generated using the same procedure as the test images.

A.1.1 Training Stage

With the reference image in the middle, users were instructed to place one random image to the left and one different random image to the right. They were asked to look closely at the reference image (with instruction to consider that as "perfect" quality by definition) and then look at the others and decide which looks "better" with respect to the reference. The two distorted images were then put aside and two new distorted images were placed beside the reference. This procedure was repeated for a total of 6 image pairs (3 blur-distorted pairs and 3 grid-distorted pairs) with no restrictions on pairings (i.e., blur-distorted images could be compared against grid-distorted images).

A.1.2 Ordering Stage

All photos were placed in random order on a large table. Users were provided with the reference image and instructed to place it as one end of the table (either left or right, as preferred by the user). Each user then arranged the other images in order of quality, with the "best" images on one end near the reference image and the "worst" images farther

away from the reference. Extra care was taken to avoid effects of glare from lighting sources when comparing images.

A total of 10 subjects provide subjective scores for this study though 1 subjects results were not considered due to a misunderstanding of the instructions (this led to improvements in the instructions for the subsequent formal user studies).

A.2 Initial Formal User Study [29]

Our initial formal user study used a modified version of the single-stimulus method from the ITU-R BT.500 recommendation. We recruited 27 subjects from undergraduate and graduate engineering programs and showed them a series of images, each of which was given a subjective quality score by every viewer. No visual acuity testing was performed on viewers, with verbal assurance of 20/20 vision accepted from each subject.

A.2.1 Equipment

All images were displayed using a 27" ASUS VG278H LCD monitor set to its native resolution of 1920×1080 and factory default settings (no explicit calibration of the monitor was performed and the 3D capabilities of the monitor were not used). Subjects were seated at a fixed distance (approximately three times the screen height) from the display in a windowless room with typical office lighting.

A.2.2 Images

We used 26 reference images for our initial study: 25 from the Kodak Lossless True Colour Image Suite [14] and one custom image created using OpenStreetMap [1]. Each source image was corrupted by three different grid distortions for a total of 78 grid-distorted images. Each grid distortion had a width of two pixels and a pseudo-random intensity from one of three ranges: black [0,85], grey [86,170], or white [171,255]. The grid width of two pixels was selected to model a gap of roughly 1mm on a tiled display (assuming a dot pitch of roughly 0.5mm).

Of the images used, 20 were also used in the LIVE IQA database [43, 44]. We applied two levels of blur distortion to each of these images (equivalent to the levels applied in the LIVE database) and included these images alongside the grid-distorted images. Each subject evaluated a total of 144 images: 26 source, 78 grid-distorted, and 40 blur-distorted.

A.2.3 Methodology

Each session of our initial user study was divided into three parts:

1. Instruction:
    - Subjects were provided written and verbal instructions for the session.
2. Training:
    - Identical to "Experiment" (see below) except for a shorter duration (i.e., fewer images) and use of different reference images. Subjects were encouraged to ask any questions during this phase.
3. Experiment:
    - We used a methodology similar to that used in the LIVE IQA database, which in turn was based upon methods from the ITU-R BT.500 recommendation [4] for the subjective assessment of television picture quality and the VQEG final reports [2, 3] for validation of objective video quality models. We used a modified single-stimulus (SS) test with references included. Each subject selected a quality score for each image using the slider of a Java application similar to that shown in Fig. A.1. The scores were input on a continuous scale with the following labels: Bad, Poor, Fair, Good, Excellent. Subjects were required to score each image before the next could be displayed. Images were shown in pseudo-random order with the restriction that no consecutive images could share the same reference (source) image. All sessions were completed in under 30 minutes, as recommended in the ITU standard.

A.3 Extended User Study

The equipment and methodology of the expanded user study were identical to the first. All differences were in the recruitment of subjects and the images used.

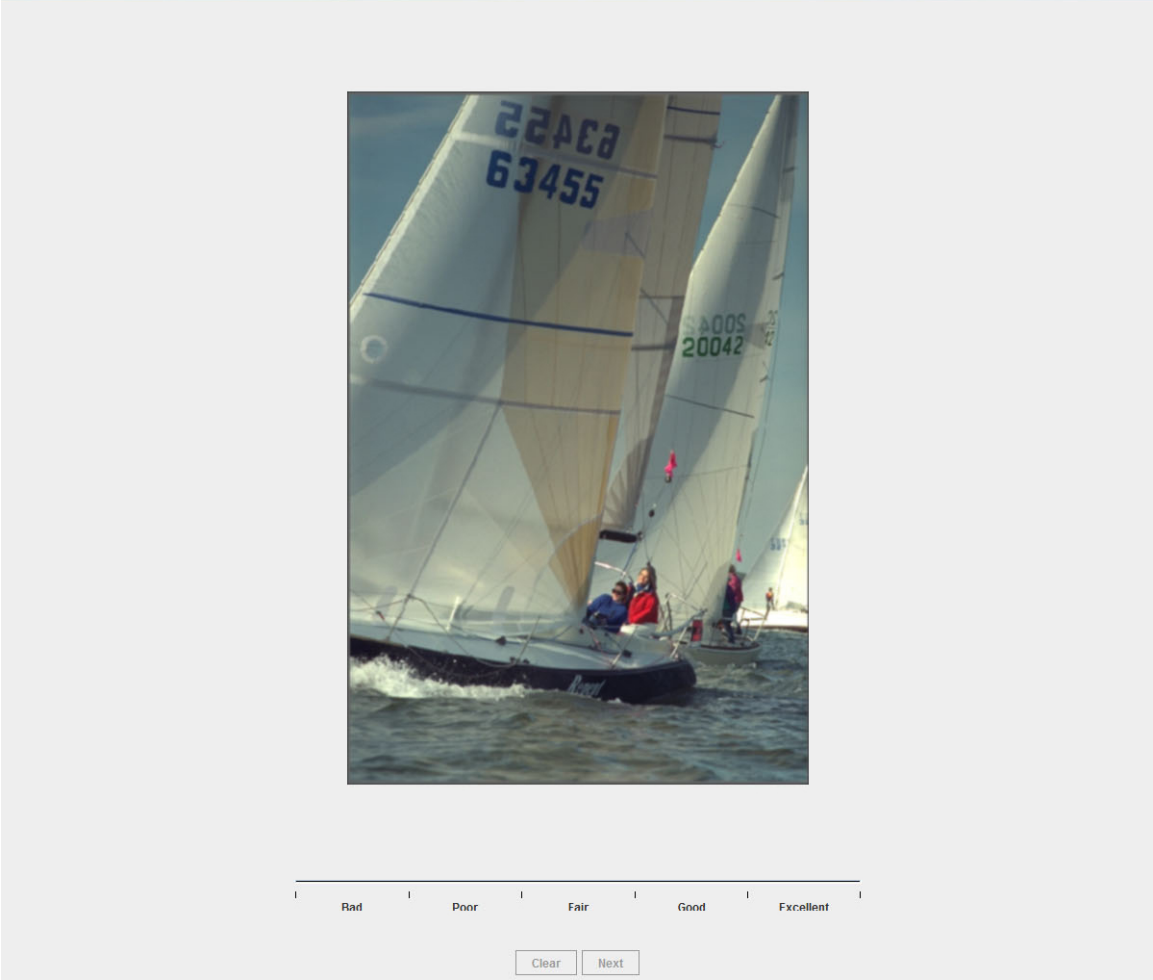


Figure A.1: The interface for the first and second formal user studies. The 'Next' button is shown inactive because the subject must select a score before moving to the next image. There is no explicit identification of unmodified reference (source) images.

A.3.1 Subject Recruitment

Our expanded user study increased the number of viewers from 27 to 33 (an increase of more than 20%). Recruitment was changed to gather volunteers from all programs of our university, rather than only engineering as in the first study. This improved the study by contributing to better gender representation (near-even male/female split) and lowering the abnormally high percentage of "expert viewers" in our sample.

Table A.1: Inter-item correlations for first two formal user studies.

User Study 1: Blur	0.8742
User Study 1: Grid	0.4204
User Study 2	0.3881

A.3.2 Images

We increased the number of reference images from 26 to 34 with the addition of eight new images from the Tecnick Testimages archive [48]. We removed the blur "control data" (used in the first study to correlate results against the LIVE database) to make room for more grid distortions. All reference images were now distorted by five levels of grid distortion: black [0,50], dark-grey [51,101], grey [102,152], light-grey [153,203], and white [204,255]. This gave the expanded user study a total of 204 images evaluated by each viewer (170 grid-distorted and 34 reference); a 40% increase over the first user study. In spite of the increased number of images, all sessions still complied with the 30 minute guideline of the ITU standard.

A.3.3 Internal Consistency

We calculated average inter-item correlations (average correlations for the scores from each possible subject pair) for our two formal user studies and the results are shown in Table A.1. The high value for the blur-distorted images from the first user study verifies the reliability of our study, while the lower values for the tiled images (roughly consistent between the two studies) suggest less agreement among subjects regarding the quality impact of the grid distortion. This further supports our assertion that the grid distortion is unique compared to traditional distortions.

A.4 Image-Correction User Study

Supplemental data for the image-correction user study described in Chapter 9.

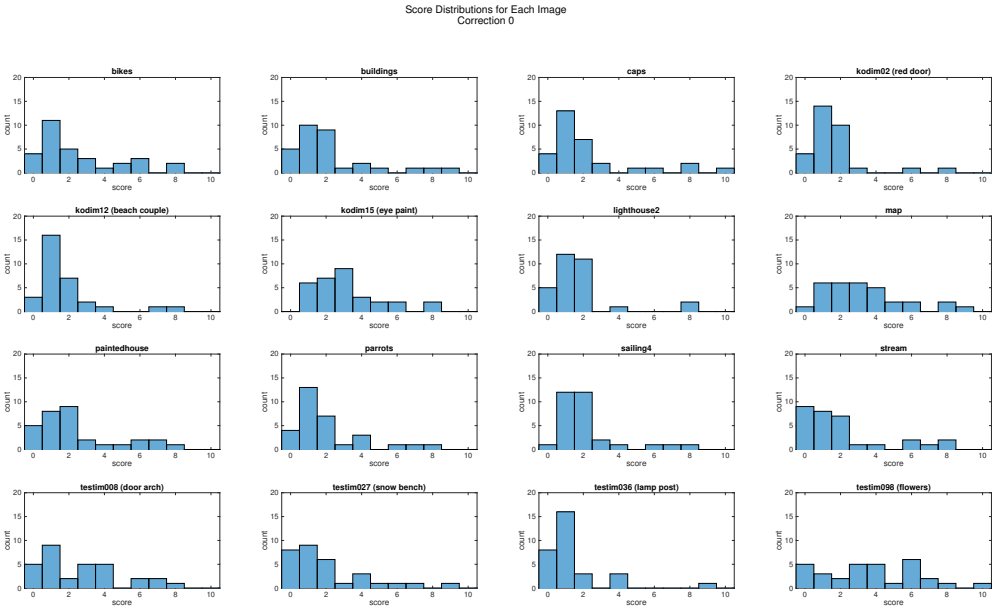


Figure A.2: Detailed score distribution for each image with correction algorithm 0.

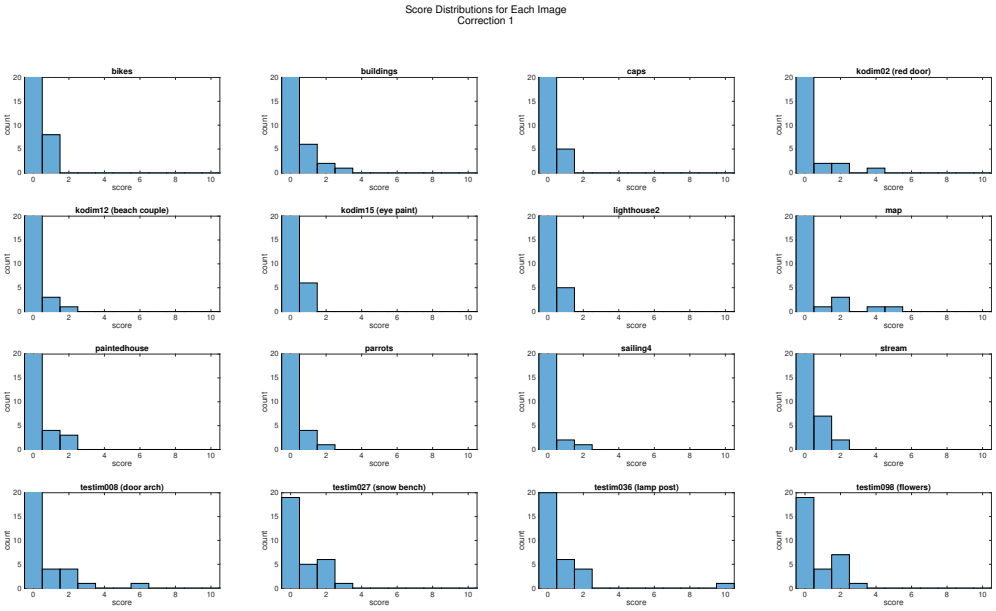


Figure A.3: Detailed score distribution for each image with correction algorithm 1.

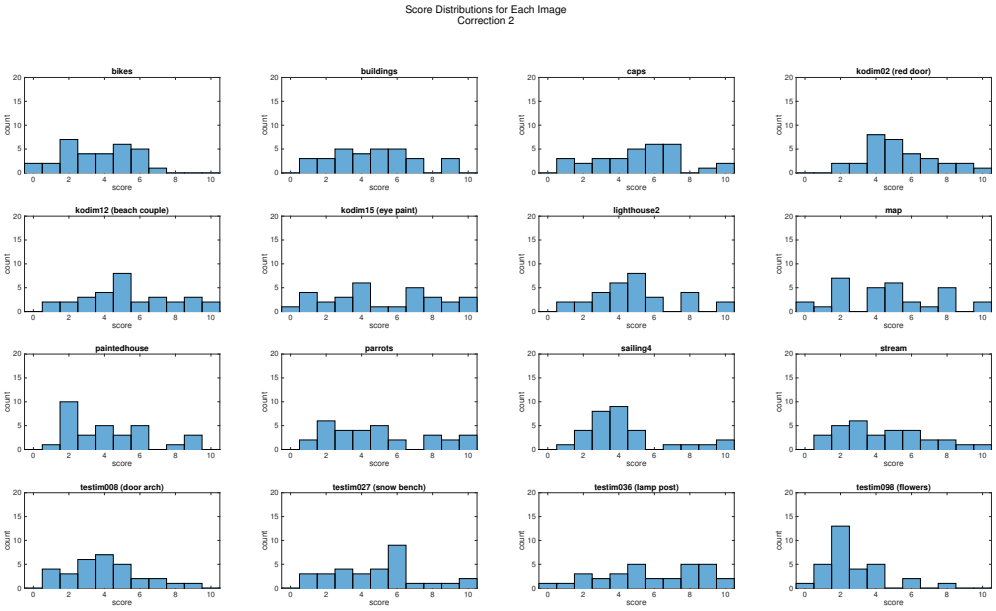


Figure A.4: Detailed score distribution for each image with correction algorithm 2.

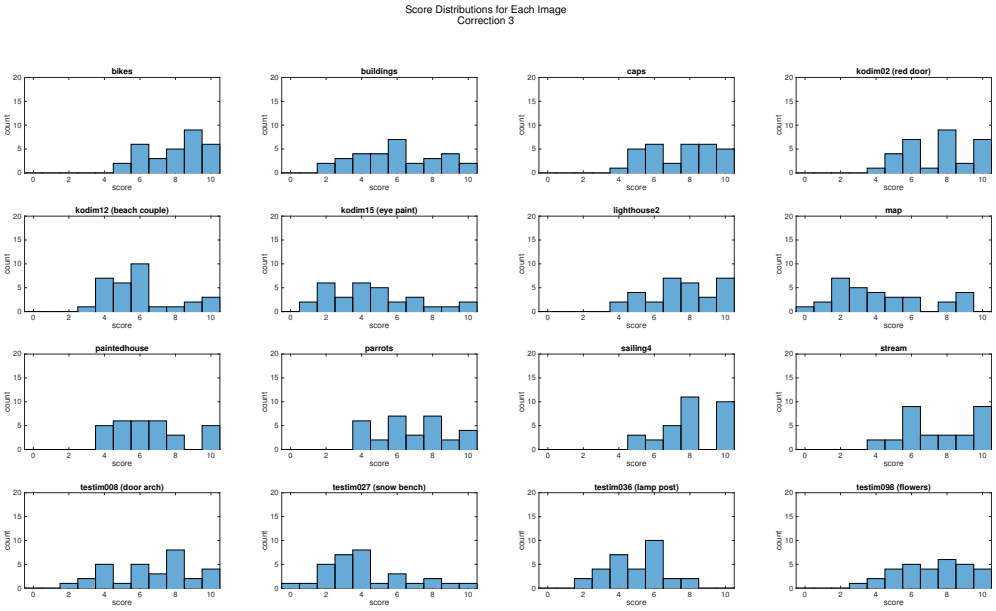


Figure A.5: Detailed score distribution for each image with correction algorithm 3.

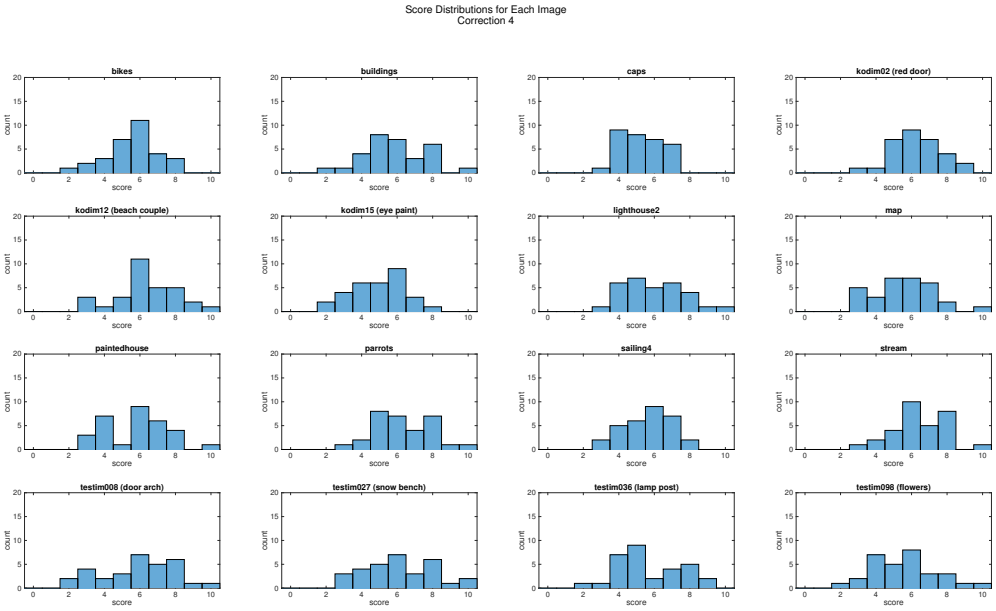


Figure A.6: Detailed score distribution for each image with correction algorithm 4.

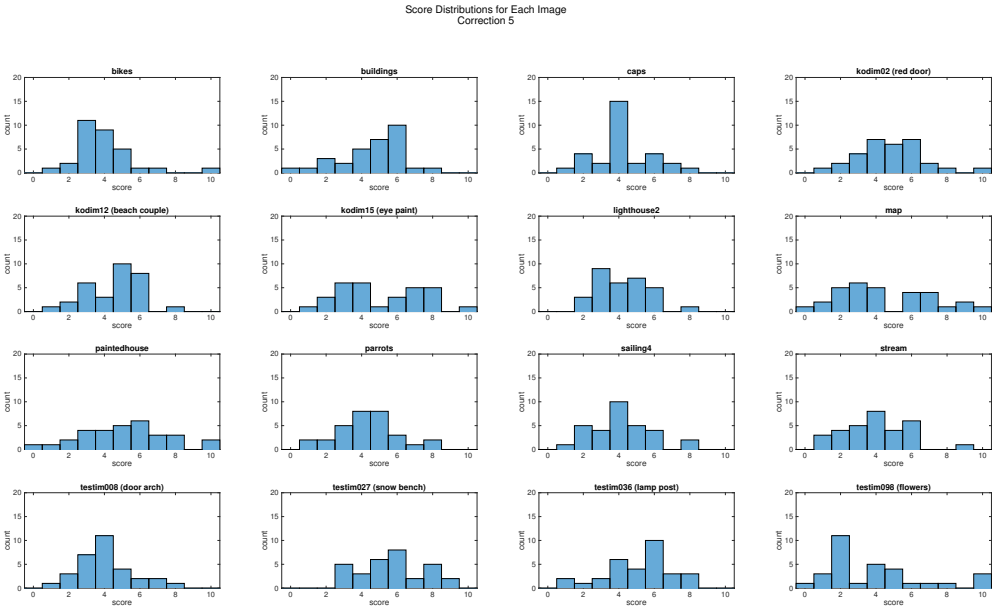


Figure A.7: Detailed score distribution for each image with correction algorithm 5.

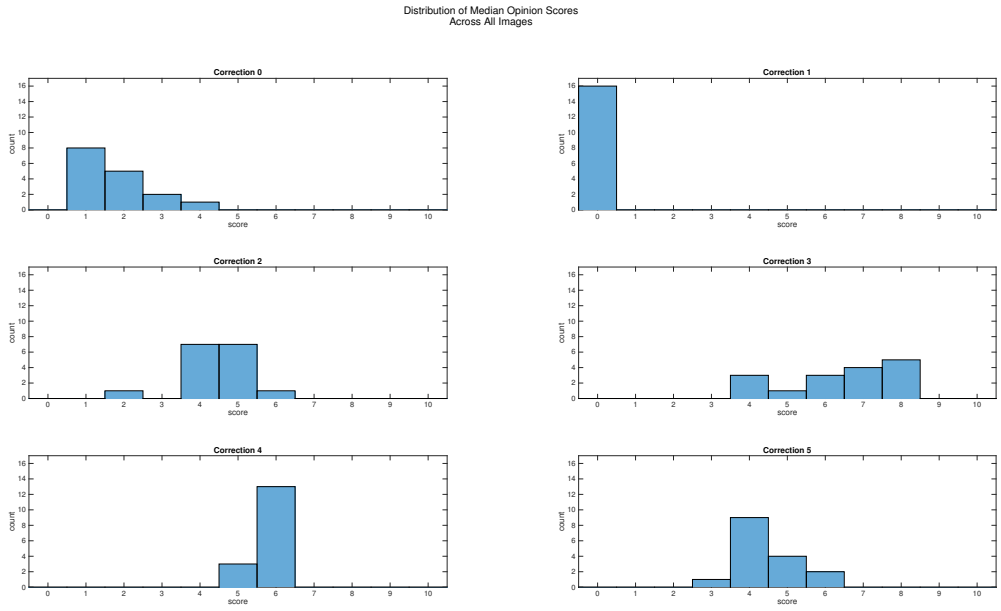


Figure A.8: Distribution of median opinion scores for each correction across all images.

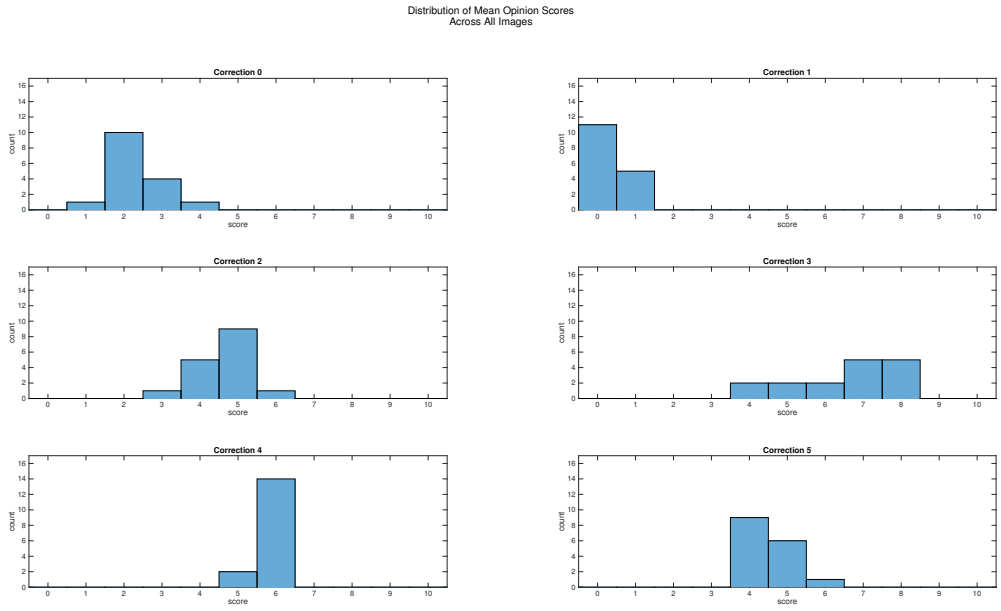


Figure A.9: Distribution of mean opinion scores for each correction across all images.

Appendix B

User Study Data Processing

This section describes details of our methods for processing data in our first two formal user studies. Sections B.1 and B.2 describe our conversion of raw subjective quality scores from the user studies into DMOS scores and Sec. B.3 details our approach to combining IQA performance scores from the separate user studies.

B.1 Raw Data Processing

We calculated DMOS values using the following procedure:

1. Raw scores were converted to raw difference scores

$$d_{ij} = r_{iref}(j) - r_{ij} \tag{B.1}$$

   for each subject i and image j.

2. Raw difference scores were converted to z-scores

$$z_{ij} = \frac{d_{ij} - \overline{d_i}}{\sigma_i} \tag{B.2}$$

   where $\overline{d_i}$ is the mean of difference scores by subject i and σ_i is the standard deviation of raw difference scores by subject i.

3. Z-scores were scaled to the range $[0, 100]$ and averaged across subjects to obtain a DMOS score for each image

$$DMOS_j = \frac{1}{N}\sum_{i=1}^{N} z_{ij} \tag{B.3}$$

where N is the total number of viewers.

B.2 Outlier and Subject Rejection

The largest difference between our methodology and that used in the creation of the LIVE IQA database is how we processed outliers. The LIVE database rejected outliers (which they defined as raw quality scores greater than σ standard deviations from the mean raw quality score) and subjects with "excessive" outlier scores (defined as a subject with greater than R outliers). A minimization algorithm was run to select values of σ and R that minimized the width of the 95% confidence interval for each distortion. We used no subject rejection or outlier removal in our studies because a) the ITU-R BT.500 standard suggests such removal only for studies with fewer than 20 subjects, and b) the optimization method used in the LIVE study's subject rejection is of questionable statistical validity. Our study monitored subjects for inattention (which we considered a valid reason to reject results) but found no cases where rejection was considered justifiable.

B.3 Combining User Study Results

While similar to each other, our first two formal user studies are different enough (e.g., inclusion of blur distortion in first study) that we avoid directly combining their results. We instead evaluate the performances of the individual studies and then combine the correlation results using the following process [50]:

1. Use Fisher's r-to-z transformation to normalize the distributions of each correlation coefficient

$$Z_F = \frac{1}{2} ln\left(\frac{1+r}{1-r}\right) \tag{B.4}$$

for correlation coefficient r.

2. Compute a weighted average based on the sample size of each user study

$$\bar{Z_F} = \frac{(N_1 - 3)Z_{F1} + (N_2 - 3)Z_{F2}}{(N_1 - 3) + (N_2 - 3)} \tag{B.5}$$

where N_x and Z_{Fx} represent the number of samples and normalized correlation coefficient, respectively, for user study x.

3. Reverse the transformation to recover the combined correlation coefficient

$$\bar{r} = \frac{e^{2\bar{Z}} - 1}{e^{2\bar{Z}} + 1} \tag{B.6}$$

where $\bar{Z}$ is the weighted average computed in Eq. B.5.

The resulting combined correlations are then treated as any other IQA performance result.

B.4 Data Processing for Image-Correction User Study

As mentioned in Section A.4, our Image-Correction User Study was based upon the TID2008 IQA database, but we elected to use a Round-Robin Tournament system instead of the Swiss Tournament system used in the TID2008 database. In this section, we compare the two systems and explain our choice of the Round-Robin Tournament system.

B.4.1 Round-Robin Tournament

In a round-robin tournament setup, each image is compared once against every other image for a total of $(N/2) \times (N - 1)$ comparisons.

- Also known as all-play-all.
- Advantage: Every image is directly compared against every other image. No transitive property is assumed between images.
- Advantage: High granularity of results. No "ties". Produces a complete ranking (as opposed to only a unique "winner" and "loser").
- Disadvantage: Many more rounds (i.e., $\mathcal{O}(n^2)$) required to rank the same number of images.

B.4.2 Swiss Tournament

In a swiss-system tournament setup, each image is not compared to every other image. Instead, images are scored based on their "wins" and "losses", and images are compared against other images with similar scores each round. A unique "winner" and "loser" are determined after $\log_2(N)$ rounds.

- Commonly used in chess tournaments to determine a winner.
- Advantage: Non-elimination principle; more rounds than elimination but fewer than round-robin.
- Advantage: Can determine a unique "winner" (and "loser") with $O(N \log_2(N))$ comparisons.
- Disadvantage: Granularity decreases quickly as one moves away from top and bottom rankings. An example of this is illustrated in Figure B.1 where we consider a ranking of 16 images (8 image pairs). The number of rounds is determined by $log_2(16)$ for a total of 4 rounds and 32 image comparisons. While this is sufficient for determining a "winner" (i.e., the "best" image), notice there are many images that are poorly ranked and "tied" with other images.
- Disadvantage: Many images are never directly compared against one another. This assumes not only a transitive property among the quality of the images, but that the differences are large enough to overcome the experimental error of viewer opinions (which is already increased by the point above).
- Disadvantage: Complexity. The "matches" of a swiss tournament are determined dynamically from the results of the previous round (the first round is random). This has the disadvantages of increasing the complexity of the user study interface (increasing the possibility of introducing errors into the study) and reducing the reproducibility of the results (i.e., it becomes more difficult to compare user results when subjects may compare an almost entirely different set of images).

B.4.3 Justification for Choice of Round-Robin Tournament

We selected the round-robin tournament system because we felt its reliability and granularity outweighed the swiss tournament's ability to handle more comparisons. Since our

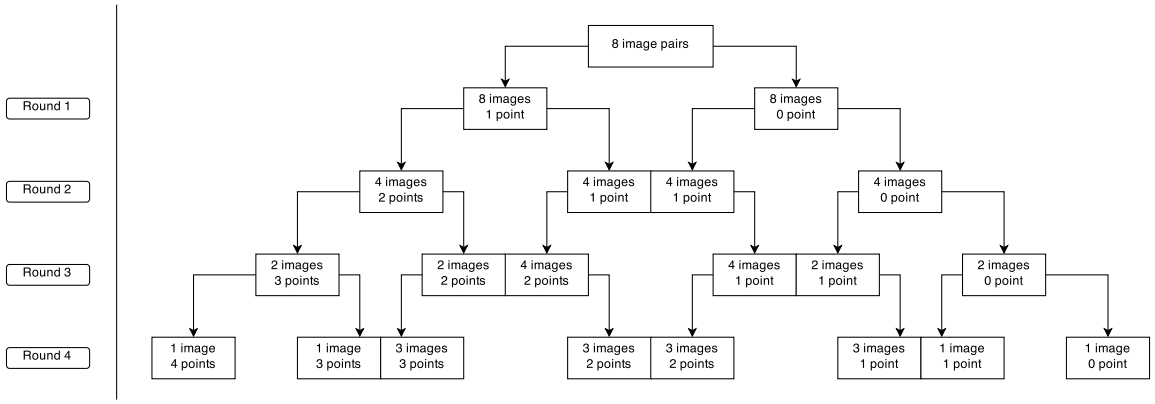


Figure B.1: A simple example of a swiss tournament ranking for 16 images. Note that while only 4 rounds are required to determine the best and worst images, the in-between image rankings are much less defined: 4 images are tied for second place, 6 images are tied for third place, and 4 images are tied for fourth place.

study required many fewer modified images than the TID2008 study, this tradeoff was feasible. The most significant consequence of this choice was the requirement to drop the number of reference images to 16 and restrict our number of correction algorithms to 6 to stay within the recommended 30 minute maximum session times.

References

[1] OpenStreetMap. http://www.openstreetmap.org/.

[2] Final report from the video quality experts group on the validation of objective models of video quality assessment, March 2000. Available: http://www.vqeg.org/.

[3] Final report from the video quality experts group on the validation of objective models of video quality assessment, phase II, Aug. 2003. Available: http://www.vqeg.org/.

[4] Methodology for the subjective assessment of the quality of television pictures, Jan. 2012. ITU-R Rec. BT. 500-13.

[5] Gerard A Alphonse and Jeffrey Lubin. Psychophysical requirements for tiled large-screen displays. In *SPIE/IS&T 1992 Symposium on Electronic Imaging: Science and Technology*, pages 230–240. International Society for Optics and Photonics, 1992.

[6] David A. Atchison and George Smith. *Optics of the human eye*, chapter 18. Butterworth-Heinemann, Boston, 2000.

[7] A.C. Bovik. *Essential guide to image processing.* Academic Press, Amsterdam, 2009.

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

[9] D.M. Chandler and S.S. Hemami. VSNR: A wavelet-based visual signal-to-noise ratio for natural images. *Image Processing, IEEE Transactions on*, 16(9):2284 –2298, sep 2007.

[10] Christie Digital Systems. *Christie Entero HB Series Specification.*

[11] Christie Digital Systems. *Christie MicroTiles Datasheet.*

[12] Scott J. Daly. Visible differences predictor: an algorithm for the assessment of image fidelity. volume 1666, pages 2–15. SPIE, August 1992.

[13] Sachin Deshpande and Scott Daly. Synchronization mismatch: vernier acuity and perception evaluation for large ultra high resolution tiled displays. In *IS&T/SPIE Electronic Imaging*, pages 75270L–75270L. International Society for Optics and Photonics, 2010.

[14] Eastman Kodak Company. Kodak lossless true color image suite. http://r0k.us/graphics/kodak/.

[15] Karen Egiazarian, Jaakko Astola, Nikolay Ponomarenko, Vladimir Lukin, Federica Battisti, and Marco Carli. New full-reference quality metrics based on hvs. In *CD-ROM proceedings of the second international workshop on video processing and quality metrics, Scottsdale, USA*, volume 4, 2006.

[16] Mark Hereld, Ivan R Judson, Joseph Paris, and Rick L Stevens. Developing tiled projection display systems. In *Proc. IPT 2000 (Immersive Projection Technology Workshop)*, 2000.

[17] International Telecommunication Union. Subjective video quality assessment methods for multimedia applications. ITU-T Recommendation P.910, 2008.

[18] Xiaodong Jiang, Jason I Hong, Leila A Takayama, and James A Landay. Ubiquitous computing for firefighters: field studies and prototypes of large displays for incident command. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 679–686. ACM, 2004.

[19] Tom Kimpe. Defective pixels in medical lcd displays: Problem analysis and fundamental solution. *Journal of Digital Imaging*, 19(1):76–84, January 2006.

[20] Tom Kimpe, Stefaan Coulier, and Gert Van Hoey. Human vision-based algorithm to hide defective pixels in lcds. volume 6057, page 60570N. SPIE, February 2006.

[21] Tom Kimpe and Yuri Sneyders. Impact of defective pixels in AMLCDs on the perception of medical images. volume 6146. SPIE, March 2006.

[22] Eric C. Larson and Damon M. Chandler. Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of Electronic Imaging*, 19(1):011006, 2010.

[23] Eric C Larson and Damon M Chandler. Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of Electronic Imaging*, 19(1):011006–011006, 2010.

[24] Patrick Le Callet and Florent Autrusseau. Subjective quality assessment IRCCyN/IVC database, 2005. http://www.irccyn.ec-nantes.fr/ivcdb/.

[25] Jason Leigh, Andrew Johnson, Luc Renambot, Tom Peterka, Byungil Jeong, Daniel J Sandin, Jonas Talandis, Ratko Jagodic, Sungwon Nam, Hyejung Hur, et al. Scalable resolution display walls. *Proceedings of the IEEE*, 101(1):115–129, 2013.

[26] Jeffrey Lubin. The use of psychophysical data and models in the analysis of display system performance. In Andrew B. Watson, editor, *Digital images and human vision*, pages 163–178. MIT Press, Cambridge, MA, USA, 1993.

[27] MATLAB. *version 8.4.0 (R2014b)*. The MathWorks Inc., Natick, Massachusetts, 2014.

[28] Theo Mayer. The 4k format implications for visualization, vr, command & control and special venue application. In *Proceedings of the 2007 workshop on Emerging displays technologies: images and beyond: the future of displays and interacton*, page 9. ACM, 2007.

[29] S.B. McFadden and P.A.S. Ward. A new image quality assessment database for tiled images. In *Image Quality and System Performance XI, Proc. SPIE*, volume 9016, pages 90160X1–90160X10, Feb. 2014.

[30] Steven B McFadden and Paul AS Ward. Selecting the proper window for ssim. In *IS&T/SPIE Electronic Imaging*, pages 82930B–82930B. International Society for Optics and Photonics, 2012.

[31] Steven B McFadden and Paul AS Ward. Towards a new image quality metric for evaluating the effects of tiled displays. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 561–565. IEEE, 2014.

[32] Dean S. Messing and Louis J. Kerofsky. Using optimal rendering to visually mask defective subpixels. volume 6057. SPIE, February 2006.

[33] K T Mullen. The contrast sensitivity of human colour vision to red-green and blue-yellow chromatic gratings. *The Journal of Physiology*, 359(1):381–400, 1985.

[34] National Center for Microscopy and Imaging Research. BioWall displays large images of the brain.

[35] Paul A Navrátil, Brandt Westing, Gregory P Johnson, Ashwini Athalye, Jose Carreno, and Freddy Rojas. A practical guide to large tiled displays. In *Advances in Visual Computing*, pages 970–981. Springer, 2009.

[36] Tao Ni, Greg S Schmidt, Oliver G Staadt, Mark A Livingston, Robert Ball, and Richard May. A survey of large high-resolution display technologies, techniques, and applications. In *Virtual Reality Conference, 2006*, pages 223–236. IEEE, 2006.

[37] Thrasyvoulos N Pappas, Robert J Safranek, and Junqing Chen. Perceptual criteria for image quality evaluation. *Handbook of image and video processing*, pages 669–684, 2000.

[38] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti. TID2008 - a database for evaluation of full-reference visual quality assessment metrics. *Advances of Modern Radioelectronics*, 10:30–45, 2009.

[39] Nikolay Ponomarenko, Flavia Silvestri, Karen Egiazarian, Marco Carli, Jaakko Astola, and Vladimir Lukin. On between-coefficient contrast masking of dct basis functions. In *Proceedings of the Third International Workshop on Video Processing and Quality Metrics*, volume 4, 2007.

[40] R.J. Safranek and J.D. Johnston. A perceptually tuned sub-band image coder with image dependent quantization and post-quantization data compression. In *Acoustics, Speech, and Signal Processing, 1989. ICASSP-89., 1989 International Conference on*, pages 1945 –1948 vol.3, May 1989.

[41] Z. M. Parvez Sazzad, Y. Kawayoke, and Y. Horita. Image quality evaluation database. http://mict.eng.u-toyama.ac.jp/database_toyama.

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

[43] H.R. Sheikh, M.F. Sabir, and A.C. Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on Image Processing*, 15(11):3440 –3451, Nov. 2006.

[44] H.R. Sheikh, Z. Wang, L. Cormack, and A.C. Bovik. LIVE image quality assessment database release 2. http://live.ece.utexas.edu/research/quality/.

[45] Lauren Shupp, Christopher Andrews, Margaret Dickey-Kurdziolek, Beth Yost, and Chris North. Shaping the display of the future: The effects of display size and curvature on user performance and insights. *Human–Computer Interaction*, 24(1-2):230–272, 2009.

[46] Joe Stellbrink. Comparison of vision-based algorithms for hiding defective sub-pixels. volume 6494. SPIE, January 2007.

[47] Richard Taylor. Interpretation of the correlation coefficient: a basic review. *Journal of diagnostic medical sonography*, 6(1):35–39, 1990.

[48] Tecnick.com. Tecnick testimages archive. http://www.tecnick.com.

[49] P.C. Teo and D.J. Heeger. Perceptual image distortion. In *Image Processing, 1994. Proceedings. ICIP-94., IEEE International Conference*, volume 2, pages 982 –986 vol.2, November 1994.

[50] Robert M. Thorndike. *Fisher's Z Transformation*, pages 361–365. Encyclopedia of Measurement and Statistics. SAGE Publications, Inc., 2007.

[51] Z. Wang and A.C. Bovik. Mean squared error: Love it or leave it? A new look at signal fidelity measures. *Signal Processing Magazine, IEEE*, 26(1):98–117, 2009.

[52] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. The SSIM Index for image quality assessment. https://ece.uwaterloo.ca/ z70wang/research/ssim/.

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

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

[55] Z. Wang, E.P. Simoncelli, and A.C. Bovik. Multiscale structural similarity for image quality assessment. In *Proc. IEEE Asilomar Conf. Signals, Syst., Comput.*, volume 2, pages 1398–1402, Nov. 2003.

[56] Zhou Wang and Alan C. Bovik. *Modern image quality assessment*. Morgan & Claypool Publishers, San Rafael, California, 2006.

[57] So Yamaoka, Kai-Uwe Doerr, and Falko Kuester. Visualization of high-resolution image collections on large tiled display walls. *Future Generation Computer Systems*, 27(5):498–505, 2011.

[58] Delia Zsivanov and Michael Perkins. Creative integration of Christie MicroTiles in tiled-display applications. In *SID Symposium Digest of Technical Papers*, volume 42, pages 240–243. Wiley Online Library, 2011.