# THE PERCEIVED USEFULNESS OF DISTORTED NATURAL IMAGES

A Dissertation
Presented to the Faculty of the Graduate School
of Cornell University
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy

by
David Marshall Rouse
January 2011

THE PERCEIVED USEFULNESS OF DISTORTED NATURAL IMAGES

David Marshall Rouse, Ph.D.

Cornell University 2011

Quality estimators aspire to quantify the perceptual resemblance but not the usefulness of a distorted image when compared to a reference natural image. However, humans can successfully accomplish tasks (e.g., object identification) using visibly distorted images that are not necessarily of high quality. This work investigates the usefulness (i.e., utility) of distorted natural images by 1) reporting methods to experimentally measure image perceived utility; 2) introducing and evaluating an objective utility estimator; and 3) generating useful but distorted natural images based on the proposed novel objective utility estimator.

Subjective experiments were conducted to verify the distinction between the quality and utility of distorted natural images. Quality scores were obtained using a standard methodology. Novel experiments were conducted to collect responses from human observers regarding the usefulness of these distorted images, too. The resulting relationship between the utility and quality scores reveals that quality does not accurately predict utility. Distortions to high-frequency signal components of natural images are observed to have the greatest impact on utility. The experiment results demonstrate that a quality evaluation of a distorted image is different from its utility evaluation, so accurate quality estimators cannot accurately estimate utility.

An understanding of the signal characteristics that distinguish utility from quality is obtained by analyzing and dismantling leading quality estimators, since no utility estimators exist. The natural image contour evaluation (NICE) is intro-

duced as a utility estimator. NICE measures contour degradations of a distorted natural image relative to a reference natural image by extracting and comparing the edges from both images. Quality estimators and NICE are assessed as both quality and utility estimators. NICE provides accurate estimates of perceived utility scores and is argued to be compatible with shape-based theories of object perception.

The perceived utility scores from the first set of experiments were found to exhibit limitations, and a novel technique that overcomes these limitations is proposed and implemented. The novel technique collects textual descriptions produced by observers viewing distorted natural images. The technique uses an observer-centric approach, so observers participating in the experiment dictate the relevant concepts that characterize image usefulness. This technique is used to obtain perceived utility scores for two collections of distorted images that simulate scenes captured by a surveillance system. The capability of both NICE and several leading quality estimators to estimate the perceived utility scores is reported. NICE is demonstrated to produce the most accurate estimates of perceived utility scores.

Last, a procedure to generate useful distorted natural images based on NICE is presented. An image independent parametric quantization table that is compatible with baseline JPEG and based on NICE is provided. The quantization table is found by using a genetic algorithm heuristic search to perform rate-distortion optimization using a baseline JPEG encoder and NICE. Rate-distortion optimization using a genetic algorithm is discussed as a tool to analyze other objective estimators.

## BIOGRAPHICAL SKETCH

David Marshall Rouse was born in Waxhaw, North Carolina on March 6, 1981. He received both a Bachelor of Science degree in May 2003 and a Master of Science degree in August 2005 in Electrical Engineering from North Carolina State University in Raleigh, NC. He joined the School of Electrical and Computer Engineering at Cornell University in the fall of 2005, where he received a Master of Science in January 2009.

This work is dedicated to the memory of my most humble grandmother, Carolyn Kirk Hargett Rouse. She always supported and praised my endeavors and accomplishments. I will always remember hearing her speak through her genuine smile to tell me that I was her favorite, youngest grandson.[1] I hope that this work receives her approval.

[1] She was ever so clever with her choice of words, for I was her youngest grandson!

## ACKNOWLEDGEMENTS

Numerous people have worked with me (and taught me) over the last five years, and I would like to recognize those individuals whose contributions made this work possible. To begin, I would like to thank Professor H. Joel Trussell for being a mentor and teaching me the fundamentals of research.

The body of work presented in this thesis would not have seen the light of day without the guidance and support of Professor Sheila Hemami, my academic advisor at Cornell. She refined and enhanced the skills imparted by Professor Trussell. Furthermore, she challenged me to tackle this seemingly unruly research problem, and at this point in time (the end), I truly appreciate the value of establishing, investigating, and communicating a research problem. I am grateful that Professor Hemami introduced me to the image processing community by presenting aspects of this work at conferences and workshops throughout the world.

I am immensely thankful to have collaborated with Professor Patrick Le Callet at the Université de Nantes. Many of the subjective experiments reported in this work were conducted in Professor Le Callet's research laboratory when I researched abroad in Nantes, France during the spring and summer of 2009. I would specifically like to recognize Romuald Pépion of Université de Nantes, who helped me run experiments and collect data.

Yiran Wang and Fan Zhang earned a Master of Engineering by helping me in the last year of my research to design and conduct subjective experiments. I am grateful for their contributions.

By attending conferences and workshops, I met Dr. Amy Reibman of AT&T Labs and Dr. John Handley of Xerox Corporation. Several of the ideas presented in this work emerged from discussions with both Dr. Reibman and Dr. Handley, and I am thankful to have had their help.

Dr. Matthew Gaubatz, Dr. Damon Chandler, and Frank Ciaramello listened to my ludicrous ideas, provided honest helpful feedback, and even endured some of my subjective experiments. I am fortunate to have had their support and phone numbers over the last five years.

I would like to recognize the support of my loving family. In particular, my mother, Cheryl, taught me the benefits of hard work and determination. I appreciate her everlasting support and unconditional love, because this would never have been possible without her support. Thanks Mom!

Finally, I would like to thank my wife, Julia, who helped me recover from research dead-ends on countless occasions. I married her, because she, in addition to her love and support, constantly offers me an unparalleled perspective of the world. Everyday with her is such a pleasure, and I am very fortunate to have married her.

# TABLE OF CONTENTS

Biographical Sketch . . . iii
Dedication . . . iv
Acknowledgements . . . v
Table of Contents . . . vii
List of Tables . . . xi
List of Figures . . . xv

**1 Introduction 1**
1.1 Prior Work . . . 4
1.2 Problem Statement and Approach . . . 7
1.3 Contribution . . . 9
1.4 Outline . . . 10

**2 Background 12**
2.1 Introduction . . . 12
2.2 Image Databases with Subjective Responses . . . 14
2.2.1 Image Databases with Perceived Quality Scores . . . 14
2.2.2 Public Safety Video Quality Database . . . 17
2.3 Objective Estimators of Subjective Scores . . . 18
2.3.1 Amplitude-Spectrum Statistics . . . 18
2.3.2 Full-Reference Image Quality Estimators . . . 19
2.3.3 Estimators Based on Properties of the Human Visual System 21
2.3.4 The Structural Similarity (SSIM) Index . . . 24
2.3.5 The Visual Information Fidelity (VIF) Criterion . . . 41
2.4 Summary . . . 46
2.5 Key Points . . . 47

**3 Subjective Estimates of Perceived Utility and Perceived Quality 48**
3.1 Introduction . . . 48
3.2 Methods: Perceived Utility Scores . . . 49
3.2.1 Reference/Distortion Image Sequences . . . 51
3.2.2 Experiment 1: Subjective Data to Derive Relative Perceived Utility Scores . . . 58
3.2.3 Experiment 2: Recognition Thresholds of Natural Images . . 64
3.2.4 Perceived Utility Scores from Subjective Data . . . 67
3.3 Methods: Perceived Quality Scores . . . 70
3.3.1 Stimuli . . . 71
3.3.2 Procedure . . . 72
3.3.3 Observers . . . 73
3.3.4 Perceived Quality Scores from Subjective Data . . . 73
3.4 Results: Quality is not a proxy for Utility . . . 73
3.4.1 Relationship Between Quality and Utility . . . 74

3.4.2 Effects of Low-frequency Content on Quality and Utility . . 85
3.4.3 Examples illustrating that quality is not a proxy for utility . 92
3.5 Limitations of Perceived Utility Scores . . . . . . . . . . . . . . . 95
3.6 Summary . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 96
3.7 Key Points . . . . . . . . . . . . . . . . . . . . . . . . . . . . 96

**4 Objective Estimates of Perceived Utility and Perceived Quality 98**
4.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . 98
4.2 NICE: Natural Image Contour Evaluation . . . . . . . . . . . . . 99
4.3 Results: Objective Estimates of Utility and Quality . . . . . . . . . 104
4.3.1 Results: Objective Estimates of Perceived Utility . . . . . . 107
4.3.2 Results: Objective Estimates of Perceived Quality . . . . . . 122
4.3.3 Results: Summary . . . . . . . . . . . . . . . . . . . . . . 133
4.4 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . . 134
4.4.1 Objective Estimators Reveal Image Characteristics that Impact Utility and Quality . . . . . . . . . . . . . . . . . . . . . 135
4.4.2 Object Recognition, Perceived Utility, and NICE . . . . . . 143
4.5 Summary . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 146
4.6 Key Points . . . . . . . . . . . . . . . . . . . . . . . . . . . . 147

**5 A Novel Technique to Acquire Perceived Utility Scores from Textual Descriptions of Distorted Natural Images 149**
5.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . 149
5.2 Methods: Measuring the Perceived Usefulness of Distorted Natural Images . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 151
5.2.1 Experiment 1 . . . . . . . . . . . . . . . . . . . . . . . . . 151
5.2.2 Experiment 2 . . . . . . . . . . . . . . . . . . . . . . . . . 156
5.3 Perceived Utility Scores from Textual Descriptions . . . . . . . . . . 162
5.3.1 Concept Vectors . . . . . . . . . . . . . . . . . . . . . . . . 162
5.3.2 Perceived Utility Loss Definition . . . . . . . . . . . . . . . 164
5.3.3 Plausibility of Utility Loss Definition . . . . . . . . . . . . 166
5.4 Analysis of Perceived Utility Scores . . . . . . . . . . . . . . . . . 168
5.4.1 Demonstrating the Need for a Utility Estimator . . . . . . . 168
5.4.2 Evidence that the Proposed Technique Yields Reproducible Results . . . . . . . . . . . . . . . . . . . . . . . . . . . . 169
5.5 Objective Estimates of Perceived Utility Scores . . . . . . . . . . . 174
5.6 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . . 177
5.7 Summary . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 178
5.8 Key Points . . . . . . . . . . . . . . . . . . . . . . . . . . . . 179

**6 Analysis of the Natural Image Contour Evaluation 181**
6.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . 181
6.2 Gradient Analysis . . . . . . . . . . . . . . . . . . . . . . . . . 182
6.2.1 Summary of NICE . . . . . . . . . . . . . . . . . . . . . . 183

6.2.2 Continuous Approximation of NICE . . . . . . . . . . . . . . . 184
6.2.3 Gradient of Continuous Approximation of NICE . . . . . . . 186
6.2.4 Sample Images . . . . . . . . . . . . . . . . . . . . . . . . . . 188
6.3 Using NICE for Rate-Distortion Optimization with JPEG . . . . . 192
6.3.1 Overview of JPEG for Grayscale Images . . . . . . . . . . . 193
6.3.2 The Rate-Distortion Optimization Problem for JPEG . . . . 194
6.3.3 Rate-Distortion Optimization using a Genetic Algorithm . . 196
6.3.4 Results: JPEG Images Optimized using NICE . . . . . . . . 198
6.3.5 JPEG Quantization Table Optimized for NICE . . . . . . . 204
6.4 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 207
6.5 Summary . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 209
6.6 Key Points . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 210

**7 Conclusion 211**

**A Recognition Thresholds of Line Drawings 214**
A.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 214
A.2 Experimental Methods . . . . . . . . . . . . . . . . . . . . . . . 215
A.2.1 Distortions . . . . . . . . . . . . . . . . . . . . . . . . . . . 215
A.2.2 Stimuli . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 218
A.2.3 Procedure . . . . . . . . . . . . . . . . . . . . . . . . . . . 221
A.2.4 Participants . . . . . . . . . . . . . . . . . . . . . . . . . . 221
A.3 Results: Recognition Thresholds . . . . . . . . . . . . . . . . . . 221
A.3.1 Raw Mean Recognition Bitrate . . . . . . . . . . . . . . . . 222
A.3.2 Normalized Mean Recognition Bitrate . . . . . . . . . . . . 225
A.4 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 228
A.5 Summary . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 229

**B Tradeoffs in Subjective Testing Methods for Image and Video Quality Assessment 230**
B.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 230
B.2 Testing Methods: ACR and SAMVIQ . . . . . . . . . . . . . . . 232
B.2.1 Absolute Category Rating (ACR) Method . . . . . . . . . . 232
B.2.2 Subjective Assessment for Video Quality (SAMVIQ) Method 233
B.3 Databases with Perceived Quality Scores for Visual Stimuli . . . . . 234
B.3.1 Still-Image Database . . . . . . . . . . . . . . . . . . . . . . 235
B.3.2 Video 1 database . . . . . . . . . . . . . . . . . . . . . . . . 236
B.3.3 Video 2 database . . . . . . . . . . . . . . . . . . . . . . . . 237
B.4 Analysis of the ACR and SAMVIQ Subjective Testing Methods . . 238
B.4.1 Observers Use of Rating Scale for ACR and SAMVIQ . . . . 238
B.4.2 Comparative Analysis of MOSs from ACR and SAMVIQ . . 240
B.5 Discussion: ACR or SAMVIQ? . . . . . . . . . . . . . . . . . . . 250
B.6 Summary . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 254

**Bibliography** **255**

# LIST OF TABLES

2.1 Statistics summarizing the correlation and accuracy of the objective estimators as quality estimators using the subjective scores from the LIVE database [116]. The objective estimator values were produced by SSIM, SSIM's individual components, SSIM's pairwise components, SSIM*, SSIM*'s individual components, and SSIM*'s pairwise components. The linear correlation $r$, Spearman rank correlation $\rho$, and the Kendall rank correlation $\tau$ are computed between the objective estimates and subjective scores (i.e., DMOSs) after nonlinear regression. The root mean square error (RMSE), outlier ratio (OR), skewness (skew), and kurtosis (kurt) are computed using the residual between the fitted objective scores and the subjective scores. . . . . . . . . . . . . . . . . . . . . . . . . . . 36

2.2 Statistics summarizing the correlation and accuracy of the objective estimators as quality estimators using the subjective scores from the LIVE database [116]. The objective estimator values were produced by MS-SSIM, MS-SSIM's individual components, MS-SSIM's pairwise components, MS-SSIM*, MS-SSIM*'s individual components, and MS-SSIM*'s pairwise components. The linear correlation $r$, Spearman rank correlation $\rho$, and the Kendall rank correlation $\tau$ are computed between the objective estimates and subjective scores (i.e., DMOSs) after nonlinear regression. The root mean square error (RMSE), outlier ratio (OR), skewness (skew), and kurtosis (kurt) are computed using the residual between the fitted objective scores and the subjective scores. . . . . . . . . . . . . . . . . . . . 37

2.3 Linear correlation coefficients between DMOS [116] and $v \times r$ for fixed $\mu_x = \mu_y = \mu$ after nonlinear regression for each artifact type in LIVE image database [111]. . . . . . . . . . . . . . . . . . . . . . . 40

2.4 Statistics summarizing the performance of VIF and VIF* as quality estimators using the LIVE database. . . . . . . . . . . . . . . . . . 46

3.1 Summary of image distortions studied. The relationship between the distortion parameter and the level of distortion is described for each distortion. For a reference image subjected to one distortion type, utility and quality are assumed to exhibit a monotonically, non-decreasing relationship with decreasing distortion level. . . . . 54

3.2 Results summarizing the relationship between perceived quality and perceived utility. Each row corresponds a subset of $n$ images either spanning a particular range of quality or corresponding to a particular distortion. The Pearson linear correlation $r$, the Spearman rank correlation $\rho$, and the Kendall rank correlation $\tau$ are computed between the perceived quality and perceived utility scores. The root mean squared error (RMSE) and the outlier ratio (OR) were computed using the utility scores and the mapped (i.e., Eq. (3.3)) quality scores. $r_{fit}$ denotes the Pearson linear correlation after applying the mapping. For the correlation statistics and OR, boldface values are statistically equivalent to the largest value for a subset of images (excluding All). Boldface RMSE values are statistically larger than the other subsets based on ANOVA. . . . . 80

4.1 Statistics summarizing 1) the ability of an objective estimator to distinguish recognizable and unrecognizable images and 2) the correlation of objective estimator values with perceived utility scores. The recognition detection accuracy is the probability that an unrecognizable image and recognizable image are correctly distinguished. The Pearson (linear) correlation coefficient $r$, the Spearman rank correlation coefficient $\rho$, the Kendall rank correlation $\tau$ are reported when the estimates are compared with the perceived utility scores for test images with perceived utility exceeding $-15$ ($n = 163$ test images). Optimal values appear in boldface with statistically equivalent values italicized. . . . . . . . . . . . . . . . . . . . . . . . . 110

4.2 Statistics summarizing the accuracy of estimators serving as utility estimators. The root mean squared error (RMSE), the outlier ratio (OR), and the resolving power $RP_\alpha$ for $\alpha = 0.1, 0.05$ are reported when the estimates are compared with the perceived utility scores for test images with perceived utility exceeding $-15$ ($n = 163$ test images). Optimal values appear in boldface with statistically equivalent values italicized. An asterisk beside the RMSE indicates that the residual variance is statistically equivalent to that of VIF according to the Brown-Forsythe-Levene test at the 95% confidence level. The skewness (Skew.) and kurtosis (Kurt.) of the residuals are italicized when the Jarque-Bera (JB) test indicates that the residuals belong to a Gaussian distribution. . . . . . . . . . . . . 111

4.3 Statistics summarizing the correlation of objective estimator values with perceived quality scores. The Pearson (linear) correlation coefficient $r$, the Spearman rank correlation coefficient $\rho$, and the Kendall rank correlation $\tau$ are reported when the estimates are compared with the perceived quality scores for all test images ($n = 243$). Optimal values appear in boldface with statistically equivalent values italicized. . . . . . . . . . . . . . . . . . . . . . 123

4.4 Statistics summarizing the accuracy of objective estimators serving as quality estimators. The root mean squared error (RMSE), the outlier ratio (OR), and the resolving power $RP_\alpha$ for $\alpha = 0.1, 0.05$ are reported when the estimates are compared with the perceived quality scores for all test images ($n = 243$). Optimal values appear in boldface with statistically equivalent values italicized. An asterisk beside the RMSE indicates that the residual variance is statistically equivalent to that of VIF* according to the Brown-Forsythe-Levene test at the 95% confidence level. The skewness (Skew.) and kurtosis (Kurt.) of the residuals are italicized when the Jarque-Bera (JB) test indicated that the residuals belong to a Gaussian distribution. . . . . . . . . . . . . . . . . . . . . . . . . . . 124

5.1 Statistics summarizing the performance of estimators serving as utility estimators using the perceived utility (PU) loss scores from the first experiment. The Pearson linear correlation $r$, the Spearman rank correlation $\rho$, the Kendall rank correlation $\tau$, the root mean-squared error (RMSE), outlier ratio (OR), the resolving power $RP_{0.05}$ are reported when the estimates are compared with the PU loss scores. The skewness (skew) and kurtosis (kurt) of the residuals are reported. RMSE values with an asterisk indicate that the accuracy of that estimator is statistically equivalent to that of MS-NICE$_4$ based on the Brown-Forsythe-Levene (BFL) at the 5% significance level. Boldface values are the optimal for that column, and italicized values are statistically equivalent to the optimal value. The correlation measures are compared using their absolute values. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 175

5.2 Statistics summarizing the performance of estimators serving as utility estimators using the perceived utility (PU) loss scores from the second experiment. Refer to Table 5.1 for an explanation of the correlation and accuracy measures. . . . . . . . . . . . . . . . . . 177

6.1 Fitted parameters for the nonlinearity $f_{u,v}(\kappa)$ defined in Eq. (6.23) to produce quantization step-sizes based on MS-NICE$_1$ for a value of $\kappa \in [0, 100]$. Smaller values of $\kappa$ correspond to less distortion as measured by MS-NICE$_1$. . . . . . . . . . . . . . . . . . . . . . . . 208

B.1 Results when comparing the ability of mean opinion scores (MOSs) from different test methods to differentiate test stimuli. For each database and a pair of test methods, the table lists the percentage of pairs of test stimuli corresponding to each of the four outcomes of a comparison test described in Section B.4.2. Considering two test methods, “A” and “B,” there are four outcomes in this comparison for each pair of stimuli: 1) the MOSs for both methods are statistically different 2) the MOSs for both methods are statistically equivalent, 3) the MOSs for method “A” are statistically different but the MOSs for method “B” are statistically equivalent, and 4) the MOSs for method “B” are statistically different but the MOSs for method “A” are statistically equivalent. As an example, MOSs obtained using SAMVIQ differentiate 6.4% pairs of test stimuli in the Still-Image database that are not differentiated by mean opinion scores obtained with ACR-5. The results illustrate that MOSs obtained with SAMVIQ better differentiate stimuli that those obtained with ACR. . . . . . . . . . . . . . . . . . . . . . . . . 247

B.2 Summary of statistical comparisons of mean opinion scores (MOSs) between different test methods for each subjective database. The Pearson linear correlation ($R$), the Spearman rank-order correlation (ROCC), the Kendall $\tau$ rank correlation coefficient ($\tau$), the root mean squared error (RMSE), and the percentage of consistent categories based on the MOSs of two test methods are computed. The RMSE is normalized to the reflect errors in terms of the proportion of the target scale range. MOSs obtained with ACR are consistent with those obtained with SAMVIQ for the Still-Image and Video 1 databases. The stimuli in the Video 2 are more difficult to distinguish in terms of perceived quality, hence the lower consistency between MOSs obtained using the two test methods (cf Figure B.5). . . . . . . . . . . . . . . . . . . . . . . . . . . 249

## LIST OF FIGURES

1.1 The usefulness or utility of a distorted natural image does not necessarily correspond to the perceived quality of that distorted natural image. A reference image and two distorted images are shown in the diagram. Images farther away from the reference image exhibit a decrease in utility and quality. The two distorted images shown in the figure were experimentally determined to have equal utility but different quality. See Chapter 3 for details about the experiment and results. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3

2.1 A full-reference objective estimator analyzes a reference and test (i.e., distorted) image to produce an output value that is consistent with the subjective score provided by human observers viewing the same two images. . . . . . . . . . . . . . . . . . . . . . . . . . . . 13

2.2 Illustration showing the values of the SSIM, the SSIM cross-correlation component $r$, and the SSIM* cross-correlation component $r^*$ as the parameter $\gamma$ is varied for a *airplane*/TS+HPF sequence. Larger values of $\gamma$ induce more texture smoothing. The objective scores are produced by comparing the test image with the *airplane*/TS+HPF image with $\gamma = 1$. The SSIM cross-correlation component $r$ is approximately one when the image is most distorted (i.e., $\gamma = 2048$). The SSIM objective value for this image is approximately 0.5, because the mean and variance components produce objective values less than one, as desired. The SSIM* cross-correlation component $r^*$ objective value is nearly zero for the *airplane*/TS+HPF distorted image with $\gamma = 2048$. Sample images from this sequence are shown in Figure 2.3. . . . . . . . . . 29

2.3 Texture smoothing plus high pass filtering distortions for the image *airplane* for several values of the TS+HPF distortion parameter $\gamma$. See Table 3.1 for more information about the TS+HPF distortions. 30

2.4 Gradient analysis of the individual SSIM components: mean $m(X, Y)$, variance $v(X, Y)$, and cross-correlation $r(X, Y)$. Images (b) – (d) have been rescaled for visibility. . . . . . . . . . . . . . 32

3.1 The original reference *airplane* image and distorted images illustrating the distortions described in Section 3.2.1. The JPEG and BLOCK distortions are introduced by quantizing coefficients of a block-based discrete cosine transform. J2K+DCQ distortions result from quantizing coefficients of a discrete wavelet transform according to the dynamic contrast-based quantization (DCQ) strategy [17]. Table 3.1 contains descriptions of each of the distortions. 53

3.2 Distorted images illustrating the distortions described in Section 3.2.1. Texture smoothing (TS) distortions are induced via total variation (TV) regularization to smooth texture regions with limited disruption to edges. A high-pass filter (HPF) that removes low-frequency signal information from images with TS distortions produces the TS+HPF distortions. Table 3.1 contains descriptions of each of the distortions. . . . . . . . . . . . . . . . . . . . . . . . . 55
3.3 Four of the nine natural images serving as reference images for the experiments. Figure 3.4 contains the four other natural images serving as reference images. The reference *airplane* image is shown in Figure 3.1(a). . . . . . . . . . . . . . . . . . . . . . . . . . . 59
3.4 Four of the nine natural images serving as reference images for the experiments. Figure 3.3 contains the four other natural images serving as reference images. The reference *airplane* image is shown in Figure 3.1(a). . . . . . . . . . . . . . . . . . . . . . . . . . . 60
3.5 Four images from the *airplane*/J2K+DCQ sequence used in Experiment 1 (Section 3.2.2). J2K+DCQ distorted images are parameterized using the encoding bitrate $R$ in bits per pixel (bpp) (see Table 3.1). The encoding bitrate of the visually lossless *airplane* image specified by the DCQ strategy is $R_{VL} = 1.85$ bpp. The perceived utility (U) scores and perceived quality (Q) scores obtained via the subjective experiments are provided for each image. . . . . 62
3.6 Quality is not a suitable proxy for utility. The scatter plots show the relationship between perceived utility scores and the perceived quality scores for nine reference images and five distortions (cf. Figure 3.1). The symbols indicate the reference image corresponding to each subjective score. The recognition threshold (RT) and the reference equivalence class (REC) are denoted on the axis corresponding to perceived utility scores. The quality adjectives are denoted on the axis corresponding to the perceived quality scores. Standard error bars have been included for both subjective scores. In each figure, the fitted nonlinear mapping from the abscissa to the ordinate is denoted by the solid curve, and the 95% prediction interval (PI) for the fitted nonlinear mapping is denoted by the dashed curves. See also Figure 3.7. . . . . . . . . . . . . . . . . . . 77
3.7 Perceived utility versus perceived quality where the symbols indicate the distortion corresponding to each subjective score. See caption of Figure 3.6. . . . . . . . . . . . . . . . . . . . . . . . 78

3.8 Perceived quality either decreases or remains the same when low-frequency content is disrupted (i.e., for TS+HPF distortions relative to TS distortions). The figures show the confidence that the perceived quality (Q) score of the TS distortions are greater than the perceived quality score for TS+HPF distortions with equal $\gamma$ as a function of the perceived quality score of the TS distortions. Confidence values either in the interval $[0, 0.025]$ or $[0.975, 1]$ indicate that the perceived quality scores are statistically different at the 95% confidence level. Confidence values greater than 0.95 (less than 0.05) indicate that the perceived quality score for the TS distorted image is greater than (less than) the perceived quality score for a TS+HPF distorted image at the 95% confidence level. $Q_{\mathrm{TS}(\gamma)}$ denotes the perceived quality $Q$ of a TS distorted image using smoothing parameter $\gamma$, and $Q_{\mathrm{TS+HPF}(\gamma)}$ is similarly defined for a TS+HPF distorted image. See Section 3.4.2 for additional details regarding the confidence analysis and its interpretation. . . . . . . 87

3.9 Disruptions to low-frequency content do not affect the perceived utility of most images. The figures show the confidence that the perceived utility (U) score of the TS distortions are greater than the perceived utility score for TS+HPF distortions with equal $\gamma$ as a function of the perceived utility score of the TS distortions. Confidence values either in the interval $[0, 0.025]$ or $[0.975, 1]$ indicate that the perceived utility scores are statistically different at the 95% confidence level. Confidence values greater than 0.95 (less than 0.05) indicate that the perceived utility score for the TS distorted image is greater than (less than) the perceived utility score for a TS+HPF distorted image at the 95% confidence level. $U_{\mathrm{TS}(\gamma)}$ denotes the perceived utility $U$ of a TS distorted image using smoothing parameter $\gamma$, and $U_{\mathrm{TS+HPF}(\gamma)}$ is similarly defined for a TS+HPF distorted image. Refer to the legend of Figure 3.8. . . . . 88

3.10 Example showing that the *skier* TS distorted image has statistically greater quality than the TS+HPF distorted image with equal $\gamma$ but statistically lower utility. Removing the low-frequency content from *skier* (i.e., the TS+HPF distorted image) introduces "halos" near edges that enhance the visibility of the skier. See also Figures 3.8 and 3.9. . . . . . . . . . . . . . . . . . . . . . . . . . . . . 90

3.11 Differences in perceived quality (Q) do not imply differences in perceived utility (U). In terms of perceived utility, the distorted images in the middle column are statistically equivalent to the distorted images in the left column. However, in terms of perceived quality the distorted images in middle column are statistically equivalent to the distorted images in the right column. . . . . . . . . . . . . 93

4.1 VIF is more sensitive to distortions at finer image scales (i.e., high spatial frequencies) over those at coarser image scales (i.e., low spatial frequencies), whereas VIF* is more sensitive to disruptions to coarser scale content than finer scale content. Figures 4.1(a) and 4.1(b) respectively show the image scale measurements computed byVIF and VIF* for the *airplane* image with J2K+DCQ ($Q = 3.8$, $U = 77$), TS ($Q = 4.0$, $U = 76$), and TS+HPF ($Q = 3.2$, $U = 69$) distortions. These images have statistically equivalent perceived utility, but the perceived quality of the TS+HPF distorted image is statistically smaller than the other two distorted images. The pooled image scale measurements for VIF reflect their similarity in perceived utility but not their differences in perceived quality. The pooled image scale measurements for VIF* reflect their differences in perceived quality not their similarity in perceived utility. . . . . 136

5.1 Six of the ten grayscale natural images serving as reference images for the experiment. The images simulate hypothetical scenes captured by a video surveillance system. In the caption for each image, "L" indicates that the image size is $640 \times 480$ pixels and "S" indicates that the image size is $320 \times 240$ pixels. The number of concepts $C$ associated with each image is provided. The remaining four images are provided in Figure 5.2. . . . . . . . . . . . . . . . . 152

5.2 Four of the ten grayscale natural images serving as reference images for the experiment. The images simulate hypothetical scenes captured by a video surveillance system. In the caption for each image, "L" indicates that the image size is $640 \times 480$ pixels and "S" indicates that the image size is $320 \times 240$ pixels. The number of concepts $C$ associated with each image is provided. The remaining six images are provided in Figure 5.1. . . . . . . . . . . . . . . . . 153

5.3 Six of the ten grayscale natural images serving as reference images for the second experiment. The images simulate hypothetical scenes captured by a video surveillance system. In the caption for each image, "L" indicates that the image size is $640 \times 480$ pixels and "S" indicates that the image size is $320 \times 240$ pixels. The number of concepts $C$ associated with each image is provided. The remaining four images are provided in Figure 5.4. . . . . . . . . . . . . . . . 158

5.4 Four of the ten grayscale natural images serving as reference images for the second experiment. The images simulate hypothetical scenes captured by a video surveillance system. In the caption for each image, "L" indicates that the image size is $640 \times 480$ pixels and "S" indicates that the image size is $320 \times 240$ pixels. The number of concepts $C$ associated with each image is provided. The remaining six images are provided in Figure 5.3. . . . . . . . . . . . . . . . . 159

5.5 The six grayscale natural images serving as novel images for the second experiment. These images act as distractors in the experiment to avoid guessing by the observers. All of the novel images were 640 × 480 pixels in size. . . . . . . . . . . . . . . . . . . . . . . 160

5.6 Given a fixed distortion parameter value (e.g., additive white Gaussian noise with noise power $\sigma_\eta^2$) does not accurately predict the perceived utility (PU) loss across a variety of images. Each bar graph shows the PU loss scores from the *first* experiment as a function of the distortion parameters. . . . . . . . . . . . . . . . . . . . . . . . . 169

5.7 Given a fixed distortion parameter value (e.g., additive white Gaussian noise with noise power $\sigma_\eta^2$) does not accurately predict the perceived utility (PU) loss across a variety of images. Each bar graph shows the PU loss scores from the *first* experiment as a function of the distortion parameters. . . . . . . . . . . . . . . . . . . . . . . . . 170

5.8 Perceived utility (PU) loss scores obtained from each experiment versus the distortion parameter for *conference* and *elevator*. The error bars indicate the 95% confidence intervals. The Pearson linear correlation $r$ is reported when the PU scores from the first experiment with those of the second experiment, interpolated to the distortion parameter values used in the first experiment. . . . . 172

5.9 Perceived utility (PU) loss scores obtained from each experiment versus the distortion parameter for *kitchen*. The error bars indicate the 95% confidence intervals. The Pearson linear correlation $r$ is reported when the PU scores from the first experiment with those of the second experiment, interpolated to the distortion parameter values used in the first experiment. . . . . . . . . . . . . . . . . . . 173

6.1 Images formed by minimizing $\phi(\boldsymbol{y};\boldsymbol{x})$ using conjugate gradient descent. The value of $\text{NICE}_1^c$ for the image formed with respect to the reference image is provided. . . . . . . . . . . . . . . . . . . . . 190

6.2 The reference image *conference* and the initial images $\boldsymbol{y}^0$ used to start the conjugate gradient descent iteration. The value of $\text{NICE}_1^c$ for the initial image with respect to the reference image is provided. 191

6.3 $\text{MS-NICE}_1$ as a function of encoding bitrate (bpp) for the images *einstein* and *figure*. A genetic algorithm was used to search for a quantization table minimizing the cost function in Eq. (6.19) using $\text{MS-NICE}_1$ as the distortion measure. The rate-distortion operating points corresponding to baseline JPEG using the example quantization table provided in the original specification are included in each figure. The rate-distortion operating points corresponding to the NICE optimized quantization table using Eq. (6.23) and the parameters in Table 6.1 are shown in each figure. . . . . . . . . . 198

6.4 Example images for rate-distortion operating points corresponding to the curves shown in Figure 6.3(a). The images in the top row were produced using quantization tables found by minimizing the rate-distortion cost function in Eq. (6.19). The images in the bottom row were produced using the example quantization tables provided with baseline JPEG specification. The images within each column have equal MS-NICE$_1$ distortion values. The images in Figures 6.4(b) and 6.4(c) have equal encoding bitrates. The MS-NICE$_1$ optimized images have encoding bitrates 39% lower than the baseline JPEG images. . . . . . . . . . . . . . . . . . . . . . . . . . . 199

6.5 Example images for rate-distortion operating points corresponding to the curves shown in Figure 6.3(b). The images in the top row were produced using quantization tables found by minimizing the rate-distortion cost function in Eq. (6.19). The images in the bottom row were produced using the example quantization tables provided with baseline JPEG specification. The images within each column have equal MS-NICE$_1$ distortion values. The images in Figures 6.5(b) and 6.5(c) have equal encoding bitrates. The MS-NICE$_1$ optimized images have encoding bitrates 25% lower than the baseline JPEG images. . . . . . . . . . . . . . . . . . . . . . . . . . . 201

6.6 MS-NICE$_4$ as a function of encoding bitrate (bpp) for the *einstein* natural image. A genetic algorithm was used to search for a quantization table minimizing the cost function in Eq. (6.19) using MS-NICE$_4$ as the distortion measure. The rate-distortion operating points corresponding to baseline JPEG using the example quantization table provided in the original specification are included in each figure. . . . . . . . . . . . . . . . . . . . . . . . . . . 202

6.7 Example images for rate-distortion operating points corresponding to the curves shown in Figure 6.6. The images in the top row were produced using quantization tables found by minimizing the rate-distortion cost function in Eq. (6.19). The images in the bottom row were produced using the example quantization tables provided with baseline JPEG specification. The images within each column have equal MS-NICE$_4$ distortion values. The images in Figures 6.7(b) and 6.7(c) have equal encoding bitrates. The MS-NICE$_4$ optimized images have encoding bitrates 34% lower than the baseline JPEG images. . . . . . . . . . . . . . . . . . . . . . . . . . . 203

6.8 The average absolute Spearman rank correlation coefficient (ROCC) between $\{Q_{u,v,k}(\lambda_j)\}_{j=1}^{N_\lambda}$ and $\{\lambda_j\}_{j=1}^{N_\lambda}$ for each DCT coefficient $(u, v)$, $u = 0, \ldots, 7$ and $v = 0, \ldots, 7$, is shown in Figure 6.8(a). The average absolute ROCC was taken across the $K = 8$ natural images. The ROCC are arranged to correspond with the basis functions corresponding to the $8 \times 8$ DCT shown in Figure 6.8(b). Let $(m, n)$ index the pixel at $m^{th}$ row and $n^{th}$ column spatial location in the $M \times N$ block of pixels. The basis function at row $v$ and column $u$ is defined as $\frac{1}{4MN}\cos(\pi(2m+1)v)\cos(\pi(2n+1)u)$ for $m = 0, \ldots, M-1$ and $n = 0, \ldots, N-1$, where $M = N = 8$. . . 206

A.1 Selected images from the *airplane*/line drawing sequence. The bitrate $R$ using the JBIG coder and the parameter $\sigma$ for the Canny edge detector are provided with each image.. . . . . . . . . . . . . . 217

A.2 Selected images from the *airplane*/J2K sequence along with the corresponding encoding bitrate $R$. . . . . . . . . . . . . . . . . . . . 219

A.3 Selected images from the *airplane*/J2K+DCQ sequence along with the corresponding encoding bitrate $R$. . . . . . . . . . . . . . . . . 220

A.4 Mean recognition bitrates for line drawings and natural images with either baseline JPEG-2000 (J2K) distortions or distortion generated using JPEG-2000 with the dynamic contrast-based quantization strategy (J2K+DCQ). Standard error bars have been included in each graph. Acronyms based on the image descriptions identify the nine natural images: *airplane* (A), *backhoe* (B), *boy and cat* (B&C), *caged birds* (CB), *guitarist* (G), *jack-o-lanterns* (J-L), *pianist* (P), *skier* (S), and *train* (T). . . . . . . . . . . . . . . . . . 223

A.5 Visually lossless bitrates $R_{VL}$ determined using the dynamic contrast-based quantization strategy with JPEG-2000 for the nine natural images (Figure A.5(a)). Mean recognition bitrates normalized by the visually lossless bitrate $R_{VL}$ (Figure A.5(b)). Standard error bars have been included in each graph. See Figure A.4 for the images corresponding to the acronyms. . . . . . . . . . . . . . . . 227

B.1 Histogram of raw observer scores from the Still-Image database using ACR5 and SAMVIQ test methods. For SAMVIQ, 20.4% of the observer scores directly correspond to the category adjectives (i.e., “Bad”, “Poor”, etc.), and 40.8% of the observer scores correspond to both the category adjectives and mid-points between categories. 239

B.2 Histogram of raw observer scores from the Video 1 database using ACR11 and SAMVIQ test methods. For ACR-11, 51.7% of the observer scores directly correspond to the category adjectives (i.e., “Bad”, “Poor”, etc.). For SAMVIQ, 23.3% of the observer scores directly correspond to the category adjectives, and 42.9% of the observer scores correspond to both the category adjectives and mid-points between categories. . . . . . . . . . . . . . . . . . . . . . . . 241

B.3 Histogram of raw observer scores from the Video 2 database using ACR5 and SAMVIQ test methods. For SAMVIQ, 18.1% of the observer scores directly correspond to the category adjectives (i.e., “Bad”, “Poor”, etc.), and 44.9% of the observer scores correspond to both the category adjectives and mid-points between categories. 242

B.4 Relationship between the average confidence interval of the mean opinion scores and the number of observers used for different test methods for the three different databases. Increasing the number of observers for ACR generates more accurate mean opinion scores, competitive with those found with fewer observers using SAMVIQ. 245

B.5 Stimuli influence the consistency of mean opinion scores (MOSs) obtained using ACR and SAMVIQ. Each scatterplot illustrates the relationship between MOSs of stimuli for a particular database obtained by using ACR-5 and SAMVIQ. The consistency of the MOSs obtained using ACR and SAMVIQ for the Still-Image and Video 1 databases indicate that the stimuli evaluated were relatively easy to discriminate in terms of perceived quality. The MOSs obtained using ACR and SAMVIQ for the Video 2 demonstrate that the stimuli were more difficult to discriminate in terms of perceived quality. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 251

CHAPTER 1

# INTRODUCTION

Imaging systems that capture, process, compress, transmit, and/or store natural images[1] supply information to humans to permit or to facilitate the performance of a particular task. For instance, people working in the public safety sector (e.g., law enforcement, fire control, and emergency services) use natural imaging systems in real-time scenarios to make immediate decisions on how best to respond to an incident [42, 43]. In another example, investigators not only examine recordings obtained with video surveillance systems but also introduce such recordings as evidence for criminal investigations [12, 28, 91].

Consumer imaging systems (e.g., digital cameras) directly used by human observers to perform a particular task capture a broad class of source content and are vulnerable to a broad class of distortions, including compression and transmission errors. When operating with limited resources (e.g., communication bandwidth or memory storage), such imaging systems can produce visibly distorted natural images. A visibly distorted image could impede a human's ability to perform a task and provoke inappropriate responses, or it could have no impact at all. Understanding the impact of distortions is clearly important to system designers, users, as well as the subjects who may be captured. Poorer task performance implies that the distorted image is less *useful* to a human observer than its undistorted counterpart: the *perceived utility* decreases. The *perceived utility* characterizes the usefulness of a distorted image as a surrogate for a reference (i.e., undistorted) natural image. For such systems and the images generated by them, an objective

[1] *Natural images* are formed using imaging devices that sense the natural environment (e.g., digital cameras, magnetic resonance imaging (MRI), etc.). Computer generated images and other types of synthetic images are not considered natural images.

estimator of perceived utility (i.e., a *utility estimator*) would facilitate current and future system design, optimization, and improvement. Furthermore, that objective estimator could be used to parameterize existing image coders to generate distorted but useful natural images.

Distorted images traditionally have been assessed with regard to how well they represent the natural world perceived by humans. Such an assessment corresponds to the *perceived quality* of the distorted image, and a *perceived quality score* quantifies a human's judgement of the severity of the visible distortion artifacts. To circumvent expensive subjective studies to acquire perceived quality scores, the image processing community has sought a signal processing tool that accurately estimates the perceived quality of a distorted image given a reference image: a *full-reference quality estimator*. A *full-reference* quality estimator would provide an estimate of the perceived quality of a distorted image through an objective analysis of the signals corresponding to both the reference and distorted images. Many full-reference quality estimators have been proposed and tested on a variety of image databases containing distorted images along with perceived quality scores obtained via subjective studies.

At a high-level, the perceived utility of a distorted image is expected to correlate with perceived quality. That is, as the level of distortion within an image is increased both the perceived utility and perceived quality are expected to decrease. Chapter 3 describes subjective experiments that were conducted to examine the relationship between perceived utility and perceived quality for a collection of distorted natural images. The resulting relationship between perceived utility and perceived quality demonstrated that an image's perceived utility does not imply that image's perceived quality and vice versa.

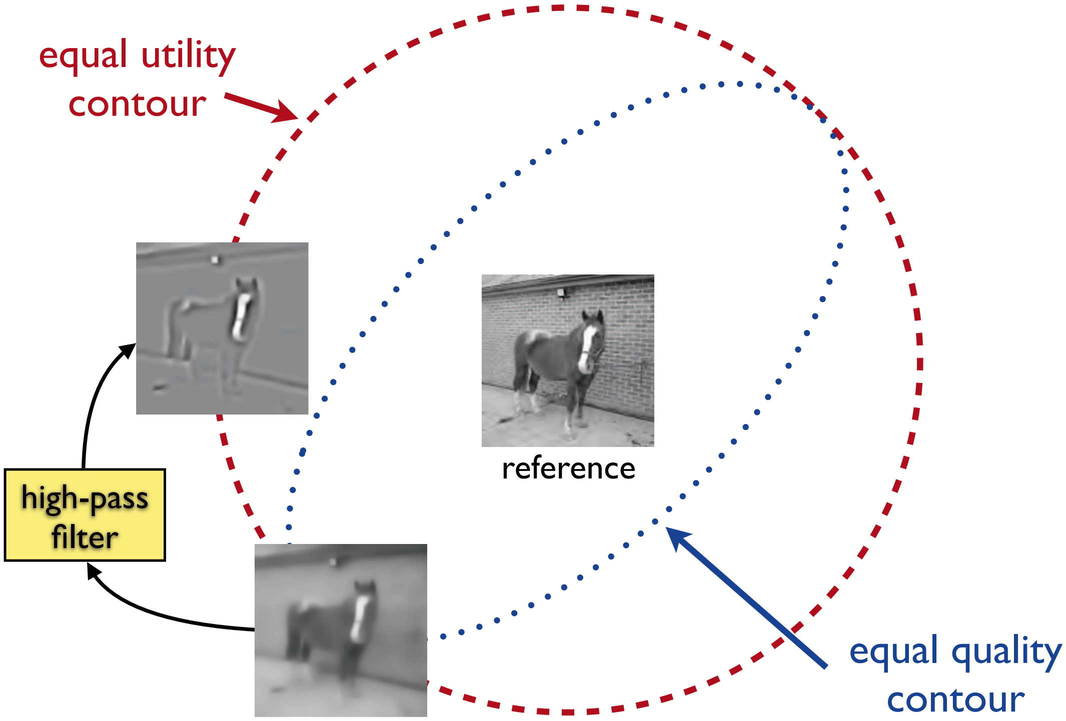


Figure 1.1: The usefulness or utility of a distorted natural image does not necessarily correspond to the perceived quality of that distorted natural image. A reference image and two distorted images are shown in the diagram. Images farther away from the reference image exhibit a decrease in utility and quality. The two distorted images shown in the figure were experimentally determined to have equal utility but different quality. See Chapter 3 for details about the experiment and results.

The diagram in Figure 1.1 illustrates the basic results from the subjective experiment discussed in Chapter 3. Consider the reference image in the center as the origin of a space of images, and images farther from the origin exhibit a decrease in utility and quality. The contours in the diagram represent distorted images with equal utility or equal quality. At the intersection of the two contours is a distorted image that was formed by smoothing (i.e., blurring) the textures in the reference image. This distorted image was filtered with a high-pass filter to produce the other distorted image, which has the same perceived utility but *lower* perceived quality. The loss of low-frequency signal information in an image impacts the quality but not the utility.

The evidence from our subjective experiments indicate that a objective estimator[2] that accurately estimates perceived quality cannot accurately estimate perceived utility. In other words, the problem of estimating perceived utility is different from the problem of estimating perceived quality. This difference motivated the work presented in this dissertation. Namely, this dissertation reports 1) experimental methods to obtain perceived utility data from human observers are developed and implemented, 2) a utility estimator that compares the edges or contours of the reference and distorted images is proposed and evaluated, and 3) a procedure to generate *distorted but useful* image based on the proposed utility estimator is presented and demonstrated.

The remainder of this chapter is organized as follows: Section 1.1 briefly reviews prior work regarding perceived utility and perceived quality. Section 1.2 summarizes the problem and approach used. The contributions of this work are listed in Section 1.3. This chapter concludes with an outline of the remaining chapters in Section 1.4.

## 1.1 Prior Work

Prior work on the perceived utility of natural images can be traced back to Johnson, who quantified task performance in terms of empirically determined sampling criteria for detection, recognition, and identification of a target object [2,58]. The sampling criteria were specified in terms of the number of resolved cycles along the

[2] A quality estimator or a utility estimator is generically referred to as an *objective estimator*.

minimum dimension of the target object and establish the level of object discrimination with respect to the distance of the target object. Johnson's criteria provide basic guidelines for the design of imaging sensors and the expected performance for a given task (i.e., target recognition).

Recent work has investigated alternatives and refinements to Johnson's criteria. For example, recognition of a target has been demonstrated to be equivalent to the detection of an equally sized circular disk, which allows for imaging devices to be characterized in terms of the smallest detectable circular disk [130]. In another example, Vollmerhausen et al. proposed a targeting task performance (TTP) metric that accounts for variations among imaging sensors and computes the integral of the square root of the product of the target contrast, the sensor frequency response, and the contrast sensitivity function of the human visual system [131]. The TTP metric was demonstrated to predict task performance more accurately than Johnson's criteria [131].

The impact of various image compression artifacts on task performance has been investigated. One study investigated the use of uncompressed and compressed synthetic aperture radar imagery captured by an airborne sensor to perform various tasks (e.g., vehicle counting and vehicle classification) and reported the relationship between task performance and the compression ratio [56]. Given the same compression ratio, Irvine et al. observed that wavelet-based compression techniques yield better task performance than standard JPEG compression [56]. Another study conducted a target identification experiment using uncompressed and compressed close-range thermal imagery containing one of a finite number of known targets [81]. O'Shea et al. demonstrate that the TTP metric can be used to predict task performance of compressed imagery using the frequency response of a parameterized Gaussian blur as the sensor frequency response in the TTP metric, where the parameters of the Gaussian blur were selected to fit the experimental results [81].

A fundamental limitation of the prior work on image utility is the use of *a priori* knowledge about the target objects imaged. The experiments conducted to measure task performance train observers to identify specific set of targets that will appear in the test images [81, 131] or prompt observers to perform specific tasks that provide information about the potential content of the image (e.g., vehicle counting) [56]. The models developed in the prior work also incorporate *a priori* knowledge about the target object(s) such as the contrast of the target [81, 131]. Practical use of such *a priori* knowledge in models requires 1) a mechanism that correctly associates known target information with the image under evaluation, which increases the complexity of the model, and 2) a database of target information, which limits the scope of images to which the model can be reliably applied. In short, the results from prior work are tailored to specific applications and provide little insight into the underlying image characteristics that allow human observers to achieve a desired task performance level for a broad class of images.

Over the past three decades, consumer imaging systems have been largely studied in the context of *perceived quality* to characterize the perceptual resemblance of a distorted image to a reference (either known or implied). Perceived quality has been historically framed as a study of “image quality” coupled with a specific human interaction with a constrained class of images [52, 99]. In more recent years, the term perceived quality refers to the perceptual fidelity of a distorted image with respect to an undistorted reference image [84]. However, studies examining the perceived quality of distorted images do not always assume that the reference image available to view (i.e., an ideal reference is implied by the viewer). In the absence, of a specific definition of perceived quality within the community,[3] this work treats perceived quality as an evaluation of distorted images in nonperformance environments, where a human interacts with an image in a recreational setting with a primary concern for aesthetic characteristics of the image (e.g., viewing broadcast television) [52].

Objective estimators of perceived quality have been proposed that are designed according to various principles (e.g., signal fidelity measures or human visual system models), and these estimators are then tuned to or trained on image databases containing distorted images with subjective scores. Such image databases contain distortions typically affecting consumer imaging systems; for example, the LIVE and CSIQ image databases [15, 111] contain images with distortions due to blur, compression, transmission errors, additive noise, and/or global contrast loss. Thus, such estimators are expected to accommodate a broad class of source content and distortions, and various estimators have achieved very good predictive performance of perceived quality for these databases.

## 1.2 Problem Statement and Approach

The work presented in this dissertation is motivated by the prior work in both image quality and utility and expands the previous narrowly studied definitions of utility in a manner that allows both a broader evaluation of utility as well as a characterization of the underlying image characteristics that impact usefulness.

[3] One could argue that the definition of perceived quality is implied in the experimental data a researcher uses to evaluate a candidate quality estimator. However, this approach merely circumvents the definition of perceived quality.

Unlike the specific tasks performed with images in prior work, the "task" is instead to report the content of an image as it is gradually improved from an initially extremely distorted and unrecognizable version to a visually lossless[4] version. No experimental methods are known exist to measure the perceived utility of distorted natural images when the task is to report the content of an image. A novel suite of experiments presented here provides utility scores for distorted images, and quality scores are collected using a standard test methodology. Distortions were strategically selected to disrupt various spatial frequencies in a broader sense than those traditionally studied in perceived quality experiments.

An analysis of the resulting relationship between perceived quality and perceived utility demonstrates that an image's perceived quality does not imply that image's usefulness and vice versa. Therefore, an *objective* estimator that accurately estimates perceived quality scores cannot accurately estimate perceived utility scores and vice versa. These results motivate a thorough analysis of the images to understand the image characteristics that produce *distorted but useful* images for human observers. The performance of several objective estimators as both quality and utility estimators is assessed. Although most of these objective estimators have been designed to estimate perceived quality, they serve as signal analysis tools not only to develop an understanding of those image characteristics that impact usefulness but also to suggest signal analysis tools for an objective utility estimator.

The natural image contour evaluation (NICE) is introduced and analyzed as a novel signal analysis tool that accurately estimates the perceived utility of distorted natural images. NICE was inspired by the importance of contour information to the human visual system for object perception [34, 67, 75]. NICE is based on

[4] A visually lossless image is visually indistinguishable from a reference image.

the hypothesis that degradations to image contours restrict the content that an image conveys to a human and decrease perceived utility. In particular, NICE estimates utility as a function of both *lost* and *introduced* contour information in a distorted image when compared with a reference image. NICE is evaluated using two separate databases of distorted images with perceived utility scores. A method is explored to generate distorted but useful images based on NICE using a baseline JPEG coder.

## 1.3 Contribution

The main contributions of this dissertation are summarized [100–106]:

- The perceived quality of a distorted natural image is not a reliable proxy to characterize the perceived utility of that image. This result implies that any quality estimator that accurately estimates the perceived quality of distorted natural images cannot also accurately estimate the perceived utility of distorted natural images across a variety of distortion artifacts. Thus, existing quality estimators cannot be assumed to reliably estimate perceived utility, rather they must be "tuned," if possible, to produce reliable estimate perceived utility. However, many quality estimators lack parameters that allow them to be tailored to different applications. [100, 105]
- The natural image contour evaluation (NICE) utility estimator is introduced and evaluated as a novel signal analysis tool that compares the edges of a distorted image to the edges of a reference image. NICE is demonstrated to accurately estimate perceived utility scores of distorted natural images. [100, 105]

- A novel technique to acquire perceived utility scores is proposed, used, and demonstrated to provide reproducible estimates of the perceived utility of distorted natural images. The performance of NICE and several quality estimators as utility estimators is evaluated. NICE is concluded to be a reliable utility estimator, robust to variety of distortion artifacts. [106]
- An image independent, parameterized quantization table compatible with a baseline JPEG coder that is based on a rate-distortion optimization with NICE is provided. This quantization table provides a means of forming *distorted but useful* natural images based on NICE. The quantization table was formed by using a genetic algorithm to search the space of quantization tables to minimize NICE for a specified rate-distortion tradeoff.

## 1.4 Outline

This dissertation is organized as follows: Chapter 2 surveys several image databases and several state-of-the-art quality estimators that are been actively used by the contemporary image processing community. Chapter 3 presents the first of two proposed experimental methodologies used to collect perceived utility scores of distorted natural images. In addition, experiments are described that were conducted to collect perceived quality scores for distorted natural images using a standard methodology. The perceived utility and perceived quality scores are compared to illustrate the relationship between the perceived utility and perceived quality.

Chapter 4 introduces the natural image contour evaluation (NICE) utility estimator and reports the performance of NICE as well as several state-of-the-art quality estimators as both utility estimators and quality estimators. Chapter 5 describes the second experimental methodology used to collect perceived utility scores of distorted natural images. This second method is a significant improvement over the first method and is demonstrated to yield reproducible results. An analysis of NICE is presented in Chapter 6, which includes an image independent, parameterized quantization table compatible with a baseline JPEG coder that is based on a rate-distortion optimization with NICE. General conclusions are provided in Chapter 7.

The appendix contains two chapters of related work. Appendix A reports the results from experiments conducted to acquire recognition thresholds for line drawings formed from reference natural images [102]. Appendix B summarizes a comparison of two testing methodologies used to collect perceived quality scores for both images and video sequences [101].

CHAPTER 2

# BACKGROUND

## 2.1 Introduction

The analysis of the usefulness of distorted natural images is motivated by prior work in the perceived quality assessment of distorted natural images. That prior work includes the development and use of testing methodologies to collect subjective responses from human observers corresponding to the perceived quality of distorted natural images as well as the design of objective quality estimators. An objective quality estimator is validated by analyzing its ability to accurately estimate the perceived quality scores obtained from subjective studies.

A generally unspoken conclusion is that an objective quality estimator that accurately estimates perceived quality scores obtained via a particular testing methodology serves as a objective model that maps an image to a subjective score. That is, an objective estimator is limited to the means with which the subjective data has been collected. At present, the methods used to obtain perceived quality scores of natural images are indirect: an observer reports her visual perception using an artificial quality scale. Thus, the observer must map her perceptual state induced by a distorted natural image to that artificial quality scale. Undeniably, the human brain employs an abundance of computational resources (i.e., neurons) to produce a perceived quality score upon viewing a distorted natural image. An objective quality estimator consolidates this computation and, hopefully, imitates this mapping from distorted image to quality score insofar as it produces output consistent with that obtained via the subjective experiments. The objective quality estimator is a theoretical model of the mapping from a distorted image to a

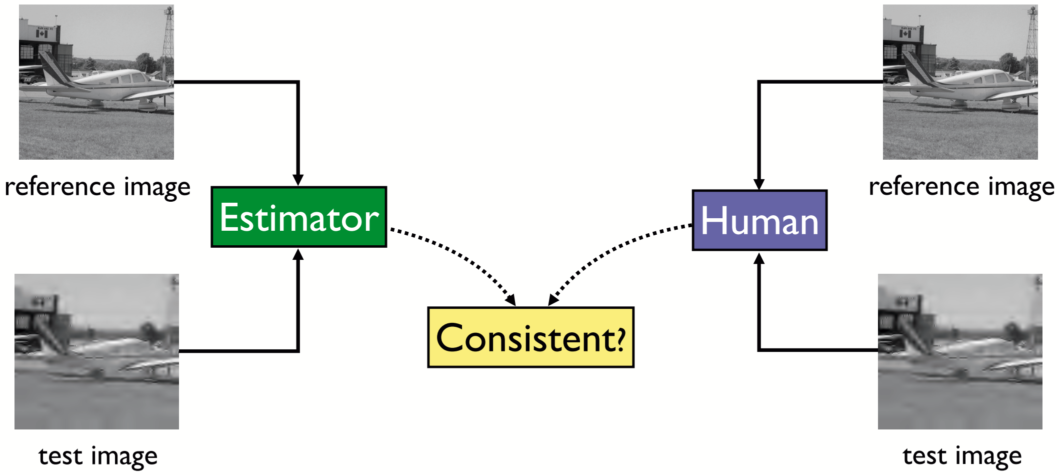


Figure 2.1: A full-reference objective estimator analyzes a reference and test (i.e., distorted) image to produce an output value that is consistent with the subjective score provided by human observers viewing the same two images.

quality score. The diagram in Figure 2.1 illustrates the conventional goal when developing a full-reference objective estimator: produce an output based on the analysis of two images that is consistent with subjective data obtained by humans viewing those images.

A variety of experiments have been conducted to form image databases linking responses from human observers to distorted images to evaluate the performance of objective quality estimators. This chapter reviews several such image databases as well as many state-of-the-art quality estimators that are actively used by members of the image processing community. The review of image databases provides insight with regard to what an objective quality estimator evaluated using that database should do as well as the types of distortions estimator should support. The review of quality estimators shows what theoretical models have been proposed to map a distorted natural image to a quality score. The "better" quality estimators tend to leverage models and/or theories of the human visual system.

## 2.2 Image Databases with Subjective Responses

A first step toward producing an objective estimator that can be used in lieu of a human observer to evaluate a distorted image is to collect responses from human observers who view a collection of distorted images. Several image databases that associate scores provided by human observers to distorted images have been formed. This section briefly reviews several image databases that contain perceived quality scores for distorted images and a video database that contains data corresponding to a human observers ability to perform a specific task with a video sequence.

### 2.2.1 Image Databases with Perceived Quality Scores

Many image databases have been created to associate perceived quality scores obtained from human observers with distorted images. This section briefly reviews a few image databases to survey the various test methods used to acquire perceived quality scores, since there is no unanimously accepted method. An investigation of the differences and similarities between perceived quality scores collected for the same stimuli using two common test methods is provided in Appendix B.

**LIVE Database**

The LIVE image database is a large collection of distorted images for which perceived quality scores have been recorded [111]. This database is frequently used as the de facto database to assess the performance of quality estimators. The database contains 29 reference 24-bits/pixel color images and 779 distorted images, which were formed from the reference images. Five types of processing artifacts were used to form the distorted images: 1) JPEG-2000 compression, 2) JPEG compression, 3) additive white Gaussian noise, 4) Gaussian blurring, and 5) simulated bitstream errors of a JPEG-2000 compressed bitstream in a fast-fading channel.

The perceived quality scores were collected to accommodate two types of quality estimators: full-reference and no-reference[1] [114]. The original difference mean opinion scores (DMOSs) of the LIVE database were obtained using a single-stimulus testing methodology, where observers rated both processed and reference images [112,116]. The reference images were hidden from the observers and used to calibrate the ratings of the processed images. The original DMOS were generated by subtracting the mean opinion scores (MOS) for the processed images from the MOS for the reference images.

A subsequent realignment experiment was conducted to calibrate observer response scales across testing sessions and processing artifacts [116]. To minimize fatigue without limiting the number of processed images evaluated, the original observer responses were collected over many testing sessions. Collecting observer responses over multiple sessions could lead to variations in observer ratings. For example, a rating of 25 in one session may not be equivalent to a rating of 25 in another session. To resolve this potential inconsistency, the original DMOS were realigned using observer responses acquired using the double stimulus continuous quality evaluation (DSCQE) test method for a subset of the distorted images in the LIVE database. For the DSCQS protocol, observers are sequentially presented with the reference and distorted images two consecutive times. The order by which the reference and distorted images are presented is random and hidden from the observer. Observers provide opinion scores for each image during the second pre-

[1] No-reference quality estimators provide quality estimates using only on the processed image

sentation of the pair of reference and processed images.

### A57 and CSIQ Databases

The A57 image database is a small collection of distorted images for which perceived quality scores have been recorded [16]. The database contains 3 reference 8-bit grayscale images and 54 distorted images, which were formed from the reference images. Six types of processing artifacts were used to form the distorted images: 1) JPEG compression; 2) additive white Gaussian noise; 3) Gaussian blurring; 4) baseline JPEG-2000 compression; 5) JPEG-2000 compression with the DCQ algorithm [20]; and 6) uniform quantization of LH subbands of a 5-level discrete wavelet transform to affect different target RMS distortion contrasts.

The A57 database reports perceived quality scores as difference mean opinion scores (DMOS). The testing method used to collect the subjective fidelity ratings is similar to the SAMVIQ protocol, which is described in detail in Appendix B. In particular, an observer indicated his opinion of the perceived quality of a distorted image relative to a reference image affixed to a table by placing the distorted image on the table such that the distance between the distorted image and the reference image reflects the perceived difference in quality. Several images were arranged on the table by the observer to produce scores naturally aligned across different distortion artifacts and reference images. Refer to [21] for additional details about the experiment.

The continuous scale image quality (CSIQ) database [15] uses a test method similar to that used for the A57 database to collect perceived quality scores. The CSIQ database contains 30 reference images and 1500 distorted images, formed from the reference images. Six types of processing artifacts were used to form the distorted images: 1) JPEG compression; 2) JPEG-2000 compression; 3) global contrast scaling; 4) Gaussian blurring; 5) additive white Gaussian noise; and 6) additive pink Gaussian noise. The perceived quality scores were formed from the responses of 35 different observers and are reported as DMOSs.

#### IVC Database

The Image Video Communication (IVC) image database [65] contains 10 reference images and 235 distorted images, which were formed from the reference images. Four types of distortion artifacts were used to form the distorted images: 1) JPEG compression; 2) JPEG-2000 compression; 3) Gaussian blur; and 4) locally adaptive resolution coding. Observer responses were collected using a double stimulus impairment scale with five impairment categories. This method consists of consecutive trials, where in each trial, an observer views the reference image followed by the distorted image and provides an impairment rating for the distorted image. Unlike a quality rating scale, which has an observer provide an opinion of perceived quality in terms of adjectives such as "Excellent," "Good," "Fair," "Poor," and "Bad," an impairment scale used to rate the distorted image uses the following adjectives: "Imperceptible," "Not annoying," "Same as reference," "Annoying," and "Very annoying." Perceived quality scores are reported as mean opinion scores.

### 2.2.2 Public Safety Video Quality Database

A public safety video quality (PSVQ) database was recently created to determine the ability of an observer to perform specific recognition tasks with distorted video sequences [43,55]. The design of the database was inspired by concepts established by the Johnson criteria [58], which was developed to characterize images in terms of an observers ability to detect, recognize, and identify targets in an image. The PSVQ database contains video sequences simulating surveillance applications and recorded video applications. Observers included law enforcement first responders and profession video analysts. The reference video sequences were standard definition video sequences formed from high-definition video sequences. The distorted video sequences simulated compression artifacts and packet loss artifacts.

Subjective responses were collected by performing two types of tasks. For one task, observers were asked to identify, for example, the object being held by a subject in a video sequence. A multiple-choice method was used to collect responses from observers. In the other task, observers were asked to report a sequence of characters in the video such as the license plate on a car. For both tasks, observers were provided with the task prior to watching a video sequence.

## 2.3 Objective Estimators of Subjective Scores

This section reviews several signal analysis tools that are designed to provide meaningful estimates of subjective scores of natural images: 1) amplitude-spectrum statistics of natural images and 2) natural image quality estimators.

### 2.3.1 Amplitude-Spectrum Statistics

A well-known characteristic of natural scenes is the relationship between the spatial frequency and the amplitude of the spatial frequency component [39]. This characterisitic is mathematical specified as $A(f) = f^{-\beta}$, where $\beta$ defines the *spec-*

*tral slope* of an image. Natural images have been reported to have spectral slope values near 1.2 on average [39, 86].

Human performance on visual discrimination tasks has demonstrated a decrease when the spectral slope of the test stimuli are artificially increased or decreased [86]. Such results motivate the use of the spectral slope as an indicator of perceived utility as a natural image is increasingly distorted. In this paper, the spectral slope $\beta$ of a test image is evaluated as a means to estimate subjective scores.

### 2.3.2 Full-Reference Image Quality Estimators

Accurately estimating the perceived quality of distorted natural images remains an open research problem, and current quality estimators could produce accurate estimates of the perceived utility scores of distorted natural images. Accordingly, full-reference quality estimators are treated as mathematical formulas and, in particular, signal analysis tools that quantify the comparison of a distorted image to a reference image. This section reviews several state-of-the-art full-reference quality estimators.

*Full-reference* quality estimators use both an explicit, external reference image $\boldsymbol{X}$ and the test image $\hat{\boldsymbol{X}}$ to estimate the subjective score of the test image. The full-reference quality estimators evaluated in this paper can be categorized as 1) conventional signal fidelity measures, 2) estimators based on properties of the HVS, and 3) estimators derived from hypothetical high-level HVS objectives. The structural similarity (SSIM) index and the visual information fidelity (VIF) criterion are quality estimators that are derived from hypothetical high-level HVS objectives. These two quality estimators are discussed at length in Sections 2.3.4

and 2.3.5, respectively.

### Conventional Signal Fidelity Measures

Mean-square error (MSE), which is used to compute the peak signal-to-noise ratio (PSNR), and root mean squared (RMS) distortion contrast provide computationally simple evaluations of signal fidelity. These measures evaluate fidelity solely in terms of the overall energy of the distortions. Root mean squared (RMS) distortion contrast $C_{rms}(\boldsymbol{E})$ measures fidelity based on the visibility of the distortions $\boldsymbol{E} = \hat{\boldsymbol{X}} - \boldsymbol{X}$ when comparing the images on a particular display device [94] and is given by

$$C_{rms}(\boldsymbol{E}) = \frac{1}{\mu_{L(\boldsymbol{X})}} \left[ \frac{1}{M} \sum_{i=1}^{M} \left( L(E_i + \mu_{\boldsymbol{X}}) - \mu_{L(\boldsymbol{E}+\mu_{\boldsymbol{X}})} \right)^2 \right]^{1/2}, \tag{2.1}$$

where $\mu_{L(\boldsymbol{X})}$ denotes the average luminance of the reference image $\boldsymbol{X}$, $L(E_i + \mu_{\boldsymbol{X}})$ denotes the luminance of the $i^{th}$ pixel of $\boldsymbol{E} + \mu_{\boldsymbol{X}}$, $\mu_{L(\boldsymbol{E}+\mu_{\boldsymbol{X}})}$ denotes the average luminance of the mean shifted distortions $\boldsymbol{E} + \mu_{\boldsymbol{X}}$, and $M$ is the total number of pixels. Eq. (2.1) normalizes the standard deviation of the luminance values $\boldsymbol{E} + \mu_X$ according to the mean luminance of $\boldsymbol{X}$. This normalization accounts for Weber's Law, which asserts that distortions of equal energy are more difficult to detect in brighter regions of an image than in darker image regions. Various other signal fidelity measures have been analyzed with regard to their performance to estimate perceived quality [1, 36].

### 2.3.3 Estimators Based on Properties of the Human Visual System

Several quality estimators capitalize on models and principles characterizing low-level HVS properties such as contrast sensitivity [31], contrast masking [17, 31, 66], and perceived contrast [5, 45]. These properties model the detection of a visual target (e.g., the distortions in an image) under a variety of conditions based on the contrast of the distortions. Many quality estimators have been proposed [14, 20, 26, 27, 47, 68, 71, 84, 87, 108, 113, 121, 125, 137–139, 142]. This section first summarizes the models of HVS properties that have been incorporated into quality estimators. Then, a subset of quality estimators, representing a variety of approaches to quality estimation, are summarized.

#### Models of HVS Properties

Quality estimators based on properties of the HVS use some measure of contrast. The conventional measure of contrast for a visual pattern is the Michelson peak-to-peak contrast given as

$$C_{Michelson} = \frac{L_{max} - L_{min}}{L_{max} + L_{min}}, \tag{2.2}$$

where $L_{max}$ and $L_{min}$ denote the respective maximum and minimum luminance values [31]. The Michelson contrast measure predicts human sensitivities to periodic visual patterns such as sine-wave gratings but fails for more complex visual patterns found in natural scenes [74].

The conventional measure of contrast varies with respect to the frequency of the sine-wave grating [31]. The contrast sensitivity function (CSF) has been derived from psychophysical experiments measuring the detection of targets as sine-wave gratings, and, hence, identifies sensitivity thresholds as a function of spatial frequency [31]. The contrast sensitivity is defined as the inverse of the physical contrast of the target when the target is at the threshold of visual detection (i.e., just visible). For a human to visually detect the target, the contrast of the target must exceed the contrast detection threshold, which is the inverse of the contrast sensitivity.

The contrast sensitivity function varies when a stimulus contains two overlapping sine-wave gratings at different frequencies. In this case, one grating is viewed as a target to be detected that is masked by the other grating. Models of contrast masking account for the variation of the contrast sensitivity due to interactions between a target and background (mask). A psychophysical experiment [66] has been conducted to investigate the effect of a masking sine-wave grating on the detectability of the target sine-wave grating when the stimuli (i.e., mask and target) are simultaneously presented. That study [66] reported that contrast thresholds increased for all spatial frequencies for high contrast masks and decreased for very low contrast masks with spatial frequencies near the target's spatial frequency. That is, targets presented against a high-contrast mask are more difficult to detect than when presented against a low-contrast mask.

When targets are suprathreshold (i.e., visible to a human observer), models account for an observer's perceived contrast. Contrast-matching experiments have been conducted that present two stimuli side-by-side, and an observer adjusts the contrast of one stimuli to match the apparent contrast of the other stimuli. Such contrast-matching experiments, using sine-wave gratings as stimuli, have revealed that as the target contrast becomes increasingly suprathreshold, the perceived contrast depends less on spatial frequency than predicted by the CSF, a result termed *contrast constancy* [5, 45]. Estimators incorporating perceived contrast adapt the CSF according to the extent that the distortions are suprathreshold.

Natural images do not resemble the sine-wave gratings used in the aforementioned experiments. Contrast threshold, contrast masking, and contrast matching experiments have been conducted to investigate the detectability of wavelet subband quantization distortions when masked by natural images [17]. The results of these experiments have been used to develop multi-scale models based on the experiment results [18]. For equal levels of distortion contrast, distributing the distortions according to the results from the contrast matching experiments for wavelet subband quantization distortions generated lower quality images than distributing the distortions according to the CSF [18]. The phenomenal appearance of the images generated by distributing distortion contrast according to image scales was consistent with the principle of global precedence [18, 76], which contends that the HVS processes a visual scene in a global-to-local order. Distortions that disrupt global precedence demonstrate a greater impact on visual quality than distortions that are spatially uncorrelated with the image [21].

### Estimators based on HVS Properties

Two quality estimators, the weighted signal-to-noise ratio (WSNR) and noise quality measure (NQM), evaluate images by incorporating HVS properties to simulate the appearance of the reference and test images to a human and compute the signal-to-noise ratio (SNR) as a function of the difference of the simulated images [27]. WSNR generates the simulated images through filtering with the contrast sensitivity function (CSF) [31]. NQM produces the simulated images through nonlinear processing based on Peli's contrast pyramid [89]. NQM's processing model accounts for the HVS properties of contrast sensitivity, contrast masking, and suprathreshold contrast perception.

Another quality estimator, the visual signal-to-noise ratio (VSNR), evaluates images according to a contrast model accounting for low-level HVS properties and the mid-level HVS property of global precedence [20, 76]. VSNR incorporates models [17] for low-level HVS properties based on experiments investigating the contrast of wavelet subband quantization distortions in natural images rather than sine-wave gratings. To evaluate visual quality, VSNR first assesses the visibility of the distortions. For subthreshold distortions, the algorithm evaluates the test image as having perfect visual quality. For suprathreshold distortions, the VSNR visual quality evaluation accounts for the HVS properties of perceived contrast and global precedence [18, 76].

The last quality estimator examined that incorporates properties of the HVS, criterion 4 (C4), assesses images using elaborate models of several processing areas of the visual cortex [14]. The models in C4 describe color vision; frequency-orientation analysis; contour detection; perceptual and localization of patterns; object discrimination; and visual memory.

### 2.3.4 The Structural Similarity (SSIM) Index

Full-reference quality estimators evaluate a test image $\hat{X}$ with respect to a reference image $X$ to quantify the visual similarity of the test image from the reference image. A challenge for quality estimators is to generate evaluations consistent with human observer opinions across a variety of image artifacts [8].

The structural similarity (SSIM) [137] index and its multi-scale extension (MS-SSIM) [138] estimate visual quality based on the premise that the human visual system (HVS) has evolved to process structural information from natural images, and, hence, a high-quality image is one whose structure closely matches that of the original. To this end, SSIM employs a modified measure of spatial correlation between the pixels of the reference and test images to quantify the degradation of an image's structure. MS-SSIM extends SSIM through a multi-scale implementation of this modified spatial correlation measure.

SSIM estimates perceptual quality using three spatially local evaluations: mean, variance, and cross-correlation. Despite its simple mathematical form, SSIM objectively predicts subjective scores as well as more sophisticated quality estimators [20, 113]. Furthermore, SSIM's simplicity has been investigated by researchers investigating how the HVS evaluates quality [8].

The three SSIM components are examined with regard to how they contribute to its quality estimation for common image artifacts. A gradient analysis illustrates the value of the SSIM cross-correlation component over the other two components. The performance of individual components and pairwise component products in estimating visual quality is assessed using the LIVE image database [111]. The objective estimates using the product of the variance and cross-correlation components match those of the complete SSIM and MS-SSIM evaluations. A computationally simple alternative to SSIM (cf. Eq. (2.13)) that ignores the mean component and sets the local average patch values to 128 exhibits a 1% decrease in linear correlation with subjective ratings to 0.934 from the complete SSIM evaluation with an over 20% reduction in the number of multiplications.

The remainder of this section has the following organization: the SSIM and

MS-SSIM index as well as the modified versions SSIM* and MS-SSIM*, which ignore the constant terms in the original specifications of SSIM and MS-SSIM are reviewed. Then, a gradient analysis of the SSIM components is presented. The results of individual and combinations of SSIM and MS-SSIM components used to estimate subjective scores of perceptual quality are reported and followed by an analysis and discussion. This section ends with a summary.

### SSIM and MS-SSIM

This section first presents the mathematical specification of SSIM and MS-SSIM. The second half of this section mathematically specifies the proposed modifications to both SSIM and MS-SSIM, denoted SSIM* and MS-SSIM*.

***Mathematical Specification of SSIM and MS-SSIM*** SSIM estimates visual quality with a similarity measure between two patches $x$ and $y$ as the product of three components: mean $m(x, y)$, variance $v(x, y)$, and cross-correlation $r(x, y)$. The two patches, $x$ and $y$, correspond to the same spatial window of the images $X$ and $Y$, respectively. The SSIM value for the patches $x$ and $y$ is given as

$$\begin{aligned}\mathrm{SSIM}(x, y) &= m(x, y)^{\alpha} \times v(x, y)^{\beta} \times c(x, y)^{\gamma} \\ &= \left(\frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}\right)^{\alpha} \times \left(\frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}\right)^{\beta} \times \left(\frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}\right)^{\gamma} \\ &= m \times v \times r \qquad (2.3)\end{aligned}$$

where $\mu_x$ denotes the mean of $x$, $\sigma_x$ denotes the standard deviation of $x$, $\sigma_{xy}$ is the cross-correlation (inner product) of the mean shifted images $x - \mu_x$ and $y - \mu_y$, and the $C_i$ for $i = 1, 2, 3$ are small positive constants. These constants combat stability issues when either $(\mu_x^2 + \mu_y^2)$ or $(\sigma_x^2 + \sigma_y^2)$ is close to zero. The positive exponents $\alpha$, $\beta$, and $\gamma$ allow adjustments to the respective component's contribution to the overall

SSIM value. The original specification for SSIM[2], set $C_3 = \frac{C_2}{2}$ and $\alpha = \beta = \gamma = 1$, which simplifies Eq. (2.3) to

$$\begin{aligned} \text{SSIM}(x, y) &= \left(\frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}\right) \times \left(\frac{2\sigma_{xy} + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}\right) \\ &= (m) \times (v \times r). \end{aligned} \tag{2.4}$$

The overall SSIM image quality index for the images $X$ and $Y$ is computed by averaging the SSIM values computed for small patches of the two images. The SSIM value is computed with $\alpha = \beta = \gamma = 1$ and after downsampling the images $X$ and $Y$ by 2 in both spatial directions [137].

MS-SSIM extends SSIM by computing the variance and cross-correlation components at $K$ image scales, where the $k^{th}$ scale image corresponds to low-pass filtering and subsampling, by a factor of 2 in both spatial directions, the original image $(k-1)$ times. The mean component is only computed at the coarsest scale, $K$. The MS-SSIM index is given by

$$\text{MS-SSIM} = m_K(X, Y)^{\alpha_K} \prod_{k=1}^{K} v_k(X, Y)^{\beta_k} r_k(X, Y)^{\gamma_k}, \tag{2.5}$$

where $m_k(X, Y)$, $v_k(X, Y)$, and $r_k(X, Y)$ respectively correspond to the mean, variance, and cross-correlation component computed and pooled across patches from scale $k$ with $k = 1$ as the full-resolution image. The exponents $\alpha_K$, $\{\beta_k\}_{k=1}^{K}$, and $\{\gamma_k\}_{k=1}^{K}$ vary according to $k$ and adjust the contribution of the components based on experimental results by Wang et al. [138] that examined perceptual image quality across scales for distortions with equal mean-squared error (MSE). The exponents are nonnegative and normalized to sum-to-one across scale (i.e. $\sum_{k=1}^{K} \beta_k = 1$). The exponents obtained from the experiment by Wang et al. [138] are $\alpha_K = 0.1333$,

[2] A Gaussian weighting function is used to compute $\mu_x$, $\mu_y$, $\sigma_x$, $\sigma_y$ and $\sigma_{xy}$ [137]. For example, $\mu_x = \sum_{j=1}^{n} w_j x_j$, where $w_j$ are weights corresponding to a circular-symmetric Gaussian function with $\sum_{j=1}^{n} w_j = 1$ and $x_j$ denotes the $j^{th}$ pixel in the patch $x$.

$\beta_1 = 0.0448$, $\beta_2 = 0.2856$, $\beta_3 = 0.3001$, $\beta_4 = 0.2363$, and $\beta_5 = 0.1333$ with $\beta_k = \gamma_k$ for $k = 1, 2, \cdots, K$.

***Mathematical Specification of SSIM* and MS-SSIM**** The constants[3] $C_i$, for $i = 1, 2, 3$, in Eq. (2.3) were introduced to counteract stability issues when either $(\mu_x^2 + \mu_y^2)$, $(\sigma_x^2 + \sigma_y^2)$, or $(\sigma_x \sigma_y)$ are very close to zero [137] When $(\sigma_{xy} \ll C_3)$ and $(\sigma_x \sigma_y \ll C_3)$, both the SSIM cross-correlation component and MS-SSIM cross-correlation component are approximately one (see Figure 2.2). This behavior occurred with very distorted images, for example, when computing the SSIM cross-correlation component of the MS-SSIM cross-correlation component between the *airplane*/TS+HPF distortion image for $\tau = 2048$, $Y$, (cf. Figure 2.3(d)) and the original, $X$, (cf. Figure 3.1(a)), since the *airplane*/TS+HPF distortion image for $\tau = 2048$ is a constant valued image (i.e., $\sigma_y = 0$). This casts doubt upon the significance of the SSIM and MS-SSIM cross-correlation component values to accurately assess the structure of images. Thus, alternative versions of SSIM and MS-SSIM, henceforth respectively identified as SSIM* and MS-SSIM*, are proposed where the positive constants $C_i$ in each component have been set to zero.

The component definitions for SSIM* and MS-SSIM* follow from straightforward consideration of the scenarios leading to the stability concerns addressed in the preceding paragraph. Suppose the constants $C_i$ have been set to zero. When both patches $x$ and $y$ have average pixel values of zero, the mean component is set to one, since the patches have identical mean values. Thus, the alternative mean component definition is given by

$$m^*(x, y) = \begin{cases} 1 & \mu_x^2 + \mu_y^2 = 0 \\ m(x, y) & \text{else} \end{cases}, \tag{2.6}$$

[3] These constants were reportedly "hand-optimized" to fit the data from the LIVE database [113].

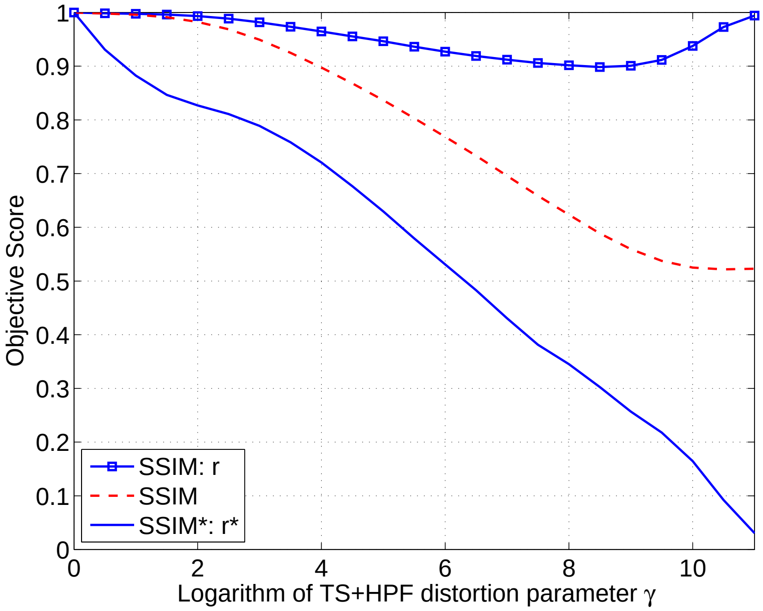
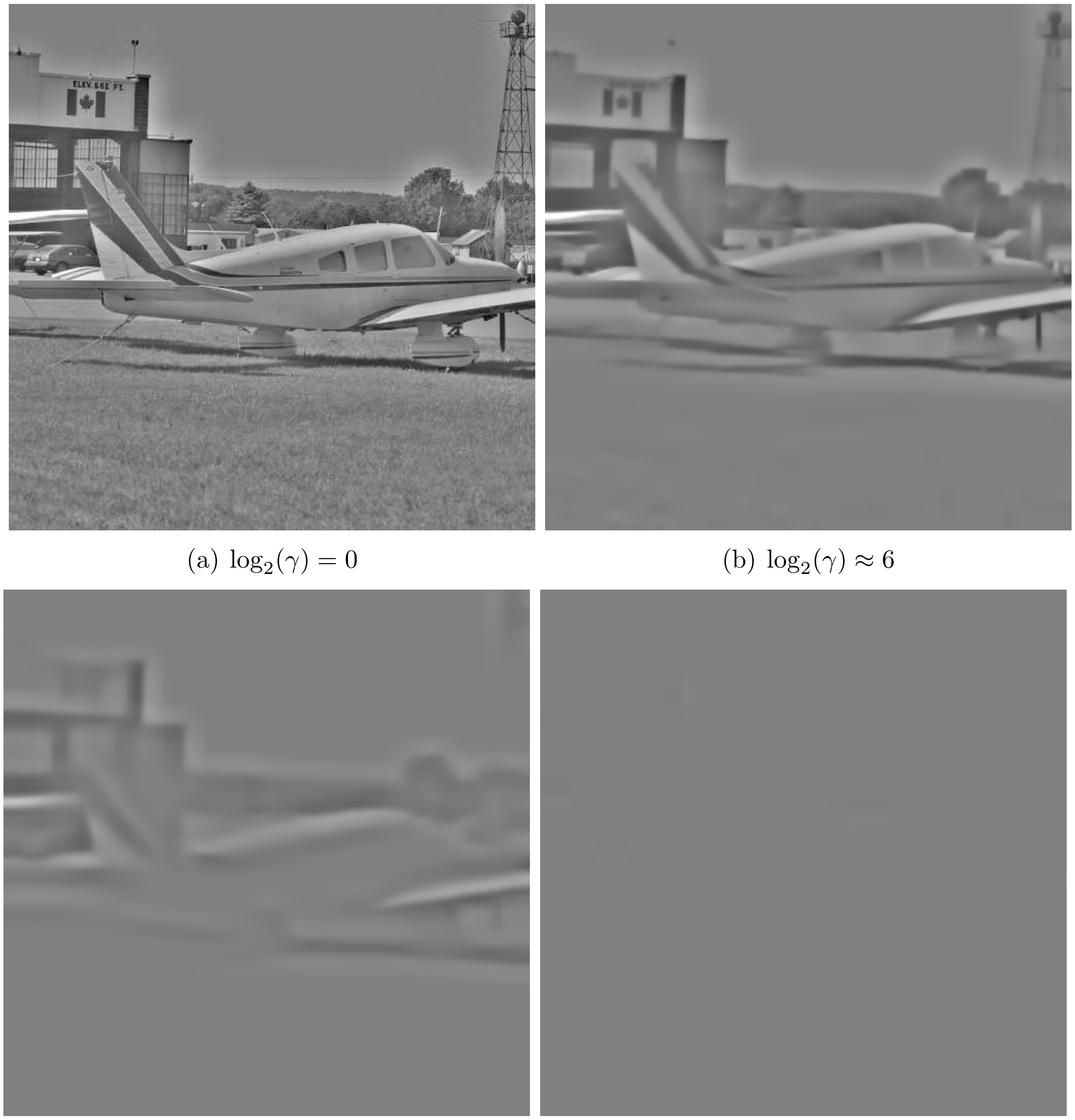


Figure 2.2: Illustration showing the values of the SSIM, the SSIM cross-correlation component $r$, and the SSIM* cross-correlation component $r^*$ as the parameter $\gamma$ is varied for a *airplane*/TS+HPF sequence. Larger values of $\gamma$ induce more texture smoothing. The objective scores are produced by comparing the test image with the *airplane*/TS+HPF image with $\gamma = 1$. The SSIM cross-correlation component $r$ is approximately one when the image is most distorted (i.e., $\gamma = 2048$). The SSIM objective value for this image is approximately 0.5, because the mean and variance components produce objective values less than one, as desired. The SSIM* cross-correlation component $r^*$ objective value is nearly zero for the *airplane*/TS+HPF distorted image with $\gamma = 2048$. Sample images from this sequence are shown in Figure 2.3.

for $m(x, y)$ as defined in Eq. (2.3) with $C_1 = 0$. Similarly, when both patches have variance zero, the variance component is set to one, since the patches have identical variances. The alternative variance component is given by

$$v^*(x, y) = \begin{cases} 1 & \sigma_x^2 + \sigma_y^2 = 0 \\ v(x, y) & \text{else} \end{cases}, \tag{2.7}$$

for $v(x, y)$ as defined in Eq. (2.3) with $C_2 = 0$. Now, suppose that $\sigma_x > 0$, and the patch $y$ is constant. Then, the variance of the patch $y$ is zero. Under this scenario,



(a) $\log_2(\gamma) = 0$

(b) $\log_2(\gamma) \approx 6$

(c) $\log_2(\gamma) \approx 8.7$

(d) $\log_2(\gamma) = 11$

Figure 2.3: Texture smoothing plus high pass filtering distortions for the image *airplane* for several values of the TS+HPF distortion parameter $\gamma$. See Table 3.1 for more information about the TS+HPF distortions.

$y$ does not correlate with $x$, so the cross-correlation component must be set to zero. When both patches have equal variance and $C_3 = 0$, the cross-correlation component must be set to one. The alternative cross-correlation component is given as

$$r^*(x,y) = \begin{cases} 0 & \sigma_x > \sigma_y = 0 \text{ or } \sigma_y > \sigma_x = 0 \\ 1 & \sigma_x = \sigma_y = 0 \\ r(x,y) & \text{else} \end{cases} . \tag{2.8}$$

for $r(x,y)$ as defined in Eq. (2.3) with $C_3 = 0$.

Combining Eqs. (2.6)-(2.8), SSIM* is given as

$$\text{SSIM}^*(X,Y) = m^*(X,Y)^{\alpha} \times v^*(X,Y)^{\beta} \times r^*(X,Y)^{\gamma}, \tag{2.9}$$

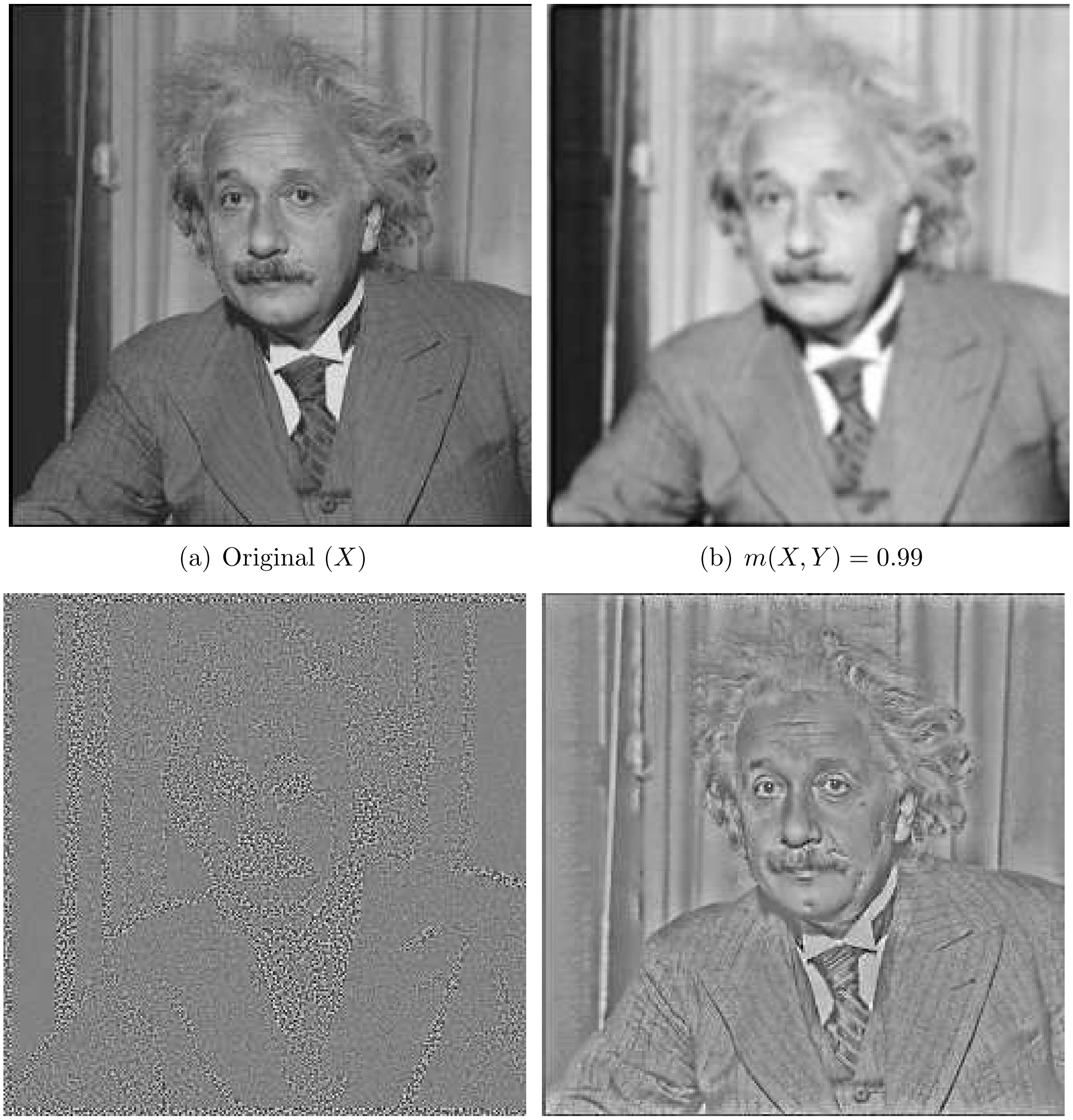
and following the extension from SSIM to MS-SSIM, MS-SSIM* is given as

$$\text{MS-SSIM}^*(X,Y) = m_K^*(X,Y)^{\alpha_K} \prod_{k=1}^{K} v_k^*(X,Y)^{\beta_k} r_k^*(X,Y)^{\gamma_k}, \tag{2.10}$$

where $m_k^*(X,Y)$, $v_k^*(X,Y)$, and $r_k^*(X,Y)$ respectively correspond to the alternative mean, variance, and cross-correlation components each computed and pooled over the image patches from scale $k$ with $k = 1$ as the full-resolution image. MS-SSIM* inherits the MS-SSIM values for the exponents: $\alpha_K$, $\{\beta_k\}_{k=1}^K$, and $\{\gamma_k\}_{k=1}^K$.

### SSIM Component Gradient Analysis

The SSIM index as given in Eq. (2.3) combines three components to estimate the visual quality of an image, but it is not immediately obvious how each component evaluates visual quality. A gradient analysis illustrated that for a fixed MSE, the total SSIM index favors an image with increased visual quality [137]. However, a gradient analysis of the individual components of SSIM was not conducted.



(a) Original $(X)$

(b) $m(X,Y) = 0.99$

(c) $v(X,Y) = 0.99$

(d) $r(X,Y) = 0.98$

Figure 2.4: Gradient analysis of the individual SSIM components: mean $m(X,Y)$, variance $v(X,Y)$, and cross-correlation $r(X,Y)$. Images (b) – (d) have been rescaled for visibility.

A gradient analysis, inspired by [137], is performed to examine the visual quality evaluation corresponding with the individual components. An original natural image $X$ is selected, and a random image $Y$ is formed whose pixel values are independently and identically drawn from a uniform distribution with mean 128 and standard deviation 1/12. For example, to optimize according to the mean component of SSIM, $m(X, Y)$, the image $Y$ is updated at iteration $k$ via gradient ascent according to

$$Y \leftarrow Y + \eta(k)\nabla_Y m(X, Y), \tag{2.11}$$

where $\eta(k)$ is the learning rate at iteration $k$ and $\nabla_Y m(X, Y)$ denotes the gradient of the mean component with respect to $Y$. Here, $m(X, Y)$ denotes the average of the individual patch means $m(x, y)$.

Figure 2.4 illustrates the effect of maximizing the individual components of SSIM for the natural image *einstein*. At first glance, using the mean component generates an image (Figure 2.4(b)) that most resembles the original in Figure 2.4(a) among the three components. However, the maximum for $m(X, Y)$ does not produce a sharp image. The optimization with the SSIM variance component yields a textured image (Figure 2.4(c)), where the textures occur along the image edges. The variance component optimization does not adequately restrict the possible pixel value configurations to produce an easily recognizable image. The image optimizing the cross-correlation component captures most of the details from the original image. For instance, notice the details in the hair, eyes and mustache in Figure 2.4(d). Moreover, the facial expression has a more accurate phenomenal appearance in Figure 2.4(a) with respect to the original than in Figure 2.4(b), where the expression appears melancholy rather than alert. The SSIM cross-correlation component clearly assesses quality according to the preservation of the reference image edges.

### Estimating Visual Quality with SSIM and MS-SSIM Components

The components of SSIM, SSIM*, MS-SSIM, and MS-SSIM* are analyzed in terms of the consistency of their objective estimates with subjective scores. The LIVE image database [111] is used to assess the performance of the components. This analysis considers the individual performance of the components and the performance of these components in pairs. That is, the analysis examines the performance of the mean; variance; cross-correlation; mean and variance; mean and cross-correlation; and variance and cross-correlation. Then, the predictive performance of $v \times r$ (cf. Eq. (2.4)) is assessed when removing the calculation of the patch means $\mu_x$ and $\mu_y$.

The SSIM and SSIM* components were computed with $\alpha = \beta = \gamma = 1$ and after filtering and downsampling the reference and test images by a factor of 2 in both spatial directions as specified by [137]. MS-SSIM and MS-SSIM* were computed with the exponents as specified in Section 2.3.4.

The LIVE image database is a large collection of distorted images for which subjective visual quality scores have been recorded [111]. The database consists of 29 reference 24-bits/pixel color images and 779 distorted images. Five types of distortions were evaluated: 1) JPEG-2000 (J2K) compression, 2) JPEG (JPG) compression, 3) additive white Gaussian noise (Noise), 4) Gaussian blurring (Blur) , and 5) simulated bitstream errors of a JPEG-2000 compressed bitstream in a fast-fading (FF) channel. Realigned difference mean opinion scores (DMOS) were used for the subjective scores [116].

The objective estimates were computed from grayscale images generated according to $Y = 0.2989R + 0.5870G + 0.1140B$, where $R$, $G$, and $B$ denote the 8-bit grayscale red, green, and blue image intensities. The nonlinear mapping of the objective estimates $a$ to the subjective scores $f$ is given as

$$f(a) = \frac{p_1}{1 + \exp(p_2(a - p_3))} + p_4. \tag{2.12}$$

The parameters $\{p_j\}_{j=1}^{4}$ were fitted to the data via a Nelder-Mead search to minimize the sum-squared error between the nonlinear mapped objective estimates and the subjective scores. The performance assessment is based on the linear correlation computed between the DMOS and the objective estimates after nonlinear regression.

***Estimation using Individual Components and Pairwise Products of Components*** The nonlinear mapping of Eq. (2.12) was fitted using the objective evaluations for the entire set of distorted images (ALL) for each component and component pair tested. Tables 2.1 and 2.2 report the statistics summarizing both the correlation and accuracy of SSIM, SSIM*, MS-SSIM, and MS-SSIM* as quality estimators as well as the individual components of each of the estimators and the pairwise products of the components of the estimators. The statistics were computed after nonlinear regression.

Individually, the SSIM cross-correlation component estimates subjective scores the best among its individual components and nearly as well as the corresponding complete SSIM definition across the six artifact types. The mean component ($m$ or $m^*$) estimates the subjective scores least accurately among the three components. The mean component alone performs poorly as a quality estimator for MS-SSIM and MS-SSIM, since the mean component is only computed for very low frequency content.

Among the pairwise combinations of the components, the product of the variance and cross-correlation components ($v \times r$) performs nearly identically to the

Table 2.1: Statistics summarizing the correlation and accuracy of the objective estimators as quality estimators using the subjective scores from the LIVE database [116]. The objective estimator values were produced by SSIM, SSIM's individual components, SSIM's pairwise components, SSIM*, SSIM*'s individual components, and SSIM*'s pairwise components. The linear correlation $r$, Spearman rank correlation $\rho$, and the Kendall rank correlation $\tau$ are computed between the objective estimates and subjective scores (i.e., DMOSs) after nonlinear regression. The root mean square error (RMSE), outlier ratio (OR), skewness (skew), and kurtosis (kurt) are computed using the residual between the fitted objective scores and the subjective scores.

| | *Components* | $r$ | $\rho$ | $\tau$ | RMSE | OR | skew | kurt |
|---|---|---|---|---|---|---|---|---|
| SSIM | $m \times v \times r$ | 0.937 | 0.948 | 0.797 | 9.614 | 0.648 | -0.083 | 2.789 |
| SSIM | $m$ | 0.833 | 0.864 | 0.683 | 15.122 | 0.770 | -0.555 | 4.090 |
| SSIM | $v$ | 0.889 | 0.918 | 0.744 | 12.524 | 0.748 | 0.149 | 2.445 |
| SSIM | $r$ | 0.930 | 0.939 | 0.776 | 10.067 | 0.675 | 0.061 | 2.550 |
| SSIM | $m \times v$ | 0.891 | 0.921 | 0.749 | 12.406 | 0.746 | 0.140 | 2.465 |
| SSIM | $m \times r$ | 0.930 | 0.939 | 0.776 | 10.066 | 0.682 | 0.049 | 2.561 |
| SSIM | $v \times r$ | 0.937 | 0.949 | 0.797 | 9.548 | 0.646 | -0.101 | 2.780 |
| SSIM* | $m^* \times v^* \times r^*$ | 0.864 | 0.854 | 0.668 | 13.759 | 0.724 | -0.689 | 4.060 |
| SSIM* | $m^*$ | 0.832 | 0.863 | 0.682 | 15.169 | 0.769 | -0.571 | 4.161 |
| SSIM* | $v^*$ | 0.837 | 0.847 | 0.658 | 14.971 | 0.736 | -0.441 | 4.204 |
| SSIM* | $r^*$ | 0.860 | 0.843 | 0.654 | 13.959 | 0.742 | -0.589 | 3.679 |
| SSIM* | $m^* \times v^*$ | 0.842 | 0.853 | 0.666 | 14.724 | 0.721 | -0.483 | 4.243 |
| SSIM* | $m^* \times r^*$ | 0.860 | 0.843 | 0.655 | 13.919 | 0.741 | -0.588 | 3.684 |
| SSIM* | $v^* \times r^*$ | 0.872 | 0.861 | 0.674 | 13.379 | 0.732 | -0.553 | 3.755 |

Table 2.2: Statistics summarizing the correlation and accuracy of the objective estimators as quality estimators using the subjective scores from the LIVE database [116]. The objective estimator values were produced by MS-SSIM, MS-SSIM's individual components, MS-SSIM's pairwise components, MS-SSIM*, MS-SSIM*'s individual components, and MS-SSIM*'s pairwise components. The linear correlation $r$, Spearman rank correlation $\rho$, and the Kendall rank correlation $\tau$ are computed between the objective estimates and subjective scores (i.e., DMOSs) after nonlinear regression. The root mean square error (RMSE), outlier ratio (OR), skewness (skew), and kurtosis (kurt) are computed using the residual between the fitted objective scores and the subjective scores.

| *Estimator* | *Components* | $r$ | $\rho$ | $\tau$ | RMSE | OR | Skew. | Kurt |
|---|---|---|---|---|---|---|---|---|
| MS-SSIM | $m \times v \times r$ | 0.934 | 0.945 | 0.793 | 9.775 | 0.623 | -0.156 | 3.260 |
| MS-SSIM | $m$ | 0.284 | 0.691 | 0.500 | 26.198 | 0.883 | 0.123 | 2.065 |
| MS-SSIM | $v$ | 0.881 | 0.909 | 0.731 | 12.909 | 0.765 | 0.138 | 2.356 |
| MS-SSIM | $r$ | 0.930 | 0.938 | 0.777 | 10.039 | 0.673 | -0.123 | 2.975 |
| MS-SSIM | $m \times v$ | 0.881 | 0.909 | 0.731 | 12.909 | 0.765 | 0.138 | 2.356 |
| MS-SSIM | $m \times r$ | 0.930 | 0.938 | 0.777 | 10.040 | 0.673 | -0.123 | 2.975 |
| MS-SSIM | $v \times r$ | 0.934 | 0.945 | 0.793 | 9.774 | 0.623 | -0.156 | 3.260 |
| MS-SSIM* | $m^* \times v^* \times r^*$ | 0.872 | 0.867 | 0.685 | 13.365 | 0.714 | -0.730 | 4.337 |
| MS-SSIM* | $m^*$ | 0.284 | 0.691 | 0.499 | 26.196 | 0.883 | 0.123 | 2.067 |
| MS-SSIM* | $v^*$ | 0.835 | 0.846 | 0.658 | 15.044 | 0.763 | -0.372 | 4.128 |
| MS-SSIM* | $r^*$ | 0.871 | 0.859 | 0.674 | 13.432 | 0.727 | -0.634 | 3.831 |
| MS-SSIM* | $m^* \times v^*$ | 0.835 | 0.846 | 0.658 | 15.042 | 0.761 | -0.372 | 4.128 |
| MS-SSIM* | $m^* \times r^*$ | 0.871 | 0.859 | 0.674 | 13.431 | 0.727 | -0.633 | 3.830 |
| MS-SSIM* | $v^* \times r^*$ | 0.872 | 0.867 | 0.685 | 13.365 | 0.714 | -0.731 | 4.337 |

corresponding complete definition that uses all three components. The product of the mean and variance components ($m \times v$) estimates subjective score well, but it is evident that the incorporation of the cross-correlation component significantly improves its objective quality estimation. Even the product of the mean and cross-correlation components ($m \times r$) accurately estimates subjective scores.

The results indicates that the cross-correlation component is the most important component among the three components when estimating the subjective scores. The absence of the cross-correlation component leads to a significant drop in both the accuracy and correlation between the objective scores and the subjective scores.

The proposed modifications to SSIM and MS-SSIM demonstrate that the constants provided with the original specifications of SSIM and MS-SSIM "tune" these estimators to the LIVE database. While SSIM* and MS-SSIM* cross-correlation components do not estimate the subjective scores as accurately as the SSIM and MS-SSIM cross-correlation components, they correctly indicate that a flat image (e.g., cf. Figure 2.3(d)) is different from the reference image. The cross-correlation component of MS-SSIM* $r^*$ has been demonstrated for the use of quality assessment of mammograms [96].

***Estimation without Computing $\mu_x$ or $\mu_y$ for SSIM*** The performance of the mean component with the LIVE image database casts doubt on its relevance in objective quality estimation for typical image artifacts.[4] However, removing the mean component $m$ from the SSIM index does not significantly reduce the computational complexity, since the variance and cross-correlation components use the terms from $m$: $\mu_x$, $\mu_y$.

[4] The LIVE database contains image artifacts representative of typical imaging applications, where there is limited variation to the luminance.

Removing or fixing the values of $\mu_x$ and $\mu_y$ produces significant computational savings. When $\mu_x$ and $\mu_y$ are computed for two patches $x$ and $y$ of $n$ pixels, the computation of $v \times r$ over $n$ pixels requires $8n + 8$ multiplications. However, if $\mu_x$ and $\mu_y$ are fixed or set to zero, the computation of $v \times r$ reduces to $6n + 8$ multiplications. For a patch of size $n = 11$, this leads to a reduction of more than 20% in the number of multiplications.

The computation of $v \times r$ with $\mu_x = \mu_y = 128$ (c.f. Eq. (2.13)) predicts subjective quality scores very well across all distortion types. Table 2.3 summarizes the linear correlation coefficients for $v \times r$ when the values $\mu_x$ and $\mu_y$ are fixed to 128. For comparison, the linear correlation of $v \times r$ is included. Moreover, the performance for $\mu_x = \mu_y = 128$ is very similar to the complete SSIM computation.

### SSIM: Analysis and Discussion

The gradient analysis of the SSIM components along with the results in Section 2.3.4 emphasizes the significance of the cross-correlation component when assessing perceptual quality. Human evaluations of perceptual quality demonstrate a preference for images that preserve image edge information across image scales [17]. This finding is consistent with the principle of global precedence, which contends that the HVS processes a visual scene in a global-to-local order [76]. The MS-SSIM cross-correlation component explicitly evaluates the pixel values across image scales, which provides a measure of how well the edges of two images match. For both SSIM and MS-SSIM, the image that maximizes the cross-correlation component with respect to a reference image possess identical edge information.

A simple analysis explains the estimation accuracy of $v \times r$ when the local average pixel values are set to 128 (cf. Table 2.3). Let $\mu$ denote a fixed mean offset

Table 2.3: Linear correlation coefficients between DMOS [116] and $v \times r$ for fixed $\mu_x = \mu_y = \mu$ after nonlinear regression for each artifact type in LIVE image database [111].

| | *Artifact Type* | | | | | |
|---|---|---|---|---|---|---|
| *Components* | ALL | J2K | JPG | Noise | Blur | FF |
| $v \times r$ | .937 | .966 | .979 | .908 | .947 | .948 |
| $\mu = 128$ | .925 | .936 | .965 | .898 | .917 | .927 |

subtracted from an image before computing the product of the SSIM variance and cross-correlation components. In terms of the SSIM definitions of $\mu_x$, $\mu_y$, $\sigma_x^2$, $\sigma_y^2$, and $\sigma_{xy}$, the product of the modified variance and cross-correlation components for a fixed mean offset $\mu$ is given as

$$\hat{v}(x,y) \times \hat{r}(x,y) = \frac{2\sigma_{xy} + C + AB}{\sigma_x^2 + \sigma_y^2 + C + A^2 + B^2}, \tag{2.13}$$

where $A = \mu_x - \mu$ and $B = \mu_y - \mu$. Eq. (2.13) is very similar to the $v \times r$ component of Eq. (2.4). The additional constant $AB$ in the numerator only shifts the objective score, and the additional constant $A^2 + B^2$ in the denominator rescales the objective score. Using the minimum MSE estimate of the mean pixel value, $\mu = 128$, ensures that on average other values of $\mu$ will demonstrate poorer predictive performance. Objective quality estimation using Eq. (2.13) does not significantly alter the linear correlation between the DMOS and the objective scores as demonstrated by the results in Table 2.3.

### SSIM: Summary

This chapter examines how the SSIM components (mean, variance, and cross-correlation) contribute to its quality estimation of common image artifacts. Modified versions of SSIM and MS-SSIM, denoted SSIM* and MS-SSIM*, are proposed whose individual components accurately reflect the difference between a very distorted image and its reference image.

An analysis of the performance of the components of SSIM, SSIM*, MS-SSIM, and MS-SSIM* as quality estimators is performed using the LIVE image database. The objective scores using the product of the variance and cross-correlation components match those of the complete SSIM and MS-SSIM calculations. A computationally simple alternative to SSIM (cf. Eq. (2.13)) that ignores the mean component and sets the local average patch values to 128 exhibits a 1% decrease in linear correlation with subjective scores to 0.934 from the complete SSIM evaluation with an over 20% reduction in the number of multiplications.

### 2.3.5 The Visual Information Fidelity (VIF) Criterion

The visual information fidelity (VIF) criterion is an extension of the information fidelity criterion (IFC) that incorporates a simple human visual system (HVS) model[5] [113,115]. VIF*, a modified version of VIF, adjusts the relative importance of distortions measured across spatial frequencies to the overall objective estimate by normalizing VIF's channel measurements before linearly pooling across image scales. VIF* provides accurate estimates of perceived quality for a broader set of distortions than VIF. A detailed mathematical description of VIF and VIF* is presented as well as the performance of each as quality estimators using the LIVE database.

[5] VIF has been argued to imitate aspects of the HVS based on its mathematical resemblance to existing models of the HVS [110].

### VIF Specification

VIF extends IFC by modeling the human visual system (HVS) as an additive Gaussian noise source that was conjectured by VIF's authors to model low-level HVS processing [113]. VIF's assessment of a test image is based on spatially local SNR measurements, computed at multiple image scales, of both the reference and test images contaminated with the modeled, low-level HVS noise.

Let the elements of the length $N_k$ vectors $\boldsymbol{C}^k$ and $\boldsymbol{D}^k$ denote the wavelet coefficients of the $k^{th}$ channel of the reference and test images, respectively.[6] The elements of the length $N_k$ vectors $\boldsymbol{E}^k$ and $\boldsymbol{F}^k$ denote the wavelet coefficients of the $k^{th}$ channel of the respective reference and test images that have been contaminated with visual noise.

VIF parses each wavelet channel into disjoint blocks composed of $P$ coefficients. The following discussion assumes only one channel, so the superscript $k$ is omitted in the subsequent discussion. Let $\vec{C}_b$ and $\vec{D}_b$ correspond to the $b^{th}$ block of $P$ spatially adjacent coefficients of $\boldsymbol{C}$ and $\boldsymbol{D}$, respectively. The $b^{th}$ block of wavelet coefficients in the channel of the reference image may be modeled as a Gaussian scale mixture [135, 136] (GSM) random vector given as $\vec{C}_b = s_b\vec{U}$, where $s_b$ is a positive random scalar and $\vec{U}$ is a zero mean Gaussian random vector of length $P$ with covariance $\boldsymbol{K}_{\vec{U}}$. The GSM model has been demonstrated to approximate the distribution of wavelet coefficients for natural images . Given $s_b$, the coefficient block $\vec{C}_b$ is a zero mean Gaussian random scalar with covariance $s_b^2\boldsymbol{K}_{\vec{U}}$, and $\vec{C}_b$ is conditionally independent of $\vec{C}_m$ for all $m \neq b$. VIF relates the $b^{th}$ block of wavelet coefficients of the test and reference images using the linear model $\vec{D}_b = g_b\vec{C}_b + \vec{V}_b$,

[6] The subscript $k$ for $N_k$ accounts for decimated wavelet decompositions, such as the steerable pyramid, whose channels in coarser image scales have fewer coefficients than channels in finer image scales.

where $g_b$ is a deterministic scalar defined for each block and $\vec{V_b}$ is a zero mean Gaussian random vector of length $P$ with covariance matrix $\sigma^2_{\vec{V_b}}\boldsymbol{I}$ specified for each block $b$. Thus, given $s_b$, the block of coefficients $\vec{D_b}$ is also a Gaussian random vector with covariance $g_b^2 s_b^2 \boldsymbol{K}_{\vec{U}} + \sigma_{\vec{V_b}}\boldsymbol{I}$.

Independent zero-mean additive Gaussian noise sources model low-level HVS noise in VIF; coefficients of the reference and test images are contaminated with visual noise. Let $\vec{E_b}$ and $\vec{F_b}$ correspond to the $b^{th}$ block of $P$ spatially adjacent coefficients of $\boldsymbol{E}$ and $\boldsymbol{F}$, respectively. The output of the HVS model for the reference image is $\vec{E_b} = \vec{C_b} + \vec{M_b}$, and the output of the HVS model for the test image is $\vec{F_b} = \vec{D_b} + \vec{N_b}$. The terms $\vec{M_b}$ and $\vec{N_b}$ are a zero mean Gaussian random vectors of length $P$ with covariance $\sigma^2_M\boldsymbol{I} = \sigma^2_N\boldsymbol{I}$, where $\sigma^2_N = \sigma^2_M$ is the HVS model parameter. Thus, given $s_b$, the block of coefficients $\vec{E_b}$ is a Gaussian random vector with covariance $s_b^2\boldsymbol{K}_{\vec{U}} + \sigma^2_N\boldsymbol{I}$, and the block of coefficients $\vec{F_b}$ is also a Gaussian random vector with covariance $g_b^2 s_b^2\boldsymbol{K}_{\vec{U}} + \sigma^2_{\vec{V_b}}\boldsymbol{I} + \sigma^2_N\boldsymbol{I}$

VIF combines two evaluations to yield an overall assessment of a test image. First, an evaluation comparing the reference coefficients before and after the HVS model value is computed. Second, an evaluation comparing the reference coefficients before the HVS model to the processed coefficients after the HVS model is computed. These two evaluations are computed for each wavelet channel. The ratio of the sum of these evaluations across the channels provides an overall assessment of the test image. Let $\boldsymbol{s}$ be a length $B_k$ vector whose $b^{th}$ element is $s_b$. Given $\boldsymbol{s}$, the VIF value is given by

$$\mathrm{VIF} = \frac{\sum_{k=1}^{K} IFC(\boldsymbol{C}^k, \boldsymbol{F}^k)}{\sum_{k=1}^{K} IFC(\boldsymbol{C}^k, \boldsymbol{E}^k)}. \tag{2.14}$$

The terms $IFC(\boldsymbol{C}^k, \boldsymbol{F}^k)$ and $IFC(\boldsymbol{C}^k, \boldsymbol{E}^k)$ are based on IFC [115] and are defined as

$$IFC(\boldsymbol{C}^k, \boldsymbol{F}^k) = \\ \sum_{b=1}^{B_k} \log_2 \left( \frac{|g_b^2 s_b^2 \boldsymbol{K}_{\vec{U}} + (\sigma_{\vec{V}_b}^2 + \sigma_N^2)\boldsymbol{I}|}{|(\sigma_{\vec{V}_b}^2 + \sigma_N^2)\boldsymbol{I}|} \right) \tag{2.15}$$

and

$$IFC(\boldsymbol{C}^k, \boldsymbol{E}^k) = \sum_{b=1}^{B_k} \log_2 \left( \frac{|s_b^2 \boldsymbol{K}_{\vec{U}} + \sigma_N^2 \boldsymbol{I}|}{|\sigma_N^2 \boldsymbol{I}|} \right), \tag{2.16}$$

where $|\cdot|$ denotes the matrix determinant and the terms $g_b$, $s_b$, $\boldsymbol{K}_{\vec{U}}$, and $\sigma_{\vec{V}_b}$ vary with $k$ and are computed from $\boldsymbol{C}^k$ and $\boldsymbol{D}^k$. For channel $k$, the term $g_b$ is estimated as the linear regression of block $\vec{D}_b$ on the block $\vec{C}_b$, and the variance of the additive zero mean Gaussian noise $\vec{V}_b$ is the mean squared error of the regression.

**VIF* Specification**

VIF emphasizes evaluations of finer image scales (i.e., higher spatial frequencies) over those of coarser image scales (i.e., lower spatial frequencies). Thus, VIF is invariant to disruptions to low frequency content (see Figure 4.1), which is functionally due to the variation in the number of coefficients blocks $B_k$ for channels at different image scales. Channels corresponding to finer image scales have more wavelet coefficients than channels corresponding to coarser image scales due to the use of a decimated wavelet transform; for a fixed block size $P$, the number of coefficient blocks is smaller for channels corresponding to coarser image scales. The proposed modifications of VIF, denoted VIF*, normalizes the channel measurements by the number of blocks $B_k$ for that channel. The same variables defined for VIF are used to mathematically specify VIF*. VIF* is defined as

$$\text{VIF}^* = \frac{\sum_{k=1}^{K} \frac{1}{B_k} IFC(\boldsymbol{C}^k, \boldsymbol{F}^k)}{\sum_{k=1}^{K} \frac{1}{B_k} IFC(\boldsymbol{C}^k, \boldsymbol{E}^k)}, \tag{2.17}$$

where $IFC(\boldsymbol{C}^k, \boldsymbol{F}^k)$ and $IFC(\boldsymbol{C}^k, \boldsymbol{E}^k)$ are defined as in Eq. (2.15). As illustrated in Figure 4.1, VIF* produces distinct scores that reflect the changes in the perceived quality scores for these images. In particular, disruptions to low-frequency content affect VIF*'s estimate, whereas VIF's estimate does not.

Statistics summarizing the performance of both VIF and VIF* as quality estimators using the perceived quality scores in the LIVE image database [111] are reported in Table 2.4. The a linear mapping was fitted to map the objective scores the perceived quality scores (i.e., DMOS) provided in the LIVE database. The standard performance statistics are reported after applying the linear mapping: the Pearson linear correlation $r$; the Spearman rank correlation $\rho$; the Kendall rank correlation $\tau$; the root mean-squared error (RMSE); the outlier ratio (OR); and the skewness and kurtosis of the residuals. The Pearson linear correlation coefficients for VIF and VIF* are statistically equivalent at the 95% confidence level. However, VIF and VIF* do not have statistically equivalent rank correlation coefficients at the 95% confidence level. The RMSE values are statistically equivalent according to the Brown-Forsythe-Levene test [10].

The improvement of VIF over VIF* in terms of the rank correlation is not surprising, since a nonlinearity is often fitted to the data to resolve any nonlinear relationship. However, VIF* provides estimates of perceived quality that are effectively as reliable as those produced by VIF. Part of the reason that VIF exhibits slightly better performance is as results of the distortions used in the LIVE database. In particular, all of the distortions impact high frequency components, which VIF is very sensitive to relative to VIF*. Thus, VIF is slightly better at resolving differences in the perceived quality of the distorted images in the LIVE databases. In Chapter 4, it is shown that VIF* can provide more accurate per-

Table 2.4: Statistics summarizing the performance of VIF and VIF* as quality estimators using the LIVE database.

| Estimator | $r$ | $\rho$ | $\tau$ | RMSE | OR | Skew./Kurt |
|---|---|---|---|---|---|---|
| VIF | -0.941 | -0.964 | -0.828 | 9.2 | 0.597 | 0.21/2.8 |
| VIF* | -0.938 | -0.934 | -0.780 | 9.5 | 0.599 | -0.08/3.7 |

ceived quality estimators for a wider variety of distortion artifacts.

## 2.4 Summary

This chapter reviewed several image databases that contain a collection distorted images along with responses from human observers. The image databases containing perceived quality scores demonstrate that a variety of test methods have been used to obtain these scores, and no standard method of acquiring perceived quality scores has been developed. The public safety video quality collects responses from observers prompted to perform specific tasks with an image, which provides the observer with prior information about the distorted image before viewing the image. Chapters 3 and 5 discuss techniques to obtain responses from human observers about the usefulness of distorted images without specific prompts prior to viewing a distorted stimulus.

This chapter also reviewed several state-of-the-art quality estimators. The structural similarity (SSIM) index and the visual information fidelity (VIF) criterion are discussed at length, as these two quality estimators are largely viewed as the "best" in the image quality assessment community. Limitations for both of these quality estimators are revealed and discussed. The image quality estimators reviewed in this chapter form a suite of baseline objective estimators that are compared with a novel utility estimator introduced in Chapter 4.

## 2.5 Key Points

- Many testing protocols have been implemented to form databases that contain perceived quality scores provided by human observers of distorted images. Several such databases that are commonly used to validate the performance of objective quality estimators were reviewed.
- Several objective estimators have been proposed and evaluated using popular image databases with perceived quality scores. Among those estimators, the structural similarity (SSIM) index is commonly adopted in lieu of mean-squared error or peak signal-to-noise ratio as a perceived quality estimator. This chapter illustrated fundamental flaws with the components of SSIM (see Figure 2.2). A modified version of SSIM is presented (see Eq. (2.9)) that has been demonstrated by other researchers as a quality estimator for mammograms [96].
- The visual information fidelity (VIF) criterion is another quality estimator that currently marketed as the leading quality estimator. A modified version, VIF*, is described (see Eq. (2.17)) that is shown to provide much more accurate estimates of perceived quality than VIF in Chapter 4.

CHAPTER 3

# SUBJECTIVE ESTIMATES OF PERCEIVED UTILITY AND PERCEIVED QUALITY

> **Hoke:** How you know the way I see, less you lookin outta my eyes?
>
> *Driving Miss Daisy*, Alfred Uhry

## 3.1 Introduction

The work presented in this chapter is motivated by the prior work in both image quality and utility and expands the previous narrowly studied definitions of utility in a manner that allows both a broader evaluation of utility as well as a characterization of the underlying image characteristics that impact usefulness. Unlike the specific tasks performed with images in prior work, the "task" is instead to report the content of an image as it is gradually improved from an initially extremely distorted and unrecognizable version to a visually lossless[1] version. A novel suite of experiments presented here provides utility scores for distorted images, and quality scores are collected using a standard test methodology. Distortions were strategically selected to disrupt various spatial frequencies in a broader sense than those traditionally studied in perceived quality experiments.

An analysis of the resulting relationship between perceived quality and perceived utility demonstrates that an image's perceived quality does not imply that image's usefulness and vice versa. Therefore, an *objective* estimator that accu-

[1] A visually lossless image is visually indistinguishable from a reference image.

rately estimates perceived quality scores cannot accurately estimate perceived utility scores and vice versa. These results motivate a thorough analysis of the images to understand the image characteristics that produce *distorted but useful* images for human observers.

To the best of our knowledge, no experimental methods exist to measure the perceived utility of distorted natural images when the task is to report the content of an image. This chapter reports the first usage of such experimental methods as well as a subsequent analysis. Section 3.2 presents the proposed experimental methodology used to collect perceived utility scores. Several standard methods are available to collect perceived quality scores for distorted natural images, and Section 3.3 reviews the experimental methodology we used to collect perceived quality scores. Experimental results illustrating the relationship between the perceived utility and perceived quality scores are presented in Section 3.4.

## 3.2 Methods: Perceived Utility Scores

For a human performing a task, a distorted natural image is a surrogate for an undistorted, reference image. A *perceived utility score* quantifies the usefulness of that distorted image with respect to the reference image for that task. More useful images provide more information about the image content to an human.

Two meaningful anchors are associated with the perceived utility of an image: the *recognition threshold equivalence class* and the *reference equivalence class.* The *recognition threshold equivalence class*, henceforth denoted the recognition threshold (RT), specifies a collection of maximally degraded images from which humans still accurately recognize the basic content of the reference image. The perceived utility score of the RT can distinguish useful distorted images from useless distorted images formed from a reference image. In particular, an image with a perceived utility score greater than that of its RT is useful, whereas an image with a perceived utility score less than that of the RT is useless. Humans recognize at least the basic content of useful images but recognize nothing in useless images.

The *reference equivalence class* (REC) specifies an equivalence class of images, including the reference image, that yield the *same* interpretation of the content as the reference image. Images in the REC may contain signal degradations that may or may not be visible to a human observer but still convey the same information as the reference image. For example, a visually lossless image could contain signal distortions yet remain visually indistinguishable from the reference image, so a visually lossless image belongs to the REC.

Two experiments[2] were conducted to obtain perceived utility scores. The first experiment acquires subjective data that were processed (cf. Section 3.2.4) to produce *relative* perceived utility scores for a collection of distorted natural images generated from each reference image. These relative perceived utility scores correspond to a unique range of values that varies for each reference image. The relative perceived utility scores for the RT and the REC of each reference image are used to map the relative perceived utility scores to a common range of values. On this common range of values, the RT is indicated by a perceived utility score of 0, and the REC is indicated by a perceived utility score of 100. The subjective data obtained in the second experiment is used to estimate the RT of each reference image. The REC did not need to be estimated from experimental data, because both the reference image and any visually lossless image belong to the REC. A

[2] The following experiments described in this section augment the experiments described in the publications [102, 103], and [100].

visually lossless image generated via JPEG-2000 compression using the dynamic contrast-based quantization (DCQ) strategy [17] defined the REC of each reference image (cf. Section 3.2.1).

The remainder of this section describes the methods used to collect subjective data and produce perceived utility scores. First, the distortion types used to construct *reference/distortion* image sequences are described. Then, the methods are reported for the experiments conducted using these sequences to acquire subjective data to 1) produce relative perceived utility scores and 2) estimate the RTs of reference images. Last, the derivation of perceived utility scores from the collected subjective data is explained.

### 3.2.1 Reference/Distortion Image Sequences

Sequences of decreasingly distorted natural images were generated from a reference natural image. Each sequence corresponds to a specific distortion and *evolves* such that subsequent images in the sequence gradually refine detail or information relative to the previous images. For brevity, such a sequence is henceforth denoted 1) generically as a *reference/distortion sequence* and 2) more specifically by explicitly indicating either the reference image name, the distortion, or both (e.g., reference/JPEG denotes a sequence of JPEG distorted images corresponding to the same undisclosed reference). The reference/distortion sequences were formed by varying a single parameter that controlled the level of distortion. For a single reference subjected to a single distortion, perceived utility is assumed to exhibit a monotonically, non-decreasing relationship with decreasing distortion level. Thus, as a reference/distortion sequence evolves toward a visually lossless image the perceived utility does not decrease. The sequences of distorted images that correspond to different distortions served as test stimuli in the experiments. Select images from the airplane/J2K+DCQ sequence are shown in Figure 3.5.

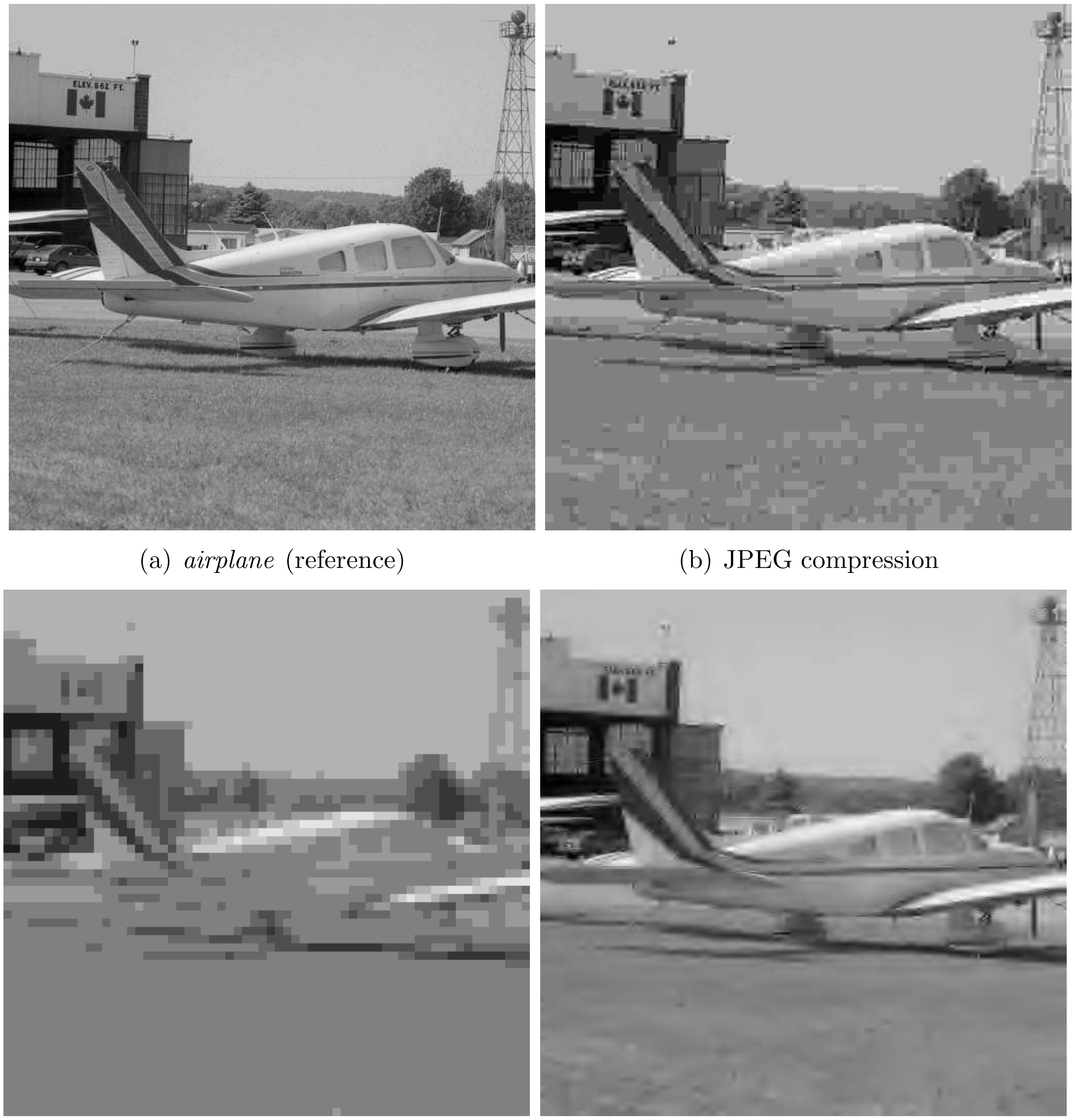
Each distortion is spatially correlated with the reference natural image and disrupts different image characteristics. The image characteristics disrupted include the spatial frequency content, contour integrity (i.e., edges), and the level of detail (i.e., textures). Example images with each distortion are shown in Figures 3.1 and 3.2, and Table 3.1 summarizes each distortion. The remainder of this section describes the five distortions evaluated in the experiments.

### JPEG: Quantized Discrete Cosine Transform Coefficients

JPEG achieves lossy compression of natural images by quantizing block-based discrete cosine transform (DCT) coefficients [90]. The quantization strategy implemented in the source code library provided by the Independent JPEG Group (IJG) [79] is used and is parameterized by $P_{jpeg} \in [0, 100]$, which scales the example luminance component quantization table suggested in the JPEG specification [124]. A sequence of images with JPEG compression artifacts evolves by increasing the parameter $P_{jpeg}$.

### BLOCK: Extreme Blocking Artifacts

Extremely low rate JPEG images effectively replace each $8 \times 8$ block of pixels with their average value. To simulate this, a reference/BLOCK sequence of images has extreme blocking artifacts and evolves by decreasing the quantization step-size $Q_{avg}$ of the average block pixel value.



(a) *airplane* (reference)

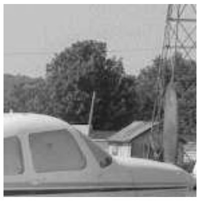
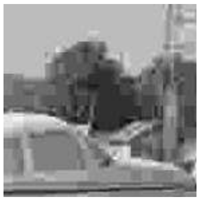
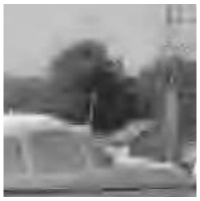
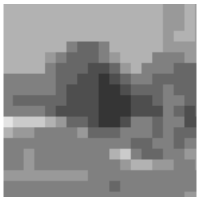
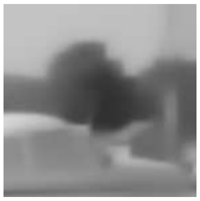
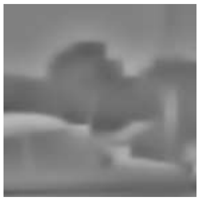
(b) JPEG compression

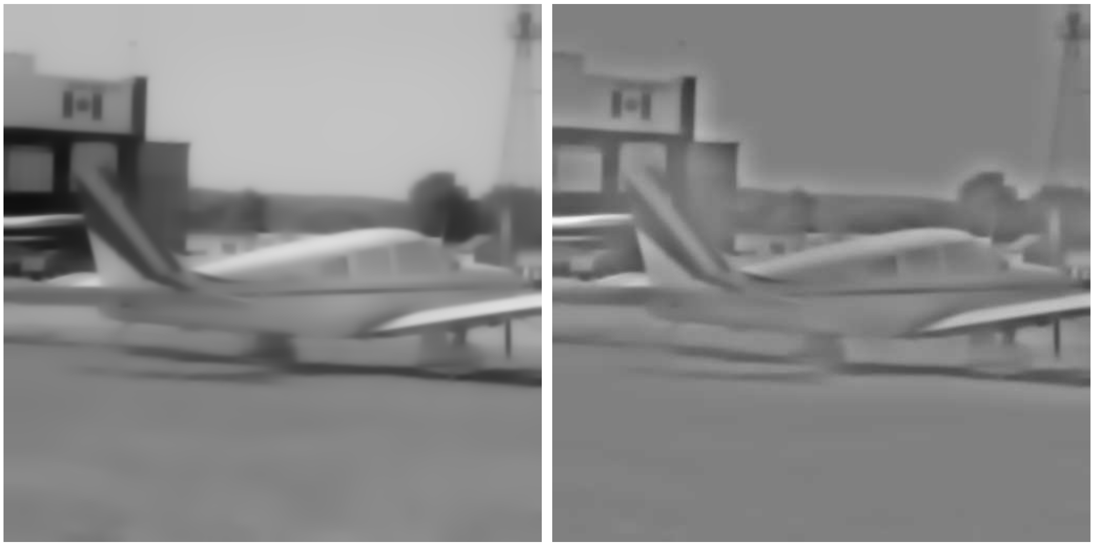
(c) BLOCK: Extreme blocking artifacts

(d) J2K+DCQ: JPEG-2000 compression using the DCQ strategy

Figure 3.1: The original reference *airplane* image and distorted images illustrating the distortions described in Section 3.2.1. The JPEG and BLOCK distortions are introduced by quantizing coefficients of a block-based discrete cosine transform. J2K+DCQ distortions result from quantizing coefficients of a discrete wavelet transform according to the dynamic contrast-based quantization (DCQ) strategy [17]. Table 3.1 contains descriptions of each of the distortions.

Table 3.1: Summary of image distortions studied. The relationship between the distortion parameter and the level of distortion is described for each distortion. For a reference image subjected to one distortion type, utility and quality are assumed to exhibit a monotonically, non-decreasing relationship with decreasing distortion level.

| *Distortion* | *Description* | *Parameter versus distortion level* | *Example* |
|---|---|---|---|
| none | Reference *airplane* image | n/a |  |
| JPEG | Quantized discrete cosine transform (DCT) coefficients according to the lossy JPEG image compression standard. Parameterized by JPEG quality parameter $P_{jpeg}$. | Increasing $P_{jpeg}$ decreases the level of distortion. |  |
| J2K+DCQ | Quantized discrete wavelet transform coefficients using quantization step-sizes specified by the dynamic contrast-based quantization (DCQ) strategy for a target encoding bitrate, $R$. | Increasing $R$ decreases the level of distortion. |  |
| BLOCK | Replace each $8 \times 8$ block of pixels by their average and quantizing this average pixel value using the quantization parameter $Q_{avg}$. | Decreasing $Q_{avg}$ decreases the level of distortion. |  |
| TS | Texture smoothing with limited disruption to image edges. Parameterized by texture smoothing parameter $\gamma$. | Decreasing $\gamma$ decreases the level of distortion. |  |
| TS+HPF | Texture smoothing (i.e., TS distortions) plus high-pass filtering. Parameterized by texture smoothing parameter $\gamma$. | Decreasing $\gamma$ decreases the level of distortion. |  |



(a) TS: Texture smoothing

(b) TS+HPF: TS plus high-pass filtering

Figure 3.2: Distorted images illustrating the distortions described in Section 3.2.1. Texture smoothing (TS) distortions are induced via total variation (TV) regularization to smooth texture regions with limited disruption to edges. A high-pass filter (HPF) that removes low-frequency signal information from images with TS distortions produces the TS+HPF distortions. Table 3.1 contains descriptions of each of the distortions.

### J2K+DCQ: Quantized Discrete Wavelet Transform Coefficients

The lossy JPEG-2000 (J2K) image compression standard represents natural images as a linear combination of wavelet basis functions [123]. Distortions are introduced by quantizing the basis function coefficients found using a discrete wavelet transform to achieve a desired encoding bitrate, $R$. The dynamic contrast-based quantization (DCQ) strategy assigns quantization step-sizes according to a measure of visual distortion parameterized by characteristics of the image, the wavelet subband coefficients, and the display. The DCQ strategy's visual distortion measure distinguishes visually lossless images from visibly distorted images, so the DCQ strategy can specify subband quantization step-sizes for lossy compression that yield a visually lossless image. A reference/J2K+DCQ sequence of images has distortions due to JPEG-2000 compression using the DCQ strategy and evolves by increasing the encoding bitrate, $R$.

### TS: Texture Smoothing

Edges distinguish objects and regions (i.e., sky and rooftop) in natural images that convey substantial meaning to human observers, whereas textures generally provide secondary information about these objects or regions. Furthermore, the extra-striate visual cortex exhibits the greatest response to images that retain contour information and lack texture information [34]. The apparent significance of edges to the human visual system inspired the evaluation of distortions that deliberately smooth texture regions in images with limited disruption to edges.

Total variation (TV) regularization traditionally has been used to remove noise from images by producing piecewise smooth images that lack textures [107]. Let $g(t)$ be a continuous signal obtained by adding noise to a reference signal for $t \in [a, b]$. TV regularization finds a restored signal $f$ from $g$ by solving an optimization problem of the form:

$$\min_f \int_a^b \left( (f(t) - g(t))^2 + \gamma \left| \frac{d}{dt} f(t) \right| \right) dt, \tag{3.1}$$

where the first term maintains the similarity between the $f$ and $g$, the second term penalizes deviations from smoothness, and $\gamma$ is a regularization parameter to control the amount of smoothing.

An alternative and equivalent approach to finding $f$ is via soft thresholding of undecimated Haar wavelet coefficients in all subbands except the coarsest LL subband [120]. The strict mathematical equivalence does not remain for 2-D signals, but sufficient visual similarities warrant this alternative for natural images [119].

Soft thresholding with thresholding parameter $\gamma$ is given by

$$S_\gamma(x) = \begin{cases} x - \gamma \operatorname{sgn}(x) & |x| > \gamma \\ 0 & |x| \leq \gamma, \end{cases} \tag{3.2}$$

where $\operatorname{sgn}(x)$ is the signum function. Given a thresholding parameter $\gamma$, thresholds are appropriately scaled for wavelet coefficients at each scale. Applying soft thresholding to a noisy image produces a similar image while reducing the presence of noise by shrinking and discarding wavelet coefficients. In the case of natural images with negligible noise, finer details such as textures act as additive noise, and soft thresholding removes those finer details. Adjusting the parameter $\gamma$ varies the level of detail removed from the reference image, where smaller values of $\gamma$ result in the removal of very few details. On the other hand, larger values of $\gamma$ induce more aggressive smoothing which may simultaneously compromise image structures (e.g., edges) important for interpretation. A 5-level undecimated Haar wavelet transform is used. A reference/TS sequence of images has distortions due to texture smoothing (TS) and evolves by decreasing a smoothing parameter $\gamma$ that controls the degree of texture smoothing induced by soft-thresholding.

### TS+HPF: Texture Smoothing plus High-Pass Filtering

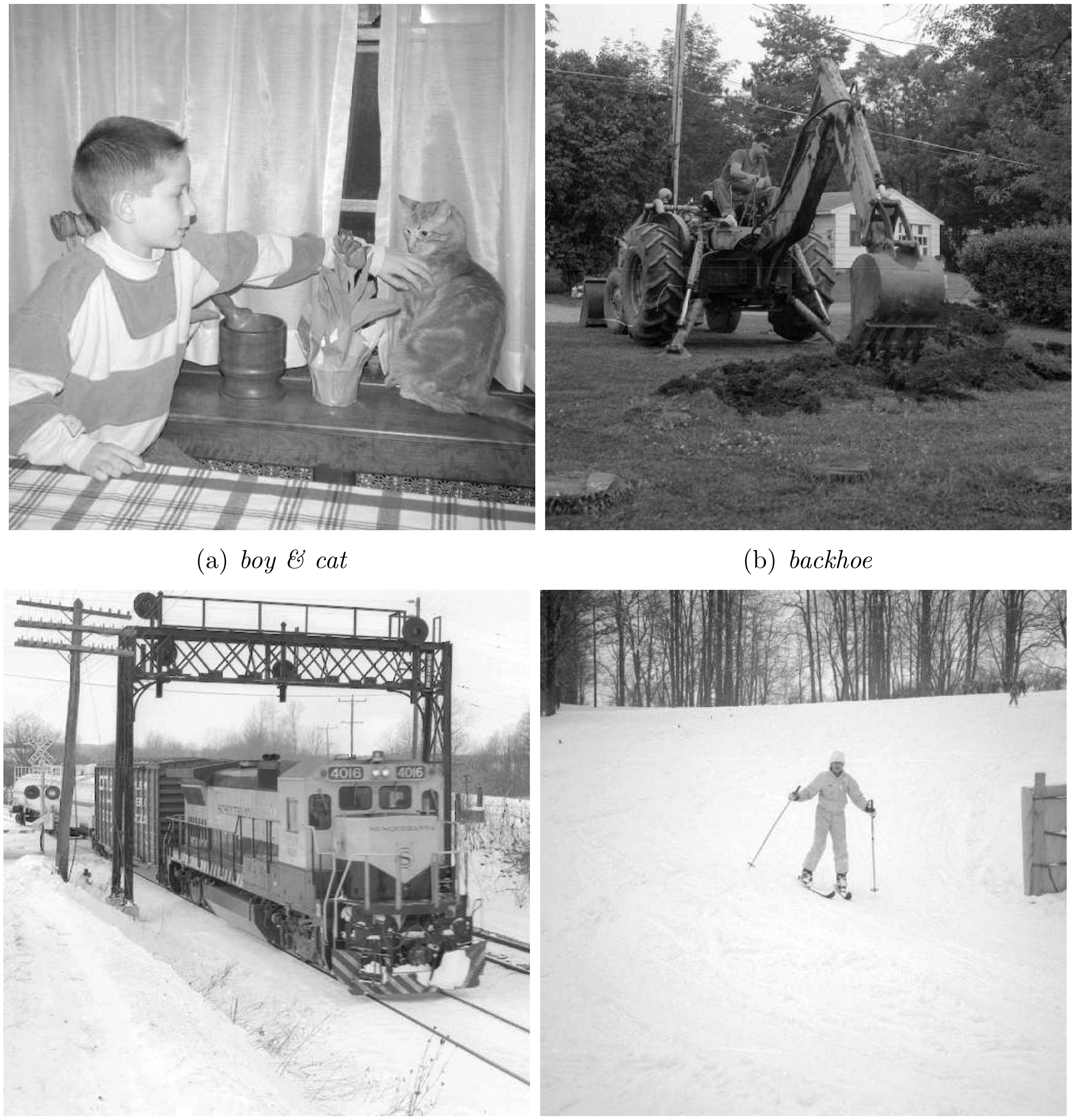
Low-frequency content is not critical to preserve the appearance of edges, which commonly coincide with object boundaries in natural images, so images subjected to texture smoothing and high-pass filtering were evaluated. When viewing high-pass filtered images, observers necessarily cannot use very low-frequency content by squinting, moving, or otherwise blurring the appearance of the stimulus to interpret the image content. A high-pass filter (HPF) that removes low-frequency content from images with TS distortions produces the TS+HPF distortions.

### 3.2.2 Experiment 1: Subjective Data to Derive Relative Perceived Utility Scores

This experiment collected subjective data that was processed to derive relative perceived utility scores of distorted images formed from the same reference image. Distorted images of the *same* reference image but subjected to *different* distortions were compared using a paired comparison test methodology. The images compared were selected from reference/distortion sequences corresponding to the same reference image but different distortions. The comparisons of images with different distortions were used to align different reference/distortion sequences for the same reference image. For example, these comparisons allow the images from both an *airplane*/J2K+DCQ sequence and an *airplane*/TS sequence to be placed in relation to one another in terms of their relative perceived utility. For the same reference image, all reference/distortion sequences corresponding to each distortion were aligned, and these aligned sequences can be merged to form a single sequence of increasingly useful images that contain *all* distorted images of the *same* reference image.

#### Stimuli

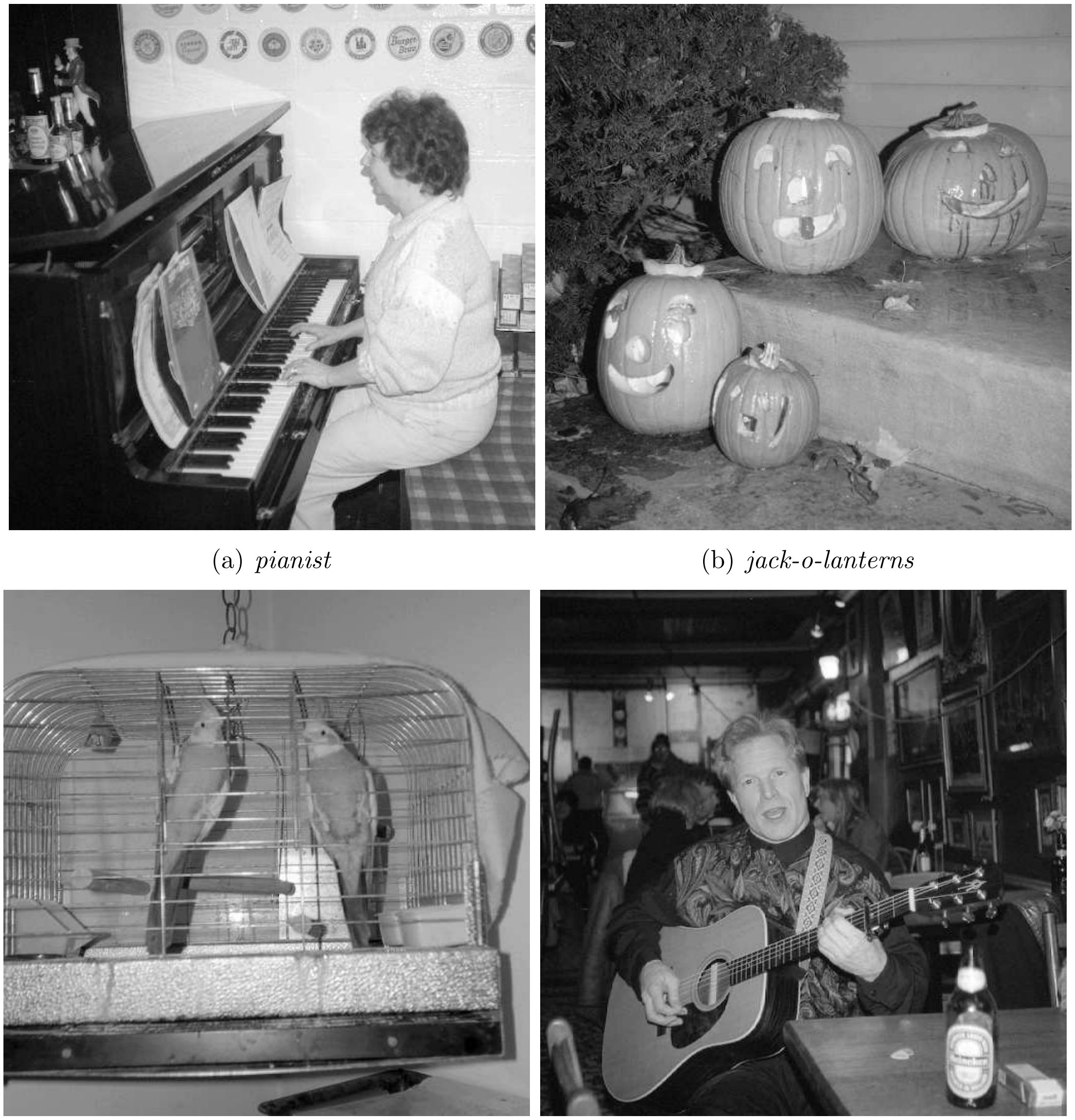
Nine grayscale natural images of size $512 \times 512$ pixels were cropped from original natural images and served as the reference images for these experiments. The content of the natural images consisted of either one or two main objects (e.g., an airplane or a boy and a cat) or a human in action (e.g., skiing or playing guitar). The nine natural images used in the experiments are shown in Figures 3.1(a), 3.3, and 3.4.



(a) *boy & cat* (b) *backhoe*

(c) *train* (d) *skier*

Figure 3.3: Four of the nine natural images serving as reference images for the experiments. Figure 3.4 contains the four other natural images serving as reference images. The reference *airplane* image is shown in Figure 3.1(a).



(a) *pianist*

(b) *jack-o-lanterns*

(c) *caged birds*

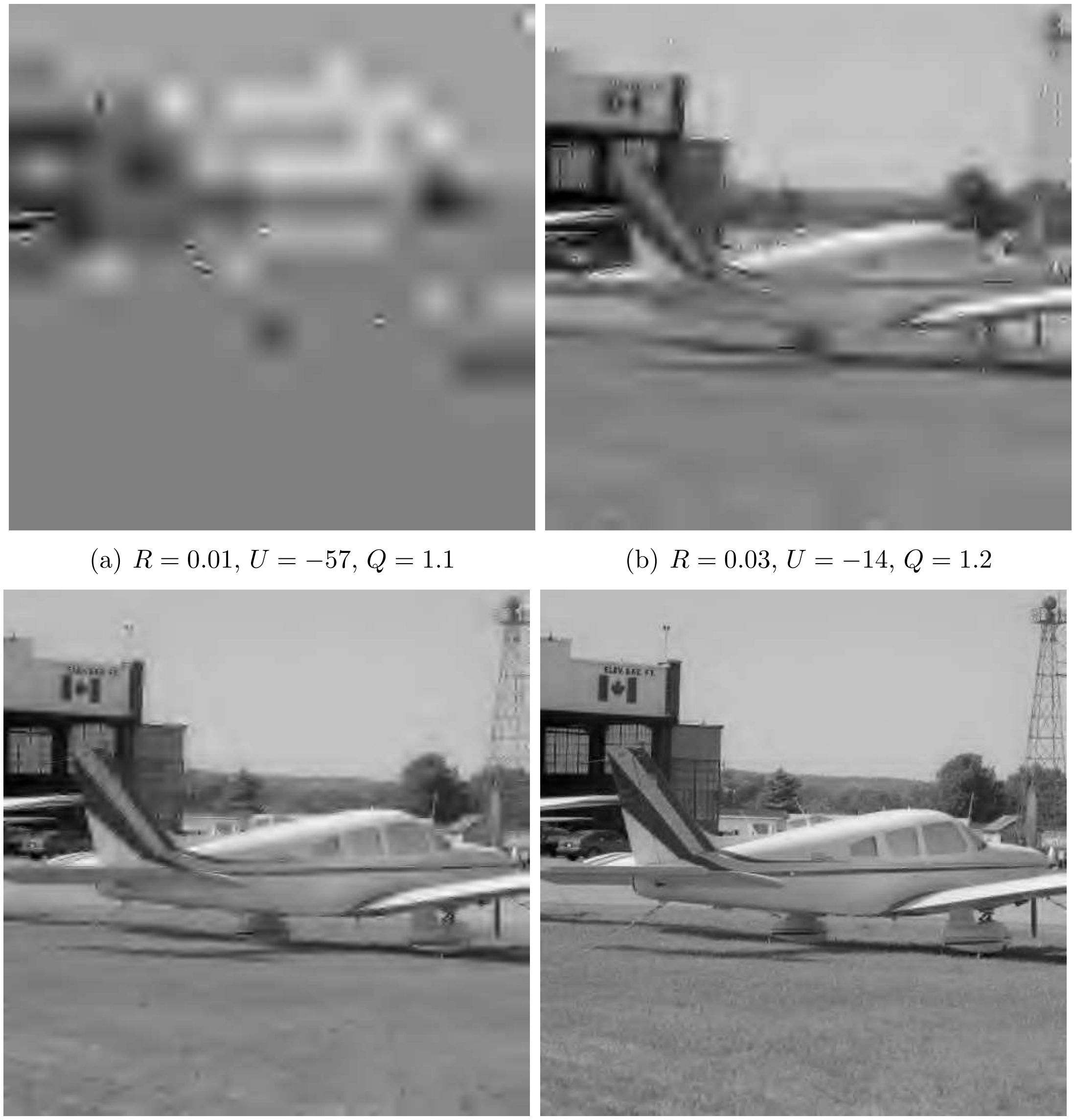
(d) *guitarist*

Figure 3.4: Four of the nine natural images serving as reference images for the experiments. Figure 3.3 contains the four other natural images serving as reference images. The reference *airplane* image is shown in Figure 3.1(a).

A collection of distorted images was formed by selecting a broad range of distortion levels from each reference/distortion sequence corresponding to each reference image and distortion. Specifically, images with JPEG distortions were formed using JPEG parameter values $P_{jpeg} = 1, 2, 5, 10, 20$, and 50. Images with BLOCK distortions were formed using quantization step-sizes $Q_{avg} = 400, 200$, and 1. Six images with J2K+DCQ distortions were formed using encoding bitrates logarithmically equally spaced from $R = 0.01$ to $R_{VL}$, where $R_{VL}$ denotes the bitrate of a visually lossless image formed using the DCQ strategy and JPEG-2000 compression. Four of the six images from the airplane/J2K+DCQ sequence are shown in Figure A.3. Images with TS and TS+HPF distortions were formed using smoothing parameters $\gamma = 2048, 446, 97, 21, 5$, and 1. The entire collection contained 243 distorted images.

### Procedure

A paired comparison testing methodology was used to collect subjective responses. Observers were asked to select an image from a pair of distorted images corresponding to the same reference image in response to the query "Which image tells you more about the content?" Most of the observers were Francophones, and for those observers, the query was presented in French as "Quelle est l'image qui donne le plus d'information sur le contenu de l'image ?" The distorted images in each pair correspond to the *same* reference image but *different* distortions (e.g., *airplane* with J2K+DCQ distortions and *airplane* with TS+HPF distortions). Each observer provided responses for a pair of images once. Certain pair comparisons were determined to be unnecessary based on responses collected in a preliminary experiment (e.g., comparing the most distorted image with J2K+DCQ distortions to the least distorted image with TS distortions), so the number of comparisons



(a) $R = 0.01$, $U = -57$, $Q = 1.1$

(b) $R = 0.03$, $U = -14$, $Q = 1.2$

(c) $R = 0.08$, $U = 15$, $Q = 1.8$

(d) $R = 0.23$, $U = 46$, $Q = 2.5$

Figure 3.5: Four images from the *airplane*/J2K+DCQ sequence used in Experiment 1 (Section 3.2.2). J2K+DCQ distorted images are parameterized using the encoding bitrate $R$ in bits per pixel (bpp) (see Table 3.1). The encoding bitrate of the visually lossless *airplane* image specified by the DCQ strategy is $R_{VL} = 1.85$ bpp. The perceived utility (U) scores and perceived quality (Q) scores obtained via the subjective experiments are provided for each image.

for each reference image was reduced.

Due to the large number of comparisons, the paired comparison tests were split into four testing sessions. Observers completed each session in approximately 30 minutes. Distorted images corresponding to the reference images *airplane*, *boy & cat*, *caged birds*, *guitarist*, and *train* were compared in the first two test sessions. J2K+DCQ, TS, and TS+HPF distorted images were included in the first session, and JPEG, BLOCK, TS, and TS+HPF distorted images were included in the second session. Both TS and TS+HPF distorted images appear in both sessions, so that the combined responses from each session also can be used to determine the relationship among J2K+DCQ distorted images and both BLOCK and JPEG distorted images via transitivity.

Distorted images corresponding to the reference images *backhoe*, *jackolanterns*, *pianist*, and *skier* were compared in the last two test sessions. The last two sessions were designed such that observers compared half of the distorted images in a single test session, and the distorted images in each session spanned the full range of distortion levels tested. All five types of distortions appeared in each of these last two test sessions.

### Observers

A total of 82 observers with verbally verified normal or corrected-to-normal acuity participated in the experiment over the four test sessions. Forty naive, Francophone observers participated in the first test session. An analysis of the results obtained from the first test session revealed that fewer observers would yield statistically equivalent results, so the remaining test sessions were conducted with fewer observers. In the second test session, ten naive, Francophone observers and ten expert, French- or English-speaking observers participated. Twenty-two naive, Francophone observers participated in the last two sessions with eleven observers per session.

### 3.2.3 Experiment 2: Recognition Thresholds of Natural Images

The experiment to estimate recognition thresholds for each of the nine reference images subjected to J2K+DCQ, TS, and TS+HPF distortions consisted of two parts. In the first part, observers called *writers* provided descriptions of the distorted images. In the second part, new observers called *readers* read these descriptions and decided which description indicated that the writer recognized the image content. Since writers typed their descriptions, response time is not a suitable indicator of recognition. The experimental methods used to estimate the recognition thresholds of the nine reference images are described.

**Stimuli**

To accurately estimate observer recognition thresholds of the reference images, reference/distortion sequences were constructed for each reference image using a dense set of distortion parameters for the J2K+DCQ, TS, or TS+HPF distortions. Reference/J2K+DCQ sequences contained 20 images corresponding to encoding bitrates $R$ that were logarithmically equally spaced from 0.01 to 0.30 bits per pixel. The choice of extremely low bitrates guarantees that unrecognizable images appear at the beginning of the sequence. Both reference/TS and reference/TS+HPF sequences contained 24 images corresponding to smoothing parameters $\gamma$ that were logarithmically equally spaced from 2048 to 1. The first image of a reference/TS sequence contains an image with only very low-frequency content, and the first image of a reference/TS+HPF sequence contains an image with a constant valued, gray image. With 9 reference images and 3 distortions, there are a total of 27 reference/distortion sequences.

### Part 1: Procedure to Collect Descriptions of Distorted Natural Images

In this part of the experiment, which is similar in design to that of Bruner and Potter [11], observers called *writers* viewed a distorted image and typed a brief description of the recognizable image content. The images that a writer viewed and described were ordered such that a writer cycled through each image of *one* reference/distortion sequence in order of decreasing distortion level. After completely viewing one reference/distortion sequence, the writer cycled through a new reference/distortion sequence corresponding to a *different* reference image and possibly a different distortion.

A writer necessarily viewed and described the images of at most 9 reference/distortion sequences, each sequence corresponding to a different reference image. The order that the reference/distortion sequences were presented to each writer was randomized. Participants completed this task in about 30 minutes.

### Part 2: Procedure to Identify Recognition Thresholds from Descriptions Collected in Part 1

In this part of the experiment, observers called *readers* who have not previously viewed the images read the descriptions produced by the writers.

This experiment consisted of consecutive trials. In each trial, a reader read all the descriptions provided by an unidentified writer for the images of a single reference/distortion sequence. The reference image corresponding to a reference/distortion sequence was simultaneously presented to the reader to compare with the descriptions, but information about the distortion viewed by the writer was hidden from the reader. The list of descriptions typed by a writer were ordered for the reader such that the first description corresponded to the first image of the reference/distortion sequence (i.e., an unrecognizable image), and the last description corresponded to the last image of the sequence. In each trial, the reader was instructed to select the first description that indicated the basic content of the reference natural image had been recognized. Trials were randomized for each reader.

This experiment was split into four sessions to alleviate observer fatigue. No time limit was imposed, and observers completed each session in approximately 30 minutes.

### Observers

A total of 49 observers with verbally verified normal or corrected-to-normal acuity participated in the experiments to estimate recognition thresholds for the nine reference images. Forty-six English-speaking observers (i.e., writers) participated in the experiment that collected descriptions of images in sequences corresponding to the different distortions. Nine to 13 observers viewed and described the distorted images in the reference/J2K+DCQ sequences for all nine reference images. Not all observers viewed a reference/J2K+DCQ sequence of images corresponding to each of the nine reference images. Twelve observers viewed and described the distorted images in the reference/TS and reference/TS+HPF sequences for all nine reference images. Three English-speaking observers (i.e., readers) participated in the experiment to identify recognition thresholds from writers' descriptions.

### 3.2.4 Perceived Utility Scores from Subjective Data

Perceived utility scores were obtained using the subjective data acquired in the two experiments described in Sections 3.2.2 and 3.2.3. The process to obtain perceived utility scores is described as three steps.

#### Relative Perceived Utility Scores from Subjective Data

Relative perceived utility scores were derived from the subjective data collected using the paired comparison test method (see Section 3.2.2). In particular, given two differently distorted images formed from the *same* reference image, the subjective data collected for the pair of images was used to estimate the actual probability that one distorted image is more useful to a human than the other.

Bradley and Terry specified a mathematical model that relates the probability that the response to stimulus $X_i$ is greater than the response to stimulus $X_j$ to a continuum of raw scale values that ranks the collection of stimuli $\{X_i\}_{i=1}^{n}$ according to some measure of merit [4]. This mathematical model was used to derive relative perceived utility scores (i.e., the raw scale values). For a reference image $X_{ref}$, let $X_i$ denote a distorted image formed from $X_{ref}$, and let $p_{ij}$ denote the probability that image $X_i$ conveys more information to a human about the content of $X_{ref}$ than image $X_j$. The Bradley-Terry model was used to map the estimates of $p_{ij}$, based on the subjective data, to relative perceived utility scores.

Distorted images subjected to the same distortion were not compared in the paired comparison test, because perceived utility is assumed to exhibit a monotonically, non-decreasing relationship as the distortion level decreased in the reference/distortion sequences. This assumption was imposed by explicitly defining the estimate of the probability $p_{ij}$ for two types of comparisons. First, for comparisons of an image with itself, the estimate of $p_{ii}$ was set to 0.5, since observers were expected to choose either image with equal probability. Second, for two different distorted images corresponding to the same reference/distortion sequence, the image with less distortion was assumed to have greater perceived utility than the image with more distortion. This second assumption was imposed by setting $p_{ij} = 0.99$ when image $X_i$ and $X_j$ belong to the same reference/distortion sequence (e.g., a JPEG distortion sequence) but the level of distortion for $X_i$ is less than that of $X_j$. The images used in the paired comparison test were broadly spaced in terms of the distortion level to accommodate this second assumption. For example, suppose $X_{R_1}$ and $X_{R_2}$ are two J2K+DCQ distorted images formed from the reference image using encoding bitrates $R_1$ and $R_2$, where $R_1 < R_2$. Because a larger encoding bitrate implies a lower level of distortion for J2K+DCQ distortions, the second assumption was imposed by setting $P(X_{R_2} > X_{R_1}) = 0.99$.

For each reference image, relative perceived utility scores for the corresponding set of distorted images were obtained from the estimates of $p_{ij}$ using a generalized linear model, which Critchlow and Flinger demonstrated is equivalent to the maximum-likelihood method used by Bradley and Terry [25]. The estimates of $p_{ij}$ were either generated from the subjective data or explicitly defined to impose the assumptions regarding the relationship among perceived utility and the distortion parameters for a single distortion. In addition to producing relative perceived utility scores, this data provides a mapping from each distortion parameter to the relative perceived utility scores for each reference image, which was used in the next step.

### Relative Perceived Utility Scores for the Recognition Threshold and the Reference Equivalence Class

The recognition threshold (RT) and the reference equivalence class (REC) of each reference image are used as anchors to map the relative perceived utility scores to the common utility scale (see Section 3.2.4). The estimates of the relative perceived utility scores for the RT and REC are described.

The subjective data from the second experiment (see Section 3.2.3) were used to estimate the relative perceived utility score coinciding with the RT of each reference image. The processed subjective data from the first experiment was used to construct mappings from each distortion parameter to the relative perceived utility scores. The RT for each reference/distortion sequence was estimated in terms of the corresponding distortion parameter based on the results from the experiments described in Section 3.2.3 (e.g., the RT for a J2K+DCQ sequence was specified in terms of the encoding bitrate $R$). The relative perceived utility score of the reference/distortion sequence's RT was found by linear interpolation using the mappings from each distortion parameter to the relative perceived utility scores. For a reference image, this yields several estimates of the relative perceived utility score for the RT, one corresponding to each distortion. The relative perceived utility score for the actual RT is estimated as the average of the relative perceived utility scores for the RT for each distortion.

Both the reference image and any visually lossless image belong to the REC. Thus, the relative perceived utility score coinciding with the minimum bitrate visually lossless image generated via JPEG-2000 compression using the dynamic contrast-based quantization (DCQ) strategy was used to define the relative perceived utility score of the REC (cf. Section 3.2.1). These visually lossless images were included in the paired comparison experiments, so the relative perceived utility scores of the REC of each reference image were directly estimated.

#### Perceived Utility Scores: Relative Perceived Utility Scores Mapped to a Common Utility Scale

Perceived utility scores were obtained by mapping the relative perceived utility scores to a common utility scale, where the recognition threshold (RT) was mapped to a perceived utility score of 0 and the reference equivalence class (REC) was mapped to a perceived utility score of 100. The relative perceived utility scores for the RT and the REC were used to define a linear mapping from relative perceived utility scores for the distorted images generated from the *same* reference image to perceived utility scores on the common utility scale.

## 3.3 Methods: Perceived Quality Scores

Human judgments of perceived quality generally indicate the perceptual resemblance of an image to a reference and are quantified by a *perceived quality score.* The reference is either 1) an explicit, external natural image that is presented to the observer or 2) an internal reference based upon observer expectations that is only accessible to the observer. Despite the vagueness of the term "quality," observers frequently attend to particular distortions (e.g., "blocky," "blurry," "sharp," etc.) to draw conclusions about the perceived quality [122].

Distorted natural images have been studied more often in the context of perceived quality than perceived utility, and several objective estimators have been developed to estimate perceived quality (see Section 2.3). The relationship between perceived quality and perceived utility is unclear; however, a poor quality image is expected to be less useful than an excellent quality image. If perceived quality accurately estimates perceived utility, then existing objective quality estimators should be suitable as utility estimators. Otherwise, those image characteristics that differentiate judgments of perceived quality from those of perceived utility need to be determined to properly design both quality and utility estimators robust to a variety of distortions.

An experiment was conducted to acquire perceived quality scores for the same images for which perceived utility scores were obtained to understand the relationship between quality and utility. The methods employed to acquire perceived quality scores are reported.

### 3.3.1 Stimuli

The nine reference images and the 243 distorted images formed from these reference images according to the methods described in Section 3.2.2 served as test stimuli in this experiment.

### 3.3.2 Procedure

The absolute category rating (ACR) [54] testing methodology [3] was used to collect perceived quality opinions of distorted images from human observers and consists of consecutive trials. In each trial, an observer was presented with a stimulus for 10 seconds. Then, the display was set to a constant gray background, and the observer was immediately requested to provide a opinion score that indicated his perceived quality of the previously displayed stimulus. The reference images were included in the test stimuli evaluated by the observer, and an observer was unaware if a stimulus was a distorted or reference image. The order of the stimuli presented was random and varied for each observer.

A discrete category rating scale was used that has five categories. Observers provide opinions of quality using the adjectives "Bad," "Poor," "Fair," "Good," and "Excellent" that define the quality categories. The observers participating in the experiment were Francophones; the rating scale respectively translated to French is "Mauvais," "Médiocre", "Assez Bon," "Bon," and "Excellent."

To alleviate observer fatigue due to prolonged evaluation sessions, the test was split into two sessions each containing roughly half of the stimuli. Observers completed each session in approximately 30 minutes and rested for five minutes between the two testing sessions.

[3] Numerical category scaling [30], adjective category scale [99], and categorical sort [60] are alternative names describing the absolute category rating (ACR) test method. The subjective assessment methodology for video quality (SAMVIQ) generally obtains more accurate perceived quality scores, but both ACR and SAMVIQ yield very similar perceived quality scores for our collection of distorted images [101].

### 3.3.3 Observers

Twenty-six naive, Francophone observers with verbally verified normal or corrected-to-normal acuity participated in the experiment, and one observer was rejected as an outlier according to criteria specified in the VQEG Multimedia Phase I report [134]. The 25 opinion scores from the remaining 25 observers were used to produce perceived quality scores for each stimulus.

### 3.3.4 Perceived Quality Scores from Subjective Data

Observers provided quality judgements that correspond to one of the five category levels (i.e., “Bad,” “Poor,” “ Fair,” “Good,” and “Excellent”). These five levels were mapped to the integers on the range 1 to 5 and yield observer opinion scores. The perceived quality score [4] for each test image was computed by averaging the corresponding observer opinion scores.

## 3.4 Results: Quality is not a proxy for Utility

The subjective data collected in Sections 3.2 and 3.3 provide perceived utility scores and perceived quality scores for a collection of distorted natural images. An analysis of the resulting relationship between the perceived quality scores and the perceived utility scores is reported and followed by a summary of the image characteristics that appear to influence human judgments of quality and utility, respectively, based on an analysis of the distortions. Example images that illustrate

[4] Prior work in the context of perceived quality often denotes a perceived quality score as a mean opinion score (MOS).

that quality is not a proxy for utility are then presented and discussed.

### 3.4.1 Relationship Between Quality and Utility

Perceived quality scores lie on the closed interval $\mathcal{Q} = [1, 5]$, whereas perceived utility scores lie on $\mathbb{R}$ with 0 denoting the recognition threshold and 100 denoting the reference equivalence class. Images with perceived utility scores less than 0 are unrecognizable and useless, and images with perceived utility scores greater than 100 are more useful than the reference image.

The relationship between quality and utility was analyzed only for those images whose perceived utility scores lie on the closed interval $\mathcal{U} = [-15, 115]$. No images had perceived utility scores greater than 115, but many images ($n = 80$) had perceived utility scores less than $-15$. Differences between perceived utility scores for images well below the recognition threshold convey less information about utility, since these values result from comparisons of two unrecognizable images. Furthermore, unrecognizable images were rated as having "Bad" quality: the perceived quality scores for these images have small standard deviation and both mean and median approximately equal to 1 [5]. Images whose perceived utility scores fall just below the recognition threshold were included, because Bruner and Potter reported that human observers, especially adults, tend to maintain incorrect hypotheses about the actual content when viewing reference/distortion sequences beginning with a very distorted, unrecognizable images as compared to observers that first view a reference/distorted sequence beginning with a less distorted image [11]. Our experiments to estimate recognition thresholds had ob-

[5] The perceived quality of unrecognizable images with perceived utility scores less than $-15$ range from 1 to 1.4 with the average, standard deviation, and median being 1.07, 0.089, and 1.04, respectively.

servers first view very distorted unrecognizable images in the reference/distortion sequences, so including images whose perceived utility scores lie on the interval $[-15, 0]$ accounts for possible overestimates of the recognition thresholds due to the phenomenon reported by Bruner and Potter.

To test whether quality is a robust proxy for utility, both correlation and accuracy statistics were used. Specifically, quality is not a robust proxy for utility if 1) perceived quality scores and perceived utility scores are weakly correlated and 2) perceived quality scores inaccurately estimate perceived utility scores. The Pearson linear correlation $r$, the Spearman rank correlation $\rho$, and the Kendall rank correlation $\tau$ are used to quantify the relationship between perceived quality scores and perceived utility scores [118]. The rank correlation measures, the $\rho$ and $\tau$, quantify the discrepancies between the rank order of the two sets of subjective scores. Neither $\rho$ nor $\tau$ are affected by a monotonic, nonlinear mapping.

The root mean squared error (RMSE) and the outlier ratio (OR) were chosen to quantify the accuracy with which perceived quality scores estimate perceived utility scores. The RMSE was computed after fitting the perceived quality scores and the perceived utility scores to a monotonic, nonlinear mapping (cf. Eq. (3.3)). The OR is the proportion of nonlinearly mapped quality scores (i.e., the utility score estimated from quality) that lie outside the 95% confidence interval of the perceived utility score.

Monotonic nonlinear functions were fitted to the subjective scores and used to map perceived quality scores to the utility range, since perceived quality exhibits a nonlinear relationship with perceived utility (cf. Figure 3.6). Let $\mathcal{Q} = [1, 5]$ denote the domain of the quality range, and let $\mathcal{U} = [-15, 115]$ denote the domain of the utility range. Let $q_i$ and $u_i$ respectively denote the perceived quality score and perceived utility score of image $i$. The nonlinear function $f : \mathcal{Q} \rightarrow \mathcal{U}$ given as

$$f(q) = a \log(q) + b \tag{3.3}$$

maps perceived quality scores to the utility range, and the parameters $\{a, b\}$ were found by minimizing the sum of the squared error based on the residuals $\{f(q_i) - u_i\}_{i=1}^{n}$, where $n$ is the number of images with both perceived quality and perceived utility scores. The fit was considered sufficient if the residuals exhibit a Gaussian distribution. The Jarque-Bera normality test determines if a collection values come from an unspecified Gaussian distribution [57], was applied to the set of residuals $\{f(q_i) - u_i\}_{i=1}^{n}$, and concluded that they did come from an unspecified Gaussian distribution at the 95% confidence level.

The two scatter plots in Figures 3.6 and 3.7 illustrate the nonlinear relationship between quality and utility for the nine reference images and five distortions with perceived utility indicated on the left ordinate. In each scatter plot, the quality adjectives delineating the quality rating scale have been provided on the top abscissa, and the two anchors, the recognition threshold (RT) and the reference equivalence class (REC), associated with perceived utility are indicated on the right ordinate. The symbols in the Figures 3.6 and 3.7 distinguish subjective scores according to the reference image and the distortion, respectively. The solid curve in each figure corresponds to the fitted nonlinear mapping from the abscissa to the ordinate (i.e., Eq. (3.3)), and the dashed curves define the 95% prediction interval (PI) for the fitted nonlinear mapping.

The nonlinear relationship between utility and quality indicates that the quality of a test image generally does not accurately predict its usefulness. The slope of the nonlinear relationship between utility and quality is positive and decreases with increasing quality, which indicates that variations in quality correspond to smaller

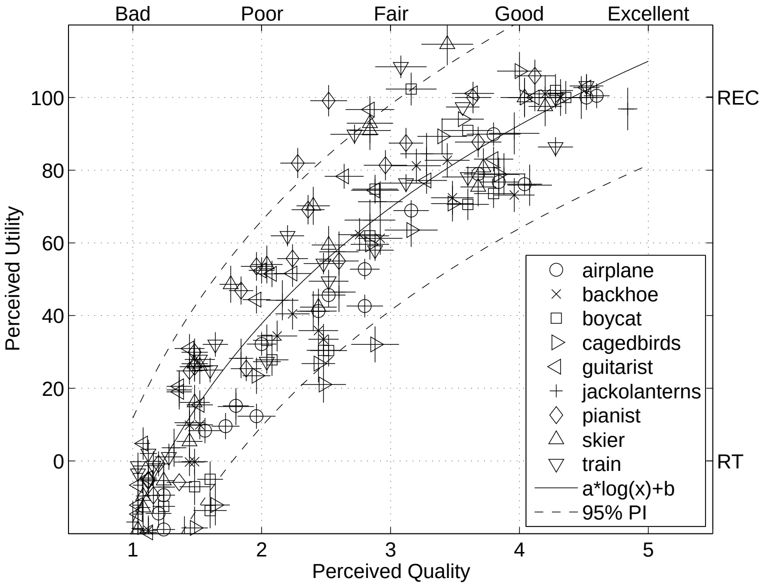


Figure 3.6: Quality is not a suitable proxy for utility. The scatter plots show the relationship between perceived utility scores and the perceived quality scores for nine reference images and five distortions (cf. Figure 3.1). The symbols indicate the reference image corresponding to each subjective score. The recognition threshold (RT) and the reference equivalence class (REC) are denoted on the axis corresponding to perceived utility scores. The quality adjectives are denoted on the axis corresponding to the perceived quality scores. Standard error bars have been included for both subjective scores. In each figure, the fitted nonlinear mapping from the abscissa to the ordinate is denoted by the solid curve, and the 95% prediction interval (PI) for the fitted nonlinear mapping is denoted by the dashed curves. See also Figure 3.7.

variations in utility as quality increases. For example, there are test images rated as having perceived quality ranging from “Fair” to “Excellent” that have high perceived utility.

The relationship between quality and utility was analyzed for the entire collection of distorted images as well as subsets of the collection that were formed by treating both 1) quality and 2) distortion type as factors. The quality range spans the interval $[1, 5]$, and three “levels” of the quality factor were defined: low

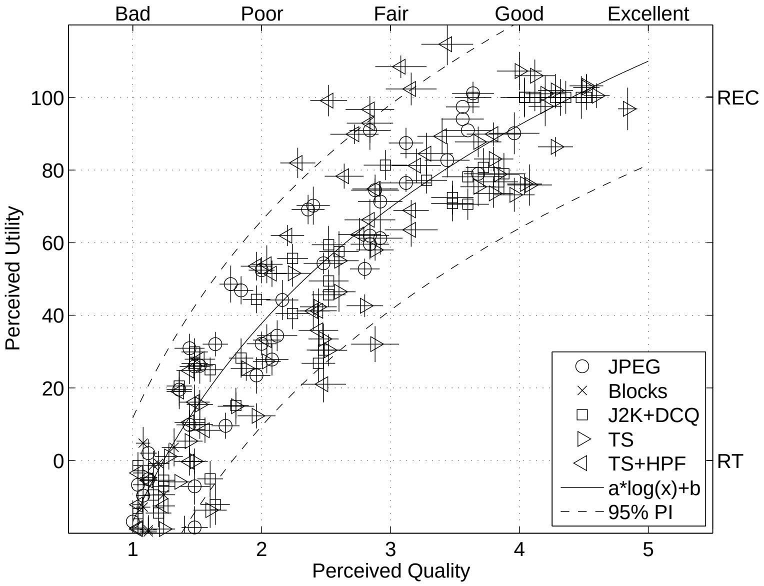


Figure 3.7: Perceived utility versus perceived quality where the symbols indicate the distortion corresponding to each subjective score. See caption of Figure 3.6.

quality $[1, 2.25)$, medium quality $[2.25, 3.75]$, and high quality $(3.75, 5]$. Subsets of distorted images spanning different regions of quality were analyzed, because the distorted images used in the experiment span distortion levels ranging from unrecognizable to visually lossless. The five distortion types correspond to the "levels" of the distortion type factor: JPEG, BLOCKS, J2K+DCQ, TS, and TS+HPF. Subsets of distorted images corresponding to different distortion types were analyzed, because each distortion type disrupts different image characteristics.

Statistical differences in either correlation or accuracy among the different levels of a factor (i.e., quality region or distortion type) preclude a reliable predictive relationship between perceived quality and perceived utility. Statistical differences between two correlation values were determined using a $z$-test after applying the Fisher transformation to the correlation values [32, 41]. Statistical differences between accuracy statistics were identified by analyzing the squared errors

$\{(f(q_i) - u_i)^2\}_{i=1}^n$ using a one-way analysis of variance (ANOVA) to determine if any of the mean squared errors statistically differ for a particular factor [32]. If ANOVA indicated that the accuracy differed according to a particular factor, then Tukey's multiple comparison procedure was used to identify which levels (e.g., high quality or J2K+DCQ) of that factor had statistically different mean squared errors. The comparison results are reported as $p$-values, where $p$-values greater than 0.05 indicate that at the 95% confidence level the mean squared errors differ among the two levels of the factor that are compared. The outlier ratio (OR) is a binomial random variable, and statistical differences between two OR values are determined via a $z$-test at the 95% confidence level using the Gaussian approximation of a binomial random variable [32].

Table 3.2 summarizes the correlation and accuracy statistics for all images and subsets of distorted images when either the quality region or the distortion is considered as a factor. The monotonic, nonlinear mapping (i.e., Eq. (3.3)) affects the Pearson linear correlation between the subjective scores. The Pearson linear correlation computed before applying the nonlinearity is denoted $r$, and it is denoted $r_{fit}$ when computed after applying the nonlinearity. For each statistic, values in boldface are statistically greater than those of the other levels within that factor. The remainder of this section summarizes key observations, which appear in boldface, followed by statistical justifications and interpretations.

**Quality does not consistently and accurately predict utility for different regions of quality.** The *entire* collection of distorted images range from unrecognizable to visually lossless, and a strong global correlation is observed, which implies that a poor quality image is less useful than an excellent quality image. However, the 95% prediction interval for the fitted nonlinear mapping between

Table 3.2: Results summarizing the relationship between perceived quality and perceived utility. Each row corresponds a subset of $n$ images either spanning a particular range of quality or corresponding to a particular distortion. The Pearson linear correlation $r$, the Spearman rank correlation $\rho$, and the Kendall rank correlation $\tau$ are computed between the perceived quality and perceived utility scores. The root mean squared error (RMSE) and the outlier ratio (OR) were computed using the utility scores and the mapped (i.e., Eq. (3.3)) quality scores. $r_{fit}$ denotes the Pearson linear correlation after applying the mapping. For the correlation statistics and OR, boldface values are statistically equivalent to the largest value for a subset of images (excluding All). Boldface RMSE values are statistically larger than the other subsets based on ANOVA.

| *Factor* | *Image Subset* | $n$ | $r$ | $\rho$ | $\tau$ | RMSE | $r_{fit}$ | OR |
|---|---|---|---|---|---|---|---|---|
| | All | 163 | 0.909 | 0.919 | 0.750 | 14.2 | 0.925 | 0.58 |
| *Quality Region* | Low Quality | 72 | **0.819** | **0.791** | **0.606** | 12.5 | **0.812** | **0.60** |
| | Med. Quality | 63 | 0.620 | 0.625 | 0.458 | **17.5** | 0.627 | **0.68** |
| | High Quality | 28 | 0.603 | 0.583 | 0.402 | 9.28 | 0.614 | 0.32 |
| *Distortion* | JPEG | 39 | **0.931** | **0.938** | **0.795** | 11.9 | **0.939** | **0.59** |
| | BLOCKS | 6 | 0.228 | 0.116 | 0.138 | 7.71 | 0.221 | 0.17 |
| | J2K+DCQ | 42 | **0.953** | **0.953** | **0.825** | 11.9 | **0.955** | 0.48 |
| | TS | 38 | **0.964** | **0.934** | **0.769** | **14.3** | **0.957** | **0.58** |
| | TS+HPF | 38 | 0.884 | 0.868 | **0.690** | **18.6** | 0.894 | **0.66** |

utility and quality (i.e., Figure 3.6) indicates that a perceived quality score corresponds to a broad range of perceived utility scores, and the range of the perceived utility scores varies for different regions of quality (e.g., the prediction interval is wider in the medium quality region than the low quality region). An analysis of the relationship between the perceived utility scores and the perceived quality scores for individual quality regions provides more insight into the relationship between quality and utility.

For different quality regions, both the correlation and accuracy between the perceived utility scores and the nonlinearly mapped perceived quality scores vary. The perceived utility scores and perceived quality scores exhibit the most linear relationship ($r = 0.82$) for images with low quality (i.e., rated as having either "Bad" or "Poor" perceived quality). Variations in perceived quality scores explain 67% (i.e., $100r^2$%) of the variation in perceived utility scores in this quality region. However, for the other quality regions, the correlation between perceived utility scores and perceived quality scores is statistically significantly smaller ($r < 0.62$), which indicates that variations in the perceived quality scores explain no more than 40% of the variation in the perceived utility scores in the medium and high quality regions.

The quality region was found to be a factor that influences the squared errors between the perceived utility scores and the nonlinearly mapped perceived quality scores based on a one-way ANOVA ($F(2, 160) = 7$, $p < 0.01$). The mean squared error between the perceived utility scores and the mapped perceived quality scores for distorted images in the medium quality region is statistically larger than that of the other two quality regions ($p \leq 0.01$).

The significant variation in both the correlation and accuracy statistics for different regions of quality demonstrate that quality does not generally provide a reliable estimate of utility. The observed relationship between quality and utility is discussed for each quality region.

*Variations in quality for distorted images in the low quality region largely coincide with variations in utility.* The slope of the overall relationship between utility and quality decreases as quality increases and is steepest within the low quality region, which indicates that small changes in perceived quality in the low quality region affect perceived utility more than small changes in quality for other regions of quality. Consider, for example, a reference/distortion sequence beginning with an unrecognizable image and evolving toward a useful image with medium perceived quality. Subsequent images in the sequence will contain less distortion than the previous images, and the sequence will evolve from unrecognizable to recognizable within the low quality region. The strong correlation ($r = 0.82$) as well as the steep slope between utility and quality within this region reflect the dramatic perceptual changes coinciding with the evolution of images from unrecognizable to recognizable in this sequence. In other words, the observed relationship between quality and utility in the low quality region suggests that observers largely judge lower quality images in terms of their ability to interpret the content.

*Distorted images in the medium quality region are useful but visibly distorted and nearly span the full range of utility* [21, 115]. Twenty percent of the distorted images in the medium quality region have very high utility (i.e., perceived utility scores greater than 90) and span nearly the entire range of the medium quality region $[2.5, 3.7]$. This clearly demonstrates that high utility does not necessarily imply high quality, since these images all have medium quality. Therefore, very useful images can contain a moderate amount of visible distortions (i.e., have medium quality). Further analysis revealed that most of the images with medium quality and high utility are TS+HPF distorted images, which suggests that removing low-frequency content can form a perceptually different image (i.e., decrease quality) without affecting the image's usefulness.

*Distorted images in the high quality region contain few visible distortions and* *span a narrow range of utility* [73, 108]. In addition, more than 60% of the distorted images have very high utility (i.e., perceived utility scores greater than 90) with quality as low as 4 (i.e., "Good" quality). Furthermore, both low correlation with and low RMSE between the perceived utility scores and the nonlinear mapped perceived quality scores was observed for distorted images in the high quality region. In other words, as the level of distortion decreases utility saturates before quality saturates, and refinements in quality for high quality images have little effect on utility.

The interpretation of the relationship between utility and quality must be qualified with respect to the natural images used in the experiments. In particular, the usefulness of the natural images was determined by an object or objects that generally occupy a large portion of the image, which led to useful images despite the presence of visible of distortions (i.e. images in the medium quality region). Had the usefulness of the images been dictated by either a smaller or less conspicuous object (e.g., recognition of the flower pot in the *boy & cat* image), the relationship between utility and quality could differ. For example, image usefulness dictated by a smaller, inconspicuous object is expected to require a higher quality image than if the usefulness is dictated by a larger, conspicuous object. Such variations in image usefulness reflect tasks that repurpose the original intent of the images. In this paper, the task was to report the content of each natural image, and the content of the images selected for the experiment is dictated by one or two conspicuous objects.

**Utility is not accurately estimated using quality for TS+HPF distorted images.** Both the accuracy with which perceived utility scores are estimated from mapped perceived quality scores as well as the correlation between the perceived utility scores and the perceived utility scores varies among the different distortion types [6]. The squared errors between the perceived utility scores and the mapped perceived quality scores were influenced by the distortion type factor based on a one-way ANOVA ($F(4, 158) = 3.39$, $p = 0.01$). The mean squared errors for estimates of perceived utility scores from perceived quality scores for TS+HPF distortions were found to be statistically larger than those for JPEG ($p = 0.02$) and J2K+DCQ ($p = 0.02$) distortions but not TS distortions ($p = 0.23$) .

TS+HPF distortions disrupt both high-frequency content via texture smoothing and low-frequency content via high-pass filtering, whereas JPEG, J2K+DCQ, and TS distortions primarily disrupt high-frequency content before low-frequency content. The perceived utility scores exhibit very strong correlation ($r > 0.93$) with the perceived quality scores for the JPEG, J2K+DCQ, and TS distorted images, and the highest correlation is observed for the TS distorted images ($r = 0.96$). The very strong correlation between the perceived utility scores and the perceived quality scores for JPEG, J2K+DCQ, and TS distorted images indicates that distortions to high-frequency content affect both utility and quality. However, the correlation between the perceived utility scores and the perceived quality scores is statistically lower for the TS+HPF distorted images than the TS distorted images ($p = 0.01$), yet the TS+HPF distorted images only lack the low-frequency content of the TS distorted images. The weak correlation as well as the large RMSE between the perceived utility scores and the mapped perceived quality scores for TS+HPF distorted images indicate that distortions to low-frequency content affect utility differently than they affect quality.

[6] Only six BLOCK distorted images have perceived utility scores greater than $-15$, so results corresponding to the BLOCK distorted images provide little insight into the relationship between quality and utility. Furthermore, these images have perceived quality scores in the range $[1, 1.3]$ (i.e., “Bad” quality) and perceived utility scores in the range $[-13, 4]$ (i.e., effectively useless).

Overall, the analysis of the relationship between utility and quality demonstrate that an image with low quality also has low utility, and an image with high quality also has high utility. However, distorted images with quality in the medium region correspond to a wide range of perceived utility scores, including high utility. In other words, high utility does not imply high quality. The perceived utility scores of TS+HPF distorted images are less accurately estimated from the perceived quality scores than for the other distortions, especially when the TS+HPF distorted image has quality in the medium region and suggests that low-frequency content affects quality differently than utility.

### 3.4.2 Effects of Low-frequency Content on Quality and Utility

JPEG, BLOCKS, J2K+DCQ, and TS distortions largely disrupt high-frequency content with limited disruption to low-frequency content. However, TS and TS+HPF distorted images with the same smoothing parameter $\gamma$ only differ with regard to the inclusion of low-frequency content. The perceived utility scores and perceived quality scores for TS and TS+HPF distorted images were compared to determine the influence of low-frequency content on both utility and quality.

For each reference image, the subjective scores for TS and TS+HPF distorted images with equal smoothing parameters $\gamma$ are tested for statistical differences when $\gamma = 1, 5, 21, 97, 446$, and 2048. Statistical differences in the subjective scores imply that the disruption to low-frequency content influences the subjective scores. For TS and TS+HPF distorted images formed from the same reference image using smoothing parameter $\gamma$, let $S_{\mathrm{TS}(\gamma)}$ and $S_{\mathrm{TS+HPF}(\gamma)}$ denote the subjective scores, respectively, and let $\sigma_{S_{\mathrm{TS}(\gamma)}}$ and $\sigma_{S_{\mathrm{TS+HPF}(\gamma)}}$ respectively denote the standard deviation of $S_{\mathrm{TS}(\gamma)}$ and $S_{\mathrm{TS+HPF}(\gamma)}$. $Z$-tests were used to determine if two scores are statistically different using the test statistic

$$z_{stat} = \frac{S_{\mathrm{TS}(\gamma)} - S_{\mathrm{TS+HPF}(\gamma)}}{\sqrt{\sigma^2_{S_{\mathrm{TS}(\gamma)}} + \sigma^2_{S_{\mathrm{TS+HPF}(\gamma)}}}}. \tag{3.4}$$

The results of the $z$-test are reported as the confidence that $S_{\mathrm{TS}(\gamma)}$ is greater than $S_{\mathrm{TS+HPF}(\gamma)}$ (i.e., $P(z \leq z_{stat})$, where $z$ is a zero mean Gaussian random variable with unit variance) and is denoted as $Conf(S_{\mathrm{TS}(\gamma)} > S_{\mathrm{TS+HPF}(\gamma)}) \in [0, 1]$.

Figures 3.8 and 3.9 show $Conf(S_{\mathrm{TS}(\gamma)} > S_{\mathrm{TS+HPF}(\gamma)})$ as a function of the perceived quality score and the perceived utility score of a TS distorted image, respectively. Values of $Conf(S_{\mathrm{TS}(\gamma)} > S_{\mathrm{TS+HPF}(\gamma)})$ less than 0.025 and greater than 0.975 indicate that the subjective scores for TS and TS+HPF distorted images with equal $\gamma$ are statistically different at the 95% confidence level (i.e., a two-sided $z$-test). Values of $Conf(S_{\mathrm{TS}(\gamma)} > S_{\mathrm{TS+HPF}(\gamma)})$ less than 0.05 indicate that the subjective score for the TS distorted image is statistically smaller than the subjective score for a TS+HPF distorted image formed from the same reference image using the same $\gamma$ at the 95% confidence level (i.e., a one-sided $z$-test). Similarly, values of $Conf(S_{\mathrm{TS}(\gamma)} > S_{\mathrm{TS+HPF}(\gamma)})$ greater than 0.95 indicate that the subjective score for the TS distorted image is statistically greater than the subjective score for a TS+HPF distorted image with the same $\gamma$. Key observations appear in boldface and are followed by a statistical justification and interpretation.

**For the same reference image, a TS+HPF distorted image never is of higher quality than a TS distorted image with the same** $\gamma$. Over all levels of quality, loss of low-frequency content led to an average decrease in perceived quality of 0.53, and in most cases, the perceived quality of a TS distorted image is statistically greater than that of a TS+HPF distorted image formed from the

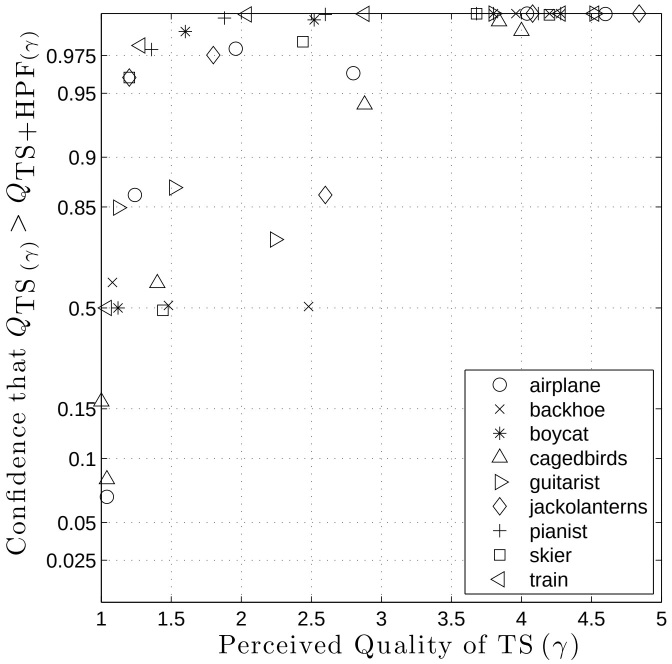


Figure 3.8: Perceived quality either decreases or remains the same when low-frequency content is disrupted (i.e., for TS+HPF distortions relative to TS distortions). The figures show the confidence that the perceived quality (Q) score of the TS distortions are greater than the perceived quality score for TS+HPF distortions with equal $\gamma$ as a function of the perceived quality score of the TS distortions. Confidence values either in the interval $[0, 0.025]$ or $[0.975, 1]$ indicate that the perceived quality scores are statistically different at the 95% confidence level. Confidence values greater than 0.95 (less than 0.05) indicate that the perceived quality score for the TS distorted image is greater than (less than) the perceived quality score for a TS+HPF distorted image at the 95% confidence level. $Q_{\mathrm{TS}(\gamma)}$ denotes the perceived quality $Q$ of a TS distorted image using smoothing parameter $\gamma$, and $Q_{\mathrm{TS+HPF}(\gamma)}$ is similarly defined for a TS+HPF distorted image. See Section 3.4.2 for additional details regarding the confidence analysis and its interpretation.

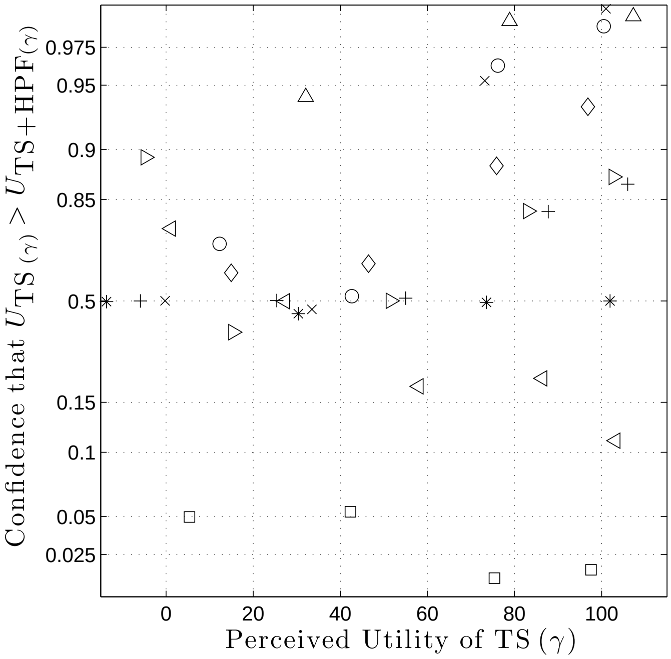
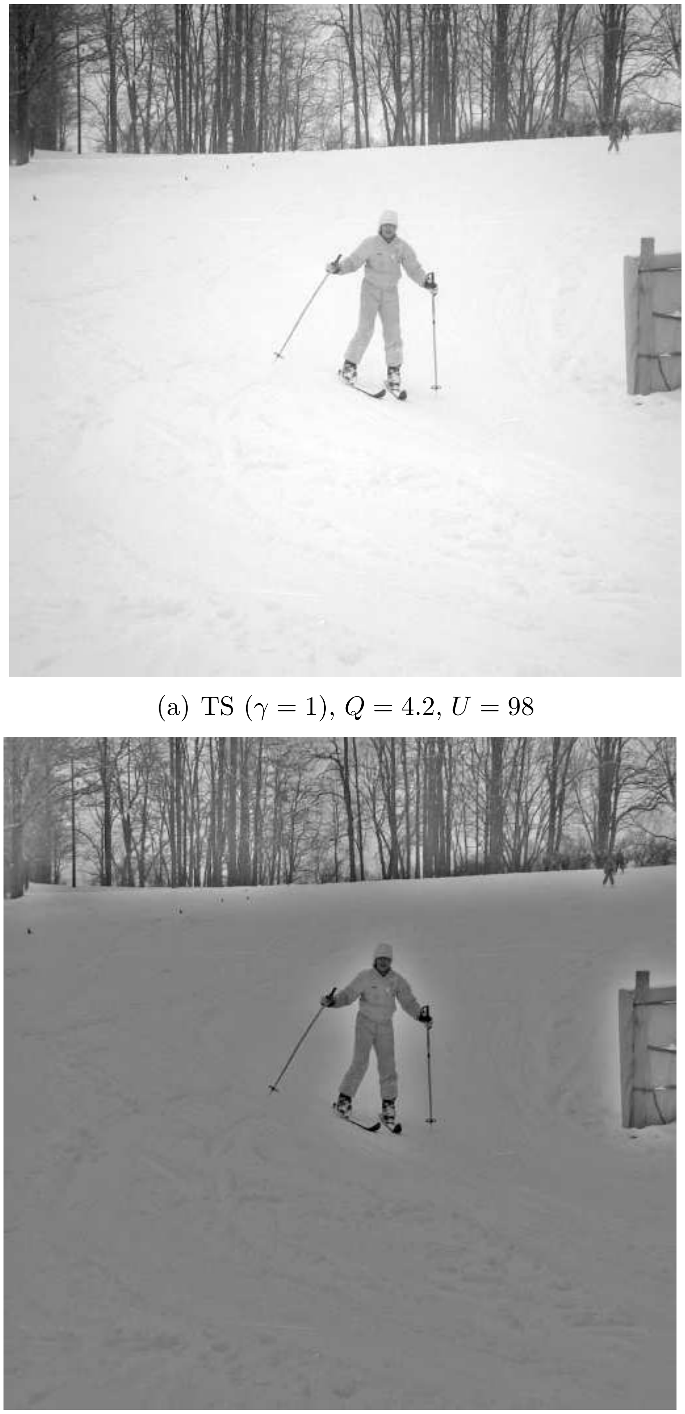


Figure 3.9: Disruptions to low-frequency content do not affect the perceived utility of most images. The figures show the confidence that the perceived utility (U) score of the TS distortions are greater than the perceived utility score for TS+HPF distortions with equal $\gamma$ as a function of the perceived utility score of the TS distortions. Confidence values either in the interval $[0, 0.025]$ or $[0.975, 1]$ indicate that the perceived utility scores are statistically different at the 95% confidence level. Confidence values greater than 0.95 (less than 0.05) indicate that the perceived utility score for the TS distorted image is greater than (less than) the perceived utility score for a TS+HPF distorted image at the 95% confidence level. $U_{\mathrm{TS}(\gamma)}$ denotes the perceived utility $U$ of a TS distorted image using smoothing parameter $\gamma$, and $U_{\mathrm{TS+HPF}(\gamma)}$ is similarly defined for a TS+HPF distorted image. Refer to the legend of Figure 3.8.

same reference image using the same $\gamma$. For some images, the perceived quality of a TS and TS+HPF distorted image with equal $\gamma$ are statistically equivalent but only when the perceived quality of the TS distorted image is less than 3 (i.e., the quality is "Fair" or worse). In short, because poorer quality images are very heavily distorted, additional distortions that affect the low-frequency content of poorer quality images have little influence on the perceived quality.

**The relationship between the utility of TS and TS+HPF distorted images with the same $\gamma$ formed from the same reference image varies for each reference image.** For many of the reference images, disruptions to low-frequency content (i.e., TS and TS+HPF distorted images with equal $\gamma$) do not affect perceived utility. However, disruptions to the low-frequency content of the images *skier*, *airplane*, *backhoe*, and *caged birds* did affect utility when the TS distorted image has high utility (i.e., perceived utility score greater than 70).

The image *skier* has a statistically greater perceived utility score when low-frequency content is disrupted (i.e., for TS+HPF distorted images) than when the low-frequency content is not disrupted (i.e., the TS distorted images). Moreover, a *skier* TS+HPF distorted image with medium quality has a perceived utility score statistically greater than 100: this image is more useful than the reference image. Removing the low-frequency content from *skier* introduces "halos" near edges that enhance the visibility of the skier and other objects (see Figure 3.10). The increased visibility of the skier could explain why removing the low-frequency content (i.e., a TS distorted image versus a TS+HPF distorted image with the same $\gamma$) increased the perceived utility. However, the observer responses do not indicate what criteria the observers used to choose the TS+HPF distorted image over the TS distorted image (see Section 3.5).



(a) TS ($\gamma = 1$), $Q = 4.2$, $U = 98$

(b) TS+HPF ($\gamma = 1$), $Q = 3.4$, $U = 115$

Figure 3.10: Example showing that the *skier* TS distorted image has statistically greater quality than the TS+HPF distorted image with equal $\gamma$ but statistically lower utility. Removing the low-frequency content from *skier* (i.e., the TS+HPF distorted image) introduces "halos" near edges that enhance the visibility of the skier. See also Figures 3.8 and 3.9.

Among TS distorted images with high utility (i.e., greater than 70), the perceived utility scores of the images *airplane*, *backhoe*, and *caged birds* were statistically smaller for TS+HPF distorted images than TS distorted images for the same $\gamma$. Because a paired comparison test methodology without ties was used, observers were forced to choose one of the images in each pair presented. The binary responses collected from observers to obtain perceived utility scores preclude a definitive explanation for why the TS distorted images were chosen over TS+HPF distorted images, but there are two possible explanations for this result:

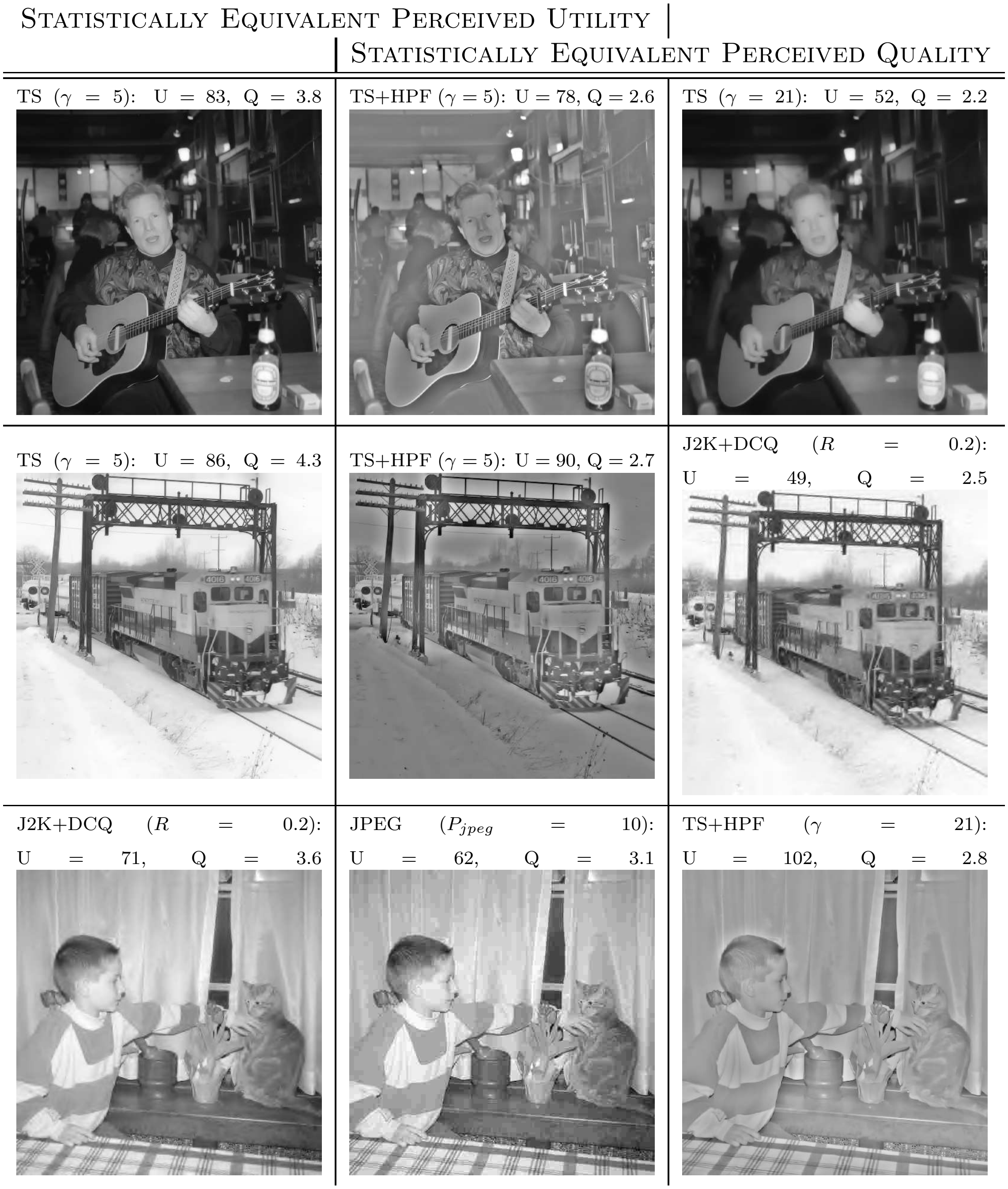
- *Relative to the TS+HPF distorted images, the low-frequency content of TS distorted images may convey useful information about the content to observers.* For example, in the *airplane* image, the removal of the low-frequency content darkens many regions of the image (e.g., the sky and the airplane). The sky similarly darkens in the *backhoe* image when low-frequency content is removed. These perceptual differences may cue different interpretations about the scene to observers, and the interpretation for the TS distorted image appears more accurate. The appearance of the specular reflections of the bird cage, which may provide an observer with information about the brightness of the room, are reduced in the *caged birds* TS+HPF image relative to its TS distorted version. Such features correspond to additional information about the image content beyond the visibility of the objects' spatial details, which would be primarily conveyed by high-frequency content (e.g., edges).
- *Observers may have found both TS and TS+HPF distorted images formed from the same reference using the same $\gamma$ equally useful and more often reverted to judgments of quality to choose an image.* This would suggest that quality is a secondary criteria to utility. In other words, given images with equal utility, observers generally preferred the higher quality TS distorted

image, except when the lower quality TS+HPF distorted image conveyed sufficiently more information about the content (e.g., the *skier* image). For many of the reference images, the values of $Conf(S_{\text{TS}(\gamma)} > S_{\text{TS+HPF}(\gamma)})$ show evidence of a slight, though not statistically significant, bias toward observers choosing the TS distorted image over the TS+HPF distorted image with equal $\gamma$.

We conjecture that the second explanation (i.e., observers revert to quality judgements) is more plausible; however, different observers may have used different criteria to make a decision (see Section 3.5).

### 3.4.3 Examples illustrating that quality is not a proxy for utility

The analysis of the relationship between perceived utility scores and perceived quality scores demonstrates that quality does not accurately predict utility, and Figure 3.11 illustrates several cases when the relationship between two distorted images based on quality does not reflect the relationship between those two images in terms of utility and vice versa. Each row of Figure 3.11 corresponds to a different reference image, and for each row the images are arranged such that 1) the distorted image on the left and the distorted image in the middle have statistically equivalent perceived utility scores but statistically different perceived quality scores and 2) the distorted image in the middle and the distorted image on the right have statistically equivalent perceived quality scores but statistically different perceived utility scores.



Statistically Equivalent Perceived Utility

Statistically Equivalent Perceived Quality

TS ($\gamma = 5$): U = 83, Q = 3.8

TS+HPF ($\gamma = 5$): U = 78, Q = 2.6

TS ($\gamma = 21$): U = 52, Q = 2.2

TS ($\gamma = 5$): U = 86, Q = 4.3

TS+HPF ($\gamma = 5$): U = 90, Q = 2.7

J2K+DCQ ($R = 0.2$): U = 49, Q = 2.5

J2K+DCQ ($R = 0.2$): U = 71, Q = 3.6

JPEG ($P_{jpeg} = 10$): U = 62, Q = 3.1

TS+HPF ($\gamma = 21$): U = 102, Q = 2.8

Figure 3.11: Differences in perceived quality (Q) do not imply differences in perceived utility (U). In terms of perceived utility, the distorted images in the middle column are statistically equivalent to the distorted images in the left column. However, in terms of perceived quality the distorted images in middle column are statistically equivalent to the distorted images in the right column.

The first two rows of the first two columns in Figure 3.11 illustrate the relationship between TS and TS+HPF distorted images. The texture smoothing parameter $\gamma$ must be increased (i.e., increasing the level of texture smoothing) for a TS distorted image to exhibit the same perceived quality observed as a TS+HPF distorted image, but the resulting TS distorted image will have lower perceived utility than the TS+HPF distorted image (first row of Figure 3.11). Similarly, a J2K+DCQ distorted image that exhibits the same perceived quality as a TS+HPF distorted image also has lower perceived utility (second row of Figure 3.11). In other words, high-frequency content must be disrupted to form a distorted image with equal quality to an image that lacks low-frequency content.

The last row of Figure 3.11 contains three images that respectively have J2K+DCQ, JPEG, and TS+HPF distortions. High-frequency content is disrupted for both J2K+DCQ and JPEG distorted images with limited disruption to low-frequency content. For the TS+HPF distorted image, the low-frequency content is lost with little disruption to the high-frequency content. The TS+HPF distorted image has "Fair" perceived quality (statistically equivalent to the JPEG distorted image) but perceived utility corresponding to the reference equivalence class.

These examples illustrate that distorted images corresponding to a specific level of utility can significantly vary in terms of quality, and distorted images corresponding to a specific level of quality can significantly vary in terms of utility. Thus, quality does not reliably predict of utility. Furthermore, the observed relationship between utility and quality implies that any objective estimator that accurately estimates perceived quality (utility) scores cannot also accurately estimate perceived utility (quality) scores across a variety of distortion types.

## 3.5 Limitations of Perceived Utility Scores

Relative perceived utility scores of distorted images were obtained using a paired comparison methodology that has two limitations. The subjective responses lack information about the specific content actually recognized by the observers viewing the distorted images, because the test method only collected binary responses (i.e., a choice) from observers in response to the query "Which image tells you more about the content?". This precludes an analysis of the data based on the actual criteria that led observers to their responses.

The second limitation is that observers may have used a secondary factor such as perceived quality to choose an image when both images appeared equal with regard to their perceived usefulness. For example, for the *airplane*, *backhoe*, and *caged birds* images, the TS distorted images had higher perceived utility than the TS+HPF distorted image with the same $\gamma$. If observers consistently rely on a secondary factor to choose an image, then the perceived utility scores will be intermixed with these secondary factors. Because TS distorted images have greater perceived quality than TS+HPF distorted images, the perceived quality is the most likely secondary factor to influence an observers decision.

Despite the limitations with the current method used to obtain relative perceived utility scores, the results still illustrate a distinction between perceived quality and perceived utility, and any improvements to the test methodology used to obtain relative perceived utility scores are expected to reveal greater differences between perceived quality and perceived utility.

## 3.6 Summary

This chapter described a suite of experiments that were conducted to obtain perceived utility and perceived quality scores for a collection of distorted natural images. An analysis comparing the the results from the experiments reveals that perceived quality is not a proxy for perceived utility. A distortion that removes low-frequency content from an image demonstrated that perceived utility is largely based on the fidelity of high-frequency content and is less affected by distortions to low-frequency content. Perceived quality, on the other hand, is affected by distortions to both low and high frequency content. The observed relationship between utility and quality implies that accurate objective quality (utility) estimators will not accurately estimate perceived utility (quality) for a broad class of distortions.

## 3.7 Key Points

- A suite of experiments were conducted to obtain perceived utility and perceived quality scores for a set of distorted natural images. The resulting relationship between the perceived utility and perceived quality scores indicates that quality is not a suitable proxy for utility.
- Distorted images spanning a broad selection of distortion types were used in the experiments, and some distortions (i.e., the texture-smoothing (TS) and TS plus high-pass filtering distortions) have not been formally examined in terms of perceived quality. Distortions to the low-frequency content of natural images were not found to affect their perceived utility, whereas distortions to high-frequency content did affect their perceived utility. Furthermore, quality is affect by distortions to both low and high frequency

content.

- Limitations with regard to the perceived utility scores were noted. First, the subjective responses lack information about the content actually recognized by observers viewing the distorted images, because observers provided responses to the query “Which image tells you more about the content?”. Second, observers likely relied on a secondary factor to resolve potential ties (i.e., equal usefulness), and that factor may have been perceived quality. An improved methodology that addresses these issues is presented in Chapter 5.
- Despite the limitations with the experimental data, the results still illustrate a distinction between perceived quality and perceived utility. In short, the observed relationship between utility and quality implies that accurate objective quality (utility) estimators will not accurately estimate perceived utility (quality) for a broad class of distortions.

CHAPTER 4

# OBJECTIVE ESTIMATES OF PERCEIVED UTILITY AND PERCEIVED QUALITY

## 4.1 Introduction

The psychometric evidence presented in Chapter 3 establishes that any objective estimator that accurately estimates perceived quality scores cannot accurately estimate perceived utility scores. The performance of several objective estimators as both quality and utility estimators is assessed. Although most of these objective estimators have been designed to estimate perceived quality, they serve as signal analysis tools not only to develop an understanding of those image characteristics that impact usefulness but also to suggest signal analysis tools for an objective utility estimator.

Two objective estimators are shown to accurately estimate utility. The first is an objective estimator that is customarily used as a quality estimator. A modified version of this estimator, in which the modifications adjust the relative importance of distortions across spatial frequencies to the overall objective estimate, is shown to generate the most accurate estimates of perceived quality among the objective estimators evaluated.

The second objective estimator is the newly proposed natural image contour evaluation (NICE) utility estimator, which was inspired by the importance of contour information to the human visual system for object perception [34, 67, 75]. NICE is based on the hypothesis that degradations to image contours restrict the content that an image conveys to a human and decrease perceived utility. In particular, NICE estimates utility as a function of both *lost* and *introduced* contour information in a distorted image when compared with a reference image.

Chapter 2 reviewed several objective estimators that are assessed as both utility and quality estimators of distorted natural images in this chapter. The natural image contour evaluation (NICE) utility estimator is introduced and described in Section 4.2. Results summarizing the performance of NICE as well as several objective estimators are reported and described in Section 4.3. The results from both of the subjective experiments reported in Chapter 3 and the analysis of objective estimators as utility and quality estimators are discussed in Section 4.4. A summary is provided in Section 4.5.

## 4.2 NICE: Natural Image Contour Evaluation

Processing in the human visual system (HVS) parses a visual stimulus into meaningful pieces that facilitate the perception of objects. The primary visual cortex extracts local, oriented edge information from a visual stimulus. This information is later processed by cortical regions of the HVS that have been associated with object perception [49]. Cells within in the extra-striate cortex, in particular V4, have been functionally described as shape descriptors [75]. The extra-striate visual cortex has been shown to exhibit an increased activation in response to images that contain contour information [34]. Thus, the evidence suggests that the HVS uses contour information for object perception.

A degradation to image contours is hypothesized to inhibit object perception. Furthermore, we hypothesize that the perceived usefulness of a distorted image is related to a human's ability to recognize objects within that image. Biderman and

Ju reported that human observers can recognize objects from line drawings nearly as efficiently as photographs [3], and the authors of the present paper have shown elsewhere that humans can recognize image content from contour information detected using a Canny edge detector operating at different image scales [102]. The fidelity of contour information from a test image with respect to a reference image may be a reliable indicator of perceived utility, and, specifically in this paper, a human's ability to extract information from the test image.

The natural image contour evaluation (NICE) utility estimator compares the contours identified in a test image to those identified in the reference image to produce a numerical score indicating the estimated utility score of the test image [100, 105]. Image contours or edges, defined by sudden intensity changes in pixel values, can be identified by the presence of an absolute maximum magnitude in the gradient of an image [95].

Image contours can be detected from a single image scale or across multiple image scales. For example, the Sobel edge detector analyzes image content from a single image scale to identify contours. However, energy from edges span multiple image scales, and the HVS does not strictly analyze one image scale of visual information [31]. A wavelet decomposition coarsely approximates the multi-scale, multi-orientation analysis conducted by the primary visual cortex, and can be used to identify contours at multiple image scales. The Sobel edge detector is computationally efficient, but multi-scale contour identification uses visual information from multiple image scales that would be available to the HVS. The performance of NICE was evaluated using both single- and multi-scale contour identification methods. The computation that NICE conducts using identified contours is described and followed by individual descriptions of the single-scale and multi-scale contour identification methods used for NICE.

### Contour Comparison

An objective score with NICE is computed by comparing the contours of the reference and test images, which are represented as binary images. Before the contours of the reference and test images are compared, binary images representing the contour maps are individually subjected to morphological dilation with a $3 \times 3$ "plus-sign" shaped structuring element $E$ [46]. Morphological dilation accommodates local registration errors between the reference and test contour maps introduced by distortions in the test image that should not be quantified as errors.

The contours of the reference and test images are compared across $S$ image scales, and $B_s$ and $\hat{B}_s$ respectively denote the contours of the reference and test images at scale $s$. The overall NICE score for the test image is

$$\text{NICE} = \frac{\sum_{s=1}^{S} d_H(B_s \oplus E, \hat{B}_s \oplus E)}{\sum_{s=1}^{S} N_{B_s}}, \tag{4.1}$$

where $N_{B_s}$ is the number of non-zero elements of $B_s \oplus E$, $d_H(X, Y)$ denotes the Hamming distance[1] between the two binary vectors $X$ and $Y$, and $B \oplus E$ denotes the dilation of the binary image $B$ using the morphological structuring element $E$. The Hamming distance quantifies 1) the number of pixels corresponding to contours in the reference image that have been *lost* in the test image due to the distortions and 2) the number of pixels corresponding to contours in the test image *introduced* by the distortions that were absent in the reference image. Since the content of natural images vary, the proportion of pixels corresponding to contours will vary. The factor $N_B$ accounts for this variability by adaptively scaling the

[1] The Hamming distance counts the number of dissimilar elements between two vectors [50].

raw score $d_H(B \oplus E, \hat{B} \oplus E)$ according to the extent of the contour information identified in the reference image.

### Single-scale Contour Identification with Classical Edge-Detectors

Numerous image processing tools have been designed to detect edges in natural images [13, 72, 95]. These are used to generate the binary images $B_1$ and $\hat{B}_1$ corresponding to contours of the finest image scale of the respective reference and test images for the single-scale implementation of NICE (i.e., $S = 1$ in Eq. (4.1)). Edge-detectors incorporate a filtering operation that approximates the first-derivative of the image. The Sobel and Canny edge-detectors were used for the single-scale version of NICE.

The Sobel edge-detector filters an image with two $3 \times 3$ linear filters, one that approximates a horizontally-oriented derivative and another that approximates a vertically-oriented derivative. If $G_x$ and $G_y$ correspond to the approximated horizontal and vertical derivatives of the original image, respectively, then an edge-intensity image, given as $G = G_x^2 + G_y^2$, is subjected to hard-thresholding, using a threshold given as twice the average value of $G$, to produce a binary image identifying image contours.

The Canny edge-detector filters the image with the derivative of a Gaussian specified for a particular $\sigma > 0$ and applies thresholding to generate a binary image [13]. The parameter $\sigma$ in the Canny filter controls the suppression of high-frequency content (i.e., textures and uncorrelated noise) before detecting edges, and NICE was implemented with the Canny edge-detector for $\sigma = 1$.

### Multi-scale Contour Identification

A wavelet representation of an image provides multi-scale directional derivatives of that image, which can be used to identify image contours at different image scales. Both the reference and test images are represented using an *undecimated* implementation of the steerable pyramid [117] using $D$ orientations and $S$ scales[2]. Let $W_{s,\theta}(i)$ and $\hat{W}_{s,\theta}(i)$ denote the $i^{th}$ wavelet coefficient of the respective reference and test images in the subband corresponding to scale $s \in \{1, 2, \ldots, S\}$ and orientation $\theta \in \{0, \frac{\pi}{D}, \frac{2\pi}{D}, \ldots, \frac{\pi(D-1)}{D}\}$.

For each image scale $s$, the local modulus maxima (LMM) [70] of wavelet coefficient scales correspond to image contours for the reference and test images. The LMM are determined from gradient vectors formed from wavelet subbands corresponding to derivatives in horizontal and vertical spatial directions [70]. Define $G_s(i) = W_{s,0}(i) - \mathrm{j}W_{s,\frac{\pi}{2}}(i)$ and $\hat{G}_s(i) = \hat{W}_{s,0}(i) - \mathrm{j}\hat{W}_{s,\frac{\pi}{2}}(i)$ as the gradient of the respective reference and test images at scale $s$, where $\mathrm{j} = \sqrt{-1}$. For image scale $s$, let $M_s(i) = |G_s(i)|$ and $A_s(i) = \angle G_s(i)$ denote the respective modulus and angle of the gradient of the reference image. Similarly, define $\hat{M}_s(i) = |\hat{G}_s(i)|$ and $\hat{A}_s(i) = \angle \hat{G}_s(i)$ for the test image. LMM of the reference image correspond to points of $M_s(i)$ greater than the two adjacent neighbors in the direction indicated by $A_s(i)$, and for the test image, LMM are similarly identified using $\hat{M}_s(i)$ and $\hat{A}_s(i)$. For scale $s$, let $\mathcal{I}_s$ and $\hat{\mathcal{I}}_s$ denote sets of indices $i$ corresponding to LMM of the respective reference image and test images.

Binary images represent image contours of the reference and test images. Thresholds used to identify contours are independently calculated for the reference and test images based on the energy of the combined horizontal and vertical

[2] The high-pass residual generated by the steerable pyramid is not used.

subbands (i.e., $M_s$ and $\hat{M}_s$). Specifically, the image contours at scale $s$ of the reference and test images are identified as LMM that exceed the respective thresholds $\beta_s = \frac{4}{P}\sum_{i=1}^{P} M_s^2(i)$ and $\hat{\beta}_s = \frac{4}{P}\sum_{i=1}^{P} \hat{M}_s^2(i)$, where $P$ is the number of wavelet coefficients. $B_s(i)$ and $\hat{B}_s(i)$, the reference and test binary images for scale $s$, are defined as

$$B_s(i) = \begin{cases} 1 & M_s(i) > \beta_s \text{ and } i \in \mathcal{I}_s \\ 0 & else \end{cases}. \tag{4.2}$$

$\hat{B}_s(i)$ is similarly defined using $\hat{M}_s$, $\hat{\mathcal{I}}_s$, and $\hat{\beta}_s$.

## 4.3 Results: Objective Estimates of Utility and Quality

Subjective experiments are reliable but prohibitively expensive methods to estimate either utility or quality, but an objective estimator that is consistent with subjective responses for either utility or quality can be used in lieu of the subjective experiments. This section evaluates each objective estimator described in Section 2.3 as both a utility estimator and a quality estimator. Specifically, the objective estimates are evaluated using the perceived utility and perceived quality scores from the subjective experiments. Objective estimators that provide accurate and reliable estimates of the subjective scores also serve as signal analysis tools that can be analyzed to understand what image characteristics impact the subjective scores. For example, an objective estimator that reliably estimates perceived utility scores can be dismantled to understand the image characteristics that affect utility.

The implementations of all the objective estimators were obtained from the respective authors and are available in the Metrix Mux compilation of objective estimators [44]. Single-scale implementations of NICE are evaluated using the Sobel and Canny edge-detector, respectively denoted as $\mathrm{NICE}_{\mathrm{Sobel}}$ and $\mathrm{NICE}_{\mathrm{Canny}}$. Multi-scale implementations of NICE are evaluated using up to four scales (i.e., for $S = 1, 2, 3, 4$ in Eq. (4.1)), where each implementation is denoted $\text{MS-NICE}_S$ (i.e., $\text{MS-NICE}_3$ denotes MS-NICE using the first three image scales).

A monotonic, nonlinear mapping between objective estimates and subjective scores is often recommended before analyzing the performance of an objective estimator [132]. However, the nonlinear mapping *functionally* compensates for objective estimator's shortcomings and obscures the relationship between the image characteristics analyzed by that objective estimator and those that affect the subjective scores. Thus, a linear mapping between the objective estimates and the subjective scores was used to avoid drawing erroneous conclusions from the results that are due to the nonlinear mapping and not the objective estimator. Furthermore, objective estimators that estimate either utility or quality using only a linear mapping are preferred, since training data is not needed to calibrate the nonlinear mapping associated with the objective estimator (see also Appendix VI.3 of [133]).

An affine linear function $h_{\mathcal{E}}$ that maps the objective estimates to the range of values corresponding to the subjective scores that lie in the domain $\mathcal{E}$ was fitted to the data. The parameters of $h_{\mathcal{E}}$ were found by minimizing the sum of the set of squared residuals $\{(h_{\mathcal{E}}(d_i) - e_i)^2\}_{i=1}^{n}$ for the $n$ images, where $d_i$ and $e_i$ respectively denote an objective estimate and a subjective score for image $i$.

To test the performance of an objective estimator as a utility estimator and a quality estimator both correlation and accuracy statistics were used to quantify the relationship between the its objective estimates and the respective subjective scores. Specifically, 1) the objective estimates and the subjective scores must be strongly correlated and 2) the objective estimator must accurately estimate the subjective scores.

The same correlation and accuracy statistics used in Section 3.4.1 to evaluate and quantify the relationship between perceived quality scores and perceived utility scores are used to evaluate the objective estimators. The root mean squared error (RMSE) is computed between the subjective scores and the linearly mapped objective estimates. The resolving power ($RP_\alpha$) is another accuracy statistic that is used to specify the smallest difference in fitted objective scores for a pair of test images such that the difference is significant based on the estimated error of the subjective scores at the $100(1-\alpha)\%$ confidence level [7].

The skewness and kurtosis of the set of residuals $\{h_{\mathcal{E}}(d_i) - e_i\}_{i=1}^{n}$ are also reported. Values of skewness and kurtosis that differ from 0 and 3, respectively, suggest that the residuals do not come from a Gaussian distribution. The best performing objective estimators will have residuals that come from a Gaussian distribution with a small standard deviation (i.e., small root mean squared error (RMSE)). Such estimators analyze important image characteristics that reliably explain the variation in the subjective scores.

The correlation and accuracy statistics are estimated from the data and, therefore, are random variables, so statistical differences, not absolute differences, indicate significant differences in the statistics. Statistical differences between correlation statistics and the outlier ratio (OR) are determined using the same methods described in Section 3.4.1.

Statistical differences in accuracy are determined by comparing the variance of the residuals corresponding to different objective estimators. An $F$-test frequently is used to compare the variance of the residuals corresponding to different objective estimators, but an assumption with the $F$-test is that the residuals come from a Gaussian distribution [32, 132]. For most objective estimators, the residuals did not come from a Gaussian distribution according to the Jarque-Bera normality test [57], so the Brown-Forsythe Levene (BFL) test [10], rather than the $F$-test, was used to compare the variance of the residuals for different objective estimators, with results reported by the corresponding $p$-value. With the BFL test, $p$-values greater than 0.05 indicate at the 95% confidence level that the variance of the residuals for two estimators are statistically equivalent.

The results that characterize the performance of the objective estimator as both 1) utility estimators and 2) quality estimators are reported separately. A general summary of the results is presented.

### 4.3.1 Results: Objective Estimates of Perceived Utility

A utility estimator should both detect recognizable images and provide accurate estimates of perceived utility.

#### Determining if Test Images Are Recognizable

Objective estimators can be used to determine if test images are recognizable by applying an appropriate threshold to the score generated by that estimator.

An image is either recognizable or unrecognizable. Cast as a two-class detection problem, the performance of a estimator as a detector can be characterized by its receiver operating characteristic (ROC) [38, 48, 51]. An ROC curve summarizes the relationship between the proportion of true positives and false positives for a given estimator using a range of threshold values. The area under the ROC curve (AUC) collapses the performance of an objective estimator to a single number. Given a pair of test images belonging to each class (i.e., one recognizable and one unrecognizable), the AUC quantifies the probability that an estimator correctly distinguishes recognizable images from unrecognizable images.

The objective estimators were evaluated as recognition detectors by applying a threshold to the objective estimates to classify an image as either recognizable or unrecognizable. A total of 1000 thresholds were tested ranging from 0.95 of the minimum objective estimate to 1.05 times the maximum objective estimate. For each threshold, the true positive rate (i.e., the proportion of times an image was correctly classified as recognizable) and the false positive rate (i.e., the proportion of times an image was incorrectly classified as recognizable) were recorded. ROC curves were generated from the recorded pairs of true positive and false positive rates. The AUC was estimated by the trapezoidal rule [51]. The AUC is a statistic estimated from available data and is therefore a random variable, so the 95% confidence intervals for the estimates of the AUC were computed [51]. The first column of Table 4.1 lists the AUC as the *recognition detection accuracy* for each objective estimator that was used to detect recognizable images across all distortions.

**VIF, VIF*, NICE$_{\mathbf{Sobel}}$, NICE$_{\mathbf{Canny}}$, and all versions of MS-NICE correctly distinguish recognizable images from unrecognizable images with statistically greater probability than the other objective estimators.** All of the other objective estimators correctly rank two such images with probability greater than chance. In Table 4.1, the absolute maximum value of the recognition detection accuracy is shown in boldface, and values that are statistically equivalent with 95% confidence are italicized. The subjective experiments revealed a linear relationship between perceived quality scores and perceived utility scores for low quality distorted images, so an objective estimator that produces accurate estimates of perceived quality scores should also accurately detect recognizable images. All the other objective estimators exhibit poor recognition detection accuracy, because these estimators severely underestimate the perceived utility scores of TS+HPF distorted images. Specific details about the performance of these estimators are discussed alongside the results presented in Section 4.3.1.

### Estimating the Perceived Utility of Recognizable Test Images

A utility estimator should accurately estimate the perceived utility of a test image deemed recognizable. Only those test images with perceived utility scores exceeding $-15$ ($n = 163$ test images) are used to evaluate an estimator's performance as a utility estimator, since accurate estimates of perceived utility scores for unrecognizable images are unnecessary. Tables 4.1 and 4.2 summarize the correlation and accuracy statistics for all the objective estimators when analyzing their linearly mapped objective estimates with respect to the perceived utility scores. The $p$-value for the BFL test $\text{BFL}_p$ is reported when the residuals of each objective estimator were compared with the residuals of VIF, since residuals for VIF exhibited the smallest variance when VIF was evaluated as a utility estimator.

The remainder of this section reports the key results, which appear in boldface, followed by a summary of the results for subsets of objective estimators that exhibit similar performance with headings for the subsets of estimators appearing in italics. Statistical justifications, general interpretations, and specific remarks about the objective estimators are reported.

Table 4.1: Statistics summarizing 1) the ability of an objective estimator to distinguish recognizable and unrecognizable images and 2) the correlation of objective estimator values with perceived utility scores. The recognition detection accuracy is the probability that an unrecognizable image and recognizable image are correctly distinguished. The Pearson (linear) correlation coefficient $r$, the Spearman rank correlation coefficient $\rho$, the Kendall rank correlation $\tau$ are reported when the estimates are compared with the perceived utility scores for test images with perceived utility exceeding $-15$ ($n = 163$ test images). Optimal values appear in boldface with statistically equivalent values italicized.

| | | *Recognition* | | | |
|---|---|---|---|---|---|
| | | *Detection* | *Correlation Measures* | | |
| | *Estimator* | *Accuracy* | $\rho$ | $\tau$ | $r$ |
| *Spectral Slope* | $\beta$ | 0.729 | 0.751 | 0.535 | 0.730 |
| *Sig. Fidelity* | PSNR | 0.768 | 0.520 | 0.422 | 0.414 |
| *Measures* | $C_{rms}(\boldsymbol{E})$ | 0.792 | 0.521 | 0.404 | 0.211 |
| *Estimators* | WSNR | 0.766 | 0.485 | 0.372 | 0.415 |
| *Based on* | NQM | 0.796 | 0.509 | 0.401 | 0.422 |
| *HVS* | VSNR | 0.790 | 0.530 | 0.436 | 0.541 |
| *Properties* | C4 | 0.830 | 0.661 | 0.517 | 0.651 |
| *Estimators* | SSIM | 0.924 | 0.862 | 0.682 | 0.845 |
| *Based on* | MS-SSIM | 0.935 | 0.731 | 0.585 | 0.652 |
| *Hypothesized* | VIF | *0.978* | **0.959** | **0.821** | **0.943** |
| *HVS* | VIF* | *0.973* | 0.928 | *0.768* | *0.924* |
| *Objectives* | $\text{NICE}_{\text{Sobel}}$ | *0.980* | *0.951* | *0.804* | *0.924* |
| *Proposed* | $\text{NICE}_{\text{Canny}}$ | *0.980* | *0.937* | *0.785* | *0.935* |
| *Utility* | $\text{MS-NICE}_1$ | *0.979* | *0.956* | *0.816* | *0.923* |
| *Estimators* | $\text{MS-NICE}_2$ | *0.980* | *0.959* | *0.821* | *0.911* |
| | $\text{MS-NICE}_3$ | *0.980* | *0.958* | *0.817* | 0.902 |
| | $\text{MS-NICE}_4$ | **0.981** | *0.947* | *0.794* | 0.901 |

**Estimators that analyze distortions to low-frequency content perform poorly as utility estimators.** *The spectral slope, signal fidelity measures, and the objective estimators based on HVS properties perform poorly as utility estimators.* Estimates from these estimators weakly correlate with and inaccurately estimate perceived utility scores. Specifically, the linear correlation between the values from these estimators and the perceived utility scores ($0.21 < r < 0.73$) imply that these

Table 4.2: Statistics summarizing the accuracy of estimators serving as utility estimators. The root mean squared error (RMSE), the outlier ratio (OR), and the resolving power $RP_\alpha$ for $\alpha = 0.1, 0.05$ are reported when the estimates are compared with the perceived utility scores for test images with perceived utility exceeding $-15$ ($n = 163$ test images). Optimal values appear in boldface with statistically equivalent values italicized. An asterisk beside the RMSE indicates that the residual variance is statistically equivalent to that of VIF according to the Brown-Forsythe-Levene test at the 95% confidence level. The skewness (Skew.) and kurtosis (Kurt.) of the residuals are italicized when the Jarque-Bera (JB) test indicates that the residuals belong to a Gaussian distribution.

| | | *Estimating Perceived Utility: Accuracy Measures* | | | | | |
|---|---|---|---|---|---|---|---|
| | *Estimator* | RMSE | OR | $RP_{0.10}$ | $RP_{0.05}$ | Skew. | Kurt. |
| *Spectral Slope* | $\beta$ | 25.6 | 0.748 | 52.3 | 64.4 | 0.51 | 2.8 |
| *Sig. Fidelity* | PSNR | 34.1 | 0.859 | 57.3 | 57.3 | *-0.19* | *2.6* |
| *Measures* | $C_{rms}(\boldsymbol{E})$ | 36.6 | 0.877 | 37.6 | 38.2 | 0.11 | 1.8 |
| *Estimators* | WSNR | 34.0 | 0.847 | 56.3 | 57.6 | *-0.22* | *2.4* |
| *Based on* | NQM | 33.9 | 0.847 | 54.1 | 54.1 | *-0.28* | *2.4* |
| *HVS* | VSNR | 31.5 | 0.742 | 41.5 | 83.9 | -0.51 | 3.0 |
| *Properties* | C4 | 28.4 | 0.785 | 69.9 | 75.9 | -0.74 | 3.9 |
| *Estimators* | SSIM | 20.0 | *0.595* | 41.5 | 55.2 | *-0.12* | *3.8* |
| *Based on* | MS-SSIM | 28.4 | 0.828 | 49.6 | 66.4 | *0.01* | *2.4* |
| *Hypothesized* | VIF | **12.4** | 0.595 | 20.8 | **26.6** | *0.04* | *2.9* |
| *HVS* | VIF* | 14.3* | *0.497* | 30.9 | 41.1 | -0.53 | 4.2 |
| *Objectives* | $\text{NICE}_{\text{Sobel}}$ | 14.3* | 0.564 | 24.0 | 33.6 | -0.37 | 4.1 |
| *Proposed* | $\text{NICE}_{\text{Canny}}$ | 13.3* | **0.454** | 29.0 | 39.1 | -0.36 | 5.2 |
| *Utility* | $\text{MS-NICE}_1$ | 14.4* | 0.583 | 22.2 | 33.0 | -0.35 | 3.7 |
| *Estimators* | $\text{MS-NICE}_2$ | 15.4* | 0.577 | 19.8 | 33.4 | *-0.15* | *3.6* |
| | $\text{MS-NICE}_3$ | 16.2 | 0.601 | **19.3** | 34.0 | *-0.06* | *3.5* |
| | $\text{MS-NICE}_4$ | 16.3 | 0.601 | 21.8 | 34.5 | *0.03* | *3.3* |

estimators account for no more than 53% (i.e., $100r^2$%) and as little as 4% of the variation in utility. Their low rank correlation with the perceived utility scores ($\rho < 0.75$, $\tau < 0.54$) indicate that a monotonic, nonlinear mapping would not significantly improve the performance of these estimators as utility estimators.

The RMSE between each estimator's linearly mapped estimates and the perceived utility scores exceeds 26, which corresponds to approximately one-quarter of the total utility scale. When the estimates from each these estimators are used as utility estimates, the residual variance is statistically larger than that of VIF, which has the smallest residual variance, according to the BFL test. The outlier ratios for these estimators indicates that their estimates of perceived utility for 74% of the distorted images lie outside the 95% confidence intervals of those distorted images' respective perceived utility scores. Last, the resolving power statistics for these estimators imply that these estimators only reliably distinguish distorted images with gross differences in perceived utility scores (i.e., in most cases exceeding 50) at both the 10% and 5% significance levels.

The TS+HPF distorted images largely influence the performance of these estimators. When each estimator was analyzed as a utility estimator with the TS+HPF distorted images removed, all estimators except the spectral slope exhibited significantly better performance as utility estimators. The performance improvements when the TS+HPF distorted images are removed indicate that these estimators operate with the assumption that distortions do not compromise the integrity of the low-frequency content without also severely distorting the high-frequency content. Such an assumption is consistent with the behavior of lossy image compression methods but becomes problematic when an image is also susceptible to transmission errors (e.g., packet loss) that arbitrarily distort an image.

The spectral slope quantifies the shape of the distorted image's frequency response. The J2K+DCQ, TS, and TS+HPF distortions primarily disrupt and suppress high-frequency content before low-frequency content as the level of distortion increases, which leads to a significant decrease in the spectral slope (i.e., $\beta$ increases in $A(f) = 1/f^{-\beta}$). JPEG distortions simultaneously disrupt, suppress, and introduce high-frequency content (e.g., blocking artifacts) and lead to a modest increase in $\beta$ relative to the other distortions as the level of distortion increases. As a result, the relationship between the spectral slope and perceived utility varies with each distortion type, and the spectral slope is observed to be an unreliable indicator of utility, since its relationship with perceived utility scores varies with distortion type.

The signal fidelity measures as well as the estimators based on HVS properties generate objective estimates that are entirely or in part a function of energy measurements of the reference and test images. PSNR and $C_{rms}(\boldsymbol{E})$ measure the global energy of the difference image $\boldsymbol{X} - \hat{\boldsymbol{X}}$ in the pixel and luminance domains, respectively. VSNR analyzes the visibility of the global contrast of the difference image across several image scales. The other estimators based on HVS properties apply different filters to suppress frequency content less sensitive to the HVS and compare the global energy of the filtered reference and test images in the frequency domain. All of these estimators account for distortions to low-frequency content, and the loss of low-frequency content significantly decreases the energy of the distorted image relative to the reference image. Consequently, each of these estimators underestimate the perceived utility scores for TS+HPF distorted images.

*SSIM performs satisfactorily as a utility estimator but MS-SSIM does not.* Among all the estimators evaluated, estimates from SSIM both moderately correlate with and accurately estimate perceived utility scores of distorted images. Specifically, the linear correlation between SSIM's estimates and the perceived utility scores ($r = 0.84$) imply that SSIM accounts for 71% of the variation in utility. Despite the large Spearman rank correlation statistic ($\rho = 0.86$), further inspection of the relationship between SSIM's estimates and perceived utility indicated that a monotonic, nonlinear mapping would not significantly improve SSIM's performance as a utility estimator.

The RMSE between SSIM's linearly mapped estimates and the perceived utility scores is 20, which corresponds to approximately one-fifth of the total utility scale. Furthermore, the variance of the residuals when SSIM's estimates are used as utility estimates are statistically larger than that of VIF according to the BFL test. The outlier ratio indicates that SSIM's estimates for 60% of the distorted images lie outside the 95% confidence interval of the respective distorted images' perceived utility scores. Last, the resolving power for SSIM implies that it only reliably distinguishes distorted images with differences in perceived utility scores as small as 42 and 55 at the 10% and 5% significance levels, respectively.

In contrast with SSIM, estimates from MS-SSIM weakly correlate with and inaccurately estimate perceived utility scores. Specifically, the linear correlation between MS-SSIM's estimates and the perceived utility scores ($r = 0.65$) imply that MS-SSIM accounts for only 43% of the variation in perceived utility. MS-SSIM's Spearman rank correlation statistic ($\rho = 0.73$) indicates that a monotonic, nonlinear mapping would not significantly improve MS-SSIM's performance as a utility estimator.

The RMSE between MS-SSIM's linearly mapped estimates and the perceived utility scores is 28, which corresponds to approximately one-third of the total utility scale. The BFL test concluded that the variance of the residuals when MS-SSIM's estimates are used as utility estimates are statistically larger than that of VIF. The outlier ratio indicates that MS-SSIM's linearly mapped estimates for 83% of the distorted images lie outside the 95% confidence interval of the respective distorted images' perceived utility scores. Last, the resolving power for MS-SSIM implies that it only reliably distinguishes distorted images with differences in perceived utility scores as small as 50 and 66 at the 10% and 5% significance levels, respectively.

Both SSIM and MS-SSIM incorporate an analysis of low-frequency content via a comparison of the spatially local mean pixel values of the reference and test images. In addition to MS-SSIM's local mean comparison of the reference and test images, MS-SSIM compares the variance of spatially local pixel values of the reference and test images across multiple image scales. Thus, both MS-SSIM's mean and variance comparisons analyze the low-frequency content of the reference and test images, whereas only SSIM's mean comparison analyzes the low-frequency content of the reference and test images.

Because SSIM and MS-SSIM analyze low-frequency content, each estimator underestimates the perceived utility scores of TS+HPF distorted images. As with the spectral slope, signal fidelity measures, and the objective estimators based on HVS properties, the interpretation of SSIM and MS-SSIM as utility estimators changes when the TS+HPF distorted images are removed from the database: both SSIM and MS-SSIM produce more accurate estimates of perceived utility scores when the TS+HPF distorted images are removed.

SSIM and MS-SSIM were modified by removing the comparisons of the reference and test images that quantify disruptions to low-frequency content, and both modified estimators exhibited better performance as utility estimators than their original implementations across all five distortion types. The linear correlation and RMSE between SSIM's estimates and perceived utility significantly improve to 0.92 and 15, respectively, when SSIM operates without the local mean comparison (i.e., when SSIM ignores disruptions to low-frequency content). The linear correlation and RMSE between MS-SSIM's estimates and perceived utility modestly improve to 0.73 and 25, respectively, when MS-SSIM operates without both the local mean and variance comparisons across multiple image scales. Even when the local mean and variance comparisons have been removed, MS-SSIM's multi-scale analysis necessarily quantifies distortions to low-frequency content and explains its modest performance improvement. However, the significant improvement demonstrated with SSIM when the local mean comparisons are removed relative to the original implementation of SSIM suggests that an analysis of high-frequency content provides reliable estimates of perceived utility[3].

*VIF* produces unreliable estimates of perceived utility, especially for TS+HPF distortions with high perceived utility.* Estimates from VIF* strongly correlate with and accurately estimate perceived utility scores, and most of VIF*'s correlation and accuracy statistics are statistically equivalent to those of VIF. However, the outlier ratio for VIF* is statistically larger than that of VIF and indicates that VIF*'s linearly mapped estimates for 60% of the distorted images lie outside the 95% confidence interval of the respective images' perceived utility scores. Furthermore, VIF*'s resolving power implies that it reliably distinguishes distorted images with differences in perceived utility scores as small as 30 at the 10% significance level, whereas VIF reliably distinguishes distorted images with differences in perceived utility scores as small as 26 at the 5% significance level. In short, VIF* less reliably distinguishes distorted images with smaller differences in perceived utility than VIF.

VIF* underestimates the perceived utility of TS+HPF distorted images with high perceived utility, because, unlike VIF, VIF* has a greater sensitivity to disruptions to low-frequency content. The negative skewness of VIF*'s residuals are a

[3] The local variance comparison used by SSIM corresponds to an analysis of high-frequency content and does not need to be removed.

consequence its poor estimates of the perceived utility scores for TS+HPF distorted images. The subjective experiments demonstrate that disruptions to low-frequency content do not consistently affect perceived utility scores. Therefore, VIF*'s poorer performance as a utility estimator, especially for TS+HPF distorted images, is expected, because VIF* is sensitive to disruptions to low-frequency content.

**Estimators that strictly analyze distortions to high-frequency content and measure degradations to image contours accurately estimate perceived utility.** *VIF, $NICE_{Sobel}$, $NICE_{Canny}$, and MS-$NICE_{S \leq 2}$*[4] *outperform the other objective estimators as utility estimators.* Relative to the other estimators evaluated, estimates from these estimators strongly correlate with the perceived utility scores. Specifically, the linear correlation between the estimates from these estimators and the perceived utility scores ($r > 0.92$) imply that these estimators account for between 83% and 89% of the variation in utility with VIF accounting for the greatest variation in utility. Despite the large Spearman rank correlation statistics for these estimators ($\rho > 0.93$), the large linear correlation statistics indicate that a monotonic, nonlinear mapping will not significantly improve the performance of these estimators as utility estimators.

Estimates from these objective estimators accurately estimate the perceived utility scores relative to the other estimators. The RMSE between each estimator's linearly mapped estimates and the perceived utility scores is less than 15.4, which corresponds to approximately one-seventh of the total utility scale. According to the BFL test, the variance of the residuals when estimates from these estimators are used to estimate utility scores are statistically equivalent to that of VIF, which has the smallest residual variance.

[4] The notation MS-$NICE_{S \leq 2}$ is used to refer to both MS-$NICE_1$ and MS-$NICE_2$.

The outlier ratios for these estimators indicate that their estimates of perceived utility scores for between 45% and 60% of the distorted images lie outside the 95% confidence intervals of the distorted images' respective perceived utility scores. $\text{NICE}_{\text{Canny}}$ and VIF were observed to have the smallest and statistically equivalent outlier ratios and indicate that at most half of their estimates lie outside the 95% confidence intervals of the distorted images' respective perceived utility scores.

The resolving power for these estimators show no consistent trend favoring one estimator over on another: estimators with smaller resolving powers relative to the other estimators at 10% significance level have relatively larger resolving powers than the other estimators at the 5% significance level. Overall, the resolving powers imply that all of these estimators reliably distinguish distorted images with differences in perceived utility scores as small as 41, and some of these estimators reliably distinguish distorted images with difference in perceived quality as small as 20.

VIF, $\text{NICE}_{\text{Sobel}}$, $\text{NICE}_{\text{Canny}}$, and $\text{MS-NICE}_{S\leq 2}$ strictly analyze the high-frequency content of the reference and test images. $\text{NICE}_{\text{Sobel}}$, $\text{NICE}_{\text{Canny}}$, and $\text{MS-NICE}_{S\leq 2}$ primarily analyze disruptions to contours, whereas VIF analyzes any disruption to high-frequency content (i.e., both contours and textures). Most importantly, all of these estimators do not analyze disruptions to low-frequency content, which contributed to the poorer performance of many of the other objective estimators as utility estimators. A detailed discussion that compares VIF to NICE is presented in Section 4.4.1.

*Among the various implementations of NICE and MS-NICE, estimates from* $\textit{NICE}_{\textit{Canny}}$ *most accurately estimate the perceived utility scores.* The RMSE for $\text{NICE}_{\text{Canny}}$ is smallest among the various implementations of NICE and MS-NICE but not statistically significant. However, the residuals for $\mathrm{NICE}_{\mathrm{Canny}}$ exhibit much higher kurtosis that those for the other implementations of NICE and MS-NICE. Residuals exhibiting high kurtosis indicate that most of the estimates from $\mathrm{NICE}_{\mathrm{Canny}}$ are very accurate with respect to the perceived utility scores and poorly estimated for only a few distorted images. Further inspection of the relationship between estimates from $\mathrm{NICE}_{\mathrm{Canny}}$ and the perceived utility scores revealed that $\mathrm{NICE}_{\mathrm{Canny}}$ poorly estimates the perceived utility scores for distorted images formed from the *skier* and *caged birds* images. Removing distorted images formed from the *skier* and *caged birds* images, both significantly increases the linear correlation and significantly reduces the RMSE to 0.97 and 9.3, respectively. The interpretation of none of the other estimators changes as significantly when these distorted images are removed; even the RMSE for VIF only reduces to 11.

$\mathrm{NICE}_{\mathrm{Canny}}$ underestimates the perceived utility scores for the *skier* distorted images. The Canny edge detector identifies contours within the snow region below the skier in the *skier* image. Because all of the distortions blur the pixel values in the snow region of the image, $\mathrm{NICE}_{\mathrm{Canny}}$ no longer detects most of these contours in the snow region in any of the distorted images at the lowest level of distortion. Consequently, $\mathrm{NICE}_{\mathrm{Canny}}$ measures a large degradation to image contours in these slightly distorted images. Furthermore, a majority of the contours detected in the reference image correspond to the snow region of the image, so additional degradations to contours have a small impact on the estimate from $\mathrm{NICE}_{\mathrm{Canny}}$. The Sobel edge detector did not identify any contours in the snow region of the image, and thus removing *skier* distorted images from the dataset did not change the interpretation of its performance as a utility estimator.

$\mathrm{NICE}_{\mathrm{Canny}}$ overestimates the perceived utility scores for the *caged birds* distorted images. The cage in the *caged birds* image blocks the two birds, and the bars of the cage contribute strong edges that are identified by the Canny edge detector. As this image is distorted, the strong edges corresponding to the bars of the cage are not significantly suppressed, and thus, $\text{NICE}_{\text{Canny}}$ only measures a small overall degradation to the image contours. Because the cage partially occludes the birds, a higher-level, more complex analysis is necessary to distinguish the birds from the cage and measure the degradation of their respective contours. We hypothesize that the human observers primarily attend to the birds with an awareness of the cage, and perceived utility is gauged by the detail of the birds. $\text{NICE}_{\text{Canny}}$ does not separately measure the degradation of contours corresponding to the birds and the cage within this image.

For the remaining distorted images, $\text{NICE}_{\text{Canny}}$ outperforms the other implementations of NICE and MS-NICE, and these different implementations largely vary with respect to the edge detector used. The Sobel, Canny, and wavelet-based edge-detectors used by NICE were evaluated using the publicly available Berkeley Segmentation Dataset and Benchmark to determine which method identifies contours that best corresponds with those identified by humans [73]. The wavelet-based edge-detector was tested using only its finest scale contour maps (i.e., $s = 1$), since $\text{MS-NICE}_1$ exhibits the smallest residual variance among the four versions of MS-NICE. The Canny edge-detector ranked highest among the three methods, which suggests that its contour maps correspond best with those formed by humans. NICE is designed assuming that degradation to contours coincide with a decrease in utility, and better correspondence between the objectively identified contours and those identified by a human should improve the performance of NICE. Thus, $\text{NICE}_{\text{Canny}}$ performs best as a utility estimator, because its contours maps correspond best with those identified humans among the edge-detectors used with NICE.

*MS-NICE$_3$ and MS-NICE$_4$ inaccurately estimate utility without an additional monotonic, nonlinear mapping.* In particular, the RMSE between the linearly mapped estimates from both MS-NICE$_3$ and MS-NICE$_4$ and the perceived utility scores is approximately 16, and the variances of the residuals when estimates from both MS-NICE$_3$ and MS-NICE$_4$ are used as utility estimates are statistically larger than that of VIF. The outlier ratios for MS-NICE$_3$ and MS-NICE$_4$ indicate that their linearly mapped estimates for 60% of the distorted images lie outside the 95% confidence intervals of the distorted images' respective perceived utility scores. Last, the resolving powers for both MS-NICE$_3$ and MS-NICE$_4$ are similar to that of VIF's at the 10% significance level but greater than that of VIF's at the 5% significance level.

Estimates from both MS-NICE$_3$ and MS-NICE$_4$ strongly correlate with perceived utility scores ($r > 0.9$, $\rho > 0.94$, $\tau > 0.79$), and their rank correlation statistics are statistically equivalent to those of VIF. However, the linear correlation statistics for both MS-NICE$_3$ and MS-NICE$_4$, though strong, are statistically smaller than that of VIF. The statistical equivalence of the rank correlation statistics but statistically different linear correlation statistics for MS-NICE$_3$, MS-NICE$_4$, and VIF suggest that a monotonic, nonlinear mapping could improve the performance of both MS-NICE$_3$ and MS-NICE$_4$ as utility estimators.

An analysis of NICE operating with content from the coarser image scales 3 and 4 (i.e., for $s = 3$ and $s = 4$ in Eq. (4.1)) revealed that the computations for these coarser image scales primarily vary for images with lower perceived utility. In contrast, MS-NICE$_{S\leq 2}$ exhibits strong linear correlation with perceived utility, and the comparison of the reference and test images at coarser image scales with

MS-NICE$_S$ for $S > 2$ exhibit an increased sensitivity to low-frequency distortions of distorted images with lower perceived utility. In other words, estimates from MS-NICE$_S$ exhibit an increasingly monotonic, nonlinear relationship with the perceived utility scores as coarser image scales are analyzed (i.e., as $S$ increases).

A monotonic, nonlinear mapping, which does not affect $\rho$ and $\tau$, improved both the linear correlation and accuracy between estimates from both MS-NICE$_3$ and MS-NICE$_4$ and the perceived utility scores. This nonlinear mapping primarily compresses differences among the objective estimates for distorted images with low perceived utility scores (i.e., near the recognition threshold). Although the nonlinearity improves their performance as utility estimators, the nonlinear mapping introduces a stage of processing that was not incorporated into MS-NICE$_S$ and illustrates that MS-NICE$_S$'s analysis of the reference and test images for $S > 2$ without the monotonic, nonlinearity degenerates as utility decreases. In particular, MS-NICE$_S$ becomes increasingly sensitive to disruptions to low-frequency content for distorted images with low perceived utility scores as $S$ increases and coarser image scales are analyzed.

### 4.3.2 Results: Objective Estimates of Perceived Quality

A quality estimator should produce objective estimates that are both strongly correlated with perceived quality and accurately estimate perceived quality. All test images ($n = 243$ test images) were used to evaluate an estimator's performance as a quality estimator, because a reliable quality estimator should accurately determine the quality of unrecognizable distorted images, even though they have “Bad” quality. Tables 4.3 and 4.4 summarize the statistics for each objective estimator when analyzing the linearly mapped objective estimates with respect to the per-

Table 4.3: Statistics summarizing the correlation of objective estimator values with perceived quality scores. The Pearson (linear) correlation coefficient $r$, the Spearman rank correlation coefficient $\rho$, and the Kendall rank correlation $\tau$ are reported when the estimates are compared with the perceived quality scores for all test images ($n = 243$). Optimal values appear in boldface with statistically equivalent values italicized.

| | | *Correlation Measures* | | |
|---|---|---|---|---|
| | *Estimator* | $\rho$ | $\tau$ | $r$ |
| *Spectral Slope* | $\beta$ | 0.518 | 0.331 | 0.585 |
| *Sig. Fidelity* | PSNR | 0.598 | 0.477 | 0.656 |
| *Measures* | $C_{rms}(\boldsymbol{E})$ | 0.627 | 0.480 | 0.401 |
| *Estimators* | WSNR | 0.582 | 0.443 | 0.648 |
| *Based on* | NQM | 0.600 | 0.461 | 0.666 |
| *HVS* | VSNR | 0.607 | 0.466 | 0.738 |
| *Properties* | C4 | 0.822 | 0.636 | 0.832 |
| *Estimators* | SSIM | 0.870 | 0.696 | 0.883 |
| *Based on* | MS-SSIM | 0.865 | 0.679 | 0.733 |
| *Hypothesized* | VIF | *0.929* | *0.774* | *0.950* |
| *HVS* | VIF* | *0.938* | *0.799* | **0.959** |
| *Objectives* | $\text{NICE}_{\text{Sobel}}$ | *0.932* | *0.780* | 0.885 |
| *Proposed* | $\text{NICE}_{\text{Canny}}$ | *0.914* | *0.746* | 0.934 |
| *Utility* | $\text{MS-NICE}_1$ | *0.935* | *0.784* | 0.875 |
| *Estimators* | $\text{MS-NICE}_2$ | *0.937* | *0.789* | 0.860 |
| | $\text{MS-NICE}_3$ | *0.940* | *0.796* | 0.855 |
| | $\text{MS-NICE}_4$ | **0.946** | **0.810** | 0.855 |

ceived quality scores. The difference between VIF*'s estimates and the perceived quality scores exhibited the smallest variance (i.e., smallest RMSE), so the $p$-value for the BFL test is reported when the residuals of estimates from each objective estimator when used as quality estimates were compared with that of VIF*.

The remainder of this section reports the key results, which appear in boldface, followed by a summary of the results for subsets of objective estimators that exhibit similar performance with headings for the subsets of estimators appearing in italics. Statistical justifications, general interpretations, and specific remarks about the objective estimators are reported.

Table 4.4: Statistics summarizing the accuracy of objective estimators serving as quality estimators. The root mean squared error (RMSE), the outlier ratio (OR), and the resolving power $RP_\alpha$ for $\alpha = 0.1, 0.05$ are reported when the estimates are compared with the perceived quality scores for all test images ($n = 243$). Optimal values appear in boldface with statistically equivalent values italicized. An asterisk beside the RMSE indicates that the residual variance is statistically equivalent to that of VIF* according to the Brown-Forsythe-Levene test at the 95% confidence level. The skewness (Skew.) and kurtosis (Kurt.) of the residuals are italicized when the Jarque-Bera (JB) test indicated that the residuals belong to a Gaussian distribution.

| | | *Estimating Perceived Quality: Accuracy Measures* | | | | | |
|---|---|---|---|---|---|---|---|
| | *Estimator* | RMSE | OR | $RP_{0.10}$ | $RP_{0.05}$ | Skew. | Kurt. |
| *Spectral Slope* | $\beta$ | 0.895 | 0.835 | 1.749 | 1.902 | -0.27 | 2.1 |
| *Sig. Fidelity Measures* | PSNR | 0.833 | 0.506 | 1.720 | 1.949 | -0.81 | 2.8 |
| | $C_{rms}(\boldsymbol{E})$ | 1.011 | 0.881 | 2.407 | 2.413 | -0.61 | 2.0 |
| *Estimators Based on HVS Properties* | WSNR | 0.841 | 0.823 | 1.650 | 2.052 | -0.90 | 2.8 |
| | NQM | 0.823 | 0.831 | 1.524 | 1.911 | -0.97 | 3.0 |
| | VSNR | 0.745 | 0.794 | 1.439 | 1.760 | -1.1 | 3.6 |
| | C4 | 0.615 | 0.808 | 1.59 | 1.60 | -0.47 | 2.9 |
| *Estimators Based on Hypothesized HVS* | SSIM | 0.519 | 0.700 | 2.507 | 2.517 | *-0.12* | *2.6* |
| | MS-SSIM | 0.751 | 0.831 | 2.771 | 2.779 | -0.30 | 2.1 |
| | VIF | 0.345* | **0.531** | **0.666** | **0.828** | 0.17 | 5.4 |
| | VIF* | **0.313** | *0.568* | 1.047 | 1.056 | *0.12* | *3.0* |
| *Objectives Proposed Utility Estimators* | $\text{NICE}_{\text{Sobel}}$ | 0.515 | 0.786 | 2.066 | 2.076 | -0.64 | 2.9 |
| | $\text{NICE}_{\text{Canny}}$ | 0.394* | 0.568 | 0.90 | 1.02 | -0.29 | 3.5 |
| | $\text{MS-NICE}_1$ | 0.535 | 0.778 | 2.246 | 2.256 | -0.77 | 3.1 |
| | $\text{MS-NICE}_2$ | 0.563 | 0.765 | 2.395 | 2.405 | -0.79 | 3.1 |
| | $\text{MS-NICE}_3$ | 0.572 | 0.782 | 2.281 | 2.291 | -0.73 | 3.0 |
| | $\text{MS-NICE}_4$ | 0.572 | 0.757 | 2.245 | 2.254 | -0.69 | 3.0 |

**Estimators that overemphasize the significance of distortions to low-frequency content perform poorly as quality estimators over a variety of distortions.** *The spectral slope, signal fidelity measures, and most of the objective estimators based on HVS properties perform poorly as quality estimators over a variety of distortions.* Estimates from these estimators, excluding C4, weakly correlate with and inaccurately estimate the perceived quality scores. Specifically, the linear correlation between the estimates from these estimators and the perceived quality scores ($r < 0.74$) imply that these estimators account for no more than 55% of the variation in quality. The rank correlation statistics ($\rho < 0.63$, $\tau < 0.48$) signify that a nonlinear mapping would not significantly improve the performance of these estimators as quality estimators.

For these estimators, the RMSE between each estimator's linearly mapped estimates and the perceived quality scores exceeds 0.75. A difference of 1 in perceived quality corresponds to a different quality category (i.e., "Fair" versus "Good"). The variance of the residuals when estimates from these estimators are used to estimate quality are statistically larger than that of VIF*, which has the smallest residual variance, according to the BFL test. The outlier ratios for these estimators indicate that their estimates of perceived quality scores for no fewer than 51% and, in most cases, more than 82% of the distorted images, lie outside the 95% confidence intervals of those distorted images' perceived quality scores. Last, the resolving power for these estimators imply that these estimators reliably distinguish distorted images with differences in perceived quality no smaller than 1.44 and 1.76 at the 10% and 5% significance levels, respectively.

The TS+HPF distortions are largely responsible for the poor performance of these estimators as quality estimators. In fact, when each estimator was analyzed with the TS+HPF distortions removed from the test image set, the interpretation of the performance of these estimators changes: the correlation and accuracy statistics of these estimators improved. Apart from the spectral slope and $C_{rms}(\boldsymbol{E})$, these objective estimators previously have been evaluated as quality estimators on other image databases that do not include distortions that deliberately disrupt the low-frequency content without severely disrupting the high-frequency content [20, 64, 116]. The performance of these estimators on the current database of test images, which includes distortions that disrupt low-frequency content without severely disrupting high-frequency content (i.e., the TS+HPF distortions for small $\gamma$), demonstrates that these estimators were designed and tested under the assumption that either 1) distortions will not compromise the integrity of the low-frequency content 2) distortions to low-frequency content will coincide with severe distortions to high-frequency content or 3) distortions to low-frequency content have a negligible impact on quality. However, the current results indicate that these different assumptions do not reflect the *general* image characteristics that influence judgments of perceived quality. Namely, the loss of low-frequency content without severely disrupting high-frequency content coincides with a significant decrease in quality.

The spectral slope, as discussed in Section 4.3.1, quantifies the shape of the distorted image's frequency response, which varies for the different distortions. However, the correlation between the spectral slope and the perceived quality scores is significantly lower than the correlation between the spectral slope and the perceived utility scores. Specifically, the spectral slope accounts for 53% of the variation of utility but only 34% of the variation in quality. An analysis of the relationship between the spectral slope and the perceived quality scores revealed that TS+HPF distorted images have spectral slopes similar to TS and J2K+DCQ distorted images, but TS+HPF distorted images have significantly lower perceived quality. Thus, the spectral slope is an unreliable indicator of quality over a variety of distortions.

The signal fidelity measures as well as the estimators based on HVS properties, excluding C4, produce estimates that are a function of the energy of the reference and test images and account for distortions to low-frequency content, which, according to the subjective experiments, significantly affects quality. However, these estimators overemphasize the significance of distortions to low-frequency content and underestimate the perceived quality scores of TS+HPF distorted images.

*C4 and SSIM perform satisfactorily as quality estimators but MS-SSIM does not.* Among all the estimators evaluated, estimates from both C4 and SSIM moderately correlate with and accurately estimate the perceived quality scores. Specifically, the linear correlation between their estimates and the perceived quality scores (C4: $r = 0.83$, SSIM: $r = 0.88$) imply that these estimators account for no less than 68% of the variation in quality (C4) and no more than 77% (SSIM) .

The RMSE between C4's and SSIM's estimates and perceived quality is 0.62 and 0.52, respectively, which correspond to estimates within a quality category of the perceived quality score. The variance of the residuals when estimates from these two estimators are used as quality estimates are statistically greater than that of VIF*, which has the smallest residual variance, according to the BFL test. The outlier ratios for these estimators indicate that their estimates of perceived quality for 70% (SSIM) and 81% (C4) of the distorted images lie outside the 95% confidence intervals of those distorted images' respective perceived quality scores. Last, the resolving power for these estimators imply that these two estimators reliably distinguish distorted images with differences in perceived quality no smaller than 1.6 at both the 10% and 5% significance levels with C4 exhibiting smaller resolving powers.

In contrast with both C4 and SSIM, estimates from MS-SSIM weakly correlate with and inaccurately estimate the perceived quality scores. Specifically, the linear correlation between MS-SSIM's estimates and the perceived quality scores indicate that MS-SSIM accounts for only 55% of the variation in quality. The RMSE between MS-SSIM's estimates and the perceived quality scores is 0.75, which nearly corresponds to one quality category. The resolving power for MS-SSIM exceeds 2.7 at both the 10% and 5% significance levels and suggests that MS-SSIM can only reliably distinguish gross differences in perceived quality corresponding to nearly three quality categories (i.e., "Bad" versus "Good") over a variety of distortion types.

An analysis of the relationship between the estimates from C4, SSIM, and MS-SSIM and the perceived quality scores revealed that their accuracy decreases as quality decreases, which indicates that their analyses of the reference and test images degenerate as quality decreases. However, the strong Spearman rank correlation ($r > 0.82$) between perceived quality and the estimates from these three estimators suggest that they each exhibit a nonlinear, monotonic relationship with the perceived quality scores. Fitting the estimates from these estimators to the perceived quality scores with a monotonic, nonlinear mapping significantly changes the interpretation of their performance as quality estimators: each significantly improves as a quality estimator. Each of these estimators analyze distortions to low-frequency content, as discussed in Section 4.3.1, and the subjective experiments demonstrate that distortions to low-frequency content affect perceived quality. However, even with a nonlinear mapping these estimators overemphasize distortions to low-frequency content and underestimate the perceived quality of TS+HPF distorted images.

**Estimators that analyze all frequency content without overemphasizing the significance of distortions to low-frequency content accurately estimate perceived quality scores over a variety of distortions.** *VIF* pro-*

*duces more reliable estimates of perceived quality scores than VIF over a variety of distortions.* Estimates from VIF strongly correlate with and accurately estimate perceived quality scores, and most of VIF's correlation and accuracy statistics are statistically equivalent to those of VIF*. The linear correlation between estimates from both VIF and VIF* and the perceived quality scores indicate that each account for more than 90% of the variation in quality, and the accuracy between estimates from both VIF and VIF* and the perceived quality scores are small (RMSE $< 0.35$) and statistically equivalent. However, the resolving power for VIF is smaller than the resolving power for VIF* at both the 90% and 95% confidence levels. In particular, the resolving powers for VIF and VIF* indicate that they reliably distinguish distorted images with differences in perceived quality scores no smaller than 0.83 and 1.1 at the 5% significance level, respectively.

VIF distinguishes smaller differences among distorted images with high perceived quality more reliably than VIF*, which results in smaller resolving powers for VIF, because VIF is more sensitive to disruptions to high-frequency content than VIF*. Modest disruptions to high-frequency content (i.e., textures) affect the perceived quality of high quality yet visibly distorted images. However, distortions to low-frequency content have a greater affect on perceived quality than distortions to high-frequency components (see Section 3.4), and VIF* is more sensitive to low-frequency distortions than VIF. As a consequence, VIF* estimates the perceived quality scores of TS+HPF distortions more accurately than VIF, which results in the slightly smaller, though not statistically significant, RMSE observed for VIF* as compared to VIF. However, VIF overestimates the perceived quality scores of TS+HPF distorted images, because disruptions to low-frequency content do not affect estimates from VIF unless they accompany severe disruptions to high-frequency content. VIF*, however, analyzes the low-frequency content. In short, VIF performs well as a quality estimator for applications that do not encounter distortions such as the TS+HPF distortions that disrupt low-frequency content without severely disrupting high-frequency content. However, VIF* performs well as a quality estimator across a variety of distortions, because its modifications to VIF normalize the individual channel measurements based on the energy distribution of the reference image across image scales (see Section 2.3.5).

**Estimators that measure degradations to image contours perform poorly as quality estimators over a variety of distortions.** *$NICE_{Sobel}$ and the various implementations of MS-NICE produce unreliable estimates of perceived quality across a variety of distortions.* Estimates from these estimators strongly correlate with and estimate with moderate accuracy perceived quality scores. Specifically, the linear correlation between the estimates from these estimators and the perceived quality scores ($r > 0.85$) imply that these estimators account for at least 72% but no more than 78% of the variation in quality.

The RMSE between each estimator's linearly mapped estimates and the perceived quality scores are approximately 0.5, which indicates that errors in quality estimates correspond to less than one quality category. The variance of the residuals when estimates from these estimators are used as quality estimates are statistically larger than that of VIF*, which has the smallest residual variance, according to the BFL test. For no fewer than 76% of the distorted images, these estimators produce estimates that lie outside the 95% confidence intervals of those distorted images' respective perceived quality scores based on the outlier ratios. The resolving powers for these estimators imply that these estimators reliably distinguish distorted images with differences in perceived quality no smaller than 2 at both the 10% and 5% significance levels.

A nonlinear relationship between the perceived quality scores and the estimates from both $\text{NICE}_{\text{Sobel}}$ and $\text{MS-NICE}_{S\leq 4}$ was observed and quantified by their strong Spearman correlation statistics ($\rho > 0.93$). Further analysis of this nonlinear relationship revealed that small degradations to contours, as measured by both $\text{NICE}_{\text{Sobel}}$ and $\text{MS-NICE}_{S\leq 4}$, correspond to large changes in the perceived quality scores. In other words, distorted images with high perceived quality scores primarily exhibit visible degradations to textures, and both $\text{NICE}_{\text{Sobel}}$ and $\text{MS-NICE}_{S\leq 4}$ do not measure degradations to image textures, which influence perceived quality. Furthermore, distorted images with very low perceived quality exhibit large changes in contours, as measured by $\text{NICE}_{\text{Sobel}}$ and $\text{MS-NICE}_{S\leq 4}$, but exhibit very little change in perceived quality. Thus, heavily distorted images (i.e., very low perceived quality) exhibit strong variations in signal characteristics that correspond to very small changes in perceived quality. This follows if one considers again a reference/distortion sequence beginning with an unrecognizable image and evolving toward a useful, medium quality image. The dramatic perceptual changes in subsequent images near the recognition threshold will coincide with significant variations in the underlying signal characteristics, especially the emergence of contours, as detected by $\text{NICE}_{\text{Sobel}}$ and $\text{MS-NICE}_{S\leq 4}$. Despite these dramatic perceptual changes, the perceived quality scores of these images are still very low relative to the undistorted reference images.

For $\text{NICE}_{\text{Sobel}}$ and $\text{MS-NICE}_{S\leq 4}$, a monotonic, nonlinear mapping increases the correlation between their objective estimates and the perceived quality scores to at least 0.94 and is statistically larger for $\text{MS-NICE}_4$ ($r = 0.97$). The nonlinear mapping also reduces the RMSE to less than 0.41 and is smallest for $\text{MS-NICE}_4$ (RMSE $= 0.28$). The fitted nonlinearity expands small differences among estimates from $\text{NICE}_{\text{Sobel}}$ and $\text{MS-NICE}_{S\leq 4}$ for distorted images with high perceived quality and compresses large differences among estimates from $\text{NICE}_{\text{Sobel}}$ and $\text{MS-NICE}_{S\leq 4}$ for distorted images with low perceived quality. Among the single- and multi-scale implementations of NICE, $\text{MS-NICE}_4$ exhibits the best performance as quality estimator when fitted with a nonlinear mapping, because, as discussed in Section 4.3.1, implementations of $\text{MS-NICE}_S$ for larger S are more sensitive to low-frequency distortions than the other versions (i.e., NICE and $\text{MS-NICE}_{S\leq 2}$), which analyze distortions to high-frequency content.

Although the monotonic, nonlinear mapping changes the interpretation of the performance of $\text{NICE}_{\text{Sobel}}$ and $\text{MS-NICE}_{S\leq 4}$ as quality estimators, the parameters of this nonlinearity may vary for distortions not included in the current collection of test images. The current results cannot definitively establish that using both $\text{NICE}_{\text{Sobel}}$ and $\text{MS-NICE}_{S\leq 4}$ with a tuned nonlinear mapping provides reliable and accurate estimates of perceived quality over a variety of distortion types.

*$\text{NICE}_{\text{Canny}}$ performs poorly as a quality estimator for medium quality distorted images.* Over the *entire* collection of distorted images, estimates from $\text{NICE}_{\text{Canny}}$ exhibit correlation and accuracy statistics as a quality estimator that are statistically equivalent to those of VIF* when considering the entire collection of distorted images. However, the performance of $\text{NICE}_{\text{Canny}}$ as a quality estimator is not consistent for different regions of quality. Specifically, estimates from $\text{NICE}_{\text{Canny}}$ exhibit statistically weaker linear correlation with the perceived quality scores ($r = 0.62$) than VIF* ($r = 0.82$) for distorted images with medium quality (i.e., perceived quality scores between $[2.25, 3.75]$). Furthermore, the RMSE between estimates using both VIF* and $\text{NICE}_{\text{Canny}}$ and perceived quality scores are 0.28 and 0.42, respectively, for medium quality distorted images, and the variance of the residuals are statistically smaller for VIF* than $\text{NICE}_{\text{Canny}}$. In both the low

and high quality regions the performance statistics for VIF* and NICE$_{\text{Canny}}$ are statistically equivalent.

The relationship between NICE$_{\text{Canny}}$ and the perceived quality scores is consistent with the relationship observed between perceived quality scores and perceived utility scores: perceived utility is unreliably predicted from perceived quality for medium quality distorted images. Likewise, NICE$_{\text{Canny}}$ estimates the perceived quality less reliably for distorted images with medium quality. TS+HPF and TS distorted images with equal $\gamma$ formed from the same reference image have very similar values for NICE$_{\text{Canny}}$, which is consistent with their equal perceived utility scores yet different perceived quality scores. NICE$_{\text{Canny}}$ overestimates the quality of TS+HPF distorted images, because it does not analyze distortions to low-frequency content, whereas VIF* does and most accurately estimates the perceived quality of TS+HPF distorted images.

### 4.3.3 Results: Summary

When estimating perceived utility scores, objective estimators that analyze the high-frequency content of the reference and test images outperform those estimators that also analyze the low-frequency content of the reference and test images. Specifically, VIF, NICE$_{\text{Sobel}}$, NICE$_{\text{Canny}}$ , and MS-NICE$_{S\leq 2}$ produce the most reliable estimates of perceived utility scores. The interpretation of both SSIM and MS-SSIM as utility estimators changes when they operate without the components that analyze low-frequency content (i.e., the mean component and, in the case of MS-SSIM, also the variance component): both estimators provide more accurate estimates of perceived utility than their original implementations.

$\text{NICE}_{\text{Canny}}$ produces the most accurate estimates of the perceived utility scores when the *skier* and *caged birds* images were discarded. These images reveal two limitations of $\text{NICE}_{\text{Canny}}$: 1) detection of less visible contours (e.g., those in snow region in the *skier* image) and 2) separate analysis of relevant versus irrelevant contours (e.g., the birds versus the bars of the cage in the *caged birds* image). Despite these limitations, $\text{NICE}_{\text{Canny}}$ demonstrates that perceived utility scores can be reliably estimated from an analysis of image contour degradation.

When estimating perceived quality scores, estimates from VIF* most accurately estimate the perceived quality scores. Unlike many of the other objective estimators, VIF* analyzes both high-frequency content and low-frequency content of the reference and test images without overemphasizing disruptions to low-frequency content. Several other estimators grossly underestimate the perceived quality scores of TS+HPF distorted images, because these estimators analyze low-frequency content but overemphasize the effect of distortions to low-frequency content. VIF* weights the relative influence of distortions to low- and high-frequency content on its estimates in a manner that yields accurate estimates of perceived quality.

## 4.4 Discussion

The subjective experiments establish that perceived quality is not a suitable proxy for perceived utility. An evaluation of objective estimators as both utility and quality estimators revealed that an analysis of degradations to high-frequency content and, specifically, image contours produces accurate estimates of perceived utility, whereas a properly weighted analysis of degradations across all frequency content produces accurate estimates of perceived quality. This section discusses 1) the image characteristics revealed by objective estimators that impact perceived utility and perceived quality and 2) the relationship between object recognition, perceived utility, and the analysis conducted by NICE.[5]

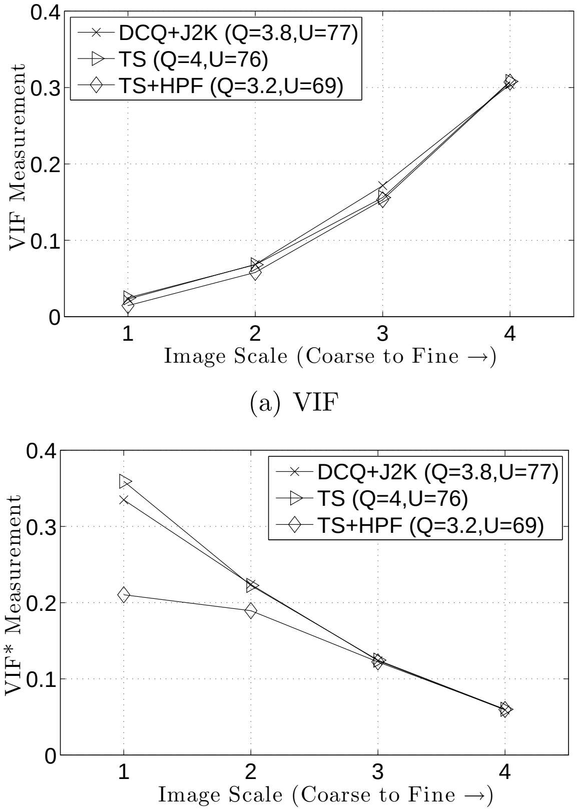
### 4.4.1 Objective Estimators Reveal Image Characteristics that Impact Utility and Quality

Among the objective estimators investigated, VIF and NICE performed best as utility estimators, and VIF* performed best as a quality estimator. First, the signal analyses conducted by VIF* and VIF are analyzed and compared, since the distinctions between VIF* and VIF reiterate the conclusion drawn from the subjective experiments that low-frequency content affect perceived utility but not quality. Second, the signal analyses conducted by VIF and NICE are analyzed and compared, since VIF and NICE illustrate different uses of high-frequency content to estimate utility. Last, the impact that an edge-detector used with NICE has on its performance as a utility estimator for other distortions is discussed.

#### VIF versus VIF*: Low-frequency Content Affects Quality

VIF and VIF* analyze the reference and test images using the steerable pyramid decomposition [117], which models the well-accepted multi-channel characterization of the analysis conducted by the human visual system in the primary visual cortex [31] (A mathematical description of VIF and VIF* is presented in Chapter

[5] We use "NICE" to generically refer to both the single-scale and multi-scale implementations of NICE, and specific implementations of NICE (e.g., $\text{NICE}_{\text{Canny}}$) will be identified when necessary.



(a) VIF

(b) VIF*

Figure 4.1: VIF is more sensitive to distortions at finer image scales (i.e., high spatial frequencies) over those at coarser image scales (i.e., low spatial frequencies), whereas VIF* is more sensitive to disruptions to coarser scale content than finer scale content. Figures 4.1(a) and 4.1(b) respectively show the image scale measurements computed byVIF and VIF* for the *airplane* image with J2K+DCQ ($Q$ = 3.8, $U$ = 77), TS ($Q$ = 4.0, $U$ = 76), and TS+HPF ($Q$ = 3.2, $U$ = 69) distortions. These images have statistically equivalent perceived utility, but the perceived quality of the TS+HPF distorted image is statistically smaller than the other two distorted images. The pooled image scale measurements for VIF reflect their similarity in perceived utility but not their differences in perceived quality. The pooled image scale measurements for VIF* reflect their differences in perceived quality not their similarity in perceived utility.

2). VIF and VIF* compute and linearly pool spatially local signal-to-noise ratios within each channel, which produces a *channel measurement* that quantifies the fidelity of the test image with respect to the reference image within that channel. The channel measurement values decrease as the fidelity of the test image with respect to the reference image within that channel decreases (i.e., the test image contains more distortion). The sum of the channel measurements from the same image scale yield *image scale measurements* that quantify the fidelity of the test image with respect to the reference image within that image scale. Because the steerable pyramid decomposition represents a coarser image scale with half as many coefficients as the next finest image scale (i.e., due to decimation), the finer image scale measurements are larger than the coarser image scale measurements. VIF linearly pools image scale measurements to produce an objective estimate for the test image, and image scale measurements at finer image scales dominate VIF's objective estimate. In contrast, VIF* normalizes each image scale measurement by the number of coefficients in that image scale, which balances the measurements from different image scale measurements, before linearly pooling. Natural images exhibit a $1/f^{\alpha}$ power spectra [39], and, as a consequence, the normalized image scale measurements at coarser image scales dominate VIF*'s objective estimate. As a result, VIF* is more sensitive to disruptions to coarser image scale content than finer image scale content.

Images from the *airplane*/J2K+DCQ, *airplane*/TS, and *airplane*/TS+HPF sequences that have statistically equivalent perceived utility are evaluated using VIF and VIF* to illustrate the differences between VIF and VIF*. The image from the *airplane*/TS+HPF sequence has the same parameter $\gamma$ as the image from the *airplane*/TS sequence and statistically has the smallest perceived quality. Figure 4.1 shows the image scale measurements from VIF and the *normalized* image scale measurements from VIF* for these three images. The image scale measurements from VIF are much larger at finer image scales (i.e., high spatial frequencies) than coarser image scales (i.e., low spatial frequencies) and exhibit very little variation among these four distorted images across all image scales. Thus, for these images, VIF's pooled image scale measurements reflect their similarity in perceived utility but not their differences in perceived quality. In contrast, the normalized image scale measurements from VIF* are larger at coarser scales than finer scales and indicate a difference between the *airplane*/TS+HPF image and the other distorted image at the coarsest image scale. Thus, for these images, VIF*'s pooled image scale measurements reflect their differences in perceived quality and not their similarity in perceived utility.

The analyses conducted by VIF* and VIF are consistent with the subjective experiments. The absence of low-frequency content (i.e., the TS+HPF distorted images versus TS distorted images with the same $\gamma$) significantly and consistently affects quality but has less consistent effects on the utility. Since VIF and the various implementations of NICE outperform the other objective estimators as utility estimators, the fidelity of low-frequency content does not strongly influence utility. The low-frequency content represents the *shading* in grayscale natural images, which forms the appearance of naturalness due to interactions between object surfaces and lighting. Natural images with undisrupted shading are visually consistent with our daily experiences with natural environments. Disruptions to an image's shading decrease its perceived quality, which the objective estimates produced by VIF*, not VIF, accurately reflect due to normalizing image scale measurements before pooling across image scales.

### Comparing VIF and NICE: Estimates of Image Contour Degradation

Fine-scale signal components describe natural image details corresponding to both object boundaries and textures, and the energy of the fine-scale signal components coincides with the visibility of these details. VIF and NICE, both of which perform best as utility estimators, specifically analyze the energy of fine-scale signal components of the reference and test images to produce an objective estimate of the test image's perceived utility. Both objective estimators[6] filter the images using two channels that separate the fine-scale signal components into horizontally and vertically oriented spatial frequency components. VIF and NICE illustrate two possible uses of the fine-scale signal components to estimate perceived utility.

VIF subjects the high-frequency channel responses for the reference and test images to a normalization mechanism functionally similar to divisive normalization (i.e., a model of gain control) that normalizes channel responses to a particular range for subsequent processing stages [113,115,141]. Divisive normalization models the relationship between the $n^{th}$ neuron's response $y_n$ to its input $t_n$ according to

$$y_n = \frac{t_n^p}{b^q + \sum_{m \in \mathcal{M}_n} w_m t_m^q}, \tag{4.3}$$

where $b$ is a positive saturation constant, $\mathcal{M}_n$ is a set of indices specifying local spatial, frequency, and orientation neuron responses to input $t_n$, the $w_m$ are weights applied to those local responses before pooling, and the exponents $p$ and $q$ are positive values that model a power-law relationship between a neuron's input and output.

VIF approximates the divisive normalization model by normalizing the channel

[6] Using the fine-scale steerable pyramid filters to identify image contours for MS-NICE lead to statistically similar performance to the single-scale implementation of NICE using the Sobel and Canny edge-detectors.

responses based on the energy (i.e., in Eq. (4.3) set $b = 0$ and $p = q = 2$) of their spatially *local* channel responses. That is, VIF performs spatially local *variance* normalization. Image contours generally elicit larger channel responses than textures, and following a spatially local variance normalization, the channel responses to both contours and textures are normalized to the same range. As a consequence of this normalization, estimates from VIF reflect *any* disruption to the high-frequency channel responses due to the distortions, so disruptions to *both* image contours and image textures affect VIF's objective estimates.

In contrast with VIF, NICE detects the edges in the reference and test images and can be viewed as performing spatially *global* variance normalization, collinear facilitation [67], and hard thresholding. NICE and MS-NICE perform global variance normalization by normalizing the channel responses based on the average channel response energy.[7] Global variance normalization reduces the magnitude of *all* the channel responses, so channel responses to image contours remain larger than those to textures.

Collinear facilitation describes the perceptual facilitation and suppression of channel responses due to interactions (i.e., connected cells) among spatially local and similarly oriented channel responses and suggests that mechanisms mediate the perception of smooth curves from line segments [92, 93]. In particular, studies of human observers report that the detection contrast of a target Gabor patch spatially flanked by two high contrast Gabor patches is highest (i.e., the target is difficult to detect) when the flanking patches are spatially very close to and have the same orientation as the target, whereas the target detection contrast is lowest (i.e., the target is easy to detect) when the spatial distance between the

[7] Applying a hard-threshold defined as $\beta$ to a signal $f$ (i.e., values of $f$ satisfying $f > \beta$ are set to 1 and otherwise zero) is viewed as normalizing the signal $f$ by $\beta$ (i.e., $f/\beta$) followed by hard-thresholding with threshold equal to 1.

flanking patches and the target is large and oriented orthogonal to the target patch [92]. Furthermore, the target detection contrast is lowest when the global orientation of the line formed by the three patches coincided with the individual patch orientations [93]. All of the edge detectors used for NICE crudely perform collinear facilitation via a thinning operation that retains local maxima.

Hard thresholding removes low energy channel responses, which largely coincide with textures, and is hypothesized to represent a decision process performed at a later stage of the human visual system corresponding to object perception. Disruptions to image textures have a negligible impact on NICE's objective score, since NICE reflects disruptions to image contours due to the distortion process.

Because NICE primarily measures degradations to image contours, we analyzed estimates of VIF when decomposed into separate fidelity measurements for contours and textures. Specifically, VIF was decomposed as

$$\mathrm{VIF} \approx \mathrm{VIF}_{\mathrm{contour}} + \mathrm{VIF}_{\mathrm{texture}}, \tag{4.4}$$

where $\mathrm{VIF}_{\mathrm{contour}}$ and $\mathrm{VIF}_{\mathrm{texture}}$ respectively represent VIF evaluated on contour and texture components of an image. Estimates from both $\mathrm{VIF}_{\mathrm{contour}}$ and $\mathrm{VIF}_{\mathrm{texture}}$ were evaluated in terms of their performance as utility estimators. The correlation statistics for $\mathrm{VIF}_{\mathrm{contour}}$ increase relative to those for VIF, whereas all of the correlation statistics for $\mathrm{VIF}_{\mathrm{texture}}$ are statistically smaller than those of VIF. The RMSE of $\mathrm{VIF}_{\mathrm{contour}}$ is 10.7, but the residual variance is statistically equivalent to that of VIF (RMSE=12.4). However, the RMSE for $\mathrm{VIF}_{\mathrm{texture}}$ is 18.3 and is statistically larger than that of VIF. In short, $\mathrm{VIF}_{\mathrm{contour}}$ accurately estimates the perceived utility scores as a function of the fidelity of the contour information.

In summary, VIF analyzes disruptions to both contours *and* textures while excluding disruptions to low-frequency content, whereas NICE primarily analyzes disruptions to contours to estimate utility. The performance of $\text{VIF}_{\text{contour}}$ as a utility estimator is parallels the performance of NICE, which corroborates the hypothesis that contour degradations coincide with decreased perceived utility.

### Edge-detectors Impact the Performance of NICE

NICE operates in conjunction with an edge detector and was assessed using three different edge detectors. As a utility estimator, NICE operating with the Canny edge detector (i.e., $\text{NICE}_{\text{Canny}}$) and excluding the *skier* and *caged birds* distorted images, outperformed NICE operating with the other edge detectors. The performance of $\text{NICE}_{\text{Canny}}$ as a utility estimator was justified in terms of the agreement of its identified edges with object boundaries identified by humans: compared with human ground truth, the Canny edge detector ranked highest among the three edge detectors (see Section 4.3.1). Despite the performance of $\text{NICE}_{\text{Canny}}$ as a utility estimator, the current database does not include distorted artifacts that are uncorrelated with the reference image (e.g., independent, additive white Gaussian noise), and the Canny edge detector frequently identifies false contours as a result of these distortion artifacts.

Correlated distortions influence a human's perception of the distortion level more than uncorrelated distortions (i.e., independent, additive white Gaussian noise) [21, 59]. Thus, uncorrelated distortions are expected to have a smaller influence on perceived utility than correlated distortions: human observers can "ignore" moderate levels of uncorrelated distortions. NICE estimates perceived utility as a function of the errors between the reference and test edge maps produced by an edge detector: an edge detected in the reference image but absent in the test image produces an error, and an edge absent in the reference image but detected in the test image produces an error. With NICE, more errors imply lower utility, and perceived utility would be underestimated when the errors are largely due to false contours that humans would "ignore." More advanced edge detectors assess various types of edge cues including pixel value discontinuities and texture boundaries [62, 69] but generally conduct a more complex analysis of an image relative to the edge detectors tested with NICE.

The distortion types used in the experiments were spatially correlated with the reference image, so the current collection of test images cannot be used to evaluate the potential vulnerabilities of the contour detection techniques used by NICE. However, the current results based on correlated distortions demonstrate the feasibility of conducting an image contour comparison to accurately estimate perceived utility. NICE operating with robust edge detectors that do not detect false contours due to uncorrelated noise sources are expected to reliably estimate perceived utility scores for such distortions.

### 4.4.2 Object Recognition, Perceived Utility, and NICE

A perceived utility score quantifies the amount of information a distorted image conveys to a human, where the *information* of a scene included the objects and activities as well as their respective details. We hypothesize that perceived utility is linked to the level of detail with which objects and activities in the scene are recognized.

Objects in the natural world can be described with varying levels of detail, and object recognition studies using images containing one object have examined the effects of simple image filtering on the level of detail accurately recognized by a human. Such object recognition studies use the taxonomy of objects proposed by Rosch to distinguish these levels of detail, which Rosch named *levels of abstraction* [97]. As an example, a *snare drum* can be identified as a *musical instrument*, a *drum*, or a *snare drum*, where Rosch's taxonomy respectively assigns these descriptions to the *superordinate*, *basic*, and *subordinate* levels of abstraction. The object recognition studies demonstrate that humans can reliably recognize an object at the basic level using only low-frequency content, whereas subordinate-level recognition requires more high-frequency content [22, 23]. Thus, humans only perceive an object's basic-level details but not its subordinate-level details in a low-pass filtered distorted image, and this result is consistent with low-pass filtering leading to a decrease in perceived utility as subordinate-level object details disappear. The object recognition studies also concluded that humans can reliably recognize an object at both the basic and subordinate levels using only high-frequency content [22, 23]. Thus, a high-pass filtered distorted image does not affect the level of detail a human perceives about the object, and this result is consistent with high-pass filtering (i.e., TS versus TS+HPF distorted images with the same $\gamma$) often negligibly affecting perceived utility.

Another recent perceptual study of object recognition used natural images containing multiple objects of varying size and demonstrated that the number and accuracy with which humans recognized objects in distorted images decreases as the level of blur increases [127]. Furthermore, the size of the objects accurately recognized decreases as the level of blur increases (i.e., disrupting high-frequency content compromises the recognition of smaller objects). These results are consistent with the criteria proposed by Johnson, which was used to design sensors and display devices [58, 98]. The Johnson criteria relates the level of object discrimination to the detectability of a bar pattern of a given spatial frequency. For object recognition, the Johnson criteria states that a human must detect a bar grating with 4 cycles across the object's minimum dimension. Increasing the number of cycles in the bar grating across the object's minimum dimension allows the object to be more accurately identified. Our perceived utility scores are consistent with this evidence, because perceived utility decreases as high-frequency content is removed or distorted.

The object recognition studies demonstrate that loss of high-frequency content but not low-frequency content impairs object recognition performance. This evidence is consistent with our subjective experiments and suggest that our perceived utility scores, rather than perceived quality scores, estimate the amount of information recognized by a human. Such studies and our perceived utility scores provide little guidance toward understanding *how* information is recognized by a human, and in particular, what underlying image characteristics impact usefulness. However, those objective estimators (i.e., VIF, NICE, and MS-NICE) that accurately estimate perceived utility were dismantled and analyzed to understand those image characteristics that impact usefulness. In particular, NICE and MS-NICE estimate utility based on a measurement of the degradation to image contours in a distorted image with respect to a reference image.

Contours form shapes, and object *shape* is hypothesized to be a primary cue for object recognition by the human visual system [128]. Humans reliably recognize objects from line-drawings [3], which provide only object shape cues, and even from degraded line drawings [82, 102]. Line drawings abstractly represent object shapes using contours, and humans quickly identify contours formed by Gabor patches aligned along a curved path placed in an image composed of an array of randomly oriented Gabor patches [40]. The ability of humans to recognize objects from abstract contour representations along with their reported ease of detecting contours among clutter support theories of shape-based object recognition.

Another object recognition study collected fMRI data for various regions of the visual cortex to understand *how* the human visual system performs object recognition. The fMRI data was collected from both the striate (i.e. primary) and extra-striate cortex when humans viewed images that contained only contour regions, texture regions, or both (i.e., the full image) [34]. In that study, the extra-striate cortex responded greatest when humans viewed images that contain only contour regions. The increased activation due to contour information corroborates theories that object recognition is largely driven by contour information (i.e., shape perception) in natural images.

In summary, NICE performs very well as a utility estimator by extracting, comparing, and quantifying the degradation to image contour information in a distorted image with respect to a reference image. Together, the theories that contour information mediates object recognition and the performance of NICE as a utility estimator demonstrate that NICE is a viable signal analysis tool that estimates the usefulness of distorted natural images.

## 4.5 Summary

This chapter introduced the natural image contour evaluation (NICE) utility estimator. NICE as well as several objective quality estimator were evaluated as utility estimators and quality estimators using the data collected in the experiments described in Chapter 3. Two estimators were shown to accurately estimate utility. One is the visual information fidelity (VIF) criterion, which is customarily used as a quality estimator. A modification to VIF denoted VIF* was proposed that outperforms VIF as a quality estimator on the current database of distorted images. The signal analyses conducted by VIF and VIF* are consistent with the observations from the subjective experiments. Specifically, VIF primarily analyzes disruptions to high-frequency content and accurately estimates perceived utility but not perceived quality, whereas VIF* exhibits increased sensitivity to low-frequency distortions relative to VIF and analyzes disruptions to all frequency content and accurately estimates perceived quality but not perceived utility.

The natural image contour evaluation (NICE) utility estimator was also shown to accurately estimate utility. NICE estimates utility as a function of both *lost* and *introduced* contour information in a distorted image when compared with a reference image. In contrast with VIF, NICE abstractly represents the reference and test images as contours and compares these contours to estimate utility. NICE was shown to be a viable signal analysis tool to estimate the usefulness of a distorted natural image. This result supports hypotheses about the importance of contour information to the human visual system for object perception.

## 4.6 Key Points

- The natural image contour evaluation (NICE) utility estimator is introduced. Single-scale and multi-scale versions of NICE are defined. The performance of various implementations of NICE as both utility estimators and quality estimators is reported. The experimental data collected in the experiments described in Chapter 3 is used to validate the performance of each objective estimator.
- The performance objective estimators summarized in Chapter 2 as both util-

ity and quality estimators is reported. Among the estimators examined, VIF*, the modified version of the visual information fidelity (VIF) criterion introduced in Chapter 2, is demonstrated to provide more accurate estimates of perceived quality than VIF.

- VIF, which largely sensitive to distortions to high-frequency content, and NICE both are shown to provide the most accurate estimates of perceived utility using the database of images described in Chapter 3. These two estimators are analyzed and shown to largely be driven by an analysis of image contours or edges. NICE is argued to be compatible with shape-based theories of object perception.

CHAPTER 5

# A NOVEL TECHNIQUE TO ACQUIRE PERCEIVED UTILITY SCORES FROM TEXTUAL DESCRIPTIONS OF DISTORTED NATURAL IMAGES

## 5.1 Introduction

Many applications value an assessment of distorted images according to their usefulness, or utility, rather than their perceptual quality. For example, the public safety sector uses imaging systems to make immediate decisions on how best to respond to an incident [42, 43]. However, quantifying distortions in such images according to perceived quality is not a proxy for perceived utility (PU) [100, 105]. For the quality task, human observers evaluate a natural image based on its perceptual resemblance to a reference. The reference may be either an explicit, external natural image or an internal reference, only accessible to the observer. For the *utility task*, the usefulness of a natural image as a surrogate for a reference is under evaluation.

Experiments to collect PU scores of visual stimuli often require observers to perform a very specific task with an visual stimulus. For example, experiments have been conducted that prompt an observer with a task (e.g., "Identify the object being held.") prior to viewing a distorted video sequence [43]. Evaluations of distorted images need not accompany a specific task, especially if the task simply is to determine the image content as is common in video surveillance applications. Perceived utility scores have been obtained using a paired comparison methodology that instructs observers to choose among two distorted images the one that "tells you more about the content" [100]. While this method suits *content interpretation* *tasks*, the responses lack information about the specific content actually seen in the distorted images.

A novel technique for acquiring PU scores is presented that collects textual descriptions produced by observers viewing distorted images. This technique uses *observer-centric* concepts for the images that emerge from an analysis of the observer descriptions. Since observers guide the creation of concepts describing useful image content, the *observer-centric* approach establishes a framework to quantify image usefulness for a broad range of tasks. *Concept vectors* that quantify the presence of concepts appearing in observer descriptions are used to generate PU scores. Two experiments are conducted to collect PU scores using the proposed technique for a combined total of 500 distorted images that simulate scenes captured by a surveillance system.

Objective estimators are sought that provide scores consistent with subjective evaluations to circumvent expensive studies to acquire PU scores. The natural image contour evaluation (NICE), which compares the contours of a test image to those of a reference image to score the test image, has been examined as a utility estimator [105]. The capability of both the natural image contour evaluation (NICE) utility estimator, which compares contours of the reference and test images, and popular quality estimators to estimate these PU is reported. The conclusions drawn from the results reported combined with results previously reported [105] establish that a multi-scale implementation of NICE (MS-NICE) is the most robust utility estimator among the estimators evaluated, since MS-NICE consistently performs as well as estimators producing the most accurate perceived utility estimates for a variety of distortion types.

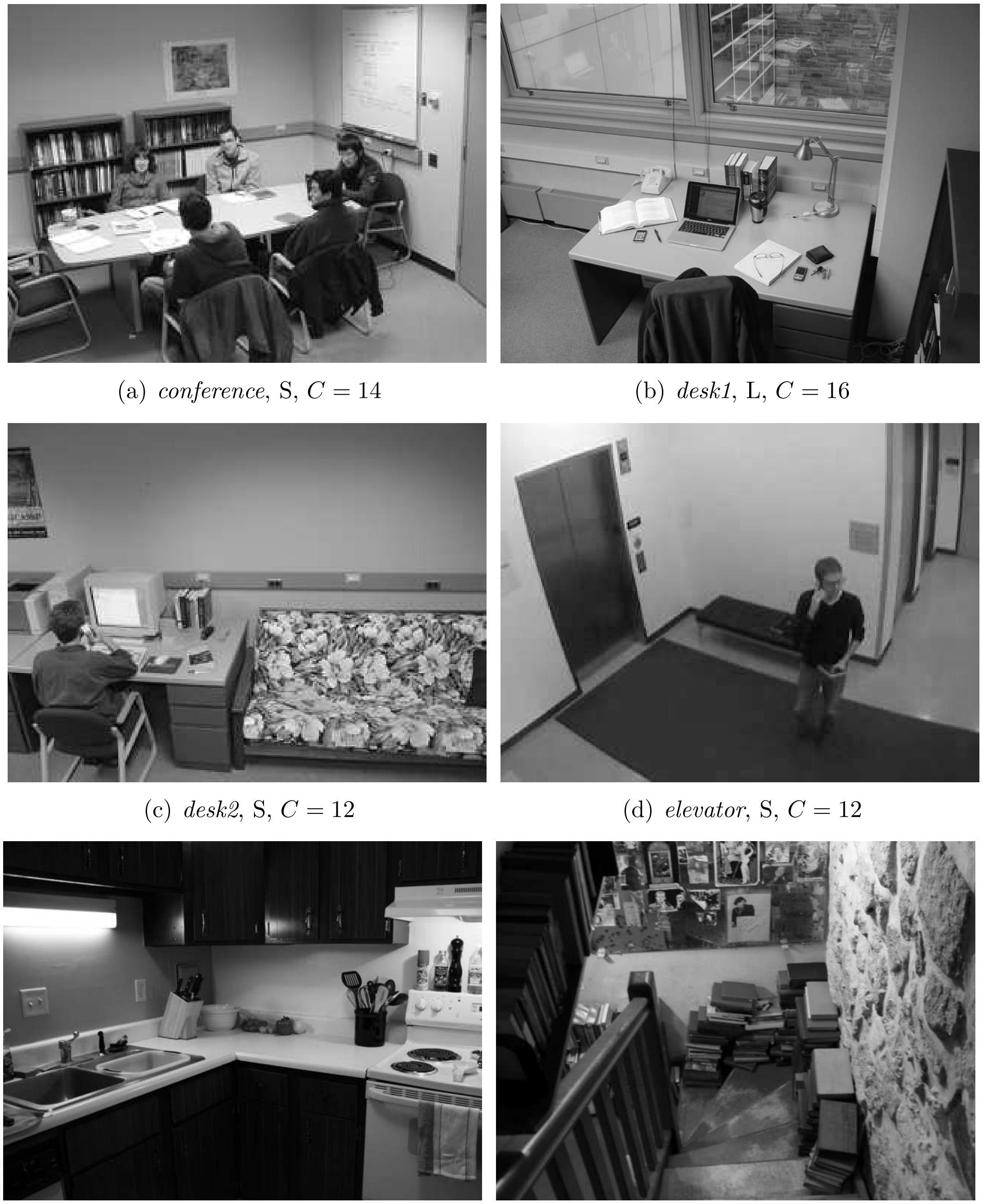
This chapter has the following organization: Section 5.2 presents the proposed methods to obtain textual descriptions from observers viewing distorted images. Section 5.3 describes the process used to generate PU scores from textual descriptions of distorted images. The PU scores obtained from the two experiments are compared in Section 5.4. The capability of NICE and several quality estimators to predict PU scores is analyzed and reported in Section 5.5, which is followed by a discussion in Section 5.6. Conclusions are presented in Section 5.7.

## 5.2 Methods: Measuring the Perceived Usefulness of Distorted Natural Images

This section describes the experimental methods used for collecting responses from observers from which perceived utility (PU) scores can be generated. Two experiments were conducted. In the first experiment, a small collection of 150 distorted images are evaluated that span three types of distortions. The second experiment contains minor procedural refinements relative to the first experiment based on exit interviews with observers in the first experiment. In addition, a large collection of 350 distorted images are evaluated that span five types of distortions. Three reference images from the first experiment are used in the second experiment, so the two experiments can be compared.

### 5.2.1 Experiment 1

The stimuli and procedures used to collect responses from observers in the first (preliminary) experiment are described.



(a) *conference*, S, $C = 14$

(b) *desk1*, L, $C = 16$

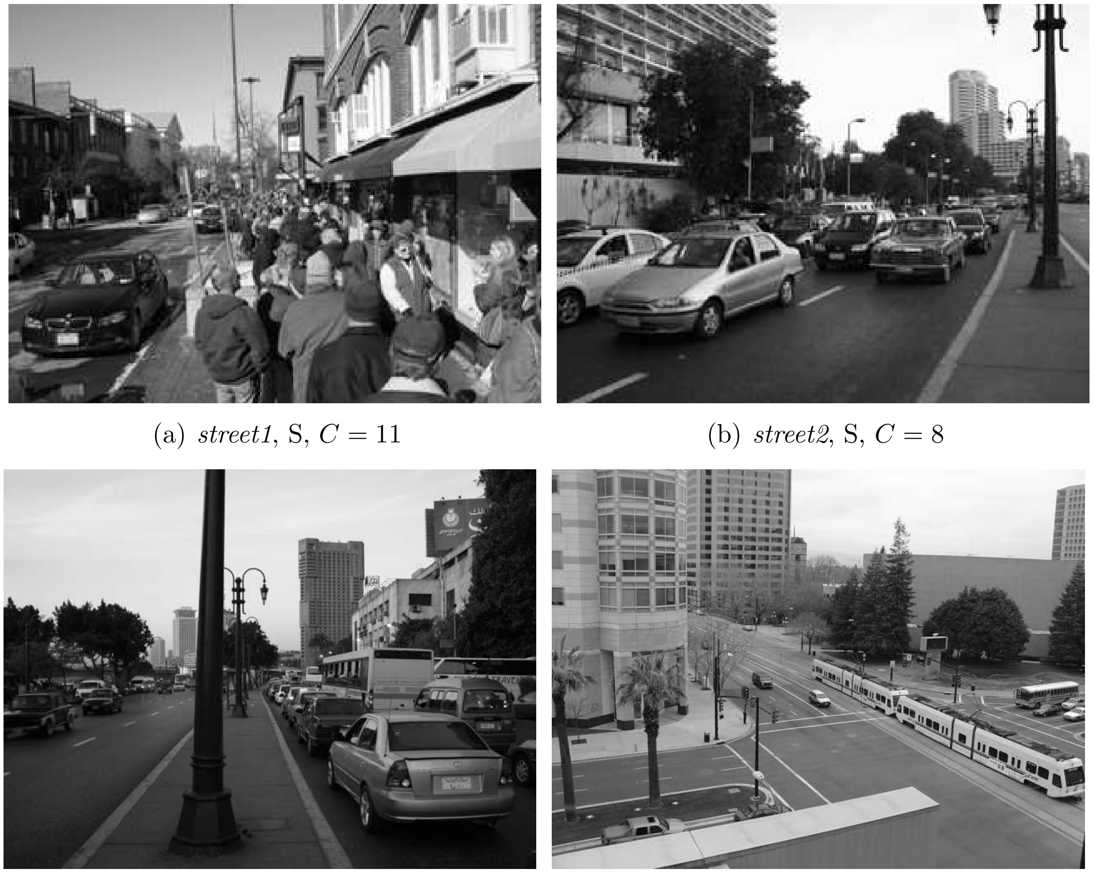
(c) *desk2*, S, $C = 12$

(d) *elevator*, S, $C = 12$

(e) *kitchen*, L, $C = 16$

(f) *stairs*, L, $C = 6$

Figure 5.1: Six of the ten grayscale natural images serving as reference images for the experiment. The images simulate hypothetical scenes captured by a video surveillance system. In the caption for each image, “L” indicates that the image size is $640 \times 480$ pixels and “S” indicates that the image size is $320 \times 240$ pixels. The number of concepts $C$ associated with each image is provided. The remaining four images are provided in Figure 5.2.



(a) *street1*, S, $C = 11$

(b) *street2*, S, $C = 8$

(c) *street3*, L, $C = 11$

(d) *tram*, L, $C = 9$

Figure 5.2: Four of the ten grayscale natural images serving as reference images for the experiment. The images simulate hypothetical scenes captured by a video surveillance system. In the caption for each image, “L” indicates that the image size is 640 × 480 pixels and “S” indicates that the image size is 320 × 240 pixels. The number of concepts $C$ associated with each image is provided. The remaining six images are provided in Figure 5.1.

**Stimuli**

Ten 8-bit grayscale natural images ($N_r = 10$) served as reference images for the first experiment (Figures 5.1 and 5.2). The images simulate scenes captured by a surveillance system. The ten reference images varied by image size: five images were 320 × 240 pixels, and the other five images were 640 × 480 pixels. The *elevator1* image is 320 × 240 pixels and was obtained directly from an AXIS 211 video surveillance camera. The other reference images were resized and cropped from larger resolution images captured by either a Sony DSC-V1 digital camera (original image size $2592 \times 1944$ pixels) or a Canon EOS Digital Rebel XTi digital camera (original image size $3888 \times 2592$ pixels). Images were resized by applying an antialiasing filter and downsampling in both horizontal and vertical directions.

From the ten reference images, 150 distorted images were formed that contained one of three types of distortions: additive white Gaussian noise; Gaussian blur; and lossy JPEG compression. For each distortion type, a *sequence of* $N_d = 5$ *distorted images* for each reference is formed such that subsequent images in the sequence contain less distortion than the preceding image, and the first image in a sequence was formed to be unrecognizable. Additive white Gaussian noise (AWGN) with noise power $\sigma_\eta^2$ was added for $\sigma_\eta = 13, 25, 50, 100$, and 200. Gaussian blur (GBLUR) was induced by convolving the reference image with a Gaussian kernel parameterized by standard deviation $\sigma_g$ for $\sigma_g = 1, 3, 5, 7$, and 9. Distortion attributed to lossy JPEG compression was induced for specific quality parameters $QP_{jpeg} = 16, 8, 4, 2$, and 1 using the quantization matrix defined in the original JPEG specification [90].

### Procedure

In the experiment,[1] observers produce typed descriptions of images. For each reference image, an observer views both the reference image and $N_d = 5$ distorted images generated from that reference image for one distortion type (e.g., an ob-

[1] The Psychophysics Toolbox is used to conduct experiments [6].

server only views *desk1* contaminated with JPEG distortions).[2] Distortion artifacts are randomly paired with reference images. An observer views and describes $(N_d + 1) \times N_r = 60$ images ($N_d = 5$ distorted images generated plus the reference image) in a testing session.

A testing session is composed of consecutive trials. In each trial, an observer views a fixation mark for 1 second, then the image to be described is displayed for $t$ seconds, and last an image of filtered Gaussian noise is displayed for 1 second. Observers provide a typed description of the image viewed for $t$ seconds,[3] where $t = 4$ or $t = 8$. The finite viewing interval forces an observer to prioritize his examination of the content and report the most important content in his description. The next trial begins after submitting a description.

The correct information about a heavily distorted image reported by an observer is sought. The trials are ordered such that an observer views the most distorted versions of the reference images first. More precisely, view a test session as being composed of $N_d + 1$ sets of $N_r$ trials, where $N_d$ is the number of distorted images generated for each of the $N_r$ reference images. In the $d^{th}$ set of $N_r$ trials, an observer views the $d^{th}$ image from the sequence of distorted images formed from each reference image. The observers view and describe the references in the last set of $N_r$ trials.

Seventeen observers participated in the experiment. The observers (one female and 16 males) were undergraduate and graduate students between the ages of 20 and 28.

[2] Observers cannot repeat the experiment for the same reference images for different distortion types, since observers may recognize *image content* in heavily distorted images from previous viewings.

[3] For 13 observers, $t = 4$ seconds, and for four observers, $t = 8$ seconds.

### 5.2.2 Experiment 2

The second experiment is an expanded version of the first experiment. Five rather than three distortion types are evaluated, and reference/distortion sequences containing $N_d = 7$ distorted images rather than $N_d = 5$ distorted images. In addition, a collection of *novel images* are mixed into the stimuli that observers view to discourage guessing based on previously viewed images.

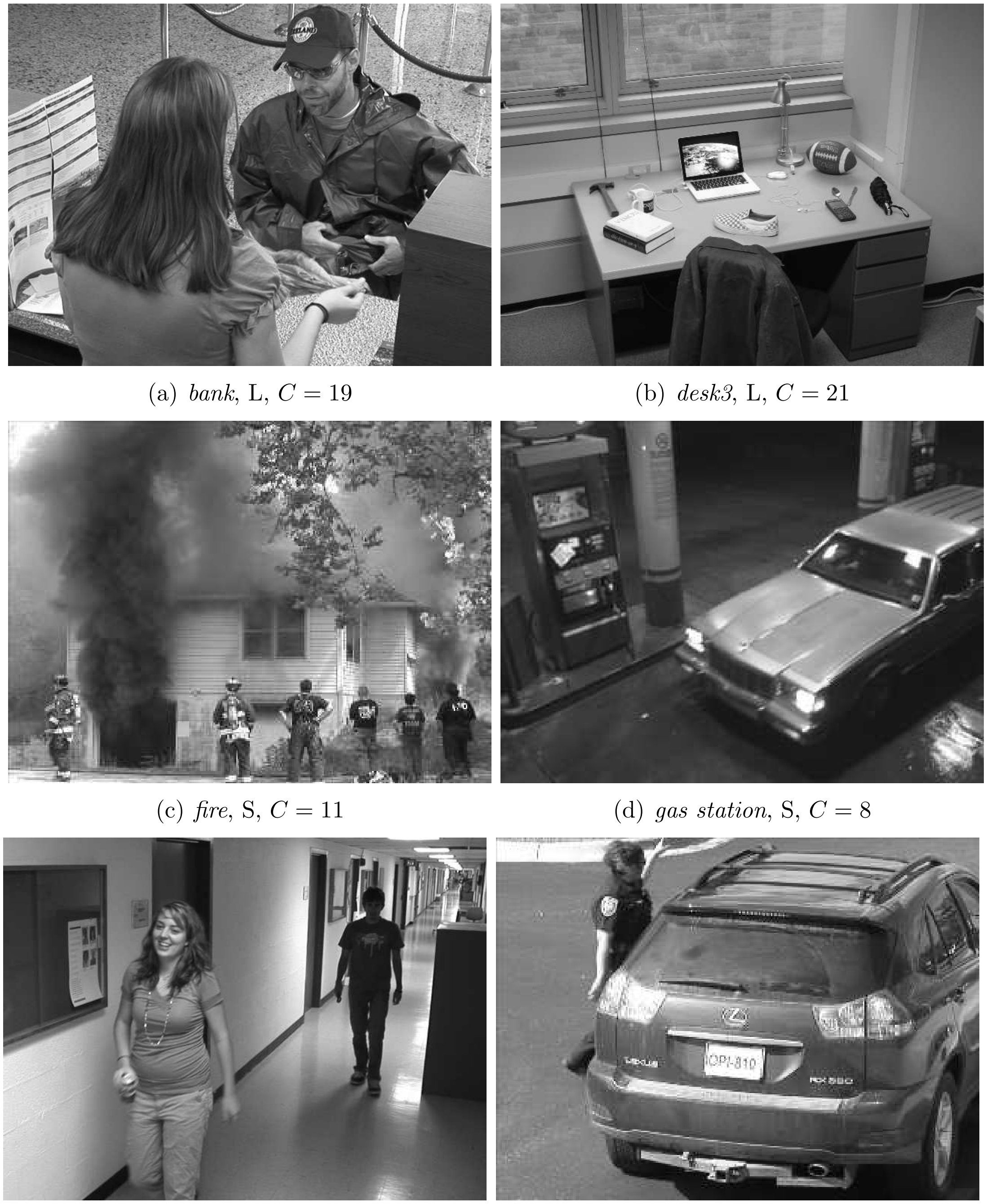
#### Stimuli

Ten 8-bit grayscale natural images ($N_r = 10$) served as reference images for the second experiment (Figures 5.3 and 5.4). The images simulate scenes captured by a surveillance system. The ten reference images varied by image size: five images were $320 \times 240$ pixels, and the other five images were $640 \times 480$ pixels. The images *conference*, *elevator*, and *kitchen* were reused from the first experiment to validate the repeatability of the proposed test methodology. The images *bank*, *fire*, *gas station*, *hallway*, *police*, and *shop* were extracted from videos in the publicly available Consumer Digital Video Library [77]. These 6 images were resized and cropped from the original versions. The image *desk3* was resized and cropped from an image captured with a Sony DSC-V1 digital camera (original image size $2592 \times 1944$ pixels). Images were resized by applying an antialiasing filter and downsampling in both horizontal and vertical directions.

From the ten reference images, 350 distorted images were formed that contained one of five types of distortions: additive white Gaussian noise; Gaussian blur; lossy JPEG compression; lossy JPEG-2000 compression using the dynamic contrast-based quantization (DCQ) strategy (J2K+DCQ); and texture-smoothing (TS). For each distortion type, a *sequence of* $N_d = 7$ *distorted images* for each reference is formed such that subsequent images in the sequence contain less distortion than the preceding image, and the first image in a sequence was formed to be unrecognizable. Such a sequence is henceforth named a *reference/distortion sequence.*

Reference/additive white Gaussian noise (AWGN) sequences were formed by varying the noise power $\sigma_\eta^2$ of the Gaussian noise added to the reference image for $\sigma_\eta = 10, 30, 50, 80, 120, 160$, and 200. Reference/Gaussian blur (GBLUR) sequences were formed by convolving the reference image with a Gaussian kernel parameterized by standard deviation $\sigma_g$ for $\sigma_g = 1, 2, 4, 6, 10, 15$, and 20. Reference/JPEG sequences were formed by compressing the reference image using baseline JPEG compression for quality parameters $QP_{jpeg} = 90, 50, 20, 8, 4, 2$, and 1 using the example quantization table provided in the original JPEG specification [79,90]. Reference/J2K+DCQ sequences were formed by compressing the reference image using a JPEG-2000 encoder with quantization step-sizes specified by DCQ to achieve encoding bitrates $R = 0.5, 0.2, 0.1, 0.05, 0.03, 0.02$ and 0.01. Reference/TS sequences were formed by smoothing textures in the reference image through soft-thresholding of Haar wavelet coefficients using the smoothing parameter $\gamma$. The reference/TS sequences contained 7 images corresponding smoothing parameters $\gamma$ that were logarithmically equally spaced from 2 to 2048.

In addition to the 350 test images, a collection of *novel images* were used in the experiment to keep the observers interested in the experiment. In addition, these images prevent the observers from immediately anticipating that an image corresponds to a previously viewed image. Seven ($N_n = 7$) novel images were downloaded from the online public image database Flickr [78]. The novel images are shown in Figure 5.5. Reference/distortion sequences containing $N_d = 7$ dis-



(a) *bank*, L, $C = 19$

(b) *desk3*, L, $C = 21$

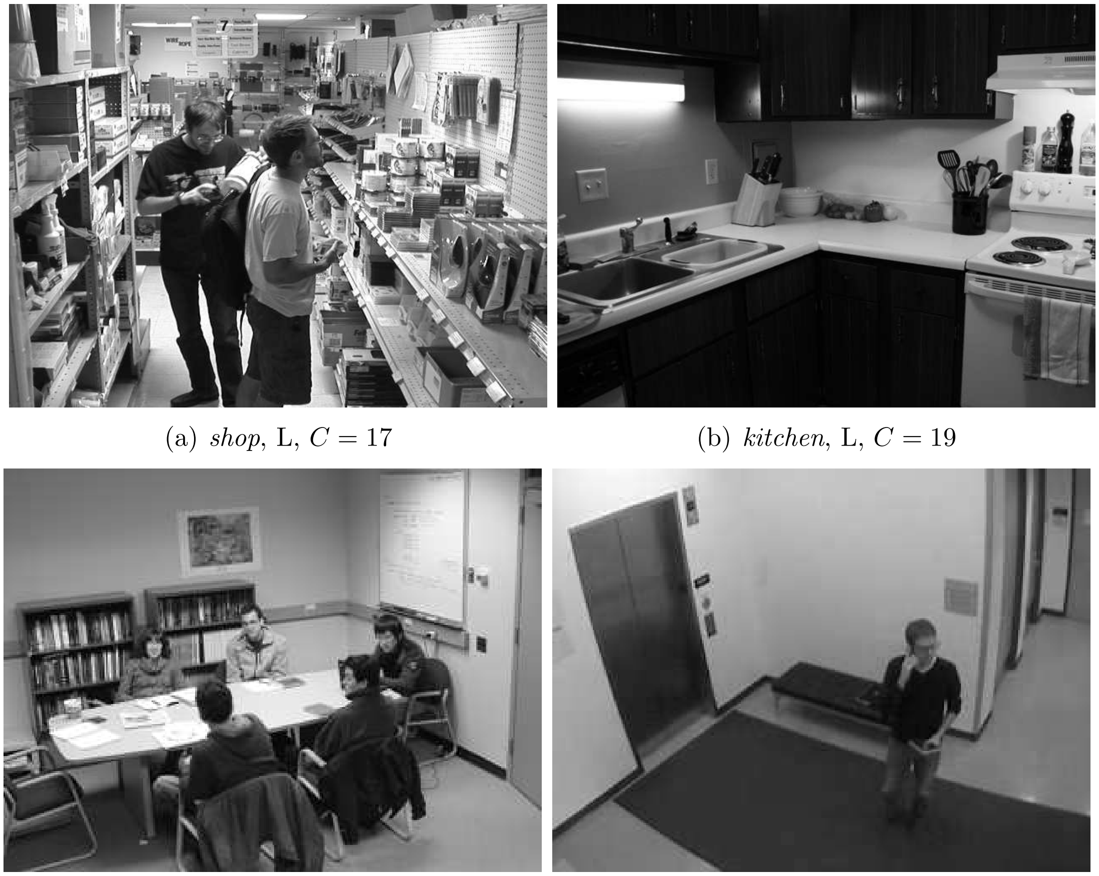
(c) *fire*, S, $C = 11$

(d) *gas station*, S, $C = 8$

(e) *hallway*, L, $C = 16$

(f) *police*, S, $C = 14$

Figure 5.3: Six of the ten grayscale natural images serving as reference images for the second experiment. The images simulate hypothetical scenes captured by a video surveillance system. In the caption for each image, "L" indicates that the image size is 640 × 480 pixels and "S" indicates that the image size is 320 × 240 pixels. The number of concepts $C$ associated with each image is provided. The remaining four images are provided in Figure 5.4.



(a) *shop*, L, $C = 17$

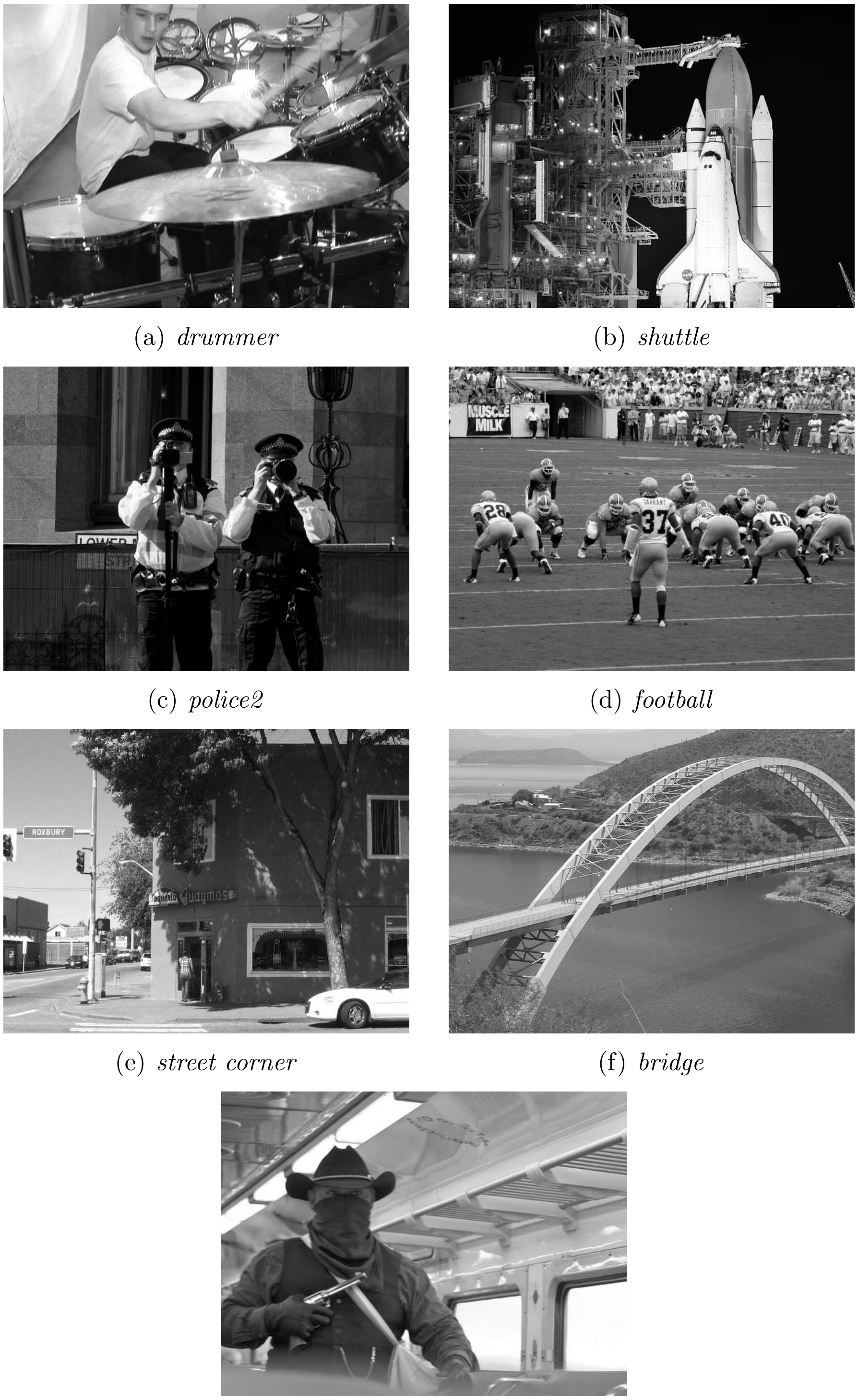
(b) *kitchen*, L, $C = 19$

(c) *conference*, S, $C = 15$

(d) *elevator*, S, $C = 13$

Figure 5.4: Four of the ten grayscale natural images serving as reference images for the second experiment. The images simulate hypothetical scenes captured by a video surveillance system. In the caption for each image, "L" indicates that the image size is $640 \times 480$ pixels and "S" indicates that the image size is $320 \times 240$ pixels. The number of concepts $C$ associated with each image is provided. The remaining six images are provided in Figure 5.3.

torted images were formed for each novel image for the five different distortion types.



(a) *drummer*

(b) *shuttle*

(c) *police2*

(d) *football*

(e) *street corner*

(f) *bridge*

(g) *cowboy*

Figure 5.5: The six grayscale natural images serving as novel images for the second experiment. These images act as distractors in the experiment to avoid guessing by the observers. All of the novel images were 640 × 480 pixels in size.

### Procedure

As in the first experiment,[4] observers produce typed descriptions of images. For each reference image, an observer views both the reference image and $N_d = 7$ distorted images generated from that reference image for one distortion type (e.g., an observer only views *elevator* contaminated with JPEG distortions).[5] Distortion artifacts are randomly paired with reference images. Due to the large number of distorted image, the experiment was split into two sessions to alleviate observer fatigue. In each session, an observer views and describes $(N_d + 1) \times N_r/2 + N_n = 187$ images ($N_d = 7$ distorted images generated plus the reference image plus the $N_n$ novel images) in a testing session.

A testing session is composed of consecutive trials. In each trial, an observer views a fixation mark for 1 second, then the image to be described is displayed for $t = 10$ seconds, and last an image of filtered Gaussian noise is displayed for 1 second. Observers provide a typed description of the image viewed for $t = 10$ seconds. The finite viewing interval forces an observer to prioritize his examination of the content and report the most important content in his description. The next trial begins after submitting a description.

The correct information about a heavily distorted image reported by an observer is sought. The trials are ordered such that an observer views the most distorted versions of the reference images first. More precisely, view a test session as being composed of $N_d$ sets of $N_r/2 + 1$ trials plus one final set of $N_r/2$ trials, where $N_d$ is the number of distorted images generated for each of the $N_r/2$

[4] The Psychophysics Toolbox is used to conduct experiments [6].

[5] Observers cannot repeat the experiment for the same reference images for different distortion types, since observers may recognize image content in heavily distorted images from previous viewings.

reference images plus one novel image per set. In the $d^{th}$ set of $N_r/2$ trials, an observer views the $d^{th}$ image from the sequence of distorted images formed from each reference image plus a random novel image with a random level of distortion. The observers only view and describe the references in the final set of $N_r/2$ trials.

Twenty-four observers participated in the experiment. The observers (11 female and 13 males) were undergraduate and graduate students between the ages of 18 and 34. Observers were either paid $10 for their participation or received credit for a cognitive psychology course.

## 5.3 Perceived Utility Scores from Textual Descriptions

Perceived utility scores are derived from the textual descriptions produce by observers that viewed the distorted images in both experiments described in Section 5.2. This section describes the process used to obtain perceived utility scores from the raw observer responses.

### 5.3.1 Concept Vectors

Perceived utility loss[6] scores are obtained by comparing *concept vectors* based on the observer descriptions of the images. *Concept vectors* are generated for both distorted and reference images. Each element of a *concept vector* is a number between 0 and 1 indicating the proportion of observers whose description included words or phrases related to a specific concept associated with the reference image

[6] Perceived utility *loss* with respect to the reference is similar to difference mean opinion scores (DMOS) reported in the perceived quality task.

content. The concepts are *observer-centric* and emerge from an analysis of the collected observer descriptions. Images formed from the same reference image use the same concepts; the number of concepts $C$ for each reference image are provided in Figures 5.1, 5.2, 5.3, and 5.4. The concepts that emerged from the observer responses from the novel images in the second experiment were not used.

To illustrate the formation of the *observer-centric* concepts, consider the *street2* image in Figure 5.2(b). Among other features described, several observers described the streetlamp appearing on the right-hand-side of the image. Observers used the following words and phrases to describe the streetlamp: “lamp post,” “lamp,” “lampost,” “lamppost,” “lightpost,” “light post,” “pole,” “post,” “streetlamp,” “streetlight,” and “streetlights.” Thus, the “streetlamp” *emerges* from the descriptions, and all these words or phrases coincide with the “streetlamp” concept, so the observer descriptions can be automatically scanned for the presence of this concept.

The concept vectors were adjusted in two steps. First, the concept vectors corresponding to the reference images were examined for the presence of infrequent concepts. Infrequent concepts, defined as those concepts that appeared in fewer than 15% of the observer descriptions, where removed from the *concept vectors.*

Second, elements of the concept vectors corresponding to the distorted images occasionally contained values indicating that more observers described a concept in a distorted image than the reference image. Given a particular distortion type and reference image, the values of the concept vectors are expected to monotonically increase as the distortion level decreases with the reference image yielding the largest value. Due to the test design, all 17 observers viewed the reference images and only a subset of these observers viewed images subjected to a particular distortion type. Consequently, it is possible that all the observers viewing the *kitchen2* image contaminated with AWGN described the knife block on the counter, but not all the observers viewing this image degraded by JPEG compression mentioned the knife block, even when viewing the reference image. This would lead to a smaller proportion of observers describing the knife block concept in the reference image relative to the proportion describing this concept in the *kitchen2* subjected to the least amount of AWGN, disrupting the desired monotonicity. To restore monotonicity, distorted image concept vectors were multiplied (element-wise) by the corresponding reference image concept vectors.[7]

### 5.3.2 Perceived Utility Loss Definition

The perceived utility loss score is defined such that experimental errors could be estimated.[8] The infrequent concepts described in Section have been removed, leaving $C$ concepts for a given natural image. Let $\boldsymbol{p}$ and $\boldsymbol{q}$ denote the concept vectors corresponding to the reference image and distorted image, respectively. The $c^{th}$ element of the concept vector $\boldsymbol{p}$ is denoted $p_c$ and defines the probability that concept $c$ is recognized.[9] Assuming that $1 \geq p_c \geq q_c \geq 0$ for $c = 1, 2, \ldots, C$, the perceived utility loss (UL) is based on two concept vectors $\boldsymbol{p}$ and $\boldsymbol{q}$ and is given by

$$\mathrm{UL}(\boldsymbol{p}, \boldsymbol{q}) = \frac{100}{A} \sum_{c=1}^{C} p_c \left(1 - q_c\right), \tag{5.1}$$

[7] The reference image concept vector is a vector of ones when all observers describe all concepts in the reference images, and in that case, no adjustment to ensure monotonicity is needed.

[8] In [106], the perceived utility loss score did not accommodate a straightforward estimate of the experimental errors.

[9] The experiment is conducted to collect data from observers to estimate the values of the concept vectors $\boldsymbol{p}$ and $\boldsymbol{q}$

where $A = \sum_{c=1}^{C} p_c$ is treated normalizing constant[10] introduced to fix the maximum value of UL to 100. Note that UL $\in [0, 100]$, since $1 \geq p_c \geq q_c \geq 0$ for $c = 1, 2, \ldots, C$, and UL increases as the perceived utility decreases. UL $= 0 \iff \boldsymbol{p} = \boldsymbol{q}$ and indicates that for all concepts the same proportion of observers describe the concepts in the reference image as in the distorted image. UL $= 100$ when all observers described the concepts in the reference image but no observers described any of the concepts in the distorted image.

Eq. (5.1) has a straightforward interpretation. First, suppose all concepts were recognized by all observers, then $p_c = 1$ for $c = 1, \ldots, C$. Then, according to Eq. (5.1), any concept recognized with probability less than 1 in a distorted image constitutes a utility loss. Now, suppose $p_c$ is less than 1, then Eq. (5.1) introduces a weight on the importance of each concept: smaller values of $p_c$ correspond to less important concepts according to the sample of people participating in the experiment.

The standard error associated with the second utility loss definition in Eq. (5.1) is based on approximating the proportions $p_c$ and $q_c$ by Gaussian random variables and assuming that a proportion $p_c$ is independent of proportion $p_d$ for $c \neq d$. The assumption recognition of concept $c$ is independent of recognition of concept $d$ for $c \neq d$. First, observe that Eq. (5.1) can be rewritten as

$$\mathrm{UL}(\boldsymbol{p}, \boldsymbol{q}) = \frac{100}{A} \sum_{c=1}^{C} p_c - \frac{100}{A} \sum_{c=1}^{C} p_c q_c \tag{5.2}$$

Let $N_p$ and $N_q$ denote the number of observers that viewed the reference and distorted images, respectively, and let $n_{p,c}$ and $n_{q,c}$ be binomial random variables counting the number of times observers recognized concept $c$ in the reference and distorted images, respectively. Then, $\hat{p}_c = \frac{n_{p,c}}{N_p}$ and $\hat{q}_c = \frac{n_{q,c}}{N_q}$ approximate the true

[10] Although the values $p_c$ are estimated in from the experimental data, the term $A$ is assumed to be deterministic to simplify the approximation of the standard deviation.

values of $p_c$ and $q_c$. The binomial random variables $n_{p,c}$ and $n_{q,c}$ are approximated as Gaussian random variables. For example, the binomial random variable $n_{p,c}$ can be approximated by a Gaussian random variable with mean $N_p p_c$ and variance $N_p p_c(1 - p_c)$. This can be transformed to approximate $p_c$ as a Gaussian random variable with mean $n_{p,c}N_p^{-1}$ and variance $n_{p,c}(N_p - n_{p,c})N_p^{-2}$. The term $p_c q_c$ is approximated by a Gaussian random variable with mean $n_{p,c}n_{q,c}(N_p N_q)^{-1}$ and variance $n_{p,c}n_{q,c}(N_p N_q - n_{p,c}n_{q,c})N_p^{-1}N_q^{-2}$. The variance of $p_c q_c$ only accounts for variation due to $q_c$ (i.e., division by $N_p N_q^2$ and not $N_p^2 N_q^2$), since $p_c$ is a weight introduced to restore monotonicity of the values of $q_c$. Since it is assumed that recognition of concepts are independent of one another, the utility loss is modeled as a Gaussian random variable with mean

$$E\{\mathrm{UL}(\boldsymbol{p}, \boldsymbol{q})\} \approx \frac{100}{A} \sum_{c=1}^{C} \hat{p}_c - \frac{100}{A} \sum_{c=1}^{C} \hat{p}_c \hat{q}_c \tag{5.3}$$

and variance

$$Var\{\mathrm{UL}(\boldsymbol{p}, \boldsymbol{q})\} \approx \frac{10000}{A^2} \sum_{c=1}^{C} \left( \frac{\hat{p}_c(1 - \hat{p}_c)}{N_p} + \frac{\hat{p}_c \hat{q}_c (1 - \hat{p}_c \hat{q}_c)}{N_q} \right), \tag{5.4}$$

where $\hat{p}_c = \frac{n_{p,c}}{N_p}$ and $\hat{q}_c = \frac{n_{q,c}}{N_q}$. Thus, the estimate of UL is unbiased, and the variance of the estimate decreases as the number of observers viewing the reference and distorted images increases.

### 5.3.3 Plausibility of Utility Loss Definition

The plausibility of the methods leading to Eq. (5.1) to produce meaningful PU scores is evaluated by analyzing the consistency of the scores corresponding to image *recognition thresholds*. The *recognition threshold* splits a sequence of distorted images into useful distorted images that provide relevant information about the reference scene content to an observer and useless images. The PU loss score for the recognition threshold should not vary across distortion types.

Five people reviewed all the descriptions obtained in the first experiment and selected the first description relevant to the reference scene content. PU loss scores computed using Eq. (5.1) corresponding to the selected images were averaged to determine the average recognition threshold in terms of the PU loss scores.

The methods leading to Eq. (5.1) were determined to provide meaningful PU scores. A one-way analysis of variance (ANOVA)[11] revealed that image size accounts for variations among recognition thresholds ($F_{1,848} = 46.8$, $p \approx 0$), and for only four of the ten reference images the distortion type accounts for differences among recognition thresholds ($F_{2,82} > 3.8$, $p < 0.03$). An inspection of the sequences of distorted images for the four images (*kitchen2*, *stairs1*, *street2*, and *street3*) whose recognition thresholds differed across distortion types revealed that these sequences sparsely sampled the distortion levels near the recognition threshold, which merely casts doubt on the accuracy of the estimated recognition thresholds for these four images. The recognition thresholds for the remaining six reference images varied according to the image size ($F_{1,508} = 18.6$, $p \approx 0$) and not the distortion type ($F_{2,507} = 0.02$, $p = 0.98$). Objective estimators that accurately estimate PU loss scores inherit the ability to estimate recognition thresholds, and only an appropriate threshold based on the image size needs to be determined.

[11] ANOVA uses an F-test that compares 1) the variance of the $m$ mean recognition threshold (RT) PU scores for each subset (e.g., images with the same distortion type) with the global mean RT PU score to 2) the total the variance of all $m \times n$ RT PU scores. The test statistic $F_{m-1,m(n-1))}$ and its p-value are reported.

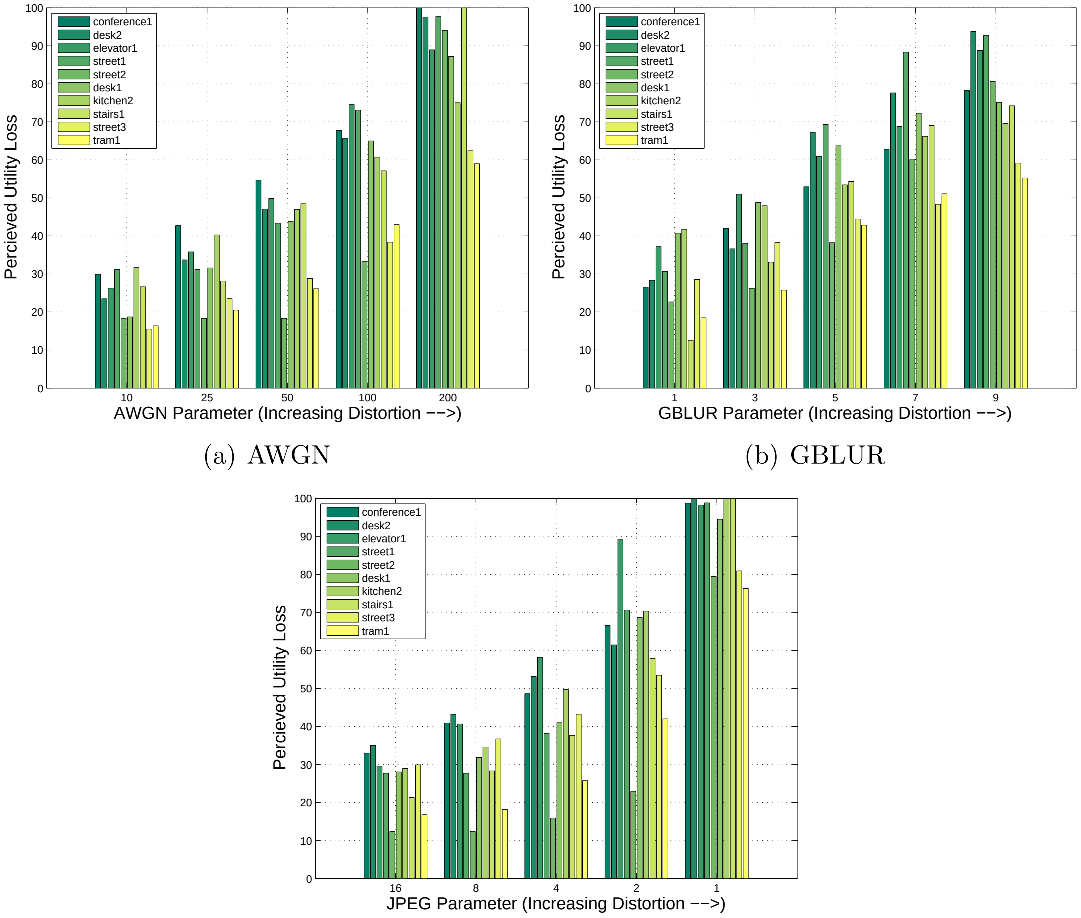
## 5.4 Analysis of Perceived Utility Scores

In this section, the perceived utility (PU) scores acquired in both are analyzed to 1) demonstrate the need for a utility estimator and 2) illustrate that the proposed method provided reproducible PU loss scores.

### 5.4.1 Demonstrating the Need for a Utility Estimator

Applications for which the images used are vulnerable to a variety of distortions need a utility estimator to predict the perceived utility (PU) loss of distorted images. However, applications often use images vulnerable to a specific class of distortions (e.g., additive white Gaussian noise). If the perceived utility loss of a given image can be reliably predicted based on that distortion parameter (e.g., additive white Gaussian noise with power $\sigma_\eta^2$), then a utility estimator would be unnecessary in that application.

The PU loss scores obtained in both experiments demonstrate that PU loss scores vary among distorted images contaminated with the same type and level of distortion. Figures 5.6 and 5.7 show, for each distortion type, the relationship between the PU loss scores as a function of the distortion parameters for all the images tested. From the figure it is evident that for a fixed value of a distortion parameter the PU loss varies according to the reference image, and the distortion parameter to poorly estimates the PU loss of a given distorted image. Thus, a utility estimator that accurately estimates the PU loss is necessary.



(a) AWGN

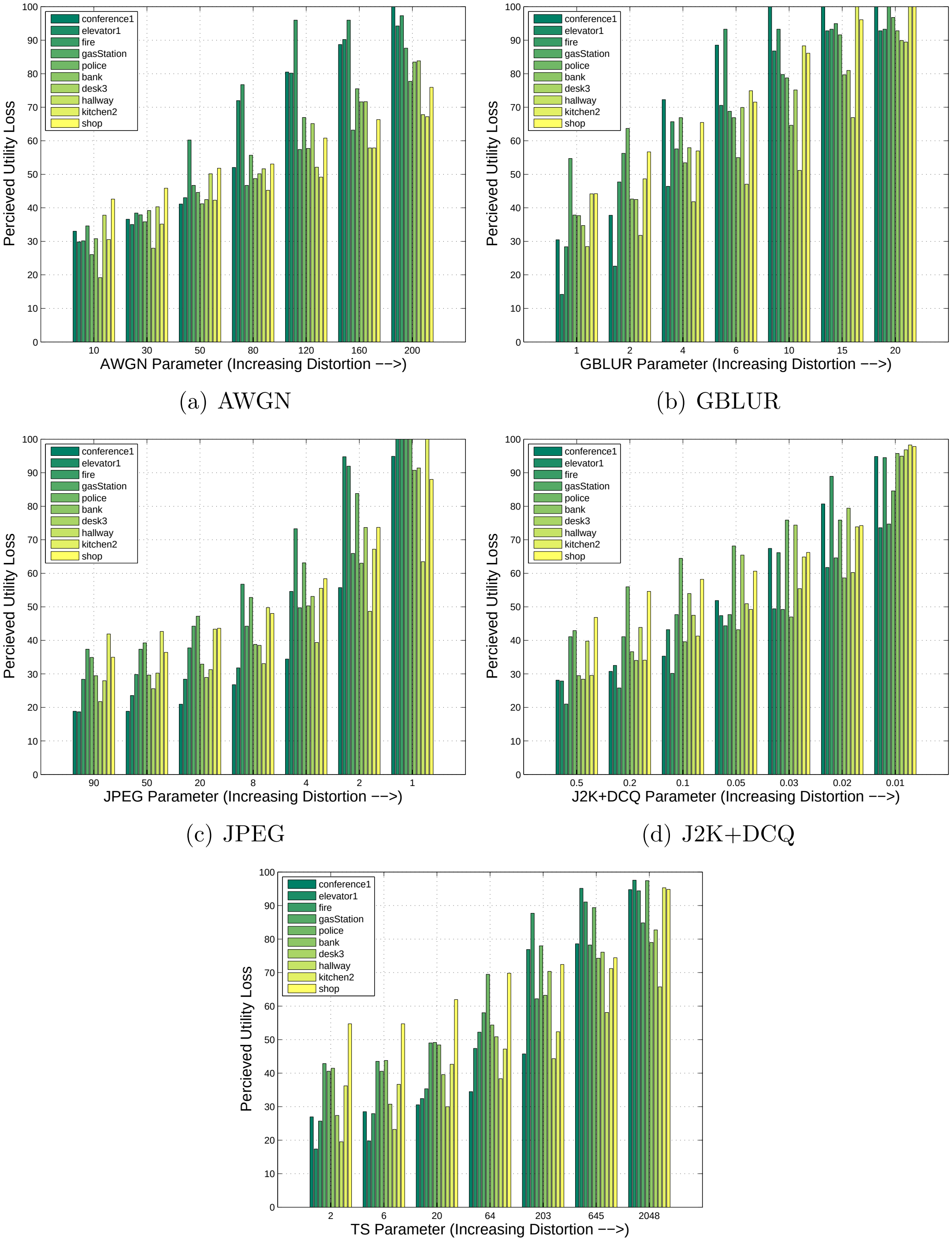
(b) GBLUR

(c) JPEG

Figure 5.6: Given a fixed distortion parameter value (e.g., additive white Gaussian noise with noise power $\sigma_\eta^2$) does not accurately predict the perceived utility (PU) loss across a variety of images. Each bar graph shows the PU loss scores from the *first* experiment as a function of the distortion parameters.

### 5.4.2 Evidence that the Proposed Technique Yields Reproducible Results

An experimental method is only useful if its results are reproducible. The images *conference*, *elevator*, and *kitchen* were used in both experiments to examine the extent that the proposed technique to obtain perceived utility (PU) scores is



(a) AWGN

(b) GBLUR

(c) JPEG

(d) J2K+DCQ

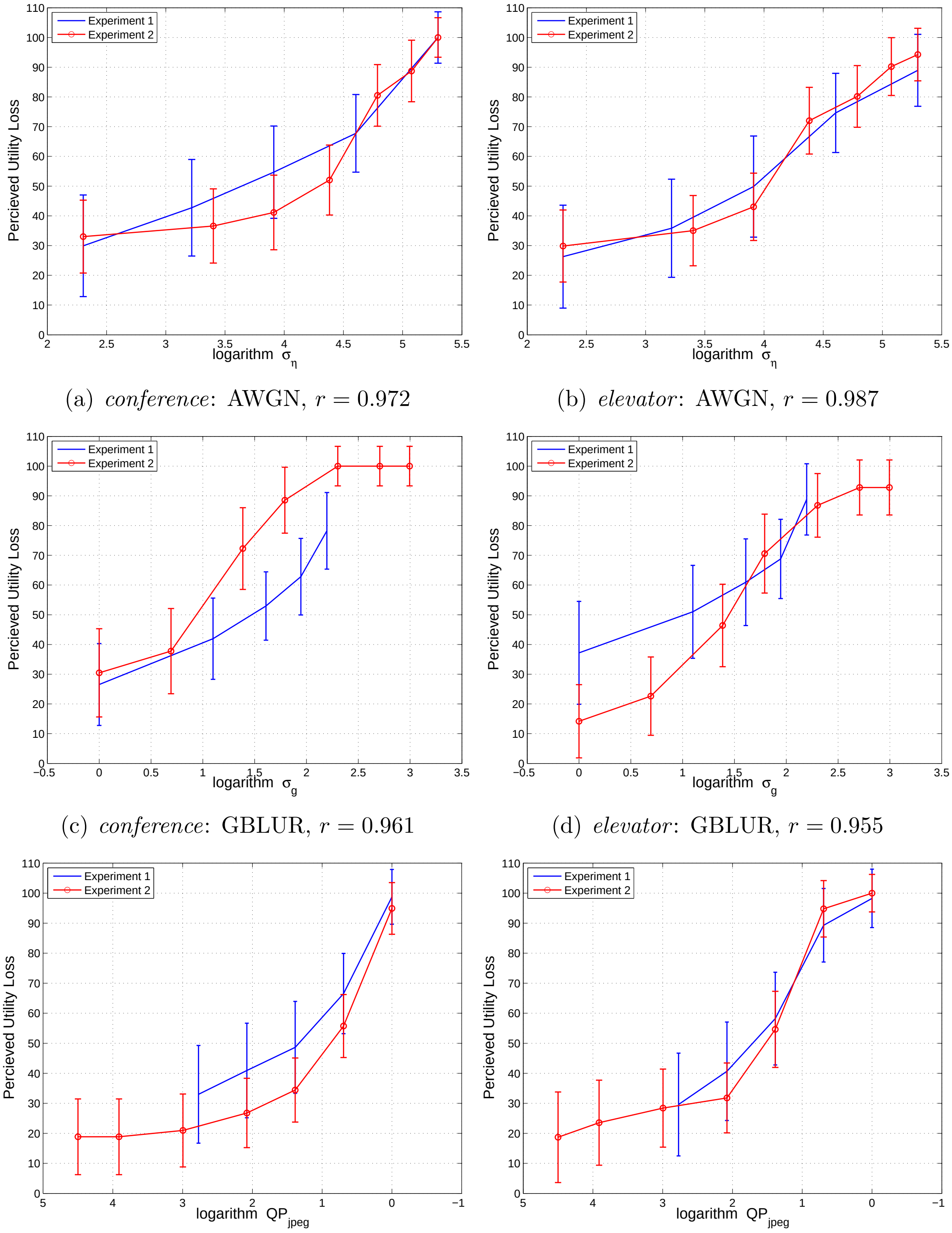
(e) TS

Figure 5.7: Given a fixed distortion parameter value (e.g., additive white Gaussian noise with noise power $\sigma_\eta^2$) does not accurately predict the perceived utility (PU) loss across a variety of images. Each bar graph shows the PU loss scores from the *first* experiment as a function of the distortion parameters.

reproducible.

For both experiments, distorted images were formed by subjecting the reference images to additive white Gaussian noise (AWGN), Gaussian blur (GBLUR), and baseline JPEG compression. However, different distortion parameters were used in each experiment. Specifically, the reference/distortion sequences contained $N_d = 5$ images in the first experiment, whereas $N_d = 7$ images appeared in the reference/distortion sequences formed for the second experiment. To compare the degree of correlation between the PU loss scores from each experiment, the PU loss scores from the second experiment were approximated for the distortion parameters used in the first experiment via linear interpolation.

The PU loss scores obtained from one experiment capture at least 90% of the variation in the PU loss scores observed from the other experiment. The relationship between the PU loss scores and the distortion parameters (e.g., $\sigma_\eta$ for AWGN) obtained in each experiment are shown in Figures 5.8 and 5.9. The error bars in the figures indicate the 95% confidence intervals corresponding to the PU loss scores. In most cases, the error bars corresponding to the PU loss scores from each experiment overlap, which indicates that the PU loss scores are statistically equivalent. However, in some cases the PU loss scores are statistically different. Despite the statistical differences in some of the PU loss scores, the PU loss scores the two experiments are strongly correlated. The Pearson linear correlation $r$ is calculated between the PU loss scores from the first experiment and the PU loss scores from the second experiment after interpolating the scores corresponding to the distortion parameters used in the first experiment. The values of $r$ are provided in Figures 5.8 and 5.9, and all correlation values exceed 0.95, which indicate that at least 0.90% of the variation in the PU loss scores from one experiment are captured



(a) *conference*: AWGN, $r = 0.972$

(b) *elevator*: AWGN, $r = 0.987$

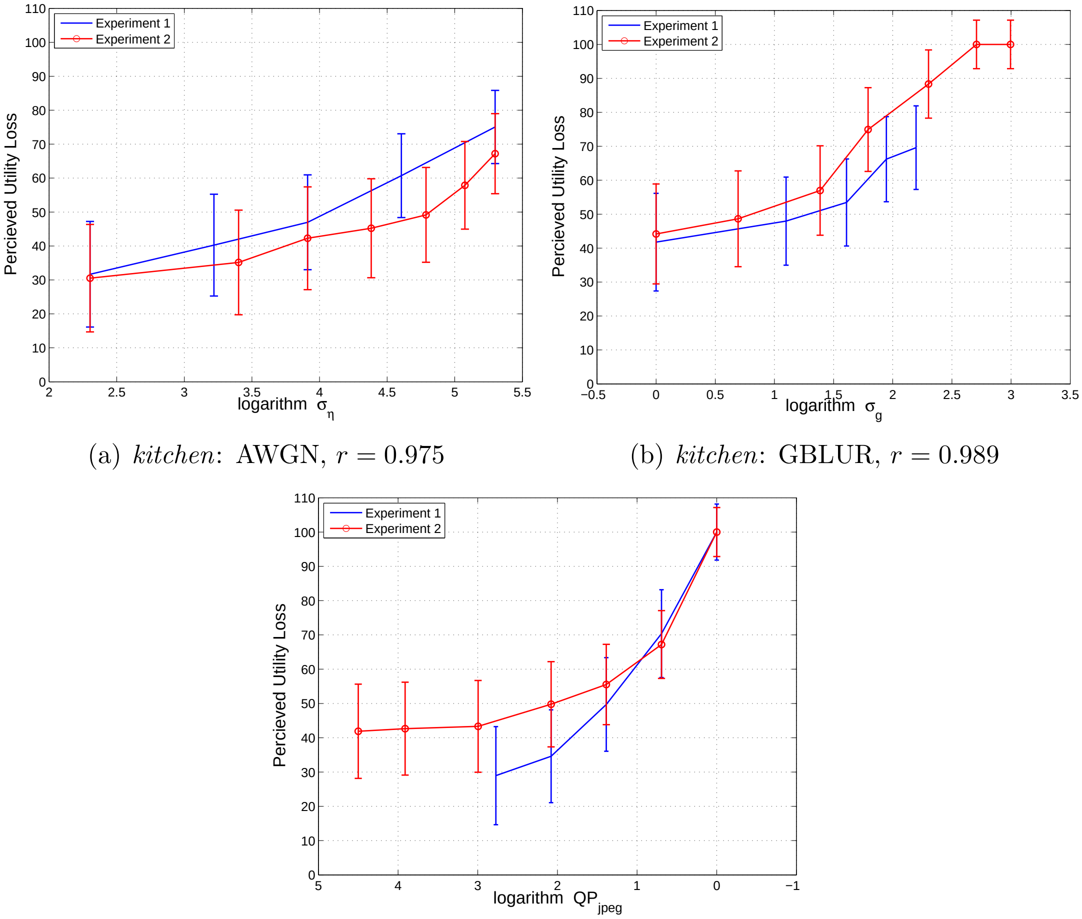
(c) *conference*: GBLUR, $r = 0.961$

(d) *elevator*: GBLUR, $r = 0.955$

(e) *conference*: JPEG, $r = 0.997$

(f) *elevator*: JPEG, $r = 0.993$

Figure 5.8: Perceived utility (PU) loss scores obtained from each experiment versus the distortion parameter for *conference* and *elevator*. The error bars indicate the 95% confidence intervals. The Pearson linear correlation $r$ is reported when the PU scores from the first experiment with those of the second experiment, interpolated to the distortion parameter values used in the first experiment.



(a) *kitchen*: AWGN, $r = 0.975$

(b) *kitchen*: GBLUR, $r = 0.989$

(c) *kitchen*: JPEG, $r = 0.979$

Figure 5.9: Perceived utility (PU) loss scores obtained from each experiment versus the distortion parameter for *kitchen*. The error bars indicate the 95% confidence intervals. The Pearson linear correlation $r$ is reported when the PU scores from the first experiment with those of the second experiment, interpolated to the distortion parameter values used in the first experiment.

in the other experiment.

## 5.5 Objective Estimates of Perceived Utility Scores

This section examines the capability of both the single-scale and multi-scale NICE utility estimators as well as popular quality estimators to estimate the perceived utility (PU) loss scores obtained from each experiment described in Section 5.2. Specifically, the following quality estimators are examined: the following the peak signal-to-noise ratio (PSNR); the weighted signal-to-noise ratio (WSNR) [27]; the noise quality measure (NQM) [27]; the visual signal-to-noise ratio (VSNR) [20]; the structural similarity index (SSIM) [137] and its multi-scale extension (MS-SSIM) [138]; the visual information fidelity (VIF) criterion [113]; and a modified implementation of VIF, denoted VIF*, that adjusts the weights used to pool the objectives scores produced by VIF across image scales[12] [100]. All of these objective quality estimators are available in the MeTriX MuX toolbox [44].

The correlation between the objective scores and the PU loss scores is measured using the Pearson (linear) correlation $r$, the Spearman rank correlation $\rho$, and the Kendall rank correlation $\tau$. The scores produced by many of the estimators exhibit a nonlinear relationship with the PU scores, and a nonlinearity is often fitted to the data to resolve this nonlinear relationship. However, the coefficients attributed to the nonlinearity augment the objective estimator and obscure the true contribution made by the objective estimator. Therefore, an linear mapping $f(x) = ax + b$ is fitted to the data to minimize the sum-squared error between the mapped objective scores $f(x)$ and the PU loss scores. This mapping simply shifts and rescales the objective scores to the range of PU loss scores. The accuracy of the linearly mapped objective scores are evaluated with respect to the PU loss scores using the root mean squared error (RMSE) and the outlier ratio (OR). In

[12] VIF* multiplies the individual subband calculations corresponding to $I(\vec{C}^N; \vec{E}^N | s^N)$ and $I(\vec{C}^N; \vec{F}^N | s^N)$ in Eqs. (12) and (13) of [113] by $\frac{1}{N}$.

Table 5.1: Statistics summarizing the performance of estimators serving as utility estimators using the perceived utility (PU) loss scores from the first experiment. The Pearson linear correlation $r$, the Spearman rank correlation $\rho$, the Kendall rank correlation $\tau$, the root mean-squared error (RMSE), outlier ratio (OR), the resolving power $RP_{0.05}$ are reported when the estimates are compared with the PU loss scores. The skewness (skew) and kurtosis (kurt) of the residuals are reported. RMSE values with an asterisk indicate that the accuracy of that estimator is statistically equivalent to that of MS-NICE$_4$ based on the Brown-Forsythe-Levene (BFL) at the 5% significance level. Boldface values are the optimal for that column, and italicized values are statistically equivalent to the optimal value. The correlation measures are compared using their absolute values.

| | *Correlation Measures* | | | *Accuracy Measures* | | | |
|---|---|---|---|---|---|---|---|
| *Estimator* | $r$ | $\rho$ | $\tau$ | RMSE | OR | skew/kurt | $RP_{0.05}$ |
| PSNR | -0.571 | -0.588 | -0.409 | 19.9 | 0.507 | 0.26/2.5 | 42.5 |
| WSNR | *-0.799* | *-0.825* | *-0.627* | 14.6* | *0.327* | -0.23/2.8 | 41.1 |
| NQM | *-0.814* | *-0.835* | *-0.637* | 14.1* | *0.340* | -0.25/2.8 | 41.1 |
| VSNR | -0.745 | -0.771 | -0.559 | 16.2 | 0.427 | 0.21/2.5 | 41.8 |
| SSIM | -0.367 | -0.378 | -0.266 | 22.6 | 0.540 | 0.42/2.4 | 28.2 |
| MS-SSIM | -0.744 | -0.762 | -0.554 | 16.2 | 0.433 | 0.17/3.0 | 49.8 |
| VIF | -0.750 | *-0.820* | *-0.607* | 16.1 | *0.353* | 0.03/2.6 | 45.9 |
| VIF* | *-0.805* | *-0.838* | *-0.633* | 14.4* | *0.320* | -0.22/2.6 | 40.6 |
| NICE$_{\text{Sobel}}$ | 0.762 | *0.801* | *0.592* | 15.7 | *0.353* | -0.03/2.6 | 39.3 |
| NICE$_{\text{Canny}}$ | 0.371 | 0.600 | 0.436 | 22.6 | 0.540 | 0.57/2.5 | 42.6 |
| MS-NICE$_4$ | **0.864** | **0.866** | **0.670** | **12.2** | **0.273** | -0.23/2.4 | 36.7 |

addition, the skewness and kurtosis of the residuals are reported along with the resolving power ($RP_\alpha$), which specifies the smallest difference in fitted objective scores for a pair of test images such that the difference is significant based on the estimated error of the subjective scores at the $100(1-\alpha)\%$ confidence level [7]. Comparisons between the accuracy (i.e., the residual variance) of MS-NICE$_4$ and the other objective estimators are made using the Brown-Forsythe-Levene (BFL) test [13] [10].

Overall, MS-NICE$_4$ provides the most reliable estimates of PU loss among the objective estimators evaluated. The results based on the PU loss scores from the first experiment show that both MS-NICE$_4$ and VIF* provide statistically equivalent errors in terms of accuracy based on a comparison of residual variances using the BFL test. However, the errors are statistically larger for VIF* than MS-NICE$_4$ when the PU loss scores from the second experiment are used. The second experiment contains more distortions and more images than the first experiment, so the data for the second experiment provides better evidence with regard to the performance of the objective estimators than the first experiment. Tables 5.1 and 5.2 contain the statistics summarizing the performance of the objective estimators using the PU loss scores from the first and second experiments, respectively. An asterisk appears next to the RMSE values for objective estimators with statistically equivalent residual variance to MS-NICE at the 5% significance level. Optimal values corresponding to the correlation measures, RMSE, and outlier ratio appear in boldface, and italicized values are statistically equivalent. Correlation measures are compared using their absolute values.

[13] The BFL test does not require that residuals follow a Gaussian distribution, which is assumed for the $F$-test.

Table 5.2: Statistics summarizing the performance of estimators serving as utility estimators using the perceived utility (PU) loss scores from the second experiment. Refer to Table 5.1 for an explanation of the correlation and accuracy measures.

| | *Correlation Measures* | | | *Accuracy Measures* | | | |
|---|---|---|---|---|---|---|---|
| *Estimator* | $r$ | $\rho$ | $\tau$ | RMSE | OR | skew/kurt | $RP_{0.05}$ |
| PSNR | -0.647 | -0.717 | -0.526 | 17.6 | 0.509 | 0.14/2.2 | 35.3 |
| WSNR | -0.765 | *-0.854* | *-0.651* | 14.8 | 0.440 | 0.14/2.3 | 29.2 |
| NQM | -0.797 | *-0.868* | *-0.670* | 13.9 | *0.400* | -0.05/2.5 | 32.1 |
| VSNR | -0.719 | -0.786 | -0.572 | 16.0 | 0.477 | 0.07/2.2 | 37.0 |
| SSIM | -0.479 | -0.603 | -0.430 | 20.2 | 0.571 | 0.34/2.4 | 32.9 |
| MS-SSIM | -0.705 | -0.775 | -0.574 | 16.3 | 0.449 | 0.07/2.6 | 45.2 |
| VIF | -0.731 | **-0.880** | **-0.686** | 15.7 | 0.489 | 0.23/2.1 | 43.6 |
| VIF* | *-0.816* | **-0.880** | *-0.684* | 13.3 | *0.397* | 0.06/2.2 | 32.4 |
| $\text{NICE}_{\text{Sobel}}$ | 0.757 | 0.788 | 0.584 | 15.0 | *0.406* | -0.20/2.8 | 37.9 |
| $\text{NICE}_{\text{Canny}}$ | 0.369 | 0.659 | 0.481 | 21.4 | 0.597 | 0.36/2.2 | 41.5 |
| $\text{MS-NICE}_4$ | **0.848** | *0.858* | *0.658* | **12.2** | **0.337** | -0.08/2.5 | 29.4 |

## 5.6 Discussion

The overall conclusions establish that $\text{MS-NICE}_4$ provides the most accurate estimates of perceived utility loss across a variety of distortions. However, $\text{NICE}_{\text{Canny}}$, which produced very accurate estimates of PU scores in Chapter 4, does not perform well on the data from the two experiments described in this chapter. A closer examination of the results for $\text{NICE}_{\text{Canny}}$ revealed that it provided very unreliable estimates of PU loss for distorted images with additive white Gaussian noise (AWGN). AWGN is typically problematic for classical edge detectors, since these distortions introduce contours to which the Canny edge detector is sensitive. Among the single-scale implementations of NICE evaluated, $\text{NICE}_{\text{Sobel}}$ provides better performance than $\text{NICE}_{\text{Canny}}$ in terms of estimating PU loss scores.

The conclusions drawn about VIF and VIF* appear to contradict those conclusions reported in Chapter 4. In Chapter 4, VIF was reported to provide more accurate estimates of perceived utility (PU) scores than VIF*, yet according to the current results, VIF* provides more accurate estimates than VIF for both datasets (i.e., Tables 5.1 and 5.2) . VIF exhibits greater sensitivity to high-frequency distortions than low-frequency distortions [100]. Consequently, VIF poorly resolves differences in the PU of heavily distorted images, where observers necessarily extract information about the image content using cues conveyed by lower-frequency signal information due to the degradation of high-frequency signal information. VIF* weights distortions across images scales based on the reference image's energy distribution, and thus, VIF* exhibits greater sensitivity to low-frequency distortions than high-frequency distortions due to the $1/f^{\alpha}$ power spectra of natural images [39]. For this reason, VIF* resolves differences in PU of heavily distorted images better than VIF.

## 5.7 Summary

A novel technique to acquire perceived utility (PU) scores is presented that collects textual descriptions produced by observers viewing distorted natural images. This technique uses *observer-centric* concepts for the images that emerge from an analysis of the observer descriptions. *Concept vectors* that quantify the presence of concepts appearing in observer descriptions are used to generate PU scores.

Since observers guide the creation of concepts describing useful image content, the *observer-centric* approach establishes a framework to quantify image usefulness for a broad range of tasks.

Two experiments are conducted using this novel technique with distorted images that simulate scenes captured by a surveillance system. The first experiment collected PU loss scores for 150 distorted images spanning three types of distortions, and the second experiment collected PU loss scores for 350 distorted images spanning five types of distortions. The distorted images in each experiment were formed from ten reference images, and three reference images were used in both experiments (i.e., a total of 17 reference images were used across both experiments). The three reference images that were common to both experiments were used to compare the results from the each experiment and demonstrate the proposed technique provides reproducible results.

The capability of both the NICE utility estimator and popular quality estimators to estimate these PU is reported. The conclusions drawn from the results reported in Tables 5.1 and 5.2 establish that MS-NICE is the most robust utility estimator among the estimators evaluated, since MS-NICE consistently performs as well as estimators yielding the most accurate PU estimates for a variety of distortion types.

## 5.8 Key Points

- A novel technique to collect perceived utility scores is presented that overcomes the limitations of the perceived utility scores collected using the techniques described in Chapter 3.

- The novel technique uses *observer-centric* concepts that emerge from an analysis of the observer descriptions. Because observers guide the creation of concepts describing useful image content, the *observer-centric* approach establishes a framework to quantify image usefulness for a broad range of tasks.
- Two experiments were conducted using this novel technique to obtain perceived utility scores for distorted images that simulate scenes captured by a video surveillance system. Perceived utility scores were collected for a total of 500 distorted images, spanning 5 types of distortion artifacts. Three reference images were common to both experiments, and the perceived utility scores for the distorted images formed from those reference images were compared to demonstrate the reproducibility of the technique.
- The capability of both the NICE utility estimator and popular quality estimators to estimate these PU is reported. The multi-scale implementation of NICE (MS-NICE) performs the best among all the estimators evaluated, since MS-NICE consistently performs as well as estimators yielding the most accurate PU estimates for a variety of distortion types.

CHAPTER 6

# ANALYSIS OF THE NATURAL IMAGE CONTOUR EVALUATION

## 6.1 Introduction

The natural image contour evaluation (NICE) was introduced and analyzed in Chapter 4 as a utility estimator. Chapter 5 demonstrated the use of NICE on a new database of images with perceived utility scores obtained using a novel test method. The combined results from those two chapters support the designation of NICE as a utility estimator for a variety of distortions.

This chapter describes a two part analysis of NICE. In the first part, a gradient analysis is conducted to illustrate those image features that minimize NICE. The gradient analysis demonstrates that restoring the visibility of edges in images minimizes NICE, which is consistent with the intended behavior of NICE described in previous chapters. In addition, the results confirm observations in previous chapters that distortions to low-frequency components (e.g., shading) and very high-frequency components (e.g., textures) have little impact on the score produced by NICE.

In the second half of the chapter, a novel method is proposed and executed to produce *distorted but useful* images compatible with an existing image codec but based on the distortion measure used by NICE. Specifically, baseline compatible JPEG quantization tables, which specify step-sizes used to quantize the discrete cosine transform (DCT) coefficients, are found via rate-distortion optimization using NICE as a distortion measure with a JPEG coder [90].

The formula for NICE was not developed for use with the block-based DCT used by JPEG. Such an incompatibility between an objective estimator and coder is common, and often the problem of tuning a coder to a objective estimator is abandoned. This chapter describes a novel approach to tune a coder to an objective estimator. In particular, JPEG quantization tables optimized with respect to NICE were found using a genetic algorithm (GA) heuristic search [109] to minimize a cost function based on the rate-distortion optimization problem. The resulting quantization tables produce images whose encoding bitrates are 34% lower than images yielding the same score with NICE but produced using the example quantization tables provided in the original JPEG specification [90]. In addition to providing JPEG quantization tables optimized with respect to NICE, the GA-based approach to rate-distortion optimization is discussed as a technique to “tune” other existing estimators to existing image codecs.

This chapter is organized as follows: A gradient analysis of NICE is presented in Section 6.2, which includes a description of a continuous approximation of NICE. Section 6.3 reports the use of NICE to perform rate-distortion optimization with JPEG. In particular, methods for producing quantization tables optimized for NICE that are compatible with JPEG are provided. The use of a genetic algorithm to perform rate-distortion optimization for other objective estimators is discussed in Section 6.4. The chapter is concluded with a summary in Section 6.5.

## 6.2 Gradient Analysis

A gradient analysis, which is similar to that conducted with the components of SSIM in Section 2.3.4, is conducted using NICE to illustrate those image features that minimize NICE. NICE contains two nonlinear components: hard-thresholding and the Hamming distance. To find the gradient of NICE, a continuous approximation of NICE is found, and the gradient of the continuous approximation is determined.

This section first summarizes NICE, which is followed by a formulation of the continuous approximation of NICE. Next, the gradient of the continuous approximation of NICE is derived. This section concludes with sample images formed by minimizing NICE using conjugate gradient descent. Various distorted images were used to initiate the conjugate gradient descent iteration.

### 6.2.1 Summary of NICE

NICE compares the contours of the reference and test images, which are represented as binary images. Before the contours of the reference and test images are compared, binary images representing the contour maps are individually subjected to morphological dilation with a 3×3 "plus-sign" shaped structuring element $E$ [46]. Morphological dilation accommodates local registration errors between the reference and test contour maps introduced by distortions in the test image and should not be quantified as errors.

The contours of the reference and test images are compared across $S$ image scales, and $b_s[m,n]$ and $\hat{b}_s[m,n]$ respectively denote the contours of the reference and test images at scale $s$ and spatial index $(m,n)$. The NICE score for the test image is computed as

$$\mathrm{NICE}_S = \frac{\sum_{s=1}^{S} d_H(b_s[m,n] \oplus E, \hat{b}_s[m,n] \oplus E)}{\sum_{s=1}^{S} N_{b_s}}, \tag{6.1}$$

where $N_{b_s}$ is the number of non-zero elements of $b_s$, $d_H(X,Y)$ denotes the Hamming distance[1] between the two binary vectors $X$ and $Y$, and $b[m,n] \oplus E$ denotes the dilation of the binary image $b[m,n]$ using the morphological structuring element $E$. The Hamming distance quantifies 1) the number of pixels corresponding to contours in the reference image that have been *lost* in the test image due to the distortions and 2) the number of pixels corresponding to contours in the test image *introduced* by the distortions that were absent in the reference image. Since the content of natural images vary, the proportion of pixels corresponding to contours will vary. The factor $N_B$ accounts for this variability by adaptively scaling the raw score $d_H(b_s[m,n] \oplus E, \hat{b}_s[m,n] \oplus E)$ according to the extent of the contour information identified in the reference image.

### 6.2.2 Continuous Approximation of NICE

A continuous approximation to NICE is formed by first removing the morphological dilation operation in NICE (cf. Eq. (6.1)). This morphological dilation operation is intended for edge-detectors that produce 1-pixel width edges.[2] The morphological dilation operation is ignored, because there is no thinning process to "localize" edges for the continuous approximation of NICE. Thus, the continuous approximation of NICE is based on NICE when it is specified as

$$\mathrm{NICE}_S = \frac{\sum_{s=1}^{S} d_H(b_s[m,n], \hat{b}_s[m,n])}{\sum_{s=1}^{S} N_{b_s}}. \tag{6.2}$$

For simplicity suppose that $S = 1$, and the continuous approximation to NICE is based on the single-scale definition of NICE:

$$\mathrm{NICE}_1 = \frac{d_H(b[m,n], \hat{b}[m,n])}{N_b}. \tag{6.3}$$

[1] The Hamming distance counts the number of dissimilar elements between two vectors [50].

[2] Many edge detectors identify edges by apply a threshold to a continuous valued signal. Then, these edge detectors perform a "thinning" procedure to localize edges. Removing the morphological dilation has no effect when the edges are not subjected to "thinning."

The signals $b[m,n]$ and $\hat{b}[m,n]$ are determined from the reference image $x[m,n]$ and a test image $y[m,n]$. Specifically, $b[m,n]$ and $\hat{b}[m,n]$ are defined as

$$b[m,n] = f_{a,b}\left(\left[\sum_{k,l} x[k,l]h_v[m-k,n-l]\right]^2 + \left[\sum_{k,l} x[k,l]h_h[m-k,n-l]\right]^2 - \tau_x\right)$$
$$\hat{b}[m,n] = f_{a,b}\left(\left[\sum_{k,l} y[k,l]h_v[m-k,n-l]\right]^2 + \left[\sum_{k,l} y[k,l]h_h[m-k,n-l]\right]^2 - \tau_y\right),$$

where $h_v[m,n]$ and $h_h[m,n]$ are impulse responses that approximate either the first-derivative or second-derivative of an image in the vertical (i.e., subscript "v") and horizontal (i.e., subscript "h") directions, $f_{a,b} : \mathcal{R} \to [-\frac{1}{2}, \frac{1}{2}]$ parameterized with scalars $a$ and $b$ is a continuous monotonic nonlinearity that represents the threshold operation. The scalars $\tau_x$ and $\tau_y$ are thresholds defined as

$$\tau_x = \frac{2}{P}\sum_{m,n}\left(\left[\sum_{k,l} x[k,l]h_v[m-k,n-l]\right]^2 + \left[\sum_{k,l} x[k,l]h_h[m-k,n-l]\right]^2\right)$$
$$\tau_y = \frac{2}{P}\sum_{m,n}\left(\left[\sum_{k,l} y[k,l]h_v[m-k,n-l]\right]^2 + \left[\sum_{k,l} y[k,l]h_h[m-k,n-l]\right]^2\right),$$

where $P$ is the number of pixels in the images $x[m,n]$ and $y[m,n]$. The threshold operation is approximated by the sigmoid (i.e., hyperbolic tangent):

$$f_{a,b}(z) = a\tanh(bz). \tag{6.4}$$

For $a = 0.5$, as $b \to \infty$ then $f_{a,b}(z-\tau) \to u(z-\tau) - \frac{1}{2}$, where $u(x)$ is the unit-step (i.e., heaviside) function.

The sigmoid function $f_{a,b}$ is used to form the signals $b[m,n]$ and $\hat{b}[m,n]$ with values that lie on the interval $[-\frac{1}{2}, \frac{1}{2}]$. For an appropriate value of $b \gg 1$, the nonlinearity will make most of these elements very close the boundaries of this interval, so the $L_2$-norm can be substituted for the Hamming distance.[3] In other

[3] Notice that the $L_1$ and $L_2$ norms are equivalent to the Hamming distance if the values of $b[n]$ and $\hat{b}[n]$ belong to the set $\{-\frac{1}{2}, \frac{1}{2}\}$, which implies that $|b[n] - \hat{b}[n]| \in \{0, 1\}$.

words,

$$\text{NICE}_1 = \frac{d_H(b[m,n], \hat{b}[m,n])}{N_b} \approx \frac{\sum_n \left| b[m,n] - \hat{b}[m,n] \right|}{\sum_n |b[m,n]|} \approx \frac{\sum_n \left| b[m,n] - \hat{b}[m,n] \right|^2}{\sum_n |b[m,n]|^2}, \tag{6.5}$$

and the approximations are due to the fact that most of the elements of $b[m,n]$ and $\hat{b}[m,n]$ equal 0 or 1. The continuous approximation of $\text{NICE}_1$ is denoted by a superscript “c” and is defined as

$$\text{NICE}_1^c = \frac{\sum_n \left| b[m,n] - \hat{b}[m,n] \right|^2}{\sum_n |b[m,n]|^2}. \tag{6.6}$$

For $S > 1$, the continuous approximation of $\text{NICE}_S$ is defined as

$$\text{NICE}_S^c = \frac{\sum_{s=1}^{S} \sum_n \left| b_s[m,n] - \hat{b}_s[m,n] \right|^2}{\sum_{s=1}^{S} \sum_n |b_s[m,n]|^2}. \tag{6.7}$$

### 6.2.3 Gradient of Continuous Approximation of NICE

The gradient of $\text{NICE}_1^c$ is provided for the case of 1-D signals $x[n]$ and $y[n]$ with support $n = 0, 1, \ldots, N-1$ for simplicity. The gradient of $\text{NICE}_1^c$ for to 2-D signals is a straightforward extension of the 1-D case. Let the length $N$ vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ respectively denote the reference and test signals, where $x[n]$ is used to denote the $n^{th}$ element of $\boldsymbol{x}$. The denominator of $\text{NICE}_1^c$ is based on the reference signal $\boldsymbol{x}$, so it can be omitted when determining the gradient. Thus, a minimum is sought for the cost function $\phi(\boldsymbol{y}; \boldsymbol{x})$:

$$\phi(\boldsymbol{y}; \boldsymbol{x}) = \sum_n \left| b[n] - \hat{b}[n] \right|^2 = \sum_n \left( f_{a,b}(w[n] - \tau_x) - f_{a,b}(\hat{w}[n] - \tau_y) \right)^2. \tag{6.8}$$

The thresholds $\tau_x$ and $\tau_y$ are defined as

$$\tau_x = \frac{2}{N} \sum_{n=0}^{N-1} w[n] \qquad \tau_y = \frac{2}{N} \sum_{n=0}^{N-1} \hat{w}[n],$$

and for some impulse response $h[n]$ approximating either the first derivative or second derivative of a signal, the signals $w[n]$ and $\hat{w}[n]$ are given as[4]

$$\begin{aligned} w[n] &= [h[n] * x[n]]^2 = \left[\sum_k x[k]h[n-k]\right]^2, \\ \hat{w}[n] &= [h[n] * y[n]]^2 = \left[\sum_k y[k]h[n-k]\right]^2. \end{aligned}$$

An obvious minimum for $\phi(\boldsymbol{y};\boldsymbol{x})$ is $\boldsymbol{y} = \boldsymbol{x}$. However, there are likely other minima, since NICE is not a metric. Performing gradient descent to minimize NICE using random initial points $\boldsymbol{y}$ may reveal some of the idiosyncrasies implied by the definition of NICE.

The gradient of $\phi$ with respect to $y[m]$ is

$$\frac{\partial \phi(\boldsymbol{y};\boldsymbol{x})}{\partial y[m]} = -2 \sum_{n \in D_m} \left[ (f_{a,b}(w[n]-\tau_x) - f_{a,b}(\hat{w}[n]-\tau_y)) \frac{\partial}{\partial y[m]} f_{a,b}(\hat{w}[n]-\tau_y) \right],$$

where $D_m = \{m-L, m-(L-1), \ldots, m, m+1, \ldots, m+L\}$ is determined by assuming that the support of the signal $h[n]$ is $\{-L, -L+1, \ldots, L\}$. The partial derivative $\frac{\partial}{\partial y[m]} f_{a,b}(\hat{w}[n]-\tau_y)$ is[5]

$$\frac{\partial}{\partial y[m]} f_{a,b}(\hat{w}[n]-\tau_y) = ab\left(1 - \tanh^2(b\hat{w}[n] - b\tau_y)\right) \times \frac{\partial}{\partial y[m]}\left(b\hat{w}[n] - b\tau_y\right)$$

To determine the partial derivative $\frac{\partial}{\partial y[m]}\left(b\hat{w}[n] - b\tau_y\right)$, the following two partial derivatives are needed:

$$\begin{aligned} \frac{\partial}{\partial y[m]} \hat{w}[n] &= \frac{\partial}{\partial y[m]} \left[\sum_j y[j]h[n-j]\right]^2 \\ &= 2\left(\sum_j y[j]h[n-j]\right) h[n-m] \end{aligned}$$

[4] The signals $x[n]$ and $y[n]$ are symmetrically extended at the boundaries based on the length of the filter $h[n]$, and the signals $w[n]$ and $\hat{w}[n]$ are defined for $n = 0, 1, \ldots, N-1$.

[5] The derivative of $f_{a,b}(z) = a\tanh(bz)$ with respect to $z$ is $\frac{\partial}{\partial z} f_{a,b}(z) = ab\left(1 - \tanh^2(bz)\right)$.

and

$$
\begin{aligned}
\frac{\partial}{\partial y[m]}\tau_y = \frac{\partial}{\partial y[m]}\frac{2}{N}\sum_n \hat{w}[n] &= \frac{2}{N}\frac{\partial}{\partial y[m]}\sum_n \left[\sum_j y[j]h[n-j]\right]^2 \\
&= \frac{4}{N}\sum_n \left(\sum_j y[j]h[n-j]\right) h[n-m]
\end{aligned}
$$

Using these two partial derivatives then, $\frac{\partial}{\partial y[m]}\left(b\hat{w}[n]-b\tau_y\right)$ is

$$
\begin{aligned}
\frac{\partial}{\partial y[m]}\left(b\hat{w}[n]-b\tau_y\right) &= 2b\left(\sum_j y[j]h[n-j]\right) h[n-m] \\
&- \frac{4b}{N}\sum_k \left(\sum_j y[j]h[k-j]\right) h[k-m]. \qquad (6.9)
\end{aligned}
$$

Therefore,

$$
\frac{\partial}{\partial y[m]} f_{a,b}(\hat{w}[n]-\tau_y) = ab\left(1-\tanh^2\left(b\hat{w}[n]-b\tau_y\right)\right) \times \frac{\partial}{\partial y[m]}\left(b\hat{w}[n]-b\tau_y\right),
$$

where $\frac{\partial}{\partial y[m]}\left(b\hat{w}[n]-b\tau_y\right)$ is defined in Eq. (6.9).

### 6.2.4 Sample Images

A gradient analysis is performed to investigate the idiosyncrasies of NICE. A reference image $\boldsymbol{X}$ is selected, and several different initial distorted images $\boldsymbol{Y}$ are formed. To optimize NICE, the image $\boldsymbol{Y}$ is represented by a vector $\boldsymbol{y}$ by stacking the columns of the matrix representing $\boldsymbol{Y}$. NICE is specified using the cost function $\phi(\boldsymbol{y};\boldsymbol{x})$ defined in Eq. (6.8). The vector $\boldsymbol{y}$ is updated at iteration $k > 1$ using conjugate gradient descent [33] according to

$$
\boldsymbol{y}^{k+1} = \boldsymbol{y}^k + s^k \Delta \boldsymbol{y}^k, \qquad (6.10)
$$

using

$$\begin{aligned} s^k &= \arg\min_s \phi(\boldsymbol{y}^k + s\Delta\boldsymbol{y}^k; \boldsymbol{x}) && (6.11) \\ \Delta\boldsymbol{y}^k &= -\nabla_{\boldsymbol{y}}\phi(\boldsymbol{y}^k; \boldsymbol{x}) + \beta^k \Delta\boldsymbol{y}^{k-1} && (6.12) \\ \beta^k &= \frac{\nabla_{\boldsymbol{y}}\phi^T(\boldsymbol{y}^k; \boldsymbol{x})\nabla_{\boldsymbol{y}}\phi(\boldsymbol{y}^k; \boldsymbol{x})}{\nabla_{\boldsymbol{y}}\phi^T(\boldsymbol{y}^{k-1}; \boldsymbol{x})\nabla_{\boldsymbol{y}}\phi(\boldsymbol{y}^{k-1}; \boldsymbol{x})}, && (6.13) \end{aligned}$$

where $\nabla_{\boldsymbol{y}}\phi(\boldsymbol{y}; \boldsymbol{x})$ denotes the gradient of $\phi$ with respect to $\boldsymbol{y}$ and $\boldsymbol{y}^T$ denotes the matrix transpose. To start the conjugate gradient descent using an initial vector $\boldsymbol{y}^0$, the vector $\boldsymbol{y}^1$ is defined as

$$\boldsymbol{y}^1 = \boldsymbol{y}^0 + s^0\Delta\boldsymbol{y}^0 \qquad (6.14)$$

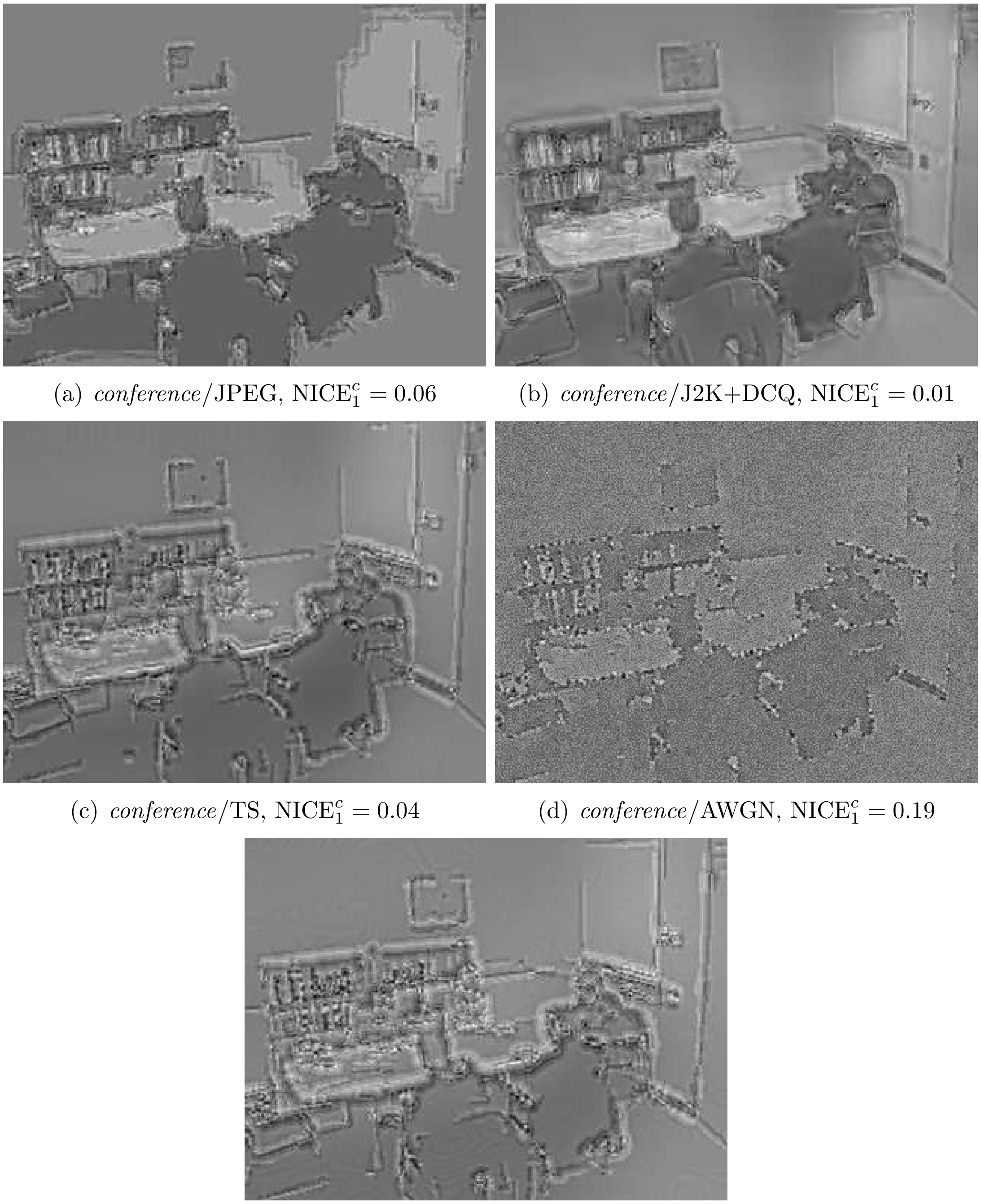
$$(6.15)$$

using

$$\begin{aligned} \Delta\boldsymbol{y}^0 &= -\nabla_{\boldsymbol{y}}\phi(\boldsymbol{y}^0; \boldsymbol{x}) && (6.16) \\ s^0 &= \arg\min_s \phi(\boldsymbol{y}^0 + s\Delta\boldsymbol{y}^0; \boldsymbol{x}). && (6.17) \end{aligned}$$

Figure 6.1 contains several images formed by minimizing $\phi(\boldsymbol{y}; \boldsymbol{x})$ with $\boldsymbol{x}$ given as the natural image *conference* and different initial images $\boldsymbol{y}$ shown in Figure 6.2. It is immediately obvious that $\text{NICE}_1^c$ does not account for distortions to very low frequency content, since all of the images lack the dynamic contrasts between light and dark regions. The edges detected by $\text{NICE}_1^c$ are determined by the responses to the band-pass filters corresponding to the steerable pyramid [117], so perturbations to low-frequency image components have no affect on $\text{NICE}_1^c$.

In each of the images, the appearance of the image edges are enhanced relative to the initial images in Figure 6.2. In particular, strong edges (e.g., the boundary between the bookshelf and the wall) are enhanced rather than weak edges (e.g., edges due to the fold in the coat in the person nearest to the camera). The finer



(a) *conference*/JPEG, $\text{NICE}_1^c = 0.06$

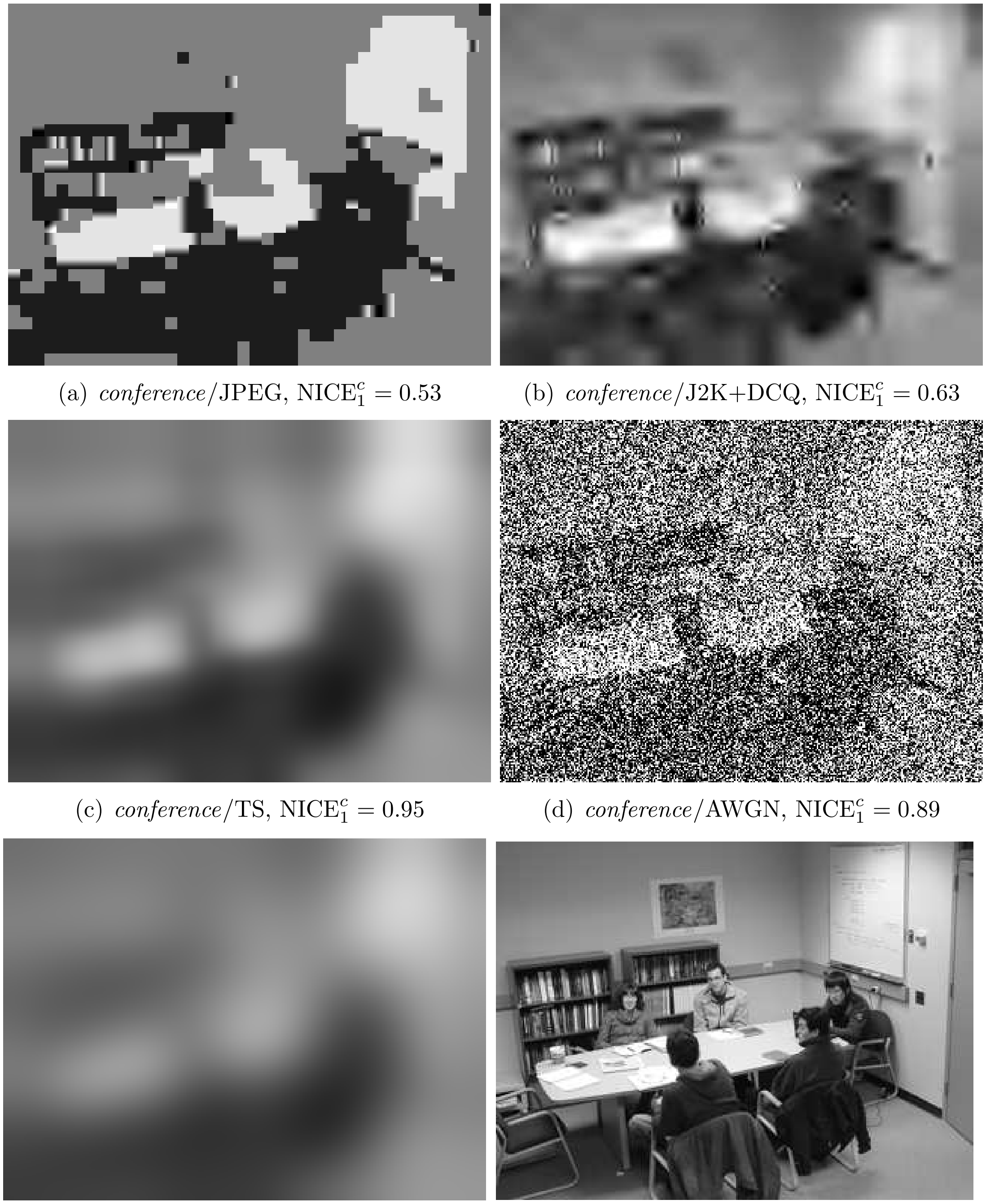
(b) *conference*/J2K+DCQ, $\text{NICE}_1^c = 0.01$

(c) *conference*/TS, $\text{NICE}_1^c = 0.04$

(d) *conference*/AWGN, $\text{NICE}_1^c = 0.19$

(e) *conference*/GBLUR, $\text{NICE}_1^c = 0.03$

Figure 6.1: Images formed by minimizing $\phi(\boldsymbol{y};\boldsymbol{x})$ using conjugate gradient descent. The value of $\text{NICE}_1^c$ for the image formed with respect to the reference image is provided.



(a) *conference*/JPEG, $\text{NICE}_1^c = 0.53$

(b) *conference*/J2K+DCQ, $\text{NICE}_1^c = 0.63$

(c) *conference*/TS, $\text{NICE}_1^c = 0.95$

(d) *conference*/AWGN, $\text{NICE}_1^c = 0.89$

(e) *conference*/GBLUR, $\text{NICE}_1^c = 0.98$

(f) *conference* (reference)

Figure 6.2: The reference image *conference* and the initial images $\boldsymbol{y}^0$ used to start the conjugate gradient descent iteration. The value of $\text{NICE}_1^c$ for the initial image with respect to the reference image is provided.

image details in the reference image are not improved. For example, the writing on the white board is not visible in any of the images.

The gradient analysis illustrates that restoring the visibility of object boundaries (i.e., edges) minimizes NICE, whereas distortions to both low-frequency and high-frequency components have little impact on NICE.

## 6.3 Using NICE for Rate-Distortion Optimization with JPEG

In previous chapters, NICE has been examined as a reliable utility estimator. That is, given a reference image and a test image, NICE provides an estimate of the usefulness of the test image relative to the reference image. Now, suppose that a reference image needs to be transmitted to a user, and a distorted image must be transmitted to satisfy an imposed rate constraint. This section examines the implementation of NICE within a rate-distortion optimization setting to form "useful" distorted natural images as defined by NICE within the JPEG framework. First, a brief overview of the JPEG coder is presented. Next, the use of genetic algorithms to perform rate-distortion optimization with a Lagrangian cost function based is described, where NICE is used as the distortion measure in the cost function. The results of the rate-distortion optimization with NICE for several images are presented and compared with images generated using the baseline JPEG coder. An example quantization table derived from the rate-distortion optimization results is described.

### 6.3.1 Overview of JPEG for Grayscale Images

JPEG achieves lossy compression of natural images by quantizing block-based discrete cosine transform (DCT) coefficients [90]. More specifically, an image is first decomposed into non-overlapping 8 × 8 blocks of pixels, and each block of pixels is transformed using the DCT. The transform coefficients in each block are quantized according to a quantization table that specifies the quantization step-size for each of 64 DCT coefficients. One quantization table is used for an image. The blocks of quantized coefficients are encoded in a raster scan order. The first DCT coefficient in each block (i.e., the DC coefficient) represents the average pixel value within that block, and this coefficient is predicted from the previously encoded block. The difference of the current DC coefficient relative to the DC coefficient of the previously block is encoded. The remaining 63 DCT coefficients of each block are processed on a block-by-block basis. Within each block, the 63 coefficients are processed in a "zig-zag" order that roughly orders the coefficients from lowest to highest spatial frequency. A run-length entropy coder is used to encode the sequence of 63 coefficients. A run-length code specifies the number of zero-valued coefficients between the current and previous nonzero valued coefficients and the value of the current nonzero valued coefficient. The run-length code is a simple entropy coder for sparse sequences. A Huffman code is used to compress the run-length encoded DCT coefficients [24].

### 6.3.2 The Rate-Distortion Optimization Problem for JPEG

The encoding bitrate of a JPEG image formed from a reference image $X$ can be determined by the choice of the quantization table $Q$ used to quantize the DCT coefficients.[6] To remain compatible with baseline JPEG, the quantization table for the luminance channel contains 64 integer values, each ranging from 1 to 255. Let $\mathcal{Q}$ denote the set of quantization tables that are compliant with baseline JPEG. The rate control problem corresponds to choosing a quantization table $Q$ for each image that will affect the encoding bitrate $R(X,Q)$ and the distortion $D(X,Q)$. For a given bitrate $R_{max}$, a quantization table $Q \in \mathcal{Q}$ is sought that solves the optimization problem:

$$\min_{Q\in\mathcal{Q}} D(X,Q) \quad \text{subject to } R(X,Q) \leq R_{max}. \tag{6.18}$$

This optimization problem can be transformed to the following unconstrained problem using Lagrange multipliers [37, 80]

$$\min_{Q\in\mathcal{Q}} J(X,Q) = D(X,Q) + \lambda R(X,Q), \tag{6.19}$$

where the nonnegative scalar Lagrange multiplier corresponds to the tradeoff between between the encoding bitrate and distortion.

With baseline JPEG, a standard approach to "rate control" is to specify a "quality" parameter $P_{jpeg} \in [0, 100]$, where higher values of $P_{jpeg}$ generally produce images that better resemble the reference image $X$ (i.e., lower distortion). This "quality" parameter $P_{jpeg}$ produces a quantization table $Q$. The original JPEG specification contains an example quantization table [124]. By scaling the

[6] The entropy coder table can also be altered to change the encoding bitrate, but only changes to the quantization table are considered

example quantization table $Q_{example}$, a quantization table $Q$ can be formed using the following equation recommended by the Independent JPEG Group (IJG) [79]

$$Q = \left\lfloor \frac{p}{100} Q_{example} + \frac{1}{2} \mathbf{1}_{8\times 8} \right\rfloor , \tag{6.20}$$

where $\mathbf{1}_{8\times 8}$ is an $8 \times 8$ matrix of ones, $\lfloor X \rfloor$ denotes rounding the elements of $X$ down to the nearest integer, and $p$ is specified according to a user specified "quality" parameter $P_{jpeg} \in [0, 100]$

$$p = \begin{cases} \frac{5000}{P_{jpeg}} & P_{jpeg} < 50 \\ 200 - P_{jpeg}^2 & P_{jpeg} \geq 50 \end{cases} . \tag{6.21}$$

The example quantization table for the luminance component is given as

$$Q_{example} = \begin{bmatrix} 16 & 11 & 10 & 16 & 24 & 40 & 51 & 61 \\ 12 & 12 & 14 & 19 & 26 & 58 & 60 & 55 \\ 14 & 13 & 16 & 24 & 40 & 57 & 69 & 56 \\ 14 & 17 & 22 & 29 & 51 & 87 & 80 & 62 \\ 18 & 22 & 37 & 56 & 68 & 109 & 103 & 77 \\ 24 & 35 & 55 & 64 & 81 & 104 & 113 & 92 \\ 49 & 64 & 78 & 87 & 103 & 121 & 120 & 101 \\ 72 & 92 & 95 & 98 & 112 & 100 & 103 & 99 \end{bmatrix} . \tag{6.22}$$

The example quantization table quantizes high frequency coefficients more coarsely than low frequency coefficients.

Baseline JPEG uses a distortion measure implied by the example quantization table along with a strategy of adjusting that quantization table corresponding to a desired "quality" level $P_{jpeg}$. When a different distortion measure $D(X, Q)$ is chosen, the rate-distortion optimization problem needs to be solved using Eq. (6.18). Using NICE as a distortion measure, the cost function in Eq. (6.19) can be minimized with respect to the quantization table $Q$ for a given value of $\lambda$, and this provides the quantization table corresponding to the optimal operating point (i.e., an encoding bitrate and distortion level) for that value of $\lambda$. An exhaustive search of all possible quantization tables to determine the quantization table that minimizes the cost function $J(X, Q)$ is unrealistic, since the search space contains $255^{64} \approx 10^{154}$ possible quantization tables. Heuristic optimization techniques tend to produce practical solutions for such difficult optimization problems, and Section 6.3.3 describes the use of the genetic algorithm search technique to minimize the cost function in Eq. (6.19) when NICE is used as the distortion measure.

### 6.3.3 Rate-Distortion Optimization using a Genetic Algorithm

A genetic algorithm (GA) is an adaptive learning heuristic that operates on a collection of points within a search space called a *population* [109]. The genetic algorithm heuristic was inspired by the theory of natural selection, where it is assumed that individuals with particular characteristics will survive to transfer those characteristics to their offspring. The genetic algorithm is used to minimize the cost function in Eq. (6.19) when NICE is used as the distortion measure due to vast number of quantization tables in the search space.

Given a cost function $J$ to be minimized, the basic structure of a GA is:

1. Generate an initial population.
2. Compute the cost of each member of the population.
3. Generate a new population using *crossover* and *mutation* operations with

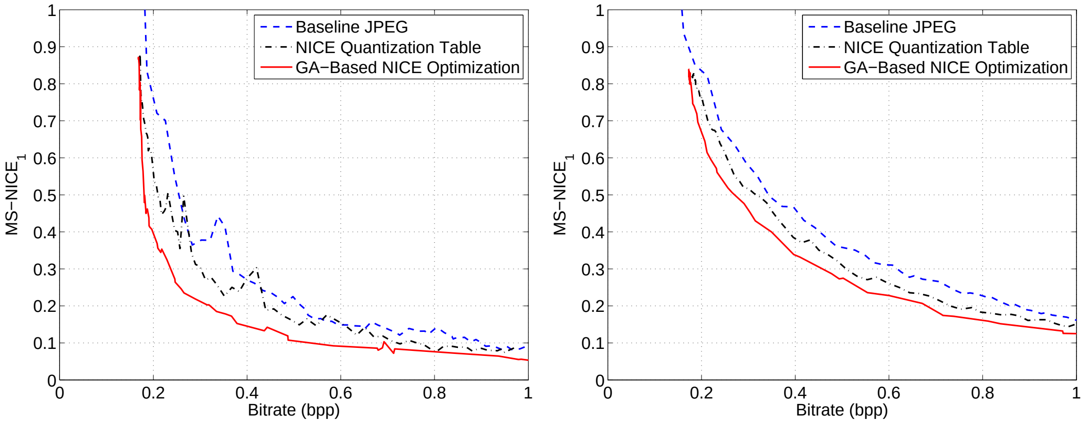
the current population.

4. Replace the old population with the new population. Go to step 2.

Each population defines a *generation*, and new generations are formed until some stopping criterion is met. The *crossover* operation generates two new members for the next population (i.e., *children*) from two members from the current population members (i.e., *parents*). The parent members were selected using the *tournament* method, where each parent member is chosen by selecting two members at random from the current population and keeping the member that yields a smaller value for the cost function. The *mutation* operation perturbs elements of members in the current population to create novel members in the next population. Last, to preserve the best members across generations a fixed number of *elite* members that yield the smallest values for the cost function among members in the current generation are carried over to the next population.

The cost function in Eq. (6.19) was minimized using a GA heuristic search to find the quantization table corresponding to the lower-bound of the rate-distortion performance for several images. MS-NICE$_1$ and MS-NICE$_4$ were separately evaluated as distortion measures. Sixty-four values of $\lambda$ were selected that were logarithmically equally spaced over the interval $[0.01, 100]$. Each value of $\lambda$ forms a unique cost function to be minimized. A population containing 640 members was used with four elite population members carried across generations. The GA was terminated when the average cost function value for the population relative to that of the previous generation was less than the threshold $\epsilon = 1 \times 10^{-9}$. When the GA terminates, the quantization table in the final population providing the smallest value for the cost function for a specific value of $\lambda$ is saved.

The GA-based rate-distortion optimization was applied to the following 8 nat-



(a) *einstein*

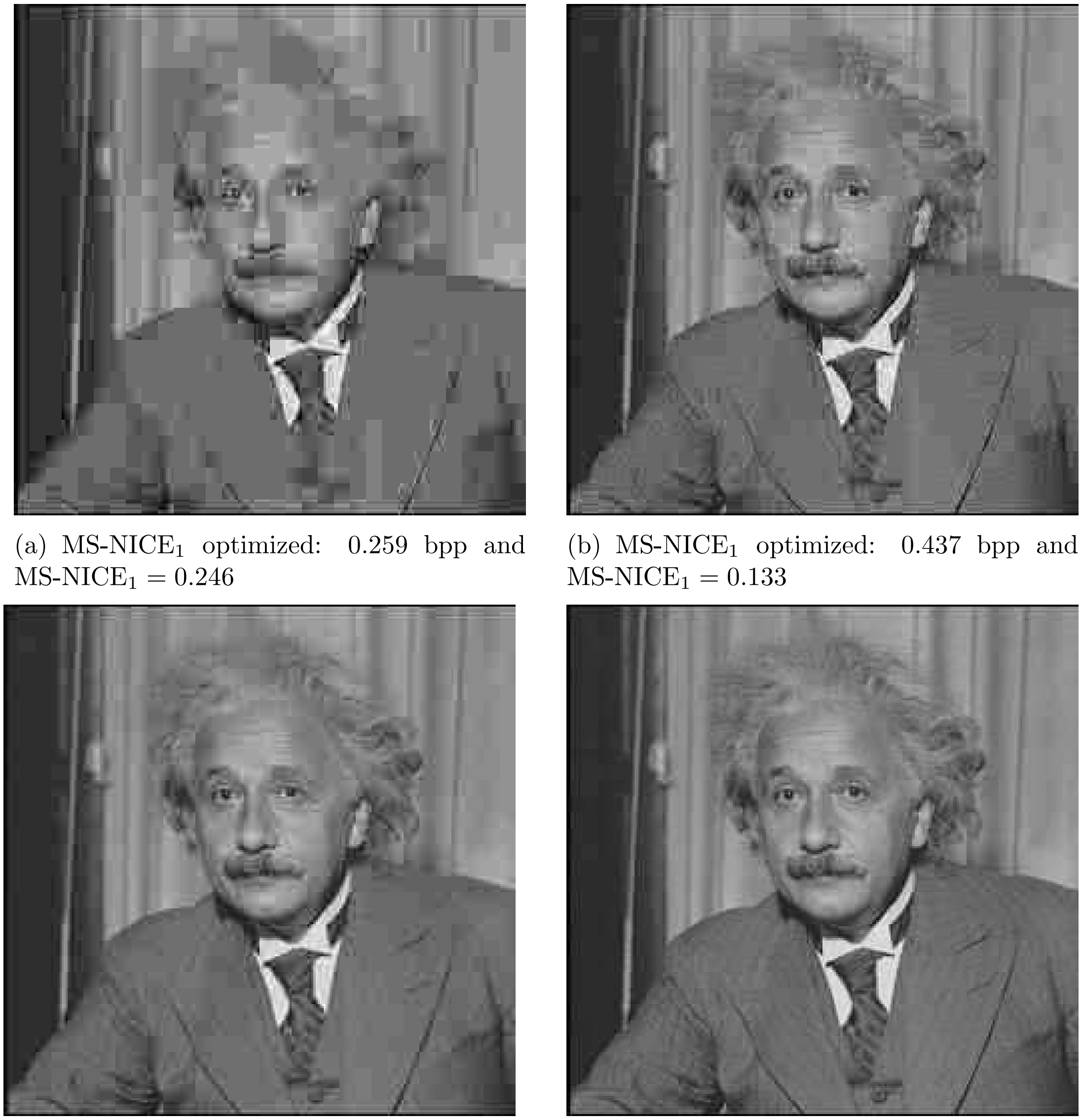
(b) *fire*

Figure 6.3: MS-NICE$_1$ as a function of encoding bitrate (bpp) for the images *einstein* and *figure*. A genetic algorithm was used to search for a quantization table minimizing the cost function in Eq. (6.19) using MS-NICE$_1$ as the distortion measure. The rate-distortion operating points corresponding to baseline JPEG using the example quantization table provided in the original specification are included in each figure. The rate-distortion operating points corresponding to the NICE optimized quantization table using Eq. (6.23) and the parameters in Table 6.1 are shown in each figure.

ural images when MS-NICE$_1$ was used as the distortion measure: *conference*, *einstein*, *elevator*, *fire*, *gas station*, *guitarist*, *pianist*, and *police*. When MS-NICE$_4$ was used as the distortion measure, the optimization was applied to the image *einstein*. These images are shown in Figures 2.4(a), 3.4, 5.3, and 5.4, which respectively appear on pages 32, 60, 158, and 159.

### 6.3.4 Results: JPEG Images Optimized using NICE

Example rate-distortion curves formed using GA-based optimization for two different natural images using MS-NICE$_1$ as the distortion measure are shown in Figure 6.3. The rate-distortion curve formed using the example quantization table pro-



(a) MS-NICE$_1$ optimized: 0.259 bpp and MS-NICE$_1$ = 0.246

(b) MS-NICE$_1$ optimized: 0.437 bpp and MS-NICE$_1$ = 0.133

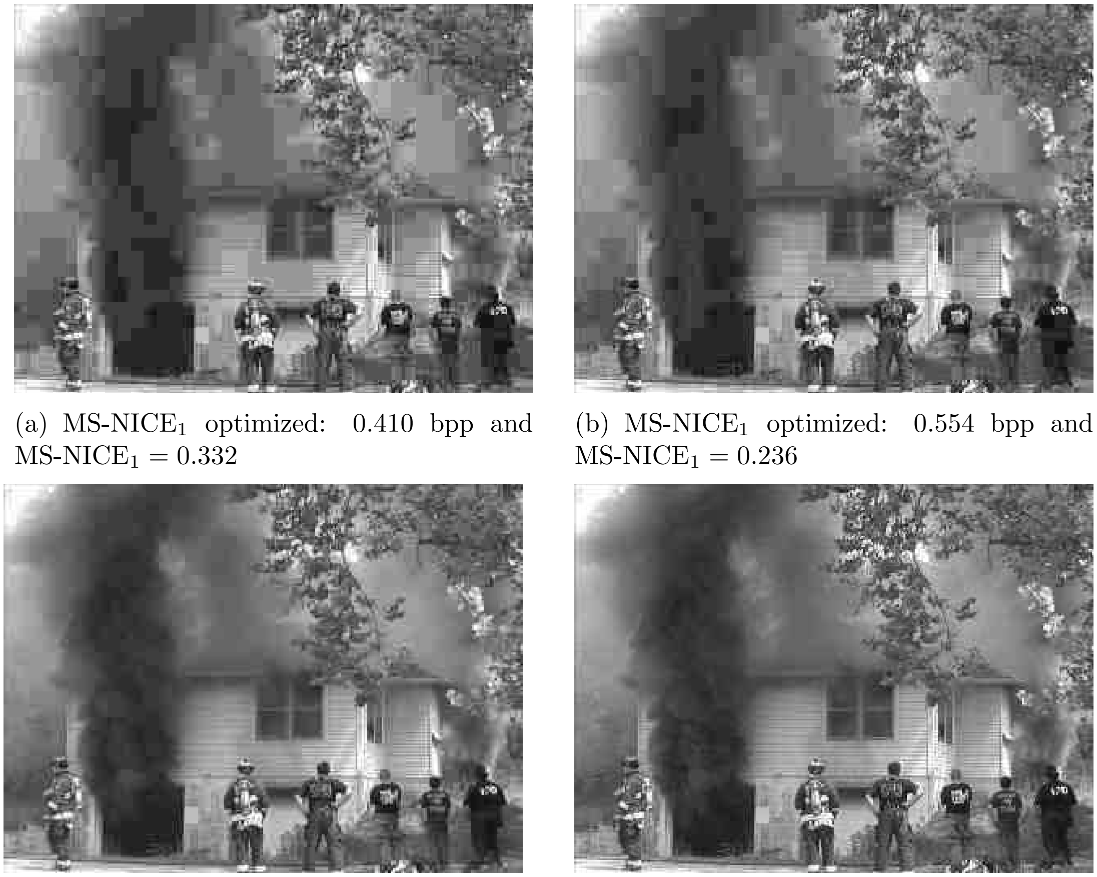
(c) Baseline JPEG: 0.434 bpp and MS-NICE$_1$ = 0.242

(d) Baseline JPEG: 0.709 bpp and MS-NICE$_1$ = 0.132

Figure 6.4: Example images for rate-distortion operating points corresponding to the curves shown in Figure 6.3(a). The images in the top row were produced using quantization tables found by minimizing the rate-distortion cost function in Eq. (6.19). The images in the bottom row were produced using the example quantization tables provided with baseline JPEG specification. The images within each column have equal MS-NICE$_1$ distortion values. The images in Figures 6.4(b) and 6.4(c) have equal encoding bitrates. The MS-NICE$_1$ optimized images have encoding bitrates 39% lower than the baseline JPEG images.

vided in the original JPEG specification and Eq. (6.20) is included for comparison in each plot. Example images corresponding to two different rate-distortion operating points along the curves for each image are shown in Figures 6.4 and 6.5. The two images in the top row of each of these two figures were formed using quantization tables that minimize MS-NICE$_1$ for a target encoding bitrate, and the two images in the bottom row were formed using the example quantization table in the original JPEG specification and Eq. (6.20). The images within the same column have equal values of MS-NICE$_1$ when compared with the reference image. The MS-NICE$_1$ optimized images have encoding bitrates 39% (*einstein*) and 25% (*fire*) lower than the baseline JPEG images with equal MS-NICE$_1$ distortion values.

The images formed using quantization tables that optimize MS-NICE$_1$ preserve the appearance of object boundaries and edges at the expense of finer image details such as textures. For example, the images in Figures 6.4(b) and 6.4(c) have the same encoding bitrate but different MS-NICE$_1$ distortion values. The MS-NICE$_1$ optimized image lacks details in the tie that are visible in the baseline JPEG image. Furthermore, the smooth shading perceived throughout the baseline JPEG image is lost in the MS-NICE$_1$ optimized image, where blocking artifacts are more visible. As another example, consider the images in Figures 6.5(b) and 6.5(c), which also have equal encoding bitrates but different MS-NICE$_1$ distortion values. The phenomenal appearance of the smoke on the left side of the baseline JPEG image is more consistent with what one experiences in the natural world as compared with the MS-NICE$_1$ optimized image. Since the smoke region contains no edges, MS-NICE$_1$ does not penalize distortions within this region. Despite the distortions within the smoke regions of the MS-NICE$_1$ optimized image, the appearance of the house, firemen, and tree branches are preserved. The detailed texture visible in the siding of the house is poorly represented in the MS-NICE$_1$, yet the impression



(a) MS-NICE$_1$ optimized: 0.410 bpp and MS-NICE$_1$ = 0.332

(b) MS-NICE$_1$ optimized: 0.554 bpp and MS-NICE$_1$ = 0.236

(c) Baseline JPEG: 0.550 bpp and MS-NICE$_1$ = 0.336

(d) Baseline JPEG: 0.758 bpp and MS-NICE$_1$ = 0.232

Figure 6.5: Example images for rate-distortion operating points corresponding to the curves shown in Figure 6.3(b). The images in the top row were produced using quantization tables found by minimizing the rate-distortion cost function in Eq. (6.19). The images in the bottom row were produced using the example quantization tables provided with baseline JPEG specification. The images within each column have equal MS-NICE$_1$ distortion values. The images in Figures 6.5(b) and 6.5(c) have equal encoding bitrates. The MS-NICE$_1$ optimized images have encoding bitrates 25% lower than the baseline JPEG images.

of the siding is maintained.

The differences among the MS-NICE$_1$ optimized images and the baseline JPEG images described in the previous paragraph are more obvious when comparing

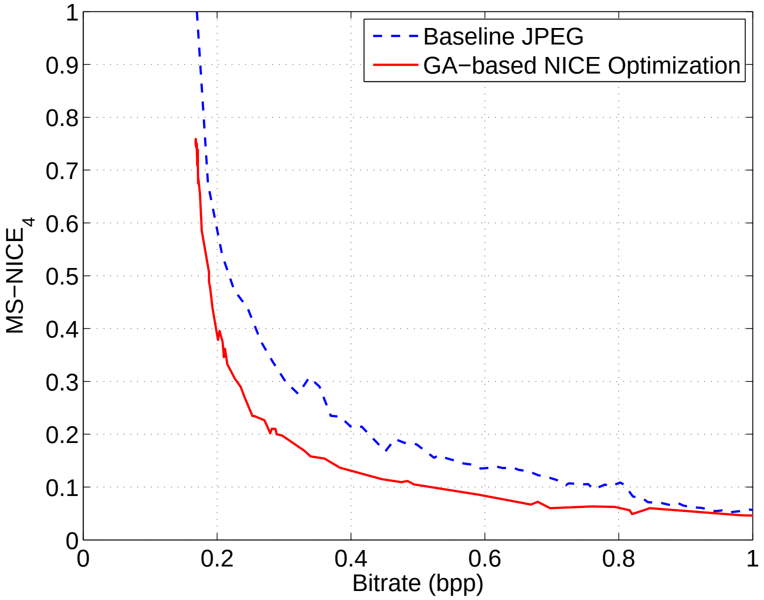
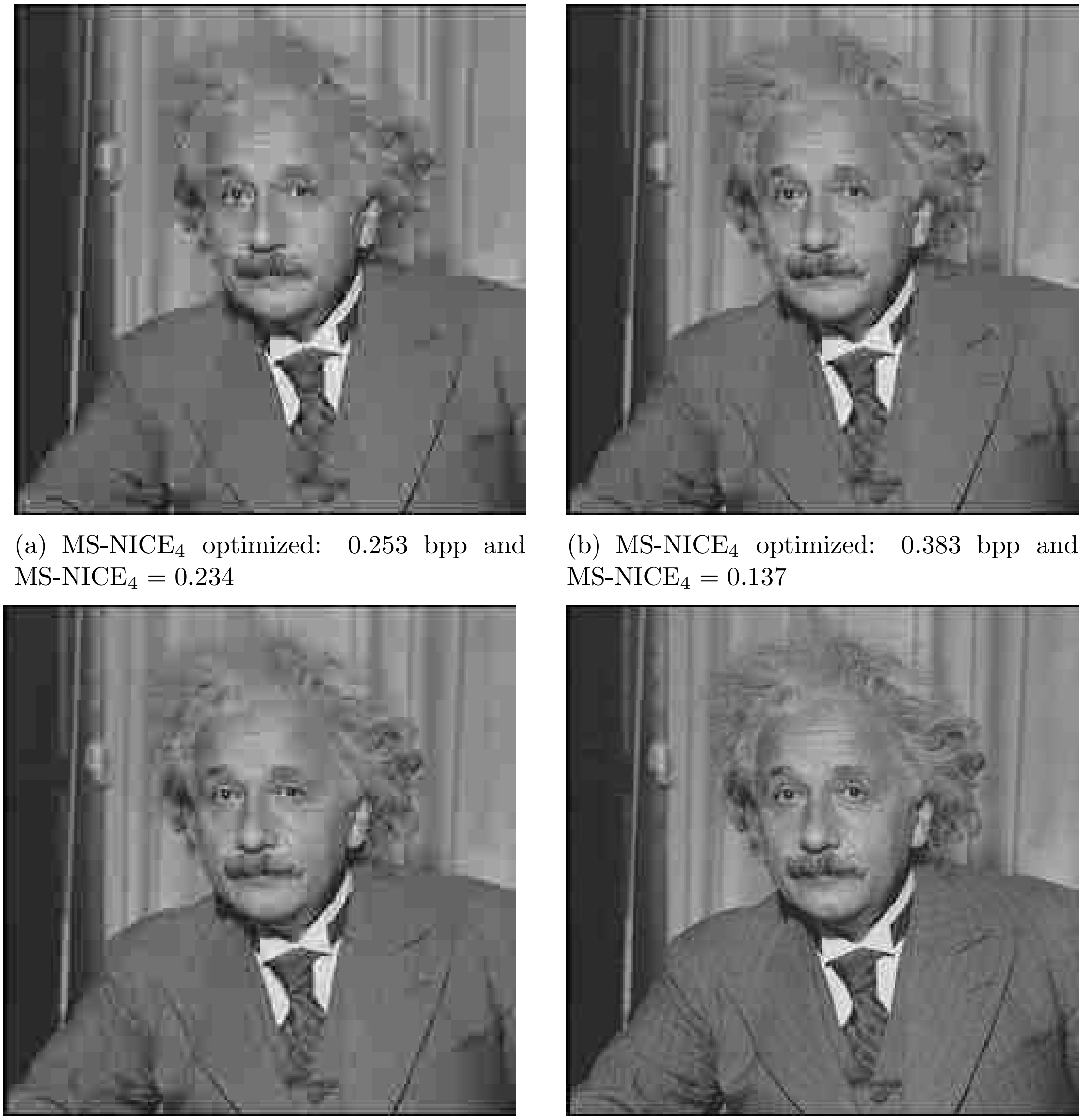


Figure 6.6: MS-NICE$_4$ as a function of encoding bitrate (bpp) for the *einstein* natural image. A genetic algorithm was used to search for a quantization table minimizing the cost function in Eq. (6.19) using MS-NICE$_4$ as the distortion measure. The rate-distortion operating points corresponding to baseline JPEG using the example quantization table provided in the original specification are included in each figure.

the images within the same column of Figures 6.4 and 6.5, which have different encoding bitrates but the same MS-NICE$_1$ distortion value. The blocking artifacts are more visible in the MS-NICE$_1$ optimized images than the baseline JPEG images for the same MS-NICE$_1$ distortion value. Thus, the loss of the textures that are visible in the baseline JPEG images but not the MS-NICE$_1$ optimized images does not impact the MS-NICE$_1$ distortion value. The MS-NICE$_1$ optimized images do preserve the appearance of the edges but not in a manner that produces a "natural" looking image. The MS-NICE$_1$ optimized images are arguable equally useful to the baseline JPEG images with the same MS-NICE$_1$ distortion level, assuming that the lack of textures does not impact their usefulness.

The rate-distortion curve formed using GA-based optimization for the *einstein*



(a) MS-NICE$_4$ optimized: 0.253 bpp and MS-NICE$_4$ = 0.234

(b) MS-NICE$_4$ optimized: 0.383 bpp and MS-NICE$_4$ = 0.137

(c) Baseline JPEG: 0.384 bpp and MS-NICE$_4$ = 0.233

(d) Baseline JPEG: 0.594 bpp and MS-NICE$_4$ = 0.135

Figure 6.7: Example images for rate-distortion operating points corresponding to the curves shown in Figure 6.6. The images in the top row were produced using quantization tables found by minimizing the rate-distortion cost function in Eq. (6.19). The images in the bottom row were produced using the example quantization tables provided with baseline JPEG specification. The images within each column have equal MS-NICE$_4$ distortion values. The images in Figures 6.7(b) and 6.7(c) have equal encoding bitrates. The MS-NICE$_4$ optimized images have encoding bitrates 34% lower than the baseline JPEG images.

image using MS-NICE$_4$ as the distortion measure is shown in Figure 6.6. The rate-distortion curve formed using the example quantization table provided in the original JPEG specification and Eq. (6.20) is included for comparison. Example images corresponding to two different rate-distortion operating points along each curve are shown in Figure 6.7. The two images in the top row of Figure 6.7 were optimized to minimize MS-NICE$_4$, and the two images in the bottom row of Figure 6.7 were formed using the example quantization table in the original JPEG specification and Eq. (6.20). The images within the same column have equal values of MS-NICE$_4$ when compared with the reference *einstein* image. The MS-NICE$_4$ optimized images have encoding bitrates 34% lower than the baseline JPEG images.

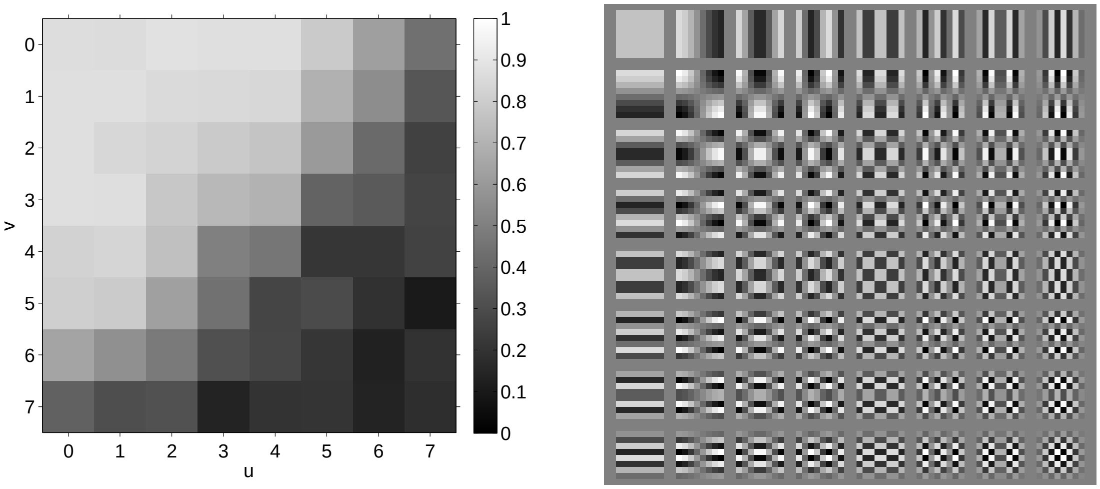
Similar to the MS-NICE$_1$ optimized JPEG images, there are striking visual differences among the images shown in Figure 6.7. Blocking artifacts are far more obvious in the MS-NICE$_4$ optimized images relative to the baseline JPEG images with equal MS-NICE$_4$ distortion values. As a consequence, the forehead region appears smoother in the baseline JPEG image in Figure 6.7(c) than the MS-NICE$_4$ optimized image in Figure 6.7(a). The MS-NICE$_4$ optimized images also lack the finer image details that are visible in the baseline JPEG images. For example, the textures in the hair and the lines in the suit are visible in the baseline JPEG image shown in Figure 6.7(d), but these details are absent in MS-NICE$_4$ optimized image in Figure 6.7(b).

### 6.3.5 JPEG Quantization Table Optimized for NICE

The Lagrange multipliers $\lambda$ specify a tradeoff between the distortion and the encoding bitrate and are used to parameterize a quantization table derived from the rate-distortion optimization when the distortion measure is MS-NICE$_1$. For each natural image, a quantization table was found that minimized the cost function in Eq. (6.19) for a specified value $\lambda$. For each value of $\lambda$, the quantization tables corresponding to each image were averaged. A mapping from $\lambda$ to each value of the quantization table was found. Thus, a total of 64 mappings were formed, each corresponding to a specific step-size in the quantization table.

Let $k$ denote a natural image, and let $Q_{u,v,k}(\lambda_j)$ denote the quantization step-size corresponding to the $(u, v)$ DCT coefficient, $u = 0, \ldots, 7$ and $v = 0, \ldots, 7$, when the cost function in Eq. (6.19) is minimized for $\lambda = \lambda_j$. Let $N_\lambda = 64$ denote the number of values of $\lambda$ for which Eq. (6.19) was minimized. For each $\lambda_j$, the average quantization step-size for the $(u, v)$ DCT coefficient across all $K = 8$ images was computed $\bar{Q}_{u,v}(\lambda_j) = \sum_{k=1}^{K} Q_{u,v,k}(\lambda_j)$.

The relationship between the average quantization step-size $\{\bar{Q}_{u,v}\}_{j=1}^{N_\lambda}$ and $\{\lambda_j\}_{j=1}^{N_\lambda}$ was observed to exhibit a nonlinear relationship for many DCT coefficients. The absolute value Spearman rank order correlation coefficients (ROCC) between $\{Q_{u,v,k}(\lambda_j)\}_{j=1}^{N_\lambda}$ and $\{\lambda_j\}_{j=1}^{N_\lambda}$ for the $(u, v)$ DCT coefficient was determined for each natural image. The average absolute ROCC across the $K = 8$ natural images was found Figure 6.8(a) shows the average absolute ROCC as an intensity image, where brighter intensities correspond to higher correlation coefficients. The average absolute ROCC values are arranged in Figure 6.8(a) to correspond with $8 \times 8$ DCT basis functions shown in Figure 6.8(b). The average absolute ROCC is greatest for DCT coefficients corresponding to either horizontal or vertical frequencies, whose basis functions can be combined to produce edges in natural images. Textures within natural images are represented using high frequency DCT coefficients (i.e., $v + u \geq 8$). The average absolute ROCC values corresponding to these



(a) Average absolute ROCC of $\{Q_{u,v,k}(\lambda_j)\}_{j=1}^{N_\lambda}$ and $\{\lambda_j\}_{j=1}^{N_\lambda}$

(b) $8 \times 8$ DCT basis functions

Figure 6.8: The average absolute Spearman rank correlation coefficient (ROCC) between $\{Q_{u,v,k}(\lambda_j)\}_{j=1}^{N_\lambda}$ and $\{\lambda_j\}_{j=1}^{N_\lambda}$ for each DCT coefficient $(u, v)$, $u = 0, \ldots, 7$ and $v = 0, \ldots, 7$, is shown in Figure 6.8(a). The average absolute ROCC was taken across the $K = 8$ natural images. The ROCC are arranged to correspond with the basis functions corresponding to the $8 \times 8$ DCT shown in Figure 6.8(b). Let $(m, n)$ index the pixel at $m^{th}$ row and $n^{th}$ column spatial location in the $M \times N$ block of pixels. The basis function at row $v$ and column $u$ is defined as $\frac{1}{4MN} \cos(\pi(2m + 1)v) \cos(\pi(2n + 1)u)$ for $m = 0, \ldots, M - 1$ and $n = 0, \ldots, N - 1$, where $M = N = 8$.

high frequency DCT coefficients are less than 0.5. The small correlations between $\lambda$ and $Q_{u,v,k}$ are consistent with NICE, since NICE does not account for distortions to textures within natural images.

The image-independent quantization table parameterized by $\kappa = 25 \log_{10} \lambda + 50$ for $\lambda > 0$ was formed using the data from the rate-distortion optimization. Specifically, a nonlinearity mapping $f_{u,v}$ defined as

$$f_{u,v}(\kappa_j) = a_{u,v} \left( \frac{1}{1 + \exp\left[-b_{u,v}(\kappa - c_{u,v})\right]} + d_{u,v} \right), \tag{6.23}$$

was fitted to the data to map the set $\{\kappa_j\}_{j=1}^{N_\lambda}$ to the set of quantization step-sizes $\{\bar{Q}_{u,v}(\kappa_j)\}_{j=1}^{N_\lambda}$. The parameters $\{a_{u,v}, b_{u,v}, c_{u,v}, d_{u,v}\}$ were chosen to minimize the sum of squared errors $\{(f_{u,v}(\kappa_j) - \bar{Q}_{u,v}(\kappa_j))^2\}_{j=1}^{N_\lambda}$. Thus, a set of 64 nonlinear mappings $\{f_{u,v}(\kappa)\}_{u=1,v=1}^{8,8}$ were fitted to the data. The fitted parameters $\{a_{u,v}, b_{u,v}, c_{u,v}, d_{u,v}\}$ corresponding to each nonlinear function $f_{u,v}(\kappa)$ are provided in Table 6.1. As $\kappa \in [0, 100]$ increases, $f_{u,v}(\kappa)$ creates quantization step-sizes that correspond to larger distortion as measured by MS-NICE$_1$.

The rate-distortion operating points for the images *einstein* and *fire* corresponding to quantization tables produced using Eq. (6.19) with the parameters in Table 6.1 are included in Figures 6.3(a) and 6.3(b). The mapping was formed by averaging the quantization tables minimizing the cost function in Eq. (6.19) for a specific value of $\lambda$ for several different natural images. As a result, the mapping does not provide the optimal quantization table for each image. However, the mapping does provide quantization tables that lead to rate-distortion operating points better suited to NICE than the rate-distortion operating points when using the example quantization tables provided with the JPEG specification.

## 6.4 Discussion

Genetic algorithms (GA) are shown to produce quantization tables that are compliant with baseline JPEG while adhering to different NICE distortion measures. The results from the rate-distortion optimization were used to produce a mapping $f_{u,v}(\kappa)$ to generate a quantization table given the parameter $\kappa$. GA have been used to determine quantization tables to optimize the "quality" of medical images, but a parameterized mapping similar to $f_{u,v}(\kappa)$ was not formed [143].

The combined GA-based optimization with a parametric mapping $f(\kappa)$ could

Table 6.1: Fitted parameters for the nonlinearity $f_{u,v}(\kappa)$ defined in Eq. (6.23) to produce quantization step-sizes based on MS-NICE$_1$ for a value of $\kappa \in [0, 100]$. Smaller values of $\kappa$ correspond to less distortion as measured by MS-NICE$_1$.

| $(u, v)$ | $a_{u,v}$ | $b_{u,v}$ | $c_{u,v}$ | $d_{u,v}$ | $(u, v)$ | $a_{u,v}$ | $b_{u,v}$ | $c_{u,v}$ | $d_{u,v}$ |
|---|---|---|---|---|---|---|---|---|---|
| (0, 0) | 239 | 0.0959 | 51.5 | 0.0724 | (4, 0) | 241 | 0.104 | 61.6 | 0.0442 |
| (0, 1) | 243 | 0.0955 | 55.8 | 0.0678 | (4, 1) | 215 | 0.112 | 56 | 0.0696 |
| (0, 2) | 243 | 0.101 | 62.6 | 0.0526 | (4, 2) | 205 | 0.107 | 48.3 | 0.0799 |
| (0, 3) | 252 | 0.0904 | 66.6 | 0.0432 | (4, 3) | 181 | 0.109 | 34 | 0.166 |
| (0, 4) | 234 | 0.0953 | 64.6 | 0.0381 | (4, 4) | 154 | 0.0812 | 26.4 | 0.298 |
| (0, 5) | 219 | 0.0998 | 54 | 0.0609 | (4, 5) | 68.8 | 0.117 | 33.4 | 1.67 |
| (0, 6) | 210 | 0.0702 | 42.9 | 0.077 | (4, 6) | 0.265 | -26.6 | 54.5 | 614 |
| (0, 7) | 0.205 | -29.4 | 48.7 | 859 | (4, 7) | 40 | 0.108 | 21.5 | 3.47 |
| (1, 0) | 241 | 0.0968 | 58.1 | 0.064 | (5, 0) | 219 | 0.0996 | 57.6 | 0.0647 |
| (1, 1) | 244 | 0.0986 | 54.5 | 0.0533 | (5, 1) | 192 | 0.139 | 46.6 | 0.0984 |
| (1, 2) | 233 | 0.116 | 58.6 | 0.0826 | (5, 2) | 163 | 0.133 | 43 | 0.217 |
| (1, 3) | 225 | 0.142 | 59.1 | 0.0744 | (5, 3) | 139 | 0.102 | 29 | 0.38 |
| (1, 4) | 228 | 0.105 | 54.7 | 0.0521 | (5, 4) | 72.7 | 0.0326 | -11.5 | 1.37 |
| (1, 5) | 206 | 0.113 | 48.8 | 0.078 | (5, 5) | 0.346 | -25.5 | 67.8 | 439 |
| (1, 6) | 152 | 0.132 | 36.7 | 0.336 | (5, 6) | -0.144 | 38.9 | 40.4 | -1090 |
| (1, 7) | 0.164 | -36.7 | 40.8 | 1000 | (5, 7) | 15.8 | 26.2 | 43.7 | 9.55 |
| (2, 0) | 247 | 0.0893 | 60.3 | 0.0452 | (6, 0) | 189 | 0.132 | 44.1 | 0.174 |
| (2, 1) | 235 | 0.116 | 58.3 | 0.0725 | (6, 1) | 159 | 0.107 | 34.7 | 0.242 |
| (2, 2) | 226 | 0.104 | 54.5 | 0.0897 | (6, 2) | 78.4 | 0.0368 | 87.5 | 1.79 |
| (2, 3) | 213 | 0.106 | 52.2 | 0.0923 | (6, 3) | 0.202 | -28.4 | 48.2 | 793 |
| (2, 4) | 210 | 0.11 | 49.5 | 0.0742 | (6, 4) | 69.7 | 0.1 | 32.4 | 1.59 |
| (2, 5) | 186 | 0.0918 | 41.5 | 0.159 | (6, 5) | 108 | 0.115 | 4.01 | 0.555 |
| (2, 6) | 45.9 | 0.0865 | 71.5 | 2.98 | (6, 6) | 5.35 | 0.551 | -7.59 | 27.2 |
| (2, 7) | 92.8 | 0.126 | 23.2 | 1.01 | (6, 7) | 15.3 | 0.058 | 47.8 | 9.49 |
| (3, 0) | 276 | 0.0745 | 71.2 | 0.0309 | (7, 0) | 131 | 0.129 | 35 | 0.6 |
| (3, 1) | 231 | 0.107 | 57.7 | 0.0544 | (7, 1) | 87.2 | 0.112 | 26.5 | 1.21 |
| (3, 2) | 215 | 0.103 | 54 | 0.0847 | (7, 2) | 26.2 | 0.0125 | 206 | 6.34 |
| (3, 3) | 195 | 0.121 | 44.9 | 0.128 | (7, 3) | 0.228 | -25.2 | 26.4 | 691 |
| (3, 4) | 182 | 0.103 | 38.3 | 0.15 | (7, 4) | -0.277 | 95.2 | 66.4 | -569 |
| (3, 5) | 139 | 0.117 | 32.7 | 0.406 | (7, 5) | -0.863 | -24.9 | 19.6 | -173 |
| (3, 6) | 81.2 | 0.124 | 29.3 | 1.28 | (7, 6) | 98 | 0.0377 | -31.4 | 0.646 |
| (3, 7) | -67.7 | -0.0629 | 22.5 | -2.69 | (7, 7) | 41.6 | 0.0603 | 42.1 | 3.21 |

be applied to other objective estimators and other image coders (e.g, JPEG-2000) as a means to investigate the "optimal" images for that objective estimator when operating within the constraints of the coder. This approach would allow existing objective estimators to be analyzed in greater detail. Furthermore, existing objective estimators may be used to improve existing coders by applying, via the GA-based optimization, the objective estimator to a coder. As a final step, the images produced using this method could be evaluated by human observers to determine if the objective estimator actually improves the coder according to some criterion such as perceived utility or perceived quality.

## 6.5 Summary

The chapter presents an analysis of the natural image contour evaluation (NICE) utility estimator. A gradient analysis is conducted based on a continuous approximation to NICE that reveals those image features that minimize NICE. This analysis demonstrates that maintaining the phenomenal appearance of object boundaries and edges coincides with minimizing NICE. The second half of the chapter presents a method to produce *distorted but useful* images compatible with a baseline JPEG coder based on NICE. In particular, an image independent parameterized quantization table is provided based the results of a rate-distortion optimization using NICE. A genetic algorithm is used to search for the optimal quantization tables for specific rate-distortion tradeoffs. This set of quantization tables is parameterized to circumvent the need for the genetic algorithm. The parameterized tables are suboptimal, since they are not tailored to the statistics of the source image. The use of genetic algorithms to perform rate-distortion optimization with other estimators is discussed.

## 6.6 Key Points

- A gradient analysis of the natural image contour evaluation (NICE) utility estimator demonstrates that the appearance of object boundaries and edges coincides with minimizing NICE.
- *Distorted but useful* images compatible with a baseline JPEG coder are formed to minimize NICE for a specific rate-distortion tradeoff. Rate-distortion optimization is performed using a genetic algorithm, since NICE was developed for use with the block-based discrete cosine transform (DCT) used by JPEG.
- An image independent parametric quantization table that is compatible with a baseline JPEG coder is presented as a suboptimal alternative to rate-distortion optimization.
- The use of genetic algorithms to perform rate-distortion optimization with other estimators is discussed.

## CHAPTER 7
# CONCLUSION

Natural images from imaging systems supply information that facilitate human observers performing various tasks. This dissertation examined human performance when performing a broad task with natural images: reporting the content of a distorted image. Novel experiments were conducted to measure the usefulness of distorted natural images in terms of this task. In addition, experiments were conducted to measure the perceived quality of these same distorted natural images. Results from both subjective experiments were compared and revealed the perceived quality does not imply an image's perceived utility. In particular, a distortion that removes low-frequency content from an image demonstrated that perceived utility is largely based on the fidelity of high-frequency content and is less affected by distortions to low-frequency content, whereas distortions to any frequency content affects perceived quality. The observed relationship between utility and quality implies that accurate *objective* quality (utility) estimators will not accurately estimate perceived utility (quality) for a broad class of distortions.

The natural image contour evaluation (NICE) utility estimator was introduced as a novel signal analysis tool to estimate the perceived utility of distorted natural images. NICE estimates utility as a function of both *lost* and *introduced* contour information in a distorted image when compared with a reference image. NICE abstractly represents the reference and test images as contours and compares these contours to estimate utility. NICE was shown to be a viable signal analysis tool to estimate the usefulness of a distorted natural image. This result supports hypotheses about the importance of contour information to the human visual system for object perception.

Along with NICE, several objective estimators, mostly designed to estimate perceived quality, were assessed in terms of their performance as utility and quality estimators. Apart from NICE, none of the objective estimators provided reliable estimates of perceived utility across a variety of distortions. The visual information fidelity (VIF) criterion, which is customarily used as a quality estimator, was modified to produce VIF* and shown to outperforms VIF as a quality estimator on the current database of distorted images. The signal analyses conducted by VIF and VIF* are consistent with the observations from the subjective experiments. Specifically, VIF primarily analyzes disruptions to high-frequency content and accurately estimates perceived utility but not perceived quality, whereas VIF* exhibits increased sensitivity to low-frequency distortions relative to VIF and analyzes disruptions to all frequency content and accurately estimates perceived quality but not perceived utility.

Recognizing the limitations to the initial experiments conducted to obtain perceived utility scores, a novel methodology was proposed, tested, and used to evaluate the performance of NICE as well as several quality estimators. The method is suitable to applications when observers are simply faced with the task of determining the content of a distorted image. For example, observers are not directed to search for specific objects being held. The method generates the definitions of utility for each image based on the observer responses by forming observer-centric concepts. This framework is flexible and can support a broad rage of tasks.

In addition to *measuring* (i.e., via experimental methods) and *estimating* (i.e., using NICE) the perceived utility of distorted natural images, a procedure is developed and used to *generate* distorted but useful natural images. An image independent parametric quantization table compatible with baseline JPEG was formed via rate-distortion optimization using NICE as a distortion measure with a JPEG coder. A genetic algorithm was used to conduct the rate-distortion optimization for specific rate-distortion tradeoffs. This procedure is discussed as at tool that could be applied to other objective estimators as well as other image coders to analyze performance of those objective estimators.

APPENDIX A

# RECOGNITION THRESHOLDS OF LINE DRAWINGS

## A.1 Introduction

The reference/texture smoothing (TS) sequences used in the experiments described in Chapter 3 were preceded by reference/line drawing sequences. This chapter describes the experiments to estimate the recognition threshold of a natural image using reference/line drawing sequences. In addition, the recognition threshold was found for sequences formed using 1) JPEG-2000 (J2K) distortions and 2) J2K distortions introduced by using step-sizes specified by the dynamic contrast-based quantization (DCQ) strategy [19].

Qualitative comparisons between the encoding bitrates corresponding to the recognition threshold from all three types of distortions suggest an underlying image characteristic that facilitates content recognition: *visual structure*. *Visual structure* loosely corresponds to the object boundaries and edges within an image. Thus, the line drawings are hypothesized to explicitly preserve the visual structure of a natural image. The J2K and J2K+DCQ distortions retain the visual structure of a natural image as long as the corresponding wavelet components are preserved. Despite the evidence that visual structure is a fundamental characteristic that predicts recognition of natural images, the disparity between the line drawing representation and the wavelet basis representation used by a JPEG-2000 encoder restricts the quantitative comparison of the results from the different types of distortions. The texture smoothing distortions, which preserve object boundaries and edges, were selected to replace line drawings, since the texture smoothing distortions afford a quantitative analysis with the J2K and J2K+DCQ distortions.

The remainder of this chapter is organized as follows: Section A.2 describes the experimental methods used. Section A.3 presents the results from the experiments, which are discussed in Section A.4. The chapter ends with a summary in Section A.5.

## A.2 Experimental Methods

An experiment was conducted to determine the bitrate corresponding to an observer's recognition threshold for reference/line drawing sequences, reference/J2K sequences, and reference/DCQ+J2K sequences.

### A.2.1 Distortions

The nine grayscale natural images used in the experiments described in Chapter 3, shown in Figure 3.1(a), Figure 3.3, and Figure 3.4, were used to generate reference/distortion sequences. The two types of distortions are described.

**Line Drawings**

Object structure is widely believed to rely on the perception of image details, such as sharp edges, which are conveyed by the high spatial frequencies [31, 72]. Edges, defined spatially by sudden intensity changes, may be identified by either the presence of an absolute maximum in the first derivative of an image or a zero-crossing in its second derivative[1] [72]. Line drawings formed from natural images

[1] Marr and Hildreth favored detecting edges by convolving the image with the Laplacian of a two-dimensional, circularly symmetric Gaussian, since it is locally optimized in both the spatial

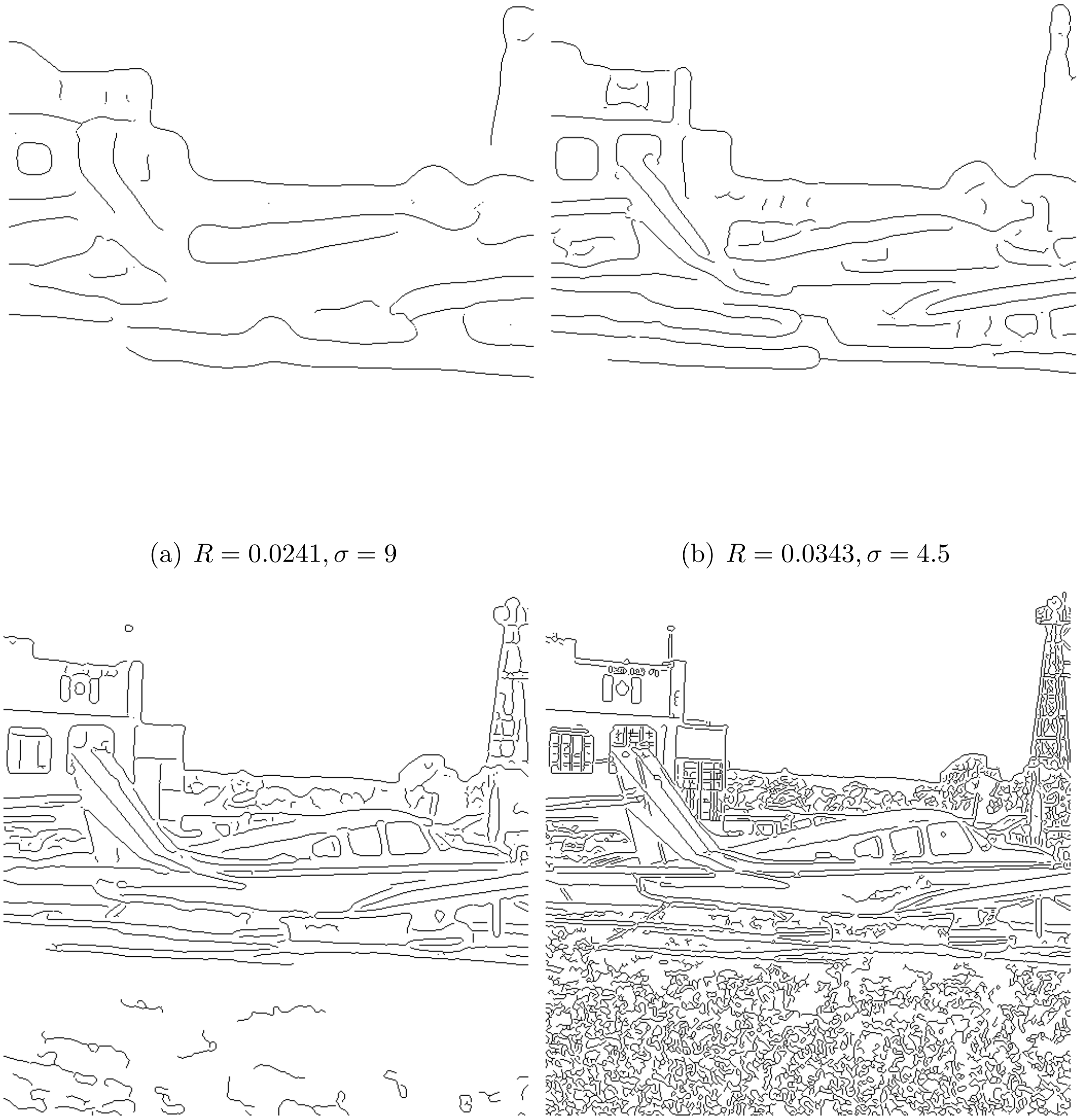
were generated with the Canny edge detector [13]. The Canny edge detector filters an image with the derivative of a Gaussian specified for a particular $\sigma > 0$ and applies thresholding to generate a binary image. The parameter $\sigma$ in the Canny filter controls the suppression of high frequencies before detecting edges. The frequency response of a derivative of a Gaussian filter is essentially a band-pass filter, suppressing both very low and very high spatial frequencies. Decreasing $\sigma$ retains more high frequency content, and the resulting line drawing will include finer image details. The bitrate $R$ of a line drawing was determined by compressing the binary image with a JBIG coder, which is the standard fax compression algorithm for bi-level images.

A reference/line drawing sequence was formed for each natural image by varying $\sigma$ in the Canny edge detector from 0.5 to 10 with an increment of 0.5, where $\sigma$ varies inversely with respect to the bitrate. Select images from the *airplane*/line drawing sequence are shown in Figure A.1.

### J2K and J2K+DCQ: Quantized Discrete Wavelet Transform Coefficients

The lossy JPEG-2000 (J2K) image compression standard represents natural images as a linear combination of wavelet basis functions [123]. Distortions are introduced by quantizing the basis function coefficients found using a discrete wavelet transform to achieve a desired encoding bitrate, $R$. The dynamic contrast-based quantization (DCQ) strategy assigns quantization step-sizes according to a measure of visual distortion parameterized by characteristics of the image, the wavelet

and frequency domains and the economy of computation of the Laplacian operator [72]. Later, Canny considered the identification of edges by convolving the image with the first derivative of a Gaussian [13]. Canny's approach demonstrated enhanced edge detection by incorporating the edge orientations.



(a) $R = 0.0241, \sigma = 9$

(b) $R = 0.0343, \sigma = 4.5$

(c) $R = 0.0684, \sigma = 2$

(d) $R = 0.2445, \sigma = 1$

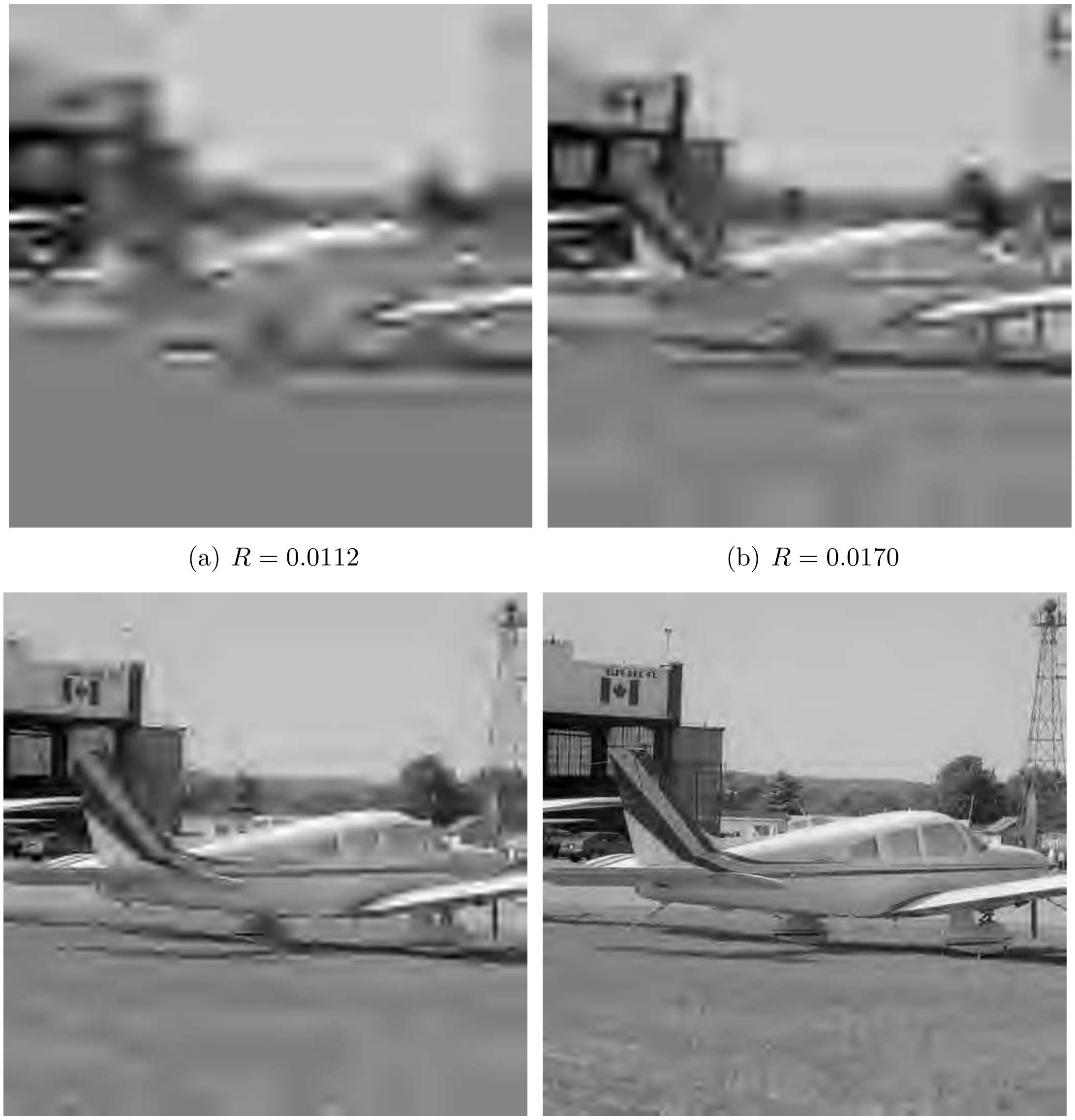
Figure A.1: Selected images from the *airplane*/line drawing sequence. The bitrate $R$ using the JBIG coder and the parameter $\sigma$ for the Canny edge detector are provided with each image..

subband coefficients, and the display [19]. The DCQ strategy's visual distortion measure distinguishes visually lossless images from visibly distorted images, so the DCQ strategy can specify subband quantization step-sizes for lossy compression that yield a visually lossless image.

A reference/J2K sequence of images has distortions due to JPEG-2000 compression using the baseline implementation and evolves by increasing the encoding bitrate, $R$. Similarly, a reference/J2K+DCQ sequence of images has distortions due to JPEG-2000 compression using the DCQ strategy and evolves by increasing the encoding bitrate, $R$. Select images from the *airplane*/J2K and *airplane*/J2K+DCQ sequences are shown in Figure A.2 and Figure A.3, where the images chosen in both sequences have approximately the same values of $R$.

### A.2.2 Stimuli

For each of the nine images, reference/distortion sequences were formed using a dense set of parameters for the line drawing, J2K, and J2K+DCQ distortions. Reference/line drawing sequences contained 20 images by varying $\sigma$ for the Canny edge detector from 0.5 to 10 with an increment of 0.5. Reference/J2K and reference/J2K+DCQ sequences contained 20 images corresponding to encoding bitrates that were logarithmically equally spaced from 0.01 and 0.3.



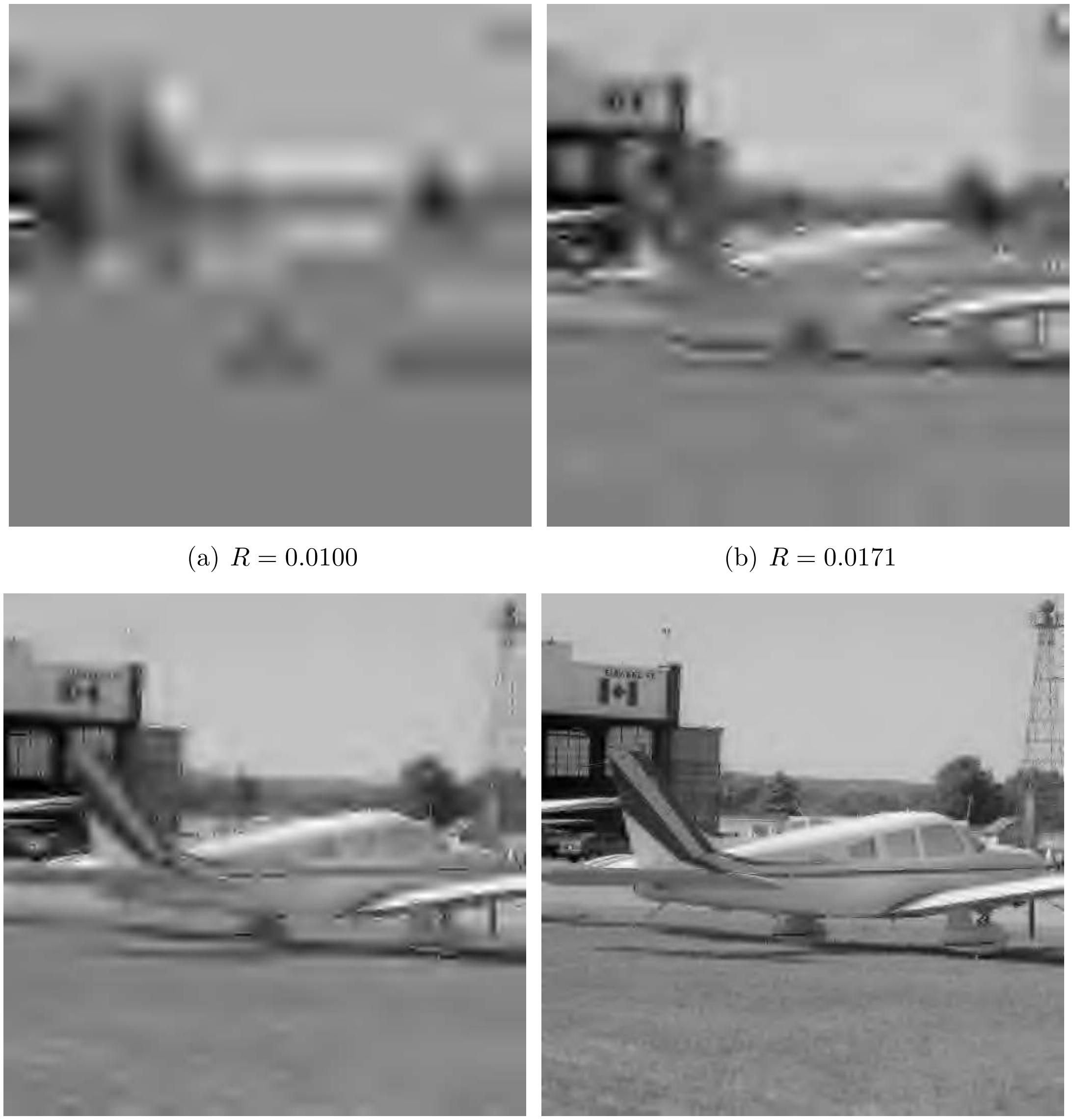
(a) $R = 0.0112$

(b) $R = 0.0170$

(c) $R = 0.0352$

(d) $R = 0.1216$

Figure A.2: Selected images from the *airplane*/J2K sequence along with the corresponding encoding bitrate $R$.



(a) $R = 0.0100$

(b) $R = 0.0171$

(c) $R = 0.0349$

(d) $R = 0.1231$

Figure A.3: Selected images from the *airplane*/J2K+DCQ sequence along with the corresponding encoding bitrate $R$.

### A.2.3 Procedure

Observers[2] viewed one reference/distortion sequence for several (if not all) of the nine natural images. Not every observer viewed each type of distortion for the natural images. Observers viewing reference/J2K+DCQ sequences also viewed reference/line drawing sequences. However, observers viewing reference/J2K sequences did not view any reference/line drawing or reference/J2K+DCQ sequences. For each image in a sequence, the observer provided a typed descriptions of the image content. The next image in the sequence was shown upon submission of a description; a time limit was not imposed. Participants typically completed the experiment session in about 30 minutes.

### A.2.4 Participants

Forty observers with normal or corrected-to-normal acuity participated in this experiment. Each series of representations was viewed by at least 9 observers and at most by 14. On average the reference/line drawing sequences were viewed by 12.1 observers. The reference/J2K+DCQ sequences were viewed by 11.8 observers on average, and the reference/J2K sequences were viewed by 14 observers.

## A.3 Results: Recognition Thresholds

An observer's point of recognition was identified when the description contained both adequate and accurate information to briefly describe the image content.[3] For

[2] These observers were called *writers* in the experiments described in Chapter 3.2.3.

[3] Judgments of observer recognition were made by the author. Additional experiments conducted later (see Chapter 3.2.3) revealed little discrepancy between the observer recognition

all nine images, all observers identified the image content before viewing either the images from the reference/J2K or reference/J2K+DCQ sequences with the largest bitrate. However, several observers did not recognize the image content for three of the reference/line drawings sequences.

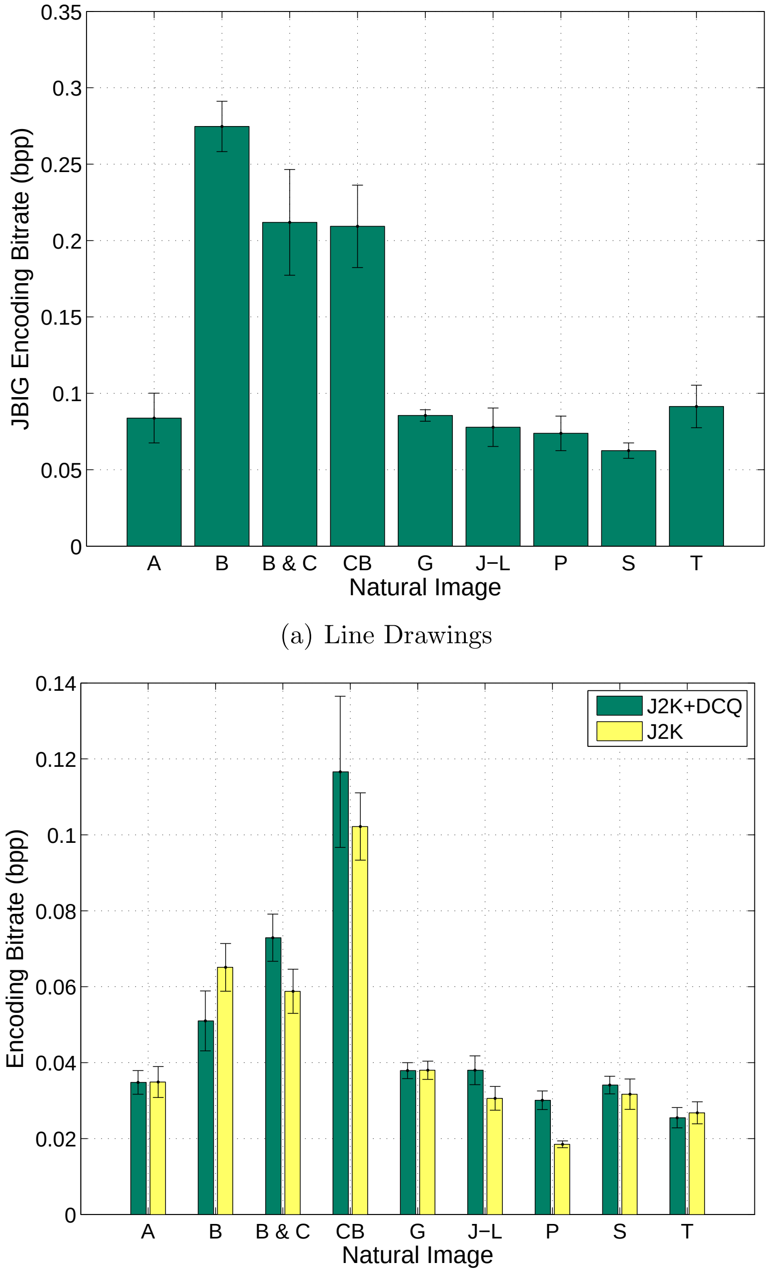
This section presents the results in two parts. The first part analyzes the raw mean recognition bitrates for both the line drawing, J2K, and J2K+DCQ distortions. The second part examines the raw mean recognition bitrates normalized by the image's visually lossless bitrate for the J2K and J2K+DCQ distortions.

### A.3.1 Raw Mean Recognition Bitrate

The average bitrate corresponding to the initial point of recognition for each natural image was noted for the line drawing, J2K, and J2K+DCQ distortions, and the standard deviation of the initial recognition bitrates was computed. Figure A.4 show the mean recognition bitrates for the nine images from Figures 3.1(a), 3.3, and 3.4 for all three distortions. Standard error bars have been included in both graphs.

Consider the mean recognition bitrates for the reference/line drawing sequences shown in Figure A.4(a). It is immediately obvious that the natural images *boy & cat*, *backhoe*, and *cagedbirds* have much larger mean recognition bitrates than those of the remaining six natural images. The required increase in bitrate reflects the nature of the content in the original natural scene. From Figures 3.1(a), 3.3, and 3.4 it is observed that the two images *boy & cat* and *cagedbirds* contain many more object boundaries than the other images. In addition, several artificial object

thresholds identified by the author and those identified by three other people.



(a) Line Drawings

(b) J2K+DCQ and J2K Distortions

Figure A.4: Mean recognition bitrates for line drawings and natural images with either baseline JPEG-2000 (J2K) distortions or distortion generated using JPEG-2000 with the dynamic contrast-based quantization strategy (J2K+DCQ). Standard error bars have been included in each graph. Acronyms based on the image descriptions identify the nine natural images: *airplane* (A), *backhoe* (B), *boy and cat* (B&C), *caged birds* (CB), *guitarist* (G), *jack-o-lanterns* (J-L), *pianist* (P), *skier* (S), and *train* (T).

boundaries appear in the structural representations for these images. For example, the stripes on the child's shirt and the pattern on the tablecloth for the image *boy & cat* lead to artificial edges corresponding to the contrast in these patterns. Likewise, the wires on the birdcage in the image *cagedbirds* occlude the birds and may distract the observer viewing the structural representation. In these two cases, the higher mean recognition bitrate is explained by the additional amount of information necessary to recognize and meaningfully represent the image content.

In the image *backhoe*, the gray levels between the foreground and background content are noticeably less distinct than for the other images and offers a different explanation for the difficulty in recognizing the content. A lower threshold for the Canny edge detector would be necessary to detect the boundary of the backhoe in the foreground. Using a lower threshold increases the number of edges detected, many of which do not belong to the primary content of the image. These additional and unnecessary edges provide additional, inaccurate, and often confusing information to the observer and make recognition much more difficult.

For the line drawing distortions, the standard error for these same three images is significantly larger than for the other images. A large standard error reflects the difficulty in recognizing the image content for the reference/line drawing sequences: different observers recognized the content at very different levels of distortion relative to the reference/line drawing sequences formed from the other natural images. In fact, these three images (*boy & cat*, *backhoe*, and *cagedbirds*) were not recognized by several of the observers who viewed the corresponding reference/line drawing sequence during the experiment. However, while not all of the observers recognized the content from the reference/line drawing sequences, the Canny edge detector successfully extracted critical information necessary for human observers to recognize the content of the remaining six natural images.

Consider the mean recognition bitrates for the J2K and J2K+DCQ distortions in Figure A.4(b). The same three natural images (*boy & cat*, *backhoe*, and *caged-birds*) had significantly larger mean recognition bitrates for both types distortions. In addition, the mean recognition bitrates for three images are also observed to have larger standard errors. This correspondence between the mean recognition bitrates and standard errors for the line drawing, J2K, and J2K+DCQ distortions suggests that a difficulty in content recognition for one representation predicts a similar difficulty for the other representation. In addition, the similarities among the mean recognition bitrates for the three distortions stresses the importance of object boundary information to observers for accurate recognition.

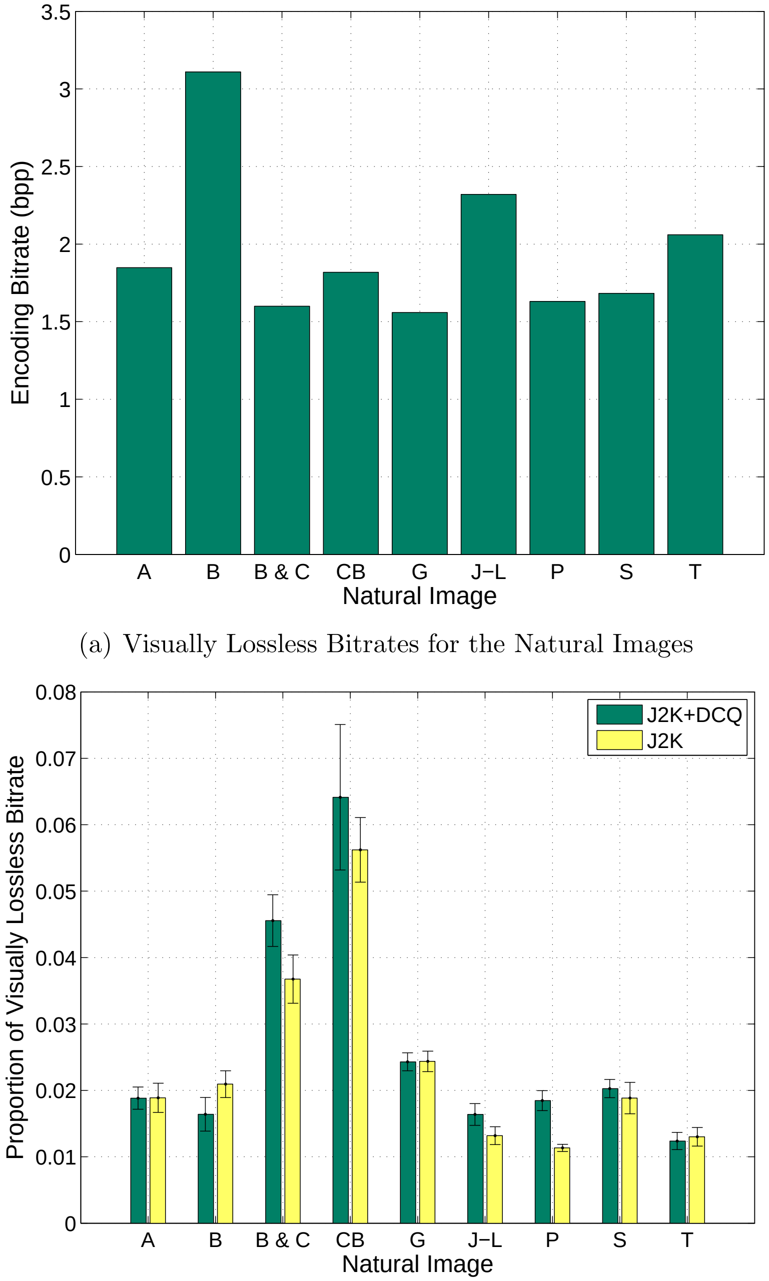
### A.3.2 Normalized Mean Recognition Bitrate

The content among the nine natural images varies, and comparing the raw mean recognition bitrates does not necessarily permit an adequate comparison. Apart from indicating the complexity of the image content in terms of both its ease of recognition and distortion, the raw mean recognition bitrate alone offers little evidence towards understanding what characteristics an image must possess to be recognizable.

The visually lossless bitrate $R_{VL}$ for an image provides a reference bitrate where recognition is unquestionable. Normalizing the mean recognition bitrate for the J2K and J2K+DCQ distortions by $R_{VL}$ specifies the mean recognition bitrate as a proportion of the visually lossless bitrate. This normalization reduces the variability among the recognition bitrates due to the differences in the original signal content of the natural image. A *visually lossless* image is visually indistinguishable from the original image. For line drawings, this definition of a visually lossless image is incompatible, since a reference line drawing is not readily available nor computable. The dynamic contrast quantization (DCQ) strategy specifies quantization step-sizes to produce a visually lossless image with (invisible) J2K+DCQ distortions.

The bitrates for the visually lossless J2K+DCQ distorted images are shown in Figure A.5(a). It is noted that the images *backhoe* and *jack-o-lanterns* have the highest visually lossless bitrates, which indicates that these images require more signal information to accurately represent the original image content than the other seven images. The normalized mean recognition bitrates are shown in Figure A.5(b) along with standard error bars. When comparing the normalized bitrates to the original bitrates shown in Figure A.4(b) it is immediately noted that only the images *boy & cat* and *cagedbirds* have significantly larger normalized mean recognition bitrates than the other seven images.

The recognition bitrate for image *backhoe* is approximately the same as those of the remaining five images upon applying the normalization. This is likely a consequence of the low contrast between the foreground and background gray levels in this image as noted in the discussion for the line drawings. Notice that the images *guitarist*, *jack-o-lanterns*, and *airplane* show similar contrasts between the gray levels for the foreground and background content. The normalized mean recognition bitrates for the remaining images are less than 0.025 of the visually lossless bitrate. This rudimentary ratio based on the visually lossless bitrate may provide a coarse indication of the recognition threshold for a given natural image.



(a) Visually Lossless Bitrates for the Natural Images

(b) Mean Recognition Bitrates Normalized by $R_{VL}$

Figure A.5: Visually lossless bitrates $R_{VL}$ determined using the dynamic contrast-based quantization strategy with JPEG-2000 for the nine natural images (Figure A.5(a)). Mean recognition bitrates normalized by the visually lossless bitrate $R_{VL}$ (Figure A.5(b)). Standard error bars have been included in each graph. See Figure A.4 for the images corresponding to the acronyms.

## A.4 Discussion

Several observers remarked on their difficulty in determining the content of the line drawings. A noted shortcoming for the reference/line drawing sequences using the Canny edge detector by varying $\sigma$ is apparent when examining Figure A.1. Without prior knowledge of the image content, the representations in Figures A.1(a) and A.1(b) do not provide adequate information to facilitate recognition of the content. Though varying $\sigma$ captures edges a different scales, larger values of $\sigma$ smooth the contours corresponding to object edges as observed in Figure A.1. Specifying $\sigma = 1$ captures the desired object boundaries in addition to other undesirable contours (e.g. the texture of the grass beneath the plane). Varying the threshold used by the Canny edge detector for fixed parameter $\sigma$ (e.g., $\sigma = 1$) could limit the amount of undesired contours that appear in the line drawing, but choosing the proper threshold is more of an art than a science.

The performance of the Canny edge detector has been shown to depend primarily on the post-processing after filtering with a derivative of a Gaussian filter [144]. The authors of that paper [144] report that modifications to the initial filter used for edge detection is likely to provide very minor improvements and recommend improvements to the post-processing procedures after filtering. New work in the field of edge detection has offered impressive improvements to the post-processing used after the initial filtering by the derivative of a Gaussian [83]. It is believed that such improvements will minimize the number of unintended edges detected and reduce the difficulty in recognizing the image content by generating better line drawings.

Alternatively, a new type of reference/distortion sequence that is more consistent with the J2K and J2K+DCQ distortions would permit a more elaborate comparison than the qualitative analysis performed using line drawings. A cartoon rendering of a natural image is a slightly more advanced version of a line drawing. The cartoon rendering is a piecewise constant representation, not a bi-level image, that omits the finer details while preserving the object boundaries of the original natural image. Such an image could be directly compared to either the J2K or J2K+DCQ distortions. The texture smoothing distortion types described in Chapter 3 were used in lieu of line drawings in subsequent experiments.

## A.5 Summary

This chapter summarized the results from an experiment investigating the use of visual structure in natural images for a recognition task. When comparing the results for the mean recognition bitrates, the same three images (*boy & cat*, *backhoe*, and *cagedbirds*) had significantly higher mean recognition bitrates for line drawings, J2K, and J2K+DCQ distortions than the other natural images. The qualitative similarity in the mean recognition thresholds for the different types of distortions indicates that the observers relied upon an image's visual structure to recognize the image content, since the line drawings are hypothesized to preserve the structure of a natural image.

APPENDIX B

# TRADEOFFS IN SUBJECTIVE TESTING METHODS FOR IMAGE AND VIDEO QUALITY ASSESSMENT

## B.1 Introduction

An objective quality estimator for either still images or video should accurately estimate the perceived quality scores of a collection of stimuli. New applications and processing techniques will introduce novel distortions that will need to be quantified in terms of perceived quality in order to confidently evaluate an objective quality estimator.

The subjective testing method used to obtain the perceived quality scores affects both the accuracy and the reliability of the data collected. Various testing methods have been used to obtain perceived quality scores for images and video sequences [30, 60, 99, 129, 140]. Among the variety of subjective methods, paired comparisons produce the most accurate and reliable estimates of perceived quality [60]. Using *paired comparisons*, an observer chooses among two stimuli the one having greater perceived quality. Raw observer responses to the paired comparison task are typically converted to scale values using either Thurstone's law of comparative judgment [126] or the Bradley-Terry model [4, 25]. Responses from many observers are needed to generate accurate scale values. Furthermore, for $N$ stimuli a total of $\frac{N(N-1)}{2}$ comparisons are necessary[3]. The large number of comparisons coupled with the need to collect responses from many observers discourage researchers from using paired comparisons to obtain perceived quality scores. Researchers generally choose alternative testing methods that provide perceived qual-

[3] Some comparisons may be omitted if the stimuli are easily distinguishable.

ity scores faster. Two common methods used to collect perceived quality scores are *absolute categorical rating* (ACR) [54] and the *subjective assessment methodology for video quality* (SAMVIQ) [35, 63].

An awareness of the potential benefits and/or costs attributed to the ACR and SAMVIQ test methods can guide researchers to choose the more suitable method for a particular application. For example, applications involving the use of high-definition displays to display stimuli would benefit from a testing method that allows observers to critically examine and reexamine stimuli (e.g., SAMVIQ), whereas stimuli for applications involving lower-end display devices, such as mobile phones, could suffice with less sophisticated testing methods (e.g., ACR). The ACR test method presents stimuli in a random order and uses a coarse resolution rating scale for evaluation. The SAMVIQ test method allows the observer to freely view several stimuli multiple times and uses a fine resolution rating scale for evaluation. Ease of implementation typically influences the adoption of ACR over SAMVIQ, since ACR accommodates more stimuli per testing session. This paper investigates the tradeoffs corresponding to the perceived quality scores obtained via these two subjective testing methods using three different subjective databases. The subjective databases contain either still-images or video sequences and perceived quality scores that have been obtained using both the ACR and SAMVIQ test methods.

This paper has the following organization: Section B.2 summarizes the two test methods compared in this paper, ACR and SAMVIQ. Section B.3 summarizes the content of the three subjective databases used to evaluate the two test methods. An analysis of the ACR and SAMVIQ test methods is presented in Section B.4. Section B.5 discusses a potential strategy to improve the SAMVIQ protocol based on observations from the analysis described in Section B.4. Section B.6 concludes this paper.

## B.2 Testing Methods: ACR and SAMVIQ

This section summarizes the two test methods investigated in this paper: absolute category rating (ACR) and the subjective assessment methodology for video quality (SAMVIQ).

### B.2.1 Absolute Category Rating (ACR) Method

The absolute category rating (ACR) [54] testing method[4] consists of consecutive trials in which an observer views and scores a stimulus. Specifically, in each trial, observers are presented with a stimulus[3], and then the display is set to a constant gray background and observers are immediately requested to provide an opinion score of the viewed stimulus. The reference stimuli are included in the test stimuli evaluated by the observer, and the observer is unaware if a stimulus is a processed or reference stimulus. The order of the stimuli presented is random and varies for each observer.

With ACR, a discrete category rating scale is used that has either five (ACR-5) or eleven (ACR-11) levels. The quality assessment task uses rating categories differentiated by the adjectives "Bad," "Poor," "Fair," "Good," and "Excellent." For ACR-5, the five levels correspond to the adjectives and are mapped to the integers on the range 1 to 5. With ACR-11, the eleven levels correspond to integers

[4] Numerical category scaling [30], adjective category scale [99], and categorical sort [60] are alternative names describing the absolute category rating (ACR) test method.

[3] Still images are displayed for 10 seconds. Video sequences are played once.

on the range 0 to 10, where "Bad" corresponds to 1, "Poor" corresponds to 3, "Fair" corresponds to 5, "Good" corresponds to 7, and "Excellent" corresponds to 9. A mean opinion score (MOS) is computed for each stimulus by averaging the corresponding observer opinion scores.

The ACR method is simple to implement and accommodates the evaluation of many stimuli in a single testing session. However, many observers are needed to minimize the influence of *contextual effects* due to presentation order in the collected opinion scores [29]. *Contextual affects* emerge when preceding stimuli influence an observer's interpretation of the quality scale, and such affects often lead to judgments of perceived quality *relative* to recently viewed stimuli [29, 85].

### B.2.2 Subjective Assessment for Video Quality (SAMVIQ) Method

The subjective assessment methodology for video quality (SAMVIQ) [35, 63] method consists of consecutive trials in which an observer freely views and scores a collection of test stimuli associated with an explicitly identified reference stimulus. Within a trial, an observer may view each stimulus multiple times and adjust an opinion score as needed. The explicit reference stimulus is available for comparison. The reference stimulus is also hidden among the test stimuli viewed by the observer. A trial is concluded by the observer only if he/she has supplied an opinion score for each stimuli, including the explicit reference stimulus.

A continuous rating scale is used with SAMVIQ. For the quality assessment task, the rating scale is categorized according to the adjectives "Bad," "Poor," "Fair," "Good," and "Excellent." The rating scale is mapped to the range 0 to

100, where “Bad” is aligned with a score of 10, “Poor” with a score of 30, “Fair” with a score of 50, “Good” with a score of 70, and “Excellent” with a score of 90. A mean opinion score (MOS) is computed for each stimulus by averaging the observer opinion scores.

The SAMVIQ method has a more complex implementation than ACR. The design of the evaluation trials restricts the number of stimuli that may be included in a single testing session. A single trial allows observers to view eleven stimuli, including both the explicit and hidden reference stimuli. Observers may view a stimulus multiple times in a trial, so fewer observers are typically needed to minimize contextual effects due to presentation order in the collected opinion scores. However, the opportunity to review a stimulus multiple times lengthens the duration of the testing session.

## B.3 Databases with Perceived Quality Scores for Visual Stimuli

This section summarizes the content of the three subjective databases used to compare the ACR and SAMVIQ test methods. Each database contains subjective scores corresponding to the quality assessment task collected according to both the ACR and SAMVIQ test methods. One database contains subjective scores for still images, and the other two databases contain subjective scores for video sequences.

### B.3.1 Still-Image Database

The Still-image database is a collection of processed images for which subjective quality scores have been recorded using both the ACR-5 and SAMVIQ [100] test methods. The database consists of five reference grayscale images and 90 processed images that were generated from the reference images. The processed images were generated from the reference images according to three different procedures: 1) JPEG-2000 compression driven by the dynamic contrast quantization algorithm [19] to choose subband quantization step-sizes for a desired bitrate, 2) a texture smoothing procedure that applies soft-thresholding for a desired threshold to undecimated Haar wavelet coefficients [103, 107, 119, 120], and 3) texture smoothing as in 2) with low-frequency signal information removed [103].

The ACR test method with five quality categories is used to collect subjective scores of the images from observers. Since there are a total of 95 images (90 test images plus 5 reference images) to be scored by the observers, a single test session lasting 25 minutes was used. Opinion scores were collected from 28 observers for all 95 images.

The SAMVIQ test method is implemented as described in Section B.2. To alleviate observer fatigue due to prolonged evaluation sessions, the processed images were partitioned into to two equally representative sets, creating two testing sessions each lasting 25 minutes. For each natural image, three anchor images, each associated with a different processing procedure and spanning the range of "quality," served as anchor images and appeared in both testing sessions to facilitate opinion score alignment across both sessions. Opinion scores were collected from 26 observers for each processed image (52 opinion scores were collected for the anchor images).

### B.3.2 Video 1 database

The Video 1 database is a collection of video sequences for which subjective quality scores have been recorded according to the ACR-5, ACR-11, and SAMVIQ test methods. This database consists of five reference video sequences recorded at a resolution of 1920 by 1080 pixels and at a frame rate of 50 progressive frames per second (1080p50). The database contains 145 processed video sequences generated from the reference video sequences to simulate 29 different broadcast applications for high-definition display devices. Perceived quality scores have been collected using both ACR-5 and ACR-11 for all video sequences in the database, but perceived quality scores were obtained using SAMVIQ for only 44 video sequences (40 processed video sequences plus 4 reference video sequences). The analysis in this paper considers only those 44 video sequences that have been evaluated using both testing methods.

Subjective scores for all five reference video sequences and 145 processed video sequences were obtained using the ACR-5 and ACR-11 test methods. Using the ACR-5 testing methodology, subjective scores were collected from 24 observers for all 150 video sequences. Using the ACR-11 test method, subjective scores were collected from 26 observers for all 150 video sequences. Perceived quality scores were collected in a single test session lasting 45 minutes using either ACR-5 or ACR-11.

Under the SAMVIQ test method, subjective scores were collected for four reference video sequences and 40 processed video sequences. Subjective scores were collected from 21 observers for the 44 video sequences. A single test session lasted 30 minutes.

### B.3.3 Video 2 database

The Video 2 database is a collection of video sequences for which subjective quality scores have been recorded according to the ACR-5 and SAMVIQ test methods [88]. This database consists of 24 reference video sequences recorded at a resolution of 1920 by 1080 pixels and at a frame rate of 50 interlaced frames per second (1080i50). The database contains 192 processed video sequences, which include the 24 reference video sequences, generated from the reference video sequences using an H.264 encoder operating at different encoding bitrates.

Subjective scores for all 192 processed video sequences were obtained using the ACR-5 test methods. Subjective scores were collected from 28 observers for all 192 video sequences. Two 30 minute test sessions, each containing 96 video sequences, were conducted to collect perceived quality scores.

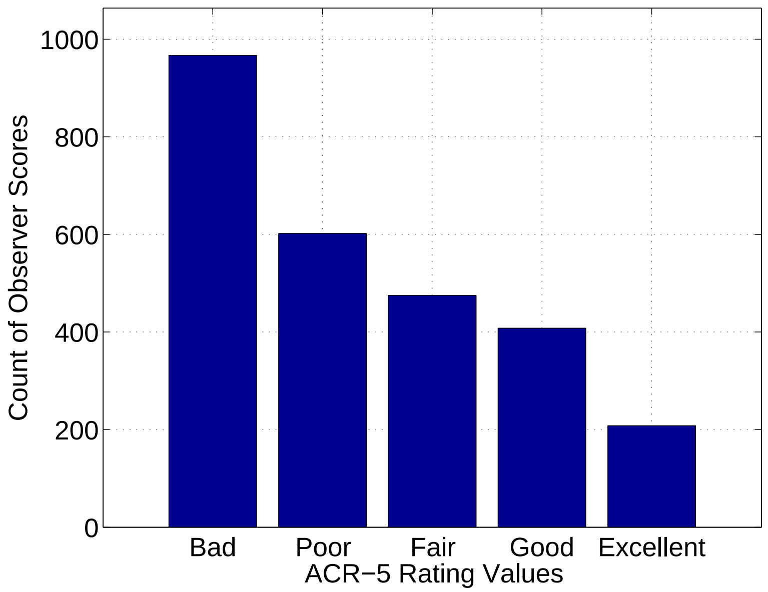
Under the SAMVIQ test method, subjective scores were collected for all 24 reference video sequences, explicitly identified as reference video sequences, and 192 processed video sequences, which include hidden reference video sequences. Only subjective scores corresponding to the set of 192 processed video sequences are used. To alleviate observer fatigue due to prolonged evaluation sessions, the processed video sequences were partitioned into to six sets, creating six testing sessions each lasting 30 minutes. In each testing session, observers provide perceived quality scores to all the processed video sequences corresponding to four reference video sequences. Among the six testing sessions, different numbers of observers provided subjective scores. Subjective scores were obtained from at least 18 and up to 39 observers for each testing session.

## B.4 Analysis of the ACR and SAMVIQ Subjective Testing Methods

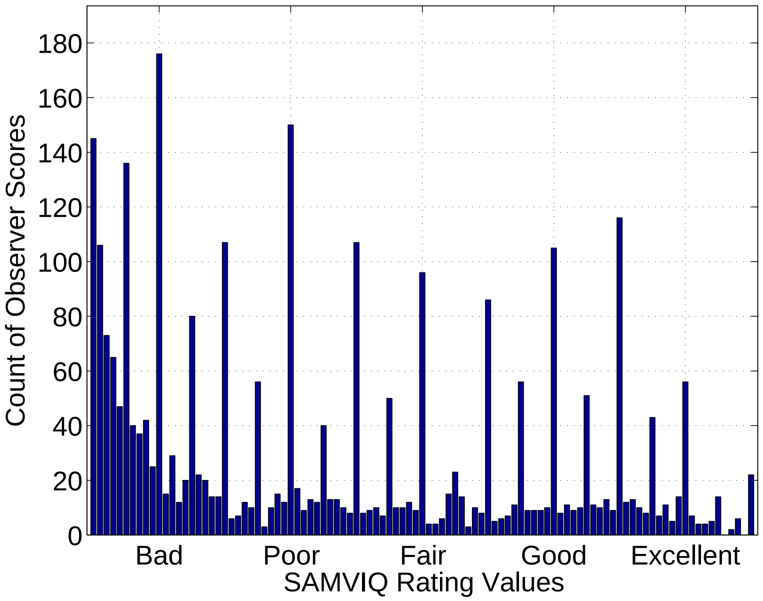
This section analyzes the subjective scores collected according to the ACR and SAMVIQ test methods for the three subjective databases described in Section B.3. This section first presents an individual analysis of the subjective scores obtained via the ACR and SAMVIQ test methods. The second part of this section compares the subjective scores obtained from the two test methods.

### B.4.1 Observers Use of Rating Scale for ACR and SAMVIQ

A distinction between the ACR and SAMVIQ test methods is the resolution of the rating scale used to score a stimuli. Despite the near continuous range of scores available to observers in the SAMVIQ test method, observers primarily submit opinion scores associated with the quality adjectives (i.e., "Bad," "Poor," "Fair," "Good," and "Excellent") and the mid-points between those adjectives. Histograms counting the frequency of the opinion scores provided by observers for each subjective database are shown in Figures B.1-B.3. For the SAMVIQ test method, it is evident from these histograms that observers quantize their scores to the the quality adjectives and the mid-points between those adjectives. Among the three databases, approximately 20% of the observer scores correspond directly to the quality adjectives, and approximately 43% of the observer scores correspond to either the quality adjectives or the mid-points between those adjectives. Specific values corresponding to the percentage of observer scores corresponding directly



(a) ACR-5



(b) SAMVIQ

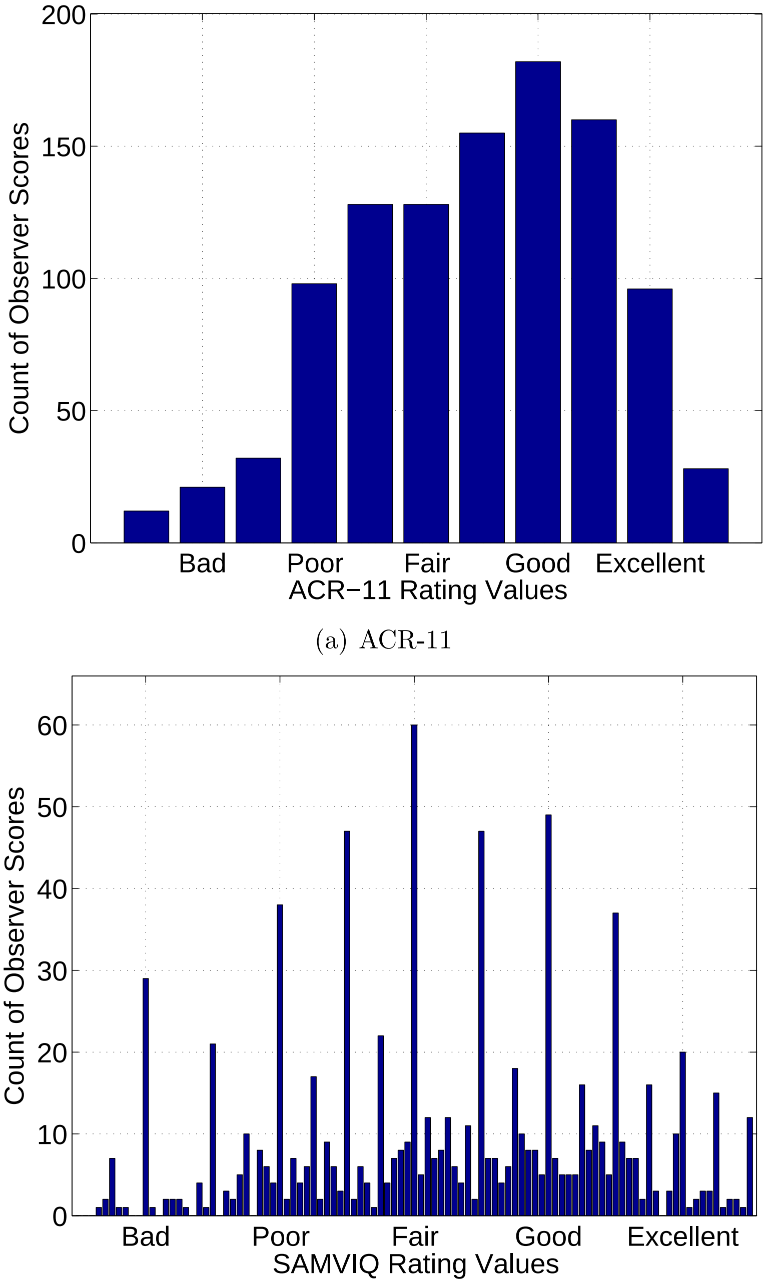
Figure B.1: Histogram of raw observer scores from the Still-Image database using ACR5 and SAMVIQ test methods. For SAMVIQ, 20.4% of the observer scores directly correspond to the category adjectives (i.e., “Bad”, “Poor”, etc.), and 40.8% of the observer scores correspond to both the category adjectives and mid-points between categories.

to the quality adjectives or to the mid-points between categories are listed in the captions of Figures B.1-B.3. In short, the fine resolution rating scale used in the SAMVIQ is superfluous, since observers typically provide scores that correspond to a coarser rating scale.

Observers frequently refrain from using the boundaries of the rating scale (i.e., scoring images as either "Bad" or "Excellent"). While the subjective scores collected for the Still-Image database appear to contradict this behavior (cf Figure B.1), this particular database contains a disproportionately large number of very low quality images, many of which deliberately contain very little recognizable content [100]. The histograms counting the frequency of opinion scores from the two video databases demonstrate that observers refrain from using the boundaries of the rating scale (cf Figures B.2 and B.3). Researchers [60] have suggested that observers "reserve" extreme ratings in test methods such as ACR in the event that stimuli viewed later in the test should demand perceived quality ratings of either "Bad" or 'Excellent." A remedy for ACR is to place a fixed set of stimuli at the beginning of the test to expose observers to the full range of "quality" to be evaluated. Responses from these initial trials are excluded from the analysis.

### B.4.2 Comparative Analysis of MOSs from ACR and SAMVIQ

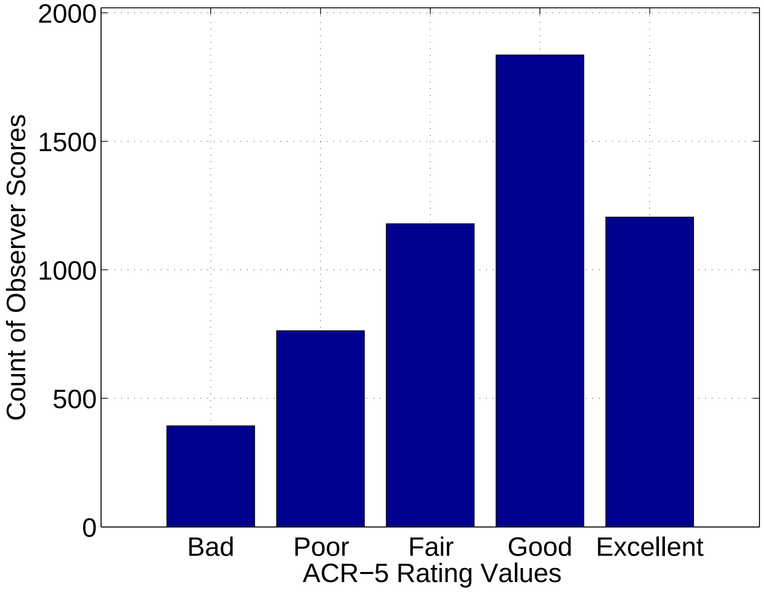
The SAMVIQ test method is resource demanding. In the three databases examined, A typical testing session duration is approximately 30 minutes and allows observers to evaluate at most 55 images (e.g. the Still-Image database) or 44 video sequences (e.g., the Video 1 database). Thus, evaluating many processed



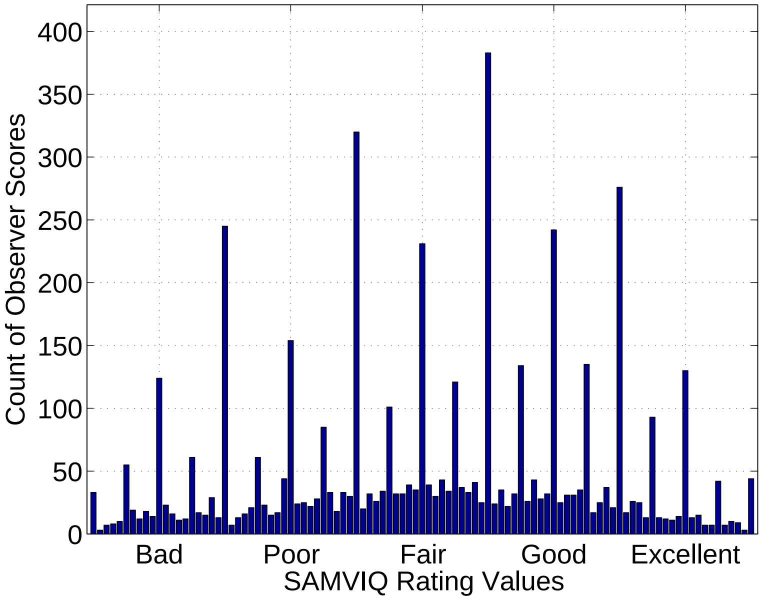
(a) ACR-11

(b) SAMVIQ

Figure B.2: Histogram of raw observer scores from the Video 1 database using ACR11 and SAMVIQ test methods. For ACR-11, 51.7% of the observer scores directly correspond to the category adjectives (i.e., "Bad", "Poor", etc.). For SAMVIQ, 23.3% of the observer scores directly correspond to the category adjectives, and 42.9% of the observer scores correspond to both the category adjectives and mid-points between categories.



(a) ACR-5



(b) SAMVIQ

Figure B.3: Histogram of raw observer scores from the Video 2 database using ACR5 and SAMVIQ test methods. For SAMVIQ, 18.1% of the observer scores directly correspond to the category adjectives (i.e., “Bad”, “Poor”, etc.), and 44.9% of the observer scores correspond to both the category adjectives and mid-points between categories.

stimuli for many different references often requires multiple testing sessions, especially when the stimuli are video sequences. Not only will this lengthen the time necessary to collect the desired subjective scores, but compensating observers for their participation in the experiment becomes expensive. The ACR test method accommodates more stimuli per testing session and may be more suitable given the available resources of the particular research group.

This section examines the tradeoffs between the ACR and SAMVIQ test methods. The first section compares the accuracy of the mean opinion scores (MOSs) collected under each test method based on the number of observers. The second section compares the ability of each test method to yield MOSs that distinguish pairs of stimuli generated from the same reference content within a subjective database. The last section examines the consistency of the MOSs between the two test methods.

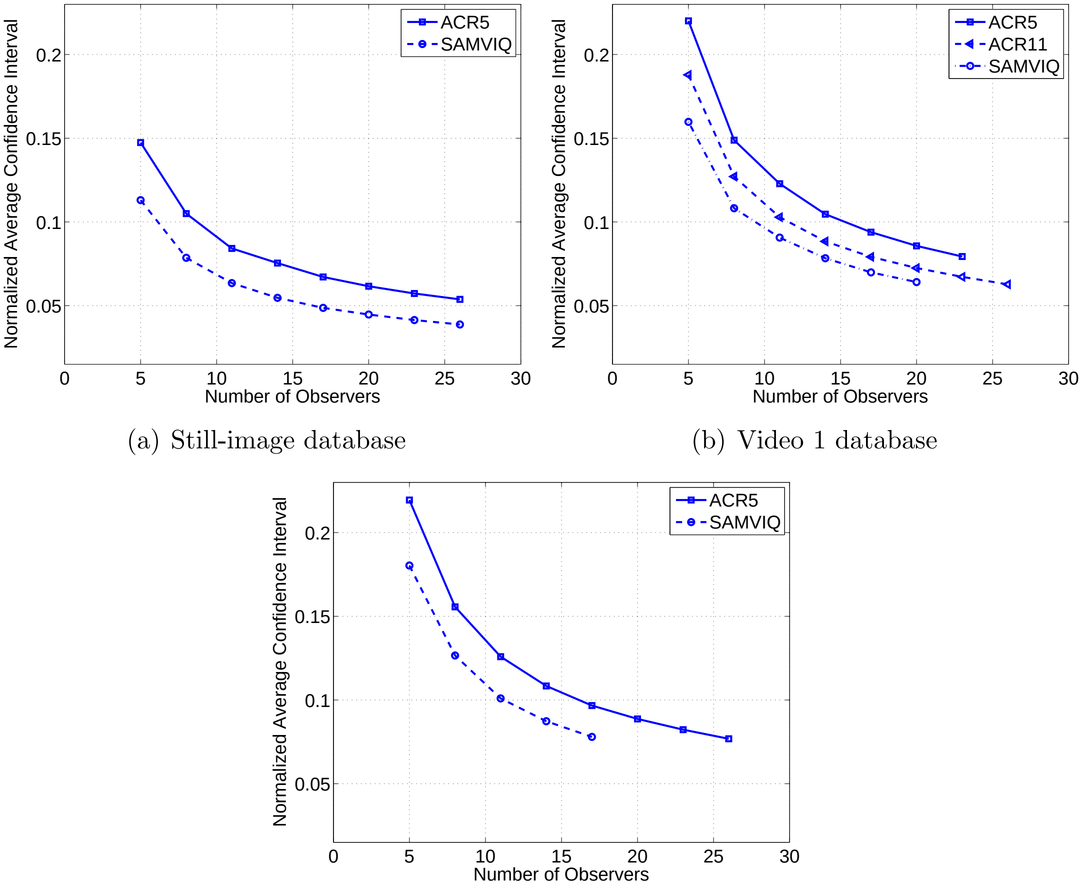
### Accuracy: SAMVIQ provides MOSs with smaller confidence intervals than ACR for the same number of observers

Among the two testing methods, SAMVIQ typically provides more accurate estimates of MOSs than ACR given the same number of observers. Figure B.4 shows the relationship between the average confidence interval of the MOS and the number of observers providing opinion scores for the different test methods for the three subjective databases. For each testing method, the MOS average confidence interval decreases as the number of observers increases, and for a fixed number of observers, SAMVIQ yields the smallest average confidence interval for the MOSs. Specifically, SAMVIQ requires between 2.8 and 11.5 fewer observers than ACR-5 for the Still-image database with an average of 5.5 (36%) fewer observers. For the Video 1 database, SAMVIQ requires between 2.4 and 9.3 fewer observers than ACR-5 and between 1 and 5.3 fewer observers than ACR-11 to achieve the same average confidence interval. On average, SAMVIQ requires 4.7 (34%) fewer observers than ACR-5 and 2.2 (17%) fewer observers than ACR-11 on the Video 1 database to achieve the same average confidence interval. For the Video 2 database, SAMVIQ requires between 1.7 and 8.4 fewer observers than ACR-5 with an average of 3.8 (27%) fewer observers to achieve the same average MOS confidence internal.

### Distinction: SAMVIQ MOSs differentiate stimuli more often than ACR MOSs

The ACR and SAMVIQ test methods use rating scales with different resolutions. As a result, MOSs for a pair of test stimuli could be statistically different using one test method but not another. This section compares the ability of the MOSs from the ACR and SAMVIQ test methods to differentiate test stimuli using the three databases. The analysis conducted in this section is similar to the resolving power for objective quality estimators proposed by Brill et al. [7].

The comparison of the ability of MOSs from two different test methods to differentiate test stimuli generated from the same reference stimulus for a particular database consists of two parts. First, for each pair of test stimuli associated with the same reference stimulus, a two-sample *t*-test with a 95% confidence level is used to determine if the corresponding MOSs, obtained using a specific test method, are statistically different. Second, for each pair of stimuli associated with the same reference stimulus, the conclusions from the *t*-test conducted for each test method are compared. Considering two test methods, “A” and “B,” there are



(a) Still-image database

(b) Video 1 database

(c) Video 2 database

Figure B.4: Relationship between the average confidence interval of the mean opinion scores and the number of observers used for different test methods for the three different databases. Increasing the number of observers for ACR generates more accurate mean opinion scores, competitive with those found with fewer observers using SAMVIQ.

four outcomes in this comparison for each pair of stimuli: 1) the MOSs for both methods are statistically different 2) the MOSs for both methods are statistically equivalent, 3) the MOSs for method "A" are statistically different but the MOSs for method "B" are statistically equivalent, and 4) the MOSs for method "B" are statistically different but the MOSs for method "A" are statistically equivalent. The outcomes corresponding to the first case will constitute the largest percentage of the comparisons, since a large proportion of the stimuli will likely yield very different levels of perceived quality.

MOSs obtained using the SAMVIQ test method regularly differentiate more test stimuli from the same reference stimulus than MOSs obtained using either ACR test method. Table B.1 summarizes the results when comparing the ability of MOSs from the ACR and SAMVIQ test methods to differentiate test stimuli in each database. For each database and a pair of test methods, the table lists the average percentage of pairs of test stimuli[3] corresponding to each of the four outcomes of the comparison described in the preceding paragraph. As an example, MOSs obtained using SAMVIQ differentiate 6.4% pairs of test stimuli in the Still-Image database that are not differentiated by MOSs obtained with ACR-5. In every database, MOSs obtained with SAMVIQ differentiate about 5 times the number of stimuli as either ACR-5 or ACR-11. For the Video 1 database, neither ACR testing method exhibits a significant advantage in terms of differentiating pairs of test stimuli.

### Consistency: Stimuli influence the consistency of MOSs from ACR and SAMVIQ

The collection of stimuli evaluated influence the consistency of MOSs collected via ACR with respect to those collected via SAMVIQ. ACR allows each observer only one opportunity to view and score a stimulus in a database. SAMVIQ allows each observer several opportunities to view and revise her score of a stimulus in a database. Consequently, if stimuli are easily discriminated by observers, then both ACR and SAMVIQ will produce consistent MOSs for the stimuli. However,

[3] Only pairs of test stimuli generated from the same reference stimulus are considered.

Table B.1: Results when comparing the ability of mean opinion scores (MOSs) from different test methods to differentiate test stimuli. For each database and a pair of test methods, the table lists the percentage of pairs of test stimuli corresponding to each of the four outcomes of a comparison test described in Section B.4.2. Considering two test methods, "A" and "B," there are four outcomes in this comparison for each pair of stimuli: 1) the MOSs for both methods are statistically different 2) the MOSs for both methods are statistically equivalent, 3) the MOSs for method "A" are statistically different but the MOSs for method "B" are statistically equivalent, and 4) the MOSs for method "B" are statistically different but the MOSs for method "A" are statistically equivalent. As an example, MOSs obtained using SAMVIQ differentiate 6.4% pairs of test stimuli in the Still-Image database that are not differentiated by mean opinion scores obtained with ACR-5. The results illustrate that MOSs obtained with SAMVIQ better differentiate stimuli that those obtained with ACR.

| | *Subjective Database* | | | | |
|---|---|---|---|---|---|
| | Still-Image | Video 1 | Video 1 | Video 1 | Video 2 |
| *Test Method A* | ACR-5 | ACR-5 | ACR-5 | ACR-11 | ACR-5 |
| *Test Method B* | SAMVIQ | ACR-11 | SAMVIQ | SAMVIQ | SAMVIQ |
| *1) Stat. diff. MOSs for both A and B* | 82.6% | 45.0% | 50.0% | 50.0% | 63.4% |
| *2) Stat. eq. MOSs for both A and B* | 6.9% | 34.4% | 31.1% | 31.7% | 10.7% |
| *3) Stat. diff. MOSs for only Method A* | 1.8% | 11.1% | 0% | 2.2% | 3.6% |
| *4) Stat. diff. MOSs for only Method B* | 6.4% | 11.1% | 15.6% | 17.8% | 14.3% |

if stimuli are more difficult to discriminate in terms of perceived quality, ACR, which only allows an observer one opportunity to view and score a stimulus, is not expected to produce MOSs that reflect differences among stimuli as well as SAMVIQ. This section investigates the consistency of the mean opinion scores obtained via the two test methods.

Since the different methods use rating scales with different ranges, a mapping between the rating scales facilitates a thorough comparison. Linear mappings between the rating scales are used that directly map the quality adjectives from one rating scale to another [88]. After linear mapping, the Pearson linear correlation [118] ($R$), the Spearman rank-order correlation [118] (ROCC), the Kendall $\tau$ rank correlation coefficient [61] ($\tau$), and the root mean squared error (RMSE) are computed using the MOSs from different testing methods for each database. Both of the rank correlation measures, the Spearman ROCC and Kendall's $\tau$, quantify the discrepancies between the rank order of two sets of MOSs; however, the Spearman ROCC accounts for the magnitude of the rank discrepancy, whereas Kendall's $\tau$ does not. The Spearman ROCC is the Pearson's linear correlation of the ranks assigned to the two sets of MOSs. The RMSE is normalized to the reflect errors in terms of the proportion of the target scale range.

For the Still-Image and Video 1 databases, the MOSs found using ACR are generally consistent with those found using SAMVIQ, indicating that the stimuli in each database are easy to discriminate in terms of perceived quality. Table B.2 summarizes the statistical analysis between MOSs obtained from different test methods after linear mapping, and the scatterplots in Figure B.5 illustrate the relationship between the MOSs found using ACR-5 and the MOSs found using SAMVIQ for the different databases. The Pearson linear correlation coefficients

Table B.2: Summary of statistical comparisons of mean opinion scores (MOSs) between different test methods for each subjective database. The Pearson linear correlation ($R$), the Spearman rank-order correlation (ROCC), the Kendall $\tau$ rank correlation coefficient ($\tau$), the root mean squared error (RMSE), and the percentage of consistent categories based on the MOSs of two test methods are computed. The RMSE is normalized to the reflect errors in terms of the proportion of the target scale range. MOSs obtained with ACR are consistent with those obtained with SAMVIQ for the Still-Image and Video 1 databases. The stimuli in the Video 2 are more difficult to distinguish in terms of perceived quality, hence the lower consistency between MOSs obtained using the two test methods (cf Figure B.5).

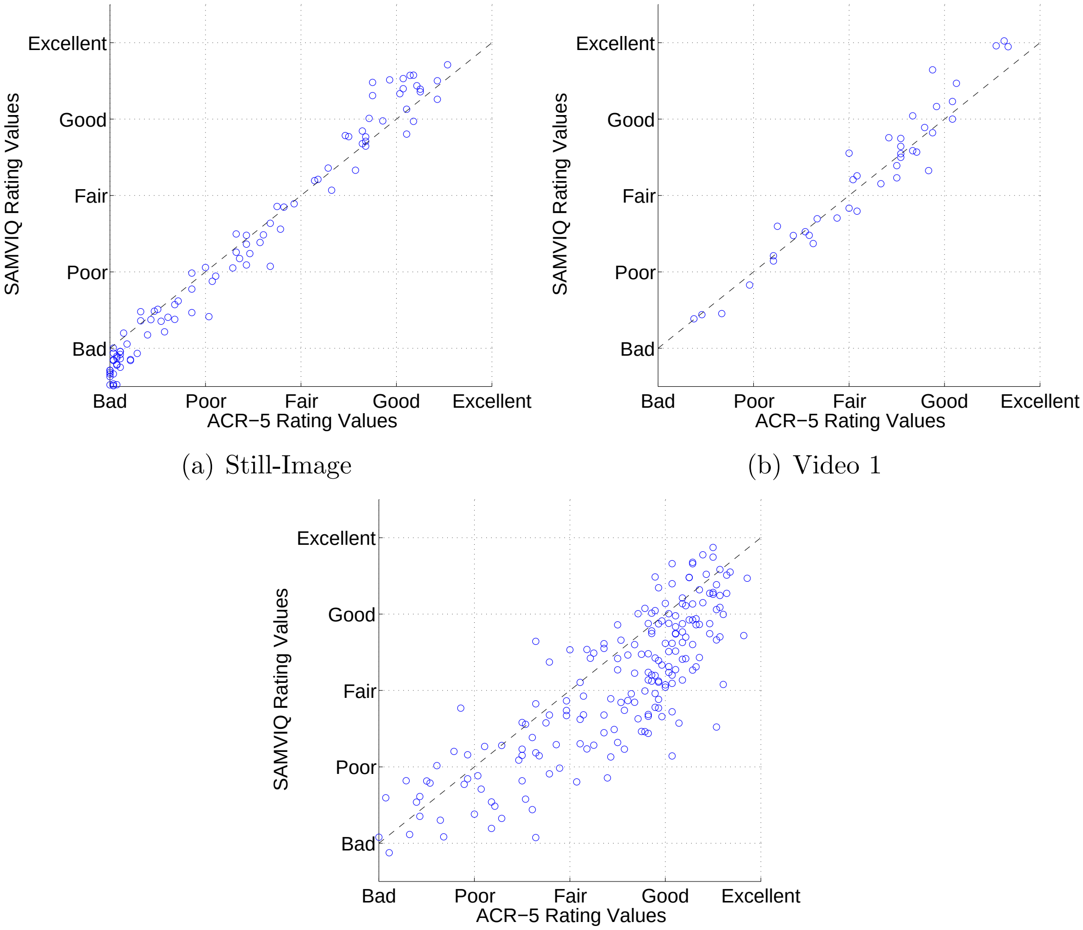
| *Database* | *Comparison* | $R$ | ROCC | $\tau$ | RMSE |
|---|---|---|---|---|---|
| Still-Image | ACR-5 to SAMVIQ | 0.989 | 0.978 | 0.883 | 0.0560 |
| Video 1 | ACR-5 to ACR-11 | 0.959 | 0.933 | 0.806 | 0.0686 |
| Video 1 | ACR-5 to SAMVIQ | 0.967 | 0.954 | 0.844 | 0.0521 |
| Video 1 | ACR-11 to SAMVIQ | 0.974 | 0.972 | 0.875 | 0.0626 |
| Video 2 | ACR-5 to SAMVIQ | 0.825 | 0.809 | 0.622 | 0.137 |

$R$ and Spearman rank-order correlation coefficients between MOSs generated via different testing methods for the Still-Image and Video 1 databases exceed 0.93, which indicates strong linear correlation and rank-order correlation. For these two databases, the values of Kendall's $\tau$ are lower than those of the Spearman ROCC. The values of Kendall's $\tau$ indicate that more than 80% of the MOSs are in rank agreement, and the high values of the Spearman ROCC indicate that the differences in the rank values of the MOSs that are not in perfect rank agreement are small. In addition to high correlation, estimating the MOSs obtained using SAMVIQ with the MOSs obtained using either ACR testing method yield RMSEs corresponding to an average error of less than 6.5% of the SAMVIQ rating scale range.

Unlike the Still-Image and Video 1 databases, the MOSs found using ACR-5 exhibit less consistency with those found using SAMVIQ for the Video 2 database, indicating that the stimuli in the Video 2 database are difficult to discriminate in terms of perceived quality. From Table B.2, all three correlation measures are markedly smaller for the Video 2 database than those for either of the other two databases. In addition, estimating the MOSs obtained using SAMVIQ with the MOSs obtained using ACR-5 yield RMSEs corresponding to an average error of 13.7% of the SAMVIQ rating scale range. Figure B.5(c) illustrates the nature of the notably decreased consistency between ACR and SAMVIQ for the Video 2 database. The scatterplot in Figure B.5(c) shows the relationship between the MOSs found using ACR-5 and the MOSs found using SAMVIQ for the Video 2 database. In Figure B.5(c), the MOSs associated with the "Good" quality adjective for ACR-5 spread a greater range of quality categories when using SAMVIQ. This indicates that MOSs obtained with ACR fail to capture subtle differences in perceived quality of stimuli, since MOSs obtained with SAMVIQ better distinguish stimuli rated as having "Good" quality under ACR. There are inconsistencies between the MOSs corresponding to the "Bad", "Poor", and "Fair" quality categories, and Section B.5 presents a discussion regarding the nature of these inconsistencies.

## B.5 Discussion: ACR or SAMVIQ?

Section B.4 compares mean opinion scores (MOSs) obtain using either the absolute category rating (ACR) testing method or the subjective assessment methodology for video quality (SAMVIQ). Despite the evidence promoting the use of SAMVIQ to obtain perceived quality scores for a collection of stimuli, ACR yields adequate



(a) Still-Image

(b) Video 1

(c) Video 2

Figure B.5: Stimuli influence the consistency of mean opinion scores (MOSs) obtained using ACR and SAMVIQ. Each scatterplot illustrates the relationship between MOSs of stimuli for a particular database obtained by using ACR-5 and SAMVIQ. The consistency of the MOSs obtained using ACR and SAMVIQ for the Still-Image and Video 1 databases indicate that the stimuli evaluated were relatively easy to discriminate in terms of perceived quality. The MOSs obtained using ACR and SAMVIQ for the Video 2 demonstrate that the stimuli were more difficult to discriminate in terms of perceived quality.

perceived quality scores for collections of stimuli evaluated in this paper (e.g., stimuli from the Still-Image database and the Video 1 database). Perhaps most interesting are the inconsistencies between the MOSs for ACR-5 and SAMVIQ corresponding to lower quality categories for the Video 2 database. This section briefly discusses the possible causes of these discrepancies.

For the Video 2 database (cf. Figure B.5(c)), MOSs corresponding to lower quality adjectives under SAMVIQ spread a greater range of quality categories under ACR-5. In other words, the MOSs obtained with ACR reflect larger differences in perceived quality among stimuli rated as either having “Bad” or “Poor” perceived quality using SAMVIQ. Two possible explanations are provided for this discrepancy: 1) a limitation with SAMVIQ and 2) the suitability of the testing method used for the application and stimuli studied.

The discrepancy between the ACR and SAMVIQ MOSs for the Video 2 database could be a consequence of the protocol for SAMVIQ: observers rate a collection of stimuli generated from the same reference stimulus in a trial. Thus, SAMVIQ encourages observers to develop separate quality scales for stimuli generated from a particular reference content. As a result, the perceived quality scores for stimuli generated from different reference content may be misaligned[1]. It has been suggest that under ACR observers develop different internal scales for stimuli generated from the same reference content [129], yet SAMVIQ does nothing to discourage this behavior. Stimuli have been evaluated following protocols that include trials that force observers to compare stimuli generated from different reference content [21, 64]. These additional trials serve to align perceived quality

[1] Suggesting that the perceived quality scores are misaligned implies that stimuli generated from different reference stimuli should be assigned a perceived quality corresponding to a global perceived quality score (as methods that measure *just-noticeable differences* (JNDs) afford [60, 140]) rather than a perceived quality scale associated with stimuli generated from a single reference content.

scores corresponding to stimuli generated from different reference content. A thorough analysis comparing the effects of these additional "alignment" trials using SAMVIQ needs to be performed.

Alternatively, the discrepancy between the ACR and SAMVIQ MOSs for the Video 2 database could be attributed to the suitability of the testing method used for the application and stimuli studied. The ACR testing method is best suited for applications with stimuli where an observer views a stimulus only once, whereas SAMVIQ is best suited for applications with stimuli where an observer may view a stimulus multiple times. The stimuli in the Video 2 database contain various distortions due to encoders operating at various bitrates, whereas the Video 1 database contains distortions due to different antialiasing filters using a broadcast application. The distortions in the Video 1 database are spatially uniform, so viewing a stimulus at any spatial location as the video is played will generally give an observer an impression of the distortions. Thus, under both the ACR and SAMVIQ testing methods, observers will very similar conclusions about the perceived quality of the video. In contrast, the distortions in the Video 2 database are spatially distributed, and all observers may not see the distortions in the video when viewing the video. The opportunity to view the video multiple times, as in SAMVIQ, increases the probability that an observer will see the distortion. Thus, the wide range of MOSs for ACR when the MOSs for SAMVIQ are in the "Bad" to "Poor" range would be explained by observers "missing" the distortions in their single view with ACR but not with SAMVIQ.

The second explanation of the discrepancy between the ACR and SAMVIQ MOSs emphasizes the need to pair the testing method with the application studied. ACR is suited to applications where observers cursorily view the stimuli, whereas

SAMVIQ is suited to applications where observers critically view stimuli.

## B.6 Summary

This paper investigated the tradeoffs of the absolute category rating (ACR) and subjective assessment for video quality (SAMVIQ) methods methods using three different subjective quality databases that have scores corresponding to each method. The subjective databases contained either still-images or video sequences.

Results are that 1) the fine resolution rating scale used by SAMVIQ is superfluous, 2) SAMVIQ scores have greater accuracy than ACR scores for the same number of observers (on average 30% fewer observers were required for SAMVIQ than ACR for the same level of accuracy), 3) SAMVIQ scores better differentiate stimuli than ACR scores, and 4) the consistency of categorical ratings between ACR and SAMVIQ is lower for databases when stimuli are more difficult to distinguish in terms of perceived quality. Increasing the number of observers for ACR generates more accurate scores, competitive with the accuracy found with fewer observers using SAMVIQ.

This analysis discloses areas of future work. First, the current analysis considers all of the subjective data in a given database. The effects of outlier rejection and deliberately reducing the number of observers could alter the conclusions drawn about the relationship between the ACR and SAMVIQ test methods. Second, the three subjective databases correspond to specific applications, and a closer analysis of suitability of different test methods should be considered [9,53]. The suitability of perceived quality scores obtained by ACR and SAMVIQ for specific applications is currently being investigated.

# BIBLIOGRAPHY

[1] İsmail Avcıbaş, Bülent Sankur, and Khalid Sayood. Statistical evaluation of image quality measures. *Journal of Electronic Imaging*, 11(2):206–233, April 2002.

[2] Lucien M. Biberman, editor. *Perception of Displayed Information.* Plenum Press, New York, NY, 1973.

[3] Irving Biderman and Ginny Ju. Surface versus edge-based determinants of visual recognition. *Cognitive Psychology*, 20:38–64, 1988.

[4] R. A. Bradley and M. E. Terry. The rank analysis of incomplete block designs I: The method of paired comparisons. *Biometrika*, 39:324–345, 1952.

[5] Nuala Brady and David J. Field. What's constant in contrast constancy? The effects of scaling on the perceived contrast of bandpass patterns. *Vision Research*, 35(6):739–756, 1995.

[6] D. H. Brainard. The psychophysics toolbox. *Spatial Vision*, 10:443–446, 1997.

[7] Michael H. Brill, Jeffrey Lubin, Pierre Costa, Stephen Wolf, and John Pearson. Accuracy and cross-calibration of video quality metrics: new methods from ATIS/T1A1. *Signal Processing: Image Communication*, 19:101–107, 2004.

[8] Alan C. Brooks and Thrasyvoulos N. Pappas. Structural similarity quality metrics in a coding context: Exploring the space of realistic distortions. In B. E. Rogowitz, T. N. Pappas, and S. J. Daly, editors, *Proc. SPIE: HVEI XI*, San Jose, CA, January 2006.

[9] Matthew D. Brotherton, Quan Huynh-Thu, David S. Hands, and Kjell Brunnström. Subjective multimedia quality assessment. *IEICE Trans. Fundam. Electron. Commun. Comput. Sci.*, E89-A(11):2920–2932, 2006.

[10] Morton B. Brown and Alan B. Forsythe. Robust tests for the equality of variances. *J. Amer. Stat. Assoc.*, 69(346):364–367, 1974.

[11] Jerome S. Bruner and Mary C. Potter. Interference in visual recognition. *Science*, 144:424–425, 1964.

[12] A. Mike Burton, Stephen Wilson, Michelle Cowan, and Vicki Bruce. Face recognition in poor-quality video. *Psychological Science*, 10(3):243–248, May 1999.

[13] John Canny. A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 8(6):679–698, November 1986.

[14] Mathieu Carnec, Patrick Le Callet, and Dominique Barba. Objective quality assessment of color images based on a generic perceptual reduced reference. *Signal Processing: Image Communication*, 2008. doi:10.1016/j.image.2008.02.003.

[15] Damon Chandler. Categorical subjective image quality (CSIQ) database. http://vision.okstate.edu/index.php?loc=csiq.

[16] Damon C. Chandler and Sheila S. Hemami. VSNR online supplement. http://foulard.ece.cornell.edu/dmc27/vsnr/vsnr.html.

[17] Damon M. Chandler and Sheila S. Hemami. Effects of natural images on the detectability of simple and compound wavelet subband quantization distortions. *J. Optical Soc. of Am. A*, 20(7), July 2003.

[18] Damon M. Chandler and Sheila S. Hemami. Suprathreshold image compression based on contrast allocation and global precedence. In B. E. Rogowitz, T. N. Pappas, and S. J. Daly, editors, *Proc. SPIE: HVEI VIII*, San Jose, CA, 2003.

[19] Damon M. Chandler and Sheila S. Hemami. Dynamic contrast-based quantization for lossy wavelet image compression. *IEEE Trans. Image Process.*, 14(4):397–410, April 2005.

[20] Damon M. Chandler and Sheila S. Hemami. VSNR: A wavelet-based visual signal-to-noise ratio for natural images. *IEEE Trans. Image Process.*, 16(9):2284–2298, September 2007.

[21] Damon M. Chandler, Kenny H. Lim, and Sheila S. Hemami. Effects of spatial correlations and global precedence on the visual fidelity of distorted images. In B. E. Rogowitz, T. N. Pappas, and S. J. Daly, editors, *Proc. SPIE: HVEI XI*, San Jose, CA, 2006.

[22] Charles A. Collin. Spatial-frequency thresholds for object categorisation at basic and subordinate levels. *Perception*, 35:41–52, 2006.

[23] Charles A. Collin and Patricia A. McMullen. Subordinate-level categorization relies on high spatial frequencies to a greater degree than basic-level categorization. *Perception and Psychophysics*, 67(2):354–364, 2005.

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

[25] Douglas E. Critchlow and Michael A. Fligner. Paired comparisons, triple comparisons, and ranking experiments as generalized linear models, and their implementation on GLIM. *Psychometrika*, 56(3):517–533, 1991.

[26] Scott J. Daly. The visible difference predictor: An algorithm for the assessment of image fidelity. In Andrew B. Watson, editor, *Digital Images and Human Vision*, chapter 14, pages 179–206. MIT Press, Cambridge, MA, 1993.

[27] Niranjan Damera-Venkata, Thomas D. Kite, Wilson S. Geisler, Brian L. Evans, and Alan C. Bovik. Image quality assessment based on a degradation model. *IEEE Trans. Image Process.*, 9(4):636–650, April 2000.

[28] Josh P. Davis and Tim Valentine. CCTV on trial: Matching video images with the defendant in the dock. *Applied Cognitive Psychology*, 23:482–505, 2009.

[29] Huib de Ridder. Current issues and new techniques in visual quality assessment. In *Proc. IEEE Intl. Conf. on Image Process.*, Lausanne, Switerzland., 1996.

[30] Huib de Ridder and Gijberta M. M. Majoor. Numerical category scaling: an efficient method for assessing digital image coding impairments. In Bernice E. Rogowitz and Jan P. Allebach, editors, *Proc. SPIE: Human Vision and Electronic Imaging: Models, Methods, and Applications*, volume 1249, Santa Clara, CA, January 1990.

[31] Russell L. De Valois and Karen K. De Valois. *Spatial Vision*. Oxford University Press, New York, 1990.

[32] J. L. Devore. *Probability and Statistics for Engineering and the Sciences*. Duxbury, fifth edition, 2000.

[33] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. Wiley, New York, second edition, 2001.

[34] Serge O. Dumoulin, Steven C. Dakin, and Robert F. Hess. Sparsely distributed contours dominate extra-striate responses to complex scenes. *NeuroImage*, pages 890–901, 2008.

[35] European Broadcast Union (EBU). SAMVIQ - subjective assessment methodology for video quality. Technical Report BPN 056, May 2003.

[36] Ahmet M. Eskicioglu and Paul S. Fisher. Image quality measures and their performance. *IEEE Trans. Commun.*, 43(12):2959–2965, 1995.

[37] H. Everett. Generalized lagrange multiplier method for solving problems of optimum allocation of resources. *Operations Research*, 11:399–417, 1963.

[38] Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27:861–874, 2006.

[39] David J. Field. Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America A*, 4(12):2379–2394, 1987.

[40] David J. Field, Anthony Hayes, and Robert Hess. Contour integration by the human visual system: evidence for a local "Association Field". *Vision Research*, 33(2):173–193, 1993.

[41] E. C. Fieller, H. O. Hartley, and E. S. Pearson. Tests for rank correlation coefficients. I. *Biometrika*, 44(3/4):470–481, 1957.

[42] Institute for Telecommunication Sciences. In C. Ford, P. Raush, and K. Davis, editors, *Video Quality in Public Safety Conference*, Boulder, CO, February4–6, 2009.

[43] Carolyn G. Ford, Mark A. McFarland, and Irena W. Stange. Subjective video quality assessment methods for recognition tasks. In B. E. Rogowitz and T. N. Pappas, editors, *Proc. SPIE: HVEI XIV*, volume 7240, January 2009.

[44] M. D. Gaubatz, D. M. Rouse, and S. S. Hemami. MeTriX MuX. http://foulard.ece.cornell.edu/gaubatz/metrix_mux.

[45] M. A. Georgeson and G. D. Sullivan. Contrast constancy: Debluring in human vision by spatial frequency channels. *Journal of Physiology*, 252:627–656, 1975.

[46] C. Giardina and E. Dougherty. *Morphological Methods in Image and Signal Process.* Prentice Hall, 1998.

[47] Douglas Granrath. The role of human visual models in image processing. *Proc. IEEE*, 69(5):552–561, 1981.

[48] David M. Green and John A. Swets. *Signal Detection Theory and Psychophysics*. Peninsula Publishing, 1988.

[49] Kalanit Grill-Spector. The neural basis of object perception. *Current Opinion in Neurobiology*, 13:159–166, 2003.

[50] R. W. Hamming. Error detecting for error correcting codes. *The Bell Systems Technical Journal*, 29(2):147–160, 1950.

[51] James A. Hanley and Barbara J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143:29–36, April 1982.

[52] B. R. Hunt and G. F. Sera. Power-law stimulus-response models for measures of image quality in nonperformance environments. *IEEE Trans. on Systems, Man, and Cybernetics*, 8(11):781–791, 1978.

[53] Quan Huynh-Thu and Mohammed Ghanbari. A comparison of subjective video quality assessment methods for low-bit rate and low-resolution video. In *Signal and Image Processing: Proceedings of the Seventh IASTED International Conference*, 2005.

[54] International Telecommunication Union (ITU) Radiocommunication Assembly. Subjective video quality assessment methods for multimedia applications, 2008. ITU-U P.910.

[55] International Telecommunication Union (ITU) Radiocommunication Assembly. Subjective video quality assessment methods for recognition tasks, 2008. ITU-U P.912.

[56] John M. Irvine, Barbara A. Eckstein, Robert A. Hummel, Richard J. Peters, and Rhonda Ritzel. Evaluation of the tactical utility of compressed imagery. *Optical Engieering*, 41(6):1262–1273, June 2002.

[57] Carlos M. Jarque and Anil K. Bera. Efficient tests for normality, homoscedas-

ticity, and serial independence of regression residuals. *Economics Letters*, 6:255–259, 1980.

[58] John Johnson. Analysis of image forming systems. In *Image Intensifier Symposium*, Fort Belvoir, VA, October 1958.

[59] Vishwakumara Kayargadde and Jean-Bernard Martens. Perceptual characterization of images degraded by blur and noise: experiments. *Journal of the Optical Society of America A*, 13(6):1166–1177, 1996.

[60] Brian W. Keelan. *Handbook of Image Quality: Characterization and Prediciton.* CRC Press, 2002.

[61] M. Kendall. A new measure of rank correlation. *Biometrika*, 30:81–93, 1938.

[62] Scott Konishi, Alan L. Yuille, James M. Coughlan, and Song Chun Zhu. Statistical edge detection: Learning and evaluating edge cues. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(1):57–74, January 2003.

[63] Franc Kozamernik, Paola Sunna, Emmanuel Wyckens, and Dag Inge Pettersen. Subjective quality of internet video codecs: Phase II evaluations using SAMVIQ. EBU Technical Review, January 2005.

[64] Eric C. Larson and Damon M. Chandler. The most apparent distortion: A dual strategy for full reference image quality. In Susan P. Farnand and Frans Gaykema, editors, *Proc. SPIE: Image Quality and System Performance VI*, volume 7242, San Jose, CA, January 2009.

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

[66] G. Legge and J. Foley. Contrast masking in human vision. *J. Opt. Soc. Am.*, 70:1458–1470, 1980.

[67] Gunter Loffler. Perception of contours and shapes: low and intermediate stage mechanisms. *Vision Research*, 48:2160–2127, 2008.

[68] 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*, chapter 13, pages 163–178. MIT Press, Cambridge, MA, 1993.

[69] Wei-Ying Ma and B. S. Manjunath. Edgeflow: A technique for boundary detection and segmentation. *IEEE Trans. Image Process.*, 9(8):1375–1388, August 2000.

[70] Stephane Mallat and Sifen Zhong. Characterization of signals from multiscale edges. *IEEE Trans. Image Process.*, 14(7):710–732, 1992.

[71] James L. Mannos. The effects of a visual fidelity criterion on the encoding of images. *IEEE Trans. Inf. Theory*, 20(4):525–536, 1974.

[72] David Marr and E. Hildreth. Theory of edge detection. *Proceedings from the Royal Society of London B*, 207(1167):187–217, February 1980.

[73] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proc. 8th Int'l Conf. Computer Vision*, volume 2, pages 416–423, July 2001.

[74] Bernard Moulden, Fred Kingdom, and Linda F. Gatley. The standard deviation of luminance as a metric for contrast in random-dot images. *Perception*, 19:79–101, 1990.

[75] Thomas M. Murphy and Leif H. Finkel. Shape representation by a network of V4-like cells. *Neural Networks*, (20):851–867, 2007.

[76] David Navon. Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology*, 9:353–383, 1977.

[77] [online]. Consumer digital video library. http://www.cdvl.org/.

[78] [online]. flickr. http://www.flickr.com/.

[79] [online]. Independent JPEG Group (IJG). http://www.ijg.org/.

[80] Antonio Ortega and Kannan Ramchandran. Rate-distortion methods for image and video compression. *IEEE Signal Processing Magazine*, 15:23–50, November 1998.

[81] Patrick D. O'Shea, Eddie L. Jacobs, and Richard L. Espinola. Effects of image compression on sensor performance. *Optical Engieering*, 47(1), January 2008.

[82] Sven Panis, Joeri De Winter, Joachim Vandekerckhove, and Johan Wagemans. Identification of everyday objects on the basis of fragmented outline versions. *Perception*, 37:271–289, 2008.

[83] Giuseppe Papari, Patrizio Campisi, Nicolai Petkov, and Alessandro Neri. Contour detection by multiresolution surround inhibition. In *Proc. IEEE Int. Conf. on Image Process. (ICIP)*, Atlanta, GA, October 2006.

[84] Thrasyvoulos N. Pappas and Robert J. Safranek. Perceptual criteria for image quality evaluation. In Alan C. Bovik, editor, *Handbook of Image and Video Processing*. Academic, New York, 2000.

[85] Allen Parducci and Douglas H. Wedell. The category effect with rating scales: Number of categories, number of stimuli, and method of presentation. *Journal of Experimental Psychology: Human perception and performance*, 12(4):496–516, 1986.

[86] C. A. Párraga, T. Troscianko, and David J. Tolhurst. The effects of amplitude-spectrum statistics on foveal and peripheral discrimination of changes in natural images, and a multi-resolution model. *Vision Research*, 45:3145–3168, 2005.

[87] William A. Pearlman. A visual system model and a new distortion measure in the context of image processing. *J. Optical Soc. Am.*, 68(3):374–386, 1978.

[88] Stéphane Péchard and Romuald Pépion. Suitable methodology in subjective video quality assessment: A resolution dependent paradigm. In *Third International Workshop on Image Media Quality and its Applications*, Kyoto, Japan, September 2008.

[89] Eli Peli. Contrast in complex images. *J. Optical Soc. of Am. A*, 7:2032–2040, 1990.

[90] William B. Pennebaker and Joan L. Mitchell. *JPEG: Still Image Data Compression Standard*. Van Nostrand Reinhold, New York, 1993.

[91] Julie K. Petersen. *Understanding Surveillance Technologies*. CRC Press, 2001.

[92] Uri Polat and Dov Sagi. Lateral interactions between spatial channels: Suppression and facilitation revealed by lateral masking experiments. *Vision Research*, 33(7):993–999, 1993.

[93] Uri Polat and Dov Sagi. The architecture of perceptual spatial interactions. *Vision Research*, 34(1):73–78, 1994.

[94] C. Poynton. The rehabilitation of gamma. In B. E. Rogowitz and T. N. Pappas, editors, *Proc. SPIE: HVEI III*, San Jose, CA, 1998.

[95] William K. Pratt. *Digital Image Processing: PIKS Inside.* Wiley-Interscience, New York, third edition, 2001.

[96] Gabriel Prieto, M. Chevalier, and E. Guibelaide. Automatic scoring of cdmam using a model of the recognition threshold of the human visual system: R*. In *Proc. IEEE Int. Conf. on Image Process. (ICIP)*, Cairo, Egypt, November 2009.

[97] Eleanor Rosch, Carolyn Mervis, Wayne Gray, David Johnson, and Penny Boyes-Braem. Basic objects in natural categories. *Cognitive Psychology*, 8:382–439, 1976.

[98] F. A. Rosell and R. H. Willson. Recent psychophysical experiments and the display signal-to-noise ratio concept. In Lucien Biberman, editor, *Perception of Displayed Information*, chapter 5, pages 167–232. Plenum Press, New York, NY, 1973.

[99] Jacques A. J. Roufs. Perceptual image quality: concept and measurement. *Philips Journal of Research*, 47:35–62, 1992.

[100] David Rouse, Romuald Pepion, Sheila Hemami, and Patrick Le Callet. Image utility assessment and a relationship with image quality assessment. In B. E. Rogowitz and T. N. Pappas, editors, *Proc. SPIE: HVEI XIV*, volume 7240, San Jose, CA, January 2009.

[101] David Rouse, Romuald Pepion, Patrick Le Callet, and Sheila Hemami. Tradeoffs in subjective testing methods for image and video quality assessment. In B. E. Rogowitz and T. N. Pappas, editors, *Proc. SPIE: HVEI XV*, volume 7527, San Jose, CA, January 2010.

[102] David M. Rouse and Sheila S. Hemami. Quantifying the use of structure in cognitive tasks. In B. E. Rogowitz, T. N. Pappas, and S. J. Daly, editors, *Proc. SPIE: HVEI XII*, San Jose, CA, January 2007.

[103] David M. Rouse and Sheila S. Hemami. Analyzing the role of visual structure in the recognition of natural image content with multi-scale SSIM. In B. E.

Rogowitz and T. N. Pappas, editors, *Proc. SPIE: HVEI XIII*, volume 6806, San Jose, CA, January 2008.

[104] David M. Rouse and Sheila S. Hemami. Understanding and simplifying the structural similarity metric. In *Proc. IEEE Int. Conf. on Image Process. (ICIP)*, San Diego, CA, October 2008.

[105] David M. Rouse and Sheila S. Hemami. Natural image utility assessment using image contours. In *Proc. IEEE Int. Conf. on Image Process. (ICIP)*, Cairo, Egypt, November 2009.

[106] David M. Rouse, Yiran Wang, Fan Zhang, and Sheila S. Hemami. A novel technique to acquire perceived utility scores from textual descriptions of distorted natural images. In *Proc. IEEE Int. Conf. on Image Process. (ICIP)*, Hong Kong, September 2010.

[107] Leonid I. Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithm. *Physica D*, 60:259–268, 1992.

[108] Robert J. Safranek and James D. Johnston. A perceptually tuned sub-band image coder with image dependent quantization and post-quantization data compression. In *Proc. of the IEEE Intl. Conf. on Acoustics, Speech, and Sig. Processing*, volume 3, pages 1945–1948, Glasgow, Scotland, May 1989.

[109] Sadiq Sait and Habib Youssef. *Iterative Computer Algorithms with Applications in Engineering.* IEEE Computer Society, Los Alamitos, CA, 1999.

[110] K. Seshadrinathan and A. C. Bovik. Unifying analysis of full reference image quality assessment. In *Proc. IEEE Int. Conf. on Image Process. (ICIP)*, San Diego, CA, October 2008.

[111] 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/.

[112] Hamid Rahim Sheikh. *Image Quality Assessment Using Natural Scene Statistics.* PhD thesis, Univ. of Texas at Austin, 2004.

[113] Hamid Rahim Sheikh and Alan Conrad Bovik. Image information and visual quality. *IEEE Trans. Image Process.*, 15(2):430–444, February 2006.

[114] Hamid Rahim Sheikh, Alan Conrad Bovik, and Lawrence Cormack. No-

reference quality assessment using natural scene statistics: JPEG2000. *IEEE Trans. Image Process.*, 14(11):1918–1927, November 2005.

[115] Hamid Rahim Sheikh, Alan Conrad Bovik, and Gustavo de Veciana. An information fidelity criterion for image quality assessment using natural scene statistics. *IEEE Trans. Image Process.*, 14(12):2117–2128, 2005.

[116] Hamid Rahim Sheikh, Muhammad Farooq Sabir, and Alan Conrad Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Trans. Image Process.*, 15(11):3441–3452, November 2006.

[117] Eero P. Simoncelli and William T. Freeman. The steerable pyramid: A flexible architecture for multi-scale derivative computation. In *Proc. IEEE Intl. Conf. on Image Process.*, Washington, D.C., October 1995.

[118] George W. Snedecor and William G. Cochran. *Statistical Methods.* Iowa State Press, eighth edition, 1989.

[119] Jean-Luc Starck, Michael Elad, and David L. Donoho. Image decomposition via the combination of sparse representations and a variational approach. *IEEE Trans. Image Process.*, 14(10):1570–1582, October 2005.

[120] Gabriele Steidl, Joachim Weickert, Thomas Brox, Pavel Mrazek, and Martin Welk. On the equivalence of soft wavelet shrinkage, total variation diffusion, total variation regularization, and SIDEs. *SIAM J. of Numerical Analysis*, 42(2):686–713, 2004.

[121] Thomas Stockham. Image processing in the context of a visual model. *Proc. IEEE*, 60(7):828–842, 1972.

[122] Dominik Strohmeier and Gerhard Tech. “Sharp, bright, three-dimensional”–open profiling of quality for mobile 3dtv coding methods. In Reiner Creutzburg and David Akopian, editors, *Proc. SPIE: Multimedia on Mobile Devices 2010*, volume 7542, San Jose, CA, January 2010.

[123] David S. Taubman and Michael W. Marcellin. *JPEG2000: Image compression fundamentals, standards, and practice.* Kluwer Academic Publishers, 2002.

[124] International Telecommunication Union (ITU): International Telegraph and Telephone Consultative Committee. Information technology–digital com-

pression and coding of continuous-tone still images: Requirements and guidelines, 1993. ITU-T T.81.

[125] Patrick Teo and David Heeger. Perceptual image distortion. In B. E. Rogowitz and Jan P. Allebach, editors, *Proc. SPIE: Human Vision, Visual Process., and Digital Display V*, volume 2179, pages 127–141, San Jose, CA, February 1994.

[126] L. L. Thurstone. A law of comparative judgement. *Psych. Rev.*, 34:273–286, 1927.

[127] Antonio Torralba. How many pixels make an image? *Visual Neuroscience*, 26:123–131, 2009.

[128] Shimon Ullman. *High-level Vision: object recognition and visual cognition.* MIT Press, Cambridge, MA, 1996.

[129] Antoon M. van Dijk and Jean-Bernard Martens. Subjective quality assessment of compressed images. *Signal Processing*, 58:235–252, 1997.

[130] A van Meeteren. Characterization of task performance with viewing instruments. *J. Optical Soc. of Am. A*, 7(10):2016–2023, 1990.

[131] Richard H. Vollmerhausen, Eddie Jacobs, and Ronald G. Driggers. New metric for predicting target acquisition performance. *Optical Engineering*, 43(11):2806–2818, November 2004.

[132] Video Quality Experts Group (VQEG). VQEG Final Report of FR-TV Phase II Validation Test, 2003.

[133] Video Quality Experts Group (VQEG). Final report from the vqeg on the validation of objective models of multimedia quality assessment, Phase I, 2008. Final Report Version 2.6.

[134] Video Quality Experts Group (VQEG). Multimedia group test plan, 2008. Draft Version 1.21.

[135] Martin J. Wainwright and Eero P. Simoncelli. Scale mixtures of gaussians and the statistics of natural images. In S. A. Solla, T. K. Leen, and K.-R. Miller, editors, *Advances in Neural Information Processing Systems*, pages 855–861. MIT Press, 2000.

[136] Martin J. Wainwright, Eero P. Simoncelli, and Alan S. Willsky. Random cascades on wavelet trees and their use in analyzing and modeling natural images. *Applied and Comp. Harmonic Analysis*, 11:89–123, 2001.

[137] Zhou Wang, Alan Conrad Bovik, Hamid Rahim Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4):600–612, April 2004.

[138] Zhou Wang, Eero P. Simoncelli, and Alan C. Bovik. Multi-scale structural similarity for image quality assessment. In *Proc. of the 37th IEEE Asilomar Conf. on Sig., Sys. and Comp.*, Pacific Grove, CA, November 2003.

[139] Andrew B. Watson. DCT quantization matrices visually optimized for individual images. In B. E. Rogowitz and Jan P. Allebach, editors, *Proc. SPIE: Human Vision, Visual Process., and Dig. Display IV*, volume 1913, pages 202–216, San Jose, CA, February 1993.

[140] Andrew B. Watson and Lindsay Kreslake. Measurement of visual impairment scales for digital video. In B. E. Rogowitz and T. N. Pappas, editors, *Proc. SPIE: HVEI VI*, volume 4299, pages 79–89, San Jose, CA, 2001.

[141] Andrew B. Watson and Joshua A. Solomon. Model of visual contrast gain control and pattern masking. *Journal of the Optical Society of America A*, 14(9):2379 –2391, 1997.

[142] Andrew B. Watson, Gloria Y. Yang, Joshua A. Solomon, and John Villasenor. Visibility of wavelet quantization noise. *IEEE Trans. Image Process.*, 6(8):1164–1175, 1997.

[143] Y.-G. Wu. Ga-based DCT quantisation table design procedure for medical images. *IEE Proc.-Vis. Image Signal Process.*, 151(5):353–359, 2004.

[144] Yang Zhang and Peter I. Rockett. The bayesian operating point of the canny edge detector. *IEEE Trans. Image Process.*, 15(11):3409–3416, November 2006.