Czech Technical University in Prague
Faculty of Electrical Engineering
Department of Radioelectronics

Université Bretagne Loire
Université de Nantes

Quality Assessment of Post-Processed Images

Doctoral Thesis

Lukáš Krasula

Prague, October 2016

Ph.D. Programme: Electrical Engineering and Information Technology
Branch of study: Radioelectronics

Doctoral School: Sciences and Technology of Information and Mathematics
Branch of study: Sciences of Information and Communication

Supervisor: Prof. Ing. Miloš Klíma, CSc.
Co-supervisor: Ing. Karel Fliegel, Ph.D.

Supervisor: Prof. Patrick Le Callet

Lukáš KRASULA

Metodologie hodnocení kvality obrazu po post-processingu

Quality Assessment Methodologies of Post-Processed Images

Abstrakt

Naprostá většina prací v oblasti hodnocení kvality byla, v posledních dvou dekádách, věnována kvantifikaci zkreslení způsobeného zpracováním obrazu. Bylo tedy vždy možné uvažovat, že původní obraz má nejlepší možnou kvalitu. V takovémto případě lze kvalitu měřit čistě jako podobnost zpracovaného obrazu vůči referenci. Některé algoritmy z oblasti post-processingu ale umožňují upravit estetické vlastnosti obrazu za účelem zvýšení vnímané kvality. Zde již reference o nejlepší možné kvalitě není předem známa a klasický přístup měření podobnosti tak nelze aplikovat. Cílem této práce je revidovat metodologie hodnocení kvality tak, aby se dokázaly vypořádat s problémy, které post-processing do hodnocení přináší. Algoritmy post-processingu, odpovídající tématu této práce, pocházejí ze dvou skupin – algoritmy vylepšení obrazu, zde zastoupené metodami doostřování, a techniky pro kompresi dynamického rozsahu (známé také jako algoritmy pro mapování tónů). Práce studuje jak subjektivní, tak objektivní metodologie hodnocení kvality v těchto oblastech a předkládá vhodná řešení překonávající doposud známé metody. Navíc je v práci navržen nový postup pro porovnání schopností objektivních metrik kvality, který řeší nedostatky v současnosti používaných metod.

Klíčová slova

Hodnocení kvality obrazu, subjektivní hodnocení kvality, objektivní metriky kvality, porovnání objektivních metrik kvality, post-processing, vylepšení obrazu, zobrazování s vysokým dynamickým rozsahem, mapování tónů.

Abstract

The vast majority of the work done in the field of quality assessment during last two decades has been dedicated to the quantification of the distortion caused by the processing of an image. The original image was, therefore, always considered to be of the best possible quality. In this kind of scenario, the notion of quality can be expressed as the fidelity of the processed version to the reference. However, some post-processing algorithms enable to adjust aesthetic properties of an image in order to enhance the perceived quality. In such cases, the best possible quality image is not available and the classical fidelity approach is no longer applicable. The goal of this thesis is to revise the quality assessment methodologies to cope with the challenges brought by the post-processing into the quality evaluation. The post-processing algorithms, relevant to the topic of this thesis, come from two groups – image enhancement, represented by image sharpening, and dynamic range compression (also known as tone-mapping) techniques. Both subjective and objective quality assessment methodologies applicable in these areas are studied and the suitable solutions, outperforming the state-of-the-art methods, are proposed. Moreover, a novel methodology for evaluating the performance of objective quality metrics, overcoming the shortcomings of the currently available methods, is presented.

Key Words

Image Quality Assessment, Subjective Quality Evaluation, Objective Quality Metrics, Performance Evaluation, Image Post-Processing, Image Enhancement, High Dynamic Range Imaging, Tone-Mapping.

Lukáš KRASULA

Méthodes d'évaluation de la qualité d'images post-traitées

Quality Assessment Methodologies of Post-Processed Images

Résumé

Ces vingt dernières années, la grande majorité des travaux réalisés dans le domaine de l'analyse de la qualité a été consacrée à la quantification de la distortion engendrée par le traitement d'une image. Par conséquent, l'image originale était toujours considérée comme étant de la meilleure qualité possible. Dans ce genre de scénario, la notion de qualité peut être exprimée comme la fidélité de la version traitée à sa version de référence. Cependant, des algorithmes de post-traitement permettent d'ajuster les propriétés esthétiques d'une image afin d'améliorer la qualité perceptible. Dans ce cas, il n'existe pas d'image ayant la meilleure qualité possible, et l'approche classique basée sur la fidélité ne peut plus être utilisée. L'objectif de cette thèse est de corriger les méthodologies d'analyse de la qualité afin de résoudre les difficultés d'évaluation de qualité que soulève le post-traitement. Les algorithmes de post-traitement, en rapport avec le sujet de cette thèse, proviennent de deux groupes : l'amélioration d'image, caractérisée par l'accentuation du contraste, et les techniques de compression de la plage dynamique (également appelée mappage tonal). Les méthodologies de l'analyse de qualité applicables dans ces domaines, tant subjectives qu'objectives, y sont étudiées, et les solutions proposées permettent de surpasser les méthodes les plus récentes. De plus, une nouvelle méthodologie est présentée afin d'évaluer les performances des indicateurs de la qualité objective, corrigeant les défauts des méthodes actuellement utilisées.

Mots clés

Évaluation de la qualité d'image, évaluation de la qualité subjective, évaluation de la qualité objective, évaluation de performance, post-traitement d'image, amélioration d'image, HDR imagerie, tone-mapping

Abstract

The vast majority of the work done in the field of quality assessment during last two decades has been dedicated to the quantification of the distortion caused by the processing of an image. The original image was, therefore, always considered to be of the best possible quality. In this kind of scenario, the notion of quality can be expressed as the fidelity of the processed version to the reference. However, some post-processing algorithms enable to adjust aesthetic properties of an image in order to enhance the perceived quality. In such cases, the best possible quality image is not available and the classical fidelity approach is no longer applicable. The goal of this thesis is to revise the quality assessment methodologies to cope with the challenges brought by the post-processing into the quality evaluation. The post-processing algorithms, relevant to the topic of this thesis, come from two groups – image enhancement, represented by image sharpening, and dynamic range compression (also known as tone-mapping) techniques. Both subjective and objective quality assessment methodologies applicable in these areas are studied and the suitable solutions, outperforming the state-of-the-art methods, are proposed. Moreover, a novel methodology for evaluating the performance of objective quality metrics, overcoming the shortcomings of the currently available methods, is presented.

Key Words

Image Quality Assessment, Subjective Quality Evaluation, Objective Quality Metrics, Performance Evaluation, Image Post-Processing, Image Enhancement, High Dynamic Range Imaging, Tone-Mapping.

Acknowledgement

First of all, I would like to thank all of my supervisors, professors Patrick Le Callet and Miloš Klíma, and Dr. Karel Fliegel for their guidance and time they dedicated to our consultations despite their busy schedules. I am really grateful for their advice and guidance. Furthermore, I want to acknowledge all the colleagues in Image and Video Communications group and Multimedia Technology group for all the fruitful discussions and creating a wonderful atmosphere in which it was a pleasure to work. It is important to appreciate the community in QUALINET as well. It has been a tremendous help and inspiration.

A special thanks goes to Lucie, Dim, Caroline, Guilleaume, and all my other friends in Nantes who helped to make it my second home. I also need to express my sincere gratitude to my family and friends who have always supported me and stood by my side. I would like to dedicate this thesis to them.

Last but not least, I want to thank the governments of the Czech Republic and France for providing me with the opportunity to study in this joint degree program and for their financial support.

This work was also partially supported by the following projects:

- SGS15/091/OHK3/1T/13 Optimization of Tone Mapping Operators Parameters for High Dynamic Range Images of the Student Grant Agency of CTU in Prague
- COST CZ LD12018 Modelling and verification of methods for Quality of Experience (QoE) assessment in multimedia systems – MOVERIQ of the Ministry of Education, Youth and Sports of the Czech Republic
- Grant No. P102/10/1320 Research and modelling of advanced methods of image quality evaluation of the Czech Science Foundation
- Grant No. 14-25251S Nonlinear imaging systems with spatially variant point spread function of the Czech Science Foundation
- SGS12/075/OHK3/1T/13 Region-of-Interest (ROI) importance with regards to the perceived image quality and implementation of an objective image quality metric based on ROI analysis of the Student Grant Agency of CTU in Prague
- UHD4U Project

Declaration

I hereby declare that I worked out the presented thesis independently and I quoted all the sources used in this thesis in accord with Methodical instructions about ethical principles for writing academic thesis.

Prague, October 2016 .....

Contents

Abbreviations

The list of abbreviations in the alphabetical order:

ACR	Absolute Category Rating
ACR-HR	Absolute Category Rating with Hidden Reference
AMOLED	Active Mask Organic Light Emitting Diode
ASDPC	Adaptive Square Design Paired Comparison
AUC	Area Under Curve
BIQI	Blind Image Quality Index
BRISQUE	Blind/Referenceless Image Spatial Quality Evaluator
BTL	Bradley-Terry-Luce model
CCFL	Cold-Cathode Fluorescent Lamp
CDF	Cumulative Distribution Function
CG	Computer Generated
CI	Confidence Interval
CPBD	Cumulative Probability of Blur Detection
CSF	Contrast Sensitivity Function
dB	Decibel
DCR	Degradation Category Rating
DCT	Discrete Cosine Transform
DL-LCD	Dual Layer Liquid Crystal Display
DLP	Digital Light Processing
DMD	Digital Micromirror Device
DMOS	Differential Mean Opinion Score
DR	Dynamic Range
DRIM	Dynamic Range Independent Metric
DRIQ	Digitally Retouched Image Quality
DSCQS	Double Stimulus Continuous Quality Scale
DSIS	Double Stimulus Impairment Scale
EDR	Extended Dynamic Range
FCM	Fuzzy C-Means
FISH	Fast Image Sharpness Metric
FMTMI	Fusion Metric for Tone-Mapped Images
FN	Feature Naturalness
FPC	Full Paired Comparison
FSIM	Feature Similarity Index
FSITM	Feature Similarity Index for Tone-Mapped Images
GCF	Global Contrast Factor
GGD	Generalized Gaussian Distribution
GSDF	Grayscale Standard Display Function
GSEA	Global Sharpening Enhancement Algorithm
GSM	Gaussian Scale Mixture

HD	High Definition
HDR	High Dynamic Range
HVS	Human Visual System
HSV	Hue Saturation Value Color Space
IFC	Information Fidelity Criterion
INLSA	Iterated Nested Least-Squares Algorithm
ITU	International Telecommunication Union
IW-SSIM	Information Weighted Structural Similarity Index
JNB	Just Noticeable Blur
JNBM	Just Noticeable Blur Metric
KLD	Kullback Leibler Divergence
KROCC	Kendall's Rank Order Correlation Coefficient
LAR	Locally Adaptive Resolution
LCD	Liquid Crystal Display
LDR	Low Dynamic Range
LED	Light Emitting Diode
LWMPA	Locally Weighted Mean Phase Angle
MAD	Most Apparent Distortion
MLE	Maximum Likelihood Estimation
MOS	Mean Opinion Score
MSE	Mean Squared Error
MSCQS	Multiple Stimulus Continuous Quality Scale
MS-SSIM	Multi Scale Structural Similarity Index
MTF	Modulation Transfer Function
MVG	Multivariate Gaussian Model
NSS	Natural Scene Statistics
OLED	Organic Light Emitting Diode
OS	Observer's score
PC	Paired Comparison
PCM	Paired Comparison Matrix
PDF	Probability Density Function
PLCC	Pearson's Linear Correlation Coefficient
PSF	Point Spread Function
PSNR	Peak Signal to Noise Ratio
QAC	Quality Aware Clustering
QoE	Quality of Experience
QoS	Quality of Service
RGB	Red, Green, Blue
RME	Root Mean Enhancement
RMS	Root Mean Square
RMSE	Root Mean Squared Error
ROC	Receiver Operating Characteristic
SDPC	Square Design Paired Comparison
SDME	Second Derivative-Based Measure of Enhancement
SDR	Standard Dynamic Range
SF	Structural Fidelity
SN	Statistical Naturalness
SNR	Signal to Noise Ratio
SROCC	Spearman's Rank Order Correlation Coefficient
SS	Single Stimulus

SSIM	Structural Similarity Index
SVD	Singular Value Decomposition
SVM	Support Vector Machine
TM	Thurstone-Moesteller model
TMO	Tone-Mapping Operator
TMQI	Tone-Mapped Images Quality Index
VDP	Visible Difference Predictor
VIF	Visual Information Fidelity
VQEG	Video Quality Experts Group

1

Introduction

For human beings as a species, visual information has always been of crucial importance. It is used as the main indicator for orientation, identification, object recognition, and many other everyday tasks. From all of the senses, humans tend to consider their sight as the most reliable one. This is the reason why human race started capturing visual information for describing the world and for future recollection of events. Everything from the first drawings on the walls of caves to the advanced digital photography, as it is known today, can be considered as capturing and reproduction of the visual information.

In recent years, the digital imagery has become an omnipresent part of the modern life. Digital cameras, computers, tablets, smartphones, and other devices able to capture and reproduce an image in the fraction of a second are affordable to virtually everyone. With this, observers' experience with image technology and thus their demands on the image quality has grown significantly. The quality of images is affected by plenty of factors in the image processing chain. These are, for example, properties of the scene (e.g. illumination, colors, etc.), components of the capturing device (e.g. lens, camera chip, etc.), image reconstruction algorithms (e.g. demosaicing, etc.), processing of the image (e.g. compression, etc.), properties of the communication channel, and many more. Some of these factors are easier to be controlled than other but all of them should be taken into account when designing any part of the image processing chain.

In order to optimize the appearance of the resulting image, it is necessary to quantify the impact of the above stated factors (either individually or overall) on the final outcome, i.e. to reliably assess the quality. A trustworthy quality evaluation tool can be used to rigorously compare different approaches in image processing (e.g. different processing algorithms) or to design and optimize new methods that can then be used within the processing chain.

Nevertheless, the notion of the quality can be seen from multiple angles. Especially in the area of telecommunications, historically most rooted understanding of quality is connected to the concept of *Quality of Service (QoS)*. It is defined in ITU-T Recommendation E.800 [1] as "The totality of characteristics of a telecommunications service that bear on its ability to satisfy stated and implied needs of the user of the service." However, since QoS is mostly focused on the physical properties of the service itself, it does not involve all the factors affecting the resulting quality as perceived by the observer. To provide more general concept expressing that the center of attention is the end-user, *Quality of Experience (QoE)* has been introduced. Despite the fact that discussions about the concept started already in 1990's, the solid definition has been missing until 2012. In this year, experts from the European Network on Quality of Experience in Multimedia Systems and Services (QUALINET)[1] released the White Paper on Definitions of Quality of Experience [2]. The working definition provided in the white paper states that "QoE is the degree of delight

[1] http://www.qualinet.eu/ (retrieved on 30/08/2016)

or annoyance of the user of an application or service. It results from the fulfilment of his or her expectations with respect to the utility and/or enjoyment of the application or service in the light of the user's personality and current state." It can be noted that the perception of quality can differ from person to person and can be changed with various conditions.

In this thesis, the **optimal quality** will be regarded as **the expression of the scene that is the most aesthetically pleasing for the observers**. This brings a fundamental change to the *classical quality assessment paradigm* where the optimum is an artifact-free representation of the real-world. Including the aesthetic aspect to the understanding of quality introduces much more diversity into the observers' opinions but, at the same time, shifts the area of quality evaluation further towards the completely user-centered model.

The aesthetic judgment, according to Immanuel Kant [3], is formed by two contradictory properties:

- *Subjectivity* which is based on a subjective feeling of pleasure or displeasure, and
- *universality* that involves an expectations or claims on the agreement of others.

Considering these two principles, a debate is still going on whether the aesthetic judgment can be at all universal. This is referred to as the "Big Question" of aesthetics [4]. The *generalists* believe that there is a clear general reason driving the aesthetic judgment while the *particularists* deny it. Aydın et al. [5] argue that, practically speaking, the truth seems to be somewhere in the middle. Although there is a significant influence of the personal taste in the observers' preferences, some general trends can definitely be identified as well. The goal of this thesis is to revise the quality assessment paradigm in order to be able to identify these general tendencies and reliably quantify them.

The remainder of the introduction will firstly describe the *classical quality assessment paradigm* in detail. Further, the image post-processing approaches and the challenges they bring to the evaluation of quality will be introduced. Finally, the objectives of this thesis will be summarized.

1.1 Classical Quality Assessment Paradigm

The absolute majority of the research regarding quality assessment has been based on the assumption that the processing introduces a distortion to the reference stimulus [6]. In other words, there is a *perfect quality* stimulus at the beginning of the processing chain and its quality is being *negatively* influenced by the processing algorithms. This assumption comes from the main application areas of quality assessment which are quantifying the impact of a compression and transmission errors. It enables to understand the quality evaluation process as a *fidelity* problem. In other words, the goal is to determine *how close* the processed (distorted) version is *to the original*. The quality assessment methodologies have, therefore, been designed for this type of scenario. Two fundamental approaches can be identified – subjective quality tests and objective quality metrics.

In subjective experiments, a group of observers is presented with a tested content and their opinions regarding its quality are collected. The evaluation always progresses under certain subjective procedure. There has been a number of procedures, suitable for different applications, being of different difficulty for the subjects, and requiring different preparation and results processing, introduced during the development of the quality assessment field. The detailed discussion regarding the subjective experimental procedures, including the in depth description of the most important methods and their applicability, advantages, and disadvantages, is provided in Chapter 2. The individual results from the observers are then pooled in order to determine a general information about the quality of the content within the test. Since the subjective experiments obtain the direct qualitative feedback from the participants, the resulting pooled opinion scores represent the most reliable reflection of the perceived quality. However, the time requirements for such tests allow only for very limited number of stimuli to be included, the experiments are also vulnerable to researchers' mistakes and biases, they mostly need specialized facilities, the participants need to be paid, etc. Moreover, the nature of the subjective tests does not allow for real-time quality evaluation and they, therefore, cannot be used to control the quality within the transfer channel or as a base for optimization.

The second approach towards the quality assessment is the use of objective quality metrics. These will be presented in Chapter 3. The idea is to develop algorithms able to automatically evaluate the perceived quality. The advantages are obvious since this approach removes most of the above stated deficiencies of subjective tests. The most crucial downside is the reliability since, unlike in case of subjective experiments, the human observers are not directly involved in the evaluation. The reliability of the metrics, thus, needs to be tested in every application before they are practically used. Objective quality criteria can be divided into three groups: Full reference, reduced reference and no reference (also known as blind) metrics.

Full-reference criteria require the presence of a whole original version of the stimulus. They are based on a comparison of a distorted version with the reference which corresponds to the *fidelity* scenario. This family is the most developed of the three and it has many subdivisions. Signal based metrics do not consider any assumptions about the two versions and simply mathematically compare the values of the two versions. Despite their rather poor performance in the terms of agreement with subjective tests, these criteria are widely used for their simplicity and speed. The widest part of the full reference metrics family consists of the metrics trying to model certain features of human visual system (HVS). These criteria are much more complex and they correspond much better to the subjective results.

Reduced-reference metrics are a kind of a compromise between full and no reference criteria. They do not need the whole original stimulus but only its certain characteristics or features (e.g. histogram, entropy, etc.). This group is the least popular and developed of the three.

The remaining class is formed by the blind quality criteria. Essentially, the quality is evaluated without having an access to any information about the original stimulus. This is much closer to the human perception because observers are able to recognize good or poor quality stimulus even without being exposed to the original version. The majority of the no reference metrics is specialized on a certain type of distortion or feature but there are also criteria based on the natural scene statistics (NSS) that attempt to evaluate overall quality regardless the processing. They can be divided into opinion aware and unaware according to the necessity to train them on subjective data.

Given the question of reliability of the objective quality criteria, their performance evaluation and comparison with respect to the ground truth data, obtained from the subjective experiments, is another important aspect of the quality assessment. The state-of-the-art methods used for this purpose are thoroughly described in Chapter 4.

1.2 Image Post-Processing and Quality Assessment

The previous section defined the classical *fidelity* approach in quality assessment. However, as described earlier, this thesis considers also the aesthetic attributes of the quality and, therefore, admits the possibility that the original version of the stimulus does not necessary has to be of the best possible quality. Thus, not only the *negative* impact of processing on the quality is recognized but the *positive* influence is considered as well. Obviously, in such scenario, the *fidelity* approach is no longer usable since the stimulus of the best possible quality is unknown.

The increase in the perceived quality can be caused by so called post-processing algorithms. For the purposes of this thesis, they can be classified into three groups:

- Restoration algorithms,
- enhancement algorithms, and
- algorithms adjusting the content for displaying.

The techniques from the first class have been designed in order to compensate for a certain type of distortion. A typical example could be algorithms for deconvolution, denoising, or deblocking. The abilities of such algorithms can, actually, be determined using a classical *fidelity* paradigm since the main concern is how well can the technique get rid of the particular artifacts. The distortion can, therefore, be applied on a

reference stimulus followed by the studied restoration algorithm, and the *fidelity* of the result to the reference can be evaluated. Another possibility is to use a no reference metric specialized on the particular distortion. Restoration techniques are not designed to increase the aesthetic properties of a stimulus and do not cause any increase of perceived quality when applied on the undistorted stimulus. Thus, this group is not relevant to the subject of the thesis.

The second group, on the other hand, is formed by the algorithms specifically designed to enhance the perceived quality of the stimulus. They are mostly specialized on a certain feature such as contrast, sharpness, colorfulness, etc. It has been proven in a subjective study [7], that human observers visually prefer content with enhanced features over the truthful, more realistic representation of the real world. To evaluate the performance and optimize the setting of the enhancement algorithms, classical quality assessment paradigm needs to be revised.

The last group has become a hot topic especially with the increase in popularity of High Dynamic Range (HDR) imaging. Here, the full range of luminance values in the real world scenes is captured and stored. However, the dynamic range of such scenes is much higher than what standard displaying devices are able to reproduce. This lead to development of dynamic range compression algorithms (which will be referred to as Tone-Mapping Operators (TMOs) throughout the thesis) which should transform the HDR image into a form displayable on the regular devices. The goal is to provide the reproduction of the scene of the highest possible quality. The difference in dynamic range of the original image and the outcome of the tone-mapping process makes the classical quality assessment paradigm inapplicable as well.

The last two classes of post-processing algorithms require revising of the quality assessment methodologies. Each of the groups introduces related, yet slightly different challenges into the quality evaluation process. For this reason, they will be treated separately further on in this thesis.

1.2.1 Challenges in Quality Assessment of Enhanced Images

As stated above, enhancement algorithms are capable of adjusting aesthetic properties of the stimulus and thus increase the perceived quality. However, the quality assessment after enhancement needs to deal with several problems as identified by Saad et al. in the Video Quality Experts Group (VQEG) e-letter [8]. Note that the challenges stated here and the main principles that will be discussed further are applicable on all of the enhancement algorithms. As a representative, image sharpening has been selected since it is one of the most common enhancement techniques having its roots already in an analog photography.

The Overshoot Effect

Most of the enhancement algorithms enable setting of one or more parameters. This way the strength and possibly other properties of enhancement effect are influenced. When applying an enhancement method, the resulting quality can, in fact, be affected in both ways – some setting the parameters enhances the original stimulus but after reaching a certain point (i.e. the optimal amount of enhancement) the perceived quality starts to drop. This happens due to the loss of naturalness. There is a thin line between the aesthetically pleasing amount of enhancement and the degree "violating" the stimulus too much. The effect is called over-enhancement or the "overshoot effect" [8]. The main problem is that the optimal level of enhancement also varies with the content, i.e. the overshoot effect appears for different set of parameters with different strength depending on the scene. The phenomenon is illustrated in Figure 1.1.

The figure shows the dependence of normalized perceived quality on the amount of enhancement (the strength which is set by parameters) for three different source contents – I1, I2, and I3. It can be seen that the optimal enhancement for one source content can lead to the significant drop in quality for the other (see ΔQ_{I1vsI3} which is a difference in perceived quality for contents I1 and I3 when the same amount of enhancement is applied). Identification of the optimal parameters is one of the goals of quality assessment methods.

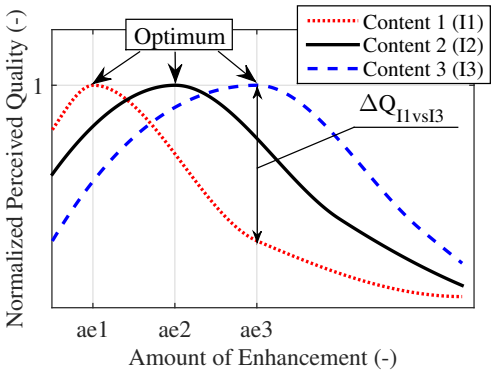
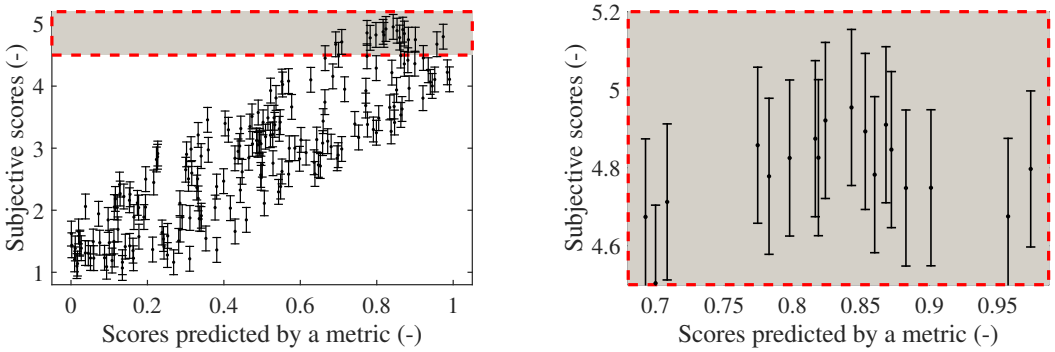


Figure 1.1: Content dependence of the overshoot effect.



(a) Overall, the predicted values correlate well with the subjective scores

(b) In the narrow range, the correct order cannot be determined from the data

Figure 1.2: The range effect.

The Range Effect

In classical quality assessment scenarios, mostly the wide range of degradations from almost imperceptible artifacts to heavy distortions is considered. In case of enhancement, on the other hand, much smaller changes in perceived quality are being considered since the main focus is on the stimuli around the optimum. This puts higher requirements on the discriminatory power of the subjective methodologies because they need to be highly effective in terms of reliable differentiation between the quality of stimuli in the narrow quality range. The difficulty of the task for the observers is also much higher.

Moreover, the classical methods for objective metrics' performance evaluation can be misleading. These methods are typically based on a correlation of subjective opinion scores and the respective objective metric's values. The Figure 1.2 depicts an example of the problem when only the small range is considered. Note that the data were created artificially for the illustration only. Overall, the scores predicted by an objective quality metric correlate well with the subjective scores obtained from the observers (result of the subjective test), as can be seen in Figure 1.2(a), but this correlation is lost when only the range highlighted by the rectangle is taken into account (Figure 1.2(b)). However, most of the scores within this range are not statistically significantly different and the correct order cannot be determined from the data. The low correlation is based on the assumption of the order which is uncertain and may be incorrect. The combination of the narrow quality range and the possible differences in personal taste of observers (some observers prefer more enhancement, other are more sensitive on the presence of the overshoot effect) increases the probability that the confidence of the estimated overall subjective scores will be low. It is, therefore, absolutely crucial to consider the uncertainty of the subjective scores in this kind of applications. This will be thoroughly discussed in Chapters 4 and 5.

1.2.2 Challenges in Quality Assessment of Tone-Mapped High Dynamic Range Images

The main motivation for HDR imaging is the capturing and reproduction of all the contrasts and details in the scene. TMOs should therefore ideally compress the dynamic range while maintaining details. However, it has been discovered [9] that human observers simultaneously require the tone-mapped image to look natural.

Maintaining Details and Naturalness

These two goals are contradictory since in an effort to reproduce all the details some naturalness corrupting artifacts often occur. It is, therefore, necessary for TMOs to find a good balance between the two entities. It is believed [10] that naturalness is a combination of different factors such as contrast, colorfulness, brightness, details, artefacts, etc. The problem is that TMOs often introduce *spatially non-uniform distortion of contrast* resulting in unnatural look.

Information about *color* is often neglected when evaluating the quality. However, it is very important in the area of HDR content processing. Number of TMOs work with color appearance models or color correction algorithms and can, therefore, produce results with different color reproduction. It is thus necessary to incorporate the chrominance information in the evaluations since it can also have a crucial influence on the naturalness.

Complexity of the Distortions

Considering all the above stated factors, versions of the same HDR scene obtained by various TMOs (also with various parameters settings) are often perceptually very different, although possibly of high quality. This makes it hard for the observers to create their mental scale for assigning quality scores or rank the stimuli in the set. Subjective experiments therefore have to be designed in a way which decreases the cognitive load as much as possible. Also the objective methods cannot be trained to recognize just a single type of distortion as in some classical quality assessment applications, since the TMOs influence multiple aspects of the image at the same time.

Here, the personal taste of the observers regarding the aesthetic attributes of an image play even more important role. To capture the common trends in the aesthetic judgment, it is, again, absolutely necessary to consider the confidence of the overall subjective scores.

1.3 Objectives of the Thesis

Regarding the above explained problems, the objectives of the thesis have been identified.

1. **Design of the performance evaluation methodology suitable for comparison of objective metrics in post-processing scenarios.** The state-of-the-art performance evaluation methods will be deeply studied and their suitability for the task will be determined. The goal is to propose a methodology able to provide fair, unbiased comparison with respect to the above described scenario.

2. **Investigation of the methodology for subjective assessment appropriate for post-processed images providing reliable evaluations.** Since the novel challenges bring more complications to the subjective quality evaluation, the appropriate robust procedure to be used in experimental studies will be carefully selected.

3. **Conducting subjective studies on challenging content to obtain a ground truth data for objective metrics training and testing.** The selected procedure will be used for creation of the representative databases enabling to test the abilities of the objective quality criteria in the given contexts.

4. **Proposing objective metrics suitable for the above described scenarios.** The final goal is to find or eventually develop metrics capable of reliable quality assessment of post-processed images.

1.4 Outline of the Thesis

Chapter 2 provides an introduction into subjective quality assessment methodologies, discusses their advantages and disadvantages, and describes the procedures for processing their results. Chapter 3 is dedicated to objective quality metrics including their classification and delineation. Chapter 4 describes the state-of-the-art methods for evaluating the performance of objective metrics with respect to the subjective results and identifies their weaknesses. Chapter 5 proposes novel methodologies overcoming the disadvantages of the current performance evaluation measures (see Objective no. 1 in Section 1.3). The discussion about the change of the role of the reference in the post-processing scenarios and the possibilities of quality assessment in this context can be found in Chapter 6. Adjustments of both subjective and objective quality assessment procedures (Objectives 2, 3, and 4 from Section 1.3) for sharpened and tone-mapped HDR images are provided in Chapters 7 and 8, respectively. Finally, Chapter 9 summarizes the contributions of this thesis and the activities that have been conducted during the Ph.D. studies. The graphical representation of the outline is depicted in Figure 1.3.

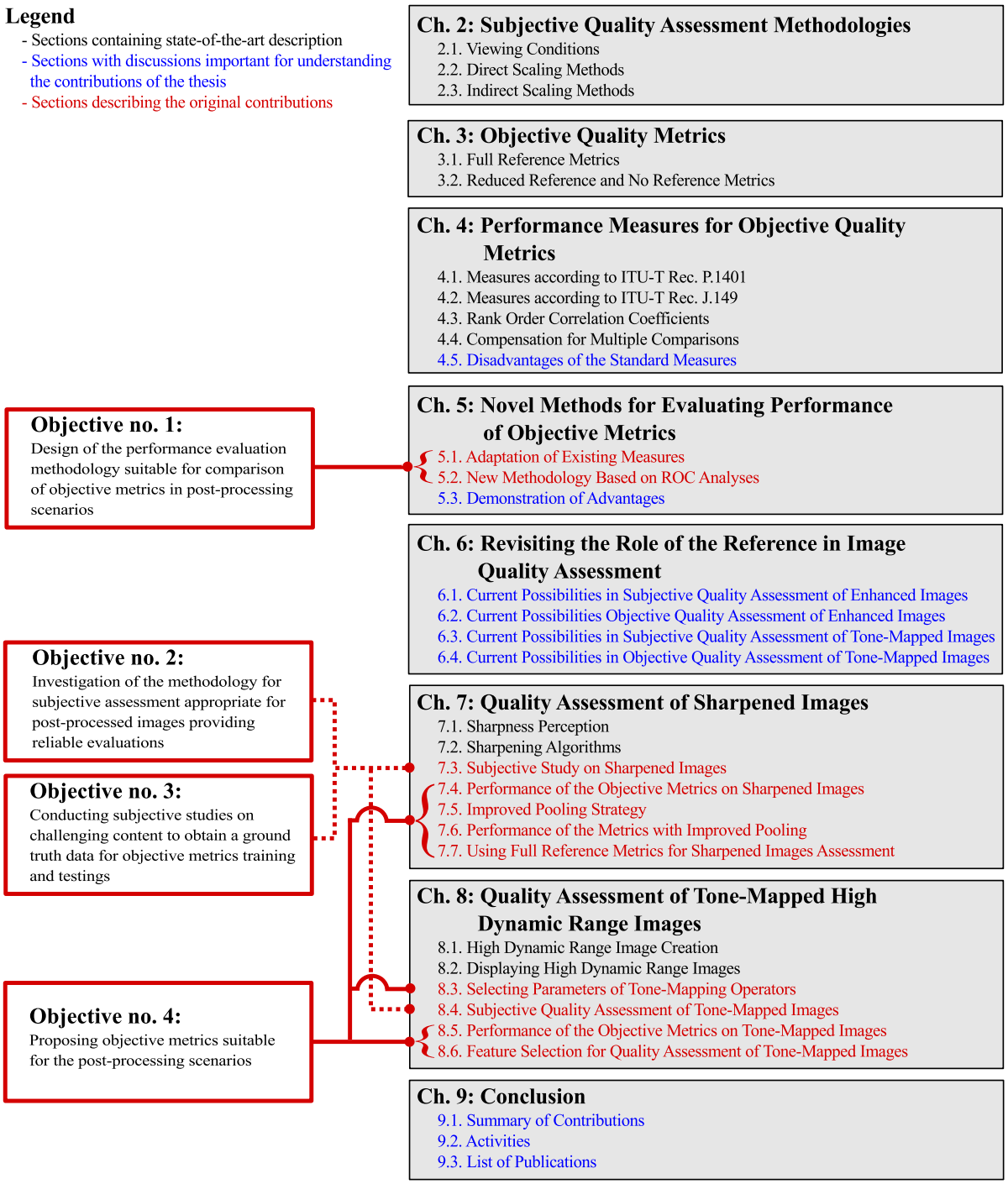


Figure 1.3: The graphical outline of the thesis.

2

Subjective Quality Assessment Methodologies

In subjective tests, a panel of human observers is asked to evaluate the quality of stimuli that are presented according to a specific procedure. The composition of the group can vary from one application area to the other. It is mostly desirable for the panel to cover as wide range with respect to age, gender, and cultural background as possible.

Subjective tests are the most reliable way to evaluate the image quality. Nevertheless, they are very much dependent on the test design and can be significantly biased when not carefully prepared, conducted, and interpreted. In order to maximize the reliability and reproducibility of the experiments, number of standards and recommendations has been issued.

The recommendations describing preparation, conditions, procedures, processing of results, and other important aspects of subjective experiments that are most relevant to the topic of this thesis are ITU-R Recommendation BT.500-13 [11] and ITU-T Recommendation P.910 [12].

The procedures for subjective tests can be divided into *direct* and *indirect* scaling methods [13]. The advantage of the *direct* scaling procedures is that the results from individual observers are directly on the ratio scale (the differences between various types of scales are summarized in Table 2.1) that is clearly defined by the particular methodology which simplifies the following processing and interpretation. On the other hand, the *indirect* procedures typically provide higher discriminatory power and can be less complicated and tiring for the observers. Several works has shown that the *indirect* scaling methods need lower number of subjects to provide the same reliability as the *direct* scaling procedures [14, 15].

Scale Type	***Properties***
Nominal	Discrete categories, not necessarily ordered
Ordinal	Scores are ordered but not numerical
Interval	Numerical (the distance between scores is known) but there is no absolute reference point, i.e. relative scale
Ratio	Numerical with respect to the reference point, i.e. absolute scale

Table 2.1: Different types of scales.

The following sections summarize the standard viewing conditions and the most popular subjective experiment procedures.

2.1 Viewing Conditions

The viewing conditions have a huge impact on the observers perception and therefore on the test itself. Most of the experiments are conducted in the laboratory conditions where the environment is strictly controlled in order to ensure the reproducibility. If the conditions are free, e.g. in case of crowdsourcing experiments [16], the number of observers needs to be much higher to compensate for the different viewing setups. Moreover, in some applications, such as psychophysical measurements, the results need to be always related to the particular conditions.

Recommendation ITU-T P.910 [12] provides the range of environmental requirements for conducting laboratory tests for multimedia purposes (see Table 2.2). Nevertheless, it is always necessary to report the exact room illumination and display types used within the experiment.

Parameter	***Setting***
Viewing distance	1 - 8 $\times$ screen height
Peak luminance of the screen	100 - 200 cd/m^2
Ratio of luminance of inactive screen to peak luminance	≤ 0.05
Ratio of the luminance of the screen, when displaying only black level in a completely dark room, to that corresponding to peak white	≤ 0.1
Ratio of luminance of background behind picture monitor to peak luminance of picture	≤ 0.2
Chromaticity of background	D_{65}
Background room illumination	$\leq$ 20 lux

Table 2.2: Viewing conditions according to ITU-T Rec. P.910.

The optimal viewing distance depends mainly on the resolution of the used display. For the full High Definition (HD) Liquid Crystal Displays (LCD), i.e. 1920 $\times$ 1080 pixels, which will be mainly used in the experimental part of this thesis, the optimal value is 3.2 times screen height [17].

The illumination of a background allows maximum detectability of distortions. For some applications the value can be determined by the application itself. The chromaticity of a background can be adapted to the chromaticity of a monitor when using a computer screen.

The above stated parameters are defined for the classical television applications. In case of HDR imaging, the conditions need to be adjusted accordingly. The specific conditions used within the conducted experiments with HDR display will be discussed in Section 8.4.4.

2.2 Direct Scaling Methods

As mentioned previously, direct scaling methods collect the opinions of observers regarding each particular stimulus directly on the ratio scale. They can be related to the *magnitude estimation* method, as introduced by Stevens [18]. The scale values depend on the selected procedure. Once the results from all observers are collected, the processing including outlier detection and averaging is applied directly on the raw data. The final outcome of the experiments has the form of *mean opinion scores* (MOS) or *differential mean opinion scores* (DMOS) with respective confidence intervals calculated from the variance of the raw scores. MOS values represent the perceptual quality of each stimulus, i.e. the stimuli of high quality reach higher MOS values. In case of DMOS, the quality of the reference stimulus is considered of the highest possible quality and the *difference* of the scores with respect to the reference scores are taken. Note that DMOS values can mostly be calculated from the MOS and vice versa.

Given the higher requirements on the observers, it is recommended to conduct a training session (with the content that does not appear in the test) where the extreme cases of quality are shown prior to the test itself. This helps participants to calibrate their mental scale to quality range within the test. The opinion collection procedures together with their respective MOS/DMOS scales are described below.

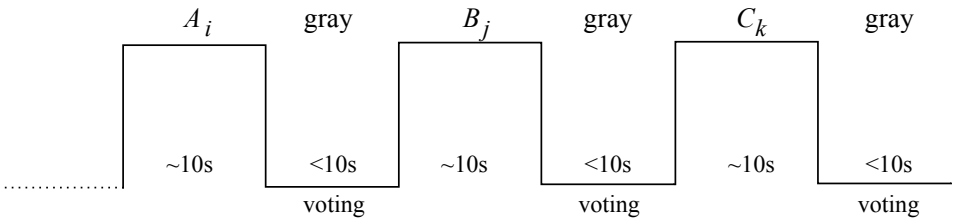


Figure 2.1: The timeline of the ACR methodology. A_i represents a content A under a test condition i, B_j a content B under a condition j, and C_k a content C under a condition k. During the voting period, the gray background is displayed.

2.2.1 Single Stimulus / Absolute Category Rating

Single Stimulus methodology (SS) [11], also known as Absolute Category Rating (ACR) [12], is the simplest subjective procedure. The stimuli are presented to the observers in a random order one at a time, followed by the mid-gray screen during the voting period. Depending on the research question, the reference can be explicitly shown to the observers or not used at all.

The participants are mostly asked to evaluate the quality of the five grade scale as:

5 Excellent

4 Good

3 Fair

2 Poor

1 Bad

The outcome of the experiment using this scale (after the results processing as described in Section 2.2.4) are the MOS scores ranging from 1 to 5.

Since the method is the least time consuming, stimuli can be also displayed repeatedly to increase the discriminatory power and reliability. The timeline of the procedure can be seen in Figure 2.1. The method has also been implemented as part of the MATLAB-based applications for image processing and quality assessment [19, 20].

ACR with Hidden Reference

The special case of the procedure with so called *hidden reference* (ACR-HR) can be used as well. Here, the reference stimuli are present within the set but the observers are not aware of this fact. This enables calculating the DMOS score for each stimulus with respect to the score of the reference as

$$DMOS(A_i) = MOS(A_i) - MOS(A_{\text{HR}}) + 5, \tag{2.1}$$

where $DMOS(A_i)$ and $MOS(A_i)$ are the DMOS and MOS scores for a content A under a condition i, respectively and $MOS(A_{\text{HR}})$ is the MOS score for the hidden reference of the content A. The higher the DMOS, the closer is the quality of the stimulus to the reference.

In case that the MOS value of the stimulus is higher than the MOS of the reference, i.e. $DMOS(A_i) > 5$, so called *crushed* DMOS value defined in [12] as

$$crushed\, DMOS(A_i) = \frac{7 \times DMOS(A_i)}{2 + DMOS(A_i)} \text{ when } DMOS(A_i) > 5, \tag{2.2}$$

is used instead.

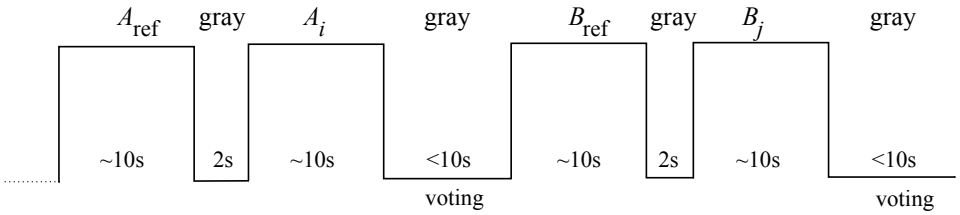


Figure 2.2: The timeline of the DCR methodology. A_i represents a content A under a test condition i, B_j a content B under a condition j, and A_{ref} and B_{ref} the reference stimuli for the contents A and B, respectively. During the voting period, the gray background is displayed.

2.2.2 Double Stimulus Impairment Scale / Degradation Category Rating

Another popular procedure is Double Stimulus Impairment Scale (DSIS) [11], also called Degradation Category Rating (DCR) [12]. As the names imply, it is assumed that the stimulus is degraded compared to the reference. The goal is to evaluate how much does the distortion affect the perceived quality. The displaying of each stimulus is always preceded by showing the reference for the same content, as shown in Figure 2.2. It is obvious that the length of the test is higher than in the case of ACR. However, if only static images are evaluated this can be compensated by showing the reference and the distorted versions simultaneously [21].

The observers are asked if the degradation is:

5 Imperceptible

4 Perceptible but not annoying

3 Slightly annoying

2 Annoying

1 Very annoying

The MOS scores (see Section 2.2.4) are therefore again in the range from 1 to 5.

2.2.3 Double Stimulus Continuous Quality Scale

The last direct scaling procedure mentioned in this thesis will be Double Stimulus Continuous Quality Scale (DSCQS) [11]. Similarly to DSIS, the stimuli are always displayed and evaluated in pairs (from the same source content). However, the observers are not explicitly told that one of the two stimuli is reference and they are supposed to evaluate both of them at the same time. The recommendation allows two different variants:

Variant I The observer is allowed to freely switch between the two stimuli and then score the quality of both of them on the scale.

Variant II The stimuli are presented twice, according to the Figure 2.3.

The observers assess the quality of both of the stimuli on the continuous scale (see Figure 2.4) which is normalized to integer values from 0 to 100 for the purpose of results processing (see Section 2.2.4). Since one of the stimuli in pair is always the reference, the DMOS values can again be obtained as the difference between the scores.

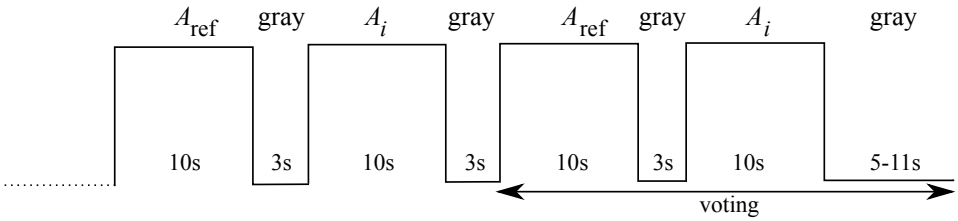


Figure 2.3: The timeline of the DSCQS methodology – Variant II. A_i represents a content A under a test condition i and A_{ref} and the reference for the content A. The voting period begins when the first stimulus is displayed for the second time. Note that the order of the reference and distorted stimulus is random.

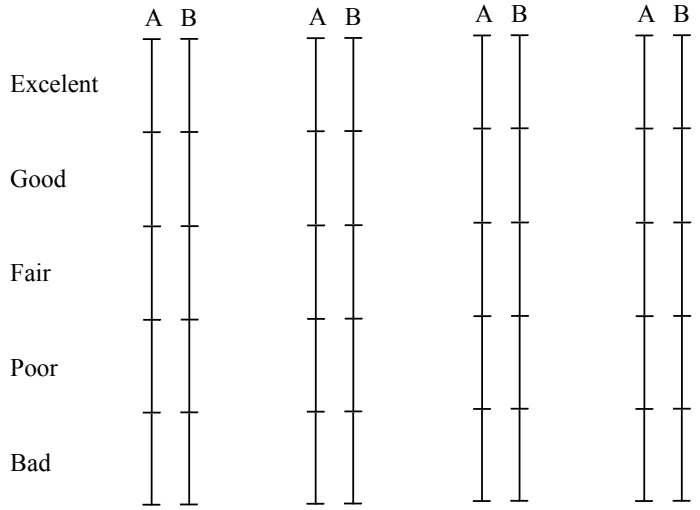


Figure 2.4: The scoring sheet for the DSCQS method.

2.2.4 Results Processing

Processing of results from direct scaling methods is thoroughly described in the Annex 2 of the ITU-R Rec. BT.500-13 [11]. The goal is to transform the collected observers' scores (OS), which are integer values in the range depending on the procedure, into the MOS/DMOS values with respective confidence intervals (CI) that will reliably reflect the general opinion. This also requires a screening of the observers in order to detect systematic outliers.

Note that the following procedures assume that the observers' opinions are drawn from the normal distribution with the mean equal to the true quality value. The distribution of data should be checked before the analysis since the central tendency of the data from different distributions than normal are better described by median instead of arithmetic mean and non-parametric methods should be used for analysis [22].

Mean Scores Calculation

Considering the i-th stimulus, the test provides $N(i)$ integer scores $OS(i,n)$, where $N(i)$ is the total number of times the particular stimulus has been evaluated. The mean score for this stimulus is obtained simply as

$$MOS(i) = \frac{1}{N(i)} \sum_{n=1}^{N(i)} OS(i,n). \tag{2.3}$$

The overall MOS for a specific test condition or content can be obtained by averaging over different contents or conditions, respectively.

Confidence Intervals Calculation

In absolute majority of cases, 95% CI are used. Such CI for the i-th stimulus is defined as

$$CI(i) = [MOS(i) - \delta(i), MOS(i) + \delta(i)], \tag{2.4}$$

where

$$\delta(i) = 1.96\frac{SD(i)}{\sqrt{N(i)}}. \tag{2.5}$$

Standard deviation of the raw scores for the i-th stimulus is given by

$$SD(i) = \sqrt{\sum_{n=1}^{N(i)} \frac{\Big(OS(i,n) - MOS(i)\Big)^2}{N(i) - 1}}. \tag{2.6}$$

The value of 1.96 in the equation (2.5) is obtained from the cumulative distribution function (CDF) of the normal distribution. In case of smaller number of observers (typically $N(i) < 30$), it is advisable to consider the Student distribution instead [22], i.e. the equation (2.5) changes to

$$\delta(i) = t\Big(0.975, N(i)\Big)\frac{SD(i)}{\sqrt{N(i)}}. \tag{2.7}$$

where the value $t\Big(0.975, N(i)\Big)$ corresponds to the a two-tailed Student distribution with $N(i) - 1$ degrees of freedom on 95% confidence level.

Screening of the Observers

The procedure for screening of the observers should not be applied more than once to the results of the same experiment and should be restricted to the cases with small number of observers (e.g. $N(i) < 20$ for $\forall i$) all of whom are non-experts [11].

The first step of the procedure is to calculate the kurtosis of the distribution of OS for each stimulus. The kurtosis for the i-th stimulus is computed as

$$\beta_2(i) = \frac{\mu_4(i)}{(\mu_2(i))^2}, \tag{2.8}$$

where $\mu_4(i)$ and $\mu_2(i)$ are the fourth and the second order moment of the distribution of votes for the i-th stimulus. The x-th order moment for the given stimulus is defined as

$$\mu_x(i) = \frac{\sum_{n=1}^{N(i)} \Big(OS(i,n) - MOS(i)\Big)^x}{N(i)}. \tag{2.9}$$

The kurtosis is used to check if the distribution of opinions is normal or not. If the value of $\beta_2(i)$ lies between 2 and 4, the distribution is considered to be (close to) normal.

In the next step, the individual votes of each observer n are considered. The goal is to find the values of P_n and Q_n for every participant. The procedure is visualized in the Algorithm 1.

If the observer is rejected, the MOS values and CI should be recalculated without considering any of his/her votes.

Algorithm 1 Procedure for the screening of the observers.

for every observer n **do**
  $P_n = 0$
  $Q_n = 0$
  for every stimulus i **do**
    if $2 \leq \beta_2(i) \geq 4$ **then**
      if $OS(i,n) \geq MOS(i) + 2 \times SD(i)$ **then**
        $P_n = P_n + 1$
      end if
      if $OS(i,n) \leq MOS(i) - 2 \times SD(i)$ **then**
        $Q_n = Q_n + 1$
      end if
    else
      if $OS(i,n) \geq MOS(i) + \sqrt{20} \times SD(i)$ **then**
        $P_n = P_n + 1$
      end if
      if $OS(i,n) \leq MOS(i) - \sqrt{20} \times SD(i)$ **then**
        $Q_n = Q_n + 1$
      end if
    end if
  end for
  if $\frac{P_n+Q_n}{N} > 0.05 \wedge \left|\frac{P_n-Q_n}{P_n+Q_n}\right| < 0.3$ **then**
    reject the observer n
  end if
end for

2.3 Indirect Scaling Methods

The indirect scaling methodologies give the observers an *ordinal task*. HVS is very good in making decisions about the correct order even in the situations where assigning numerical values becomes challenging. This phenomenon leads to the increase in reliability and lower requirements on the number of participants in the test. Moreover, the task itself is mostly more comfortable and less demanding for the observers.

However, it is only possible to transfer the results of the *ordinal task* to the *interval scale* (see Table 2.1). This was mainly regarded as a severe drawback, especially when using the subjective tests results for performance evaluation of objective metrics. Nevertheless, if special methodologies are used, this disadvantage can be compensated and the higher discriminatory power of the indirect scaling methods can be fully exploited even in these applications. This will be discussed in detail in Chapter 5.

In this thesis, two most popular approaches are described – namely *ranking* and *paired comparison (PC)*. Some other possibilities can be found e.g. in [13].

2.3.1 Ranking

The process of ranking is self-explanatory. An observer is given the set of stimuli obtained from the same source content (with or without the presence of the reference) and is asked to rank them. The task is generally pleasant for the participants for its "puzzle-like" nature [13]. The procedure is very popular especially in printing industry where it is easy to provide the subjects with access to all of the stimuli in the set at the same time.

Considering the high resolution images or videos with limited number of displays, the procedure becomes less simple to set up. The problem can be solved with smaller thumbnails that are displayed at the same time and allow to show the full resolution stimulus upon request but the simplicity of the task is then corrupted and PC methodology becomes preferable (especially its shortened form as described in the following Section).

2.3.2 Paired Comparison

Paired Comparison (PC) can be regarded as breaking down the ranking process into basic tasks where each stimulus is compared to the others. It is also standardized in ITU-T Rec. P.910 [12] and ITU-R Rec. BT.500-13 [11] (in several modifications called Stimulus Comparison methods).

In the basic PC test, every subject is presented with a pair of stimuli coming from the same source content simultaneously and is supposed to decide which of the stimuli has a higher value of the measured entity (in this case the quality). Sometimes it is possible to allow observers to state that the stimuli are the same (i.e. ternary scale is used). However, most of the PC experiments are so called forced choice tests, where the participants have to choose one of the two stimuli. It is assumed that when the stimuli are close, the probability of selecting one over the other is close to 50% and the votes will be distributed accordingly. Note that it is also possible to present the pair sequentially (i.e. not simultaneously).

The result of the PC experiment is the Paired Comparison Matrix (PCM) of size $a \times a$ for each source content, where a is the total number of stimuli within the source content. If the observer n selects stimulus A_i over A_j then $PCM_{A,n}(i,j) = 1$. Note that if the ternary scale is used and the participant evaluates the pair as not different then $PCM_{A,n}(i,j) = PCM_{A,n}(j,i) = 0.5$. The overall PCM for the content A is obtained as

$$PCM_A = \sum_{n=1}^{N} PCM_{A,n}. \tag{2.10}$$

Note that $PCM_A(i,i) = 0$ for $\forall i$ since no stimulus is compared to itself.

PC methodology provides the easiest task to the observers since only one comparison is required at the time. The obvious downside is the time requirements of the test. When considering all the possible stimuli pairs in both configurations (i.e. A_iA_j and A_jA_i), the total number of comparisons is $a(a-1)$. This number

can be halved by randomization of the order in which the stimuli are displayed (i.e. as right and left). The number of comparisons in the classical Full PC (FPC) is therefore $\frac{a}{2}(a-1)$.

It can be observed that the amount of necessary comparisons grows exponentially with a and makes the tests too long very soon. A lot of effort has been dedicated to decrease this amount while simultaneously preserving the reliability. Some solutions have been proposed already in 1960 by Dykstra [23].

In his paper, several methods enabling complete omitting of certain pairs were designed. Since every stimulus has the same frequency of occurrence in the test, the methods are known as "balanced sub-set methods". Dykstra's procedures are *group divisible designs*, *triangular designs*, *cyclic designs*, and *square designs*. This thesis will mainly exploit the last of the methods.

Square Design PC

In the square design PC methodology (SDPC), the stimuli are arranged into a square matrix. Therefore, it is required that the square root of number of stimuli a is a natural number, i.e.

$$\sqrt{a} \in \mathbb{N}. \tag{2.11}$$

The matrix of the stimuli for $a = 9$ can have the following structure:

$$\begin{matrix} A_1 & A_2 & A_3 \\ A_4 & A_5 & A_6 \\ A_7 & A_8 & A_9 \end{matrix}$$

The pairs for comparison are then created only among the stimuli in the same row and in the same column. It this case therefore: (A_1, A_2, A_3), (A_4, A_5, A_6), (A_7, A_8, A_9), (A_1, A_4, A_7), (A_2, A_5, A_8), and (A_3, A_6, A_9). Since each subset requires 3 comparisons, the total number of comparisons will be $3 \times 6 = 18$ compared to $\frac{9}{2} \times 8 = 36$ in case of FPC. The general formula for the number of necessary comparisons is $a(\sqrt{a} - 1)$.

Nevertheless, omitting certain comparisons will definitely have an influence on the reliability of the final results. To quantify this, Dykstra [23] defined a measure of *efficiency* and proved that the SDPC is a reliable method.

However, Li et al. [24, 25] pointed out that the reliability holds only if no observer errors occur which might be an overly optimistic assumption. Therefore, they came up with an adaptive scenario, ensuring that the reliability remains high even if observers make some mistakes during their evaluations.

Adaptive Square Design PC

Adaptive SDPC methodology (ASDPC) [24, 25] is based on the fact that the same error made when comparing qualitatively distant stimuli has much higher impact on the overall scores than during a comparison of the qualitatively close stimuli. Thus, it is more convenient and proof against observers' mistakes when qualitatively farther pairs are omitted. In other words, closer pairs should be arranged in the same columns and rows. Moreover, decisions regarding the qualitatively closer pairs have higher informative value. Due to the better robustness against the subjects' errors, even higher reliability than in case of classical FPC can be achieved. A different approach towards the more frequent comparing of qualitatively closer pairs was proposed by Silverstein and Farrell [26] who used a binary sorting tree for generation of the pairs.

Li et al. proposed an elegant solution for arranging the stimuli into the matrix as desired. The stimuli are sorted according to their respective quality scores using one of the models from the Section 2.3.3 and then positioned into the matrix spirally as shown in Figure 2.5.

The quality scores are, of course, not known prior to the testing (otherwise it would not have to be done). The initial matrix set up is therefore random or obtained by a pretest or some a priori information. To maximize the robustness of the procedure, quality scores are calculated after the observer has finished

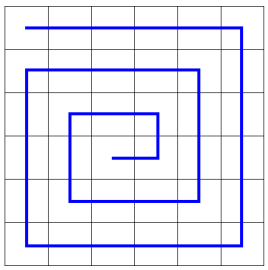


Figure 2.5: Spiral for positioning the stimuli into the matrix in ASDPC methodology.

the evaluation and the matrix is rearranged according to his/her and previous results. This makes the test adaptive and more robust.

2.3.3 Results Processing

This section introduces the ways to interpret the outcomes of the indirect scaling methodologies. The processing of PC results is thoroughly described in [27]. In case of ranking, the procedures are similar.

Firstly, two models for comparative judgment that can be used to transform PCM into the *interval scale* (see Table 2.1) will be introduced. Moreover, the way of getting useful information directly from the PCM will also be demonstrated.

Thurston-Moesteller Model

The first and most popular model for comparative judgment was proposed by L. L. Thurstone [28] and further studied by F. Moesteller [29] and is, therefore, known as Thurstone-Moesteller (TM) model. The model assumes that the decision for stimuli A_i and A_j is based on two Gaussian random variables

$$\xi_i \sim \mathcal{N}(\mu_{\xi_i}, \sigma^2_{\xi_i}),\ \xi_j \sim \mathcal{N}(\mu_{\xi_j}, \sigma^2_{\xi_j}), \tag{2.12}$$

where $\mathcal{N}$ denotes a normally distributed random variable with mean value μ and standard deviation σ. Variables ξ_i and ξ_j correspond to the stimuli A_i and A_j, respectively. Their probability density functions (PDFs) are expressed as

$$pdf_{\xi_i}(x) = \frac{1}{\sigma_{\xi_i}}\phi(\frac{x-\mu_{\xi_i}}{\sigma_{\xi_i}}),\ pdf_{\xi_j}(x) = \frac{1}{\sigma_{\xi_j}}\phi(\frac{x-\mu_{\xi_j}}{\sigma_{\xi_j}}), \tag{2.13}$$

where $\phi(x)$ is the PDF of standard normal distribution $\mathcal{N}(0,1)$ which is defined as

$$\phi(x) = \frac{1}{\sqrt{2\pi}}\exp\Big(-\frac{x^2}{2}\Big). \tag{2.14}$$

An example of PDFs for two stimuli on the quality scale is shown in Figure 2.6.

According to the model, when comparing the stimuli, a realization is drawn from each of the distributions and the stimulus with realization of higher value is selected as qualitatively better. This can also be expressed in terms of difference, i.e. A_i is selected over A_j if $\xi_i - \xi_j$ is positive, thus

$$Pr(\xi_i > \xi_j) = Pr(\xi_i - \xi_j > 0). \tag{2.15}$$

Since both variables are Gaussian, $\xi_i - \xi_j$ is also Gaussian random variable with mean $\mu_{\xi_i\xi_j}$ and standard

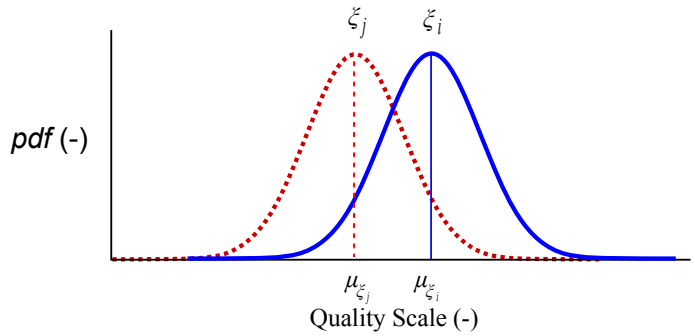


Figure 2.6: PDFs for two stimuli on the quality scale.

deviation $\sigma_{\xi_i\xi_j}$. The equation (2.15) can, therefore, be expressed as

$$Pr(\xi_i > \xi_j) = \int_0^{\infty} \frac{1}{\sqrt{2\pi\sigma^2_{\xi_i\xi_j}}} \exp\Big(\frac{-(x-\mu_{\xi_i\xi_j})^2}{2\sigma^2_{\xi_i\xi_j}}\Big) = \int_{-\mu_{\xi_i\xi_j}}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2_{\xi_i\xi_j}}} \exp\Big(\frac{-x^2}{2\sigma^2_{\xi_i\xi_j}}\Big). \tag{2.16}$$

The symmetry of the Gaussian enables to write

$$Pr(\xi_i > \xi_j) = \int_{-\infty}^{-\mu_{\xi_i\xi_j}} \frac{1}{\sqrt{2\pi\sigma^2_{\xi_i\xi_j}}} \exp\Big(\frac{-x^2}{2\sigma^2_{\xi_i\xi_j}}\Big) = \int_{-\infty}^{-\mu_{\xi_i\xi_j}} \frac{1}{\sigma_{\xi_i\xi_j}} \phi\Big(\frac{x}{\sigma_{\xi_i\xi_j}}\Big). \tag{2.17}$$

Therefore, the probability of selecting A_i over A_j can be expressed in terms of standard normal cumulative distribution function (CDF) Φ as

$$Pr(\xi_i > \xi_j) = \Phi\Big(\frac{\mu_{\xi_i\xi_j}}{\sigma_{\xi_i\xi_j}}\Big), \tag{2.18}$$

where Φ is defined as

$$\Phi(z) = \int_{-\infty}^{z} \phi(x)dx. \tag{2.19}$$

The equation (2.18) can be inverted in order to get the mean quality difference

$$\mu_{\xi_i\xi_j} = \sigma_{\xi_i\xi_j}\Phi^{-1}\big(Pr(\xi_i > \xi_j)\big), \tag{2.20}$$

where Φ^{-1} is the inverse standard normal CDF, also known as the *probit*.

According to Thurstone, the probability $Pr(\xi_i > \xi_j)$ can be estimated from the number of times the stimulus A_i was chosen over A_j during the test ($C_{\xi_i>\xi_j}$), thus

$$Pr(\xi_i > \xi_j) = \frac{C_{\xi_i>\xi_j}}{C_{\xi_i>\xi_j} + C_{\xi_j>\xi_i}}. \tag{2.21}$$

$C_{\xi_j>\xi_i}$ represents the number of cases where observers selected stimulus A_j over A_i. The estimate of mean quality difference can therefore be obtained as

$$\hat{\mu}_{\xi_i\xi_j} = \sigma_{\xi_i\xi_j}\Phi^{-1}\Big(\frac{C_{\xi_i>\xi_j}}{C_{\xi_i>\xi_j} + C_{\xi_j>\xi_i}}\Big), \tag{2.22}$$

which is known as Thurstone's Law of Comparative Judgment.

It can be seen that the value of $\sigma_{\xi_i\xi_j}$ will not be obtained from the PC test. To enable the practical computations, Thurstone proposed several simplifications. In the absolute majority of applications, **Case V model** is used. It assumes that the variables ξ_i and ξ_j are uncorrelated ($\rho_{\xi_i\xi_j} = 0$) and their standard deviations are equal, i.e. $\sigma_{\xi_i} = \sigma_{\xi_j}$. Since the variance $\sigma^2_{\xi_i\xi_j} = \sigma^2_{\xi_i} + \sigma^2_{\xi_j} - 2 \times \rho_{\xi_i\xi_j} \times \sigma_{\xi_i} \times \sigma_{\xi_j}$, it is

convenient to assume (without the loss of generality) that $\sigma^2_{\xi_i} = \sigma^2_{\xi_j} = \frac{1}{2}$. The estimate is then computed as

$$\hat{\mu}_{\xi_i\xi_j} = \Phi^{-1}\Big(\frac{C_{\xi_i>\xi_j}}{C_{\xi_i>\xi_j} + C_{\xi_j>\xi_i}}\Big). \tag{2.23}$$

To obtain the final values on the interval scale, the Least Square Method is used [29]. When a being the number of stimuli, a vector of quality scores $\mu = [\mu_{\xi_1}, \mu_{\xi_2}, \ldots, \mu_{\xi_a}]$ can be defined. An $a \times a$ matrix of the estimates of the mean quality differences D with elements $D(i,j) = \hat{\mu}_{\xi_i\xi_j}$ can then be used for optimization:

$$\hat{\mu} = \arg\min_{\mu\in\mathbb{R}^a} \sum_{i,j} (D_{i,j} - (\mu_{\xi_i} - \mu_{\xi_j}))^2. \tag{2.24}$$

This problem has a simple closed form solution. When μ_{ξ_1} is set to be 0, the individual scores are obtained as

$$\hat{\mu}_{\xi_j} = \sum_{i=1}^{a} \frac{D_{i,1}}{a} - \sum_{i=1}^{a} \frac{D_{i,j}}{a}. \tag{2.25}$$

Bradley-Terry Model

Similar comparative judgment model has been developed by Bradley and Terry [30–32]. Sometimes it is also called Bradley-Terry-Luce (BTL) model after R. D. Luce who extended the model to a multiple choice scenario [33].

Here, the probability of choosing stimulus A_i over A_j is expressed as

$$Pr(\xi_i > \xi_j) = Pr(\xi_i - \xi_j > 0) = \frac{\pi_{\xi_i}}{\pi_{\xi_i} + \pi_{\xi_j}}, \tag{2.26}$$

where $\pi_{\xi_i} = \exp(\mu_{\xi_i}/s)$ with s being a scale parameter. Therefore, the previous equation can be expressed as

$$Pr(\xi_i > \xi_j) = \frac{\exp(\mu_{\xi_i}/s)}{\exp(\mu_{\xi_i}/s) + \exp(\mu_{\xi_j}/s)} = \frac{1}{2} + \frac{1}{2}\tanh\Big(\frac{\mu_{\xi_i} - \mu_{\xi_j}}{2s}\Big). \tag{2.27}$$

From here, it has been noticed that the variable $\xi_i - \xi_j$ is a logistic variable with the mean $\mu_{\xi_i} - \mu_{\xi_j}$ and the scale parameter s. This is the main difference from the TM model, where this variable is considered to be Gaussian. It has been proven by Block and Marschak [34] that in order to fulfil the model's requirements, the variables ξ_i and ξ_j follow Gumbel distribution.

Performing the similar operations as with TM model, the estimate of quality differences can be obtained as

$$\hat{\mu}_{\xi_i\xi_j} = s\Big(\ln\Big(\frac{C_{\xi_i>\xi_j}}{C_{\xi_i>\xi_j} + C_{\xi_j>\xi_i}}\Big) - \ln\Big(1 - \frac{C_{\xi_i>\xi_j}}{C_{\xi_i>\xi_j} + C_{\xi_j>\xi_i}}\Big)\Big). \tag{2.28}$$

Since it is not necessary to compute the error function, the BTL model is computationally less demanding. Nevertheless, this advantage is irrelevant with modern computers.

The estimation of individual quality scores can be obtained by least square method as well. However, it is more common to perform Maximum Likelihood Estimation (MLE) [32]. The likelihood function has a form of

$$\mathcal{L} = \frac{\prod_i \pi_{\xi_i}^{C_{\xi_i}}}{\prod_{i<j} (\pi_{\xi_i} + \pi_{\xi_j})^{N_{\xi_i}}}, \tag{2.29}$$

where C_{ξ_i} is the total number of times where the stimulus A_i was selected in the whole test and N_{ξ_i} is the total number of evaluations where the stimulus A_i was present (during the whole test), i.e. the number of times it could have potentially been chosen.

From $\mathcal{L}$ we can get a simple function for individual scores

$$\hat{\pi}_{\xi_i} = \frac{C_{\xi_i}}{\sum_{i \neq j} N_{\xi_i} (\hat{\pi}_{\xi_i} + \hat{\pi}_{\xi_j})^{-1}}, \tag{2.30}$$

that can be solved iteratively. An effective MATLAB implementation, provided by Wickelmaier and Schmidt [35], can be used to compute the scores from the PCM.

The BTL and TM models were thoroughly compared by J. C. Handley [36], who showed that BTL model is more analytically developed. E.g. it enables a calculation of CI using large sample theory as

$$CI_{\xi_i} = \left[\ln \hat{\pi}_{\xi_i} - z_{\alpha/2} \frac{\sqrt{\psi_{ii}/N}}{\hat{\pi}_{\xi_i}},\ \ln \hat{\pi}_{\xi_i} + z_{\alpha/2} \frac{\sqrt{\psi_{ii}/N}}{\hat{\pi}_{\xi_i}} \right], \tag{2.31}$$

where α is the level of significance ($\alpha = 0.05$ for 95% CI) and ψ_{ii} is a diagonal element of $(a+1) \times (a+1)$ matrix $\hat{\Psi}$, defined as

$$\hat{\Psi} = \begin{bmatrix} \hat{\Lambda} & \mathbf{1} \\ \mathbf{1}' & 0 \end{bmatrix}^{-1}, \tag{2.32}$$

where $\mathbf{1}$ is a vector of ones and $\hat{\Lambda} = [\hat{\lambda}_{ij}]$ where

$$\begin{aligned} \hat{\lambda}_{ii} &= \frac{1}{\hat{\pi}_{\xi_i}} \sum_{i \neq j} \frac{\hat{\pi}_{\xi_j} N_{\xi_i}}{N(\hat{\pi}_{\xi_i} + \hat{\pi}_{\xi_j})^2}, \\ \hat{\lambda}_{ij} &= \frac{-N_{\xi_i}}{N(\hat{\pi}_{\xi_i} + \hat{\pi}_{\xi_j})^2}, \quad i \neq j. \end{aligned} \tag{2.33}$$

There are also several other characteristics, such as goodness of model fit or hypothesis testing, defined for the BTL model [36].

Direct PCM Processing

In some cases, it might not be necessary to put data on the continuous (interval) scale. If the only concern is which of the stimuli in each pair is of better quality or if the quality of the two stimuli is similar, it is possible to obtain this information directly from the PCM.

Every pair in the PCM can be regarded as a 2×2 contingency table as shown in Table 2.3. To determine

$C_{\xi_i > \xi_j}$	$C_{\xi_j > \xi_i}$
$C_{\xi_j > \xi_i}$	$C_{\xi_i > \xi_j}$

Table 2.3: 2×2 contingency table for stimuli A_i and A_j obtained from the PCM.

if the pair is statistically significantly different, Fisher's [37] or Barnard's [38] exact test can be used. Barnard's test is believed to be more powerful in detecting the significance.

Another possibility is to consider the data for each pair to follow a Bernoulli process $B(N, p)$, where N is the sum of all individual counts and p is the probability of success in a Bernoulli trial, which is set to 0.5, considering that, a priori, both options have the same chance of success. The binomial CDF can then be used to determine the critical region for the statistical test.

2.3.4 Factor of Influence Verification

This section describes a procedure to check, if a certain factor has a statistical influence on the evaluation. E.g. if the assessment for male and female observers statistically significantly differ or not. The method is based on permutation test and was proposed by Li et al. [39].

Given the two groups of observers $G = [g_1, g_2, \ldots, g_{N_G}]$ and $H = [h_1, h_2, \ldots, h_{N_H}]$ divided by the factor whose influence is being checked (e.g. gender). The PCM for each group can be obtained from individual observers as

$$PCM_G = \sum_{i=1}^{N_G} PCM_{g_i}, \qquad PCM_H = \sum_{j=1}^{N_H} PCM_{h_j}, \tag{2.34}$$

where PCM_{g_i} is the PCM of the i-th observer from the group G and PCM_{h_j} represents the PCM of the j-th observer from the group H.

Having calculated the two PCMs, a 2×2 contingency table can be created for each pair of stimuli A_i and A_j as where $C_{\xi_i > \xi_j, G}$ is the number of times the stimulus A_i was preferred over A_j in the group G.

$C_{\xi_i > \xi_j, G}$	$C_{\xi_i > \xi_j, H}$
$C_{\xi_j > \xi_i, G}$	$C_{\xi_j > \xi_i, H}$

Table 2.4: 2×2 contingency table for stimuli A_i and A_j obtained from the PCM_G and PCM_H.

Employing Fisher's [37] or Barnard's [38] exact test, the number of cases where the evaluation by the two groups statistically significantly differ can be determined. Significance Ratio SR can be obtained by dividing this number by total number of pairs in the test. If the factor has a significant influence, the SR should significantly differ from the case, where the groups would be divided randomly. The distribution for SR', representing the significance ratio of randomly divided observers, can be obtained by Monte Carlo Simulation. The whole process is described in Algorithm 2.

Algorithm 2 Permutation test for determining the significance of factor's influence.

Define the number of repetitions L (at least $L = 1000$)
e is the number of times where evaluation by the two groups significantly differ
K is the total number of stimuli pairs in the test
while $l < L$ **do**
  Randomly divide observers into groups G' and H'
  Calculate $PCM_{G'}$ and $PCM_{H'}$ according to equation (2.34)
  Calculate $e(l)$ for the groups by Fisher's or Barnard's exact test
  $SR'(l) = e(l)/K$
  $l = l + 1$
end while
Outcome: The distribution for SR'

The factor is considered to be of influence on the given level of significance if the initial value SR is an outlier from the distribution for SR'.

3

Objective Quality Metrics

The subjective quality tests, as discussed in the Chapter 2, have several disadvantages. They are time consuming, expensive, and require an access to a panel of naive (i.e. non-expert) and independent human observers. Moreover, they cannot be used within real time applications for controlling or optimizing the quality. For this reason, objective quality metrics are useful. These are algorithms able to measure the quality automatically. Reliable objective criteria should provide a good correspondence to the results of the subjective tests. Therefore, they always need to be benchmarked with respect to the given application [40].

Note that not all of the algorithms described in this chapter fulfil the mathematical definition of a metric[1] as defined in [41]. Strictly speaking, the term quality index is more appropriate for the majority of the following criteria. However, for the purpose of this thesis, and considering that the terminology commonly used in the literature is not strictly unified, the terms quality index, metric, measure, and model will be considered as equal. No special properties will be implied by the use of a different term.

The objective quality metrics can be divided into number of classes. Probably the most basic classification is according to the metrics' requirements with respect to the reference stimulus. Considering the subject of this thesis, the stimuli in question will always be static images. Full Reference criteria need the whole reference for the assessment. Reduced Reference metrics require only a certain characteristics of the original, such as information about edges, histogram, etc. The last group consists of No Reference or "blind" quality metrics which assess the quality based on the stimulus in question only, without any information about its original form.

Throughout the years, an enormous amount of objective criteria has been proposed. The most popular ones are summarized e.g. in [42]. A number of software packages containing implementations of various metrics are available [19,20,43,44]. The purpose of this chapter is not to provide an exhaustive overview of the existing approaches, nor to describe the metrics in full detail. It is rather to introduce the fundamental principles of the criteria that will be further used in the experimental part of this thesis. The metrics that will be described in this chapter are, together with their type and references to the literature, listed in Table 3.1.

Note that unless directly specified the following equations assume a comparison of the single channel images (i.e. gray-scale). In case of 3-channel images (RGB - according to the red, green, and blue components), metrics can be calculated either for all three channels separately and averaged, or the luminance component of the images can be extracted and compared. The second approach is more common, since the majority of the objective metrics are designed to work with the luminance component only.

[1] According to [41], the metric is defined as "a mathematical object computing the distance of two elements in a metric space." It is always non-negative, symmetrical, equal to zero if and only if the two elements are identical, and it fulfils the triangle inequality. For more information, refer to [41].

Name	*Category*	*Type*	*Reference*
Mean Squared Error (MSE)	Full reference (Section 3.1)	Signal based (Section 3.1.1)	[6]
Peak Signal to Noise Ratio (PSNR)	Full reference (Section 3.1)	Signal based (Section 3.1.1)	[6]
Structural Similarity Index (SSIM)	Full reference (Section 3.1)	HVS based (Section 3.1.2)	[45]
Multi Scale SSIM (MS-SSIM)	Full reference (Section 3.1)	HVS based (Section 3.1.2)	[46]
Information Weighted SSIM (IW-SSIM)	Full reference (Section 3.1)	HVS based (Section 3.1.2)	[47]
Visual Information Fidelity (VIF)	Full reference (Section 3.1)	HVS based (Section 3.1.2)	[48]
Feature Similarity Index (FSIM)	Full reference (Section 3.1)	HVS based (Section 3.1.2)	[49]
Most Apparent Distortion (MAD)	Full reference (Section 3.1)	HVS based (Section 3.1.2)	[50]
Visual Difference Predictor (VDP)	Full reference (Section 3.1)	HVS based (Section 3.1.2)	[51]
High Dynamic Range VDP (HDR-VDP)	Full reference (Section 3.1)	HVS based (Section 3.1.2)	[52]
HDR-VDP-2	Full reference (Section 3.1)	HVS based (Section 3.1.2)	[53]
HDR-VDP-2.2	Full reference (Section 3.1)	HVS based (Section 3.1.2)	[54]
Metric for JPEG	No reference (Section 3.2)	Compression metric (Section 3.2.1)	[55]
Metric for JPEG2000	No reference (Section 3.2)	Compression metric (Section 3.2.1)	[56]
Variance metric	No reference (Section 3.2)	Sharpness metric (Section 3.2.2)	[57]
Frequency Threshold metric	No reference (Section 3.2)	Sharpness metric (Section 3.2.2)	[58]
Gradient based metric	No reference (Section 3.2)	Sharpness metric (Section 3.2.2)	[59]
Laplacian based metric	No reference (Section 3.2)	Sharpness metric (Section 3.2.2)	[59]
Autocorrelation based metric	No reference (Section 3.2)	Blur metric (Section 3.2.2)	[59]
Histogram Frequency metric	No reference (Section 3.2)	Blur metric (Section 3.2.2)	[60]
Kurtosis based metric	No reference (Section 3.2)	Blur metric (Section 3.2.2)	[61]
Marziliano metric	No reference (Section 3.2)	Blur metric (Section 3.2.2)	[62]
HP metric	No reference (Section 3.2)	Sharpness metric (Section 3.2.2)	[63]
Kurtosis of Wavelet Coefficients	No reference (Section 3.2)	Blur metric (Section 3.2.2)	[64]
Riemannian Tensor based metric	No reference (Section 3.2)	Sharpness metric (Section 3.2.2)	[65]
Just Noticeable Blur Metric (JNBM)	No reference (Section 3.2)	Sharpness metric (Section 3.2.2)	[66]
Cumulative Probability of Blur Detection (CPBD)	No reference (Section 3.2)	Sharpness metric (Section 3.2.2)	[67]
S3 metric	No reference (Section 3.2)	Sharpness metric (Section 3.2.2)	[68]
Fast Image Sharpness (FISH)	No reference (Section 3.2)	Sharpness metric (Section 3.2.2)	[69]
Block based FISH ($FISH_{bb}$)	No reference (Section 3.2)	Sharpness metric (Section 3.2.2)	[69]
Weber contrast	No reference (Section 3.2)	Contrast metric (Section 3.2.3)	[70]
Michelson contrast	No reference (Section 3.2)	Contrast metric (Section 3.2.3)	[71]
Root Mean Squared (RMS) contrast	No reference (Section 3.2)	Contrast metric (Section 3.2.3)	[72]
Root Mean Enhancement (RME)	No reference (Section 3.2)	Contrast metric (Section 3.2.3)	[73]
Second Derivative based Measure of Enhancement (SDME)	No reference (Section 3.2)	Contrast metric (Section 3.2.3)	[74, 75]
Global Contrast Factor (GCF)	No reference (Section 3.2)	Contrast metric (Section 3.2.3)	[72]
Color Image Quality Index (CIQI)	No reference (Section 3.2)	Colorfulness metric (Section 3.2.4)	[76]
Color Quality Enhancement (CQE1)	No reference (Section 3.2)	Colorfulness metric (Section 3.2.4)	[73]
Color Quality Enhancement (CQE2)	No reference (Section 3.2)	Colorfulness metric (Section 3.2.4)	[73]
Color Saturation metric	No reference (Section 3.2)	Colorfulness metric (Section 3.2.4)	[77]
Aydın's metric	No reference (Section 3.2)	Aesthetics metric (Section 3.2.5)	[5]
Blind Image Quality Index (BIQI)	No reference (Section 3.2)	Distortion-unaware opinion-aware metric (Section 3.2.6)	[78]
BLIINDS	No reference (Section 3.2)	Distortion-unaware opinion-aware metric (Section 3.2.6)	[79]
BLIINDS-II	No reference (Section 3.2)	Distortion-unaware opinion-aware metric (Section 3.2.6)	[80, 81]
Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE)	No reference (Section 3.2)	Distortion-unaware opinion-aware metric (Section 3.2.6)	[82]
Curvelet based metric	No reference (Section 3.2)	Distortion-unaware opinion-aware metric (Section 3.2.6)	[83]
Natural Image Quality Evaluator (NIQE)	No reference (Section 3.2)	Distortion-unaware opinion-unaware metric (Section 3.2.7)	[84]
Quality Aware Clustering (QAC)	No reference (Section 3.2)	Distortion-unaware opinion-unaware metric (Section 3.2.7)	[85]
CS metric	No reference (Section 3.2)	Distortion-unaware opinion-unaware metric (Section 3.2.7)	[86]

Table 3.1: The list of the metrics described in this chapter.

3.1 Full Reference Metrics

As already stated, the full reference metrics require the presence of the whole original image. The majority of this type of criteria measure a "fidelity" of the processed version to the reference, i.e. how similar the two versions are. The assumption is that the reference is of the best possible quality. This class is the oldest and the most developed one. The full reference quality assessment algorithms provide the most reliable estimates of quality. However, there is a number of applications where the original (best quality) version is not available or known. In these cases, these metrics cannot be exploited.

3.1.1 Signal Based Metrics

The simplest criteria come from the theory of signals. They are also known as "pixel based" metrics, since they compare the original and processed images on the level of pure pixel intensity values. They are very popular for their simplicity, almost no computational requirements, and clear mathematical definition. However, they mostly do not correspond well with the subjective quality tests results, since human observers do not perceive quality as the pixel differences. Moreover, they require the two versions to be perfectly aligned, otherwise they produce very low quality estimates.

MSE

The first signal based metric is Mean Squared Error (MSE), defined as

$$MSE = \frac{1}{X \times Y} \sum_{x=1}^{X} \sum_{y=1}^{Y} (I_{\mathrm{R}}(x,y) - I_{\mathrm{P}}(x,y))^2, \tag{3.1}$$

where I_{R} is the reference image, I_{P} is its processed version, and X and Y represent the image width and height in pixels, respectively.

Since it measures the error, higher values of MSE represent lower quality.

PSNR

The most popular pixel based metric is called Peak Signal to Noise Ratio (PSNR) and is mostly expressed in decibels (dB). It is defined as

$$PSNR = 10 \log_{10} \frac{(2^{bits} - 1)^2}{MSE}, \tag{3.2}$$

where $bits$ stands for the number of bits used to express each pixel intensity value in the image, i.e. $bits = 8$ for 8-bit images.

3.1.2 Human Visual System Based Metrics

The rest of the full reference metrics that will be introduced in this thesis attempt to exploit certain properties of HVS. This enables them to better estimate the quality in terms of correspondence to the subjective results.

SSIM, MS-SSIM, IW-SSIM

Probably the most popular metric of the recent years is Structural Similarity Index (SSIM) [45]. It compares the images in terms of three bases – luminance, contrast, and structure. Practically, the similarity is calculated in small patches (11×11) pixels, creating a similarity map. The similarity in each patch u is

computed as

$$SSIM(u) = \left[l\Big(I_{\mathrm{R}}(u), I_{\mathrm{P}}(u)\Big)\right]^{\lambda_1} \left[c\Big(I_{\mathrm{R}}(u), I_{\mathrm{P}}(u)\Big)\right]^{\lambda_2} \left[s\Big(I_{\mathrm{R}}(u), I_{\mathrm{P}}(u)\Big)\right]^{\lambda_3}, \tag{3.3}$$

where λ_1, λ_2, and λ_3 are the parameters influencing the relative strength of the particular similarities. They are mostly set to be $\lambda_1 = \lambda_2 = \lambda_3 = 1$. The individual similarities are defined as

$$l\Big(I_{\mathrm{R}}(u), I_{\mathrm{P}}(u)\Big) = \frac{2 \times \overline{I_{\mathrm{R}}(u)} \times \overline{I_{\mathrm{P}}(u)} + const_1}{\overline{I_{\mathrm{R}}(u)}^2 + \overline{I_{\mathrm{P}}(u)}^2 + const_1}, \tag{3.4}$$

$$c\Big(I_{\mathrm{R}}(u), I_{\mathrm{P}}(u)\Big) = \frac{2 \times \mathrm{std}\Big(I_{\mathrm{R}}(u)\Big) \times \mathrm{std}\Big(I_{\mathrm{P}}(u)\Big) + const_2}{\mathrm{std}\Big(I_{\mathrm{R}}(u)\Big)^2 + \mathrm{std}\Big(I_{\mathrm{P}}(u)\Big)^2 + const_2}, \tag{3.5}$$

$$s\Big(I_{\mathrm{R}}(u), I_{\mathrm{P}}(u)\Big) = \frac{\mathrm{std}\Big(I_{\mathrm{R}}(u), I_{\mathrm{P}}(u)\Big) + const_3}{\mathrm{std}\Big(I_{\mathrm{R}}(u)\Big) \times \mathrm{std}\Big(I_{\mathrm{P}}(u)\Big) + const_3}, \tag{3.6}$$

where $const_1$, $const_2$, and $const_3$ are small constants for ensuring numerical stability, operators $\overline{(.)}$ and $\mathrm{std}(.)$ calculate the mean and the standard deviation in the given patch and

$$\mathrm{std}\Big(I_{\mathrm{R}}(u), I_{\mathrm{P}}(u)\Big) = \frac{1}{U-1} \sum_{i=1}^{U} \Big(I_{\mathrm{R}}(u,i) - \overline{I_{\mathrm{R}}(u)}\Big)\Big(I_{\mathrm{P}}(u,i) - \overline{I_{\mathrm{P}}(u)}\Big), \tag{3.7}$$

where U is the number of pixels in the patch u. The overall SSIM index can be calculated as an average of values obtained for each patch. A thorough analysis of the SSIM index, including its relation to MSE, can be found in [87].

Since the index is calculated for a single scale, it only covers single set of conditions, i.e. only single viewing distance. To incorporate more conditions, Wang et al. introduced multiple scales into the SSIM index calculation [46]. The resulting metric always obtains images on lower scales by low-pass filtering and downsampling with the factor of 2. The Multi Scale SSIM (MS-SSIM) for the patch u is therefore computed as

$$MS\text{-}SSIM(u) = \left[l\Big(I_{\mathrm{R}}(u), I_{\mathrm{P}}(u)\Big)\right]^{\lambda_{1,M}} \prod_{m=1}^{M} \left[c_m\Big(I_{\mathrm{R}}(u), I_{\mathrm{P}}(u)\Big)\right]^{\lambda_{2,m}} \left[s_m\Big(I_{\mathrm{R}}(u), I_{\mathrm{P}}(u)\Big)\right]^{\lambda_{3,m}}, \tag{3.8}$$

where M is the number of scales on which the index is calculated. A typical value of M is 5. The parameters in the individual scales can be obtained e.g. from the contrast senstivity function (CSF) of HVS. However, in the original paper, the parameters are calibrated on real scenes. It has been shown that introducing multiple scales into the SSIM improves the correspondence with subjective data.

Another possible improvement is to employ a more sophisticated "pooling" strategy, i.e. the method for obtaining a single value from the similarity map. There are several options such as assigning higher weights to the areas with more severe distortions, weighting according to the saliency (regions of interest), object segmentation, etc. Wang and Li [47] proposed to enhance the performance of SSIM index by using information based pooling strategy.

The procedure firstly quantifies the information that can be extracted from the original and processed image by computing the mutual information between the images and their representations after transition through the perceptual channel. If the information extracted from the reference is $\mathcal{I}_R$ and from the processed

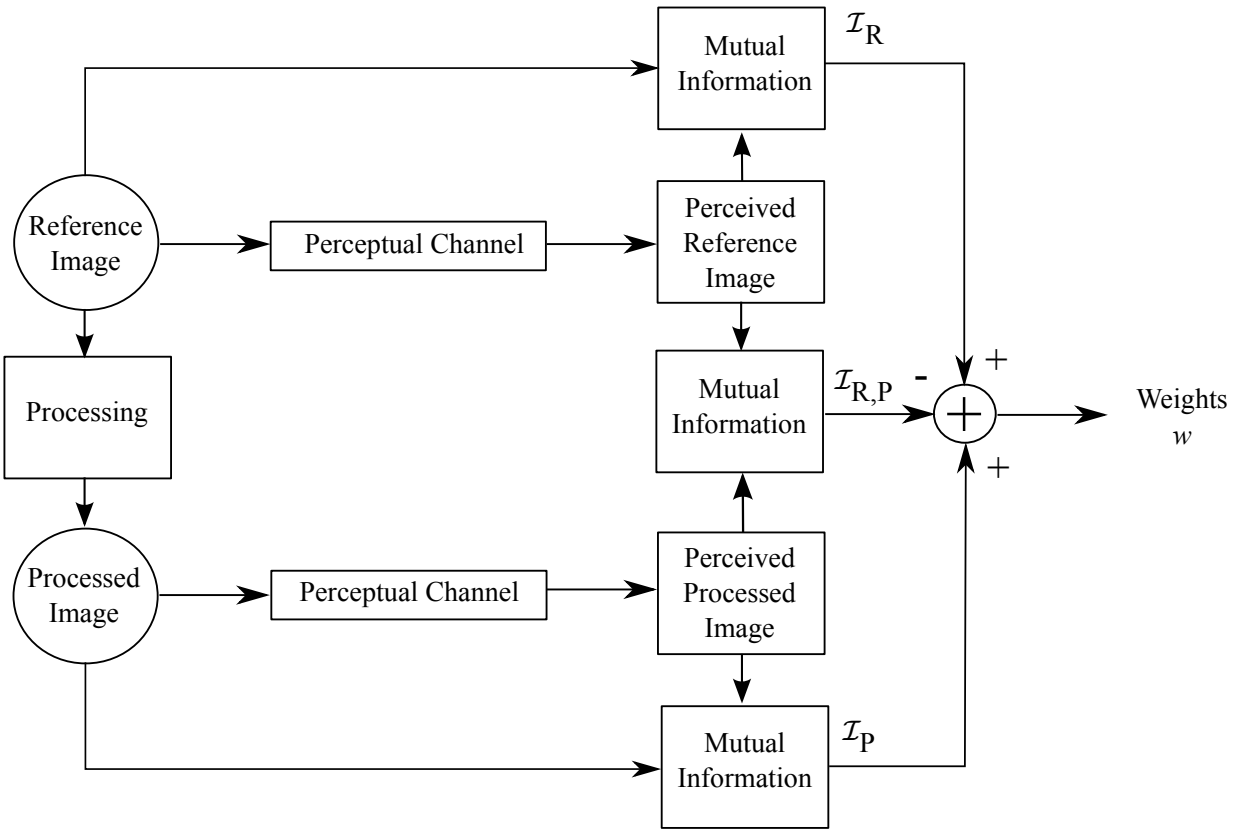


Figure 3.1: The calculation of weights in IW-SSIM [47].

image $\mathcal{I}_P$, the weights are then obtained as

$$w = \mathcal{I}_{\mathrm{R}} + \mathcal{I}_{\mathrm{P}} - \mathcal{I}_{\mathrm{R,P}}, \tag{3.9}$$

where $\mathcal{I}_{R,P}$ is the mutual information between the reference and processed image after transition through the perceptual channel. The illustration can be found in Figure 3.1.

The mutual information is calculated using Gaussian Scale Mixture (GSM) model based on the assumption that the probability density of a pixel is fully determined by its neighbors. The final index is computed on multiple scales as

$$IW\text{-}SSIM_m = \frac{\sum_i w_{i,m} c_m\Big(I_{\mathrm{R},i}, I_{\mathrm{P},i}\Big) s_m\Big(I_{\mathrm{R},i}, I_{\mathrm{P},i}\Big)}{\sum_i w_{i,m}}, \tag{3.10}$$

for $m = 1, 2, \ldots, M-1$, where i is the spatial location of a pixel in the patch. For $m = M$, also the luminance component is included, thus

$$IW\text{-}SSIM_M = \frac{1}{V} \sum_i l\Big(I_{\mathrm{R},i}, I_{\mathrm{P},i}\Big) c_M\Big(I_{\mathrm{R},i}, I_{\mathrm{P},i}\Big) s_M\Big(I_{\mathrm{R},i}, I_{\mathrm{P},i}\Big), \tag{3.11}$$

where V is the number of patches in the scale. The final index is then obtained as

$$IW\text{-}SSIM = \prod_{m=1}^{M} (IW\text{-}SSIM_m)^{\lambda_m}. \tag{3.12}$$

The values of λ are the same as in the case of MS-SSIM.

VIF

The above described information based pooling actually comes from the idea of Visual Information Fidelity (VIF) metric [48] and its predecessor Information Fidelity Criterion (IFC) [88]. Here, NSS model is used for modelling reference image's sub-band filtered coefficients. Every scalar coefficient is expressed as a random variable employing GSM model.

The informations $\mathcal{I}_{\mathrm{R}}$ and $\mathcal{I}_{\mathrm{P}}$ (see Figure 3.1) are compared in order to quantify the fidelity of the processed version to the original. The metric employs a steerable pyramid [89] for decomposition, thus the results are calculated in different scales and orientations.

A very interesting feature of VIF is its ability to recognize if the processed version is of the superior quality. In this case, the final value of VIF index is higher than one. However, this works only for images with increased contrast without an addition of noise. Nevertheless, this is the first step towards incorporating quality assessment of enhanced images into objective metrics.

FSIM

Feature Similarity Index (FSIM), proposed by Zhang et al. [49], is based on extracting and comparing low level features from reference and processed image. Namely the similarities in Phase Congruency [90] and Gradient are combined. The two features are complementary. FSIM is also one of the few metrics which have been extended to consider the information about color. In the $\mathrm{FSIM}_{\mathrm{C}}$, the images are firstly transferred to the YIQ colorspace. The classical FSIM is then calculated on the first channel and combined with similarities in the other two channels.

Another extension of the metric has been proposed by considering high level features as well [91]. When observing an image, the attention is driven by bottom-up and top-down mechanisms [92]. The first is influenced by the low-level features (such as contrast, phase congruency, etc.) while the latter is dependent on the task, observer's experience, and other high level factors. These features are much harder to be quantified. To consider both low and high level features, the metric can incorporate data from eye-tracking experiments. Although including such information was able to bring some improvement to the quality assessment, the difference in performances was not found statistically significant [93]. This suggests that, in case of static images, considering low level features only might be sufficient for the quality evaluation.

MAD

Another full-reference metric to be described is called Most Apparent Distortion (MAD) and was proposed by E. C. Larson and D. M. Chandler [50]. It is based upon a premise that HVS uses two ways of judging the quality according to the degree of distortion.

When the image is only slightly distorted, observers tend to search for distortions. Authors call this "detection-based strategy" and model it by combining the local masking model in spatial domain with local mean squared error calculated in the perceived luminance domain. Low-level properties of HVS (contrast sensitivity, non-linear perception of luminance, luminance and contrast masking) are combined into a map of locations of visible distortions. Using this map, the visibility-weighted MSE map is calculated and collapsed into the single scalar value by L_2 norm.

In case of a heavily distorted image, HVS does not have to look for the distortions, because they are dominant. That is why "detection-based strategy" is substituted by "appearance-based strategy". Here, the distortion is expressed as an extent to which the appearance of the image's subject matter is degraded. This is computed by calculating local statistics of multi-scale log-Gabor filter responses. The image is decomposed by log-Gabor filters with four orientations on five different scales. Block-based statistics, such as variation, skewness, and kurtosis, are obtained from every decomposition. These are then combined into the statistical difference map which is then again collapsed into single scalar number.

In the final calculation of the MAD index, both above described strategies are combined as

$$MAD = (d_{\text{detect}})^{\lambda}(d_{\text{appear}})^{(1-\lambda)}, \tag{3.13}$$

where d_{detect} and d_{appear} denote the values obtained by the "detection-based strategy" and "appearance-based strategy", respectively, and λ is weight chosen according to overall level of distortion. For heavily distorted images, λ should be close to 1.

Optimal procedure for selection of λ is still not defined but authors reached good results with λ calculated from d_{detect} as

$$\lambda = \frac{1}{1 + \gamma_1 (d_{\text{detect}})^{\gamma_2}}, \tag{3.14}$$

where parameters γ_1 and γ_2 are free. Optimal values for A57 database [94], are $\gamma_1 = 0.467$ and $\gamma_2 = 0.130$.

MAD is also usable on the luminance component of the image only and particularly calculation of d_{appear} is computationally demanding and has severe memory requirements.

VDP, HDR-VDP, HDR-VDP-2, HDR-VDP-2.2

Visible Difference Predictor [51] is an algorithm using model of HVS to predict where the physical difference between the original and processed version of the image is visible for human observers. It should be noted that the metric did not attempt to estimate the perceived quality but to quantify the visibility of differences. It has been extended for HDR Images by Mantiuk et al. [52].

In [53], the algorithm was significantly revised, resulting into HDR-VDP-2. The metric can be used for both SDR and HDR images. The HVS model includes simulation of optical retinal pathway with intra-ocular light scatter, photoreceptor spectral sensitivity, luminance masking, and achromatic response followed by multi-scale decomposition and neural noise model containing neural CSF and contrast masking. It requires setting of several parameters about display and viewing conditions in order to provide an estimate for the particular usage situation.

As the result a visibility map showing the probabilities of a difference detection is generated. A pooling strategy is defined in order to obtain a single value for difference visibility, as well as an estimate of MOS – i.e. image quality. The most recent version of the metric – HDR-VDP-2.2 [54] – has been calibrated on larger and more representative set of SDR and HDR images in order to provide more reliable quality predictions.

It has been shown (e.g. [95–99]) that HDR-VDP-2.2 performs well for images and videos both within and across different contents. On the other hand, its computational requirements are very high, especially for high resolution contents.

3.2 Reduced Reference and No Reference Metrics

All of the previously described metrics require the whole original image to assess the quality of the processed version. This can represent a strong limitation in certain applications. If that is the case, reduced reference and, more importantly, no reference (also know as blind) metrics are required.

Reduced reference metrics represent a compromise between full and no reference algorithms. They do not need all of the information about the reference to be present but only a certain set of its features or characteristics [100, 101]. Such information can be the edge map, entropy, histogram, etc. This approach is the least developed from the three and there are no reduced reference criteria used in the experimental parts of the thesis, since the application areas in question are mainly covered by full and no reference metrics. However, this short discussion is provided for the sake of completeness.

No reference metrics are gaining more popularity over the recent years for their universality. Moreover, HVS is, to some extent, capable of evaluating the quality of images without comparing it to the reference. A great step for the blind image quality criteria was definitely the progress in machine learning. They

can be roughly divided in terms of specialization, i.e. if they are designed specifically for certain distortion/processing/image characteristics. Such algorithms are sometimes called "distortion-aware". Another classification is in terms of the metrics' tuning. If the criterion has been tuned on the subjective data from certain database(s), it is classified among "opinion-aware" algorithms. Using any distortion-aware and/or opinion-aware metrics out of their designated application does not have to provide reliable results. Following sections will introduce some concepts from all of the previously mentioned classes.

3.2.1 Compression Metrics

The metrics in this section were designed specifically for evaluating compressed images. Considering that each compression algorithm results in different artifacts with a different impact on the final perceived quality, the criteria for quantifying this impact should vary as well. The advantage in designing such criteria is that even though there is no reference, the distortions are mostly well known and the criteria can, therefore, be tuned to detect them.

Metric for JPEG

Arguably the most popular lossy image compression scheme is JPEG.[2] A simple blind quality criterion for JPEG compressed images was proposed by Wang et al. [55]. They propose to measure blockiness and blurriness of the image by the following simple criteria.

Considering the block size in the JPEG encoder is 8×8 pixels, the blockiness can be quantified by differences on the block boundaries, i.e.

$$\begin{aligned} Bl_{\text{hor}} &= \frac{1}{X \times (\lfloor Y/8 \rfloor - 1)} \sum_{i=1}^{X} \sum_{j=1}^{\lfloor Y/8 \rfloor - 1} |d_{\text{hor}}(i, 8j)|, \\ Bl_{\text{ver}} &= \frac{1}{(\lfloor X/8 \rfloor - 1) \times Y} \sum_{i=1}^{\lfloor X/8 \rfloor - 1} \sum_{j=1}^{Y} |d_{\text{ver}}(8i, j)|, \end{aligned} \tag{3.15}$$

where $\lfloor . \rfloor$ is the floor operator (i.e. rounding to the previous integer value), X and Y stand for the width and height of the image, and

$$\begin{aligned} d_{\text{hor}}(x, y) &= I(x, y+1) - I(x, y), \\ d_{\text{ver}}(x, y) &= I(x+1, y) - I(x, y). \end{aligned} \tag{3.16}$$

The overall blockiness is then expressed as

$$Bl = \frac{Bl_{\text{hor}} + Bl_{\text{ver}}}{2}. \tag{3.17}$$

The blurriness is quantified by two criteria – average absolute difference between in-block image samples and zero-crossing rate. The first one is defined as

$$\begin{aligned} AAD_{\text{hor}} &= \frac{1}{7} \left[\frac{8}{X \times (Y-1)} \sum_{i=1}^{X} \sum_{j=1}^{Y-1} |d_{\text{hor}}(i, j)| - Bl_{\text{hor}} \right], \\ AAD_{\text{ver}} &= \frac{1}{7} \left[\frac{8}{(X-1) \times Y} \sum_{i=1}^{X-1} \sum_{j=1}^{Y} |d_{\text{ver}}(i, j)| - Bl_{\text{ver}} \right], \end{aligned} \tag{3.18}$$

and thus

$$AAD = \frac{AAD_{\text{hor}} + AAD_{\text{ver}}}{2}. \tag{3.19}$$

[2] https://jpeg.org/ (retrieved on 30/08/2016)

The zero-crossing rate can be calculated from

$$\begin{aligned} ZC_{\text{hor}} &= \frac{1}{X\times(Y-2)} \sum_{i=1}^{X} \sum_{j=1}^{Y-2} zc_{\text{hor}}, \\ ZC_{\text{ver}} &= \frac{1}{(X-2)\times Y} \sum_{i=1}^{X-2} \sum_{j=1}^{Y} zc_{\text{ver}}, \end{aligned} \tag{3.20}$$

where

$$\begin{aligned} zc_{\text{hor}}(x,y) &= \begin{cases} 1 & \text{if there is a zero-crossing at } d_{\text{hor}}(x,y) \\ 0 & \text{otherwise,} \end{cases} \\ zc_{\text{ver}}(x,y) &= \begin{cases} 1 & \text{if there is a zero-crossing at } d_{\text{ver}}(x,y) \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \tag{3.21}$$

The overall ZC is then again

$$ZC = \frac{ZC_{\text{hor}} + ZC_{\text{ver}}}{2}. \tag{3.22}$$

The final metric's score is then obtained as

$$Score_{\text{JPEG}} = \lambda_1 + \lambda_2 Bl^{\lambda_3} + AAD^{\lambda_4} + ZC^{\lambda_5}. \tag{3.23}$$

The combination parameters obtained by non-linear regression from the subjective data are: $\lambda_1 = -245.9$, $\lambda_2 = 261.9$, $\lambda_3 = -0.0240$, $\lambda_4 = 0.0160$, and $\lambda_5 = 0.0064$.

Metric for JPEG2000

Sheikh et al. [56] proposed a metric designed for JPEG2000 (or any other wavelet based) compression. Unlike in the case of regular JPEG, the artifacts introduced by these types of compression do not occur so regularly (such as blockiness) but are much more content dependent and occur mostly around strong edges. The metric employs a NSS model in the wavelet domain. More specifically, it is based on the fact that quantization of coefficients in the wavelet domain results in pushing the coefficients on finer scales towards zero.

Therefore, the amplitude of coefficients and their linear predictions from the coefficients neighboring in scale, space, and orientation are taken as quality features. The metric is computed across 6 subbands – vertical, horizontal, and diagonal on the two finest scales. Considering the content dependency, the threshold for the coefficients amplitudes is calculated from estimated mean of the subband with an offset which was found from the training set.

3.2.2 Blur / Sharpness Metrics

Great effort has been dedicated to reference-free measurement of sharpness and blur. A considerable number of such metrics is implemented in the quality assessment framework developed by Murthy and Karam [44].

Variance

This simple criterion was proposed by Erasmus and Smith in [57] as a measure for automatic focusing and astigmatism correction. It assumes that sharper image has higher variance of the pixel values than the blurred one. It is calculated as

$$VAR = \frac{1}{X \times Y} \sum_{x=1}^{X} \sum_{y=1}^{Y} \left(I(x,y) - \overline{I} \right)^2. \tag{3.24}$$

Frequency Threshold

In [58], Firestone et al. proposed another sharpness measure, based on thresholding in the frequency domain. The threshold is found experimentally and the magnitudes of all frequency components above are summed. In the framework [44], the thresholds are determined as a quarter of the number of frequency components in both dimensions.

Gradient, Laplacian

Calculation of gradient or Laplacian as a measure of sharpness was first used by Batten in his master thesis [59]. First, the gradient in a vertical ($grad_{\text{ver}}$) and a horizontal ($grad_{\text{hor}}$) direction is calculated. Then the metric value is obtained as

$$GRAD = \frac{1}{X \times Y} \sum_{x=1}^{X} \sum_{y=1}^{Y} \sqrt{grad_{\text{ver}}(x,y)^2 + grad_{\text{hor}}(x,y)^2}, \tag{3.25}$$

The procedure for the Laplacian is similar. The image is filtered by Laplacian kernel – filtered image I_{Lap} is obtained. Final index is then

$$LAP = \frac{1}{X \times Y} \sum_{x=1}^{X} \sum_{y=1}^{Y} I_{\text{Lap}}(x,y)^2. \tag{3.26}$$

Laplacian provides more accurate results but is also more sensitive to noise.

Autocorrelation

Autocorrelation based metric was also proposed by Batten [59]. The premise is that when the edges are steep, the correlation between neighboring pixels is low. Autocorrelation is calculated for two different distances along the vertical and horizontal direction. The differences are then summed into a final value which is higher for sharper images.

Histogram Frequency

This criterion was developed by Marichal et al. [60]. It is based on Discrete Cosine Transform (DCT). DCT is performed on every block of 8×8 pixels. The histogram of values in every block is then calculated. The metric counts the number of zeros in the histogram and weights them according to their position (closer to the diagonal means higher weight). This metric's value is lower for sharper images.

Kurtosis

Employing kurtosis for measuring image sharpness was proposed by Zhang et al. [61]. They consider the spectral density function to be a two-dimensional PDF of a bivariate random vector. A narrow distribution has a high kurtosis. That means that the more an image is blurred the narrower is its spectral density and the kurtosis is therefore higher. The calculation of kurtosis has already been defined in equation (2.8).

Marziliano

The metric proposed by Marziliano et al. [62], originally developed for application on JPEG2000, uses filtering by vertical Sobel kernel to localize the edges. Then, every row of the image is scanned in order to find pixels corresponding to an edge location. For these pixels, the start and the end position of an edge is found as local extrema (maximum and minimum) closest to the edge. The difference between the start and the end position is marked as an edge width. The metric is calculated as an average edge width over the

whole image. Note that the metric was designed for a measuring of blur, therefore the sharper the image the lower the metric.

HP

The HP metric was proposed by Shaked and Tastl from Hewlett Packard Laboratories [63]. First, the image is filtered by a band-pass filter and the output is thresholded to extract useful features from the image. These features are then filtered by high-pass and band-pass filters and the ratio of their outputs is calculated. The ratio is higher for sharper images because of the presence of more high frequency components in the spectrum.

Kurtosis of Wavelet Coefficients

Method proposed by Ferzli et al. [64] was especially designed for a measurement of sharpness of noisy images. It employs 3-level two-dimensional discrete dyadic wavelet transform to evaluate the image without the presence of noise. After the decomposition, sharpness metric similar to previously described Zhang's Kurtosis [61] is applied to the horizontal and vertical band. A final value is obtained by averaging outputs of the two metrics.

Riemannian Tensor

Another approach developed by Ferzli and Karam is based on Riemannian Tensor [65]. Its background is described in [102] and [103]. The image and image feature space is viewed as Riemannian manifolds in a higher dimensional space. The image manifold is the surface formed by the graph of the image. It has non-Euclidean coordinates and the distance measure is defined as

$$\begin{aligned} ds^2 &= dx^2 + dy^2 + dI^2 \\ &= dx^2 + dy^2 + (I_x dx + I_y dy)^2 \\ &= (1 + I_x^2)dx^2 + 2I_x I_y dx dy + (1 + I_y^2)dy^2, \end{aligned} \tag{3.27}$$

where I_x and I_y are the differences with respect to x and y, respectively. The Riemannian Tensor metric is then

$$r_{xy} = \det \begin{bmatrix} 1 + I_x^2 & I_x I_y \\ I_x I_y & 1 + I_y^2 \end{bmatrix}. \tag{3.28}$$

Final index is obtained as

$$RT = \frac{1}{X \times Y} \sum_{x=1}^{X} \sum_{y=1}^{Y} |r_{xy}|. \tag{3.29}$$

JNBM

The speciality of Just Noticeable Blur Metric (JNBM) developed by Ferzli and Karam [66] is that, unlike most of the others, it can measure a relative amount of blur across different content. It is based on the Just Noticeable Blur (JNB) concept described in the corresponding paper.

At the beginning, an edge detection is performed using Sobel operator. Then the number of edge pixels (edgels) in every 64×64 block is calculated. If the number is higher than the threshold (in the experimental part set to be 0.2% of a total number of pixels in the block), the block is labelled as an edge block (u_{edge}). In these blocks, a local contrast is estimated and JNB width ω_{JNB} is obtained. The edge width $\omega(edg_i)$ is then calculated for all the edges and block distortion $BD_{u_{\text{edge}}}$ is computed as

$$BD_{u_{\text{edge}}} = \left(\sum_{edg_i \in u_{\text{edge}}} \left| \frac{\omega(edg_i)}{\omega_{\text{JNB}}(edg_i)} \right|^{\lambda} \right)^{\frac{1}{\lambda}}. \tag{3.30}$$

Overall block distortion is then

$$BD = \left(\sum_{u_{\text{edge}}} \left| BD_{u_{\text{edge}}} \right|^{\lambda} \right)^{\frac{1}{\lambda}}, \tag{3.31}$$

where $\lambda = 3.6$.

Finally, the final score is obtained as

$$JNBM = \frac{U_{\text{edge}}}{BD}, \tag{3.32}$$

where U_{edge} is the total number of edge blocks.

CPBD

The Cumulative Probability of Blur Detection (CPBD) metric, developed by Narvekar and Karam [67], is an augmentation of above described JNBM [66]. The procedure is the same but no distortion is calculated. Instead, the probability of blur detection Pr_{BLUR} is estimated as

$$Pr_{\text{BLUR}} = 1 - \exp\left(- \left| \frac{\omega(edg_i)}{\omega_{\text{JNB}}(edg_i)} \right|^{\lambda} \right). \tag{3.33}$$

If $\omega(edg_i) = \omega_{\text{JNB}}(edg_i)$, then $Pr_{\text{BLUR}} = Pr_{\text{JNB}} = 63\%$. The metric value is obtained as

$$CPBD = \sum_{Pr_{\text{BLUR}}=0}^{Pr_{\text{BLUR}}=Pr_{\text{JNB}}} pdf(Pr_{\text{BLUR}}), \tag{3.34}$$

where $pdf(P_{\text{BLUR}})$ denotes the value of the PDF at given Pr_{BLUR}.

S3

The S3 algorithm was developed by Vu and Chandler [68]. It measures the perceived sharpness in the image, divided into blocks. As a result, it produces a map of the perceived sharpness. The algorithm has two stages.

The first stage is based upon the slope of the spectrum. The slope ν^* is calculated as

$$\nu^* = \arg\min_{\nu} ||\beta\theta^{-\nu} - \mathbf{s}(\theta)||_2, \tag{3.35}$$

where θ denotes the radial frequency, $\mathbf{s}(\theta)$ is the total magnitude spectrum across all the orientations, β is the scaling factor, and operator $||.||_2$ stands for the L-2 norm taken over all radial frequencies.

From the slope, the first part of the metric $S1(u)$ for the block u is obtained by

$$S1(u) = 1 - \frac{1}{1 + e^{\tau_1(\nu^* - \tau_2)}}, \tag{3.36}$$

where $\tau_1 = -3$ and $\tau_2 = 2$. This attempts to model the HVS tuning to the spectrum of natural scenes. Perceived sharpness drops slowly for the ν from 0 to 1, faster from 1 to 3 and then saturates. Details could be found in the respective paper.

For calculation of $S1(u)$, division to blocks of 32×32 pixels is used with the overlap of 24 pixels. Altogether, $S1(u)$ values create a map $S1(I)$ for the whole image I.

The second stage of the algorithm is measuring the sharpness in the spatial domain. It is based on calculation of difference

$$d(u_{\text{sub}}) = \sum_{i,j} |u_i - u_j|, \tag{3.37}$$

where u_{sub} is a 2×2 subblock of block u.

The second part of the metric is then obtained as

$$S2(u) = \arg\max_{u_{\text{sub}} \in u} d(u_{\text{sub}}). \tag{3.38}$$

The block size for this calculation is set to be 8×8 without any overlap.

These two metrics are then combined to create the final map

$$S3(u) = S1(u)^{\lambda} \times S2(u)^{-\lambda}. \tag{3.39}$$

The authors recommend to use λ parameter equal to 0.5.

For the S3 index, representing the sharpness of the image in a single value, authors in [68] suggest to take the highest value from the $S3(I)$. However, in the *Readme* file accompanying their implementation they state that the average of 1% of highest values should be calculated. Given the higher robustness against outlying values, this approach is employed.

FISH, FISH$_{\text{bb}}$

Vu and Chandler [69] also proposed another Fast Image Sharpness metric (FISH) based on wavelet decomposition. Firstly, an image is decomposed by Cohen-Daubechies-Fauraue 9/7 filters [104] into three subbands (sb_{LH}, sb_{HL}, and sb_{HH}) on three levels. Then, to quantify the amount of high frequency components in the image, log-energy $logE$ is calculated for each subband on each level as

$$\begin{aligned} logE_{\text{LH},m} &= \log_{10}\left(1 + \tfrac{1}{NC_m}\textstyle\sum_{i,j} sb^2_{\text{LH},m}(i,j)\right), \\ logE_{\text{HL},m} &= \log_{10}\left(1 + \tfrac{1}{NC_m}\textstyle\sum_{i,j} sb^2_{\text{HL},m}(i,j)\right), \\ logE_{\text{HH},m} &= \log_{10}\left(1 + \tfrac{1}{NC_m}\textstyle\sum_{i,j} sb^2_{\text{HH},m}(i,j)\right), \end{aligned} \tag{3.40}$$

where m denotes the level and NC_m is the total number of coefficients on the m-th level. The total energy on each level is obtained as

$$E_{\text{total},m} = (1-\lambda)\frac{logE_{\text{LH},m} + logE_{\text{HL},m}}{2} + \lambda\, logE_{\text{HH},m}. \tag{3.41}$$

Parameter λ was set empirically to 0.8 in order to give higher importance to the HH subband. The final sharpness index is defined as

$$FISH = \sum_{m=1}^{3} 2^{3-m} E_{\text{total},m}. \tag{3.42}$$

There is also a possibility to calculate a sharpness map by dividing the image into 16×16 blocks with 50% overlap. The decomposition than follows the procedure described in [105] and the same process as for regular FISH is applied. The index obtained from the sharpness map is called block-based FISH (FISH$_{\text{bb}}$) and is computed, similarly to S3 algorithm, as an average of 1% of highest values in the map.

3.2.3 Contrast Metrics

This section is dedicated to the metrics specialized on measuring image contrast. Perceived contrast is one of the key elements contributing to the overall perceived quality and it is, therefore, beneficial to be able to reliably measure it. Each of the following metrics firstly estimate the local contrast in a specific pixel neighborhood resulting in a map of local contrast. This map is then averaged to provide a single value representing the overall contrast of the image.

Weber

The first measure is based on the Weber's law which is defined for uniform background. It was formulated as a measure of enhancement by Agaian et al. [70] and is calculated as

$$Weber = \frac{1}{U}\sum_{u=1}^{U} 20\log\frac{I_{\max}(u)}{I_{\min}(u)}, \tag{3.43}$$

where U is the number of blocks an image is divided to, $I_{\max}(u)$ and $I_{\min}(u)$ are the maximal and the minimal intensity value within the block u.

Michelson

The second contrast measure employs a concept of Michelson contrast, adjusted by Agaian et al. [71]. It is optimal for periodic background. The computation is

$$Michelson = -\frac{1}{U}\sum_{u=1}^{U} 20\log\frac{I_{\max}(u) - I_{\min}(u)}{I_{\max}(u) + I_{\min}(u)}. \tag{3.44}$$

RMS, RME

Another widely exploited concept is Root Mean Squared (RMS) contrast [72]. It is defined as

$$RMSC = \sqrt{\frac{1}{U}\sum_{u=1}^{U}\left(I_{\text{center}}(u) - \overline{I(u)}\right)^2}, \tag{3.45}$$

where $\overline{I(u)}$ is the mean intensity of the block u and $I_{\text{center}}(u)$ is the central pixel of the block (i.e. the pixel under the consideration).

The concept has been further elaborated by Panetta et al. [73]. Properties of HVS were integrated and the original definition was modified to Root Mean Enhancement (RME) measure, calculated as

$$RME = \frac{1}{U}\sqrt{\sum_{u=1}^{U}\left|\frac{\log|I_{\text{center}}(u) - \overline{I(u)}|}{\log|I_{\text{center}}(u) + \overline{I(u)}|}\right|}. \tag{3.46}$$

SDME

A metric less sensitive to noise is Second Derivative based Measure of Enhancement (SDME), proposed by Panetta et al. [74]. It directly connects each pixel to the contrast value which can be exploited in various ways, e.g. for image contrast enhancement [75]. It is implemented as

$$SDME = -\frac{1}{U}\sum_{u=1}^{U} 20\log\left|\frac{I_{\max}(u) - 2I_{\text{center}}(u) + I_{\min}(u)}{I_{\max}(u) + 2I_{\text{center}}(u) + I_{\min}(u)}\right|. \tag{3.47}$$

GCF

A different approach has been proposed by Matković et al. as Global Contrast Factor (GCF) [106]. They firstly approximate the perceptual luminance $\mathcal{L}$ of pixels by

$$\mathcal{L} = 100\sqrt{\left(\frac{I}{255}\right)^{2.2}}. \tag{3.48}$$

Then for every pixel the average difference between itself and its neighbors above, below, on the left, and on the right is calculated. These average differences are then averaged again, providing the single contrast value GCF_m on each scale m. This procedure is repeated for nine scales. The final GCF value is obtained by weighting

$$GCF = \sum_{m=1}^{9} w_m GCF_m, \tag{3.49}$$

where w_m is the weight of the particular scale. The authors propose to calculate the weights from

$$w_m = (-0.406385\frac{m}{9} + 0.334573)\frac{m}{9} + 0.0877526. \tag{3.50}$$

3.2.4 Colorfulness Metrics

Several simple no reference metrics of image colorfulness are introduced in this section. Most of the metrics for color reproduction, known in the literature, are full reference and operate in one of the color spaces [107]. However, the reference-less approach is more relevant in the area of post-processing, since the best quality image is not known a priori.

CIQI

Color Image Quality Index (CIQI) was inspired by Hasler and Suesstrunk [108] and further modified by Fu [76]. It is defined as

$$CIQI = (\sqrt{\sigma_1^2 + \sigma_2^2} + 0.3\sqrt{\mu_1^2 + \mu_2^2})/85.59, \tag{3.51}$$

where

$$\mu_1 = \tfrac{1}{X\times Y}\textstyle\sum_{x=1}^{X}\sum_{y=1}^{Y}\Big(I_{\text{red}}(x,y) - I_{\text{green}}(x,y)\Big),$$

$$\mu_2 = \tfrac{1}{X\times Y}\textstyle\sum_{x=1}^{X}\sum_{y=1}^{Y}\Big(0.5\Big(I_{\text{red}}(x,y) + I_{\text{green}}(x,y)\Big) - I_{\text{blue}}(x,y)\Big),$$

$$\sigma_1^2 = \tfrac{1}{X\times Y}\textstyle\sum_{x=1}^{X}\sum_{y=1}^{Y}\Big(I_{\text{red}}(x,y) - I_{\text{green}}(x,y)\Big)^2 - \mu_1^2,$$

$$\sigma_2^2 = \tfrac{1}{X\times Y}\textstyle\sum_{x=1}^{X}\sum_{y=1}^{Y}\Big(0.5\Big(I_{\text{red}}(x,y) + I_{\text{green}}(x,y)\Big) - I_{\text{blue}}(x,y)\Big)^2 - \mu_2^2,$$

where I_{red}, I_{green}, and I_{blue} are the red, green and blue component of an image.

CQE1

The concept of CIQI was further elaborated by Panetta et al. [73] who introduced properties of HVS into computations. They defined CQE1 colorfulness measure which is implemented as

$$CQE1 = 0.2 \times \ln\left(\frac{\sigma_1^2}{|\mu_1|^{0.2}}\right) \times \ln\left(\frac{\sigma_2^2}{|\mu_2|^{0.2}}\right). \tag{3.52}$$

CQE2

The second extension of the CIQI measure, introduced in the same paper as CQE1 [73], is called CQE2 colorfulness and its definition is

$$CQE2_c = 0.2 \times \frac{\ln \sigma_1^2 \ln \sigma_2^2}{\ln \sigma_3^2} \times \frac{\ln \mu_1^2 \ln \mu_2^2}{\ln \mu_3^2}, \tag{3.53}$$

where

$$\mu_3 = \tfrac{1}{2}(\mu_1 + \mu_2),$$

$$\sigma_3^2 = \tfrac{1}{2\times X\times Y} \sum_{x=1}^{X} \sum_{y=1}^{Y} \left(I_{\text{r–g}}^2(x,y) - \mu_3^2) + (I_{\overline{\text{r,g}}\text{–b}}^2(x,y) - \mu_3^2) \right),$$

with

$$I_{\text{r–g}} = I_{\text{red}} - I_{\text{green}},$$

$$I_{\overline{\text{r,g}}\text{–b}} = 0.5\Big(I_{\text{red}} + I_{\text{green}}\Big) - I_{\text{blue}}.$$

Color Saturation

It is also possible to quantify the colorfulness of an image by the color saturation. The saturation channel of the image can be obtained by transferring it into the HSV [77] color space (hue, saturation, value). The higher the average of the saturation channel, the more colorfull can the image be considered.

3.2.5 Aesthetics Metrics

An interesting area in image quality assessment is quantification of aesthetic aspects. Considering the high subjectivity of the task, the experiments have to be designed with even higher level of caution. Essentially, the aesthetics plays a significant role in image post-processing. In both cases, we can see that the observers' opinions largely differ but some general agreement or trend can be identified [5, 7].

Aydın et al. [5] identified several image aspects influencing general aesthetic perception of an image. These aspects are *sharpness*, *depth*, *clarity*, *tone*, and *colorfulness*. They suggest to decompose the image using an Edge stopping pyramid [109], rather than Gaussian as is the popular approach, in order to preserve the in-focus regions intact by the decomposition. They obtain a detail layer on each level of decomposition and merge them into a multiscale contrast image. The detail layers are also used to compute a Focus map on each level providing a rough segmentation of the image according to the spatial frequency. The Focus maps and the contrast image are then used to compute the above mentioned aesthetically relevant aspects.

Since the *sharpness* increases with higher frequency components over larger area, the authors propose to measure it by averaging the absolute contrast magnitude in the first-level Focus map. This is supposed to provide more stable results.

The remaining Focus maps (i.e. levels higher than one) are searched through to estimate *depth*. The value corresponds to the level with the largest area of in-focus region. The completely out of focus image results in zero *depth*.

The *clarity* is computed from the area of out of focus region and the difference in average contrast of in focus and out of focus region. The metric increases with the presence of empty or low contrast areas.

The *tone* is obtained from the gamma corrected luminance of the image. The difference between 95% and 5% percentile of the values is taken and compensated for the over-exposed and under-exposed areas.

As the metric for colorfulness, the authors use the previously described CIQI measure in its original form as described by Hasler and Suesstrunk [108]. In order to merge the criteria into one aesthetic quality value, the calibration procedure on a subjective data was performed. For more details, refer to [5].

3.2.6 Distortion-Unaware Opinion-Aware Metrics

The next class of metrics is not designed specifically for a certain type of distortion but require training on a database(s) with observers' opinions. They mostly employ some kind of NSS model that is tuned on the images in the dataset(s).

BIQI

Blind Image Quality Index (BIQI), proposed by Moorthy and Bovik [78], is a two step framework that can be adapted to any particular set of distortions according to the training dataset. The wavelet decomposition on three scales and three orientations is firstly applied on an image, followed by parametrization of the subband coefficient by Generalized Gaussian Distribution (GGD). From this, 18-dimensional vector can be created (3 subbands $\times$ 3 orientations $\times$ 2 parameters). The distribution parameters for particular distortions can then be estimated. The upside of the metric is that it is able to detect multiple distortions in one image and provide a quality index according to their combined impact.

The available implementation has been trained on LIVE database [110] using support vector machine (SVM). It can therefore evaluate an effect of five distortions: JPEG, JPEG2000, fast fading channel, white noise, and blur. Nevertheless, the performance on JPEG images was poor, therefore, the previously described no-reference metric for JPEG compression was added for compensation.

BLIINDS, BLIINDS-II

Another approach was developed by Saad et al. as BLIINDS [79] and further revised and renamed to BLIINDS-II [80, 81]. It combines several submetrics in DCT domain, namely *GGD shape parameter*, *coefficient of frequency variation*, *energy subband ratio*, and *orientation model based feature*.

These DCT domain metrics are mutually complementary which makes them well suited for the fusion into a single quality score. Moreover, to provide more generality, the submetrics are calculated on multiple scales as well. The combination is, again, trained on the LIVE database [110] by modelling the data with multivariate GGD according to [111].

BRISQUE

Unlike the two algorithms described above, Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) [82] does not work in a transformation domain but directly in the spatial domain. It introduces a new model of statistics of a pairwise product locally normalized neighboring luminance values.

An image is first preprocessed as

$$\hat{I}(x,y) = \frac{I(x,y) - \overline{\left(I_{3\times3}(x,y)\right)}_w}{\mathrm{std}_w\left(I_{3\times3}(x,y)\right) + 1}, \tag{3.54}$$

where $\hat{I}$ is the preprocessed image, x, y are pixel coordinates, $I_{3\times3}(x,y)$ is the 3×3 neighborhood of the pixel with coordinates (x,y), and the weighted mean and standard deviation are calculated as

$$\overline{\Big(I_{3\times3}(x,y)\Big)}_w = \sum_{i=-3}^{3}\sum_{j=-3}^{3} w_{i,j} I(x+i, y+j), \tag{3.55}$$

$$\mathrm{std}_w\Big(I_{3\times3}(x,y)\Big) = \sqrt{\sum_{i=-3}^{3}\sum_{j=-3}^{3} w_{i,j}\Big[I(x+i,y+j) - \overline{\Big(I_{3\times3}(x,y)\Big)}_w\Big]^2}, \tag{3.56}$$

where $w_{i,j}$ is normalized, two-dimensional circularly-symmetric Gaussian weighting function.

18 NSS features coming from fitting the GGD to the preprocessed image and Asymmetric GGD to the pairwise coefficients are then calculated. To capture multiscale behavior, the features are also calculated on the second scale (after low-pass filtering and downsampling by the factor of 2) giving 36 features in total. A regression by SVM to the LIVE database [110] is then used to obtain the final quality predictor.

Curvelet Based Quality Metric

The next method operates in the Curvelet domain [83]. The NSS model in Curvelet domain first decomposes an image into blocks of 256×256 pixels and extract the Curvelet feature vectors for each of them. Each block is thus transformed into 5 layers of curvelet coefficients on 5 different scales. The model only considers coefficients on the finest scales, since the high frequency components better represent the image quality. The empirical PDF of the logarithm (base 10) of the magnitude of the curvelet coefficients is fitted by Asymmetrical GGD. In this way, vector of four features is obtained. Another two features describe the *orientation energy distribution* and final 6 features represent the *scalar energy distribution*. This set of 12 features is then combined by SVM regression.

3.2.7 Distortion-Unaware Opinion-Unaware Metrics

The last category of no reference metrics does not require any information about the distortion nor any subjective quality scores for training. They try to estimate the quality purely from the image features.

NIQE

The first attempt to come up with a "holistic" no-reference quality metric resulted in Natural Image Quality Evaluator (NIQE) [84]. Its base is the same as in case of above described BRISQUE metric [82]. However, it calculates the features for the salient blocks of the image only.

The preprocessed image, obtained from the equation (3.54), is divided into U blocks of 96×96 pixels. The salient blocks are selected according to the variance field δ

$$\delta(u) = \sum\sum_{i,j\in u} \mathrm{std}_w\Big(I_{3\times3}(x,y)\Big), \tag{3.57}$$

where $u = 1, 2, \ldots, B$. Only blocks with δ higher that $0.75 \times \delta_{\max}$ are considered as salient. The features in the salient blocks are then modelled using multivariate Gaussian model (MVG).

The final index is obtained as comparison of MVG of the image with MVG computed from 125 natural images[3] from publicly available Flickr database and from the Berkeley image segmentation database [112].

[3] http://live.ece.utexas.edu/research/quality/pristinedata.zip (retrieved on 30/08/2016)

The models are represented by their mean vectors ν and covariance matrices Σ. The comparison is done as

$$NIQE(\nu_1, \nu_2, \Sigma_1, \Sigma_2) = \sqrt{(\nu_1 - \nu_2)^{\mathrm{T}} \left(\frac{\Sigma_1 + \Sigma_2}{2} \right)^{-1} (\nu_1 - \nu_2)}. \tag{3.58}$$

The performance of NIQE on LIVE database is comparable to BRISQUE, even though it was trained on it, supporting the efficiency of the MVG model obtained from the natural images.

QAC

Another approach was proposed by Xue et al. [85]. They call it Quality Aware Clustering (QAC). They created a code-book of quality aware centroids in order to assess the quality of an image patch. In the learning phase, the patches were grouped according to the severity of distortion and the centroids were found. When evaluating the quality of the patch, it is classified according to the closest centroid. The quality of the whole image can then be obtained by pooling.

CS

The last metric to be described is based on Contourlet transform and Singular Value Decomposition (SVD) [86]. The Contourlet transform is used to obtain a high frequency "structural image". SVD is then applied on it and a new singular vector is created from the singular values ranging from 30 to 100. It has been observed that if a curve is drawn by linking these singular values, the area under the curve and its slope changes with the degree of a distortion. Therefore, these two entities can be used to predict the quality of the image.

4

Performance Measures for Objective Quality Metrics

If any objective quality metric is to be used as a substitute for subjective tests, its performance has to be validated, i.e. the criterion needs to be benchmarked in the given context. This performance evaluation is done on a "representative" dataset with ground truth data obtained from a subjective experiment. The term "representative" stands for the dataset's diversity in terms of content, processing algorithms assumed to be used in the particular context, and degrees of modifications (e.g. distortion) that are expected to occur in the context.

One of the most popular databases for objective image quality metrics performance evaluation is LIVE database developed by Sheikh et al. [110]. It consists of 29 natural source images distorted by five types of distortion (JPEG, JPEG2000, white noise, transmission through a fast fading channel, and Gaussian blur). The dataset contains 779 images in total. They were evaluated by 20-29 observers using ACR-HR methodology (see Section 2.2.1). The ground truth data are in the form of DMOS ranging from 0 to 100.

The CSIQ database developed by Larson and Chandler [50] with the MAD metric includes 30 source images that span five semantic categories – animals, plants, landscapes, people, and urban. These were distorted by JPEG and JPEG2000 compressions, Gaussian pink noise, Gaussian blur, and global contrast decrements resulting in 866 images. The procedure was an adjusted ranking followed by cross-content re-adjustment. The results are in form of DMOS in the 0 – 1 range, where 0 represents the best possible quality (no difference from the original).

The IVC database [113] consists of 10 source images distorted by four distortions – JPEG, JPEG2000, Locally Adaptive Resolution (LAR) coding, and blur. The total number of images is 235. DSIS methodology with 5 categories was utilized to obtain the ground truth from 15 observers. Therefore, the results span the range from 1 to 5.

Another popular dataset is Toyama [114] containing 14 source images processed by JPEG and JPEG2000 compressions in six degrees. Thus, the database contains 98 images. These were evaluated by 16 observers using ACR methodology resulting in MOS from 1 to 5.

Probably the largest available database is TID2008 developed by Ponomarenko et al. [115]. The source images are 24 natural and one artificial contents. The dataset uses 17 different types of distortions providing 1700 images in total. The PC methodology with available reference was adopted. However, the scores were transformed into MOS scores ranging from 0 to 8. In the more recent release called TID2013 [116], seven additional distortion types were introduced. Overall, 3000 distorted images are available.

Note that all of the most popular databases for objective metrics performance evaluation provide the ground truth in the form of MOS or DMOS. The reason is that the performance measures, as will be

described later on in this chapter, are defined for this kind of input only, i.e. they require the data to be on a ratio or at least interval scale (see table 2.1).

The standardized performance measures are described in ITU-T Recommendation P.1401 [117], VQEG report [118], and ITU-T Recommendation J.149 [119]. The following sections will introduce the measures and the ways to determine if the differences between performances are statistically significant. The disadvantages of the particular measures will be discussed as well.

4.1 Measures According to ITU-T Rec. P.1401

The title of the recommendation ITU-T P.1401 [117] is "Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models". It provides guidelines to the whole evaluation procedure and shares most of the criteria with VQEG report [118]. A very important part of the recommendation for performance evaluation describes a mapping of the scores predicted by a metric to a common scale with the MOS scores obtained from the subjective experiment. This should compensate for the limitations introduced by the experimental design, such as compression of the MOS scores towards the ends of the scale [118]. However, not a single standardized procedure is defined. The recommendation allows a simple linear mapping as well as other monotonic mapping procedures (in order to maintain the rank-order) such as third order polynomial mapping or logistic mapping.

Given the set of ground truth data MOS and the respective values predicted by an objective quality metric, denoted as OM, the mapped scores can be obtained as

$$OM' = \text{map}\{OM\} \text{ such that } RMSE(MOS, OM') \text{ is minimal}, \tag{4.1}$$

where $\text{map}\{.\}$ is a mapping function and $RMSE(.,.)$ is the root-mean-squared error measure calculated the same way as in Section 4.1.2. The measure for fitting optimization can be different, e.g. VQEG uses a maximization of Pearson's Linear Correlation Coefficient (PLCC).

4.1.1 Pearson's Linear Correlation Coefficient

Once the predicted scores have been mapped to the common scale, the dependency between the MOS and OM should ideally be linear. This can be measured by PLCC as

$$PLCC = \frac{\sum_{i=1}^{L}(MOS(i) - \overline{MOS}) \times (OM'(i) - \overline{OM'})}{\sqrt{\sum_{i=1}^{L}(MOS(i) - \overline{MOS})^2} \times \sqrt{\sum_{i=1}^{L}(OM'(i) - \overline{OM'})^2}}, \tag{4.2}$$

where L is the total number of stimuli in the set, i.e. length of MOS (and OM) used for performance evaluation. $PLCC = \pm 1$ represents absolute correlation, $PLCC = 0$ for totally uncorrelated series.

In order to obtain confidence intervals for the PLCC, Fisher z-transform can be used. The statistics Fz is approximately normally distributed and can be calculated as

$$Fz = \frac{1}{2}\ln\left(\frac{1 + PLCC}{1 - PLCC}\right) = \text{arctanh}(PLCC), \tag{4.3}$$

with its standard deviation

$$\sigma_{Fz} = \sqrt{\frac{1}{L-3}}. \tag{4.4}$$

The 95% confidence interval is then

$$CI_{Fz} = [Fz - 1.96\sigma_{Fz}, Fz + 1.96\sigma_{Fz}], \tag{4.5}$$

for $L > 30$. If $L \leq 30$, the value of 1.96 is substituted by the 95% percentile of the Student t-distribution with $L - dm$ degrees of freedom, where dm depends on the mapping function ($dm = 4$ for the third order polynomial mapping [117]). To obtain the CI for the PLCC, the interval needs to be transferred back from the transformation domain by the inverse Fisher z-transform, i.e.

$$CI_{PLCC} = \tanh(CI_{Fz}). \tag{4.6}$$

When comparing two PLCC values, the hypothesis testing approach is employed in order to determine the significance of the difference. Hypothesis H_0 assumes that the two coefficients are *not different*. The alternative hypothesis H_1 assumes that there is a significant difference between the PLCC values but does not discriminate which one is better. The analysis is based on calculating the FZ value as

$$FZ = \frac{Fz_1 - Fz_2 - \mu_{(Fz1-Fz2)}}{\sigma_{(Fz1-Fz2)}}. \tag{4.7}$$

Since H_0 assumes no difference, $\mu_{(Fz1-Fz2)} = 0$ and $\sigma_{(Fz1-Fz2)} = \sqrt{\sigma^2_{Fz1} + \sigma^2_{Fz2}}$. The FZ value is then compared to the 95% t-Student value for two-tailed test with $L - dm$ degrees of freedom. If it is larger, the H_0 can be rejected since the statistically significant difference between the PLCC values has been found. In the opposite case, the hypothesis cannot be rejected.

4.1.2 Root-Mean-Squared Error

Another measure described in the recommendation is RMSE. It is used to measure metrics' accuracy. The calculation is simply

$$RMSE = \sqrt{\frac{1}{L-1}\sum_{i=1}^{L}\Big(MOS(i) - OM'(i)\Big)^2}. \tag{4.8}$$

It has approximately $\chi^2(L-dm)$ distribution, where $L-dm$ is the number of degrees of freedom calculated the same way as in the case of PLCC. The higher RMSE value corresponds to worse accuracy. The range of the values depends on the common scale, i.e. on the range of MOS values within the set.

The 95% CI is obtained from this distribution as

$$CI_{RMSE} = \left[\frac{RMSE \times \sqrt{L-dm}}{\sqrt{\chi^4_{0.975}(L-dm)}}, \frac{RMSE \times \sqrt{L-dm}}{\sqrt{\chi^4_{0.025}(L-dm)}}\right]. \tag{4.9}$$

The hypothesis testing comes from the similar assumptions as in the case of PLCC. However, the statistics for the difference is defined as

$$q = \frac{RMSE^2_{\text{max}}}{RMSE^2_{\text{min}}}, \tag{4.10}$$

where $RMSE_{\text{max}}$ and $RMSE_{\text{min}}$ is the higher and lower value being evaluated, respectively. The statistics q is then compared to the 95% value from the F distribution $F(0.05, L_1 - dm, L_2 - dm)$. Since in absolute majority of the cases the compared RMSE values are obtained from the same set, $L_1 = L_2$.

4.1.3 Epsilon-Insensitive Root-Mean-Squared Error

In the case of Epsilon-Insensitive RMSE (RMSE*), the calculations are modified in order to consider the uncertainty of the ground truth data. More specifically, if the prediction error is smaller than the confidence

interval of the MOS, it is not considered as an error. It can be formalized as

$$RMSE^* = \sqrt{\frac{1}{L-1}\sum_{i=1}^{L}\left(\max\left[0, |MOS(i) - OM'(i)| - \delta(i)\right]\right)^2}, \tag{4.11}$$

where $\delta(i)$ is from the equation (2.5) or (2.7), depending on the number of observers. The rest of the calculations is similar to the classical RMSE. RMSE* inherited the dependency on the range of MOS values. It is also negatively proportional to the performance (i.e. lower RMSE* signifies higher performance).

4.1.4 Outlier Ratio

The last performance evaluation method described in this recommendation is the outlier ratio. It measures the metrics' accuracy and is defined as

$$OR = \frac{L_{\text{out}}}{L}, \tag{4.12}$$

where L_{out} is the number of the mapped predicted scores OM' which lie outside the CI, thus

$$L_{\text{out}} = \sum_{i=1}^{L} l_i, \tag{4.13}$$

with

$$l_i = \begin{cases} 0 & \text{if } OM'(i) \in CI(i), \\ 1 & \text{otherwise.} \end{cases} \tag{4.14}$$

CI is calculated from the equation (2.4). Outlier ratio is higher for worse performing metrics. Its standard deviation can be obtained from

$$\sigma_{OR} = \sqrt{\frac{OR \times (1 - OR)}{L}}, \tag{4.15}$$

and the confidence interval is again computed as

$$CI_{OR} = [OR - 1.96\sigma_{OR}, OR + 1.96\sigma_{OR}], \tag{4.16}$$

for $L > 30$. If $L \leq 30$, the 95% percentile of the Student t-distribution with $L - dm$ degrees of freedom, where dm depends on the mapping function, is used instead of 1.96.

The distribution describing outlier ratio is binomial with parameters $(OR, 1 - OR)$. The distribution of differences of two binomial variables for $L > 30$ is approximately Gaussian with $\mu_{(OR_1 - OR_2)} = \mu_{OR_1} - \mu_{OR_2} = OR_1 - OR_2 = 0$ and $\sigma_{(OR_1 - OR_2)} = \sqrt{\frac{\sigma^2_{OR_1}}{L_1} + \frac{\sigma^2_{OR_2}}{L_2}}$.

Since the null hypothesis H_0 is that the two values are not different, the equation for $\sigma_{(OR_1 - OR_2)}$ changes to

$$\sigma_{(OR_1 - OR_2)} = \sqrt{or \times (1 - or) \times (\frac{1}{L_1} + \frac{1}{L_2})}, \tag{4.17}$$

where

$$or = \frac{L_1 \times OR_1 + L_2 \times OR_2}{L_1 + L_2}. \tag{4.18}$$

The hypothesis testing is then similar to the procedure for PLCC with the statistics Z obtained as

$$Z = \frac{OR_1 - OR_2 - \mu_{(OR_1 - OR_2)}}{\sigma_{(OR_1 - OR_2)}}. \tag{4.19}$$

4.2 Measures According to ITU-T Rec. J.149

The ITU-T Recommendation J.149 [119] is entitled "Method for specifying accuracy and cross-calibration of Video Quality Metrics". It can be noted that the procedures described in this recommendation were intended for quality assessment of videos. Nevertheless, since the inputs are in the same format as in case of other quality assessment areas (i.e. MOS and OM), the methods can be adopted without any necessary modifications.

First of all, the MOS values are linearly scaled to the interval from 0 to 1 with 0 representing the best quality. The procedure can be formally described as

$$MOS' = \frac{MOS - best}{worst - best}, \tag{4.20}$$

where $best$ and $worst$ stand for the best and the worst possible score according to the scale used in the subjective quality experiment.

This is followed by fitting the results of each objective metric to the newly obtained MOS'. The mapping is the same as in equation (4.1) with the only difference being that OM is mapped to MOS' instead of raw MOS. The resulting OM' is then used to evaluate the accuracy of the metric as follows.

4.2.1 Resolving Power

The first performance measure is Resolving Power, based on the statistical analysis relative to the subjective data. Considering the mapping used on the subjective scores, their variance has to be modified as well. This is done by

$$SD'^2 = \frac{SD^2}{(best - worst)^2}, \tag{4.21}$$

where SD is the standard deviation calculated according to the equation (2.6).

In the next step, the z scores are calculated for each pair of stimuli in the set as

$$z(i,j) = \frac{MOS'(i) - MOS'(j)}{\sqrt{\frac{SD'^2(i)}{N(i)} + \frac{SD'^2(j)}{N(j)}}}, \tag{4.22}$$

with $N(i)$ being the number of observers who evaluated the i-th stimulus. Simultaneously, the differences of objectively predicted scores are obtained as

$$\Delta OM'(i,j) = OM'(i) - OM'(j). \tag{4.23}$$

It is useful for further steps to have all the values of $\Delta OM'$ positive. This can be ensured by convenient ordering in each pair. In practice, whenever the value of $\Delta OM'(i,j) < 0$, the order is reversed and therefore $\Delta OM'(i,j) = -\Delta OM'(i,j)$ and $z(i,j) = -z(i,j)$.

The z scores can then be used to determine the probability that the two stimuli are significantly different in terms of perceived quality. The probability is calculated from the CDF of the standard normal distribution, thus

$$p\Big(z(i,j)\Big) = \Phi\Big(z(i,j)\Big). \tag{4.24}$$

CDF $\Phi(z)$ is defined in the equation (2.19). Each stimuli pair (i,j) has its $\Delta OM'(i,j)$ value and the corresponding probability of difference $p\Big(z(i,j)\Big)$.

In the next step, the range of $\Delta OM'$ is equally divided into 19 bins with 50% overlap. Every bin b is then represented by the mean $\overline{\Delta OM'_b}$ and the average probability that the pairs in the bin are different $\overline{p_b}$.

The value $\overline{\Delta OM'_b}$ where $\overline{p_{bb}} > 0.95$ for $\forall bb \in [b, 19]$ is taken as the accuracy value RP on the 0.95 level of significance. The assumption is that the smaller the RP, the more accurate the metric performs. The

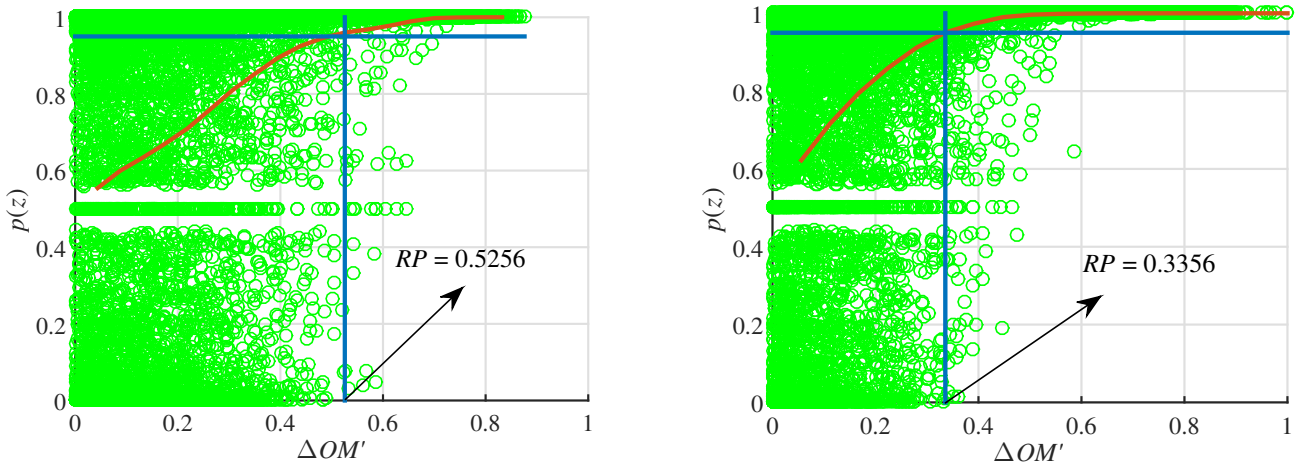


Figure 4.1: Plots showing calculation of Resolving Power for two different metrics.

Figure 4.1 shows an example with scatter plots of $p(z)$ and $\Delta OM'$ values for two different metrics. The red curve links the $\overline{p_b}$ values for each bin. The intersection with the $p(z) = 0.95$ line is taken as the RP value. The performance of the metric (b) is better with respect to the Resolving Power. If the difference in OM' values is larger than 0.3356, there is 95% probability that they are significantly different.

The disadvantage of this measure is that it does not provide any information about the performance in terms of correct decisions. If most of the objective scores are close to each other with small amount of outlying values, the Resolving Power can be low even if the performance is not very good. Moreover, no statistical comparison is defined in the recommendation.

4.2.2 Classification Plots

Another way to evaluate the performance of the objective quality metrics are the classification errors. These occur when the judgment about a stimuli pair is different in subjective and objective assessment. Considering the differences of an objective metric's scores $\Delta OM'$, a threshold of significance THR_{sig} can be selected. The pairs with the difference lower than this threshold will be classified as *qualitatively similar*. The information which pairs are truly similar and which are different can be obtained from the ground truth. Namely, if $z(i,j) \in \left(\arg\left[p\big(z(i,j)\big) = 0.05\right], \arg\left[p\big(z(i,j)\big) = 0.95\right] \right)$, the pair (i,j) is not significantly different in quality. The classification regions as shown in Figure 4.2 can then be identified.

Four outcomes can come from comparing the classifications. *Correct decision* is made when both subjective and objective data result in the same conclusion about the pair. *False tie* is the least invasive error. It occurs when the pair is significantly different in quality but the objective metric classifies it as similar. The opposite case is called *false differentiation*. The most invasive error is *false ranking* where both subjective and objective data agree on the pair being significantly different but the polarity is reversed, i.e. the worse stimulus is classified as significantly better by the objective metric.

To properly test the metrics' performance, the recommendation suggests to vary the threshold THR_{sig} and plot the relative frequencies of the classification errors and correct decision occurrence. An example for two metrics can be found in Figure 4.3. It is obvious that the metric in the case (b) reaches higher level of correct classification. False ranking occurs rarely and goes down quickly with growing THR_{sig}. This is much worse in the case (a).

The biggest issue of this type of performance evaluation is that it only allows graphical comparison of the metrics which is not very practical for comparing multiple metrics. It also does not allow for statistical analysis of differences.

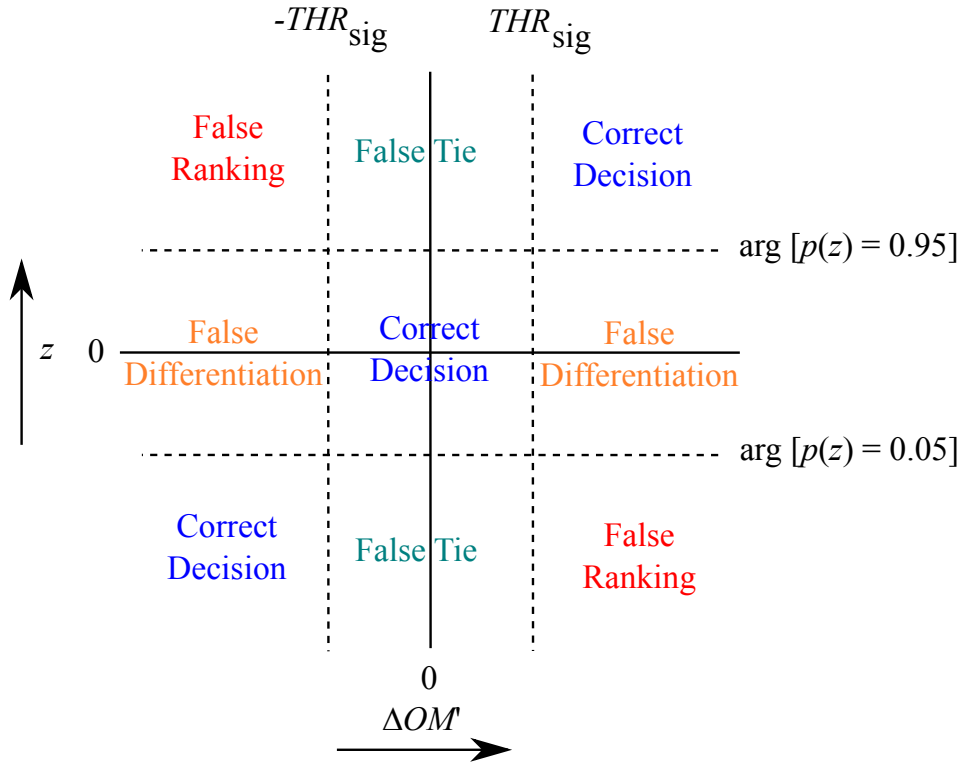


Figure 4.2: Classification regions.

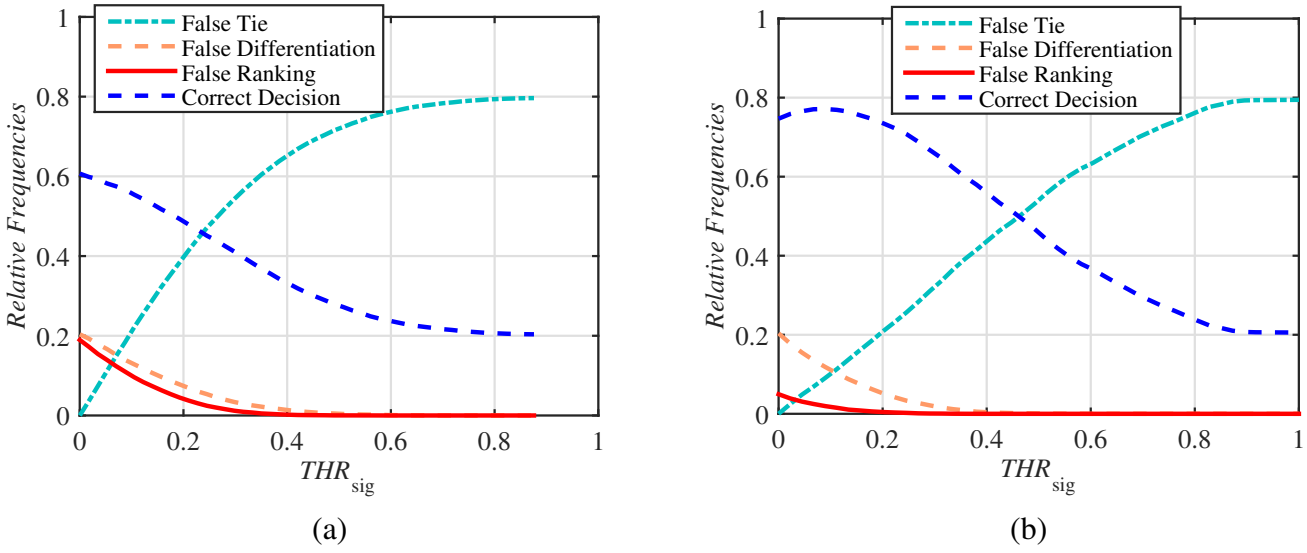


Figure 4.3: Classification plots for two different metrics.

4.3 Rank Order Correlation Coefficients

Rank order correlation coefficients are very popular non-parametric measures of objective quality metrics' performance. They are not a part of previously mentioned recommendations, however they are used in most of the studies, e.g. by VQEG [118] etc. Their advantage lies in the independence on any monotonic mapping of the objectively predicted scores. Therefore, they can be calculated directly between MOS and OM.

4.3.1 Spearman's Rank Order Correlation Coefficient

Spearman's Rank Order Correlation Coefficient (SROCC) is defined as

$$SROCC = \frac{6 \times \sum_{i=1}^{L} d_{\text{rank},i}^2}{L(L^2-1)}, \tag{4.25}$$

where $d_{\text{rank},i}$ is the difference between the rank of the i-th stimulus in subjective and objective evaluation. For example, if the i-th stimulus has the third highest MOS but fifth highest OM, $d_{\text{rank},i} = 5 - 3 = 2$. The procedure for determining the statistical significance of differences in SROCC values is similar to the case of PLCC (see Section 4.1.1).

4.3.2 Kendall's Rank Order Correlation Coefficient

To calculate Kendall's Rank Order Correlation Coefficient (KROCC), the order of each pair of stimuli in the set after both subjective and objective evaluation is checked. If the order in terms of MOS and OM agrees, the pair is considered "concordant". In the opposite case, the pair is "discordant". The final KROCC is then obtained as

$$KROCC = \frac{L_{\text{c}} - L_{\text{d}}}{\frac{1}{2}L(L-1)}, \tag{4.26}$$

where L_{c} and L_{d} are the numbers of concordant and discordant pairs in the set, respectively. The statistical significance of differences calculation is also similar to the one in Section 4.1.1.

4.4 Compensation for Multiple Comparisons

Most of the above mentioned procedures were accompanied with the possible method for determining the statistical significance of the difference in their results. However, all of these methods are defined for comparison of two values only. When comparing more than two outcomes a procedure to compensate for multiple comparison problem (also known as Type-I Error propagation problem) has to be applied [120].

The progress in multiple comparison research is nicely summarized in [121]. When testing a hypothesis, the null hypothesis H_0 is rejected if the probability that it is false is higher than the level of significance (95% throughout this thesis, i.e. the resulting p-value is smaller than 0.05). This means that there is still a 5% chance that the difference is caused purely by chance. If multiple such comparisons are performed, the probability that one of the decision is caused by chance grows. Several popular procedures to compensate for this effect are briefly introduced in this section.

There will be h null hypothesis considered, i.e. $H_0^{(1)}, ..., H_0^{(h)}$ with their respective p-values $p^{(1)}, \ldots, p^{(h)}$. In case of objective metrics performance evaluation, the hypothesis $H_0^{(1)}$ could be that the performance of the first metric is the same as the performance of the second, $H_0^{(2)}$ could mean that the second metric performs the same as the third, and $H_0^{(3)}$ could stand for the first metric performing the same as the third.

There are several arguments against using the compensation methods, e.g. in [122]. The main one is the increase in false negatives, i.e. the null hypothesis should have been rejected and was not. The decision

if to compensate or not depends on the harmfulness of false positives and false negatives. In any case, the data and all the methods should be fully reported and carefully interpreted.

4.4.1 Bonferroni Correction Procedure

The simplest method is named after Italian statistician C. E. Bonferroni. It comes directly from the Boole inequality that the probability of rejecting any true hypotheses is smaller than or equal to α [123]. Considering h hypotheses simultaneously, this can be ensured by testing them separately on the significance level $\frac{\alpha}{h}$, i.e. the null hypothesis $H_0^{(i)}$ is rejected if $p^{(i)} < \frac{\alpha}{h}$.

The advantage of this procedure is that it does not make any assumptions about the data. However, it is very conservative and has low statistical power. It is, therefore, not suitable for larger number of hypotheses.

4.4.2 Holm-Bonferroni Correction Procedure

In 1979, Holm came up with a sequentially rejective procedure which maintains no assumptions about the data while providing higher statistical power [123]. The sequential algorithm is described in Algorithm 3.

Firstly, the p-values are sorted from the smallest to the largest. They are then sequentially compared to the gradually increasing significance level. Once one of the p-values is higher, the respective hypothesis and all the following ones cannot be rejected.

Algorithm 3 Holm-Bonferroni sequentially rejective procedure [123].

Sort p-values from the smallest to the largest (i.e. $p^{(1)}$ is the smallest p-value)
for i from 1 to h **do**
  if $p^{(i)} > \frac{\alpha}{h-i+1}$ **then**
    Accept $H_0^{(i)}, \ldots, H_0^{(h)}$
    Stop the algorithm
  else
    Reject $H_0^{(i)}$
  end if
end for

4.4.3 Benjamini-Hochberg Correction Procedure

Another popular procedure was firstly proposed by Simes [124] and further described by Benjamini and Hochberg [125]. The method is also sequential but statistically more powerful than the ones previously mentioned. However, it assumes that the hypotheses are either independent or at least positively dependent. The method is introduced in Algorithm 4.

Algorithm 4 Benjamini-Hochberg sequentially rejective procedure [125].

Sort p-values from the smallest to the largest (i.e. $p^{(1)}$ is the smallest p-value)
for i from h to 1 **do**
  if $p^{(i)} < \frac{i}{h}\alpha$ **then**
    Reject $H_0^{(1)}, \ldots, H_0^{(i)}$
    Stop the algorithm
  else
    Accept $H_0^{(i)}$
  end if
end for

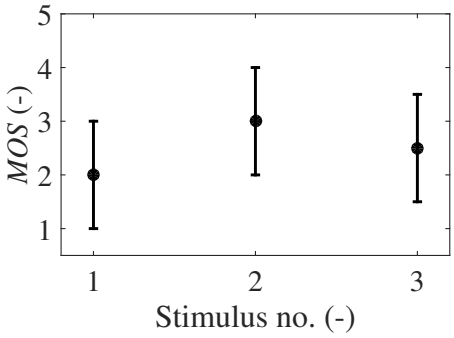


Figure 4.4: An example of three stimuli with not significantly different MOS values.

4.5 Disadvantages of the Standard Measures

All of the above described measures are being used by researchers to evaluate and compare performances of objective quality measures with respect to the subjective data. However, all of them suffer from at least two of the following disadvantages:

- They do not consider the uncertainty of MOS values, and/or
- they need a mapping to the common scale, and
- they do not allow for a simple combination of the results from multiple datasets, and
- they are only defined for MOS-like scenarios.

The remainder of this chapter is going to explain the above stated drawbacks in detail.

4.5.1 Not Considering the Uncertainty of MOS Values

This drawback is relevant to PLCC, RMSE, SROCC, and KROCC. These measures consider the MOS values obtained from the test without their respective confidence intervals. This can cause problems since if any MOS values are not statistically significantly different, it is impossible to decide from the subjective data what is the correct order of the stimuli in terms of quality. However, the objective metrics which will not result in the same rank as MOS values will be penalized.

As an example, three stimuli with MOS values and their respective 95% CI are shown in the Figure 4.4. The CI are largely overlapping and there is no statistically significant difference in the quality. Nevertheless, if a metric does not provide the highest score for the stimulus two and the lowest for the stimulus one, its performance will be considered poor, even though the correct order is not known.

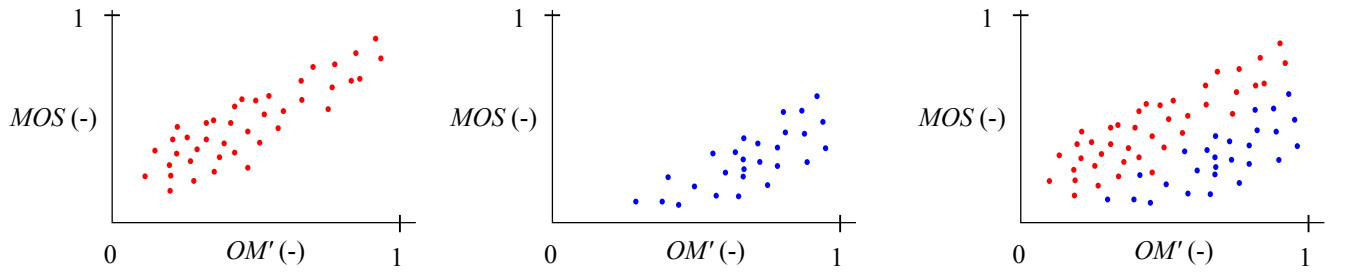
This drawback is even more severe in applications where the opinions of the observers differ largely due to various reasons such as multidimensionality of the quality, personal taste, etc. Since the image post-processing is exactly one of such applications (see Chapter 1), these measures should be used in this context with a great caution only.

4.5.2 Necessity of Mapping to the Common Scale

The mapping of scores predicted by objective metrics to the common scale with the MOS values is used for all of the measures except for the rank order correlation coefficients (SROCC and KROCC) which are not dependent on any kind of monotonic mapping. The arguments for using the mapping are mainly the compensation for the MOS range compression at the ends of the scale [118] and adaptation of the metrics to the particular scenarios [117]. However, the real applications use the metrics' scores without any mapping specific to the them. Moreover, there is no unified way of mapping that would be maintained everywhere.

	SSIM	MS-SSIM
Coefficients optimized with PLCC	**0.8575**	0.8562
Coefficients optimized with RMSE	0.8581	**0.8852**

Table 4.1: PLCC values for two different objective metrics used on CSIQ database [50] after two types of mapping.



(a) Dataset 1, $SROCC = 0.80$ (b) Dataset 2, $SROCC = 0.60$ (c) Datasets 1 & 2, $SROCC = 0.52$

Figure 4.5: An example of a metric's performance on two datasets separately and altogether.

Different mapping functions can provide different results favouring some of the metrics more than the others.

Table 4.1 shows an example of the impact of different mapping procedures on the measured performance. Two objective metrics – SSIM [45] and MS-SSIM [46] – are evaluated with respect to the CSIQ database [50] using PLCC. In both cases, a third order polynomial function is used. The first row provides the results after optimization of the mapping function according to PLCC (which is a VQEG conform procedure [118]) while in the second row, the coefficients of the mapping function are found by minimizing the RMSE (as recommended in ITU-T Rec. P.1401 [117]). We can notice a significant change in the measured performance. Note that only the function's coefficients selection procedure has been changed. If some other type of mapping (e.g. logistic) would be used, the difference could be even more severe.

Considering the above stated arguments, a methodology measuring the performance closer to the real usage of the metrics (i.e. without the mapping) would prove beneficial.

4.5.3 Complicated Combination of Multiple Datasets

As mentioned previously, the performance is always evaluated with respect to a certain dataset. In order to test metrics' abilities more generally, multiple datasets are used. To provide an overall performance evaluation, mostly the average value, weighted according to the database size, is being reported. However, this approach can be misleading.

An example of the danger of the averaging is shown in Figure 4.5. Figure 4.5(a) and (b) shows the scatter plot of MOS and OM' for two datasets with their respective $SROCC$. If the weighted average approach is used, the $SROCC_{\text{w-avg}} = 0.73$ while if the data are put together (considering the perfect mapping between the experiments), the overall $SROCC = 0.52$.

Moreover, the averaging is impossible to be used on RMSE and RMSE* since their values are dependent on the range of MOS values. Therefore, when a different procedure or scale is used in the experiments, the values do not allow for direct comparison.

To put the data from multiple databases to a common scale, Pinson and Wolf [126] proposed a mapping based on the Iterated Nested Least-Squares Algorithm (INLSA) developped by Voran [127]. All the data are firstly normalized and then mapped using the INLSA. To get a side information necessary to find a proper mapping, either subjective meta-test with content selected from all the datasets needs to be performed, or an objective quality metric is employed. Nevertheless, the mapping is then dependent on the accuracy of the used objective metric, making it impractical for the performance evaluation scenario. The inclusion of

multiple databases into the standard performance evaluation is thus complicated and should be carefully interpreted.

4.5.4 Applicability to the MOS-like Scenarios Only

An absolute majority of the subjective datasets are obtained using direct scaling methods since the data on the ratio scale are the most suitable for performance evaluation with the standard measures. Even the databases TID2008 [115] and TID2013 [116] which were developed using modified PC methodology were then artificially transferred to MOS-like scenario.

If only the interval data for each content are available (which is typical for indirect scaling methods), the measures can only be computed per content and the same problems as in case of combining multiple experiments occur. Therefore, the results of indirect scaling methods are very rarely used for benchmarking of objective quality metrics. Considering their higher discriminatory power and benefit in applications more challenging for observers, the ways to exploit such data for performance evaluation are necessary to be developed.

5

Novel Methods for Evaluating Performance of Objective Metrics

The standard methodologies for evaluating performance of objective quality metrics have been thoroughly described in Chapter 4 together with their main drawbacks (Section 4.5). As has been explained, these disadvantages are even more severe in image post-processing where the aspects like personal taste or multidimensionality of quality alteration (see Chapter 1) play an important role in the perception of quality and thus can introduce a noise into the data.

Therefore, it is desirable to develop a novel methodology that will be able to overcome the issues of the standard measures. Namely, it should:

- Consider the uncertainty of the subjective scores,
- be able to evaluate and compare the metrics' performance without the necessity of mapping,
- enable simple inclusion of different datasets in the evaluation,
- allow the benchmarking of metrics regardless the subjective procedure used to obtain the opinion scores.

Apart from these technical requirements, the new methodology should also provide the results that will be easily interpretable, i.e. it needs to be decided, *what makes an objective quality metric reliable in the given context*. This can be answered by taking into account the real usage of objective measures.

Virtually all real use case scenarios can be brought down into these two questions:

(a) Are the two stimuli significantly different in quality?

(b) If they are, which of them is of better quality?

It is convenient to evaluate the metrics' abilities to address the two above stated points separately, since some criteria could prove useful for one of the points but not the other. Certain scenarios then enable to use different models for individual tasks. For example, optimizing the bitrate while maintaining the perceived quality only requires the metric to be reliable in the case **(a)**. In enhancement, the final stimulus is desired to be noticeably different (case **(a)**) and simultaneously of higher quality (case **(b)**) than the original. If different metrics are considered for each case, conditional optimization can be used (i.e. maximizing one metric while the second metric is above the threshold for similarity).

It should be noted that *it is possible to obtain a ground truth regarding the points* **(a)** *and* **(b)** *from all the possible subjective experimental procedures*. In order to get this data from direct scaling methods, the same procedure to get $p(z)$ as in section 4.2.1 can be used. The probability $0.05 < p(z) < 0.95$ signifies that the pair is not significantly different. In case there is a significant difference, the pair with higher MOS value is of better quality. In case of indirect scaling methods the ground truth is obtained per content only using either Direct PCM Processing from section 2.3.3, or from the interval scale scores (such as statistical evaluation of BTL scores differences).

By extracting the information regarding the two identified questions (i.e. which pairs are significantly different in quality and which of the stimuli in the different pairs are the better ones), the last two problems that have been identified in Section 4.5 are eliminated, since the inputs are the same regardless the subjective procedure or format of the subjective data. Moreover, by taking into account significance of differences, the first drawback is resolved as well. It only needs to be assured, that the performance evaluation methodology does not require mapping of objective scores.

The input for the performance evaluation is therefore for each pair of stimuli (i, j):

- Information if i and j are significantly different,
- if they are, which on them is of better quality,
- the difference of the objective metric's scores $\Delta OM(i, j) = OM(i) - OM(j)$.

In the following sections, firstly the adaptability of the existing measures to work with the above described input is going to be discussed. The measures suitable for such adaptation will be adjusted accordingly. Further, a novel methodology tailored to the defined scenario will be introduced.

5.1 Adaptation of Existing Measures

Considering the input as described above, the performance evaluation measure should be able to process the *pairwise information*. Since most of the measures need the whole series of scores, they cannot be used to evaluate metrics' abilities regarding the proposed scenario. Nevertheless, three exceptions can be identified – KROCC, Resolving Power, and Classification Errors. KROCC actually divides the series of scores into concordant and discordant pairs and is therefore suitable for working with pairwise input. The Resolving Power and Classification Errors work with such input by definition. However, the Resolving Power is dependent on the mapping of objective metrics' scores into the common scale in order to compare them. This issue can be overcome in case of Classification Errors, as will be shown below.

5.1.1 Adapted KROCC

The adaptation of the KROCC is fairly straightforward. If the stimuli pair is not statistically significant, it is impossible to decide whether it is concordant or discordant. Therefore, in the adapted KROCC only the pairs with statistically significant difference in subjective votes are considered. It is calculated simply as

$$KROCC_{\text{adapted}} = \frac{L_{\text{sig,c}} - L_{\text{sig,d}}}{L_{\text{sig}}}, \tag{5.1}$$

where L_{sig} is the number of significantly different pairs out of which $L_{\text{sig,c}}$ are concordant and $L_{\text{sig,d}}$ are discordant. It can be seen that the adapted coefficient can only provide information about metrics' performance in terms of question **(b)**.

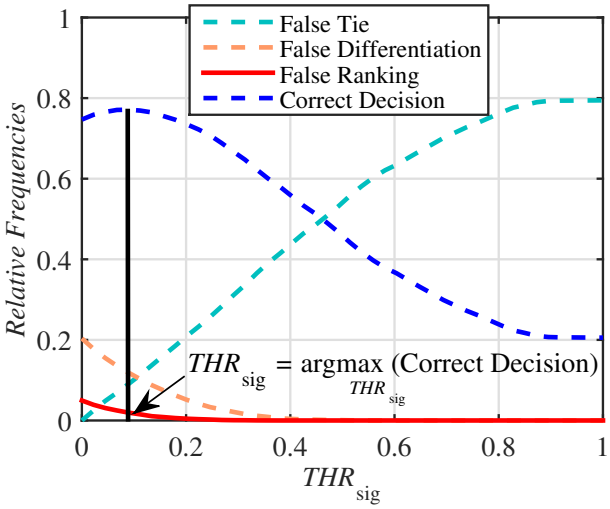


Figure 5.1: Classification plot with marked point of the interest.

5.1.2 Adapted Classification Errors

The Classification Errors are defined for differences in mapped scores $\Delta OM'$ in order to provide graphically comparable plots. However, the same procedure can also be applied on non-mapped scores ΔOM. To be able to compare the metrics' performance and overcome the impracticality of only graphical comparisons, it is possible to define the points of interest where the classification is the most desirable to be determined. This will allow to numerically compare relative frequencies which are normalized for all of the metrics.

Two points of interests will be discussed here. The first one is the point corresponding to $THR_{\text{sig}} = 0$. This represents the case where any change in a metric's score is considered to result in perceivable difference. Despite the unreasonableness of such assumption, objective metrics are often used exactly this way. This point can actually show how big mistake will be made if the measures not considering the uncertainty of the subjective scores are used.

The second point of interest is much more relevant for the metrics performance evaluation and comparison. It is the value of significance threshold maximizing the correct decision, i.e.

$$THR_{\text{sig}} = \arg\max_{THR_{\text{sig}}} (\text{Correct Decision}).$$

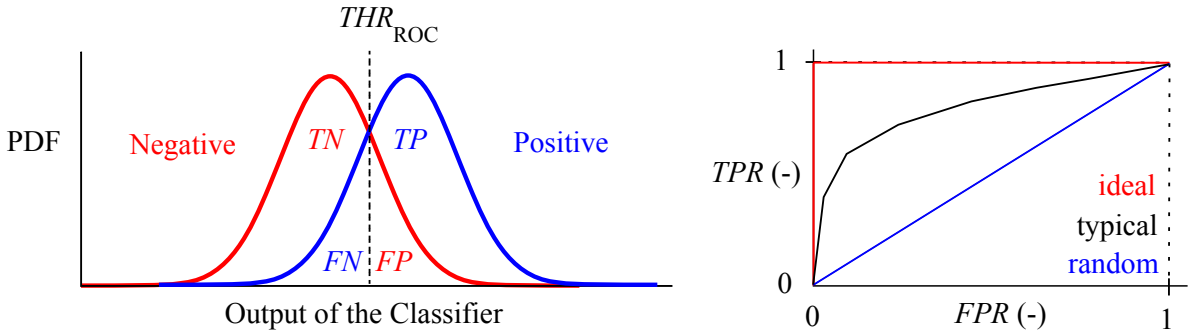
The point is marked by a black line in the Figure 5.1.

In order to compare the relative frequencies corresponding to the particular Classification Errors statistically, the binomial test similar to the one defined for Outlier Ratio (see Section 4.1.4) can be utilized. Alternatively, Fisher's [37] or Barnard's [38] exact test could be employed as well.

Although the adapted classification errors can provide valuable information and basis for objective metrics' comparison, the restriction to the points of interest can be limiting. The interpretation with respect to the above defined reliability requirements (i.e. **(a)** and **(b)**) is complicated. Therefore, it is desirable to come up with a novel methodology.

5.2 New Methodology Based on ROC Analyses

Having identified the requirements for reliable objective quality metric, the novel performance evaluation methodology able to quantify the metrics' ability to fulfil these requirements can be designed. As previously argued, it is convenient to determine the capability of metrics to address the two questions separately. Both of the questions can be understood as a problem of classification into two groups. Therefore, Receiver Operating Characteristic (ROC) analysis [128] has been selected as an appropriate tool for the performance



(a) PDFs of ratings under two hypotheses with the threshold

(b) Three different ROC curves

Figure 5.2: ROC Analyis.

evaluation. The method has been published in [129, 130].

5.2.1 ROC Analysis

ROC analysis is a popular tool to assess the performance of binary classifiers. It evaluates the ability of a classifier to assign the input into two output groups (e.g. positive / negative). If the input is classified as positive and it should be classified as negative, the outcome is called *false positive (FP)*, otherwise it is *true positive (TP)*. If the outcome is negative and should be positive, it is labeled as *false negative (FN)*, and as *true negative (TN)* in the opposite case. In the Figure 5.2(a), the PDFs of ratings under two hypotheses are shown. The threshold represents the criterion dividing the scores into four groups – *TP*, *FP*, *TN*, and *FN*.

To evaluate the decision ability, the threshold is shifted and for every position *true positive rate (TPR)* and *false positive rate (FPR)* is calculated

$$TPR = \frac{TP}{TP + FN} \quad , \quad FPR = \frac{FP}{FP + TN}. \tag{5.2}$$

The ROC curve is then the dependency of *TPR* on *FPR*. ROC curves in Figure 5.2(b) represent three different cases. The red curve would be obtained when the two PDFs are completely separated (there is no overlap between them) signifying that the classifier works ideally. The blue one represents the case when 50% of cases are correct and 50% are false, i.e. classifier is equivalent to random guessing. ROC curves for most of the classifiers will typically lie somewhere in between the two (see the black curve). The closer the ROC curve is to the upper left corner, the better.

Despite the elegance of the graphical representation, it is much more convenient to have some kind of a merit to represent the performance instead of the curve. For this purpose, the Area Under Curve (AUC) measure is defined. Its computational details can be found, for example, in [131].

$$AUC = \sum_{t=2}^{T} \frac{(TPR(t) + TPR(t-1)) \times (FPR(t) - FPR(t-1))}{2}, \tag{5.3}$$

where T is the total number of the threshold positions. In practice, T is mostly the number of the samples available from the two distributions together.

5.2.2 Statistical Comparison of ROC Analyses

Apart from the empirical definition introduced above, Bamber [132] pointed out that the AUC can also be calculated from Mann-Whitney U statistic [133] (also known as Wilcoxon statistic) which brought an opportunity to approach the analysis stochastically. This is very important since only a limited number of

samples is always available from the PDFs for both groups. The AUC can be estimated as

$$AUC = \frac{1}{N_{G1} \times N_{G2}} \sum_{i=1}^{N_{G1}} \sum_{j=1}^{N_{G2}} \mathcal{H}\Big(G_1(i) - G_2(j)\Big), \tag{5.4}$$

where N_{G1} and N_{G2} are the numbers of samples in the first and the second group, G_1 and G_2 are the vectors of independent and identically distributed samples drawn from the two populations, and $\mathcal{H}(.)$ is a Heaviside function defined in equation (6.16).

Various tests has been proposed in order to determine the statistical significance of difference between two AUC values. They can be parametric, non-parametric, or based on simulation such as permutation tests [134]. In this thesis, two non-parametric techniques proposed by Hanley and McNeil [135] and by DeLong et al. [136] will be introduced.

Hanley and McNeil Method

Hanley and McNeil [137] showed a way to calculate a standard error for an AUC as

$$SE_{\text{AUC}} = \frac{AUC(1 - AUC) + (1 - N_{G1})(Q1 - AUC^2) + (1 - N_{G2})(Q2 - AUC^2)}{N_{G1} \times N_{G2}}, \tag{5.5}$$

where $Q1$ and $Q2$ are obtained as

$$\begin{aligned} Q1 &= AUC/(2 - AUC), \\ Q2 &= 2AUC^2/(1 + AUC). \end{aligned} \tag{5.6}$$

The 95% CI of the AUC is then defined as

$$CI_{\text{AUC}} = \Big[AUC - 1.96 \times SE_{\text{AUC}}, AUC + 1.96 \times SE_{\text{AUC}}\Big]. \tag{5.7}$$

When comparing two AUC values obtained from different classifiers [135], the statistic z_{AUC} defined as

$$z_{\text{AUC}} = \frac{AUC_1 - AUC_2}{\sqrt{SE^2_{\text{AUC1}} + SE^2_{\text{AUC2}} - 2 \times cc \times SE_{\text{AUC1}} \times SE_{\text{AUC2}}}} \tag{5.8}$$

can be used. The variable cc stands for the estimated correlation between the two areas determined according to the Table I in [135].

The probability that AUC_1 is larger than AUC_2 is than obtained from the standard normal CDF as

$$Pr(AUC_1 > AUC_2) = \Phi(z_{\text{AUC}}). \tag{5.9}$$

If the probability is larger than 0.95, the difference is considered statistically significant.

DeLong Method

DeLong et al. [136] proposed another method based on *structural components*. They define two components

$$\begin{aligned} V^{(k)}_{10}(i) &= \tfrac{1}{N_{G2}} \textstyle\sum_{j=1}^{N_{G2}} \mathcal{H}\Big(G_1(i) - G_2(j)\Big), \text{ for } i = 1, \ldots, N_{G1}, \\ V^{(k)}_{01}(j) &= \tfrac{1}{N_{G1}} \textstyle\sum_{i=1}^{N_{G1}} \mathcal{H}\Big(G_1(i) - G_2(j)\Big), \text{ for } j = 1, \ldots, N_{G2} \end{aligned} \tag{5.10}$$

for each classifier k. When comparing two AUC values, two 2×2 matrices $S_{10} = [s_{10}^{(k,l)}]_{2\times2}$ and $S_{01} = [s_{01}^{(k,l)}]_{2\times2}$ can be obtained as

$$\begin{aligned} s_{10}^{(k,l)} &= \frac{1}{N_{G1}-1}\sum_{i=1}^{N_{G1}}\left[V_{10}^{(k)}(i) - AUC_k\right]\left[V_{10}^{(l)}(i) - AUC_l\right], \\ s_{01}^{(k,l)} &= \frac{1}{N_{G2}-1}\sum_{j=1}^{N_{G2}}\left[V_{01}^{(k)}(i) - AUC_k\right]\left[V_{01}^{(l)}(i) - AUC_l\right]. \end{aligned} \tag{5.11}$$

The covariance matrix is then obtained as

$$COV = \frac{1}{N_{G1}}S_{10} + \frac{1}{N_{G2}}S_{01}. \tag{5.12}$$

The approach is computationally demanding, especially for large sample sizes. Therefore, Sun and Xu [138] proposed a faster algorithm using mid-ranks instead of the Heaviside function.

The statistics z_{AUC} is here obtained as

$$z_{\text{AUC}} = \frac{AUC_1 - AUC_2}{vec^{\text{T}} \otimes COV \otimes vec}, \tag{5.13}$$

where $vec = [1, -1]$ is a column vector, vec^{T} is its transposed version, and the operator $\otimes$ is a matrix multiplication. The probability $Pr(AUC_1 > AUC_2)$ is then calculated similarly to the equation (5.9). The DeLong method should be statistically more powerful than the one proposed by Hanley and McNeil.

5.2.3 Different vs. Similar ROC Analysis

Now that the evaluation tool has been identified and the statistical comparison procedures were introduced, it is possible to define the exact way to address the two main questions identified from the real use case scenarios. In the first part, metrics' abilities with respect to the question "**(a) Are the two stimuli significantly different in quality?**" should be determined.

From the input, the distribution of the absolute metrics' scores differences $|\Delta OM|$ can be obtained for both statistically significant and not significant pairs. The assumption is that the metric scores should be close together for qualitatively similar pairs (i.e. $\Delta OM \to 0$) while for the significantly different pairs, the scores should differ more. An ideal objective criterion should, therefore, result in separated distributions. The capability of a metric in separating the two distributions is measured by ROC analysis and numerically quantified by AUC value.

Another output of the analysis can be the threshold $THR_{95\%\text{DS}}$ representing the value of ΔOM ensuring 95% probability that the stimuli are different. This value is obtained as 95% percentile of the distribution for significantly *not* different pairs. It should be noted that the value of the threshold depends on the range of the metric's scores and as such *cannot* be used for direct comparison of metrics. Nevertheless, it is an important value for the practical usage of the criteria. The pipeline of the Different vs. Similar ROC Analysis is visualized in the upper part of the Figure 5.3.

5.2.4 Better vs. Worse ROC Analysis

The second part of the proposed methodology considers the significantly different pairs only. The goal is to determine how well do metrics perform in terms of recognizing the stimulus of better quality in a pair, i.e. addressing the question "**(b) If the stimuli are different, which of them is of better quality?**"

In this part, raw differences ΔOM (not absolute) are considered. The pairs are divided according to which of the stimuli in the pair is of better quality, i.e. one group consists of pairs where the first stimulus is better (positive difference in opinion scores) and the second group contains pairs where the first stimulus is worse (negative opinion scores difference). Since the ordering of stimuli in pairs is artificial, it has been decided to consider each pair in both configurations (A_1 A_2 and A_2 A_1) to ensure that the two

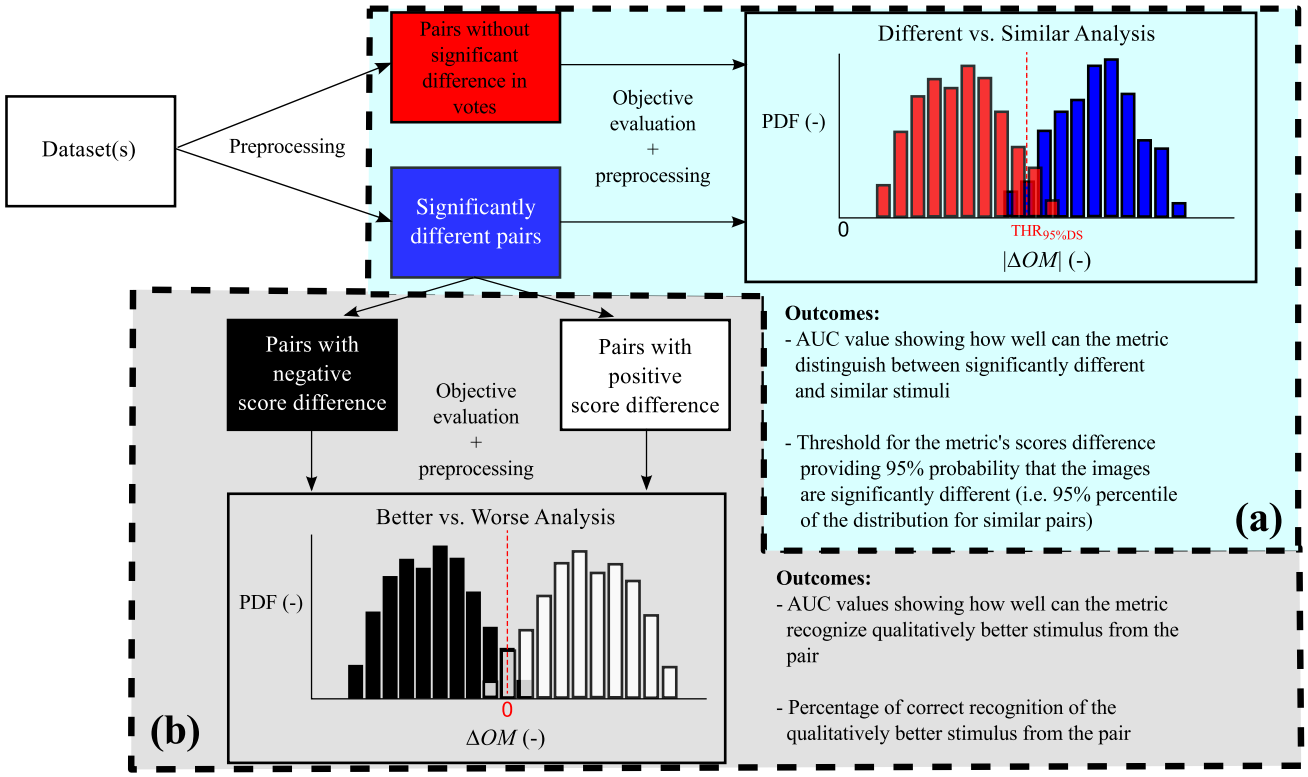


Figure 5.3: The framework of the novel performance evaluation methodology based on ROC analyses.

distributions will have the same number of samples. Nevertheless, unless the two distributions are hugely disproportional, it does not influence the performance evaluation too much. The obtained distributions for ΔOM will be symmetric around zero.

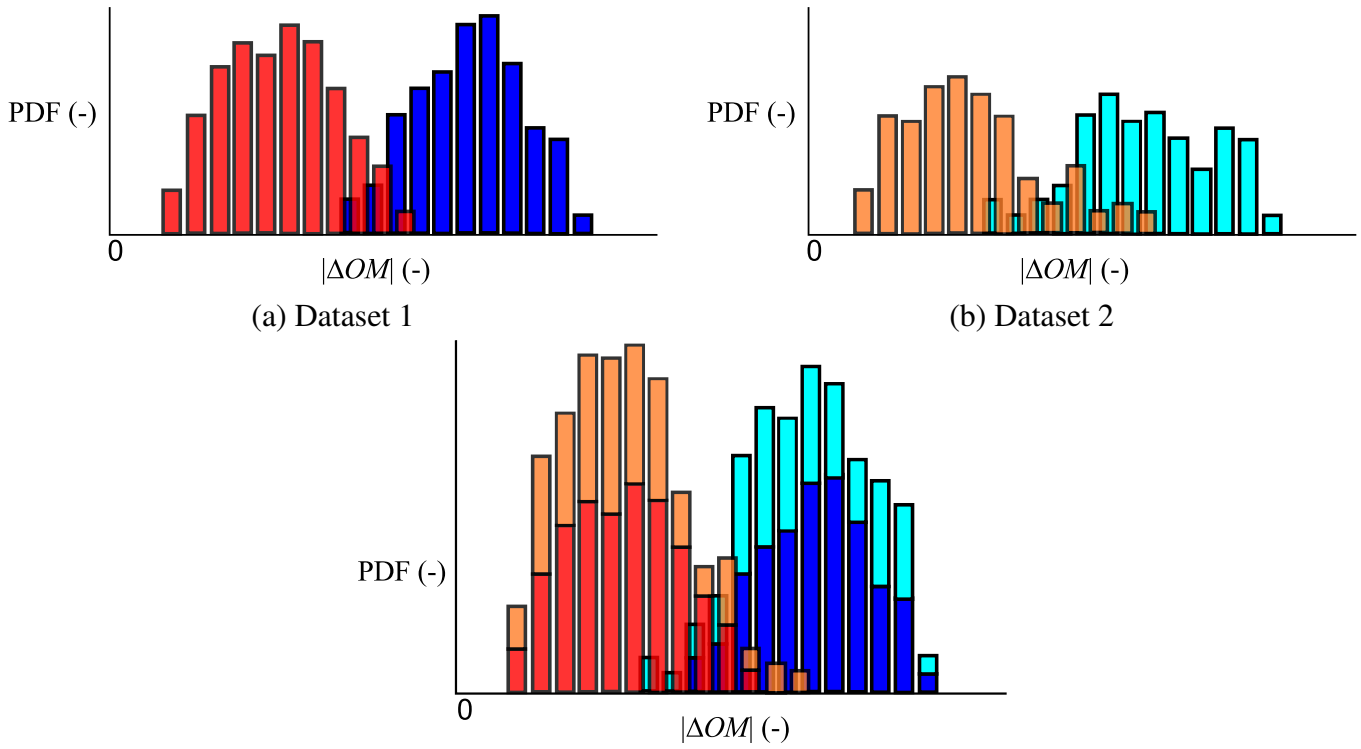
The first indicator of performance can be the percentage of correct classification in zero (C_0) showing what is the ratio between correct and incorrect recognition of the better stimulus from the pair. The relation to the adapted KROCC (Section 5.1.1) and the Classification Errors for $THR_{\text{sig}} = 0$ (Section 5.1.2) can be noted. These measures act like any difference in metrics' scores lead to a change in perceived quality which can be misleading in certain terms as will be shown later on in the example evaluation on real data.

To also consider the shape of the two distributions, ROC analysis can be performed. The resulting AUC value indicates how well are the distributions separated, providing better insight into the metrics behavior and representing the overall performance regarding the question **(b)** in a better way. The pipeline of the second part of the proposed methodology is depicted in the bottom part of the Figure 5.3.

In [130], the third analysis called Better vs. Equal-Worse Analysis is introduced. It represents a combination of the two above described parts and is bounded by them. In terms of the considered application scenario its informative value is limited and this thesis will, therefore, be restricted to the two ROC analyses. For more information about the third analysis, refer to [130].

5.2.5 Multiple Datasets Combination

As explained in Section 4.5.3, the calculation of overall performance over multiple datasets is complicated when classical evaluation measures are used. In case of the proposed methodology, the distributions for groups obtained from each dataset are always on the same scale defined by the range of a metric's scores differences ΔOM. The distributions for multiple datasets can, therefore, be easily put together as illustrated in Figure 5.4. The whole analysis (Figure 5.3) can then be run on the combined distributions (red and orange will represent one distribution while blue and cyan the second).



(a) Dataset 1 (b) Dataset 2

(c) Datasets 1 & 2

Figure 5.4: Demonstration of combining results obtained from multiple dataset using the proposed methodology.

Table 5.1: Numbering of the objective metrics.

1	2	3	4	5
PSNR	SSIM	IW-PSNR	MS-SSIM	IW-SSIM

5.3 Demonstration of Advantages

To demonstrate the advantages of the proposed methodology, the data available from the performance evaluation of IW-SSIM metric [47] has been selected. This way, no additional bias in data obtaining can be introduced and the outcomes of other performance evaluation methods are available for comparison. All the data were obtained from the supporting website[1].

Predicted scores for five objective algorithms are provided – PSNR, SSIM [45], IW-PSNR [47], MS-SSIM [46], and IW-SSIM [47]. The metrics are introduced in Section 3.1. For the sake of readability, the criteria are numbered according to the table 5.1 in the following figures. The datasets used to evaluate their performance in [47] are LIVE [110], A57 [94], IVC [113], Toyama [114], TID2008 [115], and CSIQ [50].

For the detailed analysis, IVC dataset has been selected, since it provides interesting results suitable for the demonstration.

5.3.1 Performance on IVC dataset

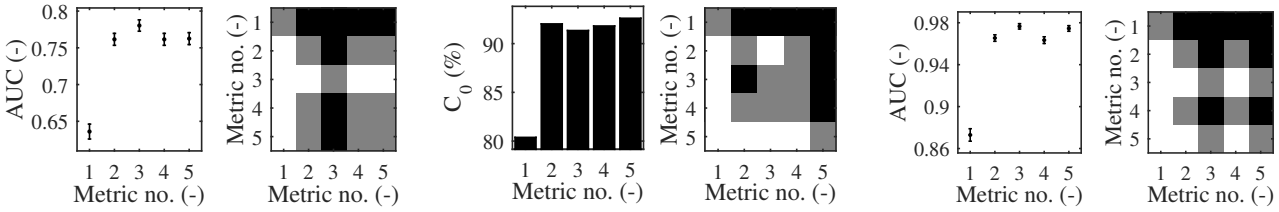
In the corresponding paper [47], four standard performance evaluation measures are used – PLCC, RMSE, SROCC, and KROCC. Their outcomes are shown in the Table 5.2.

The results of the particular analyses (sections 5.2.3 and 5.2.4), namely AUCs and percentage of correct classification (C_0) with statistical significance of differences, are depicted in Figure 5.5. The error bars

[1] https://ece.uwaterloo.ca/~z70wang/research/iwssim/ (retrieved on 30/08/2016)

	PLCC	RMSE	SROCC	KROCC
PSNR	0.6719	0.9023	0.6884	0.518
SSIM	0.9119	0.4999	0.9018	0.7223
IW-PSNR	0.8963	0.5403	0.8998	0.7165
MS-SSIM	0.9108	0.5029	0.8980	0.7203
IW-SSIM	**0.9231**	**0.4686**	**0.9125**	**0.7339**

Table 5.2: The results of standard performance evaluation measures for IVC database according to [47].



(a) Different vs. Similar (AUC) (b) Better vs. Worse (C_0) (c) Better vs. Worse (AUC)

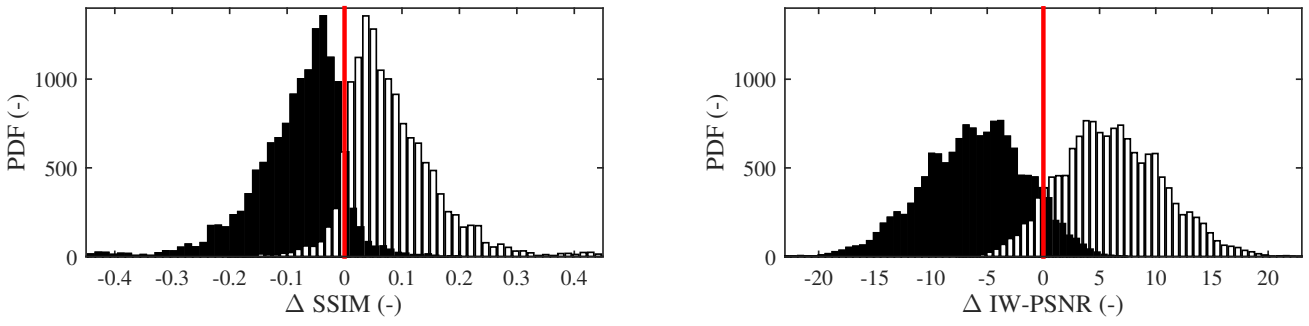
Figure 5.5: The results and statistical analysis for the IVC dataset. Significance plots show that the performance of the method in the row is either significantly better (white), worse (black), or none of the previous (gray).

represent 95% confidence intervals. The significance of the AUC values has been checked using Hanley and McNeil method [135]. For the C_0 values, Fisher's exact test [37] was employed. The multiple comparisons were compensated via Benjamini-Hochberg procedure [125]. The white boxes in the significance plots correspond to the cases when model in the row significantly outperforms the model in the column. If its performance is significantly lower, the corresponding box is black. The gray box symbolizes the case where we are not able to determine the better performing method.

It can be seen that IW-PSNR (#3) significantly outperforms all the other metrics in the Different vs. Similar analysis (Figure 5.5(a)). On the other hand, PSNR (#1) has the lowest performance, also with statistical significance.

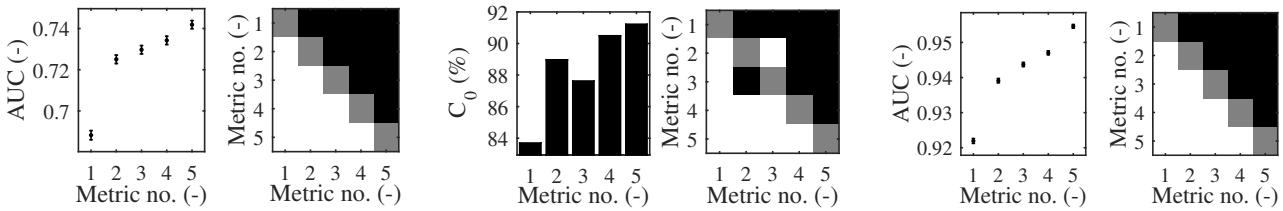
In the second analysis, an interesting phenomenon can be observed. IW-PSNR (#3) provides statistically worse classification than SSIM (#2) (Figure 5.5(b)) but reaches significantly higher AUC value (Figure 5.5(c)). To explain this, we closely studied the behavior of the two metrics. The histograms of Δ_{SSIM} and $\Delta_{\text{IW-PSNR}}$ for the two groups defined in section 5.2.4 are depicted in Figure 5.6 (the number of bins is the same for both models).

The distributions for the IW-PSNR are much broader with modes more distant from 0. This means that if we broaden the red area in the Figure 5.6, which is equivalent to not considering the pairs with small differences as being different, the performance of IW-PSNR will be dropping slower than in the



(a) SSIM (b) IW-PSNR

Figure 5.6: The distributions for the two groups in Better vs. Worse Analysis for IVC dataset.



(a) Different vs. Similar (AUC) (b) Better vs. Worse (C_0) (c) Better vs. Worse (AUC)

Figure 5.7: The results and statistical analysis for the four datasets. Significance plots show that the performance of the method in the row is either significantly better (white), lower (black), or none of the previous (gray).

AUC	**PSNR**	**SSIM**	**IW-PSNR**	**MS-SSIM**	**IW-SSIM**
IVC	0.6360	0.7615	**0.7803**	0.7615	0.7626
CSIQ	0.6827	0.6910	0.7064	0.7056	**0.7148**
LIVE	0.7250	0.7654	**0.7969**	0.7625	0.7744
Toyama	0.5954	0.7068	0.7182	0.7117	**0.7563**
ALL	0.6882	0.7251	0.7297	0.7342	**0.7419**

Table 5.3: AUC values for Different vs. Similar Analysis

case of SSIM and is therefore more robust against not considering some pairs to be significantly different. When any difference in ΔOM is considered to result in difference in perceived quality (i.e. C_0), the same conclusions as in case of standard measures can be reached (see Table 5.2).

5.3.2 Performance on Multiple Datasets Together

In this section, the performance is evaluated on four databases together (IVC [113], CSIQ [50], LIVE [110], and Toyama [114]). A57 [94] and TID2008 [115] datasets are not used, since the subjective data for the former are obtained from the seven expert subjects only, making the statistical processing not very relevant. The latter is omitted because it is not possible to determine how many observers evaluated each image from the description. Only overall number of observers is provided but not all of them evaluated all the content. Computation of z-scores would therefore be unreliable.

The results for the four databases are depicted in Figure 5.7. The statistical significance was determined the same way as in the previous section. The Tables 5.3-5.6 contain the final values obtained from the datasets separately, as well as from their combination. Note that in Table 5.4, also the thresholds for the $|\Delta OM|$ necessary to ensure the 95% probability that the stimuli in the pair are different (i.e. for 5% *FPR*) are reported. The values are dependent on the range of models' values and therefore cannot be directly used for models' comparison but their differences for particular datasets provide another insight and they are important for the practical use of the models.

Several conclusions can be drawn from the overall results. Firstly, the best performing model is IW-SSIM, followed by MS-SSIM. An important phenomenon can be seen in the Table 5.3. Even though the IW-PSNR metric reaches higher AUC value than MS-SSIM in the Different vs. Similar Analysis for each database separately, the overall performance of MS-SSIM is higher. This shows that weighted average of the particular results does not have to lead to the same conclusions as analysing all the data at the same time. The proposed methodology is therefore more convenient when multiple datasets are considered together.

Also the effect described in the Section 5.3.1 where SSIM provides better classification in the Better vs. Worse Analysis but the AUC value is higher for IW-PSNR is reflected in the overall results as well. For the explanation, refer to the stated section.

For most of the metrics, CSIQ database appears to be the most challenging. The only exception from this is PSNR which works the worst for Toyama dataset. These findings are in parallel with the standard

$THR_{95\%\text{DS}}$	**PSNR**	**SSIM**	**IW-PSNR**	**MS-SSIM**	**IW-SSIM**
IVC	8.4461	0.1285	6.4705	0.0757	0.1023
CSIQ	13.0470	0.2818	19.0379	0.2198	0.2686
LIVE	9.7317	0.2677	11.0294	0.1713	0.1710
Toyama	10.6431	0.0873	8.6840	0.0444	0.0429
ALL	12.4806	0.2677	17.9861	0.2002	0.2429

Table 5.4: Thresholds $THR_{95\%\text{DS}}$ ensuring 95% probability that the stimuli in the pair are statistically significantly different.

C_0	**PSNR**	**SSIM**	**IW-PSNR**	**MS-SSIM**	**IW-SSIM**
IVC	0.8038	0.9203	0.9135	0.9181	**0.9261**
CSIQ	0.8279	0.8721	0.8542	0.8978	**0.9049**
LIVE	0.8518	0.9081	0.8998	0.9122	**0.9190**
Toyama	0.7630	0.9069	0.8841	0.9127	**0.9386**
ALL	0.8369	0.8897	0.8762	0.9049	**0.9122**

Table 5.5: Correct Classification in Better vs. Worse Analysis

measures [47].

The last observation is the room for improvement in metrics' abilities with respect to the the Different vs. Similar Analysis. Although it is true that not all well-performing objective methods has been tested here. Nevertheless, IW-SSIM is considered to be one of the reliable criteria, outperforming other popular metrics [47], and the overall AUC value of 0.7419 is not very high.

AUC	**PSNR**	**SSIM**	**IW-PSNR**	**MS-SSIM**	**IW-SSIM**
IVC	0.8877	0.9669	**0.9795**	0.9649	0.9768
CSIQ	0.9140	0.9227	0.9265	0.9357	**0.9444**
LIVE	0.9377	0.9568	0.9657	0.9597	**0.9660**
Toyama	0.8570	0.9657	0.9613	0.9685	**0.9843**
ALL	0.9219	0.9391	0.9437	0.9470	**0.9546**

Table 5.6: AUC values for Better vs. Worse Analysis

6

Revisiting the Role of the Reference in Image Quality Assessment

As discussed in the introduction, the post-processing algorithms enable increasing the perceived quality of a stimulus by adjusting its aesthetic properties. This discards the use of classical *fidelity* approach towards quality evaluation where the similarity of the reference and processed version of the stimulus is quantified. However, the procedures described in the previous chapters have been designed with this assumption in mind. Therefore, it is necessary to revise the understanding of the reference in post-processing quality evaluation scenarios and identify, and eventually adjust, the standard assessment procedures that can possibly be used in the described context. Given the slightly different nature of the two relevant groups of post-processing algorithms (refer to Section 1.2), the discussions will be held separately.

6.1 Current Possibilities in Subjective Quality Assessment of Enhanced Images

Considering the subjective experimental methodologies as described in Chapter 2, there are several procedures that can be used even in the scenarios where the best possible quality image is not available. Namely, these are:

- Absolute Category Rating (ACR) / Single Stimulus (SS) [11, 12]
- Double Stimulus Continuous Quality Scale (DSCQS) [11]
- Ranking [13]
- Paired Comparison (PC) [12]

The application of SS, Ranking, and PC is straightforward since there is no comparison with the reference whatsoever at any point during the procedure. In case of DSCQS, the presence of the original image can actually help to see if the enhanced image is enhanced or over-enhanced. The role of the reference is, therefore, no longer to represent the perfect quality but to serve as kind of an anchor to compare with.

Quality assessment of enhanced images is not very well covered topic in literature. Subjective studies containing enhanced images are, therefore, quite rare. Some of the popular databases, namely LIVE [110], CSIQ [50], TID2008 [115], and TID2013 [116], were developed using the above identified procedures and

include some images with contrast variations but there are only few cases where the quality is enhanced. The over-enhancement is not covered at all. The same is valid for datasets CID2013 [139] and CCID2014 [140] which are specialized on contrast. Their goal is to cover the whole spectrum from low to high contrast but they do not consider the possibility of over-enhancement as well.

Bouzit and MacDonald [141] used PC strategy to compare the perceived sharpness and overall observers' preferences for four different sharpening techniques. However, their dataset does not systematically search for over-sharpening. Moreover, the dataset is not publicly available and, thus, cannot be used in further studies.

Zhang et al. [7] employed ranking of printed images in order to identify the thresholds for detection of sharpness changes and the degree of most preferred sharpening. Their data are also not available but it is the only work studying over-sharpening in a systematic manner and as such provides some valuable insights for the goals of this thesis.

The only publicly available database focused specifically on image enhancement is Digitally Retouched Image Quality (DRIQ) database proposed by Vu et al. [142]. It consists of images with enhancement in sharpness, overall contrast, and color saturation. Nevertheless, the amount and the exact way of enhancement is not very well specified since the images were just retouched using Photoshop according to the taste of one of the authors. It also does not include any over-enhanced content. The subjective procedure included within-image ranking, within-image multiple stimulus continuous quality scale (MSCQS) evaluation which is a modification of DSCQS where not two but more images are evaluated simultaneously, and across-image MSCQS. From these experiments, DMOS scores were calculated.

6.2 Current Possibilities in Objective Quality Assessment of Enhanced Images

This section discusses the applicability of objective quality metrics for enhanced images assessment. In case of no reference metrics, their suitability is obvious. However, most of them have been trained or at least tested on datasets prepared under the classical quality assessment paradigm. Their verification in this new context is, therefore, required.

For full reference metrics, and a majority of reduced reference criteria, the assumption is that the reference is of the best possible quality. Therefore, the metrics do not allow for the case that the quality could be enhanced by the processing. The only exception is VIF metric [48]. Although it requires the presence of the whole original image, it also has the ability to evaluate the processed image as qualitatively better than the reference if there was a contrast increase without the amplification of noise. This capability makes it a possible candidate for the assessment of enhanced (and especially sharpened) images.

Vu et al. [142] proposed the way to employ full-reference metrics for the given task. The approach reverses the order of the evaluated images, i.e. the original image is considered to be the distorted version (DIS) and the processed image is taken as the reference (REF). The strategy is visualized in Figure 6.1.

Nevertheless, this is possible only for the "asymmetric" criteria providing different values for different order of images (in Figure 6.1 Q1 $\neq$ Q2). It excludes the usage of some popular similarity metrics such as PSNR, SSIM, FSIM, etc. which only quantify the difference between images.

Moreover, they also came up with augmentation of MAD [50] metric, specifically designed to evaluate image enhancement. In the augmented version of the metric, only the "appearance-based strategy" (for more information refer to Section 3.1.2 or the respective paper) of MAD was employed because it exhibited better results for this kind of images. Additionally, three features were investigated – contrast, sharpness, and saturation.

A local contrast map was obtained by dividing the image into 8×8 blocks with 50% overlap and calculating RMS contrast according to [72]. A local sharpness map was calculated using S3 algorithm [68]. And finally, the image was converted into HSV color space and its S component was taken as a local saturation map. For every feature, difference d_i was calculated as

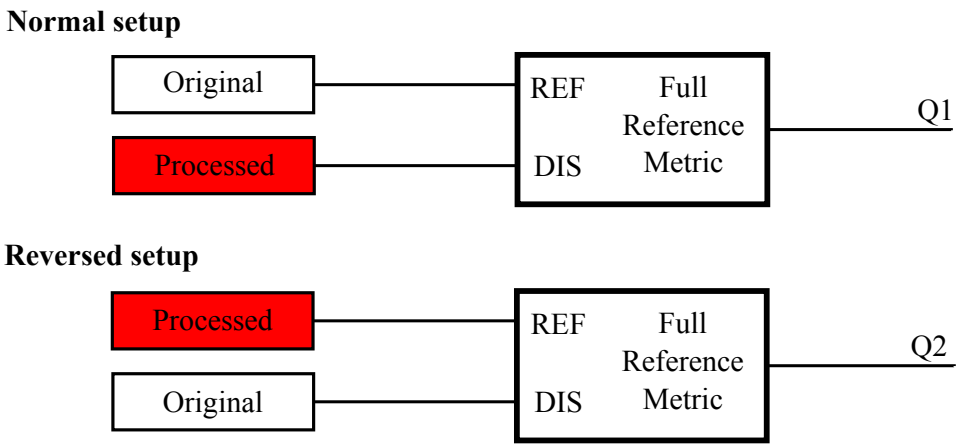


Figure 6.1: Reversed strategy for full-reference metrics proposed in [142].

$$d_i = ||q_i(I_{\mathrm{P}})||_2 - ||q_i(I_{\mathrm{R}})||_2, \tag{6.1}$$

where i = {contrast, sharpness, saturation}, q_i stands for particular feature maps, I_{P} and I_{R} denote the processed version and the reference image, respectively, and $||.||_2$ means L_2-norm.

The overall change d is calculated simply by adding the components together, thus

$$d = d_{\text{contrast}} + d_{\text{sharpness}} + d_{\text{saturation}}. \tag{6.2}$$

Finally, the augmented version of MAD is defined as

$$Augmented\ MAD = d \times MAD_a. \tag{6.3}$$

Since both the reversed strategy for full reference metrics and Augmented MAD have only been tested on the DRIQ database which does not consider over-enhancement, their applicability in the whole post-processing scenario needs to be validated.

6.3 Current Possibilities in Subjective Quality Assessment of Tone-Mapped Images

Subjective experiments evaluating performances of different TMOs are, compared to the case of image enhancement, much more numerous. A thorough overview can be found e.g. in [143]. The applicable procedures are SS [11], ranking [13], and PC [12]. Given the difference in dynamic range of the reference, DSCQS (or any other procedure using a reference) in its classical form cannot be used. However, some of the previous studies "injected" the reference into the otherwise reference-less procedures by adding it to the evaluation setup. Therefore, three fundamental approaches can be identified:

- Evaluation with a real world reference,
- evaluation with a reference on HDR display, and
- evaluation without any reference.

In the first two cases, the reference is shown together with the tested images providing the information about the real scene.

Apart from the setup, the question according to which observers evaluate the stimuli also differs. Subjects are either asked to consider the fidelity to the reference or their overall preference. There have been some contradictory results reported by studies measuring the influence of different experimental designs

on the outcome. Ashikhmin and Goyal [144] did not find a significant difference between preference and fidelity scores when no reference is provided. However, when they showed a real world reference to the observers, the outcomes from fidelity experiment differed. On the other hand, Kuang et al. [145] found the fidelity to the reference and preference results strongly correlated. Čadík et al. [10] also did not discover any statistical difference between experiments with and without real world reference. These contradictions indicate that the influence of the reference may be triggered by particular conditions and, when drawing conclusions from different experiments, it should be checked if such effects occur or not.

Since the notion of quality in this thesis involves also the aesthetic properties, the concern is more about observers' preferences than about fidelity to the real world scenes. The reference, therefore, does not represent the best possible quality rather than a link to the naturalness of the real world. The possible influence of the presence of such reference on the preferences is interesting to study.

6.4 Current Possibilities in Objective Quality Assessment of Tone-Mapped Images

In case of using objective metrics for evaluating tone-mapped images, the situation is similar to the enhanced images (see Section 6.2). The assessment is restricted either to no reference criteria, which were mostly trained and tested in different context, or the specially adjusted full reference metrics.

All of the full reference image quality metrics, described in Section 3.1, assume that the dynamic range of the original and processed image are the same. However, in case of tone-mapped HDR images, the dynamic range between the two versions differ (i.e. I_{R} is an HDR image while I_{P} is LDR). The following metrics were designed in order to overcome this issue.

DRIM

The first metric enabling comparison of images regardless their dynamic ranges was Dynamic Range Independent Metric (DRIM) [146]. It uses the detection model from HDR-VDP [52], calibrated on the data from ModelFest dataset [147], to indicate which regions contain visible contrast in HDR image and its tone-mapped version.

The metric then creates three distortion maps showing the regions where the contrast was either lost (i.e. contrast change is perceivable in HDR but imperceivable in LDR image), amplified (the opposite case), or reversed (the polarity is changed – mostly caused by halo artifacts) by tone-mapping process. The framework of the metric is visualized in Figure 6.2. DRIM inherited high computational requirements and the necessity to specify viewing conditions and display parameters. The metric was designed for visualization of perceived distortions regarding the image contrast, therefore it does not allow for calculating a single quality value.

TMQI, TMQI-II

Another approach was introduced by Yeganeh and Wang as Tone-Mapped images Quality Index (TMQI) [148]. This quality criterion consists of two parts – Structural Fidelity (SF) and Statistical Naturalness (SN). SF is obtained by modification of the previously described MS-SSIM index [46] for comparing HDR and SDR images. This modification does not penalize the difference in signal strength if they are both under or over the visibility threshold. This is determined by non-linear mapping of signals' standard deviations according to CSF. SF for a patch u is obtained as

$$SF\Big(I_{\mathrm{R}}(u), I_{\mathrm{P}}(u)\Big) = \frac{2\mathrm{std}'\Big(I_{\mathrm{R}}(u)\Big)\mathrm{std}'\Big(I_{\mathrm{P}}(u)\Big) + const_1}{\mathrm{std}'\Big(I_{\mathrm{R}}(u)\Big)^2 + \mathrm{std}'\Big(I_{\mathrm{P}}(u)\Big)^2 + const_1} \times \frac{\mathrm{std}\Big(I_{\mathrm{R}}(u), I_{\mathrm{P}}(u)\Big) + const_2}{\mathrm{std}\Big(I_{\mathrm{R}}(u)\Big)\mathrm{std}\Big(I_{\mathrm{P}}(u)\Big) + const_2}, \quad (6.4)$$

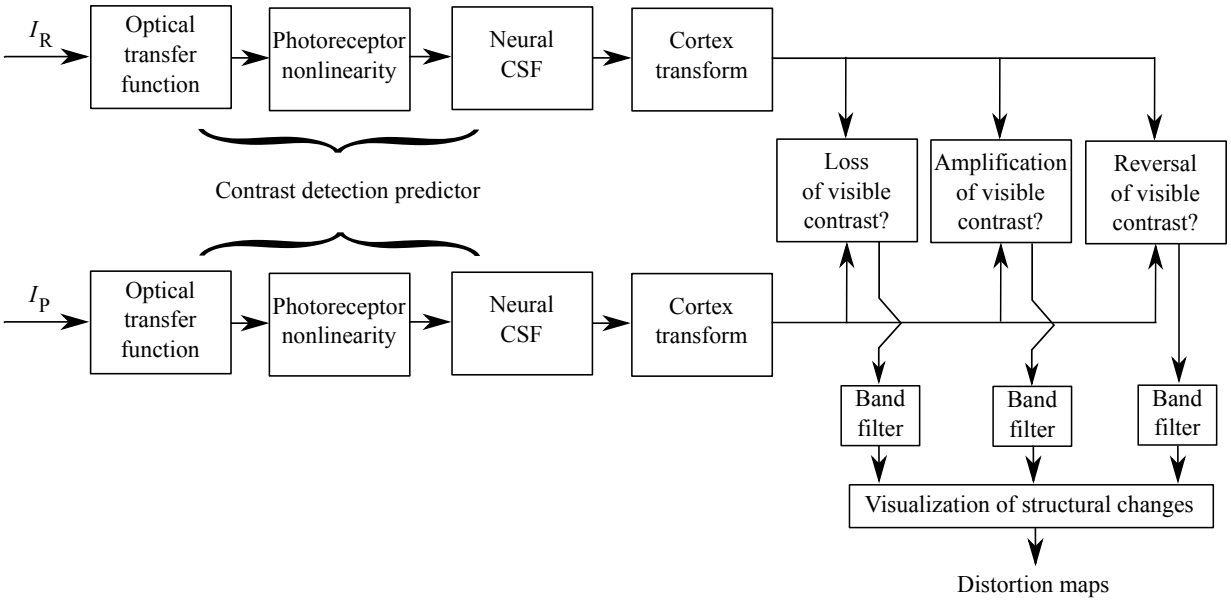


Figure 6.2: The framework of DRIM [146].

where std′(.) is the standard deviation of the patch after the non-linear mapping. Note that the luminance component is missing, compared to SSIM definition, but the structural element (i.e. the second fraction in equation 6.4) remains the same.

The mapping is defined as

$$\text{std}' = \frac{1}{\sqrt{2\pi}\theta_{\text{std}}} \int_{-\infty}^{\text{std}} \exp\Big(- \frac{(x - \tau_{\text{std}})^2}{2\theta_{\text{std}}^2} \Big) dx, \tag{6.5}$$

where

$$\theta_{\text{std}}(f) = \frac{\tau_{\text{std}}(f)}{k}, \tag{6.6}$$

with f being a spatial frequency and k representing a constant obtained from Crozier's law [149], typically ranging from 2.3 to 4. The authors propose to use $k = 3$. The threshold for the signal's standard deviation is

$$\tau_{\text{std}}(f) = \frac{\overline{\mu}}{\sqrt{2} \times const \times CSF(f)}, \tag{6.7}$$

where $\overline{\mu}$ is the mean intensity value (set to 128 by the authors) and $const$ is a constant used to fit the physiological data. TMQI uses CSF as introduced by Mannos and Sakrison [150] and fit to the data measured by Kelly [151]. The map of SF is averaged on each scale and the final SF index is obtained in the same way as in case of MS-SSIM.

The second measure implemented in TMQI – i.e. SN – does not use reference and is based on the assumption that naturalness of an image can be modelled by probability distributions of brightness and contrast (means and standard deviations) in natural gray-scale images. They discovered by analysing 3,000 images that the means and standard deviations follow Gaussian and Beta distribution respectively. The measured distributions are $\mathcal{N}(115.94, 27.99)$ and $\mathcal{B}(4.4, 10.1)$. Assuming that brightness and contrast are mutually independent, the probability that the image is natural is then expressed as

$$SN(I_P) = \frac{1}{K(I_P))} \times pdf_{\mathcal{N}(115.94,27.99)}(I_P) \times pdf_{\mathcal{B}(4.4,10.1)}(I_P), \tag{6.8}$$

where K is a factor used for normalization, thus

$$K(I_P) = \max\{pdf_{\mathcal{N}(115.94,27.99)}(I_P), pdf_{\mathcal{B}(4.4,10.1)}(I_P)\}. \tag{6.9}$$

The final TMQI is a combination of the two measures defined as

$$TMQI\Big(I_{\mathrm{R}}, I_{\mathrm{P}}\Big) = \lambda_1 SF\Big(I_{\mathrm{R}}, I_{\mathrm{P}}\Big)^{\lambda_2} + (1-\lambda_1) SN\Big(I_{\mathrm{P}}\Big)^{\lambda_3}. \tag{6.10}$$

The combination parameters λ_1, λ_2, and λ_3 were obtained from the subjective data ($\lambda_1 = 0.8012$, $\lambda_2 = 0.3046$, and $\lambda_3 = 0.7088$).

Later on, Ma et al. revised both of the terms and proposed the new version of the index called TMQI-II [152]. Namely, the contrast visibility model for HDR images has been adapted to the local luminance when calculating SF-II. The estimate of contrast in the HDR reference is therefore computed as the standard deviation in a patch divided by the local mean. The SN term is modified much more severely.

The authors argue that the distribution of means and standard deviations leading to the natural image depends on the mean and the standard deviation of the original image (obtained from the logarithm of the HDR reference). Therefore, they designed a subjective experiment and let people adjust the means and standard deviations of 60 natural images, in order to find the lower and upper bounds for naturalness. These bounds were then fitted with a linear model. The transitions from the boundary to the mean are expressed by Gaussian CDFs. SN-II is then obtained as a product of the probabilities that the image is natural in terms of its mean and its standard deviation.

TMQI-II is then obtained as

$$TMQI\text{-}II\Big(I_{\mathrm{R}}, I_{\mathrm{P}}\Big) = \frac{1}{2} SF\text{-}II\Big(I_{\mathrm{R}}, I_{\mathrm{P}}\Big) + \frac{1}{2} SN\text{-}II\Big(I_{\mathrm{R}}, I_{\mathrm{P}}\Big). \tag{6.11}$$

FSITM

The last full-reference metric that will be discussed here is a Feature Similarity Index for Tone-Mapped images (FSITM) [153]. It uses the phase congruency features to calculate the difference between original and tone-mapped version of the image. More specifically, it computes the Locally Weighted Mean Phase Angle (LWMPA) to compute the phase congruency. The main advantage of this feature is its robustness against noise.

Let the $lG^{\mathrm{e}}_{\rho,r}$ and $lG^{\mathrm{o}}_{\rho,r}$ be a quadratic pair of log-Gabor wavelets, i.e. evenly and oddly symmetric, on the scale ρ and orientation r. The responses for a two-dimensional signal (e.g. an image I) are obtained as

$$[e_{\rho,r}(I), o_{\rho,r}(I)] = [I * lG^{\mathrm{e}}_{\rho,r}, I * lG^{\mathrm{o}}_{\rho,r}], \tag{6.12}$$

where the operator $*$ stands for a convolution. The LWMPA is then computed as

$$LWMPA(I) = \arctan2\Big(\sum_{\rho,r} e_{\rho,r}(I), \sum_{\rho,r} o_{\rho,r}(I)\Big). \tag{6.13}$$

The operator arctan2(.) is defined as

$$\arctan2(x,y) = 2\arctan\frac{x}{\sqrt{x^2+y^2}+y}. \tag{6.14}$$

The values of $LWMPA(I)$ range from $-\pi/2$ to $\pi/2$. The binary phase congruency map PCG can then be obtained as

$$PCG(I) = \mathcal{H}\Big(LWMPA(I)\Big), \tag{6.15}$$

where $\mathcal{H}(.)$ is a Heaviside (unit-step) function. The definition is

$$\mathcal{H}(t) = \begin{cases} 1 & t > 0 \\ \frac{1}{2} & t = 0 \\ 0 & t < 0. \end{cases} \tag{6.16}$$

The authors propose to calculate the FSITM for each channel ch. The similarity of the congruency maps for a channel ch is computed as

$$SCG^{ch}(I_{\mathrm{R}}, I_{\mathrm{P}}) = \frac{\left| PCG\left(I_{\mathrm{R}}^{ch}\right) \cap PCG\left(I_{\mathrm{P}}^{ch}\right) \right|}{X \times Y}, \tag{6.17}$$

with X and Y being the image width and height, respectively.

The final index can then be obtained from

$$FSITM^{ch}(I_{\mathrm{R}}, I_{\mathrm{P}}) = \lambda\, SCG^{ch}\Big(I_{\mathrm{R}}, I_{\mathrm{P}}\Big) + (1-\lambda) SCG^{ch}\Big(\ln(I_{\mathrm{R}}), I_{\mathrm{P}}\Big), \tag{6.18}$$

where the parameter λ was set experimentally.

7

Quality Assessment of Sharpened Images

The enhancement algorithms will be represented by image sharpening. Nevertheless, all of the principles, except for the specialized algorithms, are applicable to any of the other enhancement methods (such as enhancement of global contrast, colorfulness, etc.). The main challenges introduced to the quality assessment by image enhancement have been summarized in Section 1.2.1.

Image sharpening is one of the fundamental tools in current digital photography. Plenty of material on the topic, written by professional photographers, can be found in works of literature (e.g [154]) or in a variety of free online tutorials [155–161].

Sharpening is the process of making images "look sharper" or "crunchier". The nomenclature is slightly problematic because the process of deblurring the image is called sharpening, as well as the increase of contrast on the edges, which makes image details more visible and image appears sharper. The issue of sharpness and its perception by human observers will be discussed more thoroughly in Section 7.1.

In almost every image processing software, like *Photoshop*[1], *Lightroom*[2], or freeware *GIMP*[3], there are some sharpening algorithms implemented. Commercial software specialized directly in the sharpening is also available – for example *Google Nik Collection*[4], *PhotoKit Sharpener*[5] by PixelGenius, or PhotoWiz's *FocalBlade*[6]. These algorithms are also implemented directly into digital cameras, scanners, or printers.

In [154], Fraser and Shewe divide the sharpening process into three stages, creating, what they call, the sharpening workflow. The first part of the workflow is *Capture Sharpening*. In this stage, a photographer should sharpen the image so that it looks like the original scene he/she was capturing. That means that it should compensate the Modulation Transfer Function (MTF) of the lens and also other processes causing blurring of the original scene. These could involve going through the anti-aliasing filter of the camera, or demosaicing. In other words, this stage is basically a deblurring of the image captured by the camera, i.e. *image restoration*. *Capture Sharpening* should be used for every image taken and, in most cases, it is incorporated directly in the camera itself (but it is usually better to switch this function off and perform *Capture Sharpening* using some of the sharpening tools). This stage is also necessary for the following two to produce decent results.

The second part is called *Creative Sharpening*. It provides a space for creativity, artistic intentions, etc. It can be used on the whole image, on specific areas of the image, or enables to highlight certain objects.

[1] http://www.adobe.com/Photoshop (retrieved on 30/08/2016)
[2] http://www.adobe.com/products/photoshop-lightroom.html (retrieved on 30/08/2016)
[3] http://www.gimp.org (retrieved on 30/08/2016)
[4] https://www.google.com/nikcollection/ (retrieved on 30/08/2016)
[5] http://pixelgenius.com/sharpener2/ (retrieved on 30/08/2016)
[6] http://thepluginsite.com/products/photowiz/focalblade/ (retrieved on 30/08/2016)

Zhang et al. [7] showed that despite the variance of the opinions, there is a trend of human observers preferring images which have been sharpened more than to look exactly like a captured scene. HVS seems to appreciate when the contrast is slightly higher than in natural scenes. The sharpening is here done by local adjustment of contrast on the edges. This is the *image enhancement* stage, therefore, the main interest of this thesis regarding sharpening will be focused here.

The last part of the workflow is *Output Sharpening*. This stage is dependent on the form the final image will be in. Here, the parameters should be adjusted for the particular device considering that there is a difference between the image on the electronic screen and the final printed version of it. The images meant for printing should usually be slightly over-sharpened on the screen but this really depends on the device itself. This should be, of course, done as at the end of the process, right before the printing. If the image is meant for the electronic screen, this particular stage can be omitted.

Very important thing is not to perform any sharpening before the image is in its final size. Resizing will corrupt and in most cases completely ruin the sharpening effect and the image has to be re-sharpened. The viewing distance should also be taken into account.

In this chapter, the basic concepts of sharpness perception, and sharpening algorithms will be introduced. Further, the most appropriate subjective procedures for sharpened images evaluation will be selected and an extensive subjective test will be described. Selected state-of-the-art objective metrics will be tested on the ground truth subjective data and an adjusted pooling strategy significantly improving the performance of the best performing metrics will be proposed and tested. Finally, a novel method for using full-reference image quality metrics for image enhancement will be described and verified in the context of image sharpening. The chapter builds upon and extends the work that has been done within the preparation of the author's master's thesis [162] and has further been published in [163].

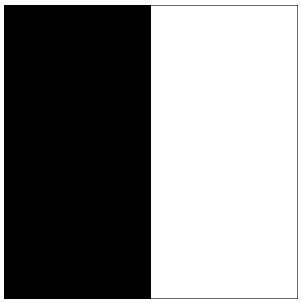
7.1 Sharpness Perception

Although the term *sharpness* is widely used in the image processing community, there is not a single, unified definition available. Most of the studies however come from the idea of *sharpness* as the outcome of two factors – **resolution** and **acutance** [155, 161]. **Resolution** is the ability of the camera to capture differences in detail. It influences the capability to distinguish between close objects in the scene. Resolution is entirely determined by the chip of the camera. It is the better known one of these two factors of sharpness and it is sometimes falsely considered to be the only one.

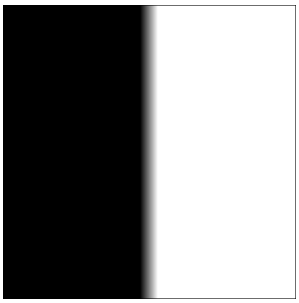
The second factor influencing the sharpness is **acutance**. It indicates the steepness of the edge. The edge with high and low acutance can be seen in Figures 7.1(a) and 7.1(b), respectively. Their corresponding profiles are in Figures 7.1(c) and 7.1(d). It is obvious that the shorter the transition between the luminance levels is, the sharper the edge appears.

The edges like the one in Figure 7.1(a) are unfortunately impossible to be found in real (i.e. non-artificial) images. It is because the acutance is affected by the factors which are never ideal. These factors are the properties of the lens, image reconstruction algorithms, motion blur caused by photographer's shaking hands, insufficient illumination of the scene and many others. This is basically the reason why a huge resolution does not ensure sharp and good quality photos. From image post-processing's point of view, acutance is much more important property of sharpness than resolution, because it is incomparably easier to be enhanced. This will be discussed in detail in the next section.

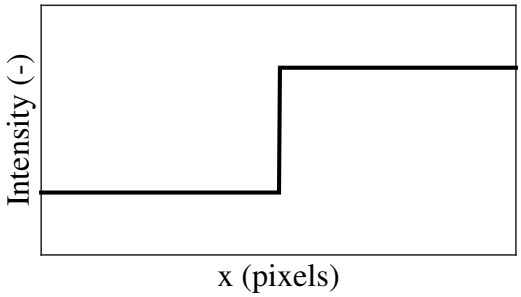
The problem of measuring perceived sharpness is its dependence on the scale, viewing distance, displaying device, and other things. The changes of perceived sharpness caused by different image sizes were investigated for various scenes by Park et al. in [164]. Three levels of image size were created – small, medium and large. It was discovered that the changes in sharpness were not proportional to the changes in size. The degree of change of perceived sharpness seemed to be higher for small-medium image set, than for medium-large. Also the sharpness of images containing moving objects or other kinds of blur seemed not to be affected that much by the difference in size. They also appeared sharper in the images



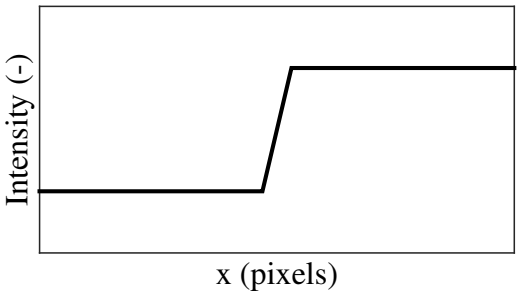
(a) Edge with high acutance



(b) Edge with low acutance



(c) Profile of the edge with high acutance



(d) Profile of the edge with low acutance

Figure 7.1: Example of edges with high and low acutance.

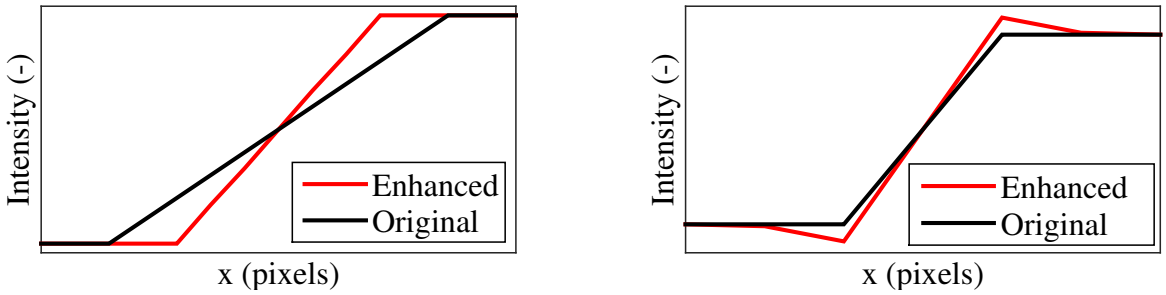
of smaller sizes. On the other hand, images containing text or repeating patterns suffered the most and appeared sharper when their size was larger.

Crete et al. [165] discovered that the same amount of blur applied on the sharper and less sharp image has a different impact on the blur perception. When the sharper image is blurred, the difference in the perceived blur is much larger. This fact is also exploited in the blur metric [165].

The impact of the sharpness on visual attention when observing the scene was examined by Vuori and Olkkonen [166]. The behavior of observers' gaze can be divided into two phases. First phase is called *fixation*. During *fixations*, the gaze is focused on a particular small region and most of the visual information about the scene is gained. The second phase is *saccade* and it represents the transitions between *fixations*. Almost no information is obtained during *saccades*. Several mechanisms influence the *saccades'* durations and *fixations'* positions. A drag of the attention based on characteristics of the scene as regions with high contrast, strong illumination, and other low-level features is called a "bottom-up" mechanism. A "top-down" mechanism is based on the task given to the observer (e.g. "look for the traffic signs") and/or his/her personal experience. This "top-down" approach is very hard to model because of its complexity. In [166], eye-tracking experiment was conducted in order to record observer's eye-movements while watching images with different levels of sharpness. The duration of *saccades* in the experiment with images of lower quality were significantly longer. Apparently, it is harder for the observer to find areas for fixations in blurred scenes. This could be explained by the fact that blurring decreases the contrast on the edges.

Extensive tests of perceived sharpness were done by Zhang et al. [7]. The study dealt with the threshold for detection of sharpness changes and the degree of most preferred sharpening. The results suggested that the detection threshold is much less dependent on the content and subjects, than the preference. However, the degree of applied sharpening for the most preferred version was consistently higher than the degree for the detection which means that some level of sharpening is generally preferred.

Very interesting effect can be achieved by the presence of noise. Despite being mostly undesirable, noise can have, under certain circumstances, positive influence on the perceived sharpness. If the picture is rather smooth, a proper amount of noise can make it look "crunchier". That is why professional photographers sometimes add some noise or film grain into their photographs. An example of such an effect can be found in [155].



(a) Decreasing of edge transition slope length (b) Edge contrast enhancement

Figure 7.2: Edge profiles with enhanced acutance.

7.2 Sharpening Algorithms

This section provides an introduction into the post-processing techniques used for increasing the perceived sharpness. As stated in the previous section, two main properties of the photograph influence how sharp it is perceived – resolution and acutance. Both of these factors can be enhanced by post-processing.

Methods used for increasing the resolution of an image are known as *super-resolution techniques*. These algorithms are able to calculate the high-resolution image from the sequence (or video) of low-resolution frames. The procedure is rather complicated because it demands the correct registration of images in the sequence, interpolation into the high-resolution sampling grid, and reconstruction concerning warping, blurring, and noise suppression. More information about the subject can be found e.g. in [167] or [168]. It is obvious that these techniques are not well-suited for the enhancement of single photographs.

Enhancement of acutance is much more suitable for the sharpening of images. In general, there are two possible approaches towards acutance enhancement. First group is formed by algorithms trying to **decrease the edge transition slope length**. This length is defined as the distance between minimum and maximum image intensity values in the neighborhood of the edge. The simplified example of edge profiles with such an enhancement is shown in Figure 7.2(a). This is applicable particularly well for image restoration of heavily blurred images where the second approach fails completely. One of the methods, developed by Arad and Gotsman [169], uses image dependent warping. As most of the other techniques, it also has problems with noise amplification and compression artifacts. Shavemaker et al. [170] proposed different approach based on morphological filtering. It takes the intensity values in the slope and substitutes half of them with the minimum and the other half with the maximum value found in the slope, thus creating the ideal step edge (like the one in Figure 7.1(a)). This is, however, not very convenient because such edges look unnatural. Augmentation of this method for enlarged images can be found in [171], or [172], where instead of creating the step edge, the values are substituted by some kind of contrast stretching function between minimum and maximum values in a slope.

The second approach towards perceived sharpness increase is the **enhancement of contrast on the edges**. The length of the transition slope remains unchanged but the difference in intensities between extrema is enlarged, creating the overshoots. This makes the images look sharper. An example of edge profile processed in this way is shown in Figure 7.2(b). Basically, the high-pass component of the image is augmented. The most popular way of doing so is *unsharp mask sharpening*. This method will be presented later in a more detailed view.

The problem of the overshoots lies in the fact that when they are too big, they produce ringing artifacts (also called *halo artifacts*) which could be very annoying for the observer and with their presence the perceived quality drops significantly (see Section 1.2.1). It is mainly the responsibility of the photographer to control this effect when sharpening the images.

Bruna et al. [173] proposed an algorithm trying to introduce more reliable overshoot control in sharpening. Firstly, it analyses image activity by high-pass filtering to distinguish between homogenous and textured regions. The image is then filtered by bank of directional filters. The sharpening gain (i.e. how

much the pixel should be sharpened) is then computed for edge pixels. After that, the ringing control is applied, in which user can regulate overshoots from complete disabling to no control at all.

The effect of sharpening can be also achieved by a simple increase of high frequency components in the image spectrum. Bouzit and MacDonald [141] proposed the sharpness enhancing technique by weighting the image spectrum according to the CSF.

More complex method of sharpening for printing 6 inch photos was proposed by Safonov et al. [174]. They resize the image to the required size and then extract certain features from the image to estimate its sharpness. The photo is then classified, employing AdaBoost algorithm, into one of four groups – sharp, slightly blurred, blurred, and strongly blurred. Every group is sharpened with a different intensity. This technique combines the two main aforementioned sharpening approaches – firstly it uses contrast stretching between minimum and maximum values in a slope to decrease the length of edge transitions and then *unsharp mask via bilateral filter* for a local contrast increase. For more information, refer to the respective paper.

Another approach is to decompose an image into the base and detail layers. The enhancement is then applied on the detail layer only. This has been exploited e.g. by Paris et al. [175] who used classical Laplacian pyramid with coefficients classification for decomposition. The classification was employed in order to prevent halo artifacts. Bae et al. [176] utilized bilateral filtering [177] in order to obtain the layers.

Other possible methods include, for example, locally adaptive bilateral clustering [178], multiscale retinex theory [179], or fuzzy logic [180, 181].

The remainder of this section provides a deeper description of the algorithms that are used in the following subjective study. Note that since the concern of the paper is connected to the *Creative Sharpening* (see the introduction of this Chapter), all methods are based on the enhancement of the contrast on the edges.

7.2.1 Unsharp Mask

Image sharpening using unsharp mask is by far the most popular technique. It has its roots in analog photography where the blurred version of the photo used to be printed together with the negative in the form of registered sandwich. This would increase edge sharpness as well as suppress the noise caused by the film grain which is random for both versions. The name of the procedure comes from the fact that the *unsharp* version is used as a *mask*. This is very well described in [182]. Some information can be also found in [154]. But the massive popularity of unsharp mask came with the digital photography.

The principle is to extract the high frequency components by subtracting the blurred version from the original as

$$I_{\text{high}} = I - h_\sigma * I, \tag{7.1}$$

where I_{high} represents the high frequency components, I stands for the original image, h_σ is the Gaussian kernel with the respective standard deviation σ used for the blurring, and operator $*$ denotes convolution.

This subtraction is then added to the original:

$$I_{\text{sharp}} = I + \lambda I_{\text{high}}, \tag{7.2}$$

or also

$$I_{\text{sharp}} = (1 + \lambda)I - \lambda(h_\sigma * I), \tag{7.3}$$

where I_{sharp} is the sharpened output image and λ is the weighting factor influencing the amount of sharpening.

As the equations (7.1), (7.2), and (7.3) suggest, the final result of sharpening is affected mainly by parameters σ and λ. Parameter σ influences the size of the neighborhood in which the sharpening is performed and λ determines how much the image will be sharpened.

Unlike its analog ancestor, digital unsharp mask unfortunately does not suppress the noise but, as a matter of fact, amplifies it. It is also better to perform sharpening only on the luminance component of the image, avoiding undesirable color shifts.

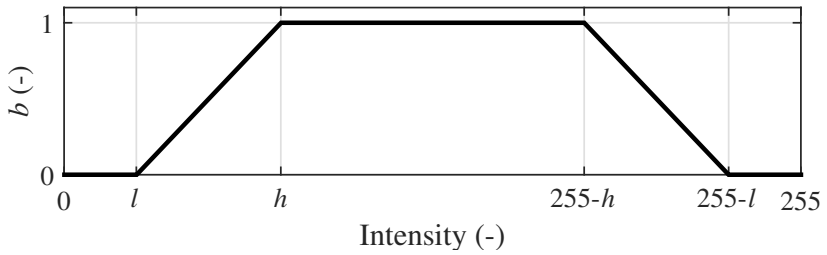


Figure 7.3: Clipping protection function b.

There have been plenty of attempts to enhance the performance of unsharp mask sharpening. Most of them are trying to solve the problem with noise using different filters to extract the high-frequency component. Some augmentations employ nonlinear filters [183–185] (e.g. quadratic [186], or morphological [187]), or adaptive filters [188]. These are compared within the same conditions in [189]. More recent method for medical images was proposed by Agaian and McClendon [190]. The unsharp mask is here used in cascade. The previously mentioned Safonov's algorithm [174] employs bilateral filter instead of Gaussian in the unsharp mask computation.

7.2.2 Enhanced Unsharp Mask

In the image processing programs like *Adobe Photoshop*, or *GIMP*, mostly the ordinary unsharp mask filter is used (new versions of *Photoshop* enables to choose between Gaussian or Lens blur). To boost the performance, photographers use multiple layers. This enhancement, integrated also in *Lightroom* [154], was, for the purpose of this work, implemented in MATLAB to enable higher control over the parameters and to be able to follow what exactly is happening during every step.

Firstly, the original image I is sharpened by an ordinary unsharp mask (equations (7.2) and (7.3)) with parameters σ and λ, obtaining the image I_{sharp}. One of the problems of unsharp masking is the loss of information in highlights and shadows due to *halo* artifacts [161]. This could be avoided using *clipping protection mask*. This mask ensures that pixels with too small, or too high intensity values, respectively, will be affected less by the sharpening. The mask C is calculated as

$$C(x,y) = \frac{1}{2}\Big(b\big(I(x,y)\big) + b\big(I_{\text{sharp}}(x,y)\big)\Big), \tag{7.4}$$

where x, y are the pixel coordinates and b is the protective, piecewise linear function shown in the Figure 7.3. Thresholds l and h set the amount of protection.

Image with clipping protection I_C is obtained as

$$I_C(x,y) = C(x,y) \times I_{\text{sharp}}(x,y) + \big(1 - C(x,y)\big) \times I(x,y), \tag{7.5}$$

where operator $\times$ represents the pixel-wise multiplication.

This ensures that most of the sharpening is done in the mid-tones and thus solves the picture loss problem. But the algorithm still amplifies the noise in homogenous regions. Also some sharpened textures, like contours of the skin, are unpleasant for the observers. There are several ways to suppress this effect. The simpler approach is to set the threshold value $thres$, used as

$$I_{thres} = \begin{cases} I_C & \text{if} |\nabla I| > thres, \\ I & \text{otherwise}, \end{cases} \tag{7.6}$$

where ∇ is the gradient operator and I_{thres} is the sharpened image with thresholding. As can be seen, the homogenous regions will be avoided because of small gradient value of pixel intensity in this region. This solution is, however, not ideal because the transitions between pixels taken from the original and the

sharpened image can look unnatural.

More complex solution is the usage of an edge mask. This can be obtained by one of the edge operators. In the implementation, vertical and horizontal kernels of Sobel operator were employed [191]. The edge mask E is then calculated as

$$E = |E_{\text{ver}}| + |E_{\text{hor}}|, \tag{7.7}$$

where E_{ver} and E_{hor} are the edges obtained by filtering by vertical and horizontal kernel, respectively. The final mask is then normalized to the range from 0 to 1.

A comparison of differences in final quality when using different operators could be a topic for further study. Canny edge detector [192] is particularly interesting because it treats all the detected edges the same way (unlike the derivative-based operators where sharper edges have stronger responses, the result of a detection by Canny's detector is only a binary mask - detected/not detected).

Sometimes it can be better to calculate the Sobel mask from an image on a smaller scale which can lead to detecting only strong, relevant edges. The implemented MATLAB algorithm enables to set `scale` value. The mask will be then obtained from the image of a size proportional to the original by this value (e.g. for $\texttt{scale} = 0.5$, the image used for edge detection is half the size of the original) and then resized to the original size using bicubic interpolation. Edge images for *cameraman.tif* (Figure 7.4(a)), available in MATLAB's Image Processing Toolbox, with `scale` values 0.5 and 1 can be seen in Figure 7.4(b) and 7.4(c), respectively. The image in Figure 7.4(b) is blurred because of the interpolation but it is not a problem in this case, as will be explained later.

To "push out" the weak edges in the mask, gamma correction can be used

$$E_\gamma(x, y) = E(x, y)^\gamma. \tag{7.8}$$

The influence of gamma correction on the edge image is shown in Figure 7.4(d). The contours of the cameraman's coat are now visible but the noise is also included and that is not desirable.

The difference in importance between the stronger and the weaker edges can be modified by contrast transformation. This is done according to the cubic contrast transformation function T shown in Figure 7.5. The shape of the function is set by the parameters `il`, `ih`, `ol`, and `oh`. If $\texttt{ol} < \texttt{il}$ and $\texttt{oh} > \texttt{ih}$, influence of the stronger edges will be strengthened and influence of the weaker will be suppressed, as can be seen in Figure 7.4(e). In the opposite case, their influence will be more similar. The edge mask is processed as

$$E_{\gamma,T}(x, y) = T\big(E_\gamma(x, y)\big). \tag{7.9}$$

Finally, the edge mask is blurred by the averaging filter. In the implementation, the 5×5 filter kernel is used

$$E_{\gamma,T,\text{avg}} = E_{\gamma,T} * \frac{1}{25}\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}, \tag{7.10}$$

where $*$ is the convolution operator. $E_{\gamma,T,\text{avg}}$ is then normalized, so its values are within the interval from 0 to 1. Resulting sharpened image is then obtained as

$$I_{\text{sharp2}}(x, y) = E_{\gamma,T,\text{avg}}(x, y) \times I_C(x, y) + \big(1 - E_{\gamma,T,\text{avg}}(x, y)\big) \times I(x, y), \tag{7.11}$$

where $\times$ represents the pixel-wise multiplication. The reason why the edge mask should be blurred comes from the equation (7.11). If the mask was too sharp, the transitions between the pixels taken from the original and the sharpened image would not necessarily have to be smooth which could create disturbing artifacts and affect the quality in a bad way.

Unsharp mask enhancement could also include suppression of noise. To perform denoising only at regions which need it, inverse edge mask could be employed (mostly more blurred). Because denoising in

(a) Original image

(b) Edge mask - `scale` $= 0.5$

(c) Edge mask - `scale` $= 1$

(d) Edge mask - $\gamma = 0.6$

(e) Edge mask after contrast transformation

(f) Blurred edge mask

Figure 7.4: Sobel edge mask for *cameraman.tif*.

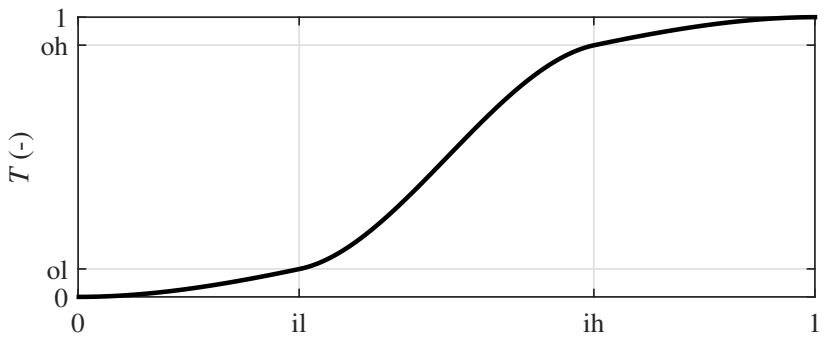


Figure 7.5: Contrast transformation function T.

luminance channel changes the image drastically, sometimes it is enough to perform denoising in chrominance channels. This, in fact, does not suppress the noise but makes it visually less annoying. More information about the subject can be found in [154].

7.2.3 Global Sharpening Enhancement Algorithm

Global Sharpening Enhancement Algorithm (GSEA) was proposed by Agaian and McClendon [190] for sharpening of medical (especially radiological) images. Considering its simplicity, it can also be successfully used for multimedia content. At first, the algorithm performs non-linear filtering using *trimmed mean* filter. This filter substitutes the pixel values by *trimmed mean* of the values inside the neighboring area of the pixel.

Intensity values inside the area are ordered and then *trimmed mean* of the ordered set is calculated as

$$\mu_{\text{trimmed}}(u) = \frac{1}{U - 2out} \sum_{k=out+1}^{U-out} I_k, \tag{7.12}$$

where U is the number of pixels in the square neighborhood (e.g. $U = 25$ for 5×5 area), out is the amount of pixels excluded from the set at the beginning and the end (in the implementation $out = \sqrt{U} - 1$), and I_i is the i-th intensity value.

Unlike the ordinary averaging filter, *trimmed mean* filter is less affected by noise. Setting the threshold out to 0 will make the filter calculate an ordinary average and setting it to the highest possible value $(U - 1)/2$ will make it calculate a median value. Filtered image $I_{\mu_{\text{trimmed}}}$ is obtained by calculating *trimmed mean* in every pixel position

$$I_{\mu_{\text{trimmed}}}(x, y) = \mu_{\text{trimmed}}(x, y). \tag{7.13}$$

For calculations on the borders, the image is mirrored.

In the next step, enhancing factor D_λ is computed

$$D_\lambda(x, y) = \left(\frac{I(x, y)}{I_{\mu_{\text{trimmed}}}(x, y)} \right)^\lambda, \tag{7.14}$$

where the fraction stands for pixel-wise division. The differences between pixel intensities in I and $I_{\mu_{\text{trimmed}}}$ should be relatively small for the flat regions and relatively big on the edges.

Finally, the enhanced image I_{GSEA} is obtained as

$$I_{\text{GSEA}}(x, y) = I(x, y) \times D_\lambda(x, y), \tag{7.15}$$

where $\times$ operator stands for pixel-wise multiplication. The main parameter influencing the result of sharpening is λ along with the size of the considered neighborhood of the pixel.

7.2.4 Sharpening Using SDME

Image enhancement using Second Derivative-like Measure of Contrast was proposed by Nercessian et al. [75]. It is based on works of DelMarco and Agaian [193] where SDME is defined as a visibility operator and Panetta et al. [74] who used it for quality assessment (see equation (3.47)).

The measure of contrast is defined as

$$CON_{\text{SDME}}(x, y) = \frac{I_{\max} - 2I(x, y) + I_{\min}}{I_{\max} + 2I(x, y) + I_{\min}}, \tag{7.16}$$

where $I(x, y)$ is the intensity value in the current pixel and $I_{\max}$ and $I_{\min}$ are the maximum and minimum intensities in the certain square neighborhood, respectively. CON_{SDME} values vary between -1 and 1. The

advantage of such a measure is that the value of the current pixel can be obtained directly from the contrast

$$I(x,y) = \left(\frac{I_{\max} + I_{\min}}{2}\right)\left(\frac{1 - CON_{\text{SDME}}(x,y)}{1 + CON_{\text{SDME}}(x,y)}\right). \tag{7.17}$$

This is the base for possible image enhancement. A new pixel value can be calculated from the enhanced contrast. Because absolute values of CON_{SDME} are in the range between 0 and 1, Nercessian et al. proposed the direct contrast enhancement by power law mapping

$$CON'_{\text{SDME}} = \text{sgn}(CON_{\text{SDME}})|CON_{\text{SDME}}|^{\lambda}, \tag{7.18}$$

where CON'_{SDME} is the enhanced contrast and λ is the enhancement factor. Contrast enhancement is yielded for $\lambda < 1$. It is important that the mapped contrast stays in the required range and thus the phase of its coefficients is preserved.

The pixel value $I_{\text{enh}}(x,y)$ of the enhanced image is then obtained as

$$I_{\text{enh}}(x,y) = \left(\frac{I_{\max} + I_{\min}}{2}\right)\left(\frac{1 - CON'_{\text{SDME}}(x,y)}{1 + CON'_{\text{SDME}}(x,y)}\right). \tag{7.19}$$

The algorithm is, again, controlled by two parameters - λ and the size of the neighborhood of the pixel under investigation. In [75], the extension of the algorithm into multiple scales is proposed. This was not included in the used implementation and a reader is encouraged to study the above mentioned paper for more information.

7.3 Subjective Study on Sharpened Images

As already explained in Chapter 6, there is a lack of datasets considering enhanced and over-enhanced content. Nevertheless, such datasets are essential for design and testing of objective metrics. Moreover, the procedures leading to reliable subjective evaluation of enhanced and over-enhanced content are not well investigated. The majority of the research in the field has been dedicated to the printed media [7]. When considering sharpening for electronic display, a revision is needed.

The goal of this section is to fill this gap and provide a selection of the most suitable subjective procedure for sharpened (or generally enhanced) images evaluation. The methodology will then be used to create a representative dataset including blurred, sharpened, and over-sharpened images applicable as a ground truth for objective metrics performance evaluation and comparison in the given context. In order to test the metrics abilities across different sharpening methods, four different techniques (Sections 7.2.1-7.2.4) will be considered for creating the database.

7.3.1 Stimuli Selection and Preparation

In the subjective experiment design there is always a tradeoff between number of source images and applied processing algorithms. Since the nature of the test required higher number of sharpening levels produced by every algorithm, smaller amount of source images had to be selected for the time plausibility of the test.

For the creation of test stimuli, four source images were used – *Caps*, *Parrots*, and *Red Hat* from the publicly available Kodak database,[7] and *Isabe* from IVC Database [113]. These particular scenes were selected for containing different textures, colorfulness, spatial frequencies, and semantics. Two of the scenes also involve humans with uncovered skin. Such content was identified as problematic by Zhang et al. [7] because observers are very sensitive about the natural appearance of the skin.

[7] http://r0k.us/graphics/kodak/ (retrieved on 30/08/2016)

(a) *Caps* (b) *Parrots* (c) *Red Hat* (d) *Isabe*

Figure 7.6: Source images used in the subjective study on sharpened images.

The images from the Kodak database were cropped to the size of 512×512 pixels. All the images also have 3 pixels wide gray frames. Smaller versions of the source images can be seen in Figure 7.6.

7.3.2 Setup of the Sharpening Methods

Since different sharpening algorithms result in different artifacts, it has been decided to include more techniques within the study. In our case, four methods are used – unsharp mask, augmented unsharp mask [154], GSEA [190], and SDME [75]. The details about the particular implementations can be found in sections 7.2.1-7.2.4.

As thoroughly discussed in section 1.2.1, the optimal amount of sharpening is content dependent. All the methods, therefore, have at least two main parameters – λ influencing the degree of sharpening and another parameter adjusting the size of the area around the pixel taken into consideration. To decrease the number of degrees of freedom, only the sharpening strength parameter λ was left free and the others were fixed creating two different settings for each method. That gives eight ways of sharpening. The specific parameter values were selected to be as universally applicable as possible. Particular parameters' setups are stated in the Table 7.1.

No.	***Method***	***Parameters***
1	Unsharp Mask	$\sigma = 0.6$ (neighborhood 3×3 pixels)
2	Unsharp Mask	$\sigma = 1.4$ (neighborhood 7×7 pixels)
3	Enhanced Unsharp Mask	$\sigma = 0.6$, no clipping protection, `scale` = 1, $\gamma = 1$, `il` $= 0.1$, `ih` $= 0.9$, `ol` $= 0.4$, `oh` $= 0.6$, 5×5 averaging filter
4	Enhanced Unsharp Mask	$\sigma = 1.4$, no clipping protection, `scale` = 1, $\gamma = 1$, `il` $= 0.1$, `ih` $= 0.9$, `ol` $= 0.4$, `oh` $= 0.6$, 5×5 averaging filter
5	GSEA	neighborhood 3×3
6	GSEA	neighborhood 7×7
7	SDME	neighborhood 3×3
8	SDME	neighborhood 7×7

Table 7.1: Setup of the sharpening methods.

7.3.3 Test Room

All the tests were conducted in laboratory conditions of the testing room within the facility of Images and Video Communications (IVC) Research Group under Institut de Recherche en Communications et Cybernétique de Nantes, École polytechnique de l'université de Nantes.

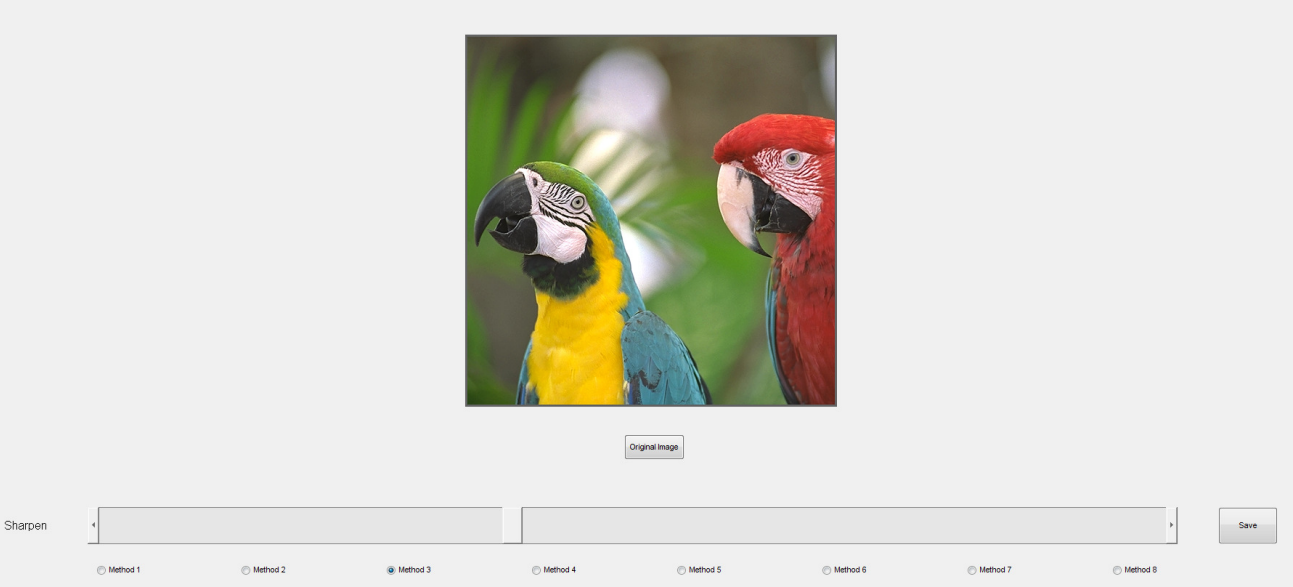


Figure 7.7: Graphical User Interface for Pretest no. 1

Color-calibrated Full HD (1920×1080 pixels) display TV Logic LVM-401 with 50 cm in diagonal was used. Viewing distance was set to be three times the height of the screen and the lighting conditions were in accordance with ITU-T Recommendation P.910 [12] (see Table 2.2).

7.3.4 Methodology Selection

In Chapter 6, the possible methodologies for subjective quality assessment of sharpened images have been summarized (SS [11], DSCQS [11], ranking [13], and PC [12]). To select the most suitable methodology, a pilot study has been conducted. Since the images in the dataset are meant for displaying on the screen, the ranking experiment, popular in printing industry, would not be an ideal solution – observers would not have an access to all of the images in their native resolution at the same time, the series was expected not to be monotonic in quality thus complicating the task, etc. Hence, only two pilot tests were designed.

Zhang et al. [7] showed in their experiment that the thresholds for the detection of change in sharpness and for the preferred amount of sharpening are subject and content dependent. The pilot tests were therefore also used to determine the thresholds for sharpness change detection and over-sharpening for each source image, so the images in the dataset would not have the same level of perceived sharpness and to include over-sharpening.

Pretest no. 1

In the first pretest, the subject used a slider to continuously adjust the level of sharpening (similar to sharpening in image processing programs) to the most preferred value. This is close to a single stimulus procedure where people actually compare the displayed image to their own virtual reference. A graphical user interface (GUI) was created in MATLAB (see Figure 7.7). Moreover, the observers had a possibility to display the original image at any time by pressing the “Original Image” button. Therefore they could compare the result to the original as in a DSCQS test using the unprocessed image as a reference. The radio buttons at the bottom of the screen enabled switching between the methods. After finishing the adjustment for all the methods, participants submitted their results using “Save” button.

After saving the results, they were asked to describe their strategy while adjusting the image. They were supposed to state the most important areas they had been focusing on and in which they detected over-sharpening artifacts when they appeared. Resulting parameters and answers were written into a text file.

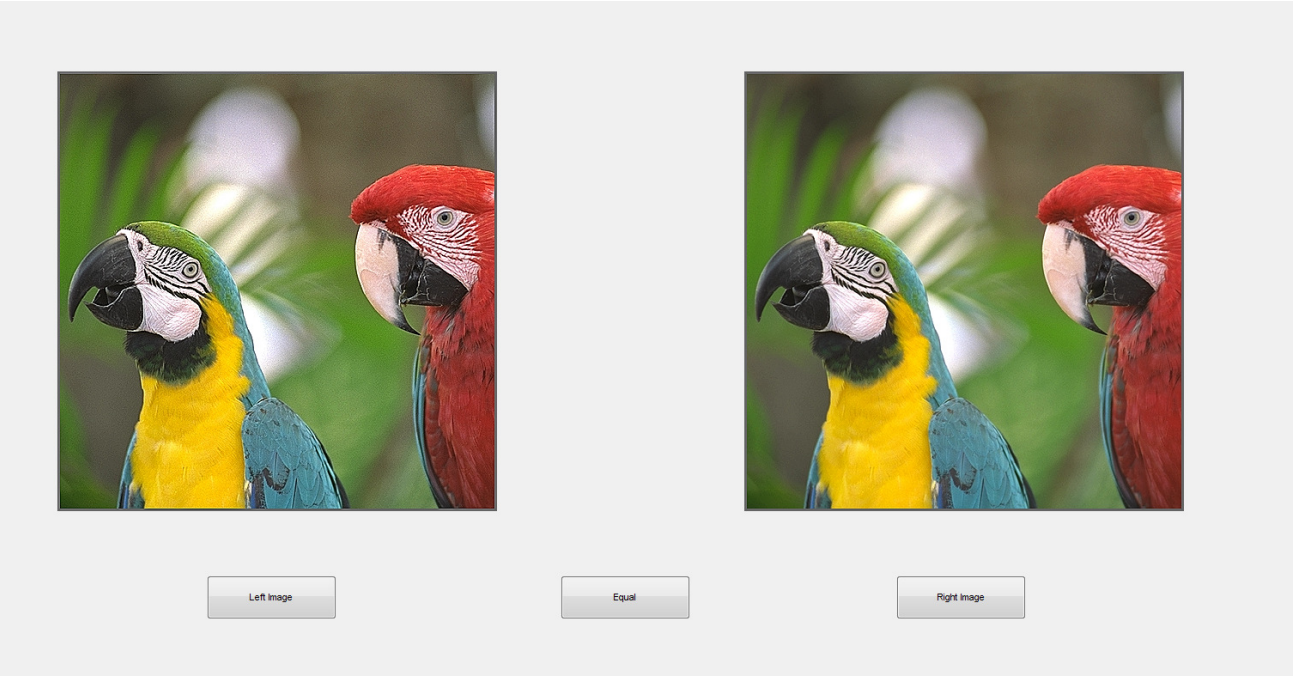


Figure 7.8: Graphical User Interface for Pretest no. 2

Pretest no. 2

The second pretest was based on the PC methodology. Its goal was to simultaneously obtain the threshold for detection of changes in sharpness and the preferred amount of sharpening. For this purpose, another MATLAB GUI was designed (see Figure 7.8).

The observers' task was to choose from two simultaneously displayed images the one they preferred. This was done by clicking either "Left Image" or "Right Image" button. The first two images were very different in terms of sharpness (different λ parameter). When participant chose one of them, the next pair was generated with

$$\lambda_{1,2} = \lambda_{\text{chosen}} \pm dist, \tag{7.20}$$

where λ_{chosen} is the parameter λ of the image chosen by the observer in the previous step, $\lambda_{1,2}$ are the parameters of the next left and right image, respectively, and $dist$ is the distance parameter ($|\lambda_1 - \lambda_2| = 2\,dist$).

After every comparison, $dist$ was lowered (by 0.25 for methods 1 - 6 and by 0.015 for 7 and 8). If $dist < 0.5$, next $dist$ was a half of the previous (for methods 7 and 8 when $dist < 0.03$). Initial setting for each method can be found in Table 7.2. The order of images (which one was displayed as left and which one as right) was randomized.

	λ_1	λ_2	$dist$
Method no. 1	2	6	3
Method no. 2	1	4	2
Method no. 3	2	6	3
Method no. 4	1	4	2
Method no. 5	2	6	3
Method no. 6	1	4	2
Method no. 7	0.7	0.9	0.2
Method no. 8	0.7	0.9	0.2

Table 7.2: Initial setting of parameters for Pre-test no. 2

The procedure was adaptive and therefore dependent on the subject. When the observer labelled the pair as equal, the λ parameter of previously chosen image and distance $dist$ were written into the text file.

Results Interpretation

Three expert observers participated in both of the pretests. The raw results can be found in the Appendix of [162]. Each subject has also been interviewed after completing the test. In the first pretest, participants described the task as very hard and were not really confident about the results. The problem lies in the continuous addition of sharpness enabling human visual system to adapt itself. This causes some kind of tolerance towards more sharpening which is undesirable. It is much easier for humans to choose from discrete levels of quality than to adjust it continuously. This suggests that it could be more comfortable and effective to substitute common sharpening sliders with a different method using direct comparisons.

In terms of strategy, the subjects were very united. This shows that the overall perceived sharpness depends much more on certain areas in the image which is in parallel with previous findings [69]. As stated also in [7], humans are very sensitive to the naturalness of human skin. Regions containing text are preferred to be sharpened more.

The second pretest was much more comfortable for the observers resulting in their higher confidence with the results. This is in agreement with the reasons stated in [194] where the problem of selecting the correct procedure for evaluating video enhancement algorithm was addressed. PC methodology was therefore found more suitable for the particular purpose.

7.3.5 Quantitative Study

The results of the pretests were also used as a base for designing the dataset for quantitative study. The goal was to include blurred, sharpened, and over-sharpened images. The particular levels of sharpening for each content were carefully selected based on the provided data in order to ensure that the over-sharpening effect is reached and the step between different sharpness levels is large enough to be noticeable.

The same four source images, eight methods, and the test room described in Sections 7.3.1-7.3.3 were used in the quantitative study. Four versions of each source image with different amount of sharpening per method were created and three different levels of Gaussian blur (GB) were also added. That gives 35 processed and one original image per each source scene. The full dataset therefore comprised 144 images.

Observers

38 naive observers, both male and female, participated in the experiment. All participants were tested for the visual acuity using Monoyer optometric table and for color perception using Ishihara's patterns. All of them had normal or corrected-to-normal vision.

Subjects were instructed to choose the image they preferred or choose the visually less disturbing one from the pair by using the left or the right arrow on the keyboard, respectively. At the beginning of each session, three training pairs were presented for the observers to get familiar with the interface.

Methodology

In the pilot study, the PC methodology [12] was found the most suitable for the evaluation. The Adaptive Square Design [24, 25] variant, as described in Section 2.3.2, has been employed in order to decrease the time requirements and increase the robustness against observers' errors.

Subjects' task was to compare different versions of the same source images, hence four matrices (each for every source image) containing 36 stimuli were created. That means that $4 \times 180 = 720$ pairs had to be compared by every subject. Considering average of 5 seconds per comparison, such session would last one hour. However, this would be very tiring for the observers and the results would not be accurate. It was therefore decided for each participant to evaluate only half of the dataset (360 pairs), decreasing the

duration to 30 minutes. In fact, the vast majority of subjects was able to finish the test in shorter time, due to the presence of easily qualitatively distinguishable pairs. The order of image pairs was randomized as well as the visualization as the right or the left image.

Results

The results were analyzed according to BTL model (see Section 2.3.3). BTL scores with their respective 95% confidence intervals for particular scenes are shown in Figures 7.9(a), 7.9(c), 7.9(e), and 7.9(g), respectively. Figures 7.9(b), 7.9(d), 7.9(f), and 7.9(h) are the visualization of statistical significance of differences between stimuli. White squares represent the cases where the stimulus in the row is significantly better than the stimulus in the column, while the black squares stand for the opposite case. If the stimuli are not significantly different, the square is gray. The image no. 1 is the original, images 2, 3, and 4 contain Gaussian blur. After that every set of four images is processed using a different method with an increasing level of sharpening.

Despite the fact that the dataset was based on the pre-tests' results and it contained images sharpened much more than what expert observers marked as the optimal value, it can be seen that in most of the cases, subjective scores of sharpened images did not drop under the score of the original. It means that even though they were over-sharpened (subjective score lower than the optimum) there was still some kind of qualitative gain.

The results also show (especially for *Caps* and *Parrots* scenes) that the preferred level of sharpening often varied and the qualitative differences between images are often not statistically significant. The reason could lie in the nature of these images. This confirms the highly subjective nature of the task and supports the use of statistical tools for further processing. Nevertheless, there are some general trends to be observed.

The *Caps* scene contains a text which is, based on the answers of experts in the first Pretest, very important for the observers. The increase in the text readability due to sharpening could cause the higher tolerance to over-sharpening, preference of sharpened images over the original, and also the higher variation among subjects.

The *Parrots* scene is very colorful and some subjects probably appreciated contrast increase despite the over-sharpening artifacts and noise amplification.

The best results, in terms of difference significance and also over-sharpening detection, were obtained for the *Red Hat* scene. As mentioned before, human observers are very sensitive to the naturalness of the skin and faces, which is probably why the participants were much more consistent in this case. Also the superiority of the Enhanced Unsharp Mask (images no. 13 - 20) has been proven because it does not adjust regions not belonging to edges (see Section 7.2.2).

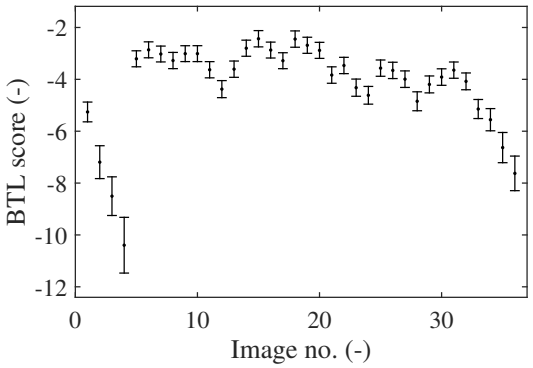
For the *Isabe* scene, the qualitative difference between the original and the sharpened versions is the largest. This could be caused by the lower quality of the original compared to the other three scenes.

The best scores were obtained for the Enhanced Unsharp Mask method, however, the differences in performance were, except for the *Red Hat* scene, not as big as expected. The reason could be that the source images were not noisy. The worst performance was given by the SDME method with the larger neighborhood (method #8).

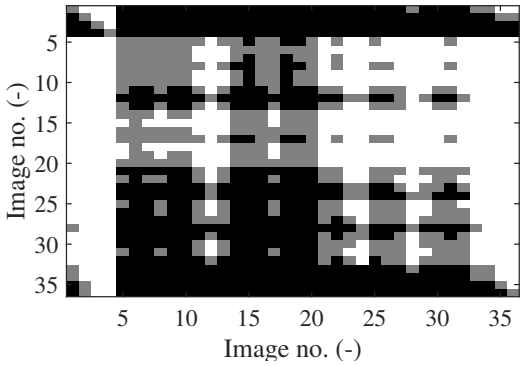
It is also interesting to note that, even though the scores did not drop under the score of the original, the tendency is definitely to decrease after reaching a certain point which proves the over-sharpening theory and thus some valid data can be obtained. The results of subjective tests were used for evaluation of objective metrics' performances which will be described later.

7.4 Performance of the Objective Metrics on Sharpened Images

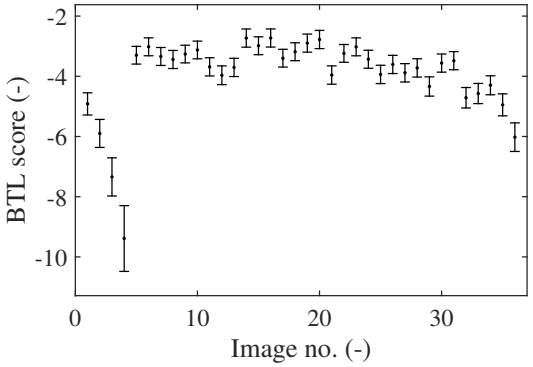
This section describes the analyses of objective quality metrics' performance with respect to the developed subjective database using the novel performance evaluation methodology from the Section 5.2. Firstly, the criteria that can be used in the context of sharpened images evaluation need to be identified.



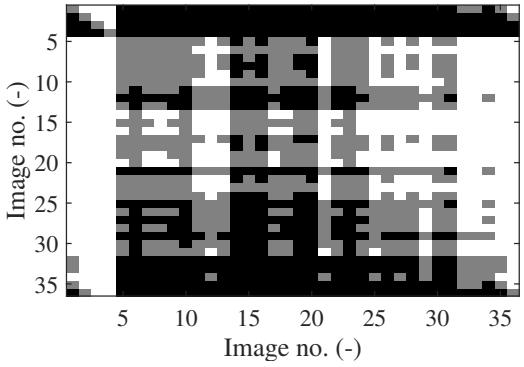
(a) BTL scores for *Caps* scene



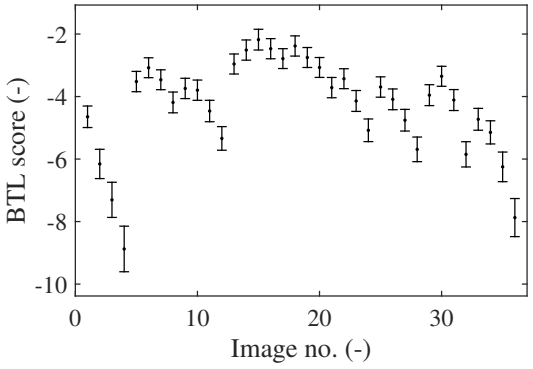
(b) Significance of differences - *Caps*



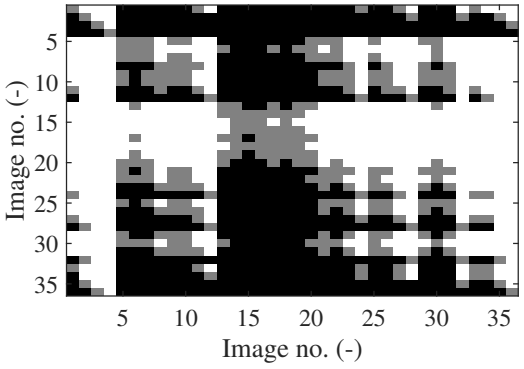
(c) BTL scores for *Parrots* scene



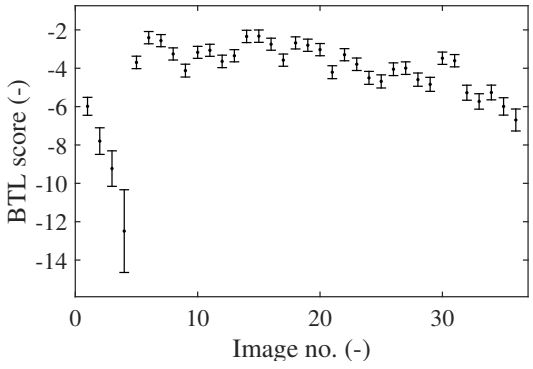
(d) Significance of differences - *Parrots*



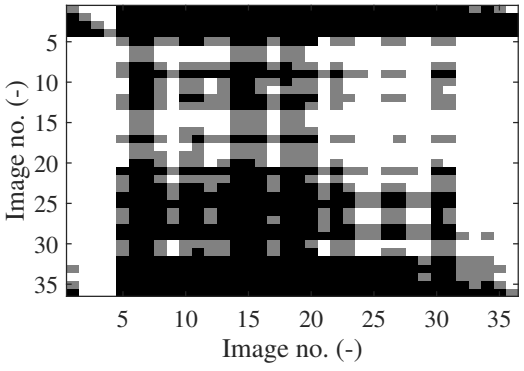
(e) BTL scores for *Red Hat* scene



(f) Significance of differences - *Red Hat*



(g) BTL scores for *Isabe* scene



(h) Significance of differences - *Isabe*

Figure 7.9: Quantitative Subjective Study Results.

The ways to use full reference quality metrics for enhanced images evaluation have been discussed in Chapter 6. The reversed strategy proposed by Vu et al. [142] provided good results on their dataset but it has not been tested on the over-enhanced content nor on different sharpening methods.

Great effort has been dedicated to the development of objective no reference sharpness/blur estimating criteria (see Section 3.2.2). However, these algorithms have been developed for the purpose of quantifying the amount of blur in the image. That means that they have been designed and tested with respect to the ideal of a clearly specified (blur-free) reference. Shaked and Tastl [63] and Zhang et al. [61] tested their criteria on sharpened images but the testing set never contained over-sharpened images, therefore only the metric's ability to assess images *positively* influenced by sharpening was investigated. The applicability of the no-reference sharpness metrics on the over-sharpened images therefore has to be verified.

The aesthetic based measures discussed in Section 3.2.5 could also be considered since they measure the appeal of the image. They have never been tested in the context of sharpening so far.

The area of distortion unaware and opinion unaware (general purpose) no-reference metrics, described in Section 3.2.7, contain another potential candidates. These criteria should be able to predict the quality regardless the processing technique and training dataset used. The rest of no-reference criteria is designed and trained for particular purposes and is therefore not suitable for evaluation of sharpened images.

In the analysis, 25 objective metrics are considered. Full-reference metrics are represented by VIF [48], reversed VIF [48], reversed MAD [50], and Augmented MAD [142]. VIF is tested in the normal mode as well for its ability to recognize the increase of quality with respect to the reference. The rest of the tested criteria consisted of no-reference measures: Variance [57], Frequency Threshold [58], Gradient [59], Laplacian [59], Autocorrelation [59], Histogram Frequency [60], Kurtosis [61], Marziliano [62], HP [63], Kurtosis of Wavelet Coefficients [64], Riemannian Tenzor [65], JNBM [66], CPBD [67], S3 [68], FISH [69], FISH_{bb}, NIQE [84], QAC [85], Aydın et al. aesthetic sharpness [5], Aydın et al. aesthetic metric [5], and CS [86].

Used MATLAB implementations of measures were obtained either from the framework created by Murthy and Karam [44], or from publicly available sources. For the sake of better readability, the criteria in the following figures are numbered according to the Table 7.3.

No.	**1**	**2**	**3**	**4**	**5**
Metric	VIF [48]	Rev. VIF [48]	Rev. MAD [50]	Augmented MAD [142]	Variance [57]
No.	**6**	**7**	**8**	**9**	**10**
Metric	Freq. Threshold [58]	Gradient [59]	Laplacian [59]	Autocorrelation [59]	Histogram Freq. [60]
No.	**11**	**12**	**13**	**14**	**15**
Metric	Kurtosis [61]	Marziliano [62]	HP [63]	Kurtosis of Wavelet Coefficients [64]	Riemannian Tenzor [65]
No.	**16**	**17**	**18**	**19**	**20**
Metric	JNBM [66]	CPBD [67]	S3 [68]	FISH [69]	FISH_{bb} [69]
No.	**21**	**22**	**23**	**24**	**25**
Metric	NIQE [84]	QAC [85]	Aydın et al. Sharpness [5]	Aydın et al. Aesthetics [5]	CS [86]

Table 7.3: Numbering of the metrics in the analyses of performance on sharpened images.

7.4.1 Results per Content

Firstly, the performance of the applicable metrics (Table 7.3) for each source image separately will be shown.

Different vs. Similar ROC Analysis

The results of the Different vs. Similar ROC Analysis for the *Caps* scene are depicted in Figure 7.10(a). The AUC values with 95% CI (left) as well as statistical significance between each pair of metrics (right) determined by Hanley-McNeil method [135] and compensated for multiple comparisons by Benjamini-Hochberg procedure [125] are reported. Note that the white square represents the case where the metrics in

the row performs significantly better than the method in the column, the black square stands for the opposite case, and the gray square signifies no statistical difference in performance.

It can be seen that the performance of the metrics is not very different. The highest AUC value is achieved by S3 metric (#18), followed by $FISH_{bb}$ (#20), Kurtosis of Wavelet Coefficients (#14), and Marziliano (#12). On the other hand, the worst performing metric is NIQE (#21).

The results for *Parrots* scene (Figure 7.10(b)) show the same winners (S3 and $FISH_{bb}$), together with FISH (#19) and VIF (#1) in the regular (not reversed) setup. Augmented MAD (#4), NIQE (#21), and QAC (#22) perform poorly. Generally, the AUC values are the highest for this particular scene.

For the *Red Hat* scene, on the other hand, none of the metrics works well and there are no statistically significant differences found among them. The results are shown in Figure 7.10(c). This is an interesting outcome since the observers' agreement was the highest for this particular scene. The image is easily over-sharpened due to the close look to the skin and the metrics seem to struggle with assessing this phenomenon.

In case of the *Isabe* scene (see Figure 7.10(d)), the highest AUC value is reached by JNBM (#16), followed by S3 (#18), CPBD (#17), $FISH_{bb}$ (#20), and Marziliano (#12). The poorest performance is provided by Augmented MAD (#4) and HP (#13).

Better vs. Worse ROC Analysis

In case of Better vs. Worse ROC analysis, the metrics' performance is generally better. That means that they are more capable to recognize which image is better in the pair than determine if the pair is significantly different. On the other hand, there are some cases where the AUC value drops under 0.5 meaning that the metric works worse than random guessing. This is caused by a systematic misclassification, i.e. evaluating the better image as the worse. The format of the results and the processing methods are the same as in the case of Different vs. Worse analysis.

The results for the *Caps* scene are visualized in the Figure 7.11(a). The group of best performing metrics is formed by S3 (#18), $FISH_{bb}$ (#20), JNBM (#16), and Marziliano (#12). S3 metric achieves the highest AUC value. The Augmented MAD (#4) and Reversed VIF (#2) work worse than random guessing.

For the *Parrots* scene (Figure 7.11(b)), the performance of S3 (#18) and $FISH_{bb}$ (#20) is very high. This scene seems to be less challenging for these metrics in terms of both analyses. Reversed VIF (#2), on the other hand, systematically assigns higher score to the worse image and provides the worst classification from the tested criteria.

The *Red Hat* scene proves to be challenging for the metrics even in this analysis, as can be seen in Figure 7.11(c). The highest AUC value, achieved by $FISH_{bb}$ (#20), is only 0.7536. The similar performance is reached also by S3 (#18) and NIQE (#21) which is surprising considering its poor results for other scenes. The worst metrics for this particular scene are Reversed MAD (#3) and Augmented MAD (#4).

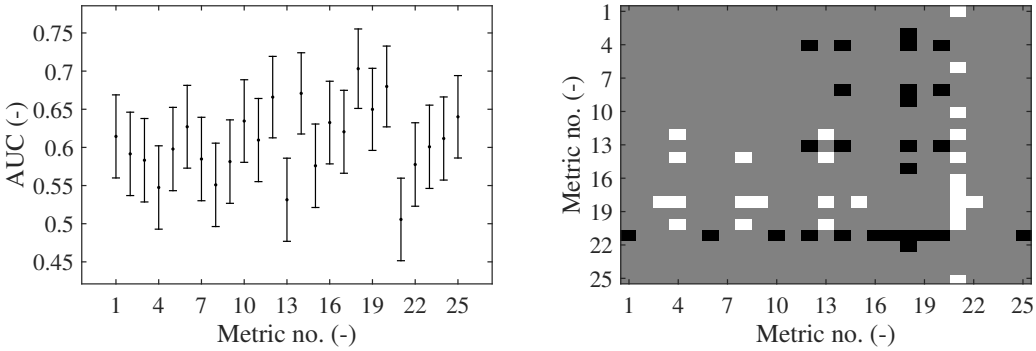
Good performance of the JNBM (#16) for the *Isabe* scene from the Different vs. Similar analysis is repeated also in case of Better vs. Worse analysis (7.11(d)). It is the best performing method together with S3 (#18), $FISH_{bb}$ (#20), and Marziliano (#12) reaches the AUC value higher than 0.9. Reversed VIF (#2) is again performing the poorest.

7.4.2 Overall Performance

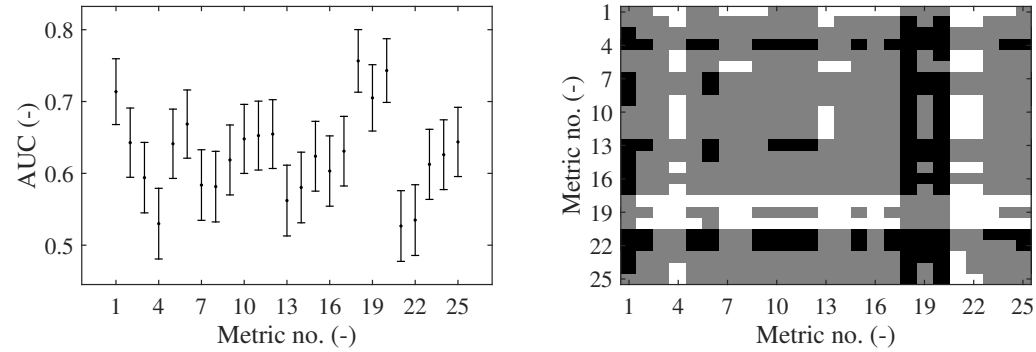
The previous section provides a valuable insight into the content-wise metrics' behavior. Nevertheless, it is also interesting to determine the overall capabilities of the criteria. Section 5.2.5 provides a way how to combine the results and run the analyses on all the data together. The results were processed and will be reported the same way as in the previous section.

Different vs. Similar ROC Analysis

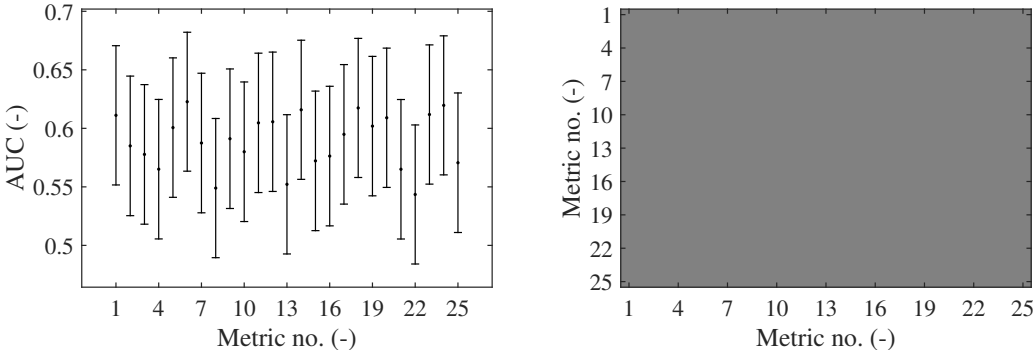
The outcome of the overall Different vs. Similar Analysis is depicted in Figure 7.12. The best performing methods are S3 (#18) and $FISH_{bb}$ (#20) with AUC values of 0.7040 and 0.6905, respectively. This does



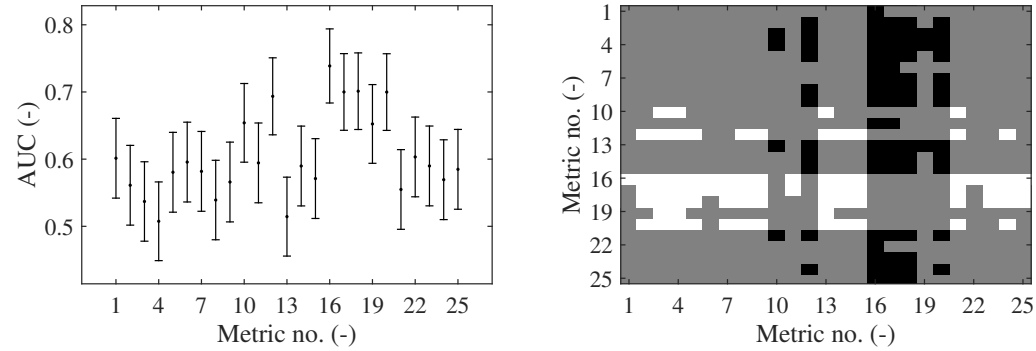
(a) AUC values with 95% CI and significance of differences for the *Caps* scene



(b) AUC values with 95% CI and significance of differences for the *Parrots* scene

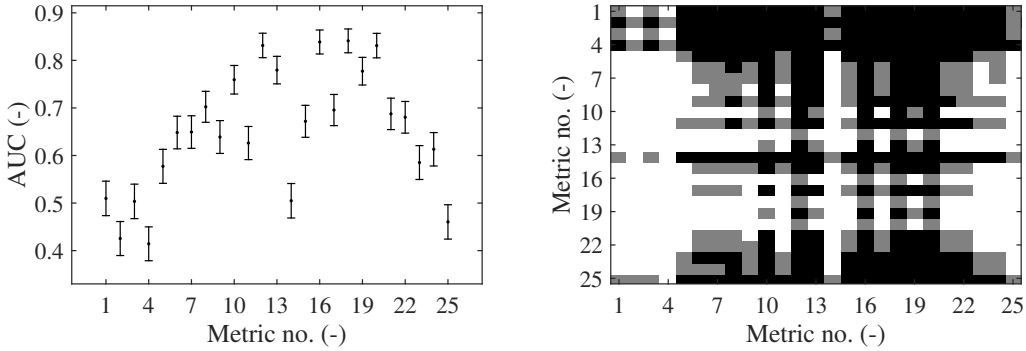


(c) AUC values with 95% CI and significance of differences for the *Red Hat* scene

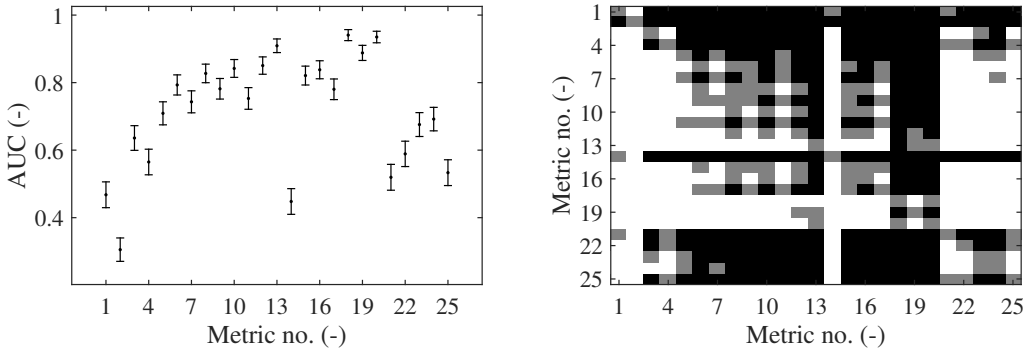


(d) AUC values with 95% CI and significance of differences for the *Isabe* scene

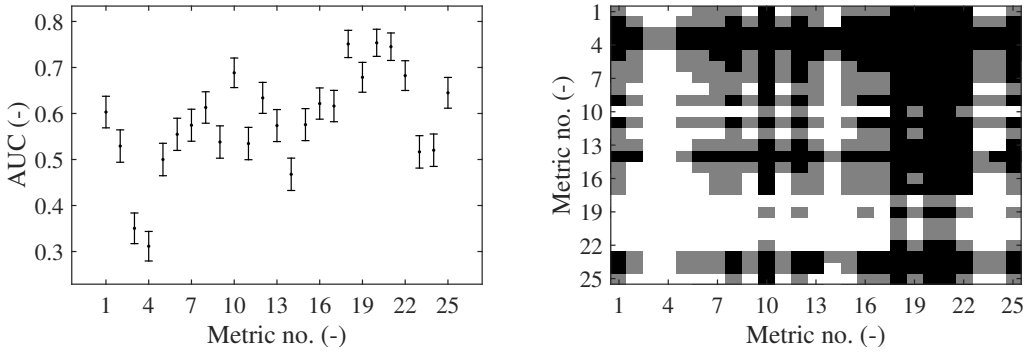
Figure 7.10: The results of the Different vs. Similar ROC Analysis for each source content.



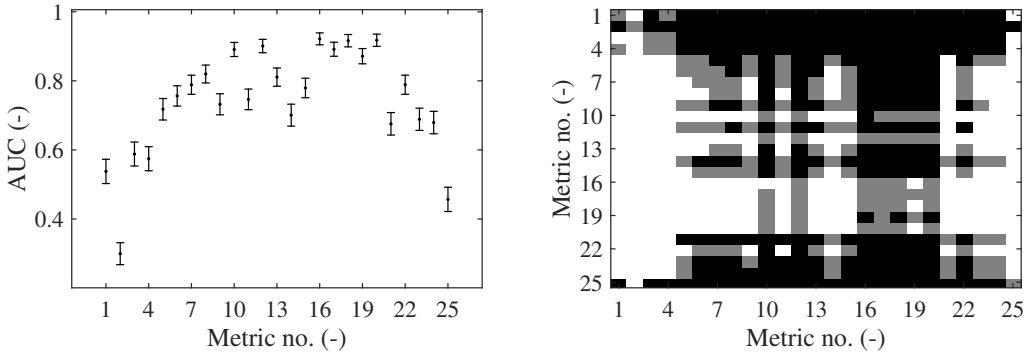
(a) AUC values with 95% CI and significance of differences for the *Caps* scene



(b) AUC values with 95% CI and significance of differences for the *Parrots* scene



(c) AUC values with 95% CI and significance of differences for the *Red Hat* scene



(d) AUC values with 95% CI and significance of differences for the *Isabe* scene

Figure 7.11: The results of the Better vs. Worse ROC Analysis for each source content.

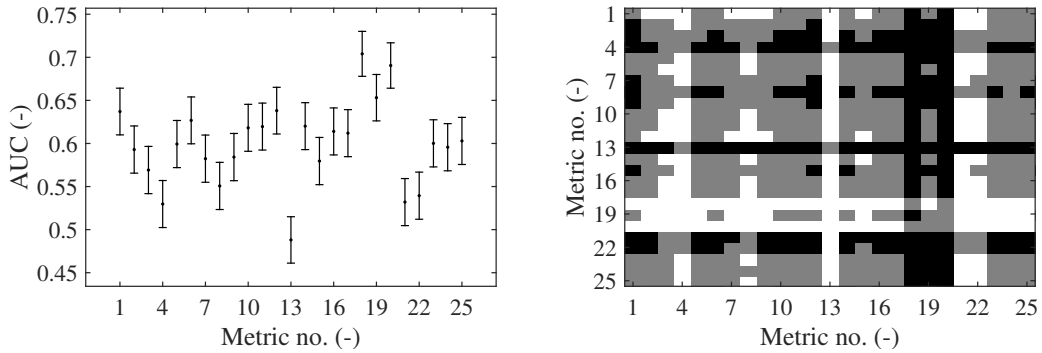


Figure 7.12: The overall results of the Different vs. Similar ROC Analysis.

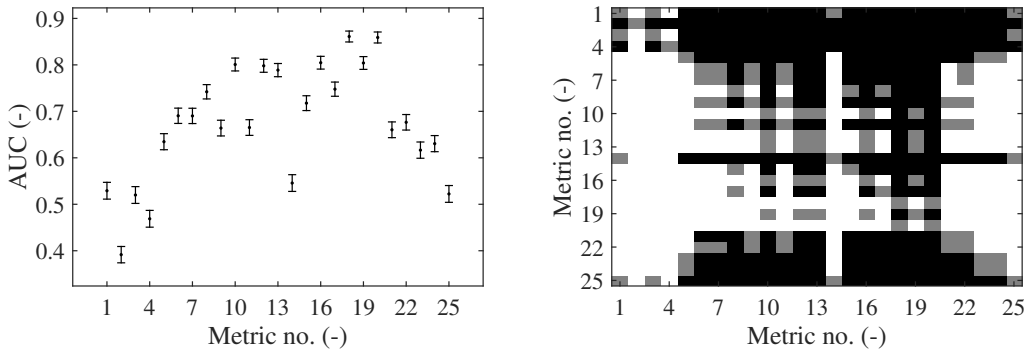


Figure 7.13: The overall results of the Better vs. Worse ROC Analysis.

not come as a surprise, since they were always among the best performing methods in the content-wise analyses. The lowest AUC value belongs to the HP metric (#13). The results suggest that there is definitely a space for improvement of objective quality assessment methods for sharpened images.

Better vs. Worse ROC Analysis

The same two metrics, i.e. S3 (#18) and FISH_{bb} (#20), reach the highest AUC values in the Better vs. Worse analysis as well (see Figure 7.13). Nevertheless, the values of 0.8611 and 0.8590 can still be improved. The worst performing metric is Reversed VIF (#2). The same conclusions can be drawn also from the percentage of correct classification in zero C_0 as shown in Figure 7.14. The statistical significance was determined using Fisher's exact test [37].

It is worth noticing that the full reference measures, even in the reversed setup, perform poorly. This

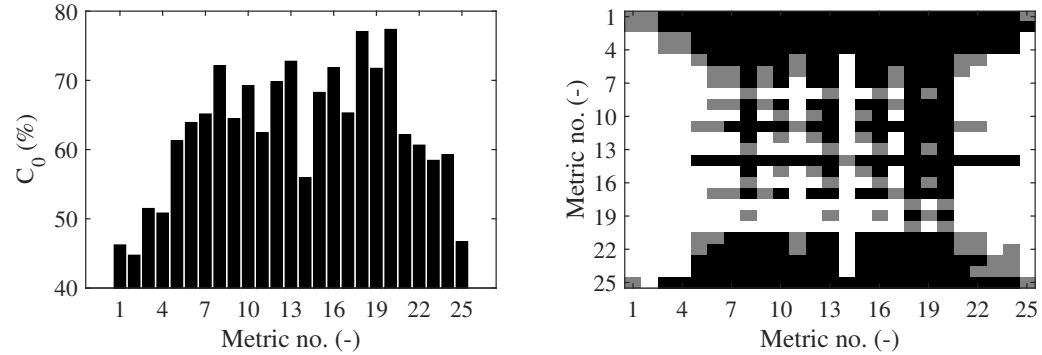


Figure 7.14: The overall correct classification of better and worse image in the pair.

suggest that the reversed setup does not provide reliable assessment when also over-enhanced content is considered. Also the holistic no reference and aesthetics based metrics do not seem to work well on the given content.

7.5 Improved Pooling Strategy

Since both of the best performing metrics (S3 [68] and FISH_{bb} [69]) result in a sharpness map which is further pooled into a single value, it is worth studying if the pooling strategy can be improved in order to enhance their performance on the sharpened images. The ability to estimate sharpness locally enables much closer investigation of metrics' behavior when evaluating sharpened images and also more ways to determine a final index from the map.

The authors proposed to obtain the metric score as the mean value of 1% of the highest values in the map. This approach considers only the image regions with the highest sharpness. That is optimal when dealing with blur but in case of sharpening, the final score is then not affected by the undesirable sharpening in other areas. Following modifications of the pooling try to boost the performance by considering also the impact of the sharpening on the regions with low estimated sharpness. In the following figures, the S3 metric will be labelled as S and FISH_{bb} as F. Particular augmentations are distinguished by the lower index in Roman numerals. The following equations will demonstrate the pooling on S3 metric. The same index means the same pooling in the case of FISH_{bb} as well.

By investigating the impact of sharpening on the S3 map it was found out that the dynamic range of the sharpness map increases at first but then it starts to decline again because of the growth in regions with lower sharpness. This is exploited in the first augmentation where the dynamic range of the sharpness map is measured, thus

$$S_{\text{I}} = \max\Big(S3_{\text{map}}(I)\Big) - \min\Big(S3_{\text{map}}(I)\Big), \tag{7.21}$$

where $S3_{\text{map}}(I)$ is the sharpness map for image I obtained by S3 algorithm.

However, the dynamic range can be influenced by outlying values. Therefore another augmentation S_{II} is proposed as

$$S_{\text{II}} = \overline{S3_{\text{map}}(I)}_{1\%\text{ high}} - \overline{S3_{\text{map}}(I)}_{1\%\text{ low}}\ , \tag{7.22}$$

where $\overline{S3_{\text{map}}(I)}_{1\%\text{ high}}$ and $\overline{S3_{\text{map}}(I)}_{1\%\text{ low}}$ are the means of 1% of the highest and the lowest values in the map, respectively.

It was also observed that the minimum value in the map is increased only after a really strong sharpening therefore the scores of these images were lowered by subtracting also the minimum value

$$S_{\text{III}} = S_{\text{II}} - \min\Big(S3_{\text{map}}(I)\Big). \tag{7.23}$$

In the fourth modification, the edge detection was employed. Even though the sharpness map, obtained by S3 metric, has the same size as the image, pixel values in every 4×4 pixels large block are the same. An edge mask E_{Canny} is therefore computed from the outcome of Canny Edge Detector [192] by dividing the edge image into blocks of 4×4 pixels and if the block contains edgel (i.e. a pixel belonging to an edge), all the pixels in the block are set to 1. The S_{IV} index is then obtained as

$$S_{\text{IV}} = \overline{S3_{\text{map}}(I) \times E_{\text{Canny}}} - \overline{S3_{\text{map}}(I) \times (1 - E_{\text{Canny}})}, \tag{7.24}$$

where operator $\times$ stands for pixel-wise multiplication and $\overline{\{.\}}$ is the mean operator. This augmentation comes from the idea that non-edge regions should not be sharpened. Since the size of the sharpness map coming from the FISH_{bb} is smaller than the image, the edge mask is directly the output of the Canny Edge Detector.

It should also be noted that (except for the cases where edge detection is involved) the improved pooling

brings no additional computational requirements.

7.6 Performance of the Metrics with Improved Pooling

The performance of the metrics with improved pooling was tested in the same way as the rest of the metrics in the Section 7.4. The results are depicted for the original S3 and FISH_{bb} algorithms (S and F) and their versions with improved pooling distinguished by the lower index in Roman numerals according to the Section 7.5.

7.6.1 Results per Content

Similarly to Section 7.4, the results per each source content are firstly reported.

Different vs. Similar ROC Analysis

Figure 7.15(a) shows the AUC values with 95% CI and significance of differences for the *Caps* scene. Although S_{I}, S_{II}, S_{III}, F_{I}, and F_{II} reached higher AUC values than original algorithms, no statistically significant change can be observed. The lowest AUC value belongs to F_{III}.

In case of the *Parrots* scene (Figure 7.15(b)), again, no statistically significant gain is detected. However, the FISH_{bb} algorithm with the proposed pooling seems to work significantly worse than original algorithms. On the other hand, all augmentations of the S3 metrics result in higher AUC values, even though without statistical significance.

On the contrary, for the *Red Hat* scene (see Figure 7.15(c)), all of the improvements of FISH_{bb} significantly outperform the original versions. The significant gain can also be observed in case of S_{III}. This is interesting since the *Red Hat* scene has been identified as the most challenging for the metrics.

No statistical gain in performance is detected also in *Isabe* scene. Only the F_{III} performs significantly worse than the metrics with original pooling strategy. The highest AUC value is reached by S_{II}.

Better vs. Worse ROC Analysis

The real added value of the proposed pooling strategies is demonstrated in Better vs. Worse ROC analysis. The results for each source content are visualized in Figure 7.16.

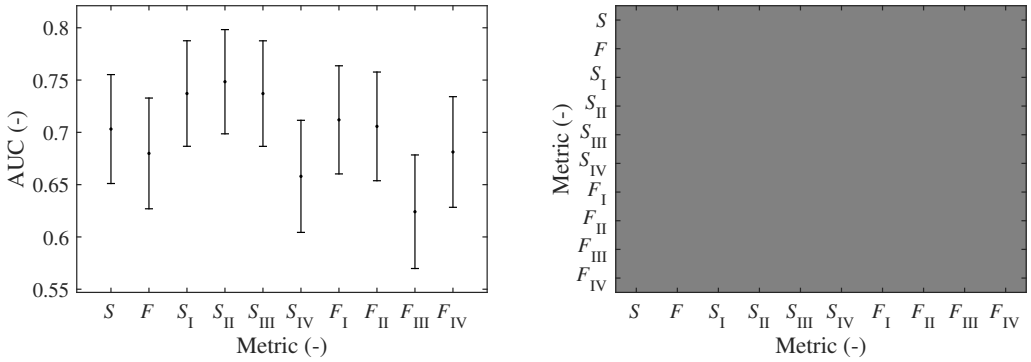
All of the augmentations provide an improved performance for the *Caps* scene (Figure 7.16(a)). Only in case of S_{IV} the difference is not statistically significant compared to S. Moreover, S_{III}, F_{I}, and F_{II} reach the AUC values higher than 0.95 and S_{II} and F_{IV} are just slightly under this value.

The proposed pooling strategies applied on the FISH_{bb} do not work well on the *Parrots* scene (Figure 7.16(b)). All of them perform significantly worse than the original pooling strategy except for F_{IV} where AUC value is higher but without statistical significance. S_{II}, S_{III}, and S_{IV}, on the other hand, provide significant improvement to the performance.

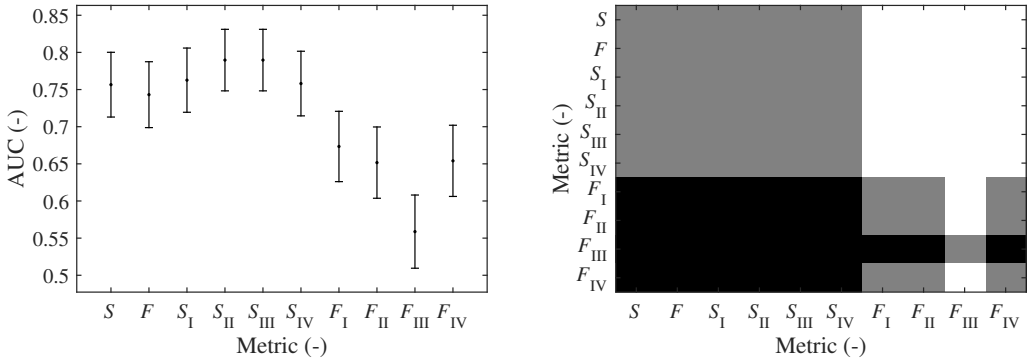
The situation is very different for *Red Hat* source image, as depicted in Figure 7.16(c). Here, all of the augmentations significantly improve the performance. Moreover, F_{I}, and F_{IV} reach the AUC value higher than 0.96. S_{I} and S_{III} perform the best form the pooling strategies applied on S3 metric with AUC values over 0.9.

In case of *Isabe* scene (Figure 7.16(d)), S_{II}, S_{III}, S_{IV}, and F_{I} significantly improve the performance. On the contrary, F_{III} performs poorly compared all of the other proposed and original pooling strategies.

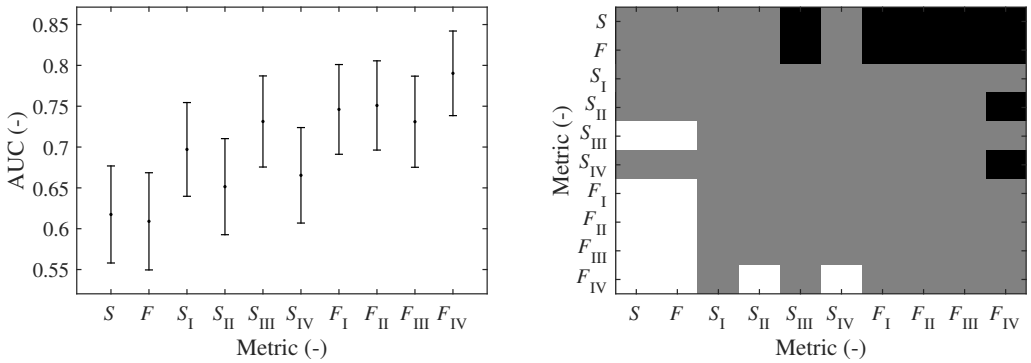
An interesting outcome of the content-wise analyses is the failure of the proposed strategies on the *Parrots* scene for the FISH_{bb} metric while providing very good performance on the challenging *Red Hat* scene. The selection of the most universal approaches will be done after considering all the data together.



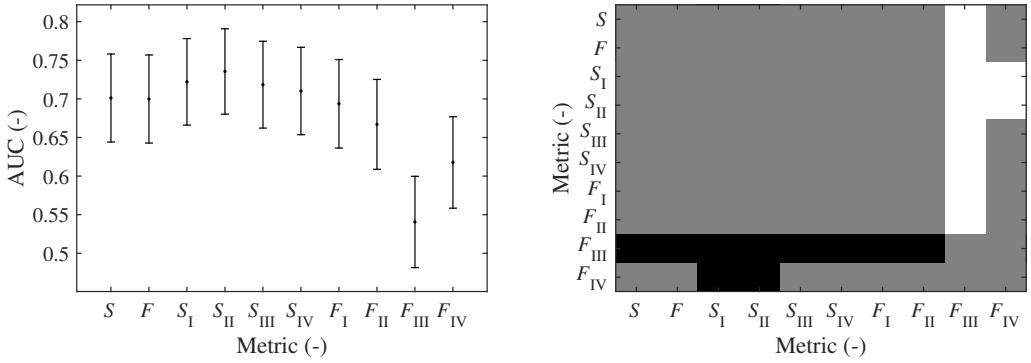
(a) AUC values with 95% CI and significance of differences for the *Caps* scene



(b) AUC values with 95% CI and significance of differences for the *Parrots* scene

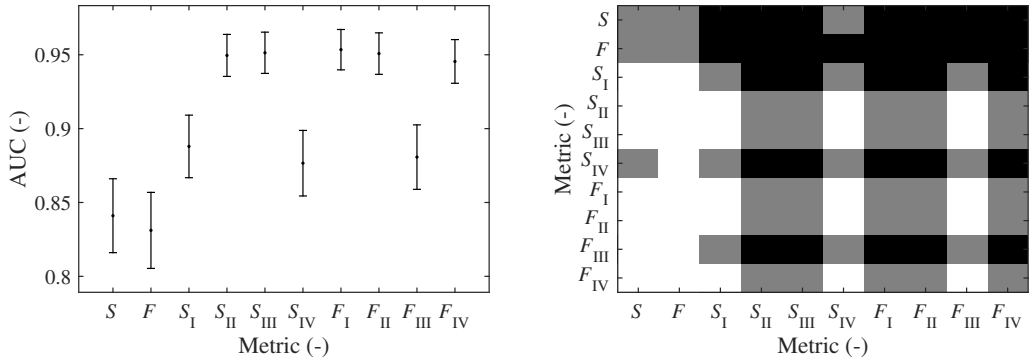


(c) AUC values with 95% CI and significance of differences for the *Red Hat* scene

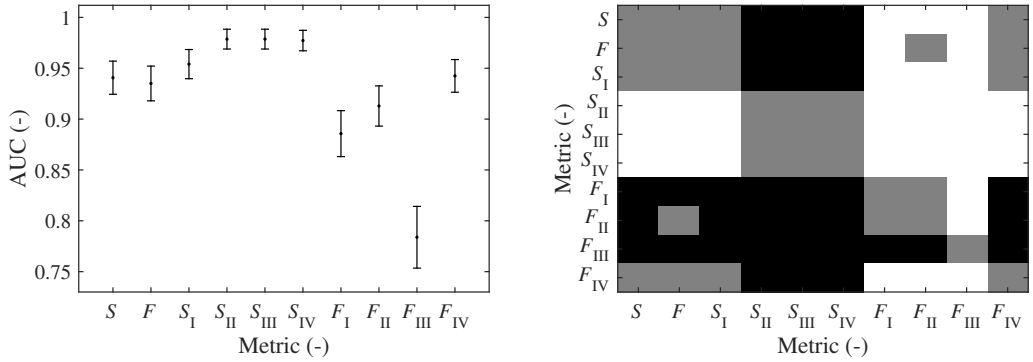


(d) AUC values with 95% CI and significance of differences for the *Isabe* scene

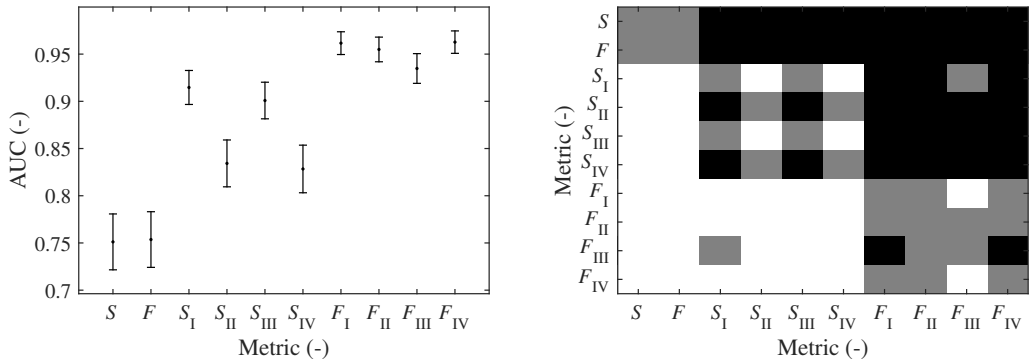
Figure 7.15: The results of the Different vs. Similar ROC Analysis for different pooling strategies for each source content.



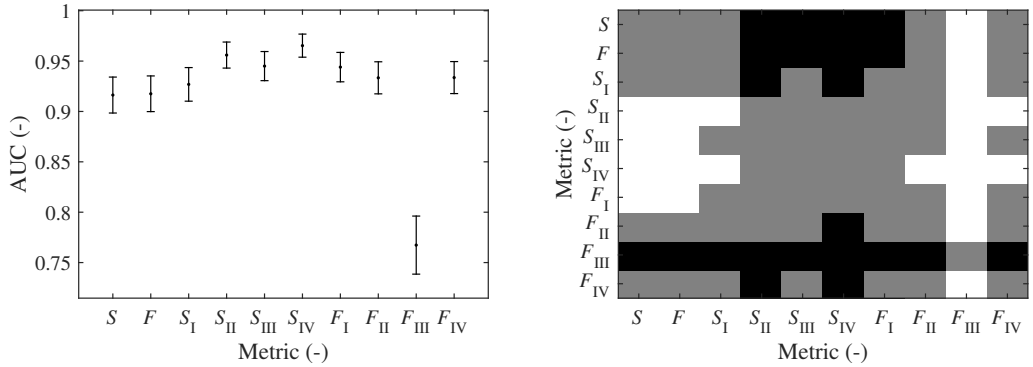
(a) AUC values with 95% CI and significance of differences for the *Caps* scene



(b) AUC values with 95% CI and significance of differences for the *Parrots* scene



(c) AUC values with 95% CI and significance of differences for the *Red Hat* scene



(d) AUC values with 95% CI and significance of differences for the *Isabe* scene

Figure 7.16: The results of the Better vs. Worse ROC Analysis for different pooling strategies for each source content.

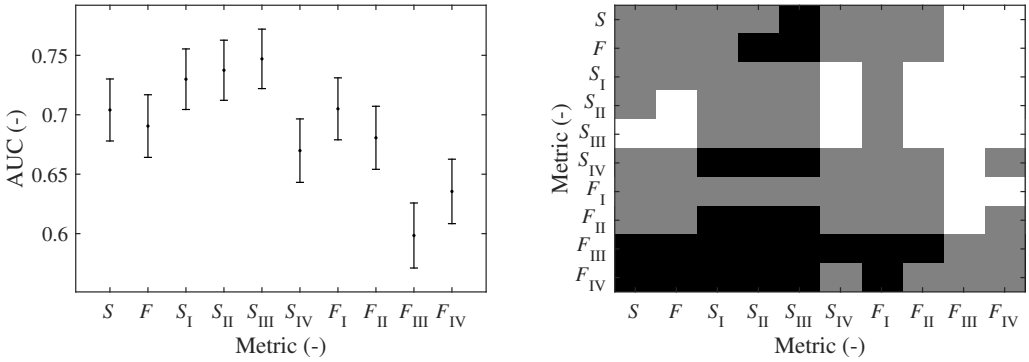


Figure 7.17: The overall results of the Different vs. Similar ROC Analysis for different pooling strategies.

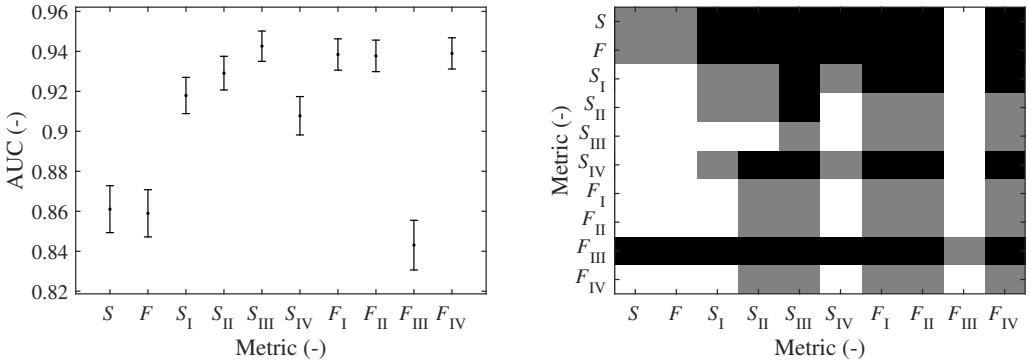


Figure 7.18: The overall results of the Better vs. Worse ROC Analysis for different pooling strategies.

7.6.2 Overall Performance

To test the universality of the metrics, all the contents are put together and the analyses are run on all the data. The remainder of this chapter provides the final results and the discussion about the abilities of the proposed pooling strategies.

Different vs. Similar ROC Analysis

The overall results of the proposed pooling strategies compared to procedures suggested by authors of the particular metrics for the Different vs. Worse analysis can be found in Figure 7.17. It can be seen that in terms of distinguishing between qualitatively close and distant pairs, S_{III} is the best of all the tested metrics (including all the criteria analyzed in Section 7.4). It is the only metric that significantly outperforms the original S3 algorithm with respect to this analysis.

Considering the pooling of the $\mathrm{FISH}_{\mathrm{bb}}$ metric, the only gain can be observed in case of F_{I} but without statistical significance. The strategies F_{III} and F_{IV} perform even statistically worse.

The strategies considering using the edge mask do not seem to work well with respect to this analysis and since they have higher computational requirements, their use does not seem to be of benefit.

Better vs. Worse ROC Analysis

The results obtained in the Better vs. Worse analysis for all the data are shown in Figure 7.18. All of the improved algorithms outperform the original versions with the exception of F_{III} which is significantly worse than the original pooling strategy. On the other hand S_{III} reaches the highest AUC value (0.9426), followed by F_{IV}, F_{I}, and F_{II}.

If only the correct classification in zero (C_0) is considered (Figure 7.19), the same conclusions can be

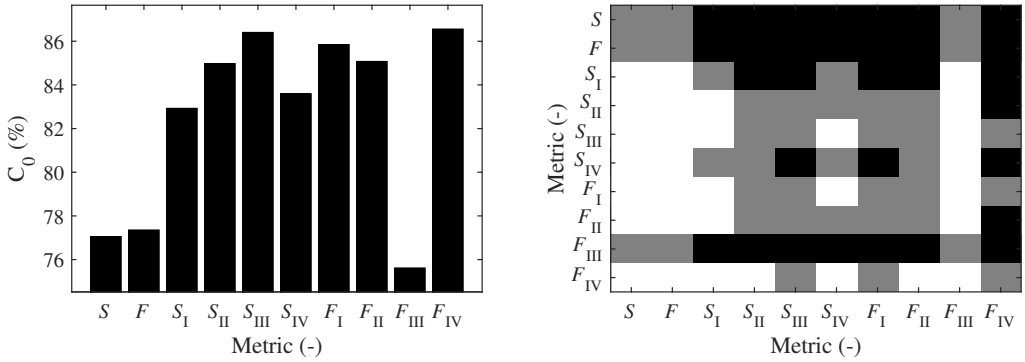


Figure 7.19: The overall correct classification of better and worse image in the pair for different pooling strategies.

made. The only difference is that the highest percentage is achieved by F_{IV} and S_{III} has the second highest value.

Considering all the provided analyses, the absolutely best performing method is S3 with improved pooling S_{III}. This metric has always been among the best performing methods for all the contents separately and has provided the best overall performance. It represents a significant improvement over the original pooling strategy and reaches high AUC values supporting its increased reliability in the considered context.

From the pooling strategies for $\mathrm{FISH}_{\mathrm{bb}}$, the most universal performance was provided by F_{I}. It is much less computationally demanding than S3 metric. However, it seems to struggle with the *Parrots* scene, where original pooling strategy works better. This is probably caused by higher threshold for over-sharpening in this scene. Nevertheless, when overall results are considered, it provides a significant improvement with respect to the original version.

7.7 Using Full Reference Metrics for Sharpened Images Assessment

The previous sections were focused on the general sharpened images quality assessment across different sharpening techniques. However, a plausible application of the objective image quality metrics is also reliable identification of the optimal level of sharpening within one sharpening technique. This section tackles this issue by using full-reference image quality metric to calculate the fidelity between the original, sharpened, and over-sharpened version of the scene. The outputs of these comparisons are then combined into the single value which can be exploited as the base for sharpening parameters' optimization.

The remainder of this chapter describes the proposed method in detail and tests its abilities on the example of popular sharpening technique using Unsharp Mask (see Section 7.2.1).

7.7.1 Method Description

Firstly, the intentionally over-sharpened "anchor image" is created from the original. This image serves as kind of "anti-reference" in the computations. The anchor image I_{anc} is obtained as

$$I_{\mathrm{anc}} = \underset{par_{\mathrm{anc}}}{method}(I), \tag{7.25}$$

where the operator $method$ represents the employed sharpening technique and par_{anc} is a vector of parameters of the given method ensuring the over-sharpening. In case of unsharp mask, the parameters would be $par_{\mathrm{anc}} = [\sigma_{\mathrm{anc}}, \lambda_{\mathrm{anc}}]$.

This image is then used to add another dimension into the calculation. The following quality scores can

I1 I2 I3 I4 I5 I6 I7 I8

Figure 7.20: Source images used in the experiment to obtain ground truth data.

be obtained

$$Q_1 = metric(I, I_{\text{sharp}}), \quad Q_2 = metric(I_{\text{sharp}}, I_{\text{anc}}), \tag{7.26}$$

where $metric$ operator stands for the used full-reference quality metric. The final score Q is then calculated as

$$Q = \tau_1 Q_1^{\tau_2} + (1 - \tau_1) Q_2^{\tau_3}, \tag{7.27}$$

where τ_1, τ_2, and τ_3 are the computational parameters from the interval [0;1]. Parameter τ_1 affects the relative importance of the partial scores, while τ_2 and τ_3 set their sensitivities. Since Q_1 and Q_2 are, for all the tested metrics (see Section 7.7.3), also bounded between 0 and 1, so is the final score Q. Once the parameters are tuned, the quality score Q is used for optimization of the sharpening (the image maximizing the Q value is optimally sharpened).

Considering the proposed approach, three main questions have to be solved:

1. How to create an anchor image?

2. Which metric is the most suitable for the purpose?

3. What are the optimal values of τ_1, τ_2, and τ_3?

To address these questions, it is necessary to have a ground-truth data from the qualitative experiment. A subjective experiment conducted to obtain such data is described in the following section. After that, the solutions to the above mentioned questions will be discussed.

7.7.2 Subjective Study for Obtaining Ground Truth

The quantitative subjective study described in Section 7.3 provided only four levels of sharpening for each sharpening technique. In order to obtain more information about the quality perception, another subjective study with more sharpening levels needed to be conducted. Moreover, since only one sharpening method is considered, more source images could be included in the test.

Stimuli

Eight images (Figure 7.20) from publicly available databases were selected. Six images were from the Kodak[8] database and two were from IRCCyN/IVC Still Image Database [113]. The particular images were selected for the stimuli to contain scenes with different spatial frequencies, colorfulness, textures, and semantics. The images from Kodak database were cropped to the size 512×512 pixels. All of the images have 3 pixels wide mid-gray frame.

Each content was processed with seven levels of applied sharpening (seven different λ parameters). Parameter σ was set to be 0.6. That makes eight versions of every content.

[8] http://r0k.us/graphics/kodak/ (retrieved on 30/08/2016)

Methodology

The PC procedure [12] has been identified as the most suitable in the context of image sharpening (refer to Section 7.3.4). This time, FPC methodology has been selected, since it enables parallel evaluation sessions. The number of the necessary evaluations in the FPC is $\frac{a(a-1)}{2}$, where a is the number of stimuli. For each content, 28 evaluations had to be done. The tests were performed in two separate runs, every observer therefore compared 112 pairs of images.

A simple MATLAB-based application was used for the evaluations. Observers selected the image by clicking on the button below it. First three pairs in the series were meant for training and getting used to the interface. They were not included in the results and a different scene was used. The order of the image pairs was randomized for every observer as well as the displaying of images as right and left.

Test Room

Multimedia Technology Group's[9] post processing laboratory within the facilities of CTU in Prague was used. The lab provides ten separate workplaces equipped with color-calibrated LCD screens with the resolution of 1600×1200 pixels.

Observers

The test was conducted in two separate runs. First four source scenes were evaluated by fifteen observers, the other four by sixteen. Both male and female subjects participated in the experiment. All of them had normal or corrected to normal vision. Before the testing itself, short introductory presentation was shown to explain the run and the purpose of the test.

Results

For the interpretation of the obtained results, BTL model (Section 2.3.3) was again used. Particular scores with 95% confidence intervals and the statistical significance of differences are depicted in the Figure 7.21. If the square between two images is gray, there is no statistically significant difference between their evaluations. A white square means that the quality of the image in the row is significantly better with respect to the image in the column and a black square marks the opposite case.

Image I5 seems to allow much more sharpening than other source contents. This is probably caused by lower quality of the original version and the increased presence of high contrast areas. The least "sharpening friendly" content is I3, where the details of the skin allow only mild sharpening.

The statistical insignificance of some images' score differences is most likely caused by the variability in the subjects' opinions and will probably not change with increasing number of observations. The results of this subjective study were used as the ground-truth for the tuning of the particular properties of the proposed approach.

7.7.3 Method Parameters Selection

In Section 7.7.1, three main problems were formulated – selection of the anchor image, objective metric, and computational parameters. The anchor image is supposed to be clearly over-sharpened (i.e. the optimal amount of applied sharpening should be smaller than the amount used to sharpen this image). The simplest way to obtain it is to use λ_{anc} (amount of sharpening) that high, that the final image will be over-sharpened, regardless the content. Firstly, the impact of different values of λ_{anc} on the metrics combinations' performance was studied. Four full-reference quality metrics were considered – SSIM [45], MS-SSIM [46], IW-SSIM [47], and VIF [48]. After the objective evaluation, the best parameters from the equation (7.27)

[9] http://mmtg.fel.cvut.cz (retrieved on 30/08/2016)

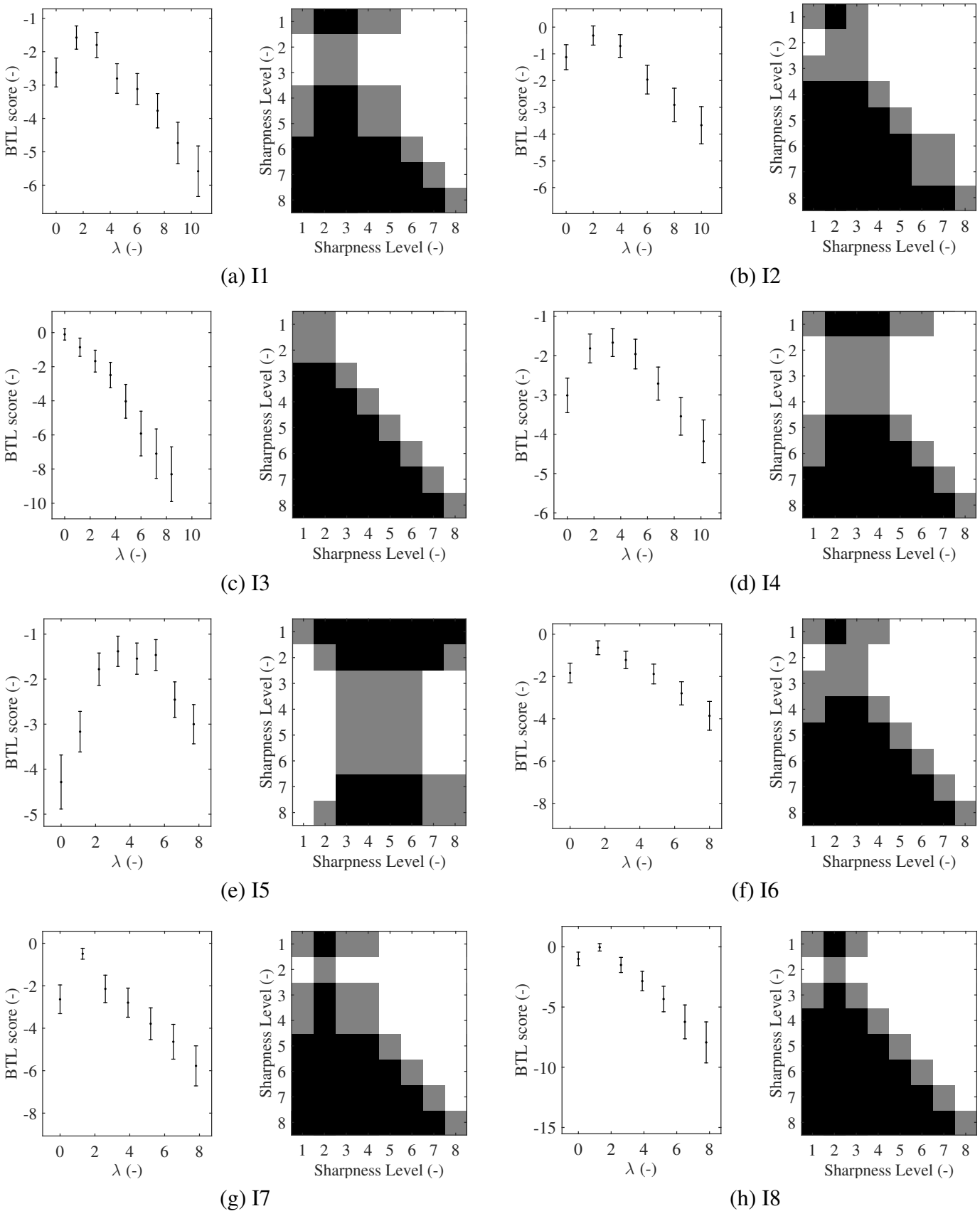
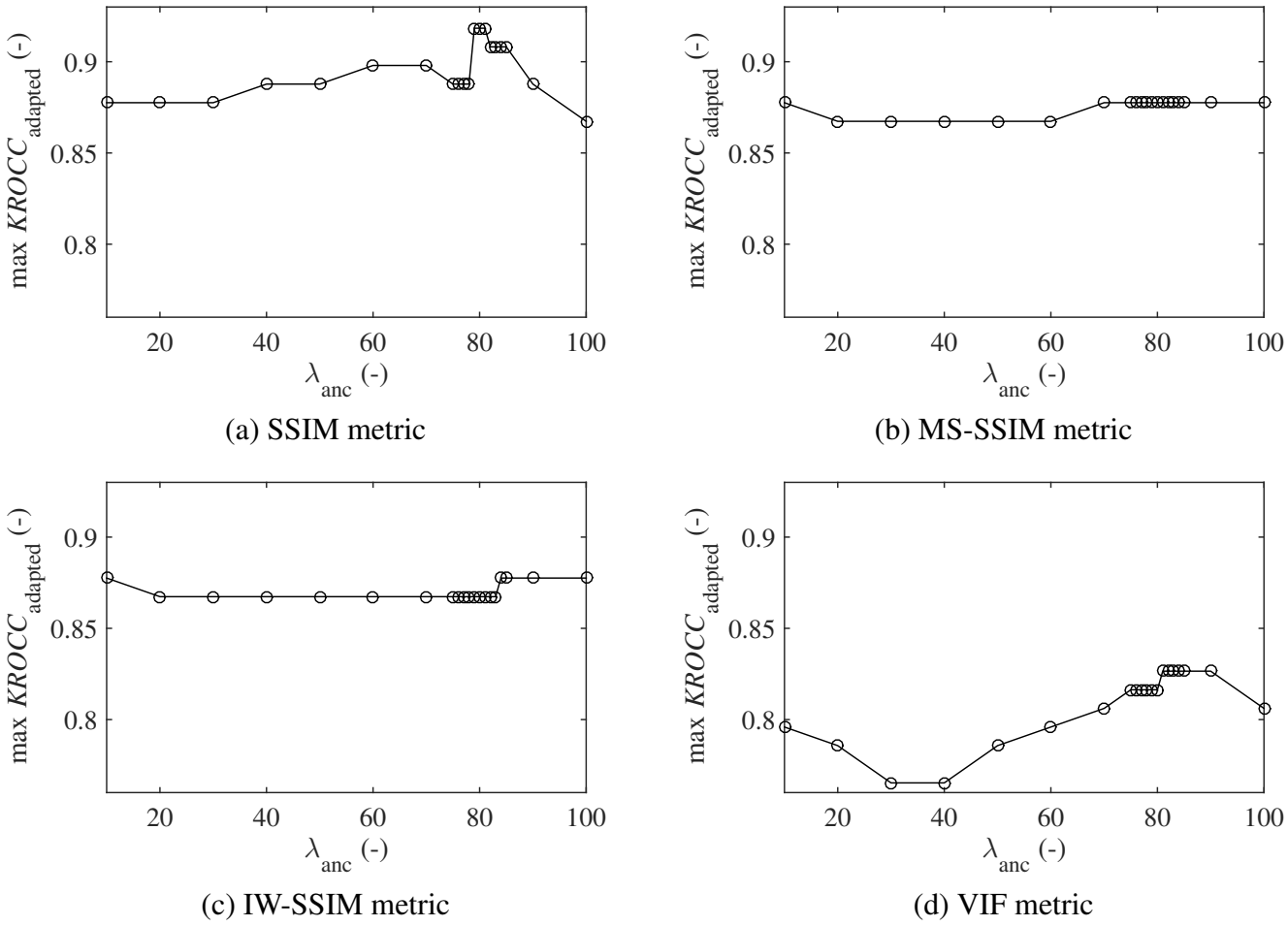


Figure 7.21: Results of the subjective experiment for obtaining ground truth data.



(a) SSIM metric

(b) MS-SSIM metric

(c) IW-SSIM metric

(d) VIF metric

Figure 7.22: Performance of metrics when different λ_{anc} is used.

were found using the optimization of adapted KROCC between the objective and subjective scores. It's calculation is described in Section 5.1.1.

The $KROCC_{\text{adapted}}$ for the parameters' setup maximizing the correlation for different λ_{anc} for particular metrics are plotted in Figure 7.22.

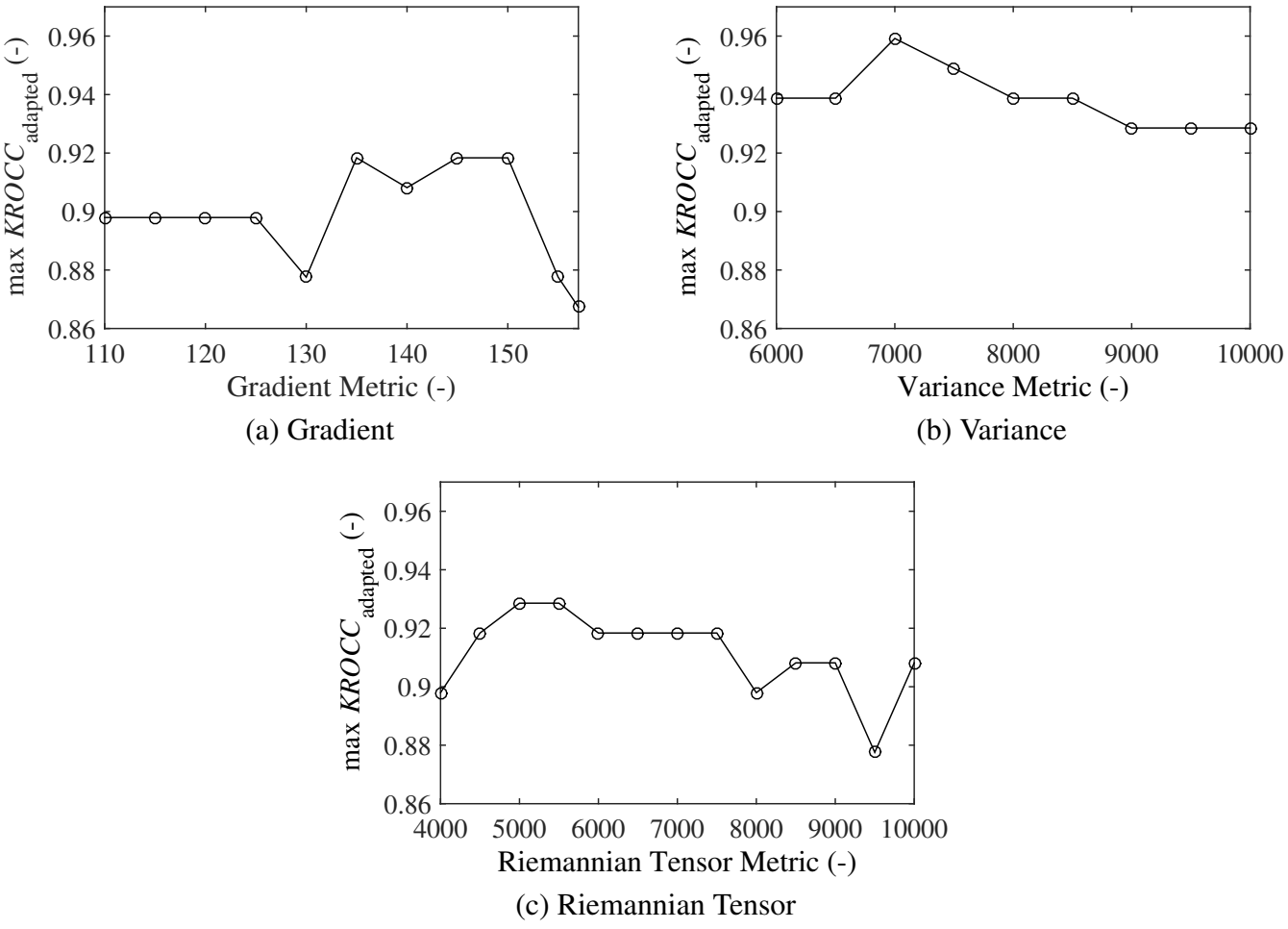
The highest coefficient values were reached for the SSIM metric. The other tested metrics had bigger problems, especially with evaluation of content I5. The local maximum of the KROCC is obtained with λ_{anc} around 80. Different computation of the anchor image does not have that significant effect on the performance of MS-SSIM, and IW-SSIM (difference in KROCC values is around 0.01). SSIM and VIF seem to be more influenced by these changes.

However, this way of creating the anchor image is not ideal because it is very much dependent on the original image quality. For example, if the image would already be sharpened, the result would be different. It is therefore more convenient to define the anchor image using some objective measure.

This measure should not need any reference to make the process of the anchor image creation as independent as possible. It also has to behave monotonically when different levels of sharpening are applied. Three no reference sharpness measures, fulfilling the requirement, were tested – Gradient [59], Variance [57], and Riemannian Tensor [65]. The implementations from the framework developed by Murthy and Karam [44] were used.

The image is sharpened, until the metric value reaches the given level. The same optimization process, maximizing the adapted KROCC values for significantly different image pairs, was applied. The results for the SSIM metric are depicted in Figure 7.23.

The best performance was obtained for the Variance metric. The final selected parameters are therefore



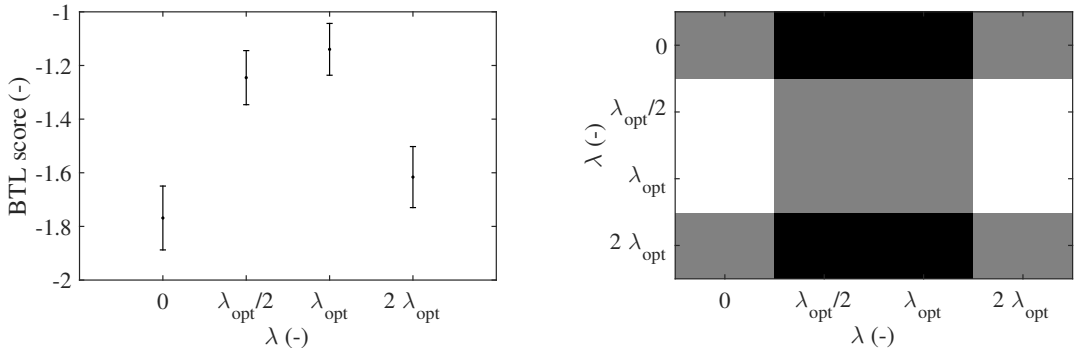
(a) Gradient

(b) Variance

(c) Riemannian Tensor

Figure 7.23: Performance of SSIM metric with the anchor image created by no reference measures.

Figure 7.24: Source images used to evaluate the performance of the proposed method.



(a) BTL score with 95% CI (b) Significance of differences

Figure 7.25: Results of the performance testing for 11 images.

– the anchor image has Variance metric value of 7000, SSIM metric is employed, and the final parameters τ_1, τ_2, and τ_3 from equation (7.27) are 0.5046, 1.0093, and 1.0444, respectively.

7.7.4 Performance Verification

To test the performance of the proposed approach, 11 images (different from the ones used for tuning of the method) were picked from the Kodak database, cropped to 512×512 pixels size and provided with the 3 pixels wide frame. Their resized versions can be seen in Figure 7.24.

The image with optimal level of sharpening, with respect to the proposed method, was obtained by maximization of the Q value from equation (7.27). Then, another two images were created using half of the optimal λ_{opt} parameter value and twice the λ_{opt}, respectively. The series for every content therefore contained 4 images, including the original. These 44 images were evaluated using the same procedure as for obtaining ground truth data.

Twenty-four observers participated in the experiment. The results for all 11 images together are depicted in Figure 7.25. The optimization based on the proposed method proves to increase the quality of images, since its score is statistically significantly higher than the score for the original image. Moreover, it also outperforms the case, when the double of the found optimal λ_{opt} parameter value was used. Note that the quality of original and version sharpened with double λ_{opt} have similar quality. Therefore, it can be assumed that the amount of applied sharpening is near the optimum. The difference between the scores for optimal λ and its half is not significant, which is probably caused by the fact that the images with these parameters applied were visually close. Nevertheless, the reached BTL score is higher for the optimal case.

The results suggest that the defined quality assessment scheme can successfully evaluate sharpened images and is suitable for the automatic image sharpening.

8

Quality Assessment of Tone-Mapped High Dynamic Range Images

Digital photography has brought the possibility to capture the real world scenes to virtually anyone. Despite the advantages which went hand in hand with the transfer from analog to digital world, such as storing, enhancement and adjustment options, and many more, there are still some areas where the state-of-the-art imaging systems cannot substitute the analog ones. One of these areas, which is now being extensively discussed, is the dynamic range (DR).

DR is defined as the ratio between the highest and the lowest luminance value in the scene. Typical DR of a real-world scene is around 10,000:1 in orders of magnitude (or higher in the direct presence of the illumination source) [195]. Most of the imaging systems produce something about 100:1 ratio, since the range of luminance values for each pixel is limited to 8-bits. This provides only 256 intensity levels which is not sufficient to reproduce all the details and contrast.

To deal with this limitations, extended dynamic range (EDR) and high dynamic range (HDR) imaging systems have been introduced. Their goal is to capture, store, and reproduce the true luminance values that occur in the real scene. The terminology for the current systems is then standard dynamic range (SDR) or low dynamic range (LDR).

In the field of scene capturing, the problem has practically been solved. Some of the cameras available on the market enable to capture and store images with higher DR. For example SpheroCam HDR by SpheronVR,[1] Ladybug spherical camera by Point Grey Research,[2] JAI AD-132GE,[3] Black Magic Pocket Cinema Camera,[4] or Red 101 [5] to name a few. Recently, Daimler company introduced the HDR capturing with a specialized chip to the automobile industry [196].

Apart from these mostly more expensive options, any camera with possibility to change the exposure time can capture HDR image. The limited dynamic range of the camera will result in presence of underexposed and overexposed regions in the photograph. If the DR of the captured scene is very high (such as the sunset, the edge between highly and poorly illuminated areas, etc.), it is impossible to fit all the details in shadows and highlights into one picture with single exposure time. The technique called *exposure bracketing* can be applied.

[1] http://www.spheron.com/ (retrieved on 30/08/2016)
[2] http://www.ptgrey.com/ (retrieved on 30/08/2016)
[3] http://www.jai.com/en/products/ad-132ge (retrieved on 30/08/2016)
[4] https://www.blackmagicdesign.com/products/blackmagicpocketcinemacamera (retrieved on 30/08/2016)
[5] http://www.red.com/learn/red-101/hdrx-high-dynamic-range-video (retrieved on 30/08/2016)

Here, the same scene is shot several times (typically sufficient number is five) with different exposure setting. It is advisable to vary the exposure time only and fix all the other factors (aperture size, focal length, ISO, etc.). Also the use of a tripod is virtually essential, since the particular images should be as matched as possible. Some small differences can be compensated as part of post-processing by following image registration, ghost removal, or lens flare removal (which could be applicable even when the HDR camera is used). These techniques are also described in [195].

Having the particular images of the same scene with different exposures available, the radiance map of the scene can be obtained and an HDR image can be created. The techniques for this will be further discussed in Section 8.1.

Once the HDR image is created, another question rises – how to present it. This issue is in detail described in Section 8.2. Here, the possibilities for direct HDR displaying and introduction of the tone-mapping operators (TMOs) used for displaying HDR content on LDR screens are provided.

The process of tone-mapping, i.e. the compression of DR to fit the regular LDR display, is the main area of interest of this chapter. The challenges introduced to the quality assessment by TMOs have already been identified in Section 1.2.2. Firstly, an optimization of TMOs' parameters (Section 8.3) in security and multimedia applications using objective quality criteria will be discussed. Further, an extensive subjective quality experiment in order to prepare a challenging dataset for testing the objective metrics' performance in the context of tone-mapping will be described in Section 8.4. In Section 8.5, applicable objective metrics will be tested and compared. Section 8.6 will then describe the selection of features relevant in the given context and their combination into a reliable quality metric.

8.1 High Dynamic Range Image Creation

Creating one HDR image from the multiple LDR images is discussed in this section. The aim is to recover the radiance map, i.e. the absolute amount of light falling onto each point of the camera sensor. Several approaches have been introduced during the years. Most of them are summarized in [197].

The idea is to estimate the irradiance as a weighted average of values $\hat{x}_i$

$$\hat{x}_i = \frac{f^{-1}(v_i)}{t_i}, \tag{8.1}$$

where f^{-1} is the camera's inverse transfer function, v_i represents the digital output obtained from the chip, and t_i stands for the i-th exposure time. The weighted average is then defined as

$$\hat{\mu} = \frac{\sum_i w(v_i)\hat{x}_i}{\sum_i w(v_i)}, \tag{8.2}$$

where w is a weighting function assigning the importance to the output values. This function varies among the approaches which will be described henceforth.

The first way how to address this issue has been proposed in 1995 by Mann and Pickard [198]. The weighting is introduced to compensate for the quantization error. The weights are obtained as

$$w = \frac{1}{\frac{d}{dv}(\log f^{-1}(v))}. \tag{8.3}$$

The logarithm makes the quantization error perceptually uniform.

In 1997, Debevec and Malik [199] used weighting with a hat function assigning more importance to the values far from the saturation regions (highlights and shadows), since they are more likely to carry some information. The function is simply defined as

$$w = \min(v - v_{\text{min}}, v - v_{\text{max}}). \tag{8.4}$$

Mitsunaga and Nayar [200] proposed a weighting based on signal to noise ratio (SNR) in their paper from 1999. The assumption here is that the noise is independent of the measured pixel value. The weight function is calculated as

$$w = \frac{f^{-1}(v)}{\frac{d}{dv} f^{-1}(v)}. \tag{8.5}$$

Tsin et al. [201] introduced a camera noise model into the weighting process in 2001. They are calculating with output standard deviation which is estimated from the images. The weight is then obtained as

$$w = \frac{t}{\hat{\sigma}_{f^{-1}(v)}}. \tag{8.6}$$

In 2003, Robertson et al. [202] extended the concept developed by Mann and Pickard [198] by including also the exposure time in the calculation

$$w = \frac{t^2}{\frac{d}{dv}(\log f^{-1}(v))}, \tag{8.7}$$

the weight grows quadratically with t.

Reinhard et al. [195], in the book from 2005, combined the approaches of Mitsunaga and Nayar [200] and Debevec and Malik [199] but using different hat function

$$w = \frac{f^{-1}(v)}{\frac{d}{dv} f^{-1}(v)} \left[1 - \left(\frac{v}{v_{\text{mid}} - 1}\right)^{12}\right]. \tag{8.8}$$

Kirk and Andersen [203] proposed another improvement in 2006. They employ camera noise model for estimation of variance. Again, the estimates are obtained directly from the output signal, thus propagating the estimation error into the weighting function. The weights are obtained as

$$w = \frac{t^2}{\left(\frac{d}{dv} f^{-1}(v)\right)^2 \sigma_v^2}. \tag{8.9}$$

The most complex approach so far has been introduced by Granados et al. [197] in 2010. They use much more complicated camera noise model, taking into account different sources and considering also spatial noise. Note that the model assumes the use of the charged-coupled device (CCD) sensor, as described by Janesick [204].

The first step is the camera calibration. They start with estimating the parameters $\hat{\mu}_{\text{r}}$ and $\hat{\sigma}_{\text{r}}$ of the *read out noise* by analyzing the distribution of the *bias frame* (image obtained with zero integration time), followed by the measurement of the saturation value v_{sat} from the saturation frame (image obtained with all the pixels at the maximum value).

Next, n flat fields (images of the spatially uniform background) are taken and the gain per pixel a_j and camera gain $\hat{g}$ are estimated.

After the LDR images are obtained, also the dark frames (images obtained with closed diaphragm) at each exposures are taken. The pixel values are estimated as

$$\hat{\mu} = \frac{\sum_i \frac{1}{\hat{\sigma}_i^2} \hat{x}_i}{\sum_i \frac{1}{\hat{\sigma}_i^2}}, \tag{8.10}$$

from where it can be seen by comparing to (8.2) that

$$w(v_i) = \frac{1}{\hat{\sigma}_i^2}, \tag{8.11}$$

where $\hat{\sigma}$ is the standard deviation of the pixel. In this step, the assumption is that it is constant for all exposures.

In the next step, it is estimated from the model and the estimated pixel value $\hat{\mu}$ as

$$w(v_i) = \frac{1}{\hat{\sigma}_i^2} = \frac{t_i^2 g^2 a_j^2}{g^2 t_i (a_j \hat{\mu} + 2\hat{\mu}_\mathrm{d}) + 2\sigma_\mathrm{r}^2}, \tag{8.12}$$

where $\hat{\mu_\mathrm{d}}$ is the mean of the pixel in the dark frame. The steps from equations (8.10) and (8.12) are iteratively repeated until convergence.

After that, bilateral filtering [177] is employed to compensate for $\hat{\sigma}$. Another details about the procedure can be found in the respective paper [197].

All of the above described procedures are employed to obtain as realistic image of the scene as possible. The final pixel values are floating point and can be stored in number of possible formats such as .hdr, .exr, .ppm, etc. The most common approaches towards HDR image encoding are introduced e.g. in [195, 205, 206]. Recently ISO/IEC JTC1/SC29/WG1 (JPEG) committee is preparing a backward compatible compression standard for HDR images called JPEG XT.

The compression uses tone-mapping to create an LDR version of the image which is encoded with legacy JPEG encoder. The residual data are encoded in a separate branch. The encoder enables adjusting of quality parameters for LDR image as well as for residual information. The impact of different settings on the resulting quality for different JPEG XT profiles is studied e.g. in [207].

8.2 Displaying High Dynamic Range Images

The presentation of the HDR content is even more challenging topic than its capture. The requirements for the technology are still very high. On the other hand, the possibility to significantly improve the quality of user's experience while watching the content makes it a hot topic for the future of the displays. The next section introduces the fundamentals of possible technologies applicable for displaying HDR images.

Since these technologies are mostly in the phase of prototypes or their price makes them unavailable to the regular customers, another possibilities are introduced. More specifically, the operators capable of adjusting the HDR content for displaying on regular LDR devices. The importance of these dynamic range compressing algorithms will probably remain even after HDR displaying solutions hit the consumer market, since they will be necessary for backward compatibility, printing industry, and other otherwise too demanding processing applications.

8.2.1 High Dynamic Range Displays

Displays capable of displaying HDR content are relatively new phenomenon compared to the whole field. Although the first attempt for HDR visualization was introduced already in 2002 by Greg Ward [208], the resulting HDR still image viewer proved to be too expensive and impractical for the real applications. On the other hand, as shown by Ledda et al. [209], it was capable of displaying images closer to the reality than regular screen (of course after tone-mapping, see Section 8.2.2). The following sections discuss possible technologies for HDR displays.

HDR Still Image Viewer

The viewer was built of three fundamental parts – a 12 V, 50 W lamp (maximum luminance cca 5,000 cd/m^2), a Large-Expanse Extra-Perspective lens (120° field of view), and two film transparencies. All of these components have to be doubled – for each eye. The diagram could be found in Figure 8.1.

The lenses were developed by Eric Howlett for NASA virtual reality experiments [210]. The problem with these lenses is that they exhibit very disturbing chromatic aberration. This could be compensated by

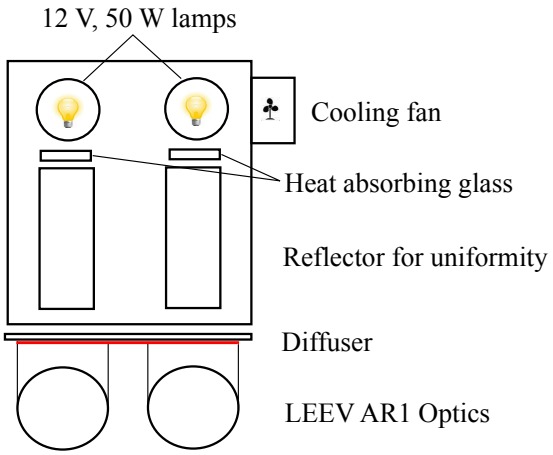


Figure 8.1: HDR viewer diagram. Redrawn from [195].

the lens with the opposite aberration or, if the images are prepared on computer, including the aberration correcting function in the process of content creation. This function scales the red channel 1.5% more than the blue one. Green channel is scaled half way in the middle. The HDR image was created by multiplication of the two film transparencies (i.e. the transparencies were put on top of each other). The position of the transparencies in the viewer is marked by red line in the diagram (Figure 8.1). The HDR images for visualization were required to be of resolution at least $2,048 \times 2,048$ pixels and had to be printed as hemispherical fisheye projection.

The background-image I_{back}, used for modulation, was created by taking the square root of the HDR image's luminance component, followed by Gaussian blurring (32×32 pixels window)

$$I_{\text{back}} = \sqrt{\mathcal{L}(I_{\text{orig}})} * h_{32\times32}, \tag{8.13}$$

where $\mathcal{L}(I_{\text{orig}})$ is the luminance component of the original image, $h_{32\times32}$ is the Gaussian kernel, and operator $*$ stands for two-dimensional convolution.

The front image, representing the color and details, was obtained as

$$I_{\text{front}} = Ch\left(\frac{I_{\text{orig}}}{I_{\text{back}}}\right), \tag{8.14}$$

where function Ch is the chromatic aberration compensation function.

The device was capable of displaying contrast up to 10,000:1 with the peak luminance of 5,000 cd/m^2 and the minimum displayable luminance of 0.5 cd/m^2. Unfortunately, as already stated above, it enabled only the visualization of still images. Moreover, four film transparencies had to be created to display one scene. The cost was, therefore, too high for virtually any application.

However, the main impact of this pioneer is the proof that it is possible to show content closer to the reality and introduction of the concept which has been taken over for other, more practical devices.

Projection-based Displays

First of the applicable principles was developed by Seetzen et al. [211] from a Canadian company Sunnybrook Technologies, later renamed to BrightSide Technologies Inc. In 2007, Dolby Laboratories bought the company and renamed it to Dolby Vision[6]. The technology is capable of displaying content dynamically (i.e. HDR video can also be visualized) and, unlike the HDR viewer, it is not limited for single observer. The diagram of the function principle can be found in Figure 8.2.

[6] http://www.dolby.com/us/en/technologies/dolby-vision.html (retrieved on 30/08/2016)

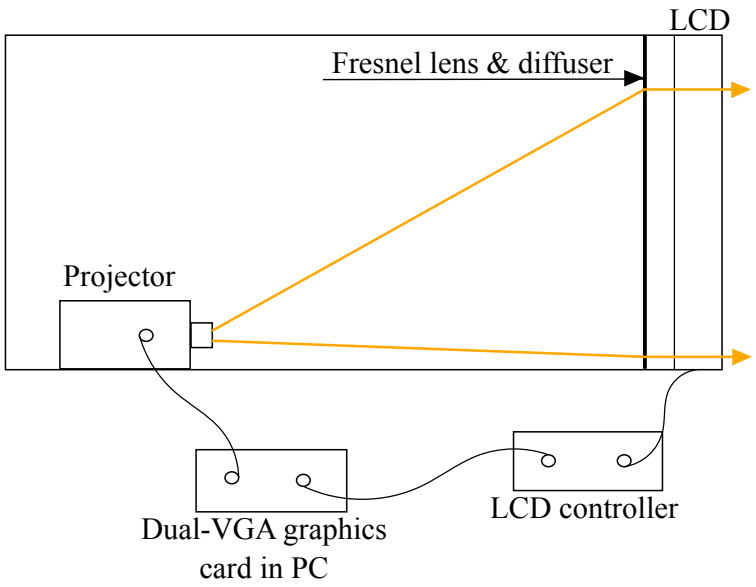


Figure 8.2: Projector-based HDR display's diagram. Redrawn from [211].

The backlight of the LCD panel is provided by Digital Light Processing (DLP) projector, using the MEMS Digital Micromirror Device (DMD) technology patented by Texas Instruments [7]. The color wheel is removed from the projector because only the luminance component is needed.

The images (the back image from the projector and the front image on the LCD panel) are obtained similarly to the images for HDR viewer (see the previous Section). Only the chromatic aberration compensation function is not used, the square-root luminance is modelled on the point spread function (PSF) of the projector, and response functions of the projector and LCD panel are involved in the processing, thus the resulting signal is linear. Some improvement to the splitting algorithm has been proposed by Luka and Ferwerda [212], increasing the gamut and saturation of the dark colors by transition from the square-root to linear function in the darker areas.

Another splitting algorithm for this setup was developed by Zhang et al. [213], using the color appearance model iCAM06 proposed by Kuang et al. [214]. The device is capable of displaying luminance up to 2,700 cd/m^2 and the minimum luminance level is supposed to be 0.054 cd/m^2 giving the dynamic range of 50,000:1.

The drawbacks of this approach are the small viewing angle, caused by the use of Fresnel lens, big power consumption and heat generation in order to produce very bright backlight, and most importantly the size of the device. Since it uses the projector, the space between the projector and the panel itself has to be around one meter long for 15 inch display. This makes the screen non-practical for consumer use.

LED-Based displays

To solve most of the drawbacks of the projector-based systems, the same company presented the LED-based HDR display. The projector is substituted by the array of white light emitting diodes (LEDs). They are arranged in the hexagonal shape, as shown in Figure 8.3.

These LEDs do not require the use of Fresnel lens to compensate for a beam spread, so the array can be positioned right behind the LCD panel's diffuser, enabling much flatter solutions than the previous approach. Also the consumption is no longer constant but dependent on visualized content which, considering the fact that most of the scenes will have less than 10% of the area covered by brightest regions, makes it comparable to the CRT displays.

On the other hand, the creation of the back image is more challenging due to the overlapping of PSFs from particular LEDs. It is resolved by down-sampling the image to the resolution of the LEDs, followed

[7] http://www.ti.com/ (retrieved on 30/08/2016)

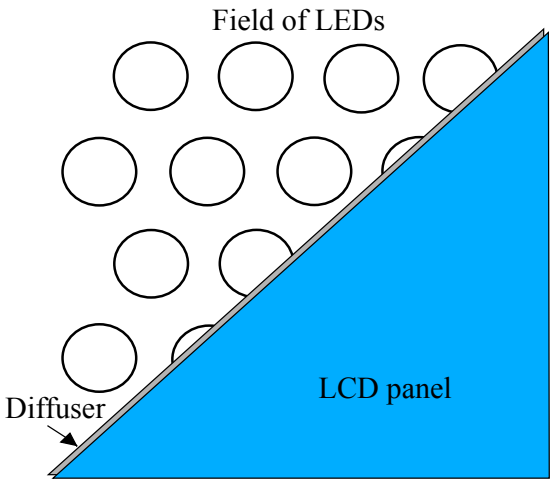


Figure 8.3: The arrangement of LEDs in the LED-based HDR display. Redrawn from [206].

by the compensation for the PSFs overlaps. Since it is a deconvolution problem, the solution is mostly unstable. In practice, the solution is approximated by Gauss-Seidel iteration. This is then reflected in the computation of back image. The implementation details can be found in [211].

The biggest drawback of the technology is the price of the high-end which is fortunately gradually decreasing. Another issue (common to all of the HDR display technologies) is the increased heat generation and therefore necessity for effective cooling. Nevertheless, this technology is probably the most promising so far and some solutions have already been introduced to the market.

Probably the first available screen using the LED-based technology was BrightSide (now Dolby) DR37-P[8]. It has a 37 LCD screen with contrast ratio 250:1 and full HD resolution (1920 $\times$ 1080). The dynamic range of the display is 200,000:1. The claimed maximum luminance value is 4,000 cd/m^2 and minimum 0.015 cd/m^2. The backlight is produced by 1,380 LEDs.

Probably the most advanced solution for HDR displaying, in terms of readiness for consumers, was introduced in 2009 by Italian company SIM2.[9] It is also based on LED-based technology and since that year, they introduced several improvements. The model that will be used in the following studies is called SIM2 HDR47E S 4K and has a 47 inch LCD screen with full HD resolution. It displays images in 16 bits per color channel. The backlight is produced by 2,202 individually controlled LEDs. The maximum luminance of the display is 4,000 cd/m^2 (the first SIM2 display could reach "only" to 2,000 cd/m^2). The contrast ratio is stated to be virtually infinite (since the backlight can be turned off but that is true for almost all of the HDR displays). The ANSI contrast (the contrast measured by the 16 $\times$ 16 checkerboard pattern simultaneously) is about 20,000:1. The lifetime of the display is estimated for cca 50,000 hours. The most advanced model of the SIM2 is HDR47E S 6MB[10] with the peak luminance of 6,000 cd/m^2.

Medical Displays Based on Dual Layer LCD

In the field of medical imaging, the problem of displaying HDR content has also been addressed. Unlike in multimedia applications, only the luminance information is mostly sufficient (i.e. the displays can be monochromatic). Here, the dual layer LCD displays (DL-LCD), firstly introduced by Visser et al. [215], is described [216].

The idea is to put second LCD panel on top of the first one. Thus the light is modulated twice, resulting in much lower leakage (the residual light coming through a completely dark pixel) and more possible steps between minimum and maximum luminance level. The backlight can remain uniform as in the case of

[8] http://www.bit-tech.net/hardware/2005/10/03/brightside_hdr_edr/1 (retrieved on 30/08/2016)
[9] http://www.sim2.com (retrieved on 30/08/2016)
[10] http://hdr.sim2.it/hdrproducts/hdr47es6mb (retrieved on 30/08/2016)

single layer LCD displays. The first generation of DL-LCDs used cold-cathode fluorescent lamps (CCFLs), the second employed array of LEDs.

The biggest drawback of this technology was the image alignment. Even if the two panels are perfectly aligned (which is mostly not the case), watching the display from not precisely on-axis position would create a parallax-error. To resolve this issue, the back image should be blurred. Actually, the similar image splitting procedure as for the previous technologies can be applied. Unfortunately, since the division is applied, the resulting front image can contain luminance higher that the maximum or lower than the minimum displayable level. This would lead to the clipping, which is unacceptable in the medical imaging field.

This lead to design of novel algorithm based on objective function optimization. Guarnieri et al. [217] objectively measure the smoothness of the back image with constraints of perfect reconstruction and clipping avoiding. The implementation is based on Multigrid methods, which do part of the calculations on image with lower resolution, bringing the important computational savings and making it applicable in the area. This method also has one disadvantage. If the original image contains step edge with luminance range larger than the range of the front panel, it cannot be visualized with smooth back image.

The same authors came up with a solution employing simplified model of human visual system (HVS) [218]. In particular, they exploit two of its mechanisms – adaptation level and veiling glare. These phenomenons influence the perception it the proximity of sharp edges with very high dynamic range. The bright light of the edge causes saturation of the electrical signal in the cells and thus makes the dark portion of the edge appear brighter. The same result is also the consequence of the light scattering inside the eye. Both of the mechanisms are described in [216]. The algorithm uses this fact to relax the constraint of perfect reconstruction up to the level, which will not be perceivable by human eye. The implementation details and performance evaluation of the algorithm are described in [219].

Last, but also very important part of the visualization process is mapping the input values onto the display luminance values. In medical imaging, this is usually done by DICOM Grayscale Standard Display Function (GSDF) [220]. However, in this application it has to be adjusted to the HDR. The adjustment considering also the ambient illumination in the diagnostic room is described in [221].

Other Technological Possibilities for HDR displays

The goal of this section is to briefly discuss some future possibilities for native HDR visualization. In projector-based HDR displays, a DMD projector is used without color wheel just to provide a backlight for the image. Theoretically speaking, the technology itself could be able to directly visualize the native HDR content, although some limitations still exist. The most important is the light scattering that happens on the edges of micromirrors, hinges, and spacings. On the other hand, novel DMD chips use some inner dark coating to minimize the reflections and thus increasing the displayable DR. Nevertheless, it is possible that the technology will hit its boundaries before reaching the capabilities necessary for HDR images.

Quite similar statement is valid also for the plasma displays. The brightness is theoretically limited only by the input power and material endurance and dark areas are not a problem, since the pixels can be completely switched off. However, the power consumption makes the plasma displays loose their popularity even in regular applications and the use of this technology for HDR displays is therefore highly improbable.

Probably the biggest potential lies in organic LED (OLED) technology. The OLED displays (mostly in active mask OLED – AMOLED setups) are already being installed in some smaller devices (such as mobile phones) and their advantages in terms of DR are obvious. The pixels can be completely switched off, cross-talk among pixels barely exists, and maximum displayable luminance is very high (dependent mostly on input power). However, the technology still has some problems to be solved for introduction of large-scale monitors capable of HDR native visualization.

8.2.2 Displaying High Dynamic Range Content on Standard Displays

In this section, a basic introduction to operators capable of compressing dynamic range to the range necessary for display on regular LDR screens is provided. Even if the cameras enable capturing of the whole dynamic range of the scene, there will always be a limitation by the presentation media. In particular, low end applications and, more importantly, printing industry will always require some form of content manipulation to be able to represent the scene as close to the original as possible, using only the dynamics available. The importance of these operators, commonly known as tone-reproduction or tone-mapping operators (TMOs), will maintain high. Note that in cases where the DR of the real scene is higher than the DR of a particular display, certain level of tone-mapping is necessary for the above mentioned HDR displays as well.

There are several ways to classify the TMOs into groups. Reinhard et al. [195] divide the operators to:

- *Global TMOs* – operators, which use the same DR compression function to map all the pixels.
- *Local TMOs* – operators modifying the DR compression function according to the pixel's neighborhood.
- *Frequency domain TMOs* – operators compressing DR according to the spatial frequency (mostly the lower spatial frequencies are extracted and compressed in DR, while high frequencies containing details are kept).
- *Gradient domain TMOs* – operators compressing DR by modifying an image derivative.

Banterle et al. [206] include also another group of *Segmentation TMOs* which firstly classify an image into different regions and based on that apply a different mapping function. Moreover, they further divide TMOs from these groups into two different categories – *Perceptual TMOs* and *Empirical TMOs*. The first are trying to model some aspects of HVS, while the latter are focusing on creating visually pleasant results exploiting the findings from other fields, such as photography.

Global TMOs

This section is dedicated to algorithms using the same mapping function for all of the pixels. These are mostly the least computationally demanding and intrusive in terms of naturalness corruption. On the other hand, for very demanding scenes, they could cause detail losses.

Perhaps the first known global TMO was proposed in 1984 by Miller and Hoffman [222]. Their work was motivated by the fact that physically based rendering algorithms in computer graphics can produce the image values in a range not displayable by the LDR display. Their TMO is designed to preserve the perceived brightness before and after mapping, so the two elements should have the same brightness ratio in the resulting tone-mapped image. The brightness is defined here as a function of luminance and they calculate it using the psychophysical model by Stevens and Stevens [223], so the algorithm is perceptual global TMO. However, the operator requires the image data to be scaled in the range od 0 to 1,000 cd/m^2 reducing its usefulness for current needs.

In 1991, Tumblin and Rushmeier [224] developed a perceptual TMO based on the same psychophysical data [223] as the previous one but with a fundamental difference. While Miller and Hoffman wanted the ratio of the brightness values remain the same, Tumblin and Rushmeier try to preserve the brightness values themselves. The concept was revised in 1993 [225].

Greg Ward introduced another concept for tone-mapping in 1994 [226] preserving contrast instead of brightness. In exploits the just noticeable difference (JND) in contrast (the smallest change in contrast that can be detected by HVS), thus only the detectable contrast steps are preserved in the LDR version. The operator uses the HVS model introduced by CIE [227] and linearly maps the input to the output.

This approach was also included in TMO by Ferwerda et al. [228] and later on made interactive by Durand and Dorsey [229]. It uses different psychophysical data, also considering the influence of scotopic

vision. On the other hand, the mapping is still linear which was proved to be less effective than non-linear TMOs.

The most simple non-linear TMOs are logarithmic and exponential, as described in [195]. The basic concepts are very simple and computationally undemanding but also sufficient for the simplest scenes only. Nevertheless the concept is extended and used in some more sophisticated algorithms.

Drago et al. [230] used logarithmic mapping but with adaptive changing of the logarithm base according to the pixel intensity. A bias function introduced by Perlin and Hoffert [231] is employed for smooth interpolation between the logarithm bases.

Reinhard and Devlin [232] model the behavior of photoreceptors in HVS by sigmoid functions (different for every color receptors, since their operation is believed to be highly independent).

Ward et al. [233] came up with a different approach, inspired by the field of image enhancement, in particular histogram adjustment. They take a logarithm of an image, which was previously downsampled to resolution corresponding to cca 1° of visual angle. The histogram is computed and used to guide mapping. Together with that, another properties of HVS are considered to provide a curve leading to the most realistic result.

Complementarily to the TMOs employing complicated HVS models, Schlick [234] proposed a simple tone-mapping method improving the basic TMOs such as logarithmic or exponential mapping. It uses rational quantization function resembling the sigmoid function. The advantages are its simplicity and computational effectiveness, the problem could be with selecting the parameters which are not calculated from any calibration model but have to be specified by user.

Mai et al. [235] looked at the TMO design from the perspective of backward compatible HDR compression [236] and tried to suggest the optimal mapping curve with respect to this application. The approach was further improved by Lauga et al. [237].

Oskarsson [238] approached tone-mapping as a clustering problem. In his paper, the k-means clustering is used to optimally map the luminance values in an HDR image into the number of luminance levels required by a display. The method is efficient, require setting of few parameters only, and can be easily extended for HDR video.

All of the global DR compression algorithms can be extended to become local operators by including e.g. bilateral filtering [177] for splitting the base and detail layer of the scene. The base layer is then compressed by the TMO and afterwards combined with details.

Local TMOs

The main drawback of the global TMOs is that in some cases, they are not able to reproduce all the fine details appearing in the HDR image. For this reason, local operators have been introduced. Unlike global TMOs, their mapping function is progressively adjusted according to the neighborhood of a particular pixel under consideration. This enables to treat images more adaptively and therefore reproduce more details. The disadvantages brought by the new concept are mainly higher computational requirements and higher probability of naturalness corruption.

The first local approach towards tone-mapping was introduced in 1993 by Chiu et al. [239]. They observed that in digital photography, the dynamic range is ostensibly increased by dimming the regions around bright objects. This technique is used within the operator. The pixel is mapped based to its value in the image and in the image filtered by a low-pass filter (according to the authors, the result is not highly dependent on the type of the filter). Although the method comes from the techniques used in photography, it introduces strong halos (contrast inversions) which are, in most cases, visually unpleasant. Also the dependence on the filter kernel size and a user operated contrast parameter is significant. Generally, the larger kernel sizes should be used to produce at least somewhat natural looking images.

Different local TMO was proposed by Rahman et al. [240] and is based on their version [241, 242] of the Retinex theory, originally developed by Land and McCann [243]. In fact, the operator is in many ways similar to the previously presented Chiu's TMO but it operates separately on each color channel and the

operations are performed in the logarithmic domain, which leads to visually better results. Moreover, the algorithm is also extended to the multiple scales (the decomposition is done by different kernel size of the Gaussian filter, particular results are then weighted by the user parameter).

Meylan and Süsstrunk [244] also used the retinex theory in their TMO. It employs principal component analysis (PCA) to decorrelate the color channels. The first channel then contains luminance information (the result of the PCA is an opponent color representation). This luminance channel is processed separately in parallel with processing of the original RGB image. Global adaptation with respect to the average luminance, calculated from the log-encoded pixel values, is applied on both branches, which are then transferred to the log-domain. The luminance branch is filtered by the modified retinex filter and put together with the chrominance components obtained from the second branch by PCA.

Another possibilities come from the area of color appearance modeling. Here, the perception of colors under different lighting conditions is studied. Popular color appearance models (CAMs) are e.g. CIECAM97 [245], CIECAM02 [246], or Hunt model [247]. The first to bring CAMs to the HDR image tone reproduction were Pattanaik et al. [248] in 1998. The model mimics HVS processing on multiple (7) scales by Gaussian filtering with different kernel sizes. It considers responses for rods and cones (separately for short, middle, and long wavelengths) with different gain controls and human spatial contrast sensitivity functions (CSFs), as measured in [249]. The calculation of chromatic and achromatic channels is according to the Hunt model [247]. For producing the result, the model is inversed with the parameters corresponding to the display used for visualization.

In the same field, Fairchild and Johnson proposed the image color appearance model (iCAM) [250, 251]. It is a modification of CIECAM02 model and operates in several color spaces. In the algorithm, the chromatic adaptation is used to push the values towards the D_{65} white point. The level of this color adaptation is influenced by the user operated parameter. This parameter also influences the amount of DR compression and haloing. Next, a more sophisticated gamma correction is performed. After these modifications, the process is (as for all CAM based tone-mapping) applied in the reversed order to map the values to match the properties of a display. The concept has been revised in 2006 by Kuang et al. [214] to be more adapted for the HDR tone reproduction applications. The algorithm inherited several attributes from its ancestor but uses bilateral filtering [177] to separate the base and details layers, which are treated separately. The gamma correction is substituted by more appropriate photoreceptor response functions and the whole framework considers both photopic and scotopic signals. The detail enhancement simulates several properties of HVS such as Stevens effect (increase in luminance causes higher perceived local contrast), Hunt effect (increase in luminance causes higher perceived colorfulness), or Bartleson-Breneman surround effect (the perceived contrast is increased when the surround changes from dark to dim to light).

Ashikhmin [252] presented another HVS based algorithm, trying to preserve the perceived local contrast. Firstly, it calculates the local adaptation level of the scene. The local neighborhood of the pixel is selected as large as possible with the constraint not to cross any strong gradients (determined by differences of Gaussians). The algorithm then maps the obtained levels to the adaptation level of the display, which is calculated as a function of the scene adaptation level using a novel concept of perceptual capacity of luminance values' range, exploiting the relativity of JND. The advantage of the operator is a lack of user parameters (except for the threshold for JND). On the other hand (as for most of the perceptual TMOs) the image has to be in absolute illuminance values to provide a correct result. If this is unknown, scaling has to be applied (which could be technically considered as a user parameter).

A typical example of a *empirical* TMO (i.e. operator not based upon an HVS model but on experience from another field) is Photographic Tone Reproduction algorithm, developed by Reinhard et al. [253]. This TMO is one of the most widely used, perhaps for its usability because authors have also provided a plausible way of its parameters estimation [254]. The method can be used as a local, as well as global. According to [254], the local aspect is to be used only if the DR is higher than 11 f-stop (logarithm of the base 2 of the scene's DR). The global version only linearly maps the values, multiplied by certain constant called *key of the scene* which should be chosen differently for dark/dim/bright scenes. This actually simulates the exposure Also the possibility to clip values higher than the threshold to the pure white is

provided. If the local component is included, a photographic technique called dodge-and-burn is employed – the exposure is varied according to the pixel neighborhood. The size of neighborhood is, similarly to Ashikmin's method [252], to be as large as possible without crossing edges (again, differences of Gaussians are used). The resolving effect is influenced by user specified sharpening parameter ϕ.

Mantiuk and Seidel [255] proposed a three step tone-mapping that can model the behavior of most of the other operators by different setting of parameters. The first step is using a four-segment sigmoid tonal curve. This is basically a global TMO. The local aspect of the operator is represented by the modulation transfer function (MTF) which determines what frequencies are going to be compressed or amplified. The third step includes a color saturation.

The characteristics of the display and ambient light has been taken into account while tone-mapping by Mantiuk et al. [256], who developed an adaptive procedure minimizing the perceptual distance between original image and the image displayed on the screen.

Frequency and Gradient Domain TMOs

The border between frequency/gradient based TMOs and local ones is very thin, since the operations varying according to the frequency or gradient are actually varying based on the neighborhood of pixel under consideration. Nevertheless, it is common in the literature [195, 206] to separate this group from the local TMOs. In fact, the oldest operator can be put into this class.

It was introduced in 1968 by Oppenheim et al. [257] and uses the concept of *homomorphic* filtering. This is based on the assumption that an image is a product of the scene illumination and the surface reflectance. The assumption is approximately valid if all the surfaces are diffuse, therefore the specular highlights or illumination sources should not be present in the scene. The idea is to separate the two components (in a logarithmic domain) and perform the DR compression on the illumination component only. The authors propose to do this by whitening filter in the logarithmic domain which preserves higher frequencies and affects only lower ones (only the low Fourier components of the image's logarithm are adjusted). The algorithm can provide plausible results but it has been overpowered by more modern procedures.

The separation of the illumination and reflectance components was also studied by Horn [258], although from the different perspective. His approach relies on the fact that low frequencies cause small gradient changes, while the changes produced by higher frequencies are much more significant. Since the gradients are computed in logarithmic domain, the results represent contrast ratios. The separation is then attained by thresholding. The integration of the gradients remaining after the thresholding requires numerical solving of the Poisson's equation.

The concept was picked up in 2002 by Fattal et al. [259]. However, instead of thresholding, they use a multiscale compressive function which should respect the fact that sharp edges result in gradients with high magnitudes, while fine textures produce much smaller magnitudes. Thus they produce a compressed gradient field which is then integrated by solving the Poisson's equation as in the case of the method's ancestor.

Tumblin and Turk [260] proposed a way of base layer obtaining using low curvature image simplifier (LCIS). It is a multilevel decomposition approach, therefore several "simplified" versions are obtained and subtracted from the previous level, resulting in several levels of detail layers. In the end, all layers are put together with different weights and forming the final tone-mapped image. The main problem is therefore the selection of proper weighting.

Another way to separate the base and detail layer is previously several times mentioned bilateral filtering [177]. It has been introduced to the HDR community by Durand and Dorsey [261]. The idea is to smoothen the image with simultaneous preservation of edges. Similarly to the previously mentioned approach, the computations are performed in logarithmic domain. The detail layer is obtained from the image by dividing it by the base layer (bilateral filtered version). Virtually any TMO can be used to compress the DR of the base layer. In the original paper, authors used the Tumblin-Rushmeier algorithm [225]. They also suggested several improvements in terms of computational time, such as filtering downsampled version of the image

(since the main concern is about low frequencies). The detail layer is obtained from the image by dividing it by the base layer (bilateral filtered version).

Choudhury and Tumblin [262] pointed out several drawbacks of the bilateral approach (smoothing across sharp gradient changes and poor performance on the areas with high-gradient or high-curvature) and proposed a way how to resolve them. The method is called *trilateral filtering* and uses two bilateral filters sequentially. After calculation of image logarithm, the gradients are obtained, smoothed by bilateral filter, and used to guide the second bilateral filter. The function preventing of filtering across sharp gradient changes is also included. The technique has one user specified parameter, setting the size of the neighborhood for the first bilateral filter smoothing the gradient (recommended values is 21 pixels).

Li et al. [263] adjusted the multiscale decomposition framework in slightly different way, designed specifically for tone-mapping and inverse tone-mapping. The activity maps are introduced on every scale to prevent artifacts to appear in the final image by tuning down the gain in the areas with high activity. The idea is inspired by neurons. From these maps, smooth gain maps are computed and used to modify particular sub-bands. The method is independent of the method used for decomposition (e.g. Laplacian pyramid, wavelet decomposition, etc.).

Segmentation TMOs

The last group of operators is based on segmentation into several regions, which are then treated differently. Similarly to the previous section, these operators are technically local operators but the use of specific segmentation techniques enables them to create a separate category.

The first TMO using categorization was proposed by Yee and Pattanaik [264]. It classifies the regions according to the logarithmic image histogram. By merging small groups of pixels, the layers are obtained (the layers differ in the number of bins used for a histogram calculation). Adaptation luminances are calculated for particular layers and used as an input to the global TMO (Banterle et al. [206] use Tumblin-Rushmeier [225] operator in their HDR toolbox). Authors recommend to use number of layers higher than 16 to prevent artifacts creation.

Krawczyk et al. [265] developed their tone reproduction algorithm inspired by the anchoring lightness perception theory, introduced by Gilchirst [266]. The theory postulates that HVS perceives the highest luminance in a field of view as white (anchor). The areas covered by the maximum luminance seem to be self-luminous. To apply the theory on the complex images, classification is performed. The TMO uses *k-means clustering* in the histogram of a logarithmic image. Soft segmentation is applied, i.e. the probability that the pixel under consideration belongs to the group is calculated for each group. Every group has its own anchor (calculated as 95% percentile of the luminance levels). The final image is obtained from the logarithm of the original image, the anchor values, and the probability functions.

Lischinski et al. [267] provided another perspective on segmentation. In their interactive TMO, they let the user classify the scene on his/her own. The segmentation is done using the brush based tool, as shown in [268] and [269], with four possibilities – basic, luminance, luma-chrome, and overexposure – each setting the different constraints on the selected pixels. The algorithm then finds a locally dependent exposure function which is employed to produce the final image.

Lauga et al. [270] designed a TMO which attempts to find an optimal segmentation of the HDR image into dark and bright regions. After that, an optimal mapping curve is find for both types of regions by minimizing MSE between logarithm of the luminance of the original and reconstructed version.

Skipping the HDR Image Creation Step

This last approach is not a tone-mapping in its true sense. Although Banterle et al. [206] classified it as a segmentation based TMO, for the following reasons it has its own category in this document. In this case, the input is not an HDR image but the stack of LDR images with different exposures (which are normally used for HDR creation as described in Chapter 8.1). The Banterle's HDR toolbox for MATLAB [206] enables also an HDR input but the images are artificially divided into differently exposed images.

The approach was firstly proposed by Mertens et al. [271]. It uses three metrics – contrast, saturation, and well-exposedness – to weight the importance of the pixel from particular LDR images in the stack. The resulting image is created by blending the exposures together with corresponding weights using Laplacian pyramid to avoid unnatural steps.

Same idea was adapted also by Song et al. [272]. They calculate the luminance values which maximize the visible contrast over different exposures and avoid the gradient inversion which causes halos in the final image. The result is obtained via a probabilistic model.

8.3 Selecting Parameters of Tone-Mapping Operators

The goal of the tone-mapping is to reproduce the HDR scene as faithfully as possible while preserving details and naturalness. However, these two aspects are mostly contradictory. It is therefore necessary to find a good balance between them.

A common property of most of the TMOs (and other post-processing algorithms) is that they have (mostly several) user-adjustable parameters. These parameters serve for adapting the operator for particular displaying scenario including viewing conditions, displaying device, and, more importantly, the particular content. In other words, optimal setting of the parameters differs with the scene. While this variability can be an advantage e.g. for artists, in some applications, where higher number of scenes needs to be processed, it is not practical to set all of the parameters manually for each scene. Therefore, some TMOs without the necessity for the tuning have been proposed, e.g. by Ward et al. [233]. Reinhard et al. [254] formulated the calculation of parameter values for the Reinhard TMO [253] directly from the scene, which makes the use easier and more suitable in the above mentioned cases. The influence of the viewing conditions on the preferred setting was investigated by Stokkermans et al. [273]. Barladian [274] introduced the estimation of parameters for the combination of TMOs proposed by Tumblin and Rushmeier [224], and Reinhard [253]. The issue is even more crucial in HDR video processing. The automatic tone-mapping for video was proposed e.g. by Kiser et al. [275].

However, not all of the TMOs provide a way to estimate the parameter values from the scene. Such variability can be a problem when comparing different TMOs. In TMO comparison studies, the parameters were mostly either left in the default setting or adjusted by authors to “the highest subjective quality”. The unfairness of such comparisons was already pointed out by Petit and Mantiuk [143] who used several different parameter settings for each TMO to increase the balance.

It seems reasonable to define an objective criterion for optimization of parameters to be used instead of the HVS in maximizing the perceived quality. This should also depend on the application – while naturalness seems to be of higher importance for Quality of Experience (QoE) in multimedia applications [9, 10], in security surveillance it becomes secondary as it is much more desirable to preserve all the details, i.e. the intelligibility is more important than QoE.

8.3.1 Tone-Mapping Operators Parameters Optimization in Security Applications

The security and surveillance systems are one of the most interesting application areas of the HDR technology. The ability to capture all the details in the scene even under challenging illumination conditions brings a considerable advantage over classical systems. Although the modern surveillance solutions use HDR security cameras, the captured video is being displayed on the regular screens. Considering the ubiquity of surveillance systems together with the high price and technological complexity of the HDR displays, this setup will most probably persist in the future systems as well. The importance of tone-mapping in such applications is therefore very high. However, the requirements on the TMOs are different than in the case of multimedia content. Specifically, it is necessary to reproduce as many details as possible because the loss of the information from the scene can have a massive impact on the credibility of the footage. Additionally, the computational power of such a system is mostly limited and therefore the simpler (probably global) TMOs are more suited in this scenario.

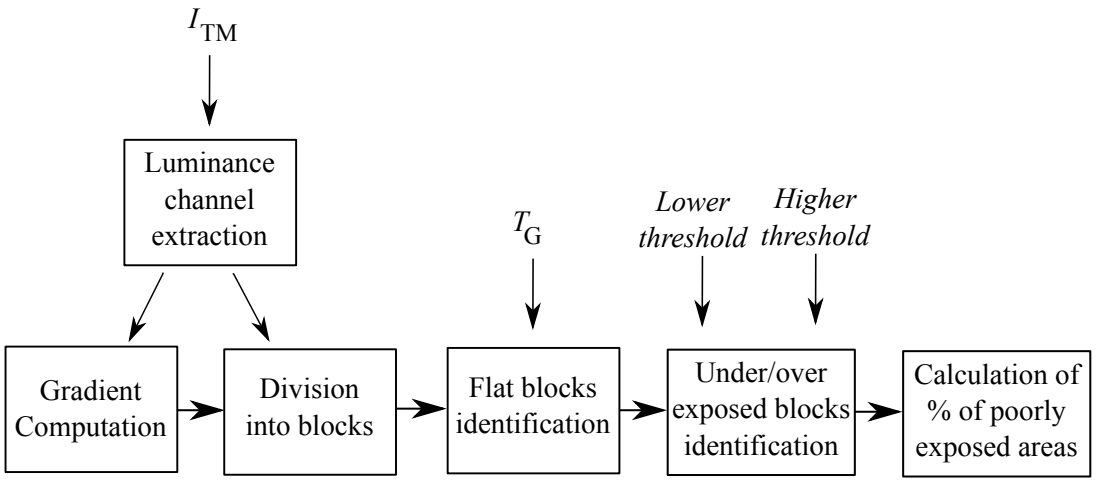


Figure 8.4: Block diagram of the proposed criterion.

Here, an optimization criterion for automatic TMO parameters tuning that is specifically geared towards security applications is described. This requires that more emphasis to be put on preserving visual details in the scene and attributes such as naturalness, perceptual quality can become secondary (although this need not always be the case). The proposed TMO parameter tuning approach is based on the assumption that the details are typically lost in the over or under exposed scene regions. Therefore, the TMO parameters are tuned in order to minimize the number of these areas. Obviously, such an approach is expected to limit detail loss due to tone-mapping operation. Another advantage of the proposed technique is that it can be used universally, regardless of the employed TMO. The performance will be demonstrated on four different operators – simple TMO using linear clipping and gamma mapping, logarithmic TMO, Reinhard global TMO [253], and iCAM06 [214]. The first two were selected for their simplicity but yet decent performance. Reinhard TMO belongs among the most popular operators and also provides the way to calculate the default parameter values, which is a good base for comparison. The last operator is the only local TMO. Its performance is very good across the content but the computational requirements are much higher. It is included to show the applicability of the proposed approach even for the more sophisticated operators. The method has been published in [276].

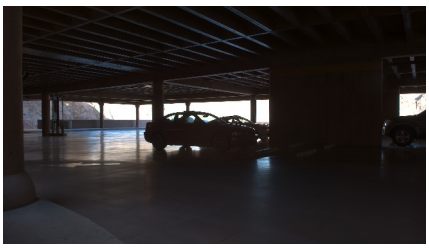
Method Description

As mentioned above, the underlying assumption is that information is lost from the image during tone-mapping, if the image contains underexposed and overexposed regions. These regions are identified as areas where the image function does not change and the mean luminance in the region is under (or over) certain threshold. The block diagram of the proposed criterion computation is in Figure 8.4.

The areas classification comes from the idea of Péchard et al. [277]. Firstly, the gradient in horizontal and vertical direction is calculated using Sobel kernel. The overall gradient is then obtained as $G = \sqrt{G_h^2 + G_v^2}$. Every block of 16×16 pixels is analyzed separately. It is classified as flat, if the sum of gradient within the block is lower than threshold T_G. This threshold was empirically set to 2550 (for 8-bit images). This value provided reasonable results across various content.

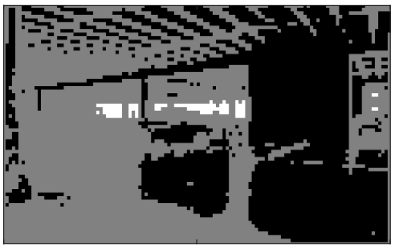
Once the block is identified as flat, the mean luminance within the block is calculated. If it is lower than 35 or higher than 220, to compensate for the influence of outlying values and a noise, the block is classified as underexposed or overexposed, respectively. It is true that the flat regions with low or high luminosity will be identified as under/overexposed even though it does not necessarily have to be the case but it can be assumed that these regions will exist in the image regardless the TMO parameters' setting and therefore do not influence the optimization process.

Finally, the percentage of the under/overexposed blocks is calculated. The optimization process finds the setting of parameters which minimizes this value. When evaluating the performance, the MATLAB



(a) Simple TMO – default parameters

(b) Simple TMO – tuned parameters



Under - 44.67% Visible - 54.12% Over - 1.21%

(c) Visibility map – default parameters

Under - 0% Visible - 100% Over - 0%

(d) Visibility map – tuned parameters

Figure 8.5: An example of the visibility map before and after parameters tuning.

function `fminsearch()` utilizing Nelder-Mead Simplex minimization [278] has been used. An example of visibility map before and after optimization of parameters for the simple TMO is depicted in Figure 8.5. Note that in this TMO, firstly a certain number of pixels from high and low luminance are saturated. As the default setting, 2.5% of pixels on both ends are clipped. After that, the dynamic range is linearly transformed between 0 and 1, followed by a gamma mapping. Default gamma is 2.2. These two parameters were tuned via the proposed approach. The image is from Fairchild's publicly available database [279]. The black regions represent the underexposed blocks, while the white label the overexposed ones.

In real surveillance applications some of the areas within the camera view can be more important than others. In these cases, the proposed method can be easily modified to incorporate the importance of those regions in calculations. This can ensure that information loss from these regions is particularly minimized.

There is an important issue that needs to be highlighted with regards to the applicability of the proposed method to HDR security video. Even though it is possible to optimize TMO parameters for each frame, this will lead to much higher computational costs. Nevertheless, the conditions within the camera view do not change rapidly. The TMO parameters can therefore be calculated and updated after certain number of frames. Such frame skipping mechanism will lower the computational and time requirements significantly thereby enabling the use of the proposed method in real-time.

Performance Evaluation

The performance of the proposed approach is tested on 9 images (see Figure 8.6 for downsized tone-mapped versions) taken from publicly available databases [279, 280]. These images were tone-mapped using the default parameter settings and using the optimization based on our criterion. To objectively compare outcomes, gradient based measure is utilized.

The details in an image are represented by changes in image function, which are reflected in the image gradient. Areas, where the gradient is equal to zero are therefore the areas without any details. Thus, we can measure the amount of details in an image as the number of nonzero gradient points. The results are depicted in Figure 8.7. The TMOs are abbreviated as S, Lo, R, and iC for simple, logarithmic, Reinhard, and iCAM06, respectively. The $\sim$ symbol marks the TMO with parameters optimized based on the proposed criterion. Since the results for the logarithmic TMO with default parameters are significantly lower than for the other TMOs, they are marked with an arrow in the bar graphs for better comparison.

The results show a clear improvement in details reproduction after the proposed optimization. The only

Figure 8.6: The images used for testing the performance of the proposed method.

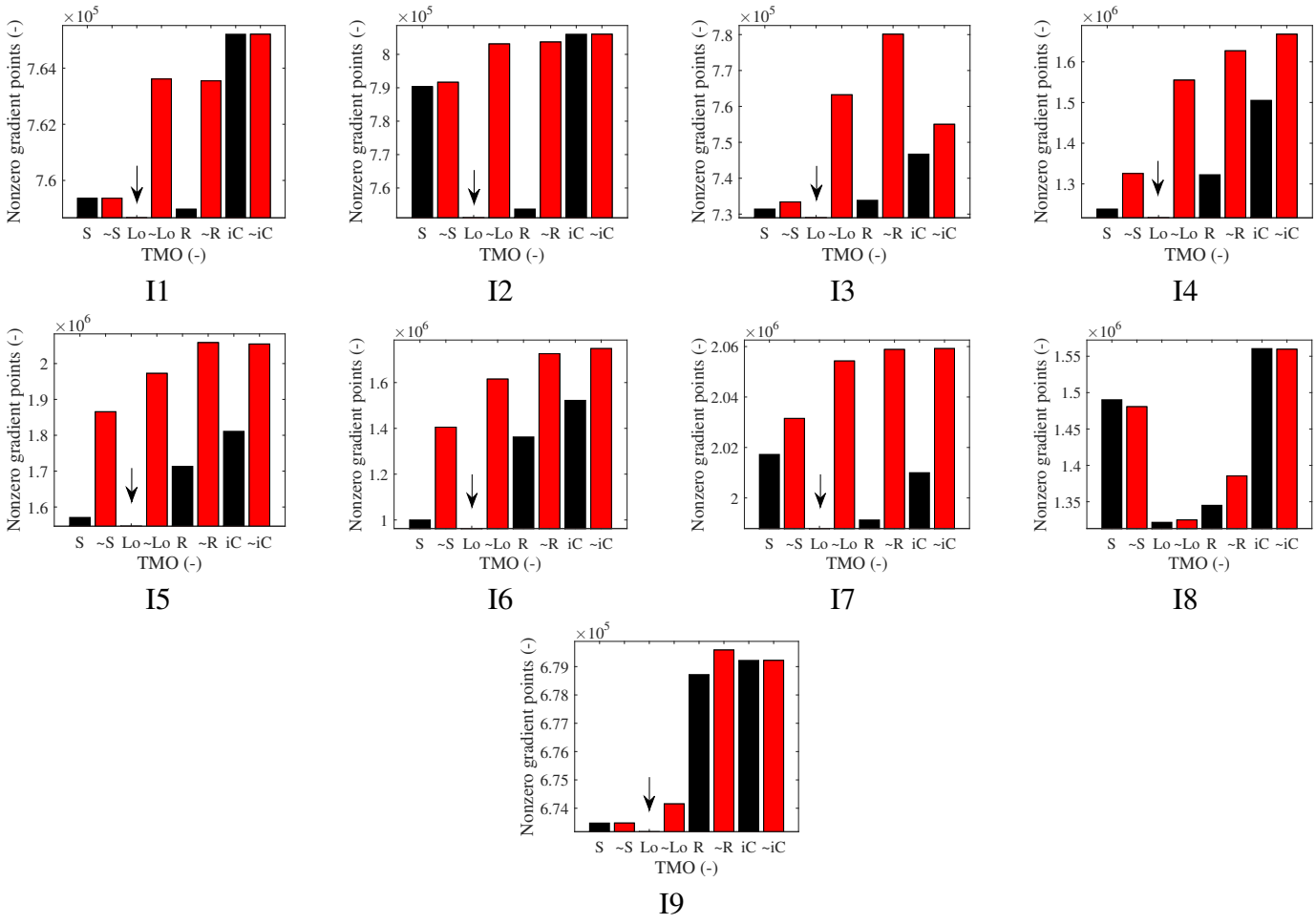
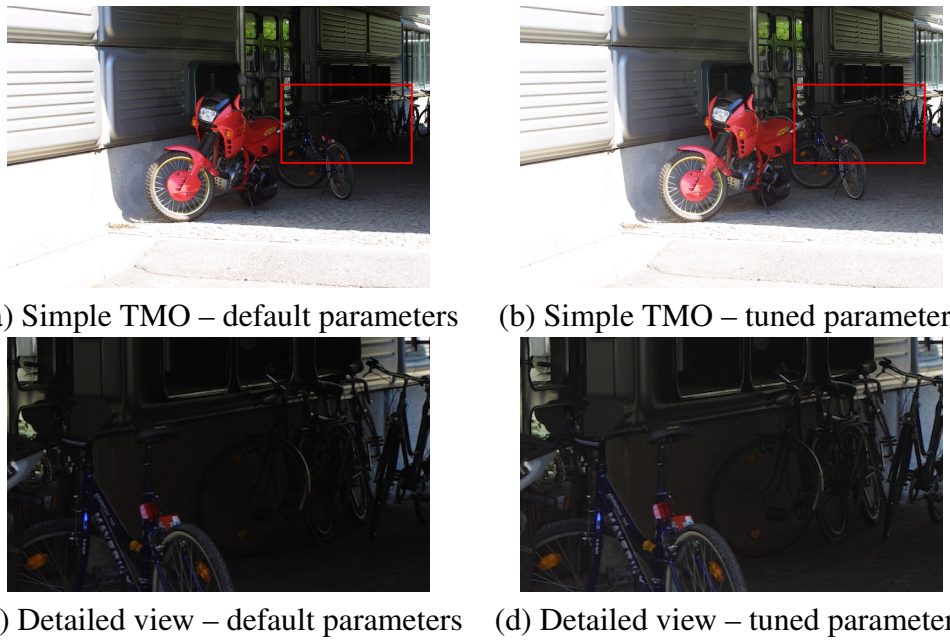


Figure 8.7: Results of objective TMO comparison.



(a) Simple TMO – default parameters (b) Simple TMO – tuned parameters

(c) Detailed view – default parameters (d) Detailed view – tuned parameters

Figure 8.8: Visible improvement in the content I8.

exception is the content I8 with the simple TMO, where the optimization actually lowered the number of nonzero gradient points. This is probably caused by the gamma correction which can, in fact, lower the overall contrast and confuse the given measure. Nevertheless if a closer look at the image itself is taken (Figure 8.8), there is a visible improvement.

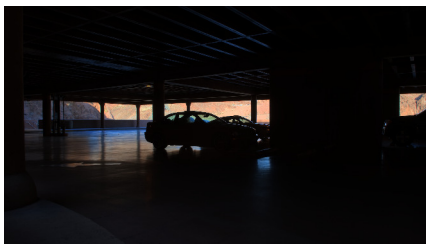
Another interesting outcome from the objective results is that after the proposed optimization, the images tone-mapped by the global operators become comparable to the local one. This is very important for the real surveillance applications.

To provide more detailed view on the performance, HDR images tone-mapped by all of the used TMOs with default and optimized parameters are provided. The results for the linear TMO for the content I5 have already been shown in Figure 8.5 (a) and (b). The outcome of the rest of the TMOs is depicted in Figure 8.9. Figure 8.10 contains the results for all of the used TMOs for the content I4. The results for all of the images can be found at `http://dbq.multimediatech.cz/users/krasula/public/spie/`.

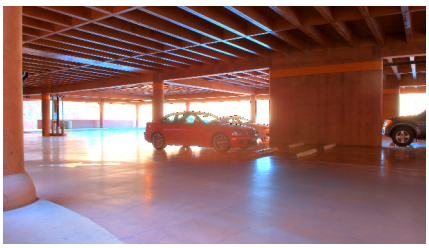
It can be seen that, especially for the logarithmic TMO, there are some overexposed areas in the resulting images. This is caused by the character of the original images which are mostly dark and thus challenging for the TMOs to reproduce all the details. Since all of the areas are considered to be of the same importance, the proposed method pushes the TMO to reproduce the details in the dark regions (because there are more of them), even though some of the details in bright areas will be lost. For this particular scenes, the linear TMO seems to work the best from the global operators, because the gamma correction enables it to reproduce the details in shadows as well as in highlights. Also the added value of the proposed approach to the local TMO is obvious.

8.3.2 Tone-Mapping Operators Parameters Optimization in Multimedia Applications

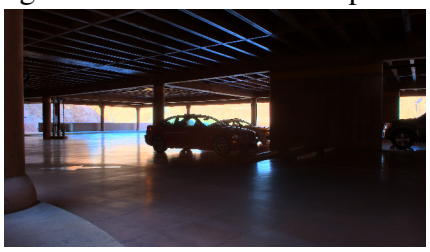
The previous section was specialized on the reproduction of details without considering the impact on the naturalness of the resulting image. In multimedia applications, however, the naturalness of the tone-mapped image plays a crucial role. It has been pointed out [9], that a single exposure image, where the naturalness is always maintained, is mostly considered as a good representation of the original real world scene and is more acceptable by the observers than the version created by TMO where the naturalness is corrupted. On the other hand, the main idea of HDR imaging systems is enhanced reproduction of details. A reliable parameters selection criterion should therefore maximize the details reproduction and



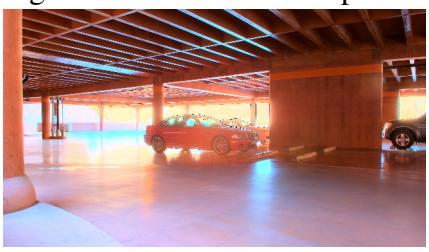
(a) Logarithmic TMO – default parameters



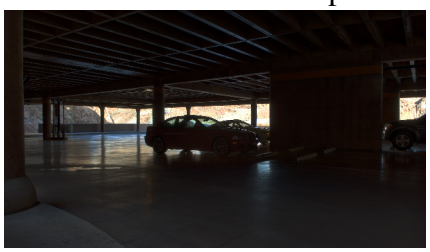
(b) Logarithmic TMO – tuned parameters



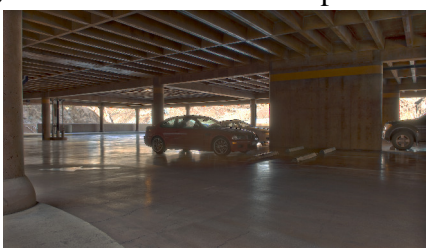
(c) Reinhard TMO – default parameters



(d) Reinhard TMO – tuned parameters



(e) iCAM06 – default parameters



(f) iCAM06 – tuned parameters

Figure 8.9: An example of the proposed method performance on the content I5.

simultaneously maintain the natural look of the scene.

To measure the reproduction of contrast, the best possibility is Dynamic Range Independent Metric (DRIM) [146]. However, its computational requirements make it impractical to be used in the optimization loop. The proposed method, therefore, focuses on simplicity and speed. It comes from the assumption that the reversal of contrast is the most undesirable artifact with respect to the contrast reproduction. The parameters selection algorithm, therefore, firstly finds the minimal value of contrast reversal possibly achieved by a given TMO. Secondly, a tolerance interval allowing some degree of contrast reversal is defined and used as a constraint. A novel metric of naturalness using combination the features most relevant for the natural look of an image is maximized within this tolerance interval. The particular parts of the parameters optimization procedures are described below. They have been published in [281].

Contrast Reversal

The criterion for contrast reversal quantification is based on calculating the gradient in both HDR and LDR versions of the scene using horizontal and vertical Sobel kernel (computations in more directions were also tested but did not provide any additional benefit in terms of performance). Note that the gradient is computed for the luminance component of the images only. The horizontal and vertical gradient images G_h and G_v (calculated the same way for the HDR image I_HDR and LDR image I_TMO) are therefore obtained as

$$\begin{aligned} G_\mathrm{h} &= \mathcal{L} * h_\mathrm{hor}, \\ G_\mathrm{v} &= \mathcal{L} * h_\mathrm{ver}, \end{aligned} \tag{8.15}$$

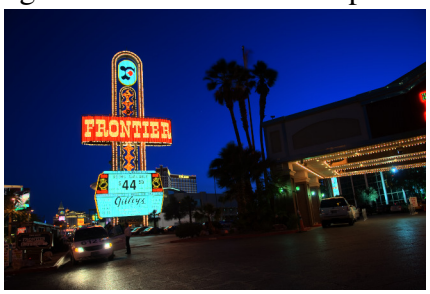
where $\mathcal{L}$ is the luminance component of the particular image, h_hor and h_ver represent the horizontal and vertical Sobel kernels, and the operator $*$ stands for the two-dimensional convolution.

(a) Simple TMO – default parameters

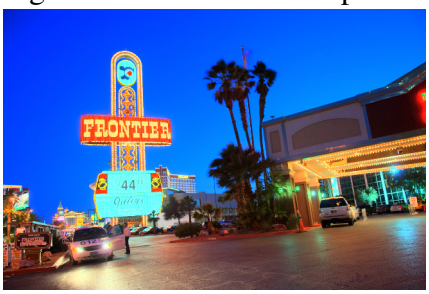
(b) Simple TMO – tuned parameters

(c) Logarithmic TMO – default parameters

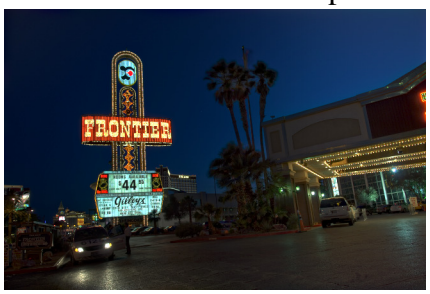
(d) Logarithmic TMO – tuned parameters



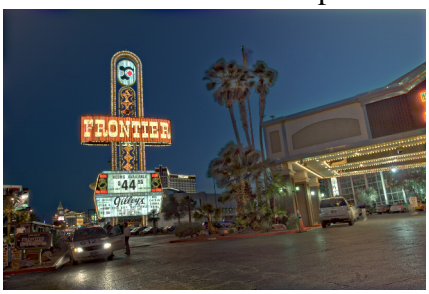
(e) Reinhard TMO – default parameters



(f) Reinhard TMO – tuned parameters



(g) iCAM06 – default parameters



(h) iCAM06 – tuned parameters

Figure 8.10: An example of the proposed method performance on the content I4.

Scene	Weber		Michelson		SDME		RMS		RME		GCF	
	KROCC	SROCC	KROCC	SROCC	KROCC	SROCC	KROCC	SROCC	KROCC	SROCC	KROCC	SROCC
1	-0.1547	-0.3168	0.0884	0.2156	-0.0221	-0.1650	0.3315	**0.5655**	-0.3315	-0.4884	**0.4199**	0.5633
2	0.3222	0.5330	0.0556	0.0705	-0.0111	-0.0308	0.6778	0.7908	0.3000	0.4427	**0.8334**	**0.9339**
3	0.3626	0.4330	-0.4066	-0.4593	0.4066	0.5209	0.7363	0.8813	0.4066	0.4681	**0.7582**	**0.8989**

Table 8.1: Performance of the selected contrast measures on the dataset developed by Čadík et al. [10] with ranks according to subjectively perceived contrast.

Scene	CIQI		CQE1		CQE2	
	KROCC	SROCC	KROCC	SROCC	KROCC	SROCC
1	0.8398	0.9483	**0.8840**	**0.9615**	0.7072	0.8801
2	-0.0663	-0.0506	**0.3094**	**0.4026**	-0.0221	-0.0330
3	0.5824	0.7582	**0.6703**	**0.8418**	0.5604	0.7451

Table 8.2: Performance of the selected colorfulness measures on the dataset developed by Čadík et al. [10] with ranks according to color representation.

Further on, the dominant gradient component in each pixel (x, y) is determined as

$$DG(x,y) = \begin{cases} 1 & \text{if } |G_{\text{h}}(x,y)| < |G_{\text{v}}(x,y)|, \\ 0 & \text{otherwise.} \end{cases} \tag{8.16}$$

The $ERR1$ is defined as the change in dominant gradient component

$$ERR1(x,y) = \begin{cases} 1 & \text{if } DG_{\text{HDR}}(x,y) \neq DG_{\text{TMO}}(x,y), \\ 0 & \text{otherwise,} \end{cases} \tag{8.17}$$

while $ERR2$ penalizes the cases where the gradient slope is reversed, i.e.

$$ERR2(x,y) = \begin{cases} 1 & \text{if sign}\{G_{\text{h,HDR}}(x,y)\} \neq \text{sign}\{G_{\text{h,TMO}}(x,y)\} \text{ or} \\ & \text{sign}\{G_{\text{v,HDR}}(x,y)\} \neq \text{sign}\{G_{\text{v,TMO}}(x,y)\}, \\ 0 & \text{otherwise.} \end{cases} \tag{8.18}$$

Note that no assumptions about the viewing conditions and HVS are made. The "worst case scenario" where all changes in luminance result in a perceivable change is considered. This is, of course, an over-simplification. Nevertheless, it enables avoiding computationally demanding modelling and necessity of a prior knowledge about the viewing conditions. Moreover, it is sufficient for the optimization criterion.

The final contrast reversal measure is a conjunction of the two errors

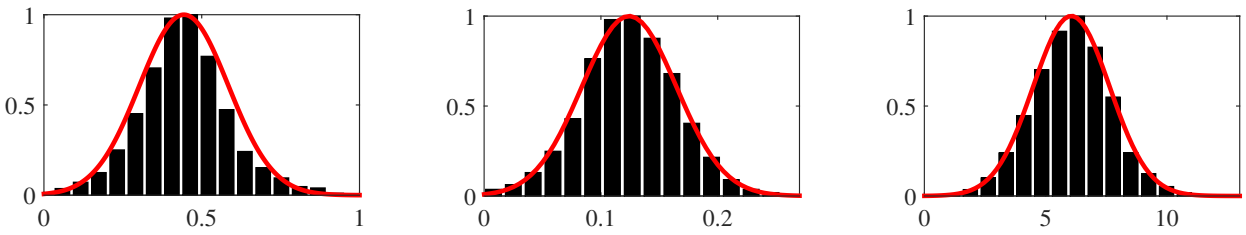
$$CR = ERR1 \cup ERR2. \tag{8.19}$$

The index used in the optimization process is a percentage of the pixels in which the contrast reversal has been identified.

Feature Naturalness

It is believed that the natural look of the image is a fusion of several factors [10]. The most important ones being overall luminance, contrast, and colorfulness. To select the most suitable metrics of these features, a dataset developed by Čadík et al. can be useful [282]. It provides the rankings of tone-mapped images not only with respect to their overall quality but also to the reproduction of brightness, contrast, color, and details.

The ability of the contrast metrics, described in Section 3.2.3, to predict the ranks according to contrast reproduction, measured by KROCC and SROCC (see Section 4.3), can be found in Table 8.1. The comparison of colorfulness metrics from Section 3.2.4 are shown in Table 8.2.



(a) Mean Intensity (b) GCF (c) CQE1 Colorfulness

Figure 8.11: Probability distributions of the particular measures in 5,000 colorful natural images.

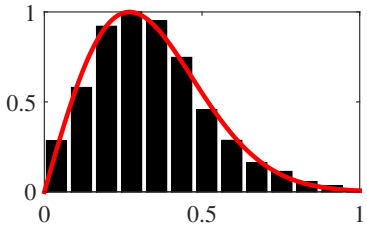


Figure 8.12: Probability distribution of the metrics' product for 5,000 colorful natural images.

The main criterion for the metrics selection was the computational simplicity and independence on the reference. It can be seen that the best predictions were provided by the GCF and CQE1 metrics, respectively. These feature metrics were, therefore, selected to be used in the novel naturalness metric.

Yeganeh and Wang [148] defined a statistical naturalness metric from the distribution of intensity and naturalness in gray-scale images. However, the measure completely omits the information about color which is one of the main factors influencing naturalness as well. The proposed metric design is inspired by the statistical naturalness part of TMQI. Firstly, 5,000 color images of different sizes and contents had been obtained from a publicly available Image Net database.[11] Then the mean intensity MI , GCF [106], and CQE1 colorfulness [73] for all of these images were calculated and their histograms were computed in order to estimate the probability distributions. It can be seen from the Figure 8.11, that all of the used metrics' histograms can be approximated by Gaussian (red curve).

In TMQI's statistical naturalness part, the assumption is that the intensity and the contrast are mutually independent, therefore no conditional distributions are necessary for joint distribution expression. However, this assumption is not valid when the colorfulness is added. Moreover, the conditional distributions would be very hard to estimate. We therefore tried to find the distribution of the criteria combination for each image (i.e. we calculated the product of the three metric values per image). The histogram is shown in Figure 8.12. The resulting histogram can be approximated by Rayleigh distribution as shown by the red curve.

The PDF of Rayleigh distribution is defined as

$$pdf_{\text{Rayleigh}}(x,\sigma) = \frac{x}{\sigma^2} e^{-\frac{x^2}{2\sigma^2}}, \tag{8.20}$$

where $x \geq 0$ and σ is a scaling factor. In this particular case $\sigma = 0.27$.

The proposed Feature Naturalness (FN) is then obtained as

$$FN = \frac{pdf_{\text{Rayleigh}}(MI \times GCF \times CQE1, \sigma)}{\max pdf_{\text{Rayleigh}}(x,\sigma)}. \tag{8.21}$$

The denominator serves for normalization. It is the global maximum of the Rayleigh PDF with respective σ from the equation (8.20). The higher the FN, the more natural should the image look.

[11] http://www.image-net.org/ (retrieved on 30/08/2016)

Optimization Procedure

The goal of the parameters optimization is to obtain the "best" representation of an HDR image which is suitable to be displayed on LDR displays. In the previous sections, two factors that play an important role in how a dynamic range-reduced HDR image will appear on LDR display were identified. Here, these two factors are used in an effort to optimize the visual appearance of the resulting tone-mapped image.

The said procedure is implemented as a constrained optimization problem. Therefore, the TMO parameters values which lead to minimum *CR* are firstly obtained. This *CR* value is employed as a constraint in the next step of the proposed method.

Let the $par = \{\tau_1, \ldots, \tau_n\}$ be the vector of TMO's parameters. The dimension of the parameter space is therefore n. Then

$$CR_{\text{min}} = \min CR(par). \tag{8.22}$$

In the second stage, the optimal TMO parameters values are found such that *FN* is maximized while the resultant *CR* is not more than $\delta\%$ larger than the minimum *CR* value (CR_{min}). The optimal set of parameters par_{opt} is therefore obtained as

$$\begin{aligned} par_{\text{opt}} &= \arg\max FN(par), \\ \text{s.t. } CR(par) &\leq CR_{\text{min}} \times (1 + \tfrac{\delta}{100}). \end{aligned} \tag{8.23}$$

The parameter δ controls the deviation allowed from the minimum *CR* value and therefore influences the relative importance of the two metrics (*CR* and *FN*). A higher value of δ gives higher importance to the *FN* metric. In this paper, we set $\delta = 5$ as it provides reasonable results across various content.

To avoid the "brute-force" approach, the Nelder-Mead downhill simplex optimization method [278] was also implemented within the algorithm. The method does not use any analytical nor numerical gradients but directly searches for the minimum in the parameter space. The dimension of this space is defined by the number of parameters.

For the purpose of TMO parameters optimization, the method has been slightly modified in order to work in the discrete parameter space rather than continuous which decreases computational requirements and the chance of ending up in the local minimum. In each step of the algorithm, when a new point of the simplex is calculated, the nearest point from the discrete space is taken. In cases where simplex becomes smaller than the step in discrete space, this could lead to false minimum detection. The stopping criterion has, therefore, also been modified. Before the minimum acceptance, all the neighboring points are tested and if any of the points results in the smaller function value, the algorithm searches again from this starting point.

Since the Nelder-Mead optimization technique is a downhill method, the calculated FN values are negated. It is also not adapted for the constrained optimization, therefore, in each step, it is checked if the CR value for the particular point is within the required range. If not, the FN value is set to ∞. In most of the cases, this approach is able to find similar results as if every point in the parameter space was searched.

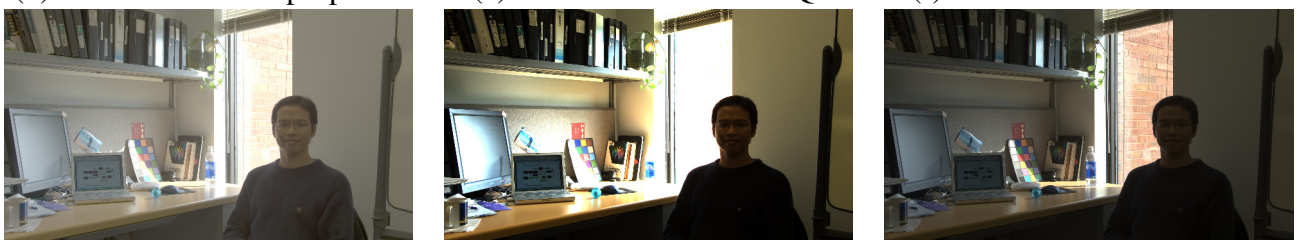
Performance Demonstration

To demonstrate the performance of the proposed approach, four TMOs are used. Global operators are represented by Drago TMO [230], Reinhard TMO [253], and simple linear tone-mapping with user adjustable clipping and gamma mapping. Selected local operator is iCAM06 [214]. Implementations of Drago and Reinhard operators are from the Banterle's HDR toolbox [206]. Recommended gamma correction was employed after tone-mapping with Drago TMO and color correction was used after Reinhard operator application. As a base for comparison, parameters maximizing TMQI [148] are also found.

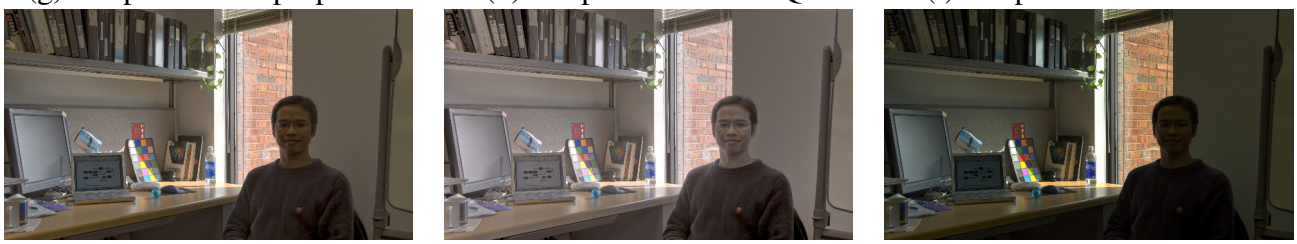
Not to bias the results by utilization of certain optimization technique, the parameter space of each TMO has been sampled and all the combinations were calculated. Minimum and maximum values for proposed method as well as for TMQI were found in these spaces. The optimal images as identified by proposed approach and TMQI for the "Willy Desk" image from the Fairchild's dataset [279] per TMO are depicted

(a) Drago TMO – proposed (b) Drago TMO – TMQI (c) Drago TMO – default

(d) Reinhard TMO – proposed (e) Reinhard TMO – TMQI (f) Reinhard TMO – default



(g) Simple TMO – proposed (h) Simple TMO – TMQI (i) Simple TMO – default



(j) iCAM06 TMO – proposed (k) iCAM06 TMO – TMQI (l) iCAM06 TMO – default

Figure 8.13: "Willy Desk" image tone-mapped by different TMOs with parameters set according to the proposed method, TMQI, and the default setup.

in Figure 8.13. Each row contains the images tone-mapped using particular TMO, while each column corresponds to the parameters selection method. Note that the image was cropped to the full HD resolution prior to the processing.

From the provided example, several things may be observed. For Drago TMO – Figure 8.13(a)-(c) – the default parameters setting results in slightly unnatural reproduction of colors and TMQI gives an image with very bright background (the wall behind the window) and darker foreground (less details on the t-shirt). The proposed method's outcome provides a good compromise between the two.

Default setup of the Reinhard TMO (Figure 8.13(d)-(f)) creates an image with a lot of underexposed areas (the foreground is very dark). The result of TMQI has again high contrast between foreground and background (even more than in the previous case). Proposed technique again maintains the details and naturalness better than the two.

The proposed method applied on the Simple TMO (linear clipping and gamma mapping) results in a small loss of contrast compared to the other methods but with the benefit of much better details preservation. The image also looks more natural.

The only local TMO employed here is iCAM06. It is also the only operator that actually influences the color components as well as the luminance. Here, the benefit of using also the information about color in the proposed method is visible. TMQI provides image with higher overall luminance but the colors are

very unnatural. The proposed technique on the other hand handles the colors in the better way and results in much more natural looking image. The TMO parameters selection approach will be further used in the quantitative subjective study design.

8.4 Subjective Quality Assessment of Tone-Mapped Images

The effort made in the area of subjective evaluation of tone-mapped images, together with discussion about the experimental setups and research questions, has been provided in Section 6.3.

The application of the subjective experiments most relevant to this thesis is testing and training of objective metrics. Publicly available datasets of HDR images processed by different TMOs including respective subjective scores are mainly restricted to the experiments performed by Čadík et al. [10] and Yeganeh and Wang [148]. The former contains 3 HDR source scenes processed by 14 TMOs and the raw scores from the rating and ranking experiments for each observer are available. It should be noted that these were collected for the printed images. In the latter, 15 source scenes and 8 TMOs were used. However, only the average ranks for each image are provided, making it hard to perform statistical analyses on the data. The TMOs employed within the experiments also do not result in color shifts and color artifacts. Moreover, none of the experiments considers rendered HDR content as an input. In the light of increased deployment of computer generated (CG) content in several applications, the metrics' abilities regarding this type of content should also be studied.

Therefore, it is meaningful to create another representative and challenging dataset for the objective criteria testing and training with natural and CG content. The following sections describe the design, conducting, and processing of the results obtained from the study [283, 284].

8.4.1 Source Content Selection

To create a representative dataset, as challenging content as possible should be selected. Even though the main attribute determining the degree of tone-mapping needed is dynamic range [285], several other properties of the scene can also play a role [286]. Two objective content selection methods have therefore been employed to identify the scenes with the most interesting parameters in the pool of HDR images (more than 150 images obtained from the publicly available databases [279, 280, 287, 288] or created within internal projects[12,13]).

Narwaria Method

The content selection method proposed by Narwaria et al. [286] is based on quantifying perceptual loss when decreasing the dynamic range. Firstly, M images with dynamic range linearly mapped in the range from 0 to maximum luminance which is always decreased with a factor Δ_m are created from i-th HDR image. For each contrast reduced image, the perceptual distance map with respect to the original is calculated using HDR-VDP-2 [53] (see Section 3.1.2 for more details). The difference between consequent maps is calculated using Kullback Leibler Divergence (KLD) based distance measure [289], defined as

$$dist_{\text{KLD}}(m,i) = \sum_{x=1}^{X}\sum_{y=1}^{Y} PDM_{\Delta_m,i}(x,y)\frac{PDM_{\Delta_m,i}(x,y)}{PDM_{\Delta_{m+1},i}(x,y)}, \tag{8.24}$$

where $PDM_{\Delta_m,i}(x,y)$ and $PDM_{\Delta_{m+1},i}(x,y)$ are the perceptual difference maps on successive levels for i-th content, and (x,y) are the pixel coordinates. Further, a clustering based analysis is performed in order to quantify the challenging nature of the content.

[12] http://www.ultrahd4u.eu/ (retrieved on 30/08/2016)
[13] http://www.images-et-reseaux.com/en/content/nevex (retrieved on 30/08/2016)

The distances for each source content are put into a vector $DIST_i = \left\{dist_{\text{KLD}}(m,i)\right\}$. The difference matrix $D = \{\ldots, DIST_i, \ldots\}$, in which the columns are the vectors $DIST_i$, is then analyzed with Fuzzy C-Means (FCM) clustering [290]. The algorithm iteratively minimizes the weighted within group sum of squared error to produce the best division into the specified number of clusters. The authors use two clusters – less and more challenging content (lower distances vs. higher distances). The result also provides a membership function specifying the degree of membership of the each content to the cluster. Therefore, the content can be ranked according to the probability of membership among the more challenging content. The higher the probability, the more challenging the content is. The algorithm is summarized in Algorithm 5.

Algorithm 5 Content selection method proposed by Narwaria et al. [286]

for $i = 1$: number of source images in the pool **do**
  for $m = 1 : M$ **do**
    Create an image with maximum luminance lowered by Δ_m
    Calculate $PDM_{\Delta_m,i}$ using HDR-VDP-2
  end for
  for $m = 1 : M - 1$ **do**
    Calculate $dist_{\text{KLD}}(m,i)$ according to equation (8.24)
  end for
  Create vector $DIST_i = \left\{dist_{\text{KLD}}(m,i)\right\}$
end for
Create the difference matrix $D = \{\ldots, DIST_i, \ldots\}$
Analyze matrix D using FCM clustering algorithm
Rank the source images according to the probability of membership in the more challenging cluster

Method Based on Under/Over Exposed Regions

The second method is based on the criterion proposed in the Section 8.3.1 of this thesis. It assumes that less challenging scene can be better represented by a single exposure image. The HDR image is therefore divided into set of single exposure images (i.e. the procedure inverse to bracketing when creating an HDR image). The procedure is implemented in Banterle's HDR Toolbox in MATLAB [206] as `GenerateExposureBracketing()` function. The minimum and maximum log2 luminance is firstly calculated and rounded (down for the minimum and up for the maximum). Minimum and maximum exposure times are then obtained as

$$\begin{aligned} t_{\text{exp,min},i} &= \left\lfloor \log_2\left(\min(\mathcal{L}_i)\right)\right\rfloor + 1, \\ t_{\text{exp,max},i} &= \left\lceil \log_2\left(\max(\mathcal{L}_i)\right)\right\rceil - 1, \end{aligned} \tag{8.25}$$

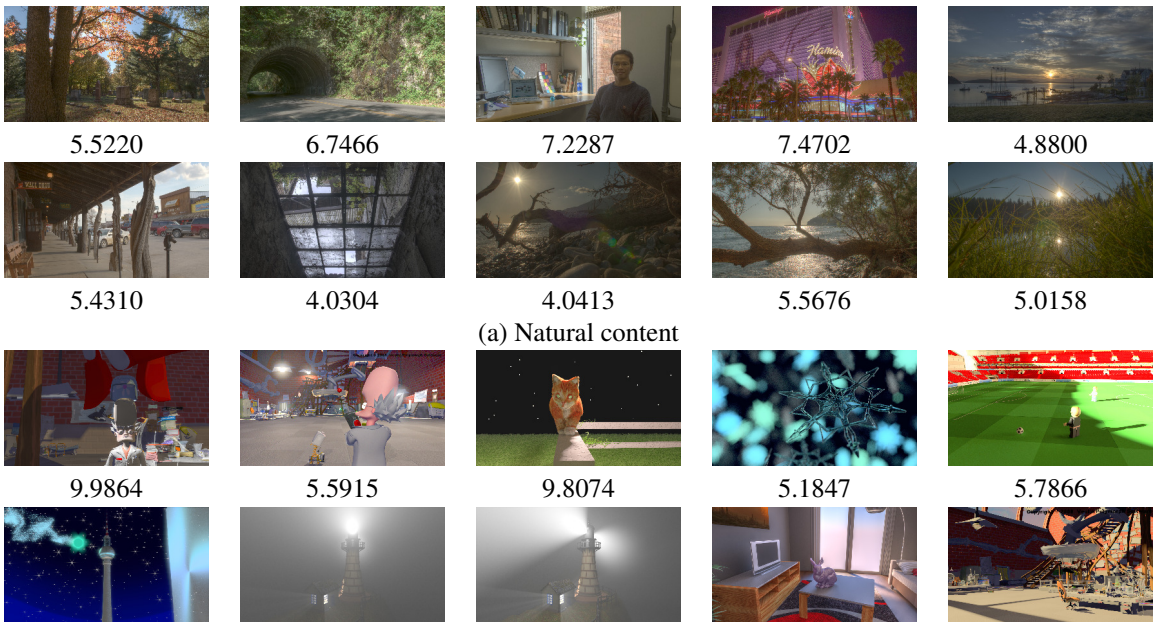
where $\mathcal{L}_i$ is the luminance of the i-th HDR image, operators $\lfloor . \rfloor$ and $\lceil . \rceil$ stand for rounding down and up. The vector T_i is then obtained by linear spacing between $t_{\text{exp,min},i}$ and $t_{\text{exp,max},i}$ with the step of 1. The m-th exposure is obtained as

$$exp_i(m) = 2^{-T_i(m)}. \tag{8.26}$$

The m-th single exposure image created from the i-th source content is then

$$I_{m,i} = \left(exp_i(m) \times \mathcal{L}_i\right)^{2.2}. \tag{8.27}$$

The area of underexposed and overexposed regions (see Section 8.3.1) is then calculated for each of them and the minimum value over the exposures is taken. The higher this value is, the more challenging the content is considered.



(a) Natural content

(b) Synthetic content

Figure 8.14: Downsized tone-mapped versions of the source content with log10 dynamic range.

Final Selection

The final content has been selected subjectively from the most challenging images as identified by the two methods. The main goal was to provide the diversity in terms of scene types, visual appeal, and low-level properties such as colors, spatial frequencies, etc.

Given the importance of the type of content and context, *10 natural* and *10 synthetic* HDR images have been selected. Since the display resolution was full HD, content with higher resolution was down sampled (using low-pass filtering followed by subsampling) and cropped such that no visible artifacts were introduced. Downsized tone-mapped versions of the source content, together with their log10 dynamic range calculated as

$$DR_{\text{log10}} = \log_{10} \frac{\max(\mathcal{L})}{\min(\mathcal{L})}, \tag{8.28}$$

are shown in Fig. 8.14. Note that the minimal non-zero luminance value is considered in equation (8.28).

Note that no highly photorealistic CG content has been included, since the mechanisms influencing the preference in such images can be affected by the level of its naturalness. To suppress the influence of naturalness as a variable, only noticeably computer generated content not attempting to reflect natural world has been selected. Such content can be found in applications such as cartoons, video games, artificial worlds in virtual reality, etc.

8.4.2 Selection of Operators and Their Parameters

To provide inter and intra TMO diversity (i.e. diversity across TMOs as well as within each TMO), four TMOs with two different parameters settings have been used. Additionally, TMO proposed by Mai et al. [235] which optimizes the mapping-curve for backward compatible HDR compression and therefore does not enable parameter adjustment has been included. Resulting 9 versions of each source content made the test well balanced in terms of time requirements and variability of images.

Global operators are represented by Drago TMO [230], Mai TMO [235], and simple TMO with linear clipping and gamma mapping. The representative of local TMOs is iCAM06 [214] and gradient based Mantiuk operator [255] is also included. The latter two can also affect color reproduction, and are thus very

1A Natural content Setup with reference	**2A** Synthetic content Setup with reference
1B Natural content Setup without reference	**2B** Synthetic content Setup without reference

Table 8.3: Description of the four parts of the experiment.

suitable for the purposes of the database. The use of the probably most popular Reinhard TMO [253] was discarded since it lead to results too similar to images produced by other TMOs.

The parameters of the operators (except for Mai TMO) were selected according to the optimization procedure described in Section 8.3.2 and by optimizing TMQI [148]. This was to create a conflict of the objective values and thus ensure that the content is challenging from the perspective of objective metrics as well. In cases where the two optimizations lead to perceptually close results (this has been checked visually), the parameters have been manually adjusted in order to provide perceptually different, yet appealing image. This happened in few cases only. None of the images in the resulting dataset are therefore visually similar. Final database is composed of 90 natural and 90 CG tone-mapped images.

8.4.3 Methodology

The possible subjective methodologies were already discussed in Chapter 6. Considering the multidimensionality of the quality (different artifacts, colors, contrast, etc.) the SS [11] procedure would be very demanding on the observers. The procedure has also been found more complicated in the context of image enhancement as described in Section 7.3.4. The reasons why the ranking [13] is impractical are stated there as well. PC [12] procedure is, therefore, again considered as the best alternative for collecting observers' opinions. To decrease the number of comparisons, ASDPC methodology is used.

Following the discussion from the Section 6.3, two setups are considered – scenario with the reference displayed on the HDR screen and no reference scenario. In the no reference scenario, the research question for the observers comes naturally – "Which image do you prefer?" On the other hand, in the scenario with the HDR display, the goal is to determine the observers' preferences (not fidelity to the reference) and see if they can be altered by the presence of the reference. To properly and clearly formulate the research question, more information about the dataset is needed.

A pretest with expert observers has been conducted in order to shed a light on observers' behavior. It has also served to determine the time requirements of the test. It has been discovered that, even with ASDPC methodology, all 20 source images (i.e. 360 comparisons) cannot be evaluated under 30 minutes which is the upper limit for observers' fatigue recommended in ITU-R Rec. BT.500 [11]. Therefore, it has been decided to separate the test in four parts according to the Table 8.3. In each part, 180 $(9 \times (\sqrt{9} - 1) \times 10)$ comparisons needed to be done.

Regarding the research questions, the expert observers were interviewed after finishing the task. All of them reported that several times they qualitatively preferred a version that was visually further from the reference, supporting the outcome of the study by Ashikhmin and Goyal [144] who discovered that when provided a reference, fidelity and preference scores may differ. Since tone-mapping is considered as a post-processing procedure in this thesis and as such, its goal is a maximization of the perceived quality, observers' preference are of more interest for the database. To check if the preference can be influenced by the presence of the HDR reference itself, it has been attempted to make the task in the setups with and without reference as similar as possible. After consultation with the expert observers, the research question in the scenarios with HDR reference (i.e. **1A** and **2A** in the Table 8.3) has been finalized to: "The real scene is shown on the central display. Which of the two side versions do you prefer?"

(a) Setup with HDR reference (**1A** & **2A**) (b) Setup without any reference (**1B** & **2B**)

Figure 8.15: Experimental setups with and without HDR reference.

8.4.4 Test Room

The test took place within the facilities of IVC group at Polytech Nantes. The room complied with the recommendations stated in ITU-T Rec P.910 [12] as shown in Table 2.2. The two viewing set-ups are shown in Figure 8.15.

For displaying the HDR images, SIM2 HDR47E S 4K display was used, which is a 16-bit, 47-inch, 1080p LCD TV with maximum and minimum displayable luminance of 4000 and 0.03 cd/m^2, respectively. The two LDR displays were 8-bit, 46-inch (Philips 46PFL9705H) with maximum displayable luminance of 200 cd/m^2. The displays were calibrated, gamma corrected, and color corrected using X-Rite i1Display Pro tool.[14] The viewing distance was set to approximately three times the height of the screen (active part). In the experiment with HDR reference, for each comparison, the observers saw three stimuli: one on the HDR display placed at the center and two on the LDR displays on either side as shown in Fig. 8.15(a). Since there are two types of displays, the room illumination was adjusted accordingly. In particular with HDR display (brighter) in the center, the illumination at the center (just above the HDR monitor) was set to form the luminance of 100 cd/m^2 while the diffused light made up the illumination for each LDR display (about 50 cd/m^2). Such a setup ensured a suitable illumination setting for the observers, and they were comfortable while viewing both HDR and LDR stimuli. A dual modulation algorithm [291] considering the PSF of the screen, local dimming, and coarse LED sampling was used for displaying HDR scenes. This mapping affects the scene's appearance which can by all means influence the results even though the observers were asked about their preference. The most accurate way to quantify the impact of the pre-processing is via calibrated subjective test wherein image (video) displayed on HDR monitor is compared with other displays (e.g. OLED). Unfortunately, this is not straightforward because the ambient light conditions in the two cases (HDR and OLED) can be different. The strategy of using different ambient light, as was used here, can be a possible trade-off to that end. On the other hand, an objective approach could also be used by employing perceptual models of visibility to quantify the loss of contrast/picture details as a result of pre-processing (the scenario leading to lower losses should ideally produce a better looking image). Such studies seem to be missing in the literature. However, for our purposes, the setup still ensured that the reference is the same for all the observers, unlike the studies with "real-world" reference.

8.4.5 Observers

20 observers participated in each part of the experiment (Tab. 8.3). Given the four separate parts, our study involved 80 observers in total. They were recruited through a university mailing list and therefore

[14] http://www.xrite.com/i1display-pro (retrieved on 30/08/2016)

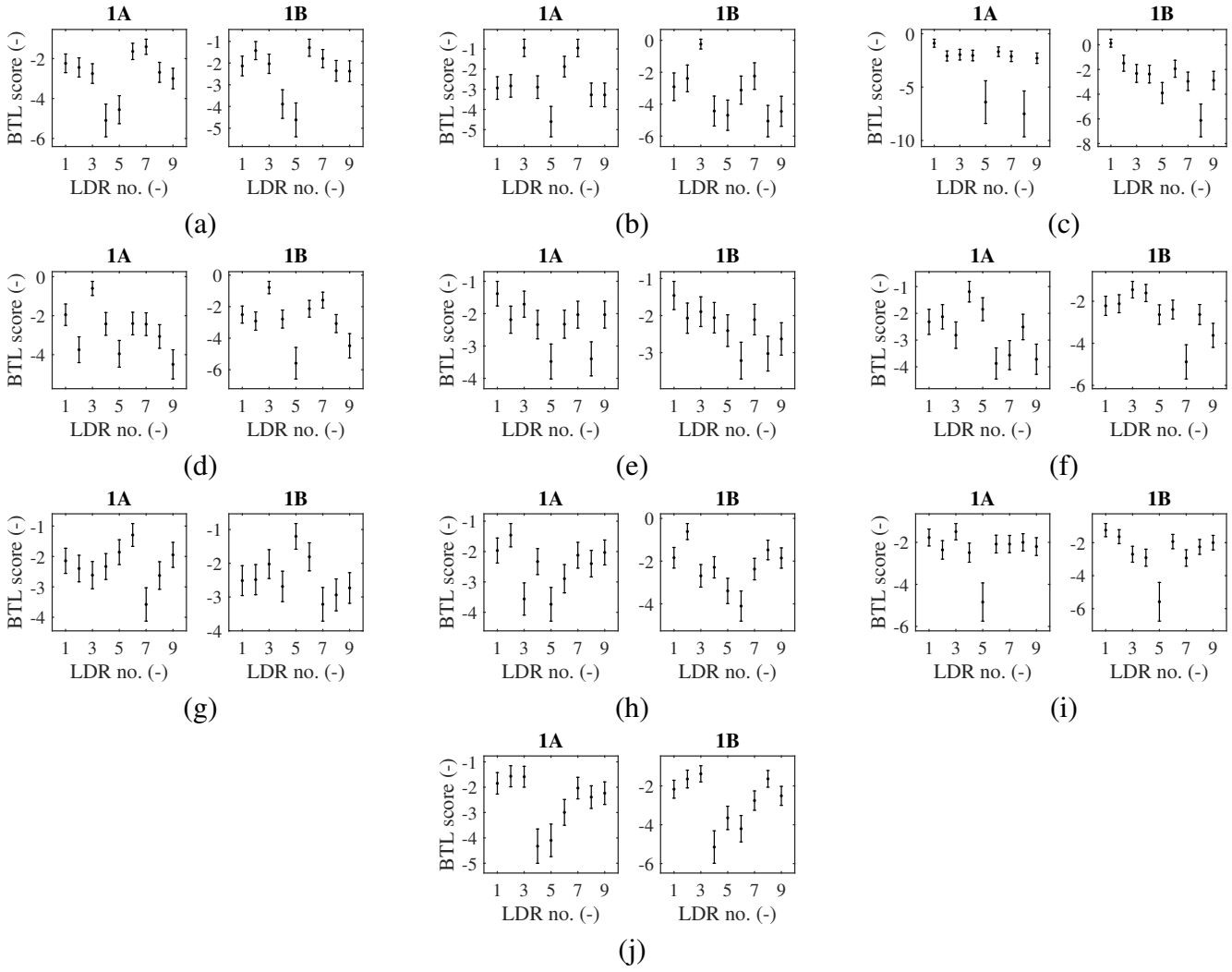


Figure 8.16: BTL scores with 95% confidence intervals for the natural content in both setups (**1A** & **1B**).

not experts in image processing. Each participant was checked for visual acuity and color perception with Monoyer and Ishihara test, respectively.

The subjects voted using a keyboard. After pressing an arrow, a green frame occurred around the corresponding image and the vote was confirmed by pressing Enter, thus minimizing the probability of error. At the beginning of the session, two random image pairs were shown, enabling observers to get familiar with the interface. The votes from these evaluations were not saved and included in the results.

8.4.6 Results

Observer's preferences were transformed to an interval scale using the BTL model [30]. Statistical significance of differences in BTL scores for each pair was also determined using the procedure described in [36]. The full results for natural and CG content can be found in Figures 8.16 and 8.17, respectively. The level of agreement between observers is higher in case of synthetic content resulting in smaller confidence intervals of BTL scores. Nevertheless, even for the natural content, clear trends can be identified.

Analysis of Differences Between Setups

Here, the influence of the HDR reference on observers' preferences is determined. To do so, the image pairs which were evaluated differently when the reference was shown and when not are found. This can be

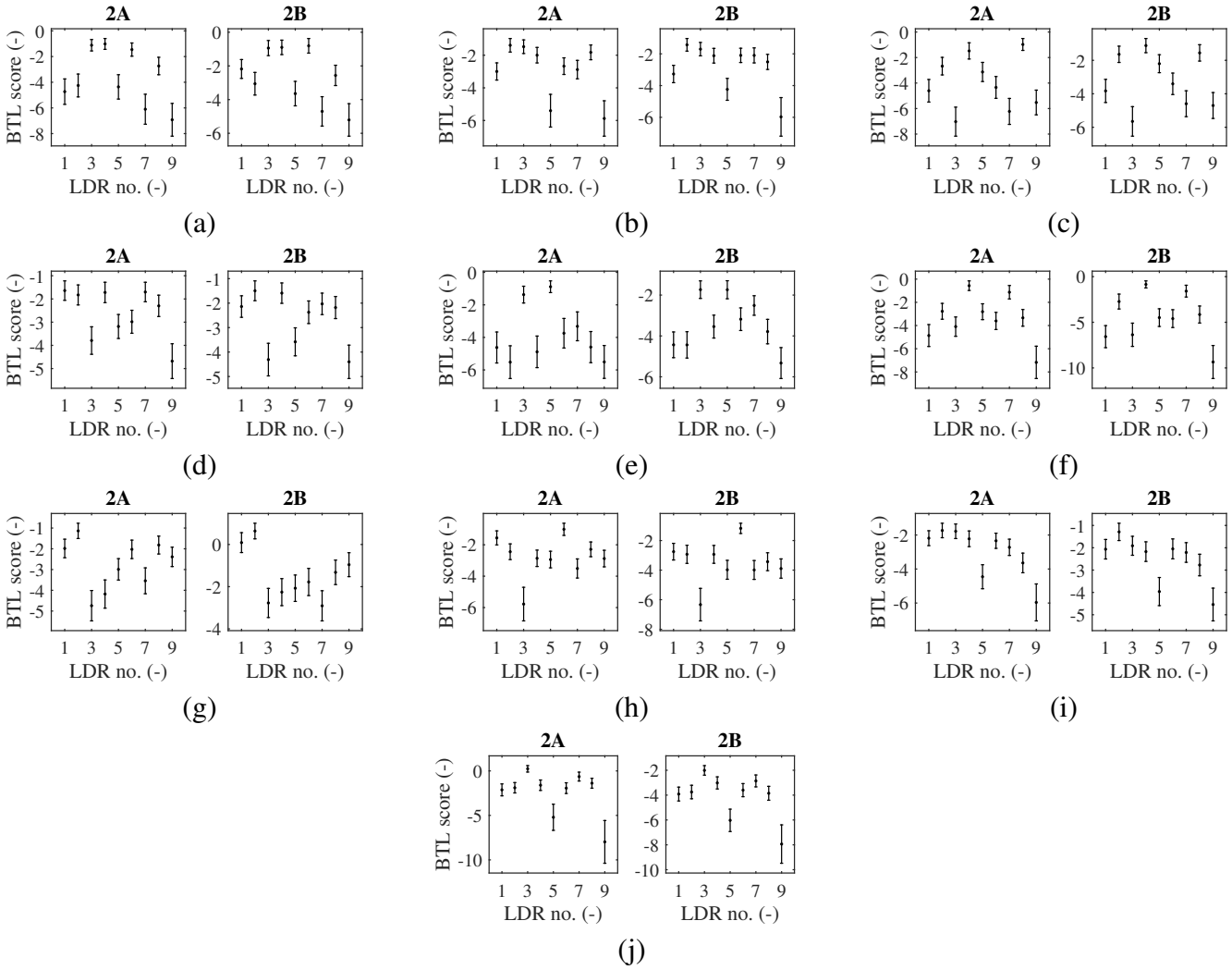


Figure 8.17: BTL scores with 95% confidence intervals for the synthetic content in both setups (**2A** & **2B**).

Natural content	**I1**	**I2**	**I3**	**I4**	**I5**	**I6**	**I7**	**I8**	**I9**	**I10**	**Σ**
$F = 1$	6	5	4	6	11	13	10	7	14	7	**83**
$F_S = 1$	0	0	0	0	1	2	1	1	2	0	**7**
Synthetic content	**I1**	**I2**	**I3**	**I4**	**I5**	**I6**	**I7**	**I8**	**I9**	**I10**	**Σ**
$F = 1$	5	5	2	6	4	1	5	2	2	4	**36**
$F_S = 1$	0	0	0	0	0	0	0	0	0	0	**0**

Table 8.4: Number of pairs identified as evaluated differently between the two setups (with and without the reference) for each source content.

quantified using a measure F, defined for images A_i and A_j as

$$F(A_i, A_j) = \begin{cases} 1 & \text{if sign}\{\Delta_{\text{BTR}}(A_i, A_j)\} \neq \text{sign}\{\Delta_{\text{BTN}}(A_i, A_j)\}, \\ 0 & \text{otherwise.} \end{cases} \tag{8.29}$$

where $\Delta_{\text{BTR}}(A_i, A_j)$ is the difference of BTL scores for stimulus A_i and A_j in the setup with the reference, $\Delta_{\text{BTN}}(A_i, A_j)$ is their BTL scores difference in the scenario without the reference, and sign$\{.\}$ stands for the signum operator. Thus, F indicates whether or not the presence of the HDR reference had an impact on the user preference, and is simply based on the sign of the difference between the BTL scores.

However, the reversal of the scores (i.e. $F = 1$) alone may not imply that the observed differences are statistically significant (i.e. it cannot be concluded from the data which image from the pair has higher perceived quality). Another measure F_S can be defined as

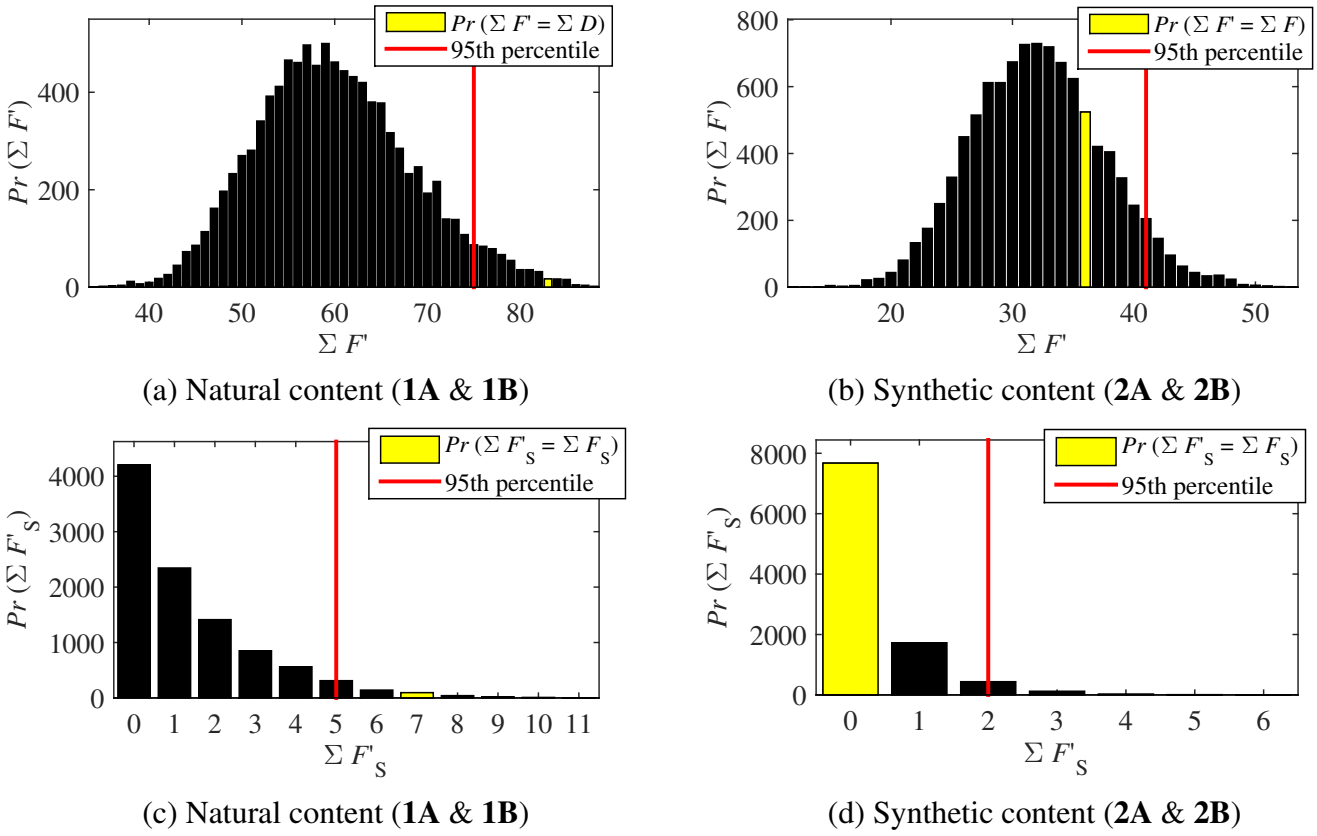
$$F_S(A_i, A_j) = \begin{cases} 1 & \text{if sign}\{\Delta_{\text{BTR}}(A_i, A_j)\} \neq \text{sign}\{\Delta_{\text{BTN}}(A_i, A_j)\} \bigwedge \\ & S_R(A_i, A_j) = 1 \bigwedge S_N(A_i, A_j) = 1, \\ 0 & \text{otherwise.} \end{cases} \tag{8.30}$$

where $S_R(A_i, A_j)$ and $S_N(A_i, A_j)$ represent the 95% significance of BTL scores differences in the scenario with and without reference, respectively. They are equal to 1 if the difference between scores is found to be statistically significant, otherwise they are 0. Thus the pair is considered to be evaluated differently (i.e. $F_S(A_i, A_j) = 1$) only if the BTL scores are reversed and the difference is statistical significant in both cases. Thus, F_S extends the quantity F by considering not only reversal of scores (i.e. sign change) but also the associated confidence levels. Hence, it represents a more reliable measure to quantify statistically the impact that the HDR reference might have on the observer preference.

Tab. 8.4 shows how many pairs were identified by the measure as differently evaluated in the two scenarios (i.e. with and without reference) for natural and synthetic content. It can be observed that the cross-scenario agreement between observers is truly higher for the second part of the experiment (i.e. with synthetic content). More importantly, there have been no cases where the scores for a pair would be reversed with statistical significance (there are no cases with $F_S = 1$). For further statistical tests towards determining if the differences in evaluation are caused by the presence of the display or just by the variations in observers' opinions, the permutation test (also known as bootstrapping) as described in Section 2.3.4 are employed.

Here, instead of the number of significantly differently evaluated pairs determined by Fisher's [37] or Barnard's [38] exact test (see Section 2.3.4), F and F_S are used to quantify the difference. Therefore, $\Sigma F'$ and $\Sigma F'_S$ is calculated in each iteration. This is more appropriate since the incomplete design of PC test was employed. The results obtained from 10,000 iterations are depicted in Figure 8.18.

From the probability distributions for $\Sigma F'$ it can, again, be inferred that the mutual agreement of participants was much higher in the case of synthetic content (median of $Pr(\Sigma F')$ and $Pr(\Sigma F'_S)$ for synthetic content is 32 and 0, compared to 59 and 1 for natural content, respectively). Nevertheless, the distributions in Figure 8.18 show that the permutation of observers in between groups influences the final result which suggest that there is an agreement among observers in both setups, i.e. the opinions are not random.



(a) Natural content (**1A** & **1B**)

(b) Synthetic content (**2A** & **2B**)

(c) Natural content (**1A** & **1B**)

(d) Synthetic content (**2A** & **2B**)

Figure 8.18: Results of Monte Carlo simulation after 10,000 permutations.

Second interesting result is that the influence of the HDR reference on the study outcome for synthetic content was not found statistically significant ($Pr(\Sigma F' < \Sigma F) = 0.7933$ and $Pr(\Sigma F'_S < \Sigma F_S) = 0.7673$) while this effect was present in case of natural content ($Pr(\Sigma F' < \Sigma F) = 0.9956$ and $Pr(\Sigma F'_S < \Sigma F_S) = 0.9925$). This is probably caused by a combination of several factors. To shed a light on perceptual mechanisms in play, a closer look to the image pairs that were evaluated differently between the two setups has been taken. Smaller versions of examples of such pairs are depicted in Figure 8.19.

The contradiction between images 8.19(a) and 8.19(b) is probably caused by the shift in color reproduction. Without the reference, the image 8.19(a) is possibly seen as more appealing for its more saturated colors. When provided the reference, observers can see that the image 8.19(b) has more natural reproduction of colors and even though they are asked about their preference, they tend to see the image 8.19(a) as much more unnatural than if the reference is not present. The HDR reference can also provide additional information such as at what daily hour was the scene captured which can also affect the judgment. The image reproducing the illumination more faithfully tend to be preferred even though the observers *were not* asked to choose the closer image. Such an example is represented by images 8.19(c) and 8.19(d). Without the reference, the image 8.19(d) is preferred for its colorfulness and details.

With regards to the CG content, it did not trigger such effects and the results from both setups can therefore be considered as equivalent. It is possible that the absence of the naturalness dimension enables observers to focus on the low level aesthetic properties (such as color saturation, contrast, clarity, etc.) only, thus depending less on the presence of the reference. However, further study with more synthetic content should be carried out in order to find out if the effects from Figure 8.19 cannot be triggered in case of CG images and if the presence of the reference does not influence the observers' judgement in case of this type of content. Since the color reproduction was one of the main factors driving the difference, it might be interesting to run the test with monochromatic natural images to determine whether the effect is maintained.

Nevertheless, the analysis confirms that the presence of the reference *can influence* the results of the study, even if the same research question is asked, and the highest care should be taken when drawing

Figure 8.19: An example of natural image pairs evaluated differently in the two experimental setups.

general conclusions from differently obtained datasets.

8.5 Performance of the Objective Metrics on Tone-Mapped Images

Compared to classical quality assessment tasks, objective evaluation of tone-mapped images brings several extra challenges. These were already discussed in Chapter 1. The full-reference metrics applicable to the tone-mapped HDR images are described in Section 6.4. Recently, Granados et al. [292] came out with two more measures for contrast loss and contrast waste, taking into account a camera's noise and human contrast perception. However, the implementation of the metric is not available.

Another way is to evaluate the quality of the tone-mapped image alone using no-reference measures. However, most of these criteria are trained on typical distortions (see Section 3.2.6) and their applicability out of the context therefore has to be verified. Nevertheless, several *distortion unaware* metrics have been introduced as well. Their description is provided in Section 3.2.7. Aydın et al. [5] pointed out that the aesthetic properties of the tone-mapped image are of high importance in the quality perception and proposed a measure based on sharpness, clarity, depth, and tone.

The majority of the mentioned objective methods have not been validated against the ground truth for tone-mapped content, and this is more so for the case of CG content. Hence, it is necessary to test them in this context. The analyses will be done on the data from each part of the experiment separately. It is also possible to merge the subjective data from particular scenarios and provide more general performance results but it is more informative to analyze the metrics with respect to each scenario, moreover, when it has been proven that the setup can have a significant impact on the results. The methodology from the Section 5.2 is again adopted.

15 existing objective methods have been selected for comparison. For the sake of brevity, the methods are denoted by numbering as described in Table 8.5 in following graphs. Since DRIM [146] only provides

2D error maps (calculated for each tone-mapped version with respect to the HDR reference), a simple averaging of those values has been adopted to obtain a single score. $DRIM_l$ is therefore the average for the contrast loss map, $DRIM_a$ for contrast amplification, $DRIM_r$ for contrast reversal, and $DRIM_{all}$ for average of all the computed maps. Note that since lower values of DRIM maps correspond to higher quality, the score differences for this metric have been inverted. The MATLAB implementation was obtained from the authors. Lastly, scores for the block based version of the FISH [69] metric ($FISH_{bb}$). Technically, it is a no-reference metric for sharpness but its performance on the aesthetics of photographs taken by different cameras [293] makes it a good candidate for the given task as well. The analysis has been published in [284].

TMQI [148]	TMQI-II [152]	$DRIM_l$ [146]	$DRIM_r$ [146]	$DRIM_a$ [146]
1	2	3	4	5
$DRIM_{all}$ [146]	$FSITM_r$ [153]	$FSITM_g$ [153]	$FSITM_b$ [153]	BLIINDS-II [80]
6	7	8	9	10
NIQE [84]	QAC [85]	CS [86]	CurveletQA [83]	$FISH_{bb}$ [69]
11	12	13	14	15

Table 8.5: The numbering of the evaluated metrics.

8.5.1 Results of Different vs. Similar ROC Analysis

Firstly, the metrics' capabilities in distinguishing between significantly different and similar pairs (i.e. to what extent the differences in metrics scores are smaller for similar pairs) are determined. The resulting AUC values for this analysis in all four scenarios can be found in Figure 8.20. Significance of differences is calculated according to the procedure described by Hanley and McNeil [135] and to compensate for multiple comparisons, Benjamini-Hochberg procedure [125] is used. In the significance plots, if the method in the row works statistically significantly better than the one in the column, the corresponding square is white. Black signifies the opposite case. Gray square marks the cases without significant difference in performance. The AUC value lower than 0.5 means that according to such metric, similar images are mostly having larger differences in scores than different images.

Based on the AUC values, we can see that none of the metrics can cope particularly well with the task of recognizing different and similar pair. Most of the performances are around AUC value 0.5, which is equivalent to a random guessing. Moreover, statistical significance plots show that the performances are mainly equivalent. For natural content, $FSITM_r$ (#7) metric reached the highest AUC values in both scenarios (see Figure 8.20(a) and (b)). The poorest performance is achieved by $DRIM_a$ (#5) in scenario **1A** (Figure 8.20(a)) and $FSITM_b$ (#9) in the scenario **1B** (Figure 8.20(b)).

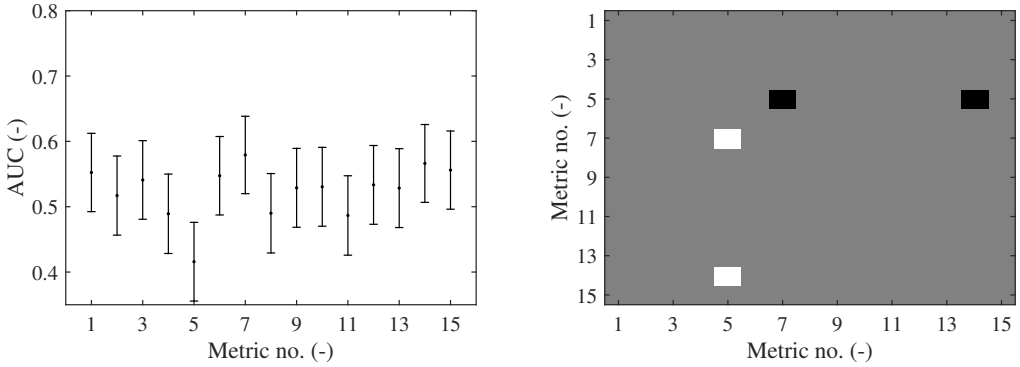
When the CG content is considered (Figure 8.20(c) and (d)), the best performing method is $DRIM_l$ (#3), closely followed by $DRIM_{all}$ (#6). TMQI-II (#2), on the other hand, performs poorly. This might be caused by its base in naturalness estimation.

The performances are generally slightly higher for the synthetic content (Figure 8.20(c) and (d)). Nevertheless, the AUC values below 0.7 suggest that none of the metrics can be used as a reliable detector of significantly different pairs.

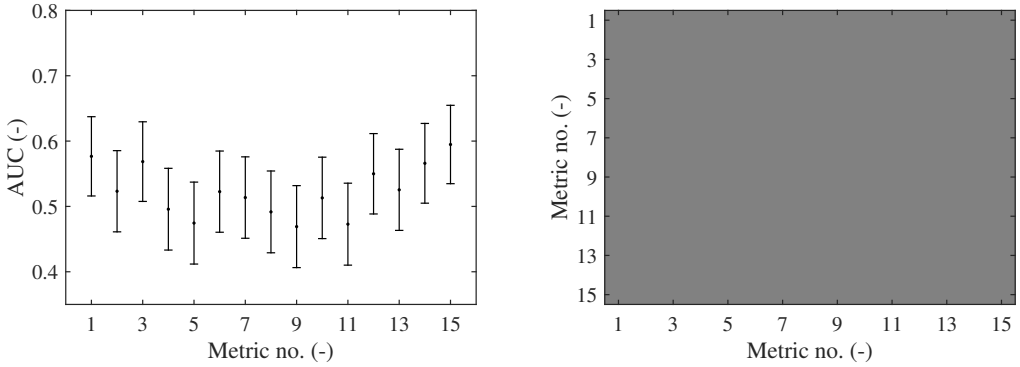
8.5.2 Results of Better vs. Worse ROC Analysis

This section tests the metrics' ability to recognize the better image in statistically significant pair. The results of ROC analysis are processed in the same way as in the previous section and are depicted in Figure 8.21.

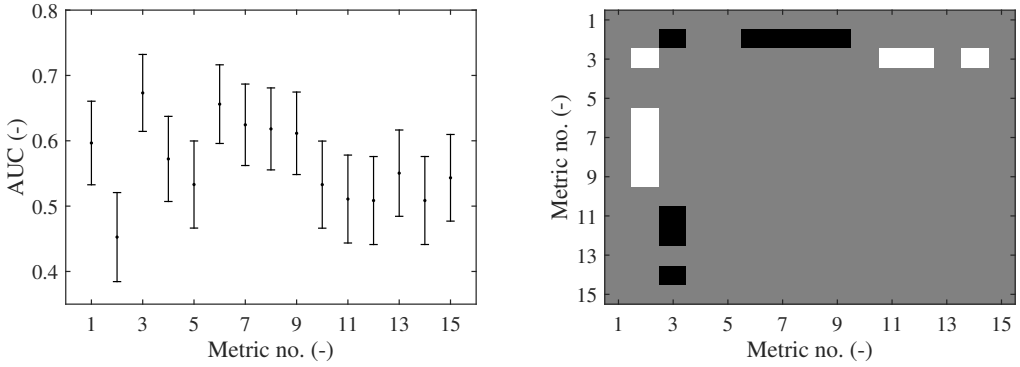
In this case, AUC values lower than 0.5 mean that the metric is systematically assigning higher score to the images with lower quality (which is again in contradiction with the real purpose of an objective method). It can also be seen that the natural content is very challenging for the tested metrics. The highest AUC value



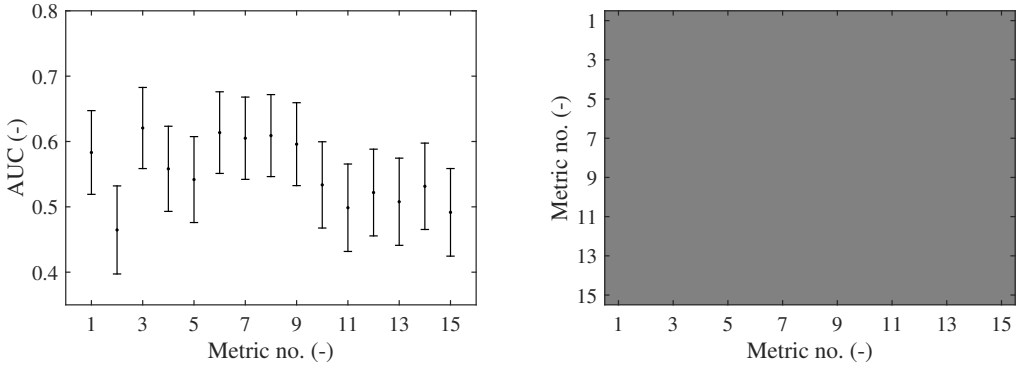
(a) AUC values with 95% CI and significance of differences for the scenario **1A**



(b) AUC values with 95% CI and significance of differences for the scenario **1B**

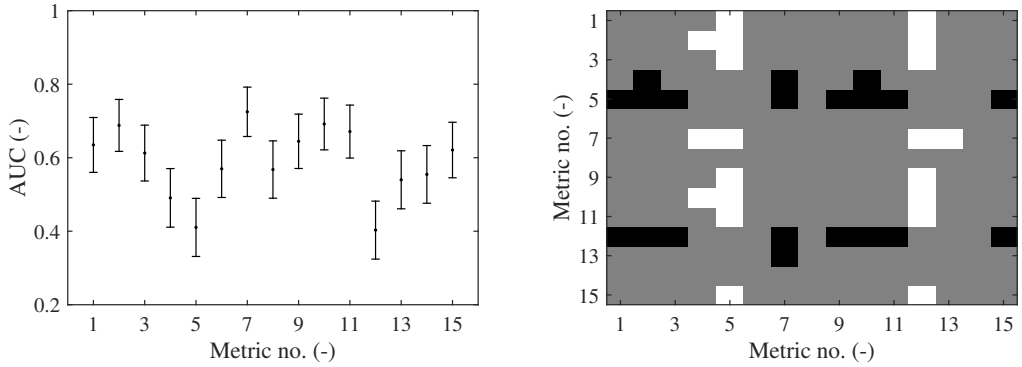


(c) AUC values with 95% CI and significance of differences for the scenario **2A**

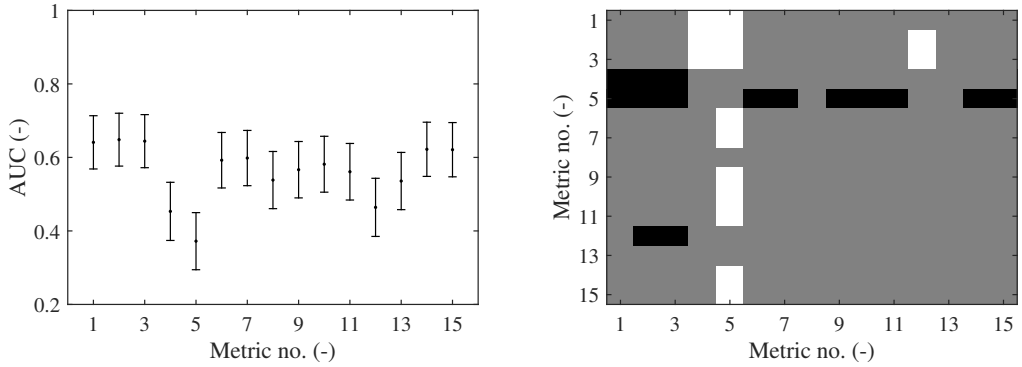


(d) AUC values with 95% CI and significance of differences for the scenario **2B**

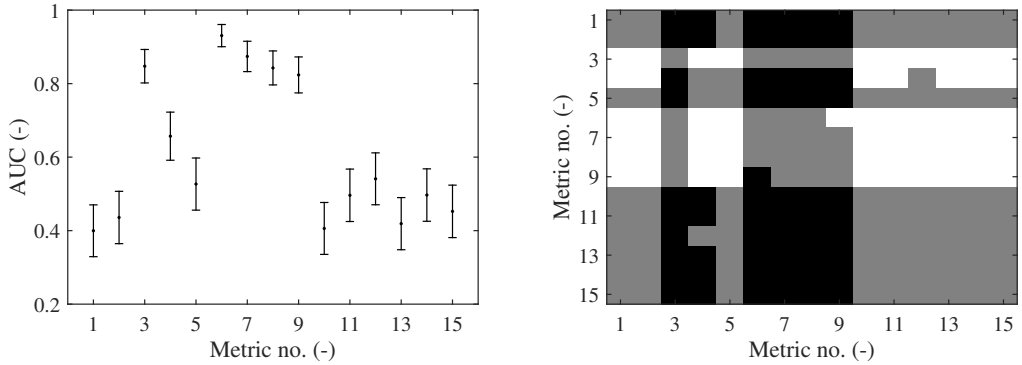
Figure 8.20: The results of the Different vs. Similar ROC Analysis for each scenario (see Table 8.3).



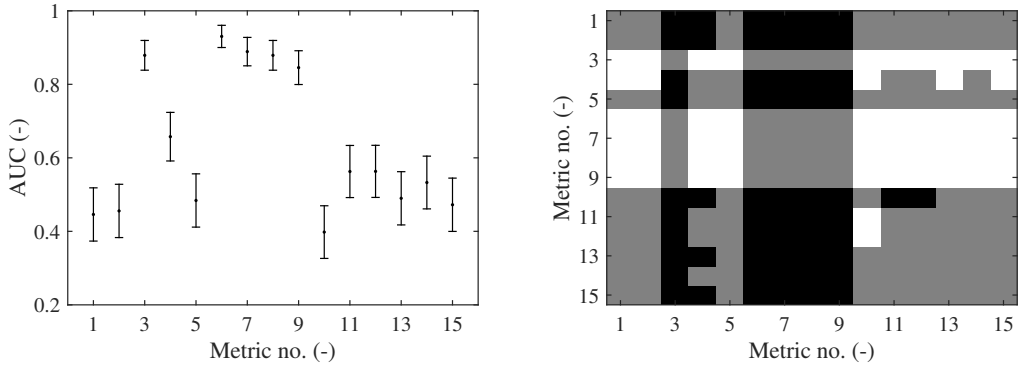
(a) AUC values with 95% CI and significance of differences for the scenario **1A**



(b) AUC values with 95% CI and significance of differences for the scenario **1B**



(c) AUC values with 95% CI and significance of differences for the scenario **2A**



(d) AUC values with 95% CI and significance of differences for the scenario **2B**

Figure 8.21: The results of the Better vs. Worse ROC Analysis for each scenario (see Table 8.3).

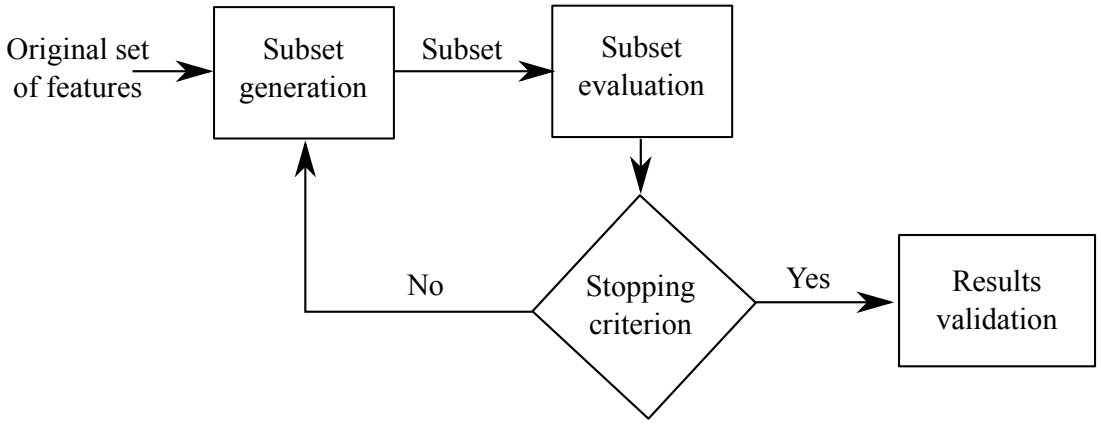


Figure 8.22: Framework of feature selection algorithms. Redrawn from [294].

in the setup **1A** (Figure 8.21 (a)) is achieved by $FSITM_r$ (#7). Nevertheless, the AUC values are around 0.7 which corresponds to the correct classification around 66% only. The worst performing methods are $DRIM_a$ (#5) and QAC (#15). However, in case of $DRIM_a$, the misclassification means that observers tend to prefer cases when the perceived contrast is amplified as compared to the reference. This is in line with findings from image post-processing that human observers prefer enhanced content against the pure depiction of the real world [7]. The scenario **1B** (Figure 8.21(b)) shows comparable performance of most of the metrics and the performance is even lower.

The synthetic content, as depicted in Figure 8.21(c) and (d), shows that the $DRIM_{all}$ (#6) provides the best assessment with very high AUC values in both scenarios (which is not surprising since there was no statistically significant difference in observers' evaluations). $DRIM_l$ (#3), and FSITM for all channels perform well as well. Such a good performance of the metrics quantifying only the differences in perceived contrast or low level features also supports the hypothesis that in case of CG content (that is not too photo-realistic), the observers' evaluation is shifted towards low level characteristics of an image, since the naturalness is no longer a factor. The poor performance of the measures where the naturalness quantification is involved (such as TMQI and TMQI-II) also supports this claim. The performance of no-reference measures is generally poor. The specific nature of the distortions introduced by tone-mapping is probably too complex and differs too much from the assumptions of such methods. It seems that in terms of correct recognition of better image in case of non photo-realistic CG content, the quality can be reliably estimated by DRIM or FSITM metrics.

8.6 Feature Selection for Quality Assessment of Tone-Mapped Images

The analyses in Section 8.5 showed a very poor performance of the available objective metrics when evaluating natural content. The purpose of this section is to identify the most relevant quality related features and combine them in a meaningful way, creating a reliable fusion metric.

The existing strategies for feature selection are thoroughly described in [294]. Such techniques were practically used in the area of image quality assessment, for example, by Nuutinen et al. [295]. Their goal was to identify the quality related features in order to reliably compare the images coming from different cameras. The purpose of the feature selection strategies is to select a subset from the full set of features that will be able to model the data most accurately.

The classical framework of feature selection algorithms is depicted in Figure 8.22. According to the ways of generating subsets, the strategies can be roughly divided into three approaches:

- *Complete search* algorithms,

- *sequential search* algorithms, and
- *random search* algorithms.

The *complete search* strategies guarantee the selection of the globally optimal subset (with respect to the given criterion). Even though some methods not requiring testing all possible combinations, such as branch and bound [296] or beam [297] algorithms, have been introduced, these strategies are mostly very time consuming and not practical for larger feature sets.

The *sequential search* algorithms either start with one feature and add one at the time, or begin with the full set and sequentially remove one feature. Which type to choose depends on the desired size of the subset. If smaller subset is more convenient, the adding of features is more appropriate [295]. The possible strategies include sequential forward selection, sequential backward selection, or bidirectional selection [298]. The disadvantage is that the approach cannot guarantee finding of the global optimum and only focuses on a certain path.

The last group is created by the *random search* algorithms. The initial subset is generated randomly. Further on, a sequential approach, adjusted to include randomness such as random-start-hill-climbing or simulated annealing [297], can be used. Alternatively, new set can also be generated completely randomly again. Such approach is also known as Las Vegas algorithm [299]. The randomness helps avoiding following a single path.

In terms of evaluating the currently selected subset, following approaches can be identified:

- *Filter* models,
- *wrapper* models, and
- *hybrid* models.

The *filter* models use a criterion independent on any mining algorithms. These criteria can be based e.g. on distance [300], information [301], dependency [302], or consistency [298]. The *wrapper* models use a performance of a mining algorithm used on the selected subset to evaluate the subset. The *wrapper* models are generally more effective but also more computationally demanding. The *hybrid* models combine the two approaches. The often used stopping criteria include:

- All the possibilities have been tested,
- a maximum number of iterations has been reached,
- a maximum number of features have been reached,
- adding more features does not provide any improvement,
- a sufficiently good subset has been found, etc.

The following sections will describe the selection procedure used for identification of the most relevant features for quality assessment of tone-mapped images, their combination into the fusion metric, and its performance verification.

8.6.1 Selection of the Most Relevant Features

Čadík et al. [10] identified the perceptual attributes contributing to overall quality perception for natural content. These are *brightness*, *contrast*, *details*, *color*, and *artifacts*. They also argue that the perceived contrast depends on lightness, chroma, and sharpness. The resulting fusion metric should, therefore, combine these perceptual attributes.

The Las Vegas algorithm [299] was used to provide an initial insight into the combinations behavior. After that the resulting data have been carefully investigated and the subset for the sequential algorithm has been selected.

The full set consisted of 60 metrics' scores. The goal was to meaningfully combine as small number of features as possible while providing good performance. In classical feature selection scenario, all the features are treated as independent entities. However, since size of the subset is supposed to be small, it is not desirable to combine multiple metrics of the same type together, i.e. the subsets including two different contrast metrics can be omitted. Therefore, several groups of criteria have been created. This enabled to focus on the more meaningful combinations only. Every subset in the Las Vegas algorithm was generated by randomly selecting the groups from which the criteria will be taken. Each group was assigned with different possibilities. Some groups enabled selection of only one criterion from the group, some enabled random selection of several metrics, and some required usage of all the criteria in the group.

Group 1 The first group consisted of full-reference criteria comparing the contrast and structure of the HDR reference and the tone-mapped version. The metrics in the group were: structural similarity from TMQI [148], structural similarity from TMQI-II [152], contrast reversal from Section 8.3.2, contrast loss DRIM_l [146], contrast reversal DRIM_r [146], and contrast amplification DRIM_a [146]. If the Group 1 was used, random amount of metrics have been randomly selected for the subset.

Group 2 The metrics in the second group estimated full-reference feature similarity in each channel – FSITM_r [153], FSITM_g [153], and FSITM_b [153]. If this group was selected, all three of the metrics were used since it had been found less meaningful to estimate the feature similarity in one or two channels only.

Group 3 The third group included metrics of contrast (see Section 3.2.3). These were GCF [106], Weber contrast [70], Michelson contrast [71], SDME [74], and RMS contrast [73]. Here, only one metric at a time could have been chosen.

Group 4 The fourth group comprised colorfulness metrics, as described in Section 3.2.4, i.e. CIQI [76], CQE1 colorfulness [73], CQE2 colorfulness [73], and color saturation. Again, only one metric from this group could have been in a subset.

Group 5 The fifth group was created by sharpness/blur metrics (refer to Section 3.2.2). The members were Variance [57], Frequency Threshold [58], Gradient [59], Laplacian [59], Autocorrelation metric [59], Histogram Frequency [60], Kurtosis [61], Marziliano [62], HP [63], Kurtosis of Wavelet Coefficients [64], Riemannian Tensor [65], JNBM [66], CPBD [67], S1 [68], S2 [68], S3 [68] with improved pooling (S_III) according to Section 7.5, FISH [69], and its block based variant FISH_bb [69]. Only one blur/sharpness metric was allowed to be in a subset.

Group 6 The sixth group was formed by Aesthetics metrics proposed by Aydın et al. [5] and described in Section 3.2.5. Any number of them could have been randomly selected into the subset.

Group 7 The seventh group consisted of saliency models. These were included since more details should provide more salient regions. The scores were therefore created from the saliency maps by averaging assuming that more salient regions will result in higher average saliency. The included models were Frequency-tuned saliency model [303], Graph based model [304], Itti-Koch model [305], Spectral residual model [306], Incremental coding length saliency model [307], and SUN [308]. Only one per subset could have been selected.

Group 8 The last group was formed by the metrics not belonging to any previous category. It included NIQE [84], CS [86], QAC [85], BIQI [78], BRISQUE [82], BLIINDS-II [81], Curvelet based metric [83], statistical naturalness from TMQI [148], statistical naturalness from TMQI-II [152], feature naturalness from Section 8.3.2, mean intensity from Section 8.3.2, percentage of under and over

exposed areas from Section 8.3.1, JPEG2000 metric [56], and JPEG metric [55]. Any number of these metrics could have been randomly selected to be included in any subset.

Optimization Function

In each iteration of the Las Vegas algorithm, the subset was randomly generated by selecting the groups that will contribute to the subset and, according to the group properties, the criteria from it. After selecting the subset, the optimization algorithm has been run to train the weights of the contribution of each metric in the subset. Only the linear combination of metrics has been allowed since it ensures the transparent insight into the contributions of each criterion. The selected model in i-th iteration was defined as

$$SCORE_i = \tau_1 \times metric_1 + \tau_2 \times metric_2 + \ldots + \tau_{k_i} \times metric_{k_i}, \tag{8.31}$$

where τ are the weights for each $metric$ in the subset of the size k_i.

Having a single dimensional ground truth data, this would be a simple regression problem. However, given the evaluation criterion described below, the parameters τ needed to be found using an optimization procedure. Considering the high dimensionality of the parameters space, the direct search methods, such as Nelder Mead downhill simplex [278], were not found suitable since they are too prone to ending up in the local minima and are very dependent on the starting point. The `ga()` function in MATLAB using a genetic optimization algorithm [309–311] has also been tested. Nevertheless, the best results were obtained using MATLAB's `patternsearch()` function [312, 313].

Evaluation Criterion

The evaluation criterion that has been used as a base for optimization, as well as for evaluating the subsets, is based upon the two analyses described in Section 5.2. The $SCORE_i$ resulting from the particular parameters selection was evaluated against the ground truth obtained from the extensive subjective study (see Section 8.4). Note that only the data from the scenario **1A** (see Table 8.3) were considered. This was found more meaningful regarding the presence of the full reference criteria in the feature set. The resulting fusion should, therefore, be able to model observers' preferences with respect to the original scene.

The resulting AUC value from the Different vs. Similar analysis $AUC_{\text{DS},i}$ and Correct Classification of Better vs. Worse stimulus $C_{0,i}$ have been averaged in order to provide an overall performance value $PERF_i$, thus

$$PERF_i = \frac{AUC_{\text{DS},i} + C_{0,i}}{2}. \tag{8.32}$$

This should ideally lead to optimizing the performance with respect to both of the performance analyses. The parameters resulting in the best performance for the i-th subset have been saved together with the overall optimized performance $PERF_{\text{opt},i}$. Note that the full database has been used so the optimal value showed how well could the combination of the features in the subset perform if trained and tested on the same data. This, by all means, leads to over-fitting. Nevertheless, for the purpose of feature selection, it provides the information about which features can be used to model the data in the best way. To provide a general combination, different training procedure has been adopted. This will be discussed in Section 8.6.2.

Final Selection

The Las Vegas algorithm has been run for 2,000 iterations. The optimal performance values with respect to the size of the subset are depicted in Figure 8.23. An interesting outcome can be that even with high number of features, the overall performance does not get over 0.83. This is probably caused by the challenging nature of the two performance analyses together. Combining the features to work universally with respect to both of the analyses would probably require more sophisticated non-linear model. Nevertheless, the goal was not to perfectly model the data. To have an insight on which features are the most relevant for the

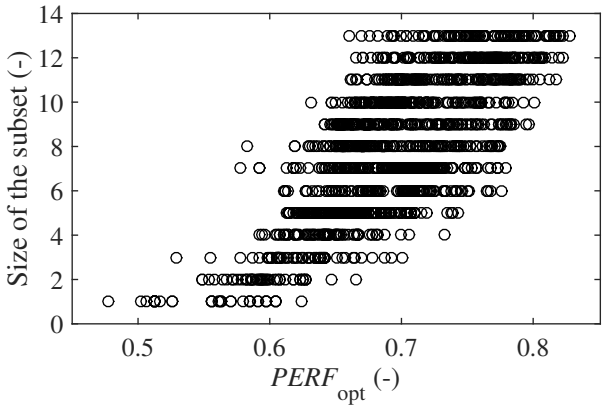


Figure 8.23: The optimal performance values with respect to the size of the subset after 2,000 iterations of the Las Vegas algorithm.

modelling, a closer look has been taken on the best performing subsets for each subset size (i.e. the subsets having the performance values $PERF_{\text{opt}}$ in the Figure 8.23 in the most right in each row).

Interestingly, four features were present in each of the best performing subsets from the size 4 to 13. These metrics were structural similarity from TMQI-II, FSITM_r, FSITM_g, and FSITM_b. Therefore, they have been selected as an initial subset for a forward selection sequential algorithm. The forward selection has been chosen since only the small number of features has been assumed to be added without jeopardizing the generality. The subsets training and evaluation remained the same as in case of Las Vegas strategy. The stopping criterion has been set to stop the algorithm when adding another feature will not result in increase of $PERF_{\text{opt}}$ and simultaneously increase of both AUC_{DS} and C_0. The rational was to avoid over-fitting the data by focusing more on one aspect of the performance than the other.

These conditions have lead to adding only one feature – feature naturalness as described in Section 8.3.2. Such subset has been found convenient for several reasons. Firstly, the size of the subset (five features) is small enough to avoid over-fitting and thus potentially provide more generality. Secondly, the combination of these features makes sense also with respect to the quality related perceptual attributes as described by Čadík et al. [10]. Structural similarity measures the reproduction of *contrast* and structure which provides an information about *details* and *artifacts*. The feature naturalness determines if the combination of *brightness*, *contrast*, and *color* of the tone-mapped version is plausible for a natural looking image. The FSITM quantifies feature similarity, therefore, it should capture changes in *details* reproduction and detect the presence of *artifacts*. Moreover, since all three channels are included, the *color* artifacts should also be found.

8.6.2 Training of the Parameters

The previous sections identified the most relevant features for quality assessment of tone-mapped images on the earlier developed dataset. This section describes the training of the parameters to provide higher generality of the proposed approach. Typically, the dataset is being repeatedly randomly divided into training and testing parts and median results are provided. However, this still tests the performance within one dataset. It has therefore been decided to train the parameters of the combination on completely different database.

For this purposes, the dataset developed by Yeganeh and Wang [148] with TMQI has been chosen. It is formed by 15 source images and 8 TMOs. Since the database contains only within content ranks and, therefore, does not allow for any other analysis, simple maximization of average KROCC has been adopted. This has been done using `patternsearch` method. The resulting Fusion Metric for Tone-

Metric	*Content no.*															*Average*	*Hit Count*	*Min*
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15			
FMTMI	0.93	0.79	**0.64**	0.71	**0.71**	0.57	**0.79**	**0.57**	0.79	0.64	0.86	**0.86**	**0.62**	**0.71**	0.93	**0.74**	**7**	**0.57**
TMQI	0.79	0.64	**0.64**	0.71	0.64	**0.93**	0.57	**0.57**	0.57	**0.86**	0.71	0.57	0.55	0.64	0.86	0.68	4	0.55
TMQI-II	0.79	0.29	0.57	0.50	0.50	**0.93**	0.71	0.50	0.71	0.79	0.71	0.43	**0.62**	0.57	0.79	0.63	2	0.29
FSITM$_r$	**1.00**	0.71	0.50	**0.79**	0.57	0.64	**0.79**	0.50	0.86	0.64	**0.93**	0.71	0.55	**0.71**	0.93	0.72	5	0.5
FSITM$_g$	0.93	**0.93**	0.50	0.71	0.57	0.36	0.64	**0.57**	0.79	0.57	0.86	0.57	0.55	0.64	0.93	0.67	2	0.36
FSITM$_b$	0.71	0.71	0.29	0.71	0.64	0.29	0.29	0.29	**1.00**	0.79	0.71	0.71	-0.25	0.50	**1.00**	0.56	2	-0.25

Table 8.6: KROCC of the metrics for the dataset developed by Yeganeh and Wang [148]. Hit count is the number of contents for which the metric performed the best.

Metric	*Content no.*															*Average*	*Hit Count*	*Min*
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15			
FMTMI	0.98	0.86	0.79	0.86	**0.83**	0.69	0.88	0.69	0.90	0.79	0.93	**0.95**	0.68	**0.88**	0.98	**0.85**	3	**0.68**
TMQI	0.90	0.79	**0.81**	0.88	0.74	**0.98**	0.69	**0.71**	0.69	**0.93**	0.88	0.71	0.68	0.74	0.95	0.81	4	**0.68**
TMQI-II	0.90	0.50	0.69	0.69	0.67	**0.98**	0.83	0.67	0.81	0.90	0.83	0.60	**0.79**	0.76	0.90	0.77	2	0.5
FSITM$_r$	**1.00**	0.76	0.64	**0.90**	0.71	0.74	**0.90**	0.57	0.93	0.79	**0.98**	0.86	0.65	**0.88**	0.98	0.82	**5**	0.57
FSITM$_g$	0.98	**0.98**	0.69	0.86	0.71	0.55	0.79	0.62	0.90	0.71	0.93	0.76	0.74	0.79	0.98	0.80	1	0.55
FSITM$_b$	0.81	0.81	0.60	0.86	0.79	0.43	0.43	0.45	**1.00**	0.86	0.86	0.86	-0.18	0.71	**1.00**	0.68	2	-0.18

Table 8.7: SROCC of the metrics for the dataset developed by Yeganeh and Wang [148]. Hit count is the number of contents for which the metric performed the best.

Mapped Images (FMTMI) is, thus, defined as

$$FMTMI = \tau_1 \times SS\text{-}II + \tau_2 \times FN + \tau_3 \times FSITM_{\mathrm{r}} + \tau_4 \times FSITM_{\mathrm{g}} + \tau_5 \times FSITM_{\mathrm{b}}, \qquad (8.33)$$

where the parameters' values are determined by maximizing average KROCC on the Yeganeh's dataset. The final values are $\tau_1 = 0.2129$, $\tau_2 = 0.0443$, $\tau_3 = 1$, $\tau_4 = 0.0621$, and $\tau_5 = 0.0931$. The highest weight has been assigned to the FSITM$_{\mathrm{r}}$, the smallest to the FN. Nevertheless, all the metrics provide a valuable contribution.

8.6.3 Performance Verification

The performance of the proposed fusion metric has been evaluated and compared to other existing criteria on three available datasets. The performance comparison procedures for Yeganeh's [148] and Čadík's database were using rank order correlation coefficients (see Section 4.3) only, since the nature of the subjective scores provided with the datasets did not allow for more sophisticated statistical analysis. Moreover, the same way has already been adopted when proposing new criteria [148, 153] and, therefore, allows for the direct comparison under the same conditions. The performance on the dataset introduced in this thesis is done by the novel ROC based framework from Section 5.2.

Note that since the parameters have been trained on the Yeganeh's database, the comparison is not completely fair and is included for the purpose of completeness only. The resulting KROCC and SROCC per content, together with an average value, number of times when the metric was among the best performing methods (denoted as hit count), and minimum value, can be found in Tables 8.6 and 8.7, respectively. In terms of KROCC, the proposed metric provides the highest average coefficient value, the highest hit count, and the highest minimum value. With respect to the SROCC, it still reaches the highest average coefficient value, but results in lower hit count than FSITM$_{\mathrm{r}}$ and TMQI, and has the same minimum value as TMQI. Overall, the performance of the proposed metric is satisfactory and can be considered superior over the state-of-the-art metrics.

The dataset developed by Čadík et al. [10] contains 3 source images processed with 14 TMOs. The values of KROCC and SROCC in the same format as previously are in Tables 8.8 and 8.9, respectively. In case of KROCC, the proposed method ranks the highest on average, results in the same hit count as TMQI-II and FSITM$_{\mathrm{g}}$, and reaches the highest minimum value together with TMQI-II. It should be noted that TMQI-II and FSITM$_{\mathrm{g}}$ did not perform well on the Yeganeh's dataset. This suggests higher universality of the proposed FMTMI. Moreover, it reaches the highest average SROCC value, as well as the highest hit count and minimum SROCC. Its performance can, therefore, again be considered superior over the other criteria.

Metric	*Content no.* 1	2	3	*Average*	*Hit Count*	*Min*
FMTMI	0.64	0.74	**0.77**	**0.72**	**1**	**0.64**
TMQI	0.56	0.77	0.62	0.65	0	0.56
TMQI-II	**0.67**	0.64	0.69	0.67	**1**	**0.64**
FSITM$_r$	0.44	0.67	0.59	0.56	0	0.44
FSITM$_g$	0.44	**0.87**	0.62	0.64	**1**	0.44
FSITM$_b$	0.41	0.56	0.74	0.57	0	0.44

Table 8.8: KROCC of the metrics for the dataset developed by Čadík et al. [10]. Hit count is the number of contents for which the metric performed the best.

Metric	*Content no.* 1	2	3	*Average*	*Hit Count*	*Min*
FMTMI	**0.80**	0.89	**0.90**	**0.86**	**2**	**0.8**
TMQI	0.71	0.91	0.77	0.80	0	0.71
TMQI-II	0.78	0.82	0.86	0.82	0	0.78
FSITM$_r$	0.64	0.71	0.74	0.70	0	0.64
FSITM$_g$	0.64	**0.92**	0.77	0.78	1	0.64
FSITM$_b$	0.64	0.77	0.86	0.76	0	0.64

Table 8.9: SROCC of the metrics for the dataset developed by Čadík et al. [10]. Hit count is the number of contents for which the metric performed the best.

Lastly, the metrics are compared on the dataset developed in this thesis (scenario **1A**, see Table 8.3). It enables the performance evaluation using the proposed ROC based analyses. The result of the Different vs. Similar ROC Analysis is depicted in Figure 8.24. It can be seen that the proposed FMTMI reaches the highest AUC value. However, the performance is statistically better only with respect to TMQI-II and FSITM$_g$. The Figure 8.25 shows the result of the Better vs. Worse ROC Analysis. Here, the proposed metric significantly outperforms all the other metrics while reaching the AUC value of 0.9243. The correct classification of the better images from pairs is 84% which is a significant improvement in reliability over the other criteria.

Considering the performance on all three publicly available databases, the proposed FMTMI showed good generality and proved to be the most reliable criterion for objective quality assessment of natural tone-mapped images so far.

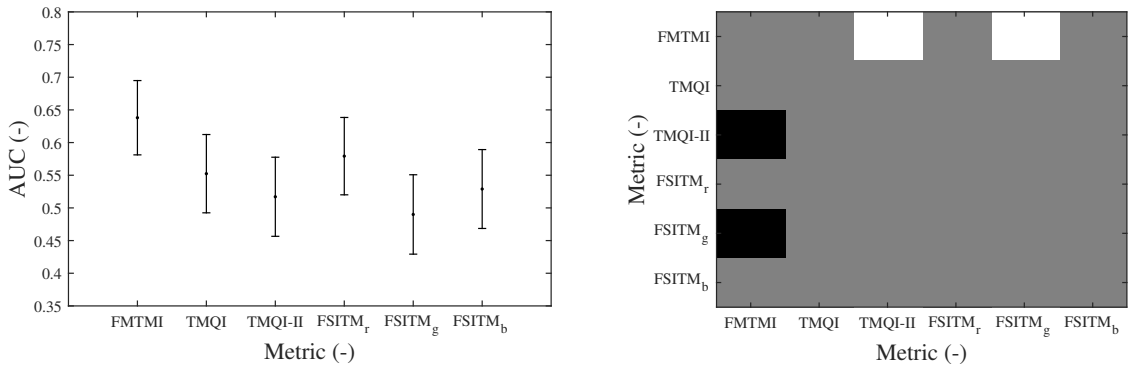


Figure 8.24: AUC values with 95% CI and significance of differences for the Different vs. Similar ROC Analysis in the scenario **1A** from the developed dataset (see Table 8.3).

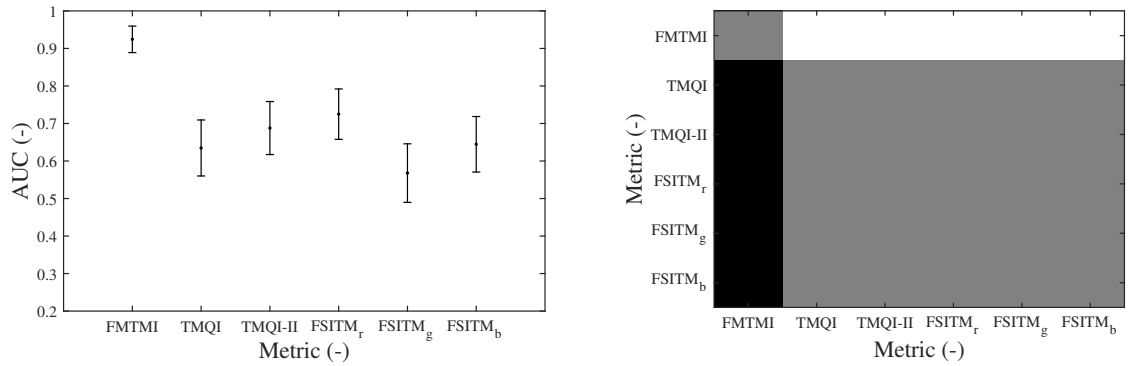


Figure 8.25: AUC values with 95% CI and significance of differences for the Better vs. Worse ROC Analysis in the scenario **1A** from the developed dataset (see Table 8.3).

9

Conclusion

In this thesis, the quality assessment paradigm is revised in order to tackle the challenges brought by image post-processing algorithms. Unlike in classical quality evaluation scenarios, the quality can no longer be quantified as a fidelity of the processed version of the stimulus to the original one. Considering the ability of post-processing techniques to adjust the aesthetic properties of the stimuli leading to increase in their subjective appeal, the notion of quality needs to be extended to include the aesthetic attributes as well. Moreover, the stimulus of the best possible quality is not known. The thesis identifies both subjective and objective methods suitable for quality assessment of two relevant groups of post-processing algorithms – enhancement techniques, represented by image sharpening, and tone-mapping of high dynamic range (HDR) images. In this way, the whole framework for objective metrics testing and design, tailored to these specific applications, is developed. It includes the guidelines for preparing ground truth datasets and a novel performance comparison methodology.

9.1 Summary of Contributions

The purpose of this section is to clearly summarize the contributions of the thesis and to explain their importance for the scientific community. Firstly, a thorough introduction into subjective experimental methodologies, including statistical methods for processing of the results, has been provided, followed by the description of the popular objective quality metrics. State-of-the-art procedures for evaluating the performance of the objective criteria, together with the methods for determining statistical significance of differences, have been introduced as well. These parts of the thesis represent an overview of the current state in the field of quality assessment and provide the basis upon which the following main contributions are built.

9.1.1 Novel Method for Evaluating the Performance of Objective Metrics

Current methods used for evaluating the performance of objective criteria suffer from at least one of the following drawbacks: They do not consider the uncertainty of the opinion scores, they require mapping to the common scale, they are applicable to the MOS-like scenarios only, and/or they do not allow for simple combination of multiple datasets. To overcome these disadvantages, a novel methodology has been proposed. It has been designed to test metrics' abilities regarding two practically oriented scenarios: (a) how well can the metric distinguish between qualitatively different and similar pairs of stimuli, and (b) how well can the metric recognize which of the stimuli in the qualitatively different pair is of better quality. The performance is quantified by two separate receiver operating characteristic (ROC) analyses enabling simple

numerical comparisons of the individual metrics including the means to calculate statistical significance of differences in performances. The pairwise approach ensures universality over various data formats. Moreover, the results from different subjective experiments can be easily put together regardless their initial format, enabling simple combination of multiple datasets without the necessity of any mapping. Last but not least, considering the statistical significance of the subjective scores makes the method suitable for the post-processing scenarios where the observers' personal taste has much more impact than in the standard applications.

9.1.2 Subjective Study on Sharpened Images Including Over-Sharpening

Despite its importance for understanding the quality perception and automatic enhancement, the phenomenon of over-enhancement is not very well studied. It is mainly due to its challenging nature caused by the variability of personal opinions. This thesis attempts to tackle this issue on the example of image sharpening. The dataset including blurred, sharpened, and over-sharpened content has been carefully prepared using four different sharpening techniques in two setups. Adaptive Square Design Paired Comparison (ASDPC) methodology has been found the most suitable for the evaluation given the simplicity of the task for the observers, robustness against participants' errors, and, most importantly, its discriminatory power which enables to capture the trends in the overall quality perception with a reasonable number of observers. In combination with the novel method for evaluating the performance of objective metrics, it creates a robust quality assessment framework capable to gain valuable data from the subjective studies and use them efficiently for objective criteria design and comparison.

The subjective study itself, although conducted on the lower number of source contents, provides several valuable insights. Firstly, it confirms that human observers prefer the images to be enhanced with respect to the original but the optimal amount of sharpening differs for each source content. Secondly, it also verifies the fact that HVS is very sensitive to the naturalness of human skin. The content including uncovered skin, especially in the close-up, typically allows for less enhancement. The best enhancement has been achieved by the most sophisticated enhanced unsharp mask method. Most importantly, the results of the study can be used to test the abilities of objective criteria in the context of sharpening and over-sharpening by different techniques.

9.1.3 Improved Pooling Strategy for Sharpness Metrics

The best performance on the developed sharpened images database has been reached by two local sharpness measures – S3 [68] and FISH_{bb} [69]. It is generally accepted that the perceived sharpness of the stimulus is affected by the sharpest regions only. The obtained sharpness maps are, therefore, pooled exclusively from these areas. However, it has been observed that when the content is over-sharpened, there is an increase in sharpness in the regions with low sharpness. This motivated the design of several pooling strategies considering the areas with low sharpness as well. The following performance comparison has shown that including these areas in the sharpness pooling can, indeed, significantly improve the reliability of the metrics. It also maintains the availability to correctly recognize blurring and does not bring any additional computational requirements.

Overall, the best metric in the given context is S3 with the proposed pooling based on the dynamic range of the sharpness map, robust against outlier values, with subtracted minimum (see equation (7.23)). It can be considered as the best currently available method for objective quality assessment of sharpness when the over-sharpening effect is also involved. In case of FISH_{bb}, the strategy based on the classical dynamic range – equation (7.21) – provides the best overall performance (significantly better than the original strategy). Nevertheless, it reaches slightly lower performance than the original metric for one of the contents.

9.1.4 Novel Method of Using Full Reference Metrics on Enhanced Images

Considering the assumption of most of the full reference quality metrics that the original image is of the best possible quality, their usage in the post-processing scenarios is not meaningful. Although Vu et al. [142] came up with the reversed strategy (refer to Figure 6.1), it has only been designed for the positive influence of the enhancement, i.e. it does not consider the possibility of over-enhancement. This has been demonstrated on the dataset introduced in this thesis. Therefore, a novel method which includes the assumption of over-enhancement has been proposed.

It uses an intentionally over-enhanced anchor image as a "negative reference" and calculates the score for the enhanced image from its similarity to the original and to this anchor. Note that the main motivation of the method is automatic image enhancement and, as such, it is designed to be reliable within the enhancement technique. Good performance across different techniques is not needed. The proposed method has been tested on image sharpening using unsharp mask. The most suitable strategy for the anchor image creation, together with other parameters, has been found from the subjective study. Another experiment has been conducted in order to verify the performance. The results suggest that the method can be successfully used for automatic image enhancement.

9.1.5 Novel Method for TMOs Parameters Selection in Security Applications

The absolute majority of tone-mapping operators (TMOs) enable setting of several parameters in order to adjust to the scene. Most of the time, the parameters are either left in their default state or set manually to maximize the quality as subjectively perceived by the user. The first way may not be able to fully exploit the potential of the given TMO, the second one is dependent on the user's abilities and personal taste, and can also be less fair when comparing different TMOs. In this thesis, an attempt is made to come up with an objective way to select the TMOs parameters that will provide a result close to the optimal quality while maintaining reproducibility and fairness.

The requirements put on the operators are dependent on the application. Given the increased number of HDR surveillance cameras, one of the hot application areas for tone-mapping is security. The surveillance systems can be expected to adapt HDR solutions in terms of capture but it is less probable that HDR displays will be used in the near future. Thus, a novel universal and computationally simple parameters selection method has been designed to optimize the tone-mapping process. It is based on minimizing the under and over exposed areas which should enable reproduction of as many details as possible for the given TMO. The method has been demonstrated to clearly outperform the default setting of the parameters on three global and one local TMO.

9.1.6 Novel Method for TMOs Parameters Selection in Multimedia Applications

In security surveillance, the focus is purely on the reproduction of details. However, in multimedia applications, the overall aesthetic quality is of interest which brings a dependency on more factors. Above all, human observers require the images to look natural. The amount of details comes after this. Good tone-mapped image should, therefore, balance the two requirements.

The proposed parameters optimization procedure attempts to maximize the naturalness while simultaneously keeping the amount of reproduced details sufficiently high. To quantify the two entities, novel criteria have been proposed. The maintaining of details is measured by detecting the reversal of contrast between HDR original and a tone-mapped version. The metric is computationally simple and, thus, suitable for the optimization. The naturalness is considered to be a product of three features – overall intensity, contrast, and colorfulness. The distribution of this product for 5,000 high quality natural images has been determined and approximated with Rayleigh distribution. The probability that the product of these features for the result of the particular TMO parameters settings belongs to this distribution quantifies its naturalness. The advantages of the proposed approach have been demonstrated on four TMOs and it has also been used for preparing the database for quantitative subjective experiment.

9.1.7 Subjective Study on Tone-Mapped Natural and Computer Generated Images

The main purpose of the subjective experiment conducted on the tone-mapped images has been to obtain a representative and challenging dataset for objective metrics training and testing. Since apart from the digital photography, an important application of HDR and tone-mapping is also computer graphics, the content includes both natural and synthetic images. Two objective content selection procedures have been employed in order to preselect from the pool of more that 150 scenes the source images that will be challenging for TMOs. The final selection consists of 10 natural and 10 computer generated (CG) images. Four TMOs, each with two different sets of parameters, accompanied by one more TMO, without the possibility to tune its parameters, have been chosen in order to provide both inter and intra TMO diversity. The parameters have been optimized using the proposed tuning method for multimedia content and by maximizing TMQI [148]. In the cases where the two methods lead to similar results, one of the settings has been manually adjusted to increase variability. All this effort has been made to come up with noticeably different images.

For the reasons similar to the subjective experiment on sharpened images, the ASDPC methodology has been adopted. However, two experimental setups have been considered – scenario with the reference displayed on the HDR screen between the two LDR displays and the no reference scenario with the two LDR displays only. The purpose of the two setups has been to determine whether the presence of the reference can have an impact on the observers' preferences. Note that observers have been asked to select the aesthetically preferred version in both scenarios.

The differences between the experimental setups have been statistically analysed and the results have shown that the presence of the reference can influence the observers. However, this effect has only been found in case of natural images. CG images have been evaluated equivalently whether the reference was displayed or not. A possible explanation is that the reference can alter the perception of naturalness and provide other information, such as daily hour (e.g sunset vs. daylight), which can influence the participants' decisions. For example, if both of the images are of good quality and one of them is closer to the reference in naturalness while the other has more dazzling colors, the preferences can be changed by the display of the original. This phenomenon might be weakened by the absence of naturalness in the synthetic content. However, it is also possible that the CG content in the dataset have not included the cases which would trigger such effect. A specific study needs to be designed in order to determine if the preferences regarding the CG content cannot be changed by the presence of the reference image.

Fifteen applicable objective metrics have been tested on the dataset. It has been found that DRIM [146] is reliable when used on CG content. Considering that it measures the reproduction of contrast, it seems that the preferences regarding this type of content are mainly driven by the low-level features. Not surprisingly, the metrics working with estimating the naturalness (such as TMQI [148] or TMQI-II [152]) perform poorly. In case of natural images, none of the metrics performs particularly well. This identified the need to design a novel objective metric for natural tone-mapped images.

9.1.8 Novel Metric for Tone-Mapped Images Based on Fusion of Features

Given the poor performance of the tested metrics on the natural tone-mapped images, a novel criterion has been designed. Firstly, a feature selection procedure, based on the modified Las Vegas algorithm followed by forward selection sequential algorithm, has been adopted to identify the most relevant features. Five features have been selected – structural similarity from TMQI-II [152], feature naturalness proposed in this thesis, and FSITM [153] for all three channels. FMTMI is then formed by the linear combination of these features, trained on the dataset developed by Yeganeh and Wang [148]. The performance of the metric has been verified against all available datasets and has proven to be the most universal and reliable from the currently available criteria.

The further work could include investigating more complicated machine learning based approaches for combining the features. Since the two separate ways for reliable quality assessment of natural and CG images have been identified, an objective way for content classification needs to be found in order to design a universal metric independent on the content type.

9.2 List of Publications

In this section, the complete list of author's publications is provided. It also includes papers not directly connected to the topic of the thesis. These papers helped to broaden the author's experience and knowledge.

The contributions of the authors on all of the publications are equal.

9.2.1 Publications Related to the Topic of the Thesis

Publications in Impacted Journals

[I] **L. Krasula**, M. Narwaria , K. Fliegel, P. Le Callet, "Preference of Experience in Image Tone-Mapping: Dataset and Framework for Objective Measures Comparison," *IEEE Journal of Selected Topics in Signal Processing (under review)*.

[II] **L. Krasula**, P. Le Callet, K. Fliegel, M. Klíma, "Quality Assessment of Sharpened Images: Challenges, Methodology, and Objective Metrics," *IEEE Transactions on Image Processing (under review)*.

[III] P. Dostál, **L. Krasula**, M. Klíma, "Influence of High Level Features of HVS on Performance of FSIM," *Radioengineering*, vol. 22, no. 4, pp. 1048-1055, December 2013.

[IV] **L. Krasula**, M. Klíma, E. Rogard, E. Jeanblanc, "MATLAB-based Applications for Image Processing and Image Quality Assessment Part II: Experimental Results," *Radioengineering*, vol. 21, no. 1, pp.154-161, April 2012.

[V] **L. Krasula**, M. Klíma, E. Rogard, E. Jeanblanc, "MATLAB-based Applications for Image Processing and Image Quality Assessment Part I: Software Description," *Radioengineering*, vol. 20, no. 4, pp.1009-1015, December 2011.

[VI] M. Klíma, K. Fliegel, P. Páta, S. Vítek, M. Blažek, P. Dostál, **L. Krasula**, T. Kratochvíl, V. Říčný, M. Slanina, L. Polák, O. Kaller, L. Boleček, "DEIMOS – An Open Source Image Database," *Radioengineering*, vol. 20, no. 4, pp.1016-1023, December 2011.

Publications Excerpted by ISI

[VII] **L. Krasula**, K. Fliegel, P. Le Callet, M. Klíma, "On the accuracy of objective image and video quality models: New methodology for performance evaluation," *International Conference on Quality of Multimedia Experience (QoMEX)*, June 2016.

[VIII] P. Hanhart, **L. Krasula**, P. Le Callet, T. Ebrahimi, "How to Benchmark Objective Quality Metrics from Paired Comparison Data?," *International Conference on Quality of Multimedia Experience (QoMEX)*, June 2016.

[IX] J. Søgaard, **L. Krasula**, M. Shahid, D. Temel, K. Brunnström, M. Razaakh, "Applicability of Existing Objective Metrics of Perceptual Quality for Adaptive Video Streaming," *Electronic Imaging, Image Quality and System Performance XIII*, pp. 1-7(7), 2016.

[X] **L. Krasula**, M. Narwaria, K. Fliegel, P. Le Callet, "Rendering of HDR content on LDR displays: An objective approach," *Proc. SPIE 9599, Applications of Digital Image Processing XXXVIII*, September 2015.

[XI] **L. Krasula**, M. Narwaria, K. Fliegel, P. Le Callet, "Influence of HDR reference on observers preference in tone-mapped images evaluation," *7th International Workshop on Quality of Multimedia Experience (QoMEX)*, 2015.

[XII] **L. Krasula**, K. Fliegel, P. Le Callet, M. Klíma, "Objective Evaluation of Naturalness, Contrast, and Colorfulness of Tone-Mapped Images," P*roc. SPIE 9217, Applications of Digital Image Processing XXXVII*, 2014.

[XIII] K. Fliegel, **L. Krasula**, P. Páta, J. Myslík, J. Pecák, M. Jícha, "System for Objective Assessment of Image Differences in Digital Cinema," *Proc. SPIE 9217, Applications of Digital Image Processing XXXVI*, 2014.

[XIV] **L. Krasula**, M. Narwaria, P. Le Callet, "An Automated Approach for Tone Mapping Operator Parameter Adjustment in Security Applications," *Proc. SPIE 9138, Optics, Photonics, and Digital Technologies for Multimedia Applications III*, 2014.

[XV] **L. Krasula**, K. Fliegel, P. Le Callet, M. Klíma, "Using Full-Reference Image Quality Metrics for Automatic Image Sharpening," *Proc. SPIE 9138, Optics, Photonics, and Digital Technologies for Multimedia Applications III*, 2014.

[XVI] P. Dostál, **L. Krasula**, M. Klíma, "HLFSIM: Objective image quality metric based on ROI analysis," *IEEE International Carnahan Conference on Security Technology (ICCST)*, pp.367-375, October 2012.

Other Related Publications

[XVII] T. Vigier, **L. Krasula**, A. Milliat, M. Perreira Da Silva, P. Le Callet, "Performance and robustness of HDR objective quality metrics in the context of recent compression scenarios," *IEEE Digital Media Industry and Academic Forum*, 2016.

[XVIII] A. Pinheiro, K. Fliegel, P. Korshunov, **L. Krasula**, M. Bernardo, M. Pereira, T. Ebrahimi, "Performance Evaluation of the Emerging JPEG XT Image Compression Standard," *IEEE 16th International Workshop on Multimedia Signal Processing (MMSP)*, 2014.

9.2.2 Publications Unrelated to the Topic of the Thesis

Publications Excerpted by ISI

[XIX] K.Fliegel, P. Janout, J. Bednář, **L. Krasula**, S. Vítek, J. Švihlík, P. Páta, "Performance evaluation of image deconvolution techniques in space-variant astronomical imaging systems with nonlinearities," *Proc. SPIE 9599, Applications of Digital Image Processing XXXVIII*, September 2015.

[XX] S. Vítek, J. Švihlík, **L. Krasula**, K. Fliegel, P. Páta, "GPU accelerated processing of astronomical high frame-rate videosequences," *Proc. SPIE 9599, Applications of Digital Image Processing XXXVIII*, September 2015.

[XXI] S. Vítek, M. Klíma, **L. Krasula**, "Video compression technique impact on efficiency of person identification in CCTV systems," *International Carnahan Conference on Security Technology (ICCST)*, pp.1-5, October 2014.

[XXII] M. Řeřábek, L. Yuan, **L. Krasula**, P. Korshunov, K. Fliegel, T. Ebrahimi, "Evaluation of Privacy in High Dynamic Range Video Sequences," *Proc. SPIE 9217, Applications of Digital Image Processing XXXVII*, 2014.

[XXIII] S. Vítek, **L. Krasula**, M. Klíma, V. Hvězda, M. Herrera Martínez, "Influence of HEVC Compression on Task Detection in Security Video Sequences," *International Carnahan Conference on Security Technology (ICCST)*, 2013.

[XVI] P. Dostál, **L. Krasula**, M. Klíma, "Can state-of-the-art HVS based objective image quality criteria be used for image reconstruction techniques based on ROI analysis?," *Proc. of SPIE Optics, Photonics, and Digital Technologies for Multimedia Applications II*, vol. 8436, 2012.

Other Unrelated Publications

[XXV] **L. Krasula**, P. Dostál, M. Klíma, "Content dependent demosaicing algorithm implementation," *22nd International Conference Radioelektronika*, pp.1-4, 17-18 April 2012.

9.3 Activities

This section lists the activities that demonstrate author's involvement in the quality assessment domain and recognition by the scientific community.

9.3.1 Achievements

The following recognitions have been received.

- **Top 6 Paper at QoMEX'2016**
  L. Krasula, K. Fliegel, P. Le Callet, M. Klíma, "On the accuracy of objective image and video quality models: New methodology for performance evaluation," *International Conference on Quality of Multimedia Experience (QoMEX)*, 2016

- **Top 10% Paper at MMSP'2014**
  A. Pinheiro, K. Fliegel, P. Korshunov, L. Krasula, M. Bernardo, M. Pereira, T. Ebrahimi, "Performance Evaluation of the Emerging JPEG XT Image Compression Standard," *IEEE 16th International Workshop on Multimedia Signal Processing (MMSP)*, 2014.

- **Outstanding Diploma Thesis Award by Preciosa Foundation**
  L. Krasula, "Quality Assessment of Sharpened Images."

- **Best Poster Award at 1st Qualinet Summer School on Quality Assessment**
  L. Krasula, "Quality Assessment of Sharpened Images."

- **Best Group Study Award at 1st Qualinet Summer School on Quality Assessment**
  A. Barri, P. Dostál, L. Krasula, A. Kumcu, P. Le Callet, "Statistical analysis of observers' behavior."

9.3.2 Standardization

The author has been part of several projects towards standardization in image and video processing field. He was a member of the expert group responsible for evaluation of JPEG XT,[1] a novel standard designed for backward-compatible compression of still HDR images, and JPEG XS,[2] a low-latency lightweight image coding standard able to support increasing resolution (such as 8K) and frame rate in a cost effective manner. He is also a member of an Ad Hoc group within JPEG preparing the Guidelines for image coding systems evaluation, providing the best practices and recommendations regarding quality assessment of still images.

The author has also been active within VQEG's Visually Lossless Quality Analysis Group.[3] The purpose of the effort is to define the technical parameters describing the meaning of the term "visually lossless

[1] https://jpeg.org/jpegxt/index.html (retrieved on 30/08/2016)
[2] https://jpeg.org/jpegxs/index.html (retrieved on 30/08/2016)
[3] http://www.its.bldrdoc.gov/vqeg/projects/visually-lossless-quality-analysis.aspx (retrieved on 30/08/2016)

transformation" and to determine if low-impairment transformations are visually lossless. To achieve this goal, detailed subjective and objective test methodology should be created.

The author is also a member of QUALINET[4] – the European network of researchers working in the domain of QoE evaluation. His activities have mostly been connected to the HDR and Databases Task Forces.

9.3.3 Reviews

The author has served as a reviewer for the following journals and conferences. He also reviewed several master and bachelor theses and supervised one successfully defended master thesis.

Journals

- IEEE Transactions on Image Processing
- Signal Processing: Image Communication
- IEEE Transactions on Circuits and Systems for Video Technology
- Journal of Electronic Imaging
- Electronics Letters
- IET Image Processing
- EURASIP Journal on Image and Video Processing

Conferences

- IEEE International Workshop on Multimedia Signal Processing (MMSP) 2015
- IEEE International Workshop on Multimedia Signal Processing (MMSP) 2016
- IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2016
- ACM Multimedia Systems (MMSys) 2016

9.3.4 Internships

The author has been enrolled in a joint-degree Ph.D. study program at the University of Nantes, France and Czech Technical University in Prague, Czech Republic, supported by the French government scholarship. During the studies, he spent 18 months in the Image and Video Communications (IVC) laboratory in Nantes and 15 months in Multimedia Technology Group (MMTG) at CTU in Prague. Apart from these stays, two internships have been done.

- **Institution:** iMinds research institute, Department of Electronics and Informatics (ETRO) at the Vrije Universiteit Brussel (VUB), Belgium
  Supervisor: Professor Peter Schelkens
  Topic: Comparison of objective metrics for tone-mapped images
- **Institution:** School of Computer Science, Bangor University, UK
  Supervisor: Dr. Rafal Mantiuk
  Topic: Models of human visual system for image and video quality assessment

[4] http://www.qualinet.eu/ (retrieved on 30/08/2016)

9.3.5 Training Schools

The author participated in the following training schools.

2012 1st Qualinet Summer School on Quality Assessment, Ilmenau, Germany

2013 2nd Qualinet Summer School on Crowdsourcing and QoE Assessment, Patras, Greece

2014 3rd Qualinet Summer School on Statistical and Learning Methods for Reliable Assessment of QoE, Nantes, France

2015 Training School of COST Action IC1005 on HDR Imaging, Brno, Czech Republic

2015 3D ConTourNet Summer School, Lisbon, Portugal

9.3.6 Grants and Projects

The author received the grant SGS15/091/OHK3/1T/13 Optimization of Tone Mapping Operators Parameters for High Dynamic Range Images of the Student Grant Agency of CTU in Prague. He also participated in the projects COST CZ LD12018 Modelling and verification of methods for Quality of Experience (QoE) assessment in multimedia systems – MOVERIQ of the Ministry of Education, Youth and Sports of the Czech Republic, Grant No. P102/10/1320 Research and modelling of advanced methods of image quality evaluation and Grant No. 14-25251S Nonlinear imaging systems with spatially variant point spread function of the Czech Science Foundation, and SGS12/075/OHK3/1T/13 Region-of-Interest (ROI) importance with regards to the perceived image quality and implementation of an objective image quality metric based on ROI analysis of the Student Grant Agency of CTU in Prague.

Bibliography

[1] ITU-T Recommendation E.800. Definitions of terms related to quality of service, 2008. 17

[2] P. Le Callet, S. Möller, and A. Perkis. Qualinet white paper on definitions of quality of experience (2012). *European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003)*, Version 1.2, March 2013. 17

[3] D. Burnham. Internet encyclopedia of philosophy – Kant's aesthetics, 2011. [Online] http://www.iep.utm.edu/kantaest/. 18

[4] N. Zangwill. *The Stanford Encyclopedia of Philosophy*, chapter Aesthetic Judgment. 2014. (Fall 2014 Edition). 18

[5] T. Aydın, A. Smolic, and M. Gross. Automated aesthetic analysis of photographic images. *IEEE Transactions on Visualization and Computer Graphics*, 21(1):31–42, 2015. 18, 40, 54, 55, 107, 158, 164

[6] S. Winkler. *Digital Video Quality: Vision Models and Metrics*. John Wiley and Sons, Ltd, 2005. 18, 40

[7] B. Zhang, J. P. Allebach, and Z. Pizlo. An investigation of perceived sharpness and sharpness metrics. In *SPIE-IS&T Electronic Imaging: Image Quality and System Performance II*, volume 5668, 2005. 20, 54, 84, 92, 93, 100, 102, 104, 162

[8] M. A. Saad, P. Le Callet, and P. Corriveau. Blind image quality assessment: Unanswered questions and future directions in the light of consumers needs. *VQEG e-letter*, 1(2):62–66, December 2014. 20

[9] M. Narwaria, M. Perreira Da Silva, P. Le Callet, and R. Pepion. Single exposure vs tone mapped high dynamic range images: A study based on quality of experience. In *22nd European Signal Processing Conference (EUSIPCO)*, 2014. 22, 138, 142

[10] M. Čadík, M. Wimmer, L. Neumann, and A. Artusi. Evaluation of HDR tone mapping methods using essential perceptual attributes. *Computers & Graphics*, 32:330–349, 2008. 22, 86, 138, 145, 149, 163, 166, 167, 168, 201

[11] ITU-R Recommendation BT.500-13. Methodology for the subjective assessment of the quality of television pictures, January 2012. 25, 27, 28, 29, 30, 32, 83, 85, 102, 152

[12] ITU-T Reccomendation P.910. Subjective video quality assessment methods for multimedia applications, 2008. 25, 26, 27, 28, 32, 83, 85, 102, 104, 119, 152, 153

[13] J. A. Ferwerda. Psychophysics 101: How to run perception experiments in computer graphics. In *ACM SIGGRAPH 2008 Classes*, SIGGRAPH '08, pages 87:1–87:60, New York, NY, USA, 2008. ACM. 25, 32, 83, 85, 102, 152

[14] R. Mantiuk, A. Tomaszewska, and R. Mantiuk. Comparison of four subjective methods for image quality assessment. *Computer Graphics Forum*, 31(8):2478–2491, 2012. 25

[15] E. Bosc, R. Pépion, P. Le Callet, M. Köppel, P. Ndjiki-Nya, M. Pressigout, and L. Morin. Towards a new quality metric for 3-D synthesized view assessment. *IEEE Journal on Selected Topics in Signal Processing*, pages J–STSP–ETVC–00048–2011, 2011. 25

[16] T. Hoßfeld, M. Hirth, J. Redi, F. Mazza, P. Korschunov, B. Naderi, M. Seufert, B. Gardlo, S. Egger, and C. Keimel. Best practices and recommendations for crowdsourced QoE – lessons learned from the Qualinet task force crowdsourcing. *European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003 Qualinet)*, 2014. 26

[17] ITU-R Recommendation BT.2022. General viewing conditions for subjective assessment of quality of SDTV and HDTV television pictures on flat panel displays, August 2012. 26

[18] S.S. Stevens. On the psychophysical law. *Psychological Review*, 64:153–181, 1957. 26

[19] L. Krasula, M. Klíma, E. Rogard, and E. Jeanblanc. MATLAB-based applications for image processing and image quality assessment part I: Software description. *Radioengineering*, 20(4):1009–1015, 2011. 27, 39

[20] L. Krasula, M. Klíma, E. Rogard, and E. Jeanblanc. MATLAB-based applications for image processing and image quality assessment part II: Experimental results. *Radioengineering*, 21(1):154–161, 2012. 27, 39

[21] K. Fliegel, L. Krasula, P. Páta, J. Myslík, J. Pecák, and M. Jícha. System for objective assessment of image differences in digital cinema. In *SPIE 9217, Applications of Digital Image Processing XXXVI*, 2014. 28

[22] F. De Simone, L. Goldmann, V. Baroncini, and T. Ebrahimi. Subjective evaluation of JPEG XR image compression. In *SPIE Optics and Photonics, Applications of Digital Image Processing XXXII*, volume 7443, 2009. 29, 30

[23] O. Dykstra. Rank analysis of incomplete block designs: A method of paired comparisons employing unequal repetitions on pairs. *Biometrics*, 16(2):176 – 188, 1960. 33

[24] J. Li, M. Barkowsky, and P. Le Callet. Analysis and improvement of a paired comparison method in the application of 3DTV subjective experiment. In *2012 IEEE International Conference on Image Processing (ICIP 2012)*, pages 1 – 4, 2012. 33, 104

[25] J. Li, M. Barkowsky, and P. Le Callet. Boosting paired comparison methodology in measuring visual discomfort of 3DTV: Performances of three different designs. In *SPIE-IS&T Electronic Imaging: Stereoscopic Display and Applications XXIV*, 2013. 33, 104

[26] D. A. Silverstein and J. E. Farrell. Quantifying perceptual image quality. In *PICS'98*, pages 242 – 246, 1998. 33

[27] K. Tsukida and M. R. Gupta. How to analyze paired comparison data. Technical report, UWEE, [online] http://mayagupta.org/publications/PairedComparisonTutorialTsukidaGupta.pdf, 2011. 34

[28] L. L. Thurstone. A law of comparative judgement. *Psychological review*, 34(4):273, 1927. 34

[29] F. Moesteller. Remarks on the method of paired comparisons: I. The least squares solution assuming equal standard deviations and equal correlations. *Psychometrika*, 16(1):3–9, March 1951. 34, 36

[30] R. A. Bradley and M. Terry. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4):324 – 345, 1952. 36, 154

[31] R. A. Bradley. Rank analysis of incomplete block designs: III. some large-sample results on estimation and power for a method of paired comparisons. *Biometrika*, 42:450 – 470, 1955. 36

[32] R. A. Bradley. 14 paired comparisons: Some basic procedures and examples. *Handbook of statistics*, 4:299 – 326, 1984. 36

[33] R. D. Luce. *Individual choice behavior: A theoretical analysis*. Wiley, 1959. 36

[34] H. Block and J. Marschak. Random orderings and stochastic theories of responses. *Contributions to Probability and Statistics*, 97, 1960. 36

[35] F. Wickelmaier and C. Schmid. A Matlab function to estimate choice model parameters from paired-comparison data. *Behavior Research Methods*, 36:29 – 40, 2004. 37

[36] J. C. Handley. Comparative analysis of Bradley-Terry and Thurstone-Mosteller paired comparison models for image quality assessment. In *IS&T's Image Processing, Image Quality, Image Capture, System Conference*, pages 108 – 112, 2001. 37, 154

[37] R. A. Fisher. On the interpretation of χ^2 from contingency tables, and the calculation of p. *Journal of Royal Statistical Society*, 85(1):87–94, 1922. 37, 38, 73, 79, 111, 156

[38] G. A. Barnard. A new test for 2x2 tables. *Nature*, 156(3954):177, 1945. 37, 38, 73, 156

[39] J. Li, M. Barkowsky, and P. Le Callet. Subjective assessment methodology for preference of experience in 3DTV. In *11th IEEE IVMSP Workshop: 3D Image/Video Technologies and Applications*, 2013. 37

[40] J. Søgaard, L. Krasula, M. Shahid, D. Temel, K. Brunnström, and M. Razaakh. Applicability of existing objective metrics of perceptual quality for adaptive video streaming. In *Electronic Imaging, Image Quality and System Performance XIII*, 2016. 39

[41] T. Richter. From index to metric: Using differential geometry to define a global visual quality metric. In *SPIE 8135, Applications of Digital Image Processing XXXIV*, 2011. 39

[42] W. Lin and C.-C. J. Kuo. Perceptual visual quality metrics: A survey. *Journal of Visual Communication and Image Representation*, 22(4):297 – 312, 2011. 39

[43] M. Gaubatz and S. S. Hemami. MeTriX MuX visual quality assessment package. [online] http://foulard.ece.cornell.edu/gaubatz/metrix_mux/metrix_mux_1.1.zip (visited 17/05/2015). 39

[44] A. V. Murthy and L. J. Karam. A MATLAB-based framework for image and video quality evaluation. In *Proceedings of QoMEX 2010*, 2010. 39, 47, 48, 107, 121

[45] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, April 2004. 40, 41, 69, 78, 119

[46] Z. Wang, E. Simoncelli, and A. Bovik. Multi-scale structural similarity for image quality assessment. In *IEEE Asilomar Conference on Signal, Systems and Computers*, volume 2, pages 1398–1402, 2003. 40, 42, 69, 78, 86, 119

[47] Z. Wang and Q. Li. Information content weighting for perceptual image quality assessment. *IEEE Transactions on Image Processing*, 20(5):1185 – 1198, May 2011. 40, 42, 43, 78, 79, 81, 119, 201, 203

[48] H. R. Sheikh and A. C. Bovik. Image information and visual quality. *IEEE Transactions on Image Processing*, 15(2):430–444, 2006. 40, 44, 84, 107, 119

[49] L. Zhang, L. Zhang, X. Mou, and D. Zhang. FSIM: A feature similarity index for image quality assessment. *IEEE Transactions on Image Processing*, 20(8):2378–2386, August 2011. 40, 44

[50] E. C. Larson and D. M. Chandler. Most apparent distortion: Full-reference image quality assessment and the role of strategy. *Journal of Electronic Imaging*, 19(1):011006–1 – 011006–21, 2010. 40, 44, 59, 69, 78, 80, 83, 84, 107, 201

[51] S. Daly. *Digital Images and Human Vision*, chapter The Visible Differences Predictor: An Algorithm for the Assessment of Image Fidelity, pages 179 – 206. MIT Press, 1993. 40, 45

[52] R. Mantiuk, S. Daly, K. Myszkowski, and H-P. Seidl. Predicting visible differences in high dynamic range images – model and its calibration. In *Human Vision and Electronic Imaging, SPIE*, 2005. 40, 45, 86

[53] R. Mantiuk, K. J. Kim, A. G. Rempel, and W. Heidrich. HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. In *ACM Transactions on Graphics (Proc. of SIGGRAPH'11)*, volume 30, 2011. 40, 45, 149

[54] M. Narwaria, R. Mantiuk, M. Perreira Da Silva, and P. Le Callet. HDR-VDP-2.2: A calibrated method for objective quality prediction of high-dynamic range and standard images. *Journal of Electronic Imaging*, 24(1):010501, 2014. 40, 45

[55] Z. Wang, H. Sheikh, and A. Bovik. No-reference perceptual quality assessment of JPEG compressed images. In *IEEE International Conference on Image Processing*, volume 1, pages I–477 – I–480, September 2002. 40, 46, 165

[56] H. R. Sheikh, A. C. Bovik, and L. Cormack. No-reference quality assessment using natural scene statistics: JPEG2000. *IEEE Transactions on Image Processing*, 14(11):1918–1927, November 2005. 40, 47, 165

[57] S. Erasmus and K. Smith. An automatic focusing and astigmatism correction system for the SEM and CTEM. In *J. Microscopy*, volume 127, pages 185–199, 1982. 40, 47, 107, 121, 164

[58] L. Firestone, K. Cook, N. Talsania, and K. Preston. Comparison of autofocus methods for automated microscopy. In *Cytometry*, volume 12, pages 195–206, 1991. 40, 48, 107, 164

[59] C. F. Batten. Autofocusing and astigmatism correction in the scanning electron microscope. Master's thesis, University of Cambridge, Cambridge, U.K., 2000. 40, 48, 107, 121, 164

[60] X. Marichal, W. Ma, and H. J. Zhang. Blur determination in the compressed domain using DCT information. In *IEEE International Conference on Image Processing*, volume 2, pages 386–390, 1999. 40, 48, 107, 164

[61] N. Zhang, A. Vladar, M. Postek, and B. Larrabee. A kurtosis-based statistitcal measure for two-dimensional processes and its application to image sharpness. In *Proceedings Section of Physical and Engineering Sciences of American Statistical Society*, pages 4730–4736, 2003. 40, 48, 49, 107, 164

[62] P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi. Perceptual blur and ringing metrics: Applications to JPEG2000. *Signal Processing: Image Communications*, 19(2):163–172, February 2004. 40, 48, 107, 164

[63] D. Shaked and I. Tastl. Sharpness measure: Towards automatic image enhancement. In *IEEE International Conference on Image Processing*, volume 1, pages 937–940, September 2005. 40, 49, 107, 164

[64] R. Ferzli, L. J. Karam, and J. Caviedes. A robust image sharpness metric based on kurtosis measurement of wavelet coefficients. In *Proceedings of the 1st International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, 2005. 40, 49, 107, 164

[65] R. Ferzli and L. J. Karam. A no reference objective sharpness metric using Riemannian tensor. In *Third International Workshop on Video Processing and Quality Metrics for Consumer Electronics VPQM-07, Scottsdale, Arizona*, pages 25–26, January 2007. 40, 49, 107, 121, 164

[66] R. Ferzli and L. J. Karam. A no-reference objective image sharpness metric based on the notion of just noticeable blur (JNB). *IEEE Transactions on Image Processing*, 18(4):717–728, April 2009. 40, 49, 50, 107, 164

[67] N. Narvekar and L. J. Karam. A no-reference image blur metric based on the cumulative probability of blur detection (cpbd). *IEEE Transactions on Image Processing*, 20(9):2678 – 2883, September 2011. 40, 50, 107, 164

[68] C. Vu, T. Phan, and D. Chandler. S3: A spectral and spatial measure of local perceived sharpness in natural images. *IEEE Transactions on Image Processing*, 21(3), 2011. 40, 50, 51, 84, 107, 112, 164, 172

[69] P. V. Vu and D. M. Chandler. A fast wavelet-based algorithm for global and local image sharpness estimation. *Signal Processing Letters, IEEE*, 19(7):423 –426, july 2012. 40, 51, 104, 107, 112, 159, 164, 172

[70] S. Agaian, K. P. Lentz, and A. M. Grigoryan. A new measure of image enhancement. In *IASTED International Conference on Signal Processing and Communications*, 2000. 40, 52, 164

[71] S. Agaian, B. Silver, and K. A. Panetta. Transform coefficient histogram-based image enhancement algorithms using contrast entropy. *IEEE Transactions on Image Processing*, 16:741 – 758, 2007. 40, 52, 164

[72] B. Moulden, F. A. A. Kingdom, and L. D. Gatley. The standard deviation of luminance as a metric for contrast in random-dot images. *Perception*, 19:79–101, 1990. 40, 52, 84

[73] K. Panetta, Ch. Gao, and S. Agaian. No reference color image contrast and quality measures. *IEEE Transactions on Consumer Electronics*, 59(3):643 – 651, August 2013. 40, 52, 54, 146, 164

[74] K. Panetta, Y. Zhou, and S. Agaian. Nonlinear unsharp masking for mammogram enhancement. *IEEE Transactions on Information Technology in Biomedicine*, 15(4):918 – 928, 2011. 40, 52, 99, 164

[75] S. Nercessian, S. S. Agaian, and K. A. Panetta. Multi-scale image enhancement using a second derivative-like measure of contrast. In *SPIE-IS&T Electronic Imaging: Image Processing: Algorithms and Systems X*, volume 8295, 2012. 40, 52, 99, 100, 101

[76] Y.-Y. Fu. *Color image quality measures and retrieval*. PhD thesis, Department of Computer Science, New Jersey Institute of Technology, 2003. 40, 53, 164

[77] M. K. Agoston. *Computer Graphics and Geometric Modeling: Implementation and Algorithms*. London: Springer, 2005. 40, 54

[78] A. K. Moorthy and A. C. Bovik. A two-step framework for constructing blind image quality indices. *IEEE Signal Processing Letters*, 17(5):513–516, May 2010. 40, 55, 164

[79] M. A. Saad, A. C. Bovik, and C. Charrier. A DCT statistics-based blind image quality index. *IEEE Signal Processing Letters*, 17(6):583–586, June 2010. 40, 55

[80] M. A. Saad, A. C. Bovik, and C. Charrier. DCT statistics model-based blind image quality assessment. In *Proceedings of the IEEE International Conference on image Processing (ICIP)*, pages 3093–3096, September 2011. 40, 55, 159

[81] M. A. Saad, A. C. Bovik, and C. Charrier. Blind image quality assessment: A natural scene statistics approach in the DCT domain. *IEEE Transactions on Image Processing*, 21(8):3339–3352, 2012. 40, 55, 164

[82] A. Mittal, A. K. Moorthy, and A. C. Bovik. Referenceless image spatial quality evaluation engine. In *45th Asilomar Conference on Signals, Systems and Computers*, November 2011. 40, 55, 56, 164

[83] L. Liu, H. Dong, H. Huang, and A. C. Bovik. No-reference image quality assessment in curvelet domain. *Signal Processing: Image Communication*, 29(4):494–505, 2014. 40, 56, 159, 164

[84] A. Mittal, R. Soundararajan, and A. C. Bovik. Making a completely blind image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013. 40, 56, 107, 159, 164

[85] W. Xue, L. Zhang, and X. Mou. Learning without human scores for blind image quality assessment. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013. 40, 57, 107, 159, 164

[86] Q. Sang, X. Wu, Ch. Li, and Y. Lu. Universal bling image quality assessment using contourlet transform and singular-value decomposition. *Journal of Electronic Imaging*, 23(6), 2014. 40, 57, 107, 159, 164

[87] R. Dosselmann and X. D. Yang. A comprehensive assessment of the structural similarity index. *Signal, Image and Video Processing*, 5:81 – 91, 2009. 42

[88] H.R. Sheikh, A.C. Bovik, and G. De Veciana. An information fidelity criterion for image quality assessment using natural scene statistics. *IEEE Transactions on Image Processing*, 14(12):2117–2128, December 2005. 44

[89] E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J. Heeger. Shiftable multi-scale transform. *IEEE Transactions Information Theory*, 38(2):587 – 607, March 1992. Special Issue on Wavelets. 44

[90] P. Kovesi. Image features from phase congruency. *Videre: Journal of Computer Vision Research*, 1(3):1–26, 1999. 44

[91] P. Dostál, L. Krasula, and M. Klíma. HLFSIM: Objective image quality metric based on ROI analysis. In *IEEE International Carnahan Conference on Security Technology (ICCST)*, 2012. 44

[92] J. Theeuwes. Top-down and bottom-up control of visual selection. *Acta Psychologica*, 135(2):77–99, 2010. 44

[93] P. Dostál, L. Krasula, and M. Klíma. Influence of high level features of HVS on performance of FSIM. *Radioengineering*, 22(4):1048–1055, 2013. 44

[94] D. M. Chandler and S. S. Hemami. VSNR: A waveletbased visual signal-to-noise ratio for natural images. *IEEE Transactions on Image Processing*, 16(9):2284–2298, September 2007. 45, 78, 80

[95] G. Valensize, F. De Simone, P. Lauga, and F. Dufaux. Performance evaluation of objective quality metrics for HDR image compression. In *SPIE 9217, Applications of Digital Image Processing XXXVII*, 2014. 45

[96] P. Hanhart, M. V. Bernardo, M. Pereira, A. M. G. Pinheiro, and T. Ebrahimi. Benchmarking of objective quality metrics for HDR image quality assessment. *EURASIP Journal on Image and Video Processing*, 2015. 45

[97] M. Azimi, A. Banitalebi-dehkordi, Y. Dong, M. T. Pourazad, and P. Nasiopoulos. Evaluating the performance of existing full-reference quality metrics on high dynamic range (HDR) video content. In *International Conference on Multimedia Signal Processing (ICMSP)*, 2014. 45

[98] M. Řeřábek, P. Hanhart, P. Korschunov, and T. Ebrahimi. Subjective and objective evaluation of HDR video compression. In *9th International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*, 2015. 45

[99] T. Vigier, L. Krasula, A. Milliat, M. Perreira Da Silva, and P. Le Callet. Performance and robustness of HDR objective quality metrics in the context of recent compression scenarios. In *IEEE Digital Media Industry and Academic Forum*, 2016. 45

[100] Z. Wang and E. Simoncelli. Reduced-reference image quality assessment using a wavelet-domain natural image statistic model. In *Human Vision and Electronic Imaging X, Proc. SPIE*, volume 5666, January 2005. 45

[101] R. Soundararajan and A. C. Bovik. RRED indices: Reduced reference entropic differencing for image quality assessment. *IEEE Transactions on Image Processing*, 21(2):517–526, February 2012. 45

[102] N. Sochen, R. Kimmel, and R. Malladi. A general framework for low level vision. *IEEE Transactions on Image Processing*, 7(2):310–318, February 1998. 49

[103] C. Sagiv, N. Sochen, and Y. Y. Zeevi. Integrated active contours for texture segmentation. *IEEE Transactions on Image Processing*, 15(6):1633–1646, 2006. 49

[104] A. Cohen, I. Daubechies, and J.-C. Feauveau. Biorthogonal bases of compactly supported wavelets. *Communications on Pure and Applied Mathematics*, 45, 1992. 51

[105] H. W. Park and H. S. Kim. Motion estimation using low-band-shift method for wavelet-based moving-picture coding. *IEEE Transactions on Image Processing*, 9(4), 2000. 51

[106] K. Matković, L. Neumann, A. Neumann, T. Psik, and W. Purgathofer. Global contrast factor – a new approach to image contrast. In *Proceedings of the First Eurographics Conference on Computational Aesthetics in Graphics, Visualization and Imaging*, Computational Aesthetics'05, pages 159–167, Aire-la-Ville, Switzerland, Switzerland, 2005. Eurographics Association. 53, 146, 164

[107] M. Luo, G. Cui, and B. Rigg. The development of the CIE 2000 colour difference formula: CIEDE2000. *Color Research & Application*, 26(5):340–350, 2001. 53

[108] D. Hasler and S. E. Suesstrunk. Measuring colorfulness in natural images. In *Electronic Imaging*, pages 87–95, 2003. 53, 55

[109] E. S. L. Gastal and M. M. Oliveira. Domain transform for edge-aware image and video processing. *ACM Transactions on Graphics (Proc. of SIGGRAPH)*, 30(4):69:1–69:12, 2011. 54

[110] H. R. Sheikh, M. F. Sabir, and A. C. Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on Image Processing*, 15(11):3440–3451, 2006. 55, 56, 59, 78, 80, 83

[111] K. Sharifi and A. Leon-Garcia. Estimation of shape parameter for generalized gaussian distributions in subband decompositions of video. *IEEE Transactions on Circuits and Systems For Video Technology*, 5(1):52–56, 1995. 55

[112] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating sementation algorithms and measuring ecological statistics. In *International Conference on Computer Vision*, volume 2, pages 416 – 423, 2001. 56

[113] P. Le Callet and F. Autrusseau. Subjective quality assessment irccyn/ivc database, 2005. http://www.irccyn.ec-nantes.fr/ivcdb/. 59, 78, 80, 100, 118

[114] Z. M. Parvez Sazzad, Y. Kawayoke, and Y. Horita. Image quality evaluation database. http://mict.eng.u-toyama.ac.jp/database_toyama/. 59, 78, 80

[115] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti. TID2008 – a database for evaluation of full-reference visual quality assessment metrics. *Advances of Modern Radioelectronics*, 10(4):30–45, 2009. [online]. Available: http://www.ponomarenko.info/tid2008.htm. 59, 70, 78, 80, 83

[116] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, and C.-C. Jay Kuo. Image database TID2013: Peculiarities, results and perspectives. *Signal Processing: Image Communication*, 30:55–77, 2015. 59, 70, 83

[117] ITU-T Reccomendation P.1401. Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models, 2012. 60, 61, 68, 69

[118] VQEG. Final report from the video quality experts group on the validation of objective models of video quality assessment. Technical report, VQEG, 2000. 60, 66, 68, 69

[119] ITU-T Reccomendation P.149. Method for specifying accuracy and cross-calibration of video quality metrics (VQM), 2004. 60, 63

[120] K. Brunnström, S. Tavakoli, and J. Søgaard. Compensating for type-I errors in video quality assessment. In *7th International Worskhop on Quality of Multimedia Experience (QoMEX)*, 2015. 66

[121] Y. Benjamini. Simultaneous and selective inference: Current successes and future challenges. *Biometrical Journal*, 52:708–721, 2010. 66

[122] J. H. McDonald. *Handbook of Biological Statistics*. Sparky House Publishing, 2008. [online] http://www.biostathandbook.com/. 66

[123] S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. 67

[124] R. J. Simes. An improved Bonferroni procedure for multiple test of significance. *Biometrika*, 73:751–754, 1986. 67

[125] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, 57(1):289–300, 1995. 67, 79, 107, 159

[126] M. H. Pinson and S. Wolf. An objective method for combining multiple subjective data sets. In *Visual Communications and Image Processing*, 2003. 69

[127] S. D. Voran. Iterated nested least-squares algorithm for fitting multiple data sets. Technical report, NASA/STI, 2002. 69

[128] J. A. Swets. *Signal Detection Theory and ROC Analysis in Psychology and Diagnostics: Collected Papers.* Mahwah, New Jersey: Lawrence Erlbaum Associates, 1996. 73

[129] L. Krasula, K. Fliegel, P. Le Callet, and M. Klíma. On the accuracy of objective image and video quality models: New methodology for performance evaluation. In *International Conference on Quality of Multimedia Experience (QoMEX)*, 2016. 74

[130] P. Hanhart, L. Krasula, P. Le Callet, and T. Ebrahimi. How to benchmark objective quality metrics from paired comparison data. In *International Conference on Quality of Multimedia Experience (QoMEX)*, 2016. 74, 77

[131] A. Slaby. ROC analysis with Matlab. In *ITI 2007 29th International Conference on Information Technology Interfaces*, June 2007. 74

[132] D. Bamber. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*, 12(4):387–415, 1975. 74

[133] H. Mann and D. Whitney. On a test of wheather one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60, 1947. 74

[134] E. S. Venkatraman. A permutation test to compare receiver operating characteristic curves. *Biometrics*, 56(4):1134–1138, 2000. 75

[135] J. A. Hanley and B. J. McNeil. A method of comparing the area under two ROC curve derived from the same cases. *Radiology*, 148:839–843, 1983. 75, 79, 107, 159

[136] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson. Comparing the areas under two or more correlated reciever operating characteristic curves: A nonparametric approach. *Biometrics*, 44(3):837–845, September 1988. 75

[137] J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143:29–36, 1982. 75

[138] X. Sun and W. Xu. Fast implementation of delong's algorithm for comparing the areas under correlated receiver operating characteristic curves. *IEEE Signal Processing Letters*, 21(11):1389–1393, 2014. 76

[139] K. Gu, G. Zhai, X. Yang, W. Zhang, and M. Liu. Subjective and objective quality assessment for images with contrast change. In *IEEE International Conference on Image Processing*, pages 383–387, 2013. 84

[140] K. Gu, G. Zhai, W. Lin, and M. Liu. The analysis of image contrast: From quality assessment to automatic enhancement. *IEEE Transactions on Cybernetics*, 46(1):284–297, 2016. 84

[141] S. Bouzit and L. W. MacDonald. Assessing the enhancement of image sharpness. In *SPIE-IS&T Electronic Imaging: Image Quality and System Performance III*, volume 6059, 2006. 84, 95

[142] C. Vu, T. Phan, P. Singh, and D. M. Chandler. On the quality assessment of enhanced images: A database, analysis, and strategies for augmenting existing methods. In *IEEE Southwest Symposium on Image Analysis and Interpretation*, 2012. 84, 85, 107, 173, 203

[143] J. Petit and R. Mantiuk. Assessment of video tone-mapping: Are cameras' s-shaped tone-curves good enough? *J. Vis. Commun. Image R.*, 24:1020–1030, 2013. 85, 138

[144] M. Ashikhmin and J. Goyal. A reality check for tone mapping operators. *ACM Trans. on Applied Perception*, 3(4):399–411, 2006. 86, 152

[145] J. Kuang, H. Yamaguchi, G. M. Johnson, and M. D. Fairchild. Testing HDR image rendering algorithms. In *IS&T/SID 12th Color Imaging Conf.*, 2004. 86

[146] T. O. Aydın, R. Mantiuk, K. Myszkowski, and H-P. Seidl. Dynamic range independent image quality assessment. In *International Conference on Computer Graphics and Interactive Techniques*, 2008. 86, 87, 143, 158, 159, 164, 174, 203

[147] A. Watson. Visual detection of spatial contrast patterns: Evaluation of five simple models. *Optics Express*, 6(1):12–33, 2000. 86

[148] H. Yeganeh and Z. Wang. Objective quality assessment of tone-mapped images. *IEEE Transactions on Image Processing*, 22(2):657–667, February 2013. 86, 146, 147, 149, 152, 159, 164, 166, 167, 174, 201

[149] W. J. Crozier. On the variability of critical illumination for flicker fusion and intensity discrimination. *Journal of General Physiology*, 19(3):503–522, 1935. 87

[150] J. L. Mannos and D. J. Sakrison. The effects of a visual fidelity criterion on the encoding of images. *IEEE Transactions on Information Theory*, 20(4):525–536, 1974. 87

[151] D. H. Kelly. Effects of sharp edges on the visibility of sinusoidal gratings. *Journal of the Optical Society of America*, 60(1):98–103, 1970. 87

[152] K. Ma, H. Yeganeh, K. Zeng, and Z. Wang. High dynamic range image compression by optimizing tone mapped image quality index. *IEEE Transactions on Image Processing*, 24(10):3086–3097, 2015. 88, 159, 164, 174

[153] H. Ziaei Nafchi, A. Shahkolaei, R. Farrahi Moghaddam, and M. Cheriet. FSITM: A feature similarity index for tone-mapped images. *IEEE Signal Processing Letters*, 22(8):1026–1029, 2015. 88, 159, 164, 167, 174

[154] B. Fraser and J. Schewe. *Real World Image Sharpening with Adobe Photoshop, Camera Raw, and Lightroom*. Peachpit Press, second edition, 2010. 91, 95, 96, 99, 101

[155] Tutorials: Sharpness. http://www.cambridgeincolour.com/tutorials/sharpness.htm. [Online] (visited on 11/06/2016). 91, 92, 93

[156] Sharpening: Unsharp mask. http://www.cambridgeincolour.com/tutorials/unsharp-mask.htm. [Online] (visited on 11/06/2016). 91

[157] Guide to Image Sharpening. http://www.cambridgeincolour.com/tutorials/image-sharpening.htm. [Online] (visited on 11/06/2016). 91

[158] B. Fraser. Out of gamut: (almost) everything you wanted to know about sharpening in photoshop but were afraid to ask. http://www.creativepro.com/article/out-gamut-almost-everything-you-wanted-know-about-sharpening-photoshop-were-afraid-ask. [Online] (visited on 11/06/2016). 91

[159] B. Fraser. Out of gamut: A two-pass approach to sharpening in photoshop. http://www.creativepro.com/article/out-gamut-a-two-pass-approach-sharpening-photoshop. [Online] (visited on 11/06/2016). 91

[160] M. Aland. Sharpen photos smartly. http://www.creativepro.com/article/sharpen-photos-smartly. [Online] (visited on 11/06/2016). 91

[161] M. Reichmann. Understanding sharpness. https://luminous-landscape.com/understanding-sharpness/. [Online] (visited on 11/06/2016). 91, 92, 96

[162] L. Krasula. Quality assessment of sharpened images. Master's thesis, Czech Technical University in Prague, May 2013. 92, 104

[163] L. Krasula, P. Le Callet, K. Fliegel, and M. Klíma. Quality assessment of sharpened images: Challenges, methodology, and objective metrics. *IEEE Transactions on Image Processing*. (under review). 92

[164] J. Y. Park, S. Triantaphillidou, R. E. Jacobson, and G. Gupta. Evaluation of perceived image sharpness with changes in the displayed image size. In *SPIE-IS&T Electronic Imaging: Image Quality and System Performance IX*, volume 8293, 2012. 92

[165] F. Crete, T. Dolmiere, P. Ladret, and M. Nicolas. The blur effect: Perception and estimation with a new no-reference perceptual blur metric. In *SPIE-IS&T Electronic Imaging: Human Vision and Electronic Imaging XII*, volume 6492, 2007. 93

[166] T. Vuori and M. Olkkonen. The effect of image sharpness on quantitative eye movement data and on image quality evaluation while viewing natural images. In *SPIE-IS&T Electronic Imaging: Image Quality and System Performance III*, volume 6059, 2006. 93

[167] A. Katsaggelos. *A Super-Resolution of Images*. Morgan and Claypool Publishers, 2007. 94

[168] S. Chaudhuri. *Super-Resolution Imaging*. Springer, 2001. 94

[169] N. Arad and C. Gotsman. Enhancement by image-dependent warping. *IEEE Transactions on Image Processing*, 8(8):1063 – 1074, August 1999. 94

[170] J. G. M. Schavemaker, M. J. T. Reinders, J. J. Gerbrands, and E. Backer. Image sharpening by morphological filtering. *Pattern Recognition*, 33(6):997 – 1012, June 2000. 94

[171] Q. Wang, R. Ward, and J. Zou. Contrast enhancement for enlarged images based on edge sharpening. In *IEEE International Conference on Image Processing*, 2005. 94

[172] H. Q. Luong and W. Philips. Sharp image interpolation by mapping level curves. In *Visual Communications and Image Processing Conference*, 2005. 94

[173] A. Bruna, A. Buemi, M. Guarnera, and G. Santoro. Adaptive directional sharpening with overshoot control. In *SPIE-IS&T Electronic Imaging: Image Processing: Algorithms and Systems VI*, volume 6812, 2008. 94

[174] I. V. Safonov, M. N. Rychagov, K. Kang, and S. H. Kim. Adaptive sharpening of photos. In *SPIE-IS&T Electronic Imaging: Color Imaging XIII: Processing, Hardcopy, and Applications*, volume 6807, 2008. 95, 96

[175] S. Paris, S. W. Hasinoff, and J. Kautz. Local Laplacian filters: Edge-aware image processing with a Laplacian pyramid. *ACM Transactions on Graphics*, 30(4), 2011. 95

[176] S. Bae, S. Paris, and F. Durand. Two-scale tone management for photographic look. *ACM Transactions on Graphics*, 25:637–645, 2006. 95

[177] C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images. In *IEEE International Conference on Computer Vision*, pages 839–846, 1998. 95, 128, 134, 135, 136

[178] K. K. V. Toh and N. A. M. Isa. Locally adaptive bilateral clustering for image deblurring and sharpness enhancement. *IEEE Transactions on Consumer Electronics*, 57(3):1227 – 1235, August 2011. 95

[179] Y. Ling, C. Yan, C. Liu, X. Wang, and H. Li. Adaptive tone-preserved image detail enhancement. *The Visual Computer*, 28:733–742, 2012. 95

[180] Z. Gui and Y. Liu. An image sharpening algorithm based on fuzzy logic. *Optik*, 122:697 – 702, 2011. 95

[181] F. Russo. An image enhancement technique combining sharpening and noise reduction. *IEEE Transactions on Instrumentation and Measurement*, 51(4):824 – 828, 2002. 95

[182] R. W. Lambrecht and CH. Woodhouse. *Way Beyond Monochrome*. Focal Press, 2011. 95

[183] T-H. Yu, S. K. Mitra, and J. F. Kaiser. A novel nonlinear filter for image enhancement. In *SPIE/SPSE Conference on Image Processing Algorithms and Techniques II*, pages 303 – 305, 1991. 96

[184] T-H. Yu and S. K. Mitra. Unsharp masking with nonlinear filters. In *Seventh European Signal Processing Conference, EUSIPCO-94*, 1994. 96

[185] G. Ramponi, N. K. Strobel, S. K. Mitra, and T-H. Yu. Nonlinear unsharp masking methods for image contrast enhancement. *Journal of Electronic Imaging*, 5(3):353 – 366, July 1996. 96

[186] N. Strobel and S. K. Mitra. Quadratic filters for image contrast enhancement. In *Conference Record of the Twenty-Eighth Asilomar Conference on Signals, Systems and Computers*, volume 1, 1994. 96

[187] P. Maragos and R. W. Schafer. Morphological filters – Part I: Their set theoretic analysis and relations to linear shift invariant filters. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 35(8):1152 – 1169, 1987. 96

[188] A. Polesel, G: F. Ramponi, and J. J. Mathews. Image enhancement via adaptive unsharp masking. *IEEE Transactions on Image Processing*, 9:505 – 510, 2000. 96

[189] M. A. Badamchizadeh and A. Aghagolzadeh. Comparative study of unsharp masking methods for image enhancement. In *Third International Conference on Image and Graphics (ICIG'04)*, 2004. 96

[190] S. Agaian and S. A. McClendon. Novel medical image enhancement algorithms. In *SPIE-IS&T Electronic Imaging: Image Processing: Algorithms and Systems VIII*, volume 7532, 2010. 96, 99, 101

[191] M. Sonka, V. Hlaváč, and R. Boyle. *Image Processing, Analysis and Machine Vision*. Thomson Learning, third edition, 2008. 97

[192] J. F. Canny. A Computional Approach to Edge Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(6):679–698, 1986. 97, 112

[193] S. DelMarco and S. Agaian. The design of wavelets for image enhancement and target detection. In *SPIE Mobile Multimedia/Image Processing, Security, and Applications*, volume 7351, April 2009. 99

[194] A. Leontaris, P. C. Cosman, and A. R. Reibman. Quality evaluation of motion-compensated edge artifacts in compressed video. *IEEE Transactions on Image Processing*, 16(4):943 – 956, April 2007. 104

[195] E. Reinhard, G. Ward, S. Pattanaik, P. Debevec, W. Heidrich, and K. Myszkowski. *High Dynamic Range Imaging: Acquisition, Display, and Image-Based Lighting*. CA: Morgan Kaufmann, 2010. 125, 126, 127, 128, 129, 133, 134, 136, 204

[196] M. Bürker, C. Rößing, and H. P. A. Lensch. Exposure control for HDR video. In *SPIE Photonics Europe*, 2014. 125

[197] M. Granados, B. Ajdin, M. Wand, C. Theobalt, H.-P. Seidl, and H. P. A. Lensch. Optimal HDR reconstruction with linear digital cameras. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 215–222, 2010. 126, 127, 128

[198] S. Mann and R. W. Pickard. Extending dynamic range by combining different exposed pictures. In *IS&T Annual Conference*, pages 442–448, 1995. 126, 127

[199] P. Debevec and J. Malik. Recovering high dynamic range radiance maps from photographs. In *ACM SIGGRAPH*, pages 369–378, 1997. 126, 127

[200] T. Mitsunaga and S. K. Nayar. Radiometric self calibration. In *IEEE Computer Vision and Pattern Recognition Conference*, pages 1374–1380, 1999. 127

[201] Y. Tsin, V. Ramesh, and T. Kanade. Statistical calibration of the CCD imaging process. In *IEEE International Conference on Computer Vision*, pages 480–487, 2001. 127

[202] M. Robertson, S. Borman, and R. Stevenson. Estimation-theoretic approach to dynamic range improvement using multiple exposures. *Journal of Electronic Imaging*, 12(2):219–228, 2003. 127

[203] K. Kirk and H. J. Andersen. Noise characterization of weighting schemes for combination of multiple exposures. In *British Machine Vision Conference*, volume 3, pages 1129–1138, 2006. 127

[204] J. Janesick. *Scientific charge-coupled devices*. SPIE Press, 2001. 127

[205] P. Debevec, E. Reinhard, G. Ward, and S. Pattanaik. High dynamic range imaging. In *SIGGRAPH'04 ACM Course Notes*, 2004. 128

[206] Francesco Banterle, Alessandro Artusi, Kurt Debattista, and Alan Chalmers. *Advanced High Dynamic Range Imaging: Theory and Practice*. AK Peters (CRC Press), Natick, MA, USA, 2011. 128, 131, 133, 136, 137, 147, 150, 204

[207] A. Pinheiro, K. Fliegel, P. Korshunov, L. Krasula, M. Bernardo, M. Pereira, and T. Ebrahimi. Performance evaluation of the emerging JPEG XT image compression standard. In *IEEE 16th International Workshop on Multimedia Signal Processing (MMSP)*, 2014. 128

[208] G. Ward. A wide field high dynamic range stereographic viewer. In *PICS 2002*, April 2002. 128

[209] P. Ledda, A. Chalmers, and H. Seetzen. HDR display: A validation against reality. In *International Conference on Systems, Man and Cybernetics*, 2004. 128

[210] S. R. Ellis, W. S. Kim, M. Tyler, M. W. McGreevy, and L. Stark. Visual enhancements for perspective displays: Perspective parameters. In *IEEE International Conference on Systems, Man and Cybernetics*, 1985. 128

[211] H. Seetzen, W. Heidrich, W. Stuerzlinger, G. Ward, L. Whitehead, M. Trentacoste, A. Ghosh, and A. Vorozcovs. High dynamic range display systems. *ACM Transactions on Graphics*, 23(3), 2004. 129, 130, 131, 204

[212] S. Luka and J. A. Ferwerda. Colorimetric image splitting for high dynamic range displays. In *Annual Conference SID 2009*, pages 1298–1301, 2009. 130

[213] D. Zhang, J. A. Ferwerda, and J. B. Phillips. Appearance-based image splitting for HDR displays. In *18th Color and Imaging Conference: Color Science and Engineering Systems, Technologies and Applications*, 2010. 130

[214] J. Kuang, G. M. Johnson, and M. D. Fairchild. iCAM06: A refined image appearance model for HDR image rendering. *Journal of Visual Communication and Image Representation*, 18:406–414, 2007. 130, 135, 139, 147, 151

[215] H. M. Visser, J. J. W. M. Rosink, N. Raman, and R. Rajae-Joordens. Tuning LCD displays to medical applications. In *International Display Research Conference (EuroDisplay 2005)*, pages 74–77, September 2005. 131

[216] Chiron Project Whitepaper. Dual layer high dynamic range display for medical applications. Technical report, Univeristy of Trieste and FIMI S.R.L., 2010. 131, 132

[217] G. Guarnieri, L. Albani, and G. Ramponi. Image-splitting techniques for dual-layer high dynamic range LCD display. *Journal of Electronic Imaging*, 17(4):043009–1–9, 2008. 132

[218] G. Guarnieri, L. Albani, and G. Ramponi. Minimum-error splitting algorithm for dual leyer LCD display – part I: Background and theory. *Journal of Display Technology*, 4(4):383–390, 2008. 132

[219] G. Guarnieri, L. Albani, and G. Ramponi. Minimum-error splitting algorithm for dual leyer LCD display – part II: Implementation and results. *Journal of Display Technology*, 4(4):391–397, 2008. 132

[220] Digital imaging and communications in medicine (DICOM) – part 14: Grayscale standard display function. Technical report, National Electrical Manufacturers Association, 2007. 132

[221] G. Guarnieri, G. Ramponi, S. Bonfiglio, and L. Albani and. Nonlinear mapping of the luminance in dual-layer high dynamic range displays. In *IS&T/SPIE Electronic Imaging*, 2009. 132

[222] G. S. Miller and C. R. Hoffman. Illumination and reflection maps: Simulated objects in simulated and real environments. In *SIGGRAPH 84 Course Notes for Advanced Computer Graphics Animation*, July 1984. 133

[223] J. C. Stevens and S. S. Stevens. Brightness function: Effects of adaptation. *Journal of the Optical Society of America*, 53(3), 1963. 133

[224] J. Tumblin and H. Rushmeier. Tone reproduction for realistic computer generated images. Technical Report GIT-GVU-91-13, Graphics, Visualization, and Useability Center, Georgia Institute of Technology, 1991. 133, 138

[225] J. Tumblin and H. Rushmeier. Tone reproduction for computer generated images. *IEEE Computer Graphics and Applications*, 6:42 – 48, November 1993. 133, 136, 137

[226] G. Ward. *Graphics Gems IV*, chapter A Contrast-based Scale Factor for Luminance Display, pages 415–421. Boston: Academic Press, 1994. 133

[227] CIE. An analytic model for describing the influence of lighting parameters upon visual performance: Vol 1, technical foundations. Technical report, CIE Pub. 19/2.1 Technical Committee 3.1, 1981. 133

[228] J. A. Ferwerda, S. Pattanaik, P. Shirley, and D. P. Greenberg. A model of visual adaptation for realistic image synthesis. In *SIGGRAPH 96*, pages 249–258, August 1996. 133

[229] F. Durand and J. Dorsey. Interactive tone mapping. In *11th Eurographics Workshop on Rendering*, pages 219–230, 2000. 133

[230] F. Drago, K. Myszkowski, T. Annen, and N. Chiba. Adaptive logarithmic mapping for displaying high contrast scenes. In *Computer Graphics Forum*, volume 22, 2003. 134, 147, 151

[231] K. Perlin and E. M. Hoffert. Hypertexture. *Computer Graphics*, 22(3):253–262, 1989. 134

[232] E. Reinhard and K. Devlin. Dynamic range reduction inspired by photoreceptor physiology. *IEEE Transactions on Visualization and Computer Graphics*, 11(1):13–24, 2005. 134

[233] G. Ward, H. Rushmeier, and C. Piatko. A visibility matching tone reproduction operator for high dynamic range scenes. *IEEE Transactions on Visualization and Computer Graphics*, 3(4):291 – 306, 1997. 134, 138

[234] C. Schlick. *Photorealistic Rendering Techniques*, chapter Quantization Techniques for the Visualization of High Dynamic Range Pictures, pages 7–20. New York: Springer-Verlag, 1994. 134

[235] Z. Mai, H. Mansour, R. Mantiuk, P. Nasiopoulos, R. Ward, and W. Heidrich. Optimizing a tone curve for backward-compatible high dynamic range image and video compression. *IEEE Transactions on Image Processing*, 20(6):1558–1571, 2011. 134, 151

[236] F. Dufaux, P. Le Callet, R. Mantiuk, and M. Mrak. *High Dynamic Range Video*. Elsevier Ltd., 2016. 134

[237] P. Lauga, G. Valensize, G. Chierchia, and F. Dufaux. Improved tone mapping operator for HDR coding optimizing the distortion/spatial complexity trade-off. In *22nd European Signal Processing Conference (EUSIPCO)*, 2014. 134

[238] M. Oskarsson. Temporally consistent tone mapping of images and video using optimal k-means clustering. *Journal of Mathematical Imaging and Vision*, pages 1–14, 2016. [Online] http://link.springer.com/article/10.1007/s10851-016-0677-1. 134

[239] K. Chiu, M. Herf, P. Shirley, S. Swamy, C. Wang, and K. Zimmerman. Spatially nonuniform scaling functions for high contrast images. In *Graphics Interface '93*, pages 245–253, 1993. 134

[240] Z. Rahman, D. J. Jobson, and G. A. Woodell. A multiscale retinex for color rendition and dynamic range compression. In *SPIE Applications of Digital Image Processing XIX*, volume 2847, 1996. 134

[241] D. J. Jobson, Z. Rahman, and G. A. Woodell. Retinex image processing: Improved fidelity of direct visual observation. In *IS&T Fourth Color Imaging Conference: Color Science, Systems, and Aplications*, volume 4, pages 124–125, 1995. 134

[242] Z. Rahman, G. A. Woodell, and D. J. Jobson. A comparison of the multiscale retinex with other image enhancement techniques. In *IS&T 50th Annual Conference: A Celebration of All Imaging*, volume 50, pages 426–431, 1997. 134

[243] E. H. Land and J. J. McCann. Lightness and retinex theory. *Journal of the Optical Society of America*, 63(1):1–11, 1971. 134

[244] L. Meylan and S. Süsstrunk. High dynamic range image rendering with a retinex-based adaptive filter. *IEEE Transactions on Image Processing*, 15(9):2820–2830, 2006. 135

[245] CIE. The CIE 1997 interim colour appearance model (simple version). Technical report, CIECAM97, 1998. 135

[246] N. Moroney, M. D. Fairchild, R. W. G. Hunt, C. J. Li, M. R. Luo, and T. Newman. The CIECAM02 color appearance model. In *IS&T 10th Color Imaging Conference*, pages 23–27, 2002. 135

[247] R. W. G. Hunt. *The Reproduction of Colour*. John Wiley and Sons, 2004. 135

[248] S. N. Pattanaik, J. A. Ferwerda, M. D. Fairchild, and D. P. Greenberg. A multiscale model of adaptation and spatial vision for realistic image display. In *SIGGRAPH '98*, pages 287–298, 1998. 135

[249] A. B. Watson and J. A. Solomon. Model of visual contrast gain control and pattern masking. *Journal of the Optical Society of America*, 14(9):2379–2391, 1997. 135

[250] M. D. Fairchild and G. M. Johnson. Meet iCAM: An image color appearance model. In *IS&T/SID 10th Color Imaging Conference*, pages 33–38, 2002. 135

[251] M. D. Fairchild and G. M. Johnson. The iCAM framework for image appearance, image differences, and image quality. *Journal of Electronic Imaging*, 13:126–138, 2004. 135

[252] M. Ashikhmin. A tone mapping algorithm for high contrast images. In *13th Eurographics Workshop on Rendering*, pages 145–155, 2002. 135, 136

[253] E. Reinhard, M. Stark, P. Shirley, and J. Fewerda. Photographic tone reproduction for difital images. *ACM Transactions on Graphics*, 21(3):267 – 276, 2002. 135, 138, 139, 147, 152

[254] E. Reinhard, G. Ward, P. Debevec, P. Sumata, and W. Heidrich. Parameter estimation for photographics tone reproduction. *Journal of Graphics Tools*, 7(1):45 – 51, 2003. 135, 138

[255] R. Mantiuk, K. Myszkowski, and H. Seidel. A perceptual framework for contrast processing of high dynamic range images. *ACM Transactions on Applied Perception*, 3(3):286–308, 2006. 136, 151

[256] R. Mantiuk, S. Daly, and L. Kerofsky. Display adaptive tone mapping. *ACM Transactions on Graphics*, 27(3):1–10, 2008. 136

[257] A. V. Oppenheim, R. Schafer, and T. Stockham. Nonlinear filter of multiplied and convolved signals. In *Proceedings of the IEEE*, volume 56, pages 1264–1291, 1968. 136

[258] B. K. P. Horn. Determining lightness from an image. In *CVGIP*, volume 3, pages 277–299, 1974. 136

[259] R Fattal, D. Lischinski, and M. Werman. Gradient domain high dynamic range compression. *ACM Transactions on Graphics*, 21(3):249–256, 2002. 136

[260] J. Tumblin and G. Turk. LCIS: A boundry hierarchy for detail-preserving contrast reduction. In *Siggraph 1999*, pages 83–90, 1999. 136

[261] F. Durand and J. Dorsey. Fast bilateral filtering for the display of high-dynamic-range images. *ACM Transactions on Graphics*, 21(3):257–266, 2002. 136

[262] P. Choudhury and J. Tumblin. The trilateral filter for high contrast images and meshes. In *Eurographics Symposium on Rendering*, pages 186–196, 2003. 137

[263] Y. Li, L. Sharan, and E. H. Adelson. Compressing and companding high dynamic range images with subband architectures. *ACM Transactions on Graphics*, 24(3):836–844, 2005. 137

[264] H. Yee and S. N. Pattanaik. Segmentation and adaptive assimilation for detail-preserving display of high-dynamic range images. *The Visual Computer*, 19(7-8):457–466, 2003. 137

[265] G. Krawczyk, R. Mantiuk, K. Myszkowski, and H.-P. Seidl. Lightness perception inspired tone mapping. In *First ACM Symposium on Applied Perception in Graphics and Visualization (APGV)*, 2004. 137

[266] A. Gilchirst, Ch. Kossyfidis, F. Bonato, T. Agostini, X. Li, J. Cataliotti, B. Spehar, V. Annan, and E. Economou. An anchoring theory of lightness perception. *Psychological Review*, 106(4):795–834, 1999. 137

[267] D. Lischinski, Z. Farbman, M. Uyttendaele, and R. Szeliski. Interactive local adjustment of tonal values. *ACM Transactions on Graphics*, 25(3):646–653, 2006. 137

[268] A. Levin, D. Lischinski, and Y. Weiss. Colorization using optimization. *ACM Transactions on Graphics*, 23(3):689–694, 2004. 137

[269] A. Agarwala, M. Dontcheva, M. Agrawala, S. Drucker, A. Colburn, B. Curless, D. Salesin, and M. Cohen. Interactive digital photomontage. *ACM Transactions on Graphics*, 23(3):294–302, 2004. 137

[270] P. Lauga, A. Koz, G. Valensize, and F. Dufaux. Segmentation-based optimized tone mapping for high dynamic range image and video coding. In *Picture Coding Symposium (PCS)*, 2013. 137

[271] T. Mertens, J. Kautz, and F. Van Reeth. Exposure fusion: A simple and practical alternative to high dynamic range photography. In *PG '07: Proceedings of the 15th Pacific Conference on Computer Graphics and Applications*, pages 382–390, 2007. 138

[272] M. Song, D. Tao, Ch. Chen, J. Bu, J. Luo, and Ch. Zhang. Probabilistic exposure fusion. *IEEE Transactions on Image Processing*, 21(1):341–357, 2012. 138

[273] M. G. M. Stokkermans, M. J. Murdoch, and U. Engelke. Preference for key parameter of tone mapping operator in different viewing conditions. *Experiencing Light*, pages 1 – 4, 2012. 138

[274] B. Barladian. Robust parameter estimation for tone mapping operator. In *13th International Conference on Computer Graphics and Vision - Graphicon 2003*, September 2003. 138

[275] C. Kiser, E. Reinhard, M. Tocci, and N. Tocci. Real-time automated tone mapping system for HDR video. In *IEEE International Conference on Image Processing*, 2012. 138

[276] L. Krasula, M. Narwaria, and P. Le Callet. An automated approach for tone mapping operator parameter adjustment in security applications. In *SPIE 9138, Optics, Photonics, and Digital Technologies for Multimedia Applications III*, 2014. 139

[277] S. Péchard, P. Le Callet, M. Carnec, and D. Barba. A new methodology to estimate the impact of H.264 artefacts on subjective video quality. In *in Proceedings of the Third International Workshop on Video Processing and Quality Metrics, VPQM2007*, 2007. 139

[278] J. A. Nelder and R. Mead. A simplex method for function minimization. *The Computer Journal*, 7(4):308–313, 1965. 140, 147, 165

[279] M. Fairchild. http://rit-mcsl.org/fairchild//hdr.html. 140, 147, 149

[280] M. Narwaria, M. Perreira Da Silva, P. Le Callet, and R. Pépion. Effect of tone mapping operator on visual attention deployment. In *SPIE optics+photonics - Applications of Digital Image Processing XXXV, San Diego*, 2012. 140, 149

[281] L. Krasula, M. Narwaria, K. Fliegel, and P. Le Callet. Rendering of HDR content on LDR displays: An objective approach. In *SPIE Applications of Digital Image Processing XXXVIII*, 2015. (accepted for publication). 143

[282] L. Krasula, K. Fliegel, P. Le Callet, and M. Klíma. Objective evaluation of naturalness, contrast, and colorfulness of tone-mapped images. In *Proc. SPIE 9217, Applications of Digital Image Processing XXXVII*, 2014. 145

[283] L. Krasula, M. Narwaria, K. Fliegel, and P. Le Callet. Influence of HDR reference on observers preference in tone-mapped images evaluation. In *7th International Workshop on Quality of Multimedia Experience (QoMEX)*, 2015. 149

[284] L. Krasula, M. Narwaria, K. Fliegel, and P. Le Callet. Preference of experience in image tone-mapping: Dataset and framework for objective measures comparison. *IEEE Journal on Selected Topics in Signal Processing*. (under review). 149, 159

[285] D. Kane and M. Bertalmio. System gamma as a function of image and monitor dynamic range. *Journal of Vision*, 16(4), 2016. 149

[286] M. Narwaria, C. Mantel, M. Perreira Da Silva, P. Le Callet, and S. Forchhammer. An objective method for high dynamic range source content selection. In *6th Int. Workshop on Quality of Multimedia Experience*, 2014. 149, 150

[287] M. Klíma et al. DEIMOS – an open source image database. *Radioengineering*, 20(4):1016–1023, 2011. 149

[288] EMPA HDR database. http://empamedia.ethz.ch/. 149

[289] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86, 1951. 149

[290] M. Yang. A survey on fuzzy clustering. *Mathematical and Computer Modelling*, 18(11):1–16, 1993. 150

[291] M. Narwaria, M. Perreira Da Silva, and P. Le Callet. *High Dynamic Range Video*, chapter 14. Dual Modulation for LED-Backlit HDR Displays. Elsevier Ltd., 2016. 153

[292] M. Granados, T. Aydın, J. R. Tena, J. F. Lalonde, and C. Theobalt. Contrast use metrics for tone mapping images. In *IEEE International Conference on Computational Photography (ICCP)*, 2015. 158

[293] T. Virtanen, M. Nuutinen, M. Vaahteranoksa, P. Oittinen, and J. Hakkinen. CID2013: A database for evaluating no-reference image quality assessment algorithms. *IEEE Transactions on Image Processing*, 24(1):390–402, 2015. 159

[294] H. Liu and L. Yu. Towards integrating feature selection algorithms for classification and clustering. *IEEE Transactions on Knowledge and Data Engineering*, 17(4):491–502, 2005. 162, 205

[295] M. Nuutinen, T. Virtanen, and P. Oittinen. Image feature subset for predicting the quality of consumer camera images and identifying quality dimensions. *Journal of Electronic Imaging*, 23(6), 2014. 162, 163

[296] P. M. Narendra and K. Fukunaga. A branch and bound algorithm for feature subset selection. *IEEE Transactions on Computers*, 26(9):917–922, 1977. 163

[297] J. Doak. An evaluation of feature selection methods and their application to computer security. Technical report, University of California at Davis, Department of Computer Science, 1992. 163

[298] H. Liu and H. Motoda. *Feature Selection for Knowledge Discovery and Data Mining*. Boston: Kluwer Academic, 1998. 163

[299] G. Brassard and P. Bratley. *Fundamentals of Algorithms*. New Jersey: Prentice Hall, 1996. 163, 164

[300] H. Almuallim and T. G. Dietterich. Learning boolean concepts in the presence of many irrelevant features. *Artificial Intelligence*, 1994. 163

[301] M. Ben-Bassat. *Handbook of Statistics-II*, chapter Pattern Recognition and Reduction of Dimensionality, pages 773–791. North Holland, 1982. 163

[302] M. A. Hall. Correlation-based feature selection for discrete and numeric class machine learning. In *17th International Conference on Machine Learining*, 2000. 163

[303] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk. Frequency-tuned salient region detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 164

[304] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *Advances in neural information processing systems*, pages 545–552, 2006. 164

[305] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998. 164

[306] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2007. 164

[307] X. Hou and L. Zhang. Dynamic visual attention: Searching for coding length increments. In *Advances in Neural Information Processing Systems*, 2008. 164

[308] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell. SUN: A Bayesian framework for saliency using natural statistics. *Journal of Vision*, 8(7):32, 1–20, 2008. 164

[309] D. E. Goldberg. *Genetic Algorithms in Search, Optimization & Machine Learning*. Addison-Wesley, 1989. 165

[310] A. R. Conn, N. I. M. Gould, and Ph. L. Toint. A globally convergent augmented Lagrangian algorithm for optimization with general constraints and simple bounds. *SIAM Journal on Numerical Analysis*, 28(2):545–572, 1991. 165

[311] A. R. Conn, N. I. M. Gould, and Ph. L. Toint. A globally convergent augmented Lagrangian barrier algorithm for optimization with general inequality constraints and simple bounds. *Mathematics of Computation*, 66(217):261–288, 1997. 165

[312] Ch. Audet and J. E. Dennis Jr. Analysis of generalized pattern searches. *SIAM Journal on Optimization*, 13(3):889–903, 2003. 165

[313] R. M. Lewis, A. Shepherd, and V. Torczon. Implementing generating set search methods for linear constraint minimization. *SIAM Journal on Scientific Computing*, 29(6):2507–2530, 2007. 165

List of Tables

List of Figures