Camera based Display Image Quality Assessment

Ping Zhao

Thesis submitted to Gjøvik University College
for the degree of Doctor of Philosophy in Computer Science

2015

Camera based Display Image Quality Assessment

Faculty of Computer Science and Media Technology
Gjøvik University College

Camera based Display Image Quality Assessment / Ping Zhao
Doctoral Dissertations at Gjøvik University College 3-2015
ISBN: 978-82-8340-019-9
ISSN: 1893-1227

The most exciting phrase to hear in science, the one that heralds new discoveries, is not 'Eureka!' but 'That's funny...'

(Isaac Asimov)

Declaration of Authorship

I, Ping Zhao, hereby declare that this thesis and the work presented in it is entirely my own. Where I have consulted the work of others, this is always clearly stated.

Signed:

(Ping Zhao)

Date:

Summary

This thesis presents the outcomes of research carried out by the PhD candidate *Ping Zhao* during 2012 to 2015 in Gjøvik University College. The underlying research was a part of the HyPerCept project, in the program of Strategic Projects for University Colleges, which was funded by *The Research Council of Norway*. The research was engaged under the supervision of *Professor Jon Yngve Hardeberg* and co-supervision of *Associate Professor Marius Pedersen*, from The Norwegian Colour and Visual Computing Laboratory, in the Faculty of Computer Science and Media Technology of Gjøvik University College; as well as the co-supervision of *Associate Professor Jean-Baptiste Thomas*, from The Laboratoire Electronique, Informatique et Image, in the Faculty of Computer Science of Université de Bourgogne.

The main goal of this research was to develop a fast and an inexpensive camera based display image quality assessment framework. Due to the limited time frame, we decided to focus only on projection displays with static images displayed on them. However, the proposed methods were not limited to projection displays, and they were expected to work with other types of displays, such as desktop monitors, laptop screens, smart phone screens, etc., with limited modifications. The primary contributions from this research can be summarized as follows:

1. We proposed a camera based display image quality assessment framework, which was originally designed for projection displays but it can be used for other types of displays with limited modifications.

2. We proposed a method to calibrate the camera in order to eliminate unwanted vignetting artifact, which is mainly introduced by the camera lens.

3. We proposed a method to optimize the camera's exposure with respect to the measured luminance of incident light, so that after the calibration all camera sensors share a common linear response region.

4. We proposed a marker-less and view-independent method to register one captured image with its original at a sub-pixel level, so that we can incorporate existing full reference image quality metrics without modifying them.

5. We identified spatial uniformity, contrast and sharpness as the most important image quality attributes for projection displays, and we used the proposed framework to evaluate the prediction performance of the state-of-the-art image quality metrics regarding these attributes.

The proposed image quality assessment framework is the core contribution of this research. Comparing to conventional image quality assessment approaches, which were largely based on the measurements of colorimeter or spectroradiometer, using camera as the acquisition device has the advantages of quickly recording all displayed pixels in one shot, relatively inexpensive to purchase the instrument. Therefore, the consumption of time and resources for image quality assessment can be largely reduced. We proposed a method to calibrate the camera in order to eliminate unwanted vignetting artifact primarily introduced by the camera lens. We used a hazy sky as a closely uniform light source, and the vignetting mask was generated with respect to the median sensor responses over only a few rotated shots of the same spot on the sky. We also proposed a method to quickly determine whether all camera sensors were sharing a common linear response region. In order to incorporate existing full reference image quality metrics without modifying them, an accurate registration of pairs of pixels between one captured image and its original is required. We proposed a marker-less and view-independent image registration method to solve this problem. The experimental results proved that the proposed method worked well in the viewing conditions with a low ambient light. We further identified spatial uniformity, contrast and sharpness as the most important image quality attributes for projection displays. Subsequently, we used the developed framework to objectively evaluate the prediction performance of the state-of-art image quality metrics regarding these attributes in a robust manner. In this process, the metrics were benchmarked with respect to the correlations between the prediction results and the perceptual ratings collected from subjective experiments. The analysis of the experimental results indicated that our proposed methods were effective and efficient. Subjective experiment is an essential component for image quality assessment; however it can be time and resource consuming, especially in the cases that additional image distortion levels are required to extend the existing subjective experimental results. For this reason, we investigated the possibility of extending subjective experiments with baseline adjustment method, and we found that the method could work well if appropriate strategies were applied. The underlying strategies referred to the best distortion levels to be included in the baseline, as well as the number of them.

Acknowledgments

Image quality assessment can be challenging. It requires both professional knowledge and practical experience in many disciplines physics, optics, color science, human perception, computer vision, image processing, statistics and psychophysics. Fortunately, I collaborated with an extended group of experts, who were always willing to coach, assistant and motivate me to move forward. In addition, many volunteers continuously supported me. Now, it is my pleasure to thank you all for helping me to bring this research into reality.

First, I would like to emphasize my gratitude to my main supervisor *Professor Jon Yngve Hardeberg*, co-supervisors *Associate Professor Marius Pedersen*, and *Associate Professor Jean-Baptiste Thomas*, for your professional guidance to keep my research on the right track, as well as for your brilliant ideas to motivate me to push the boundaries of the state-of-the-art scientific research. Without your countless hours of reviewing and commenting on the papers, we might not be able to have many excellent research outcomes. Meanwhile, I would also like to thank *Professor Ivar Farup* and *Associate Professor Philip John Green*, for the intensive lectures on the topics of image processing and color science respectively. Your professional knowledge and practical experience greatly consolidated my research foundation.

In addition, I would like to thank all other colleagues in The Norwegian Colour and Visual Computing Laboratory for your patience in answering my questions and your participation in the multiple subjective experiments, especially *Ludovic Gustafsson Coppel*, *Osamu Masuda*, *Raju Shrestha*, *Sony George*, *Aditya Suneel Sole*, *Joschua Thomas Simon-Liedtke*, *Radovan Slavuj*, *Hilda Deborah*, *Xingbo Wang* and also, the many students in the master education programs. I would also like to thank *Associate Professor Peter Nussbaum*, for your participation in my subjective experiments, as well as kindly offering me a teaching opportunity on the topic of modern display technologies. I would also like to thank the anonymous reviewers in various publication channels for your time and extensively detailed comments on the papers during the years.

For inspiring study and work environment, I would like to thank my colleagues *Hilde Bakke* for providing me the effective and efficient coordination assistance on my PhD course registrations, to thank *Rachael McCallum* for providing me excellent administrative supports, and to thank *Jingjing Yang* for providing excellent reimbursement support for my frequent travels. Finally, I would like to express my gratitude to my parents for your understanding and patience, and to my wife and daughter for your encouragement and abiding support.

Ping Zhao

Gjøvik
November 4th, 2015

The Publications

The six papers listed below constitute the core research of the present thesis, Further, two publications and their contributions are related to the present research, but are not included in this thesis.

List of Included Papers

Paper A

Ping Zhao, Marius Pedersen, Jon Yngve Hardeberg, and Jean-Baptiste Thomas. "**Camera-Based Measurement of Relative Image Contrast in Projection Displays**." In *4th European Workshop on Visual Information Processing*, 112-17. Paris, France: IEEE, June, 2013.

Paper B

Ping Zhao, Marius Pedersen, Jon Yngve Hardeberg, and Jean-Baptiste Thomas. "**Image Registration for Quality Assessment of Projection Displays**." In *21st International Conference on Image Processing*, 3488-92. Paris, France: IEEE, October, 2014.

Paper C

Ping Zhao, Marius Pedersen, Jean-Baptiste Thomas, and Jon Yngve Hardeberg. "**Perceptual Spatial Uniformity Assessment of Projection Displays with a Calibrated Camera**." In *22nd Color and Imaging Conference*, 159-64. Boston, MA, USA: Society for Imaging Science and Technology, November, 2014.

Paper D

Ping Zhao, Marius Pedersen, Jon Yngve Hardeberg, and Jean-Baptiste Thomas. "**Measuring The Relative Image Contrast of Projection Displays**." *Journal of Imaging Science and Technology* 59, no. 3 (April, 2015): 030404-1 - 030404-13.

Paper E

Ping Zhao, and Marius Pedersen. "**Measuring Perceived Sharpness of Projection Displays with a Calibrated Camera**.". Submitted to *Journal of Visual Communication and Image Representation*.

Paper F

Ping Zhao, and Marius Pedersen. "**Extending Subjective Experiments for Image Quality Assessment with Baseline Adjustments**." In *Image Quality and System Performance XII, Proceedings of 27th Annual Symposium on Electronic Imaging*, edited by Mohamed-Chaker Larabi and Sophie Triantaphillidou, 93960R-1 - 9360R-13. San Francisco, CA, USA: The International Society for Optics and Photonics, February, 2015.

List of Related Papers

Ping Zhao, Yao Cheng, and Marius Pedersen. ″**Objective Assessment of Perceived Sharpness of Projection Displays with a Calibrated Camera**.″ In *8th Colour and Visual Computing Symposium*. Gjøvik, Norway: IEEE, August, 2015 (in press).

Ping Zhao, Marius Pedersen, Jean-Baptiste Thomas, and Jon Yngve Hardeberg. **DIQTool: An Open Source Software Toolbox for Camera based Display Image Quality Assessment**. Submitted to *Image Quality and System Performance XIII, Proceedings of 28th Annual Symposium on Electronic Imaging*. San Francisco, CA, USA: The International Society for Optics and Photonics, February, 2016.

Contents

Part I

Introduction

Chapter 1

Research Introduction

> My definition of an expert in any field is a person who knows enough about what's really going on to be scared.
>
> P. J. PLAUGER

This chapter provides a brief introduction to the research. The motivation, research goals, research questions, research methodology and the outline of this thesis are presented.

1.1 Research Motivation

Image quality is characterized by quantifying and analyzing a set of measurable image quality attributes [44, 88]. The physical properties, such as screen dimension, display resolution, refreshing rate, etc., are associated with specific displays and/or their viewing conditions. They impact the image quality, but they are unlikely to vary in a typical image quality assessment cycle. In this research, we focused on content dependent image quality attributes, such as brightness, contrast, colors, sharpness and artifacts; because these attributes are essentially terms of visual perception [130], and they greatly impact the visual experience. The existing research characterizing displays, such as CRT monitors [17, 57, 56], and LCD screens [17, 57, 45, 56], were presented. The industrial communities have also established many international standards, such as IEC9241-304 [75], IEC9241-305 [76], IEC9241-307 [77], IEC61966-3 [70], IEC61966-4 [71], IEC61966-5 [74], IEC61966-6 [73], IDMS 1.03 [69], SPWG 3.8 [166], and TCO 6.0 [169]. The research and standards mentioned above were largely based on the measurements of colorimeters and/or spectroradiometers. The spectroradiometers were primarily designed to quantify the average physical response over a small spot area of displayed patches at discrete spatial locations. The measurements are known to be accurate, but it may take a long time to collect a large number of samples; especially under low light conditions, which are typical for projection displays. In addition, spectroradiometers are relatively expensive to purchase and they are likely to be unavailable in the real practice of image quality assessment. In contrast, we can take full advantage of a digital still camera to record all displayed pixels on the screen in one shot [147, 62]. In this case, using camera as the acquisition device to measure the relative image quality attributes can be a fast, inexpensive complementary to the spectroradiometer based approach. However, cameras need to be carefully calibrated in order to eliminate unwanted artifacts, which are mainly introduced by the camera's optical and electronic subsystems. Meanwhile, the acquisition settings should be optimized as well. In addition, the work-flow of processing the raw camera sensor data and evaluating the image quality with respect to the selected image quality metrics as well as the subjective ratings should be defined. Thus, a camera based display image quality assessment framework is required; especially in the cases of incorporating full reference image quality metrics, which require an exact mapping between each pair of pixels in the captured image and its original respectively. In this context, the preservation of geometrical order, as well as the intensity and chromaticity relationships between two consecutive pixels on the displays, should be maximized. A few full reference metric based image quality assessment frameworks with similar ideas with scanners have been proposed in the research domain of printing [134, 181]. In these cases, shift invariant features which were highly dependent on the image content was adopted, or a modified control point matching method was used. In the real practice of projection displays, these methods were not suitable; it was not only because both the type and amount of spatial distortion in the captured images might vary with respect to the relative position and orientation between the projector, screen and camera, but also because in many cases people want to achieve a view-dependent image quality optimization. Thus, a novel and more flexible image registration method should be proposed.

1.2 Research Goals

The first goal of this research was to develop a fast and an inexpensive camera based display image quality assessment framework. In order to increase the generalization and applicability of the proposed method, the framework should be independent from specific type of image quality metrics. The existing full reference, reduced reference and no reference based image quality metrics without any modification should be incorporated into the framework. In order to maximize the reliability, validity, and robustness of the image quality assessment, the work-flow of the proposed framework should be proceeded robustly.

The secondary goal was to implement the proposed framework for projection displays. The main motivation behind is that existing related research was prettylimited. The majority of the research conducted was based on the measurements of spectroradiometer. To our best knowledge, we were the first to propose a systematic approach of evaluating image quality of projection displays with a digital still camera. By applying the proposed framework in the field, we were able to observe, identify, and recognize the potential research problems in actions. We might not merely come up with corresponding solutions, but also simultaneously improve the framework design based on the experience learned in an iterative manner.

The third goal of this research was to identify the most important image quality attributes for projection displays, and evaluate the prediction performance of state-of-the-art image quality metrics regarding these attributes with the developed framework. We identified spatial uniformity, contrast and sharpness as the most important image quality attributes for projection displays. The objective evaluation results were correlated with the perceptual ratings collected from subjective experiments. One goal was to rank all metrics intensively within each one of the full reference, reduced reference, and no reference categories. Another goal was to have a lateral comparison between different metric categories, so that the category of metrics with the highest prediction performance can be identified with respect to the statistical analysis of experimental results.

1.3 Research Questions

With respect to the research goals described above, we initiated this research by asking several research questions:

1. What is the basic work-flow using a digital still camera to perform the image quality assessment for projection displays?

2. How to calibrate cameras in order to eliminate unwanted imaging artifacts and optimize the acquisition settings in the purpose of image quality assessment?

3. How do we incorporate existing full reference image quality metrics into the proposed framework without modifying them?

4. What are the most important image quality attributes for projection displays and how do we evaluate them by using the developed framework?

5. What are the best state-of-the-art image quality metrics? Is there a clear advantage using full reference image quality metrics over reduced reference and no reference metrics?

The first question was related to the design of the image quality assessment framework. The design involved identifying the key operational components and organizing them in a well defined work-flow. This created a starting point for the rest of the research. The second question was related to the fact that digital still cameras typically have many manual settings, such as aperture, ISO, shutter speed, etc. In order to maximize the preservation of captured image quality while minimizing the influence of artifacts introduced by the camera system, we could not simply set everything to auto and hope the camera would do its best. Instead, we should follow the well established international standards, and propose novel ideas to solve the challenges. The third question is related to the research challenge of incorporating existing full reference image quality metrics without modifying them. The full reference image quality metrics require exact registration of pairs of pixels in one captured image and its original respectively. The registration method should be marker-less and view-independent in order to maximize the flexibility and robustness of the proposed framework. The fourth question is related to using the developed framework to evaluate the most important image quality attributes. For different displays, the selection and evaluation priorities and criteria of image quality attributes can be different. The main purpose was to confirm the validity, reliability and robustness of the proposed framework in real projection environments. The fifth question was related to the use of the proposed framework to evaluate the prediction performance of the state-of-the-art image quality metrics with respect to their correlation with the perceptual ratings collected from subjective experiments. In this process, we were able to benchmark and rank the metrics vertically in one of the full reference, reduced reference, and no reference categories, as well as to compare their performance laterally between different metric categories.

1.4 Research Methodology

First, we perform a comprehensive survey of literature regarding image quality assessment involving both acquisition and assessment procedures. The purpose was to understand the typical assessment targets, acquisition devices, calibration procedures, experimental setup, viewing conditions, test charts, classifications and evaluation methods of image quality attributes, international standards, data analysis methods, and common practice. In the survey, it was found that the majority of existing research concentrated on the domains of printing and desktop monitors. There were also a few works related to projection displays, but the efforts were quite limited and none of them use cameras to acquire projections. So, there was no experience that we could learn from the past. For this reason, our research should be experiment oriented.

Then, we had to setup all equipment in the field, and performed image quality assessment accordingly. In this process, we were expected to confront many practical issues, for which we had to come up with corresponding solutions. We did not only design the image quality assessment framework for general displays, but also implemented, tested, and improved it specifically for projection displays. In order to achieve this goal, we identified three typical viewing conditions of projection displays, and decided to setup our laboratory to simulate the home-theater like darkroom environment. By simulating the environment of real projection applications and conducting experiments in the field, we could actively study the whole workflow of using a camera to perform image quality assessment in a quantitative fashion. Meanwhile, we decided to implement the framework specifically for projection displays, and used it to evaluate the state-of-the-art image quality metrics regarding the most important image quality attributes. With respect to the studies of the evaluation outcomes, we could validate our framework and benchmark the image quality metrics in a quantitative manner.

Both of the objective and subjective results were studied with respect to the statistical analysis. In this context, we performed both descriptive and analytical studies to explore the absolute values and the distribution of numerical data. In this process, we could identify and recognize the potential challenges in the research, and decide how to engage them via further extended experiments or simulations. In this process, we used and modified the framework iteratively. Then, the framewjonork was gradually tested and improved with respect to the proposed novel ideas in the calibration and image quality assessment procedures. The assessment was actually performed from both subjective and objective perspectives. The motivation was to correlate the objective results with subjective results in order to evaluate selected image quality metrics, by assuming that the subjective results approximate the actual perception of an overall average observer. In this context, the statistics based psychometric rating and scaling procedures were incorporated to minimize the impact of the variance of judgment criteria between different observers.

In a summary, we had a survey to obtain the insight of existing knowledge, methodology and practice regarding image quality assessment for general displays, proposed an image quality assessment framework based on the information obtained from the survey, implemented the framework specifically for projection displays, used the framework to conduct objective and subjective experiments in order to evaluate the state-of-art image quality metrics regarding the most important image quality attributes for projection displays, and further improved the framework with respect to both quantitative and qualitative experimental outcomes in an iterative manner.

1.5 Outline of Thesis

This thesis is intended to provide the potential readers with the understanding needed to calibrate a digital still camera and use it to assess the image quality of displays, by first introducing how to utilize both quantitative evaluation and psychophysical experiment to engage the research, and then presenting the proposed image quality assessment workflow, methodologies and related experimental results. Therefore, this thesis is divided into two parts and eleven chapters.

PART I includes Chapter 1 to Chapter 5. The motivation, goals, questions, and methodologies of this research project were presented in Chapter 1. An overview of the history and definition of image quality, as well as the existing research methodologies regarding objective and subjective experiments, were presented in Chapter 2. The research outcomes and contributions from individual publication were summarized in Chapter 3. The discussions related to individual publication and the relationship between them were presented in Chapter 4. The conclusion and perspectives of this thesis were given in Chapter 5.

PART II includes Chapter 6 to 11. It presents the research outcomes as the main contributions of this thesis via all publications.

PART III presents the experimental setup, the acquisition devices, test charts, and projectors that we used.

Chapter 2

Display Image Quality Assessment

> The use of thesis-writing is to train the mind, or to prove that the mind has been trained; the former purpose is, I trust, promoted, the evidences of the latter are scanty and occasional.
>
> SIR THOMAS CLIFFORD ALLBUTT

In this thesis, a camera based display image quality assessment is presented. Prior knowledge of fundamental principles of image quality is mandatory for fully understanding the presented work. Hence, the goals of this chapter are to present an overview of the underlying research area and provide a concise introduction to the image quality definitions, image quality attributes, measurement instruments, objective assessment methods, and subjective experimental methods.

2.1 What is Image Quality

Image quality is not a new term. The earliest history regarding ”the quality of an image” can be traced back to the beginning of 17th century, when optical instruments, the telescope and the microscope were invented [43]. At that moment, image quality was no more than an optical concept associated to the acquisition instruments. In recent years, thanks to the rapid advancement of imaging technologies and the tremendous growth in the use of digital media, the scope of image quality has been greatly extended to cover the entire imaging pipeline. For display image quality assessment, it is important to understand what we going to measure before we actually perform the assessment. Therefore, a clear definition to image quality is required. However, there is no universal and comprehensive definition yet. This is mainly because the term image quality may have significantly different meanings to people from different perspectives with different concerns. In the existing literature, several definitions of image quality have been proposed:

- Jacobson [82] defined image quality as the subjective impression found in the mind of the observer relating to the degree of excellence exhibited by an image.
- Engeldrum [43] interpreted image quality as the integrated set of perceptions of the overall degree of excellence of the image. In his theory of Image Quality Circle, the concept of image quality was associated with customer perceptual rating, customer perception, physical image parameters, and technology variables, of which the image quality assessment components formed a closed loop (Figure 2.1).
- Janssen [83] followed visuo-cognitive processes to define image quality as the degree to which the image was both useful and natural. In this case, the usefulness of an image was defined as the precision of the internal representation of the image; and the naturalness of an image was defined as the degree of correspondence between the internal representation of the image and knowledge of reality as stored in memory.

- Ridder et al. [39] divided image quality into three categories: fidelity, usefulness, and naturalness. Among them, fidelity was referred to the reproduction accuracy of an observed image in comparison to the original, which was assumed to have perfect quality. Usefulness indicated image suitability for the designed task. Naturalness was defined as a match between a reproduced image and the mental impression of an observer, affected by memory traces.
- Fairchild [46] defined image quality as the perceptible visual differences from some ideal and the magnitude of such differences.
- Yendrikhovskij [199] suggested that image quality was understood as the subjective impression of how well image content was rendered or reproduced.
- Keelan [88] defined image quality as an impression of its merit or excellence, as perceived by an observer neither associated with the act of photography, nor closely involved with the subject matter depicted.
- The International Imaging Industry Association [31] defined image quality as the perceptually weighted combination of all visually significant attributes of an image when it was considered in its marketplace or application.

Based on the proposals presented above, it is not difficult to see that image quality is commonly defined from the subjective perspective. This is mainly because humans are the ultimate visual information interpreter. The visual stimuli are acquired by the human visual system, and the corresponding signals are further decomposed and forwarded along millions of neuron pathwways in parallel to the human brain in order to interpret. Human interpretations are fuzzy in nature. One observer may make his/her own independent judgment regarding the the quality of an image even without knowing what image quality actually is. From this point of view, image quality is a subjective and relative term, because one observer may have significantly different perception criteria regarding one or more specific image quality attributes. The underlying attributes are terms of visual perception in the current concerns. These concerns may vary with respect to the imaging applications and their related contexts. For example, one picture about an extreme sports man climbing rocks was taken with a digital still camera; in this case, the motion blur for the movements, the details of the person's struggling face, and the image resolution for magazine level printing can be the main concerns. The vivid colors of the background scene may not be because the trees and sky might be completely de-focused in order to feature the climber in the foreground. Due to the probabilistic nature of the human brain and its high context dependence, ordinary people actually refer image quality as the overall quality of an image reproduction with respect to his/her own perceptual opinion regarding a set of weighted image quality attributes, while scientific researchers may refer image quality as the mean perceptual opinions with respect to one or more visually significant attributes in the concern of current application. In the qualitative approach, image quality can be defined explicitly with well written text statements, but the corresponding computation remains an open question for research discussions.

In an objective manner, image quality can be quantized as one or more numbers by applying image quality metrics to the captured images with or without referring to their originals, which are assumed to have perfect image quality. Depending on the availability of the reference images, the metrics can be broadly divided into three groups: full reference, reduced reference and no reference. In the full reference approach, image quality is defined as the magnitude of quality degradation from the original image to its reproductions with respect to one or more specific image quality attributes, and all pairs of pixels in the two images are used. In contrast, the reduced reference based metrics count only image features. The no reference based metrics determine the quality of an image blindly, absolutely without its original. In all these cases, the scales of and the interpretations to the numeric

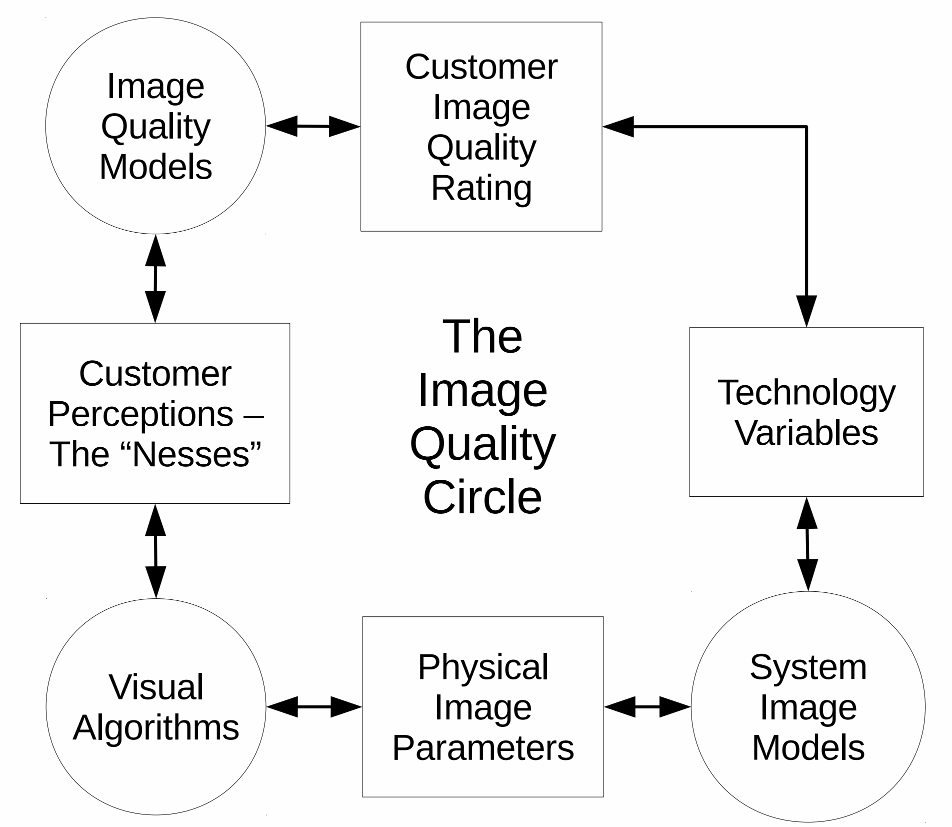


Figure 2.1: The complete Image Quality Circle with the three connecting links: System/Image Models, Visual Algorithms, and Image Quality Models. [43]

quality indication are totally different. In other words, image quality is defined implicitly with numbers in independent numeric spaces of such image quality metrics. Thus, the quality of an image is the predicted image quality and it needs to correlate well with the actual image quality in order to be claimed to be valid. In this case, the subjective image quality forms an approximation of the ground truth.

With respect to the discussions above, image quality should be defined as the subjective quality, which is an approximation of the actual image quality. The subjective image quality stands for the mean perceptual opinions obtained from the statistical regression analysis of subjective ratings, which are sampled from a specific human population with respect to a few visually significant attributes in the concern of current imaging application. This definition of image quality is used in the rest of this research.

2.2 Image Quality Attributes

Image quality requires a systematic assessment approach from both subjective and objective perspectives. In both cases, image quality is characterized based on a set of image quality attributes, which are terms of human perceptions [130]. The ultimate goal is to correlate the objective assessment results with the subjective assessment results, so that we can eventually eliminate the demand of observers. Generally, image quality can be separated into two levels: low-level/concrete attributes, which can be measured directly with instruments or estimated based on the measurement results; high-level/abstract attributes, which are abstraction of low-level attributes but they are strongly associated with observer's expertise and experience regarding the underlying image quality attributes, such as naturalness and usefulness. The difference between the two levels is not limited to the abstraction level, but it seems that the importance of high level attributes lies in their ability to inform the observer of the meaning of low-level attributes for the general quality [99].

Among the low-level attributes, physical properties such as screen dimension, display resolution, refreshing rate, have impacts on the perceived image quality, but in a typical work-flow of image quality assessment they can be assumed to be constants, since they are independent from the image content and normally do not vary over time. In this research, we used cameras as the acquisition devices. For this reason, we only focused on the assessment of low-level perceptual image quality attributes, which were image content dependent. There is a lot of research related to characterizing electronic devices based on these image quality attributes. These devices were, but not limited to, printers [130, 58, 133, 136, 135], CRT monitors [57, 17], LCD/LED monitors [57, 17, 45, 170, 67], projection displays [176, 109, 177, 114, 167]. According to these studies, the image quality attributes can be generally divided into five groups: lightness, contrast, colorfulness, sharpness, noises. Each group may include several sub-groups with respect to various classification criteria. Most of the image quality attributes were studied with associations to many other image quality attributes.

2.2.1 Lightness

Lightness stands for the perceived intensity of light coming from the image itself, rather than any property of the portrayed scene [5]. It should be used only for non-quantitative reference to physiological sensations and perception of light, so it ranges from ”light” to ”dark” [130]. Lightness is a close concept to brightness, which is also a perceptual image quality attribute. In this case, lightness is defined as the brightness of an area relative to the brightness of a similarly illuminated area that appears white or highly transmitting [105]. Lightness has a significant impact on the perceptual experience [47, 5]. The relationship between relative brightness and saturation to lightness and chroma of a surface, for a single-hue triangle in a hue-chroma-lightness space can be presented in Figure 2.2.

2.2.2 Colorfulness

Color is a human sensation and it represents the perception of incidental light acquired by the human visual system. The accuracy of color reproduction in an image can be represented by the color distance between the image reproduction and its reference in a specific color space. In most cases, when people use the term of color, they actually exclude lightness and refer the term of color to colorfulness, which is a perceptual attribute that covers the aspects of hue, saturation and gamut [197, 167, 158]. The relationship between these aspects can be demonstrated with the Munsell color system (Figure reffig:munsell). Thus, colorfulness can be defined as the attribute of a visual perception according to which an area appears to exhibit more of less of its hue. In this context, chroma is defined as the colorfulness of an area judged in proportion to the brightness of a similarly illuminated area that appears to be white or highly transmitting, while saturation is defined as the colorfulness of an area judged in proportion to its brightness [68]. In addition, hue is defined as an attribute of a visual perception according to which an area appears to be similar to one of the colors, red, yellow, green, and blue, or to a combination of adjacent pairs of these colors considered in a closed ring [15]. For electronic devices, such as scanner, printer, and displays, the color gamut stands for the entire range of colors that the device can reproduce accurately in a specific color space. The color gamut is expected to be as large as possible, but none of the known devices can reproduce all colors [155]. Nevertheless, the most pleasing color might not necessarily be the most accurate color [45].

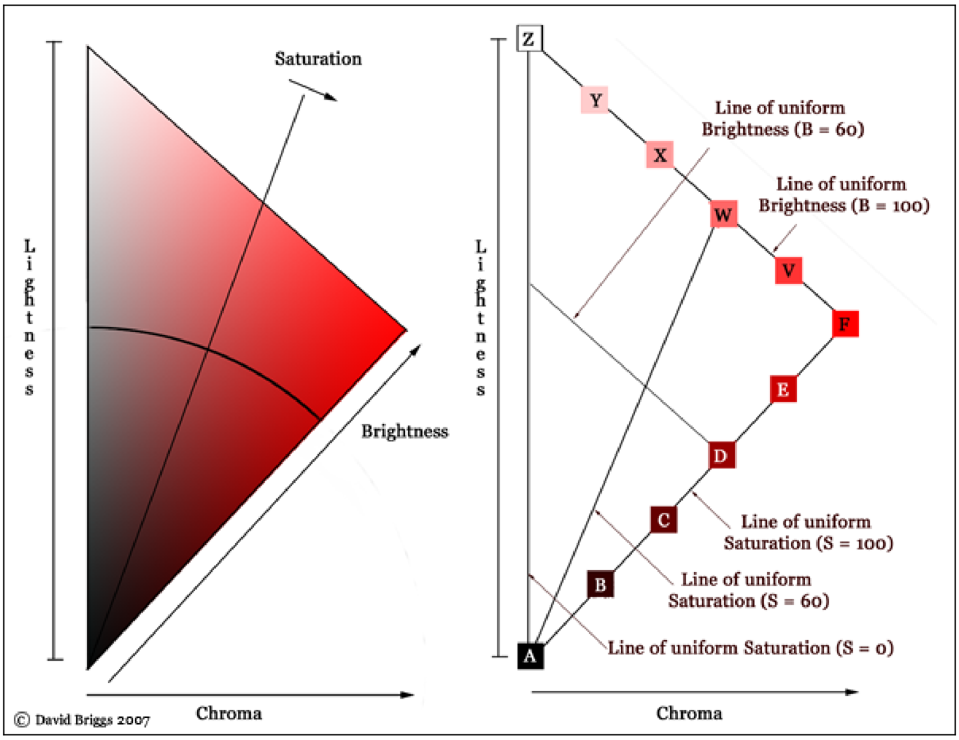

Saturation
Lightness
Brightness
© David Briggs 2007
Chroma
Z
Y
X
W
V
F
E
D
C
B
A
Line of uniform Brightness (B = 60)
Line of uniform Brightness (B = 100)
Line of uniform Saturation (S = 100)
Line of uniform Saturation (S = 60)
Line of uniform Saturation (S = 0)
Lightness
Chroma

Figure 2.2: The conceptual relationship of lightness, absolute brightness, relative brightness, chroma and saturation. [19]

2.2.3 Contrast

In most literature, contrast for an image was defined as a measure of the luminance variation relative to the average luminance in the surrounding region, however no standard definition to contrast in a complex scene has been given. One most widely adopted definition for measuring image contrast is Michelson formula [120]:

$$C_M = \frac{I_{max} - I_{min}}{I_{max} + I_{min}}, \tag{2.1}$$

where I_{max} and I_{min} stand for the maximum and minimum value of lighting respectively. Another widely adopted contrast definition is the Weber fraction specially defined for simple test patterns [140]:

$$C_W = \frac{I_s - I_b}{I_b}, \tag{2.2}$$

where I_s and I_b stand for the foreground and background lightness respectively. In the research domain of tone reproduction, contrast is defined as the rate of change of the relative luminance of image elements of a reproduction, as a function of the relative luminance of the same image elements of the original image; on log-log coordinates, contrast is the slope of the relationship between the reproduction and original [105]. In the cases for color complex scene, we may define contrast approximately as a measurement of the luminance and/or chromatic variations in one region relative to the average variance in the surrounding region in the same scene. There are two important aspects in the contrast research. One of them is related to the contrast sensitivity function (Figure 2.4). DeValois et al. [40] indicated that the contrast sensitivity of human visual system followed a certain curve with respect to the current average luminance level, and the spatial frequency of luminance variations. Thus, an optimization for the image content or size can be achieved accordingly by

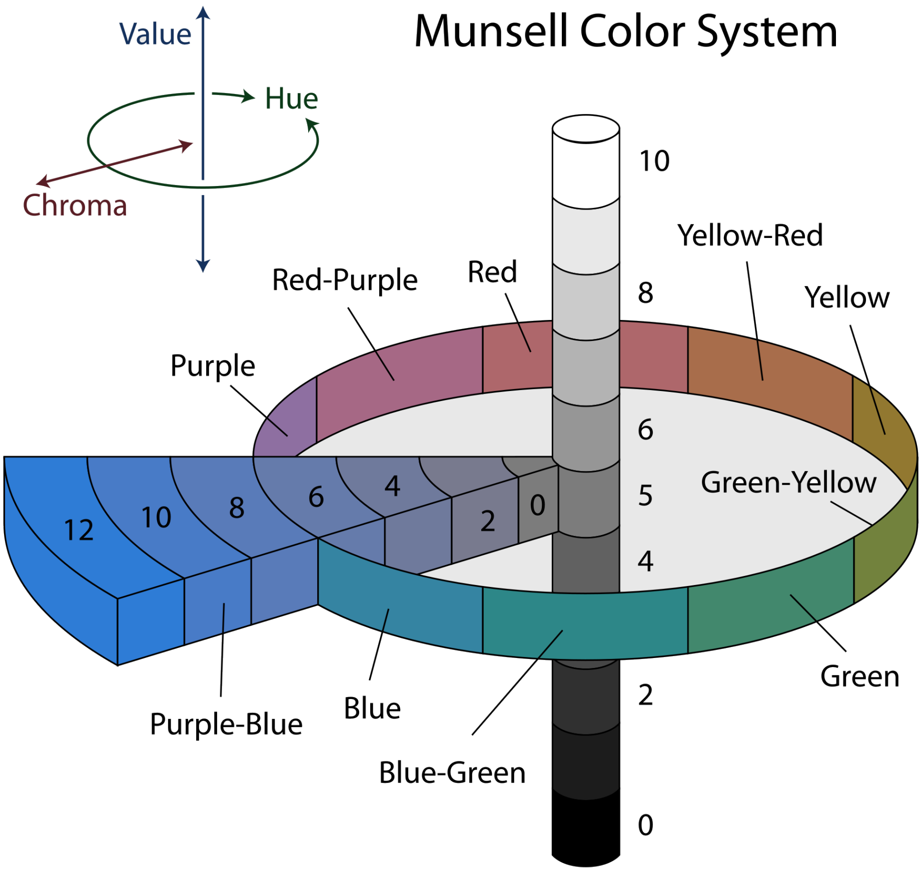


Figure 2.3: Munsell Color System [47]

keeping the high spatial frequency components in the images, while the low frequency ones are being eliminated [108, 138]. The other research aspect of contrast is related to contrast masking, which is a visual phenomenon of human visual system. The term is used commonly to refer to any de-structive interaction or interference among transient stimuli that are closely coupled in space or time [97]. Thus, the masked signal shows different visual effect under the different contrast masking signal [49]. This effect is modeled either with a threshold elevation image, or with a contrast transducer function calculated from the masking curve of contrast discrimination experiments, given that the image is decomposed into the appropriate spatial frequency bands [33].

2.2.4 Sharpness

Sharpness is an attribute defining how abrupt the boundaries are between different tones and colors [85, 12, 191]. It is commonly recognized to be an important image quality attribute for perceptual evaluation despite the technology used, and it is closely associated with other attributes, such as lightness, contrast, and blur. Since sharpness defines the amount of details the human can observe in image reproductions at a certain distance, it is commonly referred to as the counterpart of blur. The human visual system has a remarkable capability to detect image blur without seeing the original image, but unfortunately the underlying mechanism is not well understood [63]. One way to determine sharpness

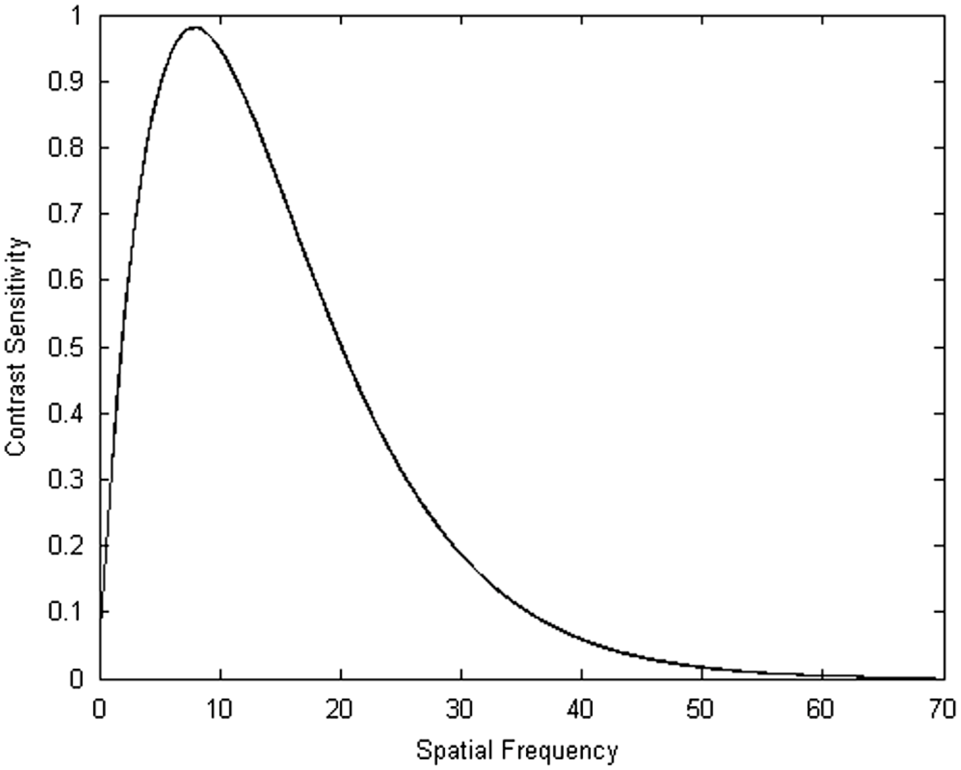


Figure 2.4: An illustration of contrast sensitivity function [29]

is measuring the rise distance of the slant edges, or calculate the density of line pairs with increasing spatial frequency, or do the corresponding analysis in the frequency domain, where frequency is measured in cycles or in line pairs per distance (millimeters, inches, pixels or degree). Specifically, the International Organization for Standardization defined ISO 12233 to standardize the procedure of measuring the resolution and spatial frequency responses (Figure 2.5) of camera lens with a special test chart [175].The existing research regarding sharpness is largely focused on the design and evaluation of reduced reference based and no reference based image quality metrics.

2.2.5 Aesthetic

Aesthetic properties related to the composition of the image (e.g. Rule of Thirds and Visually Weight Balance [87], see Figure 2.6), the photographic techniques (e.g. macro), the use of colors and light, and the pleasantness of look-and-feel are highly subjective [36, 84, 116]. In most cases, when ordinary people talk about the image quality of a picture, they actually refer to the aesthetic attribute. The corresponding assessment outcomes strongly depend on the professional knowledge and practical experience related to photography, painting and other art forms. In conventional approaches, the aesthetic properties were evaluated based on hand-crafted visual descriptors to mimic the photographic rules [116]. Aesthetic properties were largely used in real-time image retrieval systems, so the corresponding research focused on the optimization of image feature extraction, descriptor generalization, and minimizing the computational cost. In recent years, generic descriptors (e.g. Bag-of-Visual-Words [34], Fisher Vector [81]) were proposed and implemented based on support vector machines to learn the distribution of local statistics in the images. The image quality evaluation performance of these methods depends on the selection and use of training data and methods, so the actual outcomes might not be deterministic. Unlike other image

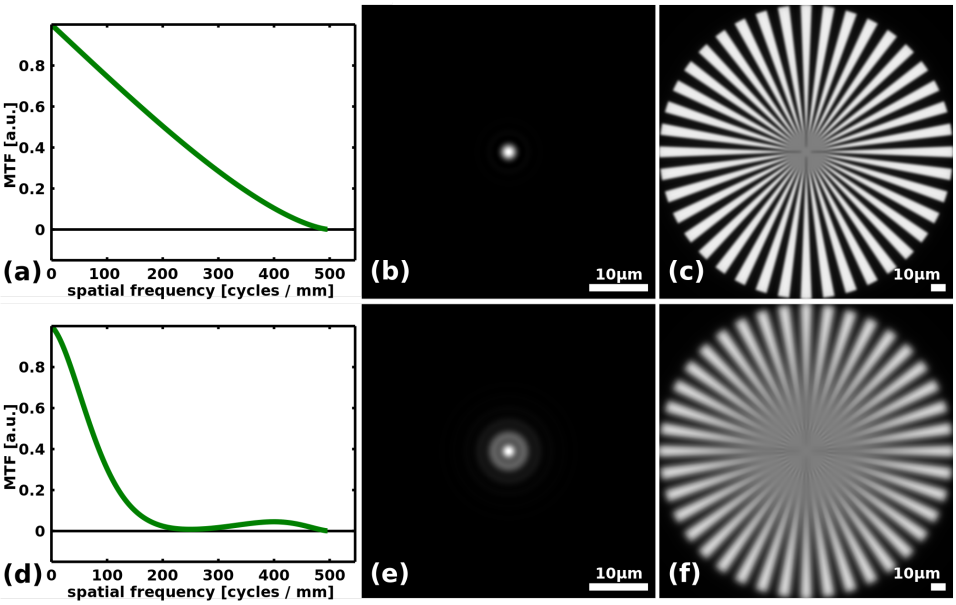


Figure 2.5: An example of spatial frequency response curves corresponding to different levels of captured sharpness [4]

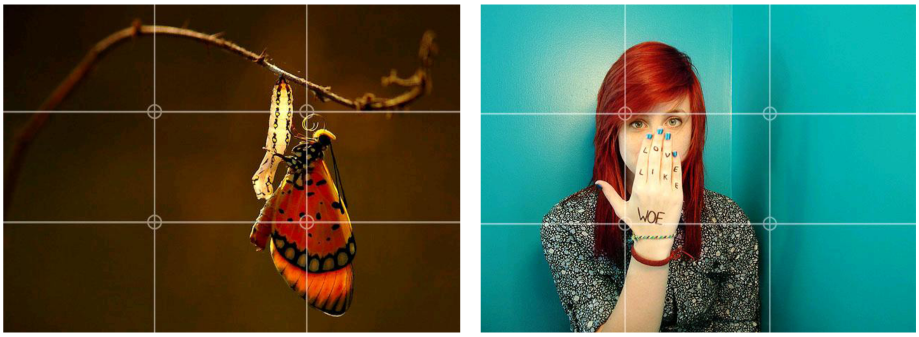


Figure 2.6: Two examples of object composition in photography according to the Rule of Thirds [115]

quality attributes, the research related to aesthetic attributes assessment is more imaging application oriented. In most cases, the attributes were largely used for either image classification or quality ratings. For example, Li el al. [101] designed a group of methods to extract features to represent both the global and local characteristics of a pointing, and correlate them with perceptual opinions. Surová et al. [180] proposed a method incorporating spatial pattern, crown condition, percent crown cover, and a tree mortality index as aesthetic features to assess the quality of forest area images captured with a false color infrared aerial photographs. Li et al. [100] proposed a framework to evaluate the aesthetic quality of people faces by incorporating both perceptual features and social relationship features. Datta et al. [37] designed an online real-time system for accepting uploaded photographs and perform both classification and quality rating simultaneously.

2.2.6 Noises

Noise, such as speckles, spikes, reseals, missing data, marks, blemishes, banding and abnormalities, are created either in expectation or unexpectedly during the processes of acquisition, transmission, and processing of image data [90, 9]. Similarly, artifact is a range of errors in the perception or representation of any visual information introduced to an image in the processes, such as optical acquisition, digital sampling, image compressing and signal processing [182, 157]. The boundary between artifact and noise is fuzzy and subjective. In this research, we think that artifact is a part of noise. Common artifacts were linked to glossary items like lens distortion [188], reflection [122], blooming [193], chromatic aberrations [161], moire pattern [210], jaggies [201], ringing [198] ghosting [14], blocking [202] and so on.

2.2.7 Relationships between Image Quality Attributes

Many types of perceptual image quality attributes were introduced in the previous sections, however they are not completely independent from each other. One image quality attribute may have connections to one or more other attributes. For example, the estimation of lightness is based on the measurements of luminance, which form a foundation for the studies of many other image quality attributes. For example, in the study of perceptual contrast of projection displays, Majumder et al. [110] emphasized that luminance is more important for perception than chrominance. In the White's illusion phenomenon, the relationship between the lightness of two gray regions was revealed to be the opposite of what is predicted by local edge ratios or contrasts [189, 152]. In other studies [151, 24, 25], lightness was also integrated into the computation of image contrast. In the study of digital printing, the banding and contouring artifacts were found to have connections with lightness [30, 93]. Ridder [38] studied the naturalness of images with respect to the saturation and lightness variations. It was found that the difference between naturalness and quality diminished with decreasing lightness. In addition, the evaluation of contrast attribute has a strong connection to the measurement of lightness, as well as colorfulness, sharpness/blur, and artifacts as well. Several studies [11, 53, 8, 107] regarding contrast sensitivity of human eye and its effects on image quality were presented, while other research [13, 49, 196] focused on the modeling of contrast masking. In addition, several studies [54, 148] for determining structural similarity or degradation based on contrast measurements were presented. In these contrast studies focus on modelings, both luminance and chrominance attributes were used.

Regarding the research of colors, a huge amount of effort has been expended. One good example is the research related to color appearance modeling [92, 48, 123]. One color appearance model includes predictors of at least the relative color appearance attributes, such as lightness, chroma and hue [47], and it can be used to predict image quality. Such a model addresses the perspectives of presented stimuli, viewing condition, colorimetry, color appearance phenomena, and chromatic adaption etc. Color appearance models incorporate chromatic adaptations as well as the predictors of brightness and colorfulness. They also adopt the color adaptation model as a module in the initial step, so this module can be selected or replaced in preference. In the post-adaption step, adapted tristimulus data and other additional data, like absolute luminance level, colorimetric data on the proximal field, background and surround, are combined to provide higher level signals in order to produce predictors of color appearance attributes.

2.2.8 Summary

Image quality can be characterized based on perceptual attributes from various perspectives, however the selection of the most important image quality attributes has different priorities in different research domains. For digital printing, the research [130, 58] suggested that lightness, contrast, sharpness, artifacts, colors, and physical attributes are all important. Lindberg [103] evaluated many image quality attributes, such as color gamut, sharpness, contrast, tone quality, detail highlights, detail shadow, gloss level, gloss variation, color shift, patchiness, mottle, and ordered noise. Among them, the print mottle and color gamut were found to account for most of the variations with respect to the factor analysis. Johnson [85] specially remarked colorfulness, sharpness, and contrast for printing. For mobile displays, Gong et al. [59] suggested that clearness was the most important, followed by naturalness, sharpness, colorfulness, contrast and brightness. In this context, clearness is a high-level attribute associated with other low-level attributes; however, the actual numeric relationship was not given in the research. In contrary, Kim et al. [91] emphasized that naturalness had a high priority than clearness related to overall image quality. For stereo displays, You et al. [200] and Lehtimaki et al. [98] pointed out that noise, sharpness and perceived depth are priorities for stereoscopic imaging. Thomas et al. [176] and Strand et al. [167] remarked lightness and colorfulness for projection displays, while Majumder et al. [108, 94] indicated that lightness is more important than the colorfulness. In this research, with respect to the literature, we can see that the image quality has strong connections to contrast and sharpness image quality attributes despite the actual display technology used. In addition, it was known that for projection displays spatial uniformity is an important image quality attribute [112, 176, 110]. Hence, in this research, we pay special attentions to contrast, sharpness and spatial uniformity by utilizing our proposed image quality assessment framework.

2.3 Objective Assessment

The first step of an image quality assessment is the image acquisition. In this case, we use one or more measurement instruments to acquire the physical responses of image reproductions on the displays. Subsequently, we determine the image quality of these displays by applying the corresponding metrics to the captured images with or without referring to their originals. In this case, the expected assessment outcome is either an image quality score or a distortion map illustrating the image quality degradation. Many instrument options are available for measuring the physical responses of image reproductions in various ways. It is important to first understand what types of acquisition instruments are available, and what procedures we should follow in order to use them. In this section, we briefly describe the most frequently used instruments for the image quality assessment.

2.3.1 Radiometer

A radiometer is an electronic device for measuring the intensity of radiant energy at a specific spot by non-contact means. In most cases, the radiometers employ only single photocell sensors to detect the emitted radiation, and it is common to incorporate an optical filter with the radiometers in order to narrow the spectrum band of the measurement interests. The optical filtering offers an adaptable and cost effective solution to the spectral measurement. The radiometers are normally used to measure either irradiance or radiance (Figure 2.7). In the latter case, the radiation of emission from a specific light source is being quantified. In addition, if the level of exposure is required, then the integrated irradiance measurement follows. Radiometers are commonly used to quantify the light which outside the visible spectrum. For example, ultraviolet light which is widely used in the industry for various applications, such as curing of photo-resists in semiconductor manufacturing, curing of emulsions for printing or plate-making, and color-fastness testing. In these cases, either radiance or irradiance measurement is conducted to quantify the range and peak of the wavelength. The radiometers are also commonly known as radiation thermometers because they can be used to measure the infrared energy of radiation emitted from the material surfaces with respect to their thermal energies.

2.3.2 Photometer

In contrast to radiometers, a photometer is an optical instrument for measuring the luminance and illuminance of visible light, specifically to compare the relative intensities of the

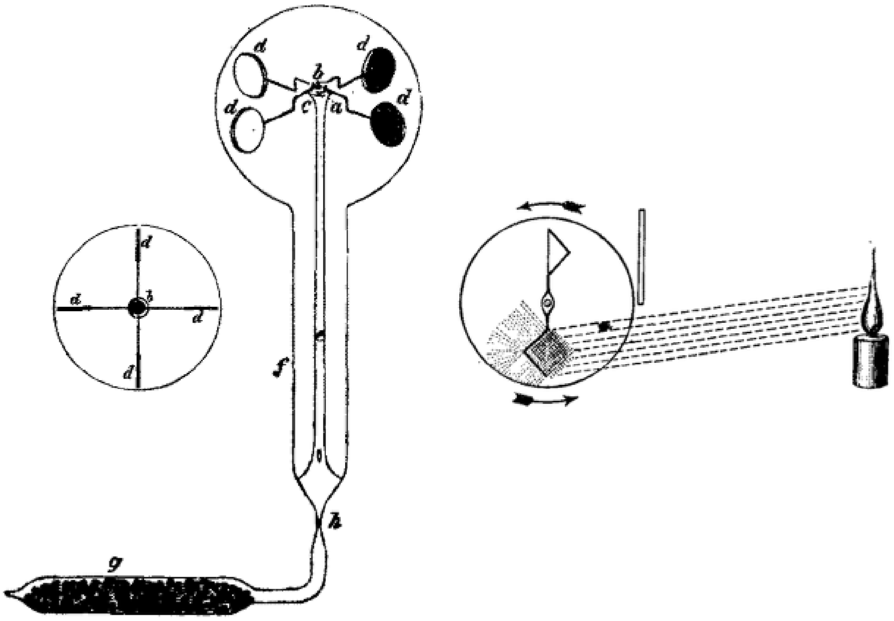


Figure 2.7: The principle of the original Crookes radiometer [26, 192].

light emitted from different sources. The photometers use luminous flux and luminous intensity meters to measure the light. However, in real practice, the existing meters might not available to meet the specific geometric requirements for the light measurement, and they have to be customized by the manufacturers. The photometers are required to have spectral responsivity to the light as a CIE standard observer (Figure 2.8), by following the CIE $V(\lambda)$ function [153]. The function describes the luminous sensitivity of a human eye in photonic conditions. Most modern photometers incorporate silicon photo-diodes with optical filters placed in front of the sensors. In these cases, the transmission of the filters and the spectral response of the sensors can be combined to approximate CIE $V(\lambda)$ function. The measurement quality of one photometer is determined with respect to the errors between the spectral responsivity of the photometer and the actual spectral power distribution of the known light source being measured. In order to quantify the errors, CIE committee defined the quality factor f_1 in order to measure the broadband light sources without spectral mismatch correction [173]. The quality factors have been used by the photometer manufacturers and the lighting industry for years, but no official methods for determining the uncertainties of the quality factors have been published [141]. For the image quality assessment of displays, photometers were commonly used to measure or calibrate the luminance outputs of displays, especially in the research domain of medical imaging of which the gray-scale needs to be very accurate [187, 195, 95, 6].

2.3.3 Colorimeter

A colorimeter is a measurement instrument which applies three or more color filters to the incident light, and it measures one or more of the following photometric properties: luminance, illuminance, luminous intensity, luminous flux, and chromacity. The spectral sensitivity of the filters also need to match the CIE tristimulus color matching functions in order to emulate the human visual system. Therefore, the colorimeters can be used in the scenes for which the photometers are required. A part of the incident light is expected to be absorbed by the light filters, thus a lower light intensity strikes the photo-diodes. The amount of light penetration and absorbency range of wavelength are important to char-

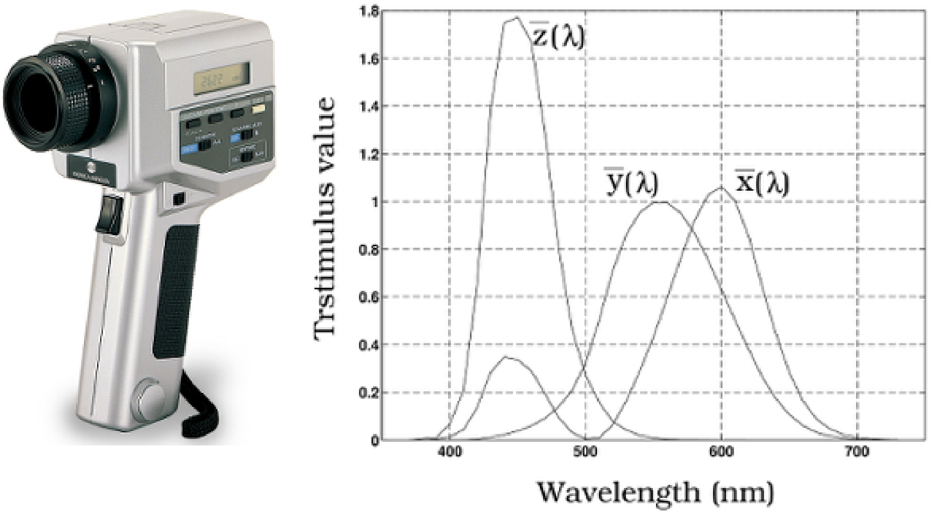


Figure 2.8: The most widely used hand-hold photometer Konica Minolta CS-100 (left), and the color matching functions for the CIE 1931 standard colorimetric observer (right) [172].

acterizing the filter transmittance, as well as calibrating the colorimeters. Eventually, the filtered light is converted by the detectors into electronic signals, which directly yield the standard CIE XYZ or CIE LAB tristimulus values as the measurement outputs. However, in this process, the matching of spectral sensitivity of the filters to the standard CIE tristimulus curves might have limited accuracy; then the quality of a colorimeter can be assessed by following the procedure defined in CIE standard [72]. The colorimeters can be alternatives to spectroradiometers, however they cannot provide detailed spectral information. For the image quality assessment of displays, colorimeters were commonly used to measure the luminance and chromacity for display calibrations. For example, a green filter is incorporated into an imaging colorimeter to perform the detection of mura or blemish artifacts for flat panel displays [145]. Jean-Baptiste [176] and Liang [102] used a colorimeter to measure the tristimulus values of a large amount of color patches, and used these measured values to build up a 3D look up table in order to rebuild the color gamut of the specific imaging device respectively. Son et al. [162] used a colorimeter to obtain the CIE XYZ values of primary colors of a time-varying mobile beam projector, determined the corresponding linear color transform matrix, and correct displayed colors accordingly. The well-known colorimeters in the consumer market are, but not limited to, Spyder series, X-Rite i1 (Figure 2.9), X-Rite ColorMunki series, and ColorHug. These colorimeters are mainly used to quickly calibrate display colors, so they do not have their own light source but they are placed directly on the top of the screen surface. For scientific research, more advanced colorimeters, such as CR series and LMT C series are used.

2.3.4 Spectroradiometer

Spectroradiometer is suitable for measuring the light source of which the spectral energy distribution is required for analysis. They measure all aspects of the radiometric, photometric, and colorimetric quantities of the light source, as well as the radiation spectrum distribution. In other words, one colorimeter can be the a faster, less expensive and more efficient alternative to an spectroradiometer, but with less measurement accuracy and without detailed spectral information. The dispersion of light is usually accomplished in the

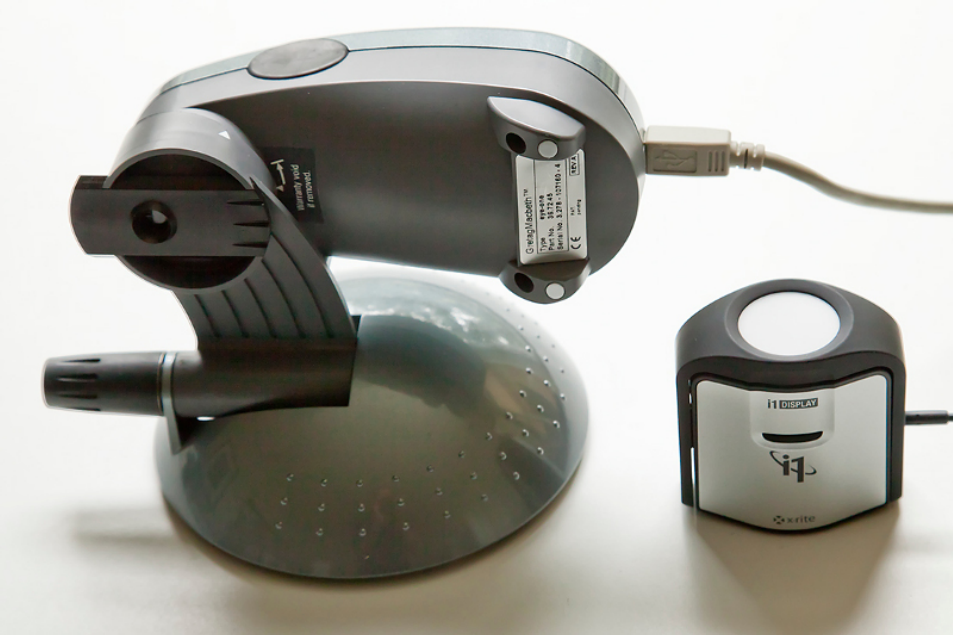


Figure 2.9: The X-Rite i1 colorimeter for measuring colors of displays.

spectroradiometer by means of prisms or diffraction gratings (Figure 2.10). In this case, the light is spread onto a linear CCD array as the energy detector. Normally, a spectroradiometer makes an additional measurement by following one measurement of the light source with its aperture closed. This procedure is called "cooling down", which is designed to estimate the thermal or random noise inside the spectroradiometer. Because the detector signal is calculated by counting the number of photon strikes and its value cannot be negative, the noise is assumed to have a Poisson distribution; the mean of the noise distribution is expected to have a positive value, therefore it can be estimated and removed by subtracting it from the actually measured signals. In addition, since the photon strike numbers are being integrated, the detector saturation must be carefully avoided with respect to the exposure control. Due to these reasons, the entire measurement process of a spectroradiometer may take a long time to finish, especially in a low light condition. The CIE $V(\lambda)$ curve and color matching curves are stored in the software, which is used to process the obtained power spectral distribution. Thus, the measurement errors associated with the photometers and filter colorimeters can be largely avoided in the spectroradiometer. So, adequate sensitivity, high linearity, low stray light, low polarization error, and a spectral band-pass resolution of 5 nm or less are essential for spectroradiometer to obtain good measurement accuracy.

2.3.5 Summary

According to the statements above, we can see that optical instruments can be used to measure one or more aspects of the light source, such as radiometric, photometric, and colorimetric properties. With respect to the experimental objectives, different instruments should be used in different scenarios. Sometimes, two or more instruments can be combined to use in one experiment. One typical case can be characterizing or calibrating the instrument with lower accuracy with respect to the same type of measurements provided by the instrument with high accuracy. Then, the common experiments can be performed much faster with the characterized or calibrated instrument but with lower cost. However, the accuracy might be limited, especially with respect to the variance and noise levels.

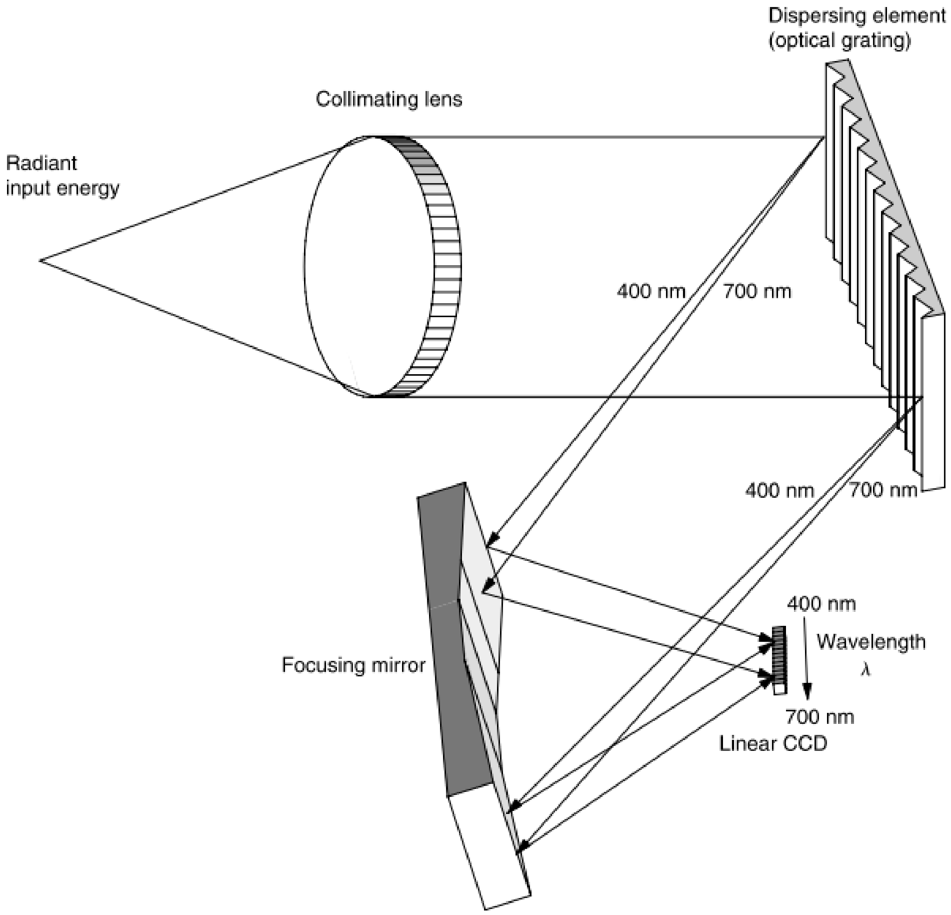


Figure 2.10: The dispersion of light accomplished in the spectroradiometer by means of prisms or diffraction gratings [179]

Nevertheless, the radiometers and photometers are normally used only for characterizing the radiance and irradiance, while the colorimeters and spectroradiometers are used in the cases of which the quantities of luminance, illuminance, and chromacity are demanded in a specific color space. Since the colorimeters and spectroradiometer are required to match the color matching functions of CIE standard observer, and their viewing angles are also specified (e.g. 10 degrees or 15 degrees), they may be treated as objective observers. However, they cannot completely replace human observers due to the fact that many perceptual capabilities are still not well understood [89].

2.4 Subjective Assessment

Image quality can be assessed either objectively or subjectively. In the former case, observers are replaced by optical instruments to observe the image reproductions. The capabilities and behaviors of the human visual system are simulated by the computational metrics. However, the numeric results given by the metrics may have values in completely different ranges with respect to different scales. Although it is possible to normalize the metric results in their own metric space, the normalized metric results with the same value have different meanings. Therefore, the most reliable way to benchmark the metric performance is to correlate the metric results with the perceptual results. However, the actual perceptual results are unknown and they are unlikely to be obtained accurately, but they

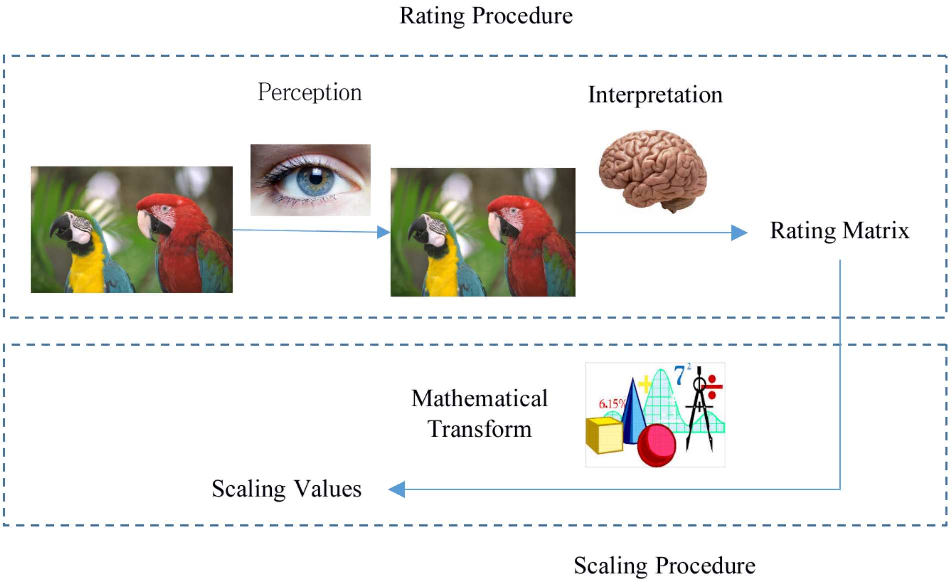


Figure 2.11: Conceptual rating and scaling procedures of psychometric model

can be estimated with respect to the perceptual results collected from a small group of observers in the subjective experiments. In this case, the estimated perceptual results are assumed to form a ground truth in the current experimental environment regarding a specific image quality attribute. The collected perceptual results are assumed to have the identical distribution as the actual perceptual results for the entire target population. Hence, the statistic conclusions made upon the small group of observers can be generalized to cover the entire target population. Until present, due to the lack of knowledge regarding biological structure of the human visual system and human brain, subjective experiment is still the most reliable way to perform image quality assessment [44]. The observer's physical condition, mental state, color experience and personal preference increase the variance of sampled data and they are somehow difficult to be quantified precisely. As a result, a large number of image stimuli and observers are required. The number of image stimuli is proportional to the amount of time used by the observers, and the length of the experimental study. There is constant trade off between the wish to have as many stimuli as possible and the acceptable resource (time, money, observers, etc.) consumption. Usually, an agreement between the number of stimuli and observers need to be found, which is likely to be a reasonable midpoint. The goal of subjective experiment is to obtain the perceptual indications regarding a specific image quality attribute or overall image quality. The typical work-flow can be generalized as a conceptual psychometric model, which is divided into two major procedures: rating and scaling (Figure 2.11).

2.4.1 Rating Procedure

In the rating procedure, the human visual system acquires the displayed images; then the brain interprets the information to generate opinions regarding the underlying image quality attribute. These implicit perceptual and cognitive processes vary largely from one observer to another, but they can be potentially influenced via the interactions with either the instructor or the experimental environment in the field. In the case of image quality assessment, the end product of the rating procedure is a matrix representing the numerical ratings of each level of image distortion from all human observers. Brown et al. [21] presented an excellent study regarding the challenges of interpreting the rating scales. They identified the research challenges and classified them into five major categories: unequal-interval judgment criterion scales, lack of inter-observer correspondence, linear difference between group average criterion scales, lack of intro-observer consistency, perceptual and criterion shifts. A good understanding of these problems is essential for advancing the design and improvements of psychometric models. Suppose that we have four human observers A, B, C, and D, and they are asked to rate three distortion levels of one test image with category judgment method; obviously, each human observer has his/her own judgment criterion scales. In this context, the challenges in the rating procedure can be briefly demonstrated in Figure 2.12. Typically, the judgment criterion scales have different origins, ranges, and intervals. The differences of origins and ranges are mainly due to the natural preference for perception and/or the temporary variations of judgment criterion scales; they can be estimated with linear transfer functions, which are formulated with psychometric scaling models. However, so far there is no effective way to quantify the interval differences, since the psychometric rating is completely an implicit perceptual and cognitive process. Empirically, well trained observers with color expertise are more likely to have above average equal intervals, while the non-experts are not. One may argue that it is possible to employ Monte Carlo like statistical analysis to estimate the judgment criterion scales, however the essential large amount of random tests are impractical to be applied to a large group of human observers. In many cases, the ratings from an individual observer can be inconsistent. To the same stimulus, regarding a specific image quality attribute, one observer may give completely different ratings in various rating sessions. If the variation can be assumed to be a random factor which follows a normal distribution around the ”true” perceived value. The real perceived value can be estimated by statistical regression, but the trade off is that the regression requires a large number of random samples of which the collection is both time and resource consuming. Generally, the rating procedure is performed with one of the following methods: rank order, category judgment, and pair comparison. The rank order basically requires observers to use numbers to indicate their preferences of ranking for a series of image distortions for one test image regarding a specific image quality attribute. In this case, the meanings of each rank number can be specifically defined by the experiment designer. In some cases, the observers are provided with a series of statements describing the meaning of the ranks. However, each of these descriptions is eventually associated with a rank number. Category judgment is proceeded in a similar fashion. The main difference between category judgment and rank order is that multiple image distortions can be classified into the same category, however it is not the case for rank order. From this point of view, rank order is designed to evaluate if there is a perceptual difference between different levels of image distortions. In this case, the observers are forced to make choices, even if they find absolutely no difference between two image distortions. The category judgment has a more flexible tolerance of small perceptual differences. Pair comparison is performed based on the comparison between each pair of the image distortions. Since all possible combination of pairs should given to the observers, the amount of workload for both experimental instructor and subject can be significant, if the number of image distortions is large.

2.4.2 Scaling Procedure

In the scaling procedure, the raw ratings are transformed in order to distinguish the perception of a stimuli and the corresponding judgment criterion scale for assigning rating to that stimuli. The outcomes indicate the relative impression of the perceived image quality attribute or overall image quality. They are meaningless without the references to the observers’ judgment criterion scales. Brown et al. [21] presented six typical scaling methods:

- Median rating: it uses median ratings over all human observers regarding a single stimuli as the scaled ratings. There is no assumption of equal intervals of judgment criterion scales. In contrast, it provides only the ordinal information of ratings.

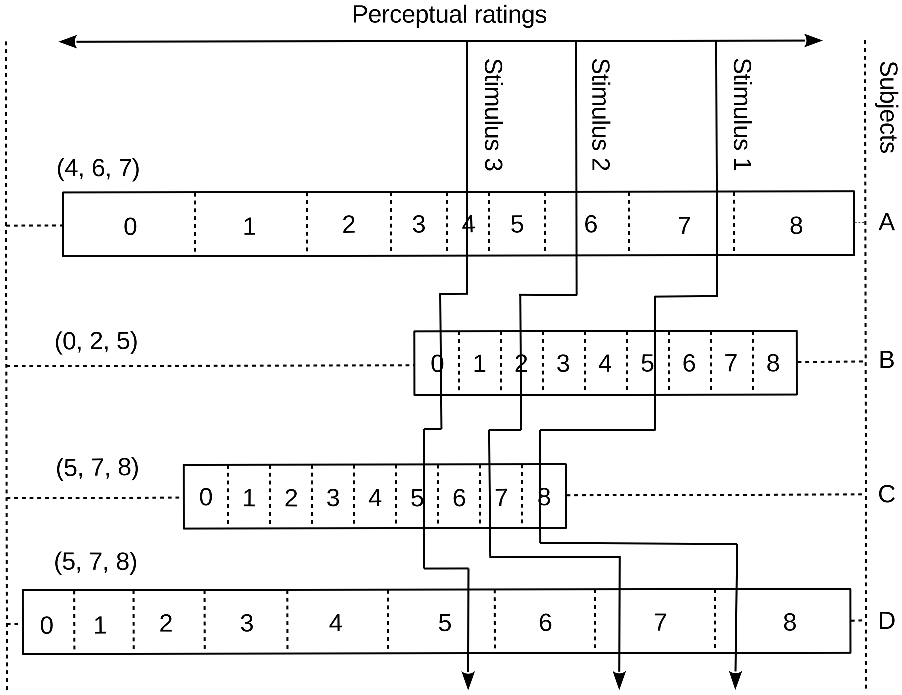


Figure 2.12: The judgment criterion scales of four human observers.

- Mean rating: it uses mean rating as the scaled output, and it requires the interval of judgment criterion scales must be equal. However, this assumption does not hold in most cases.

- Origin-adjusted rating: it removes the rating mean prior to aggregating them for each human observer and it cancels the differences of origins of judgment criterion scale, but not the differences of interval sizes.

- Z-score: it is similar to origin-adjusted rating in removing the differences of shift. In addition, it normalizes ratings with respect to their standard deviation, so the linear differences between observers are eliminated. The definition and calculation of Z-score are both simple; but the method accounts in both shift and normalization of judgment criteria, which account for most of the variance in the judgment criteria differences. So, it is very common for image quality researchers to use the Z-score method to scale the raw subjective ratings.

- Least square rating: it does not merely inherit the features of Z-score scaling, but also counts in the correlations between individual and all observers in the same group. Larger correlation indicates for larger contribution from individual observer to the same group of observers.

- Scenic beauty estimate: it was originally developed to scale ratings of scenic beauty of forest area, but the procedures are also appropriate for use with ratings of other types of stimuli. The differences in ratings are assessed by comparing an observer's rating distribution (assumed to have a normal distribution) for one landscape area against each of several other landscape areas. It features with a relative operating characteristic, where a bi-variate graph of the cumulative probability of the ratings for the selected landscape, is compared against the cumulative probability of other ratings, respectively. The scaled outcomes are generated by calculating the distance of the standardized relative operating characteristic from a positive diagonal of difference matrix.

2.4.2.1 Baseline Adjustment

In a typical working cycle of subjective image quality assessment, the case may occur, in the data post-processing phase, that the researcher realizes that it is mandatory to adopt observations on additional image distortions to draw the final conclusions. For example, a general tendency of human perception has been discovered; but the numerical distance between two consecutive distortion levels might be larger than they are expected. As a result, many perception details within these pre-defined intervals are not available. Conventionally, the researcher needs to conduct a large new subjective experiment incorporating all existing and additional image distortions. The purpose is to assess all stimuli to be rated under the same circumstances, so the unwanted experimental artifacts between possibly two or more separate sessions can be largely avoided. However, the whole process is non-trivial and it may consume considerable time and resources.

Baseline adjustment can be a potential answer to this challenge. This method introduces common stimuli to form a baseline in order to determine the comparability of ratings between different experiment sessions, and allows the computation of scale values expressed relative to responses for the baseline stimuli [23]. The basic concept is depicted in Figure 2.13. The ratings for unique stimuli in both rating sessions are scaled respectively with respect to the selected common baseline in either session, and then they are merged to generate the final ratings. Suppose that we have one human observer, who is asked to rate four stimuli 1, 2, 3, and 4 in the first session and another two additional stimuli 5 and 6 in the second session. In this case, stimuli 3 and 4 are selected to form a common baseline. Notice that the ratings for them across the two rating sessions may have different values. The ratings of unique stimuli 1 and 2 in the first session are scaled with respect to the baseline in the rectangle on the left in Figure 2.13, and the ratings of unique stimuli 5 and 6 in the second session are scaled with respect to the baseline in the rectangle on the right in Figure 2.13. Since we are merging the scaled ratings from the second rating session to the first one, then the ratings for the baseline in the first session is scaled to generate the scaled baseline which has zero mean and normalized standard deviation. Finally, all scaled ratings are combined to be used as the final ratings. It is important that the stimuli for the two sessions are rated under the same circumstances. In order to achieve this goal, the following precautions should be followed [21]:

- the observers for each session should be randomly selected from the same observer population,
- the observer groups should be sufficiently large,
- the baseline stimuli should be representative of the full set of stimuli to be rated,
- the non-baseline stimuli should be randomly assigned to the different sessions,
- all other aspects of the sessions (e.g., time of day, experimenter) should remain constant.

Baseline adjustment is a higher level of abstraction on rating scaling, it must be integrated with a specific scaling method which is mathematically well formulated. All scaling methods introduced in the previous Section 2.4.2 can be adopted as the candidates. In this paper, we choose to integrate Z-score scaling with baseline adjustment method. Z-score scaling is widely used in the psychometric modeling, mainly because of the simplicity on

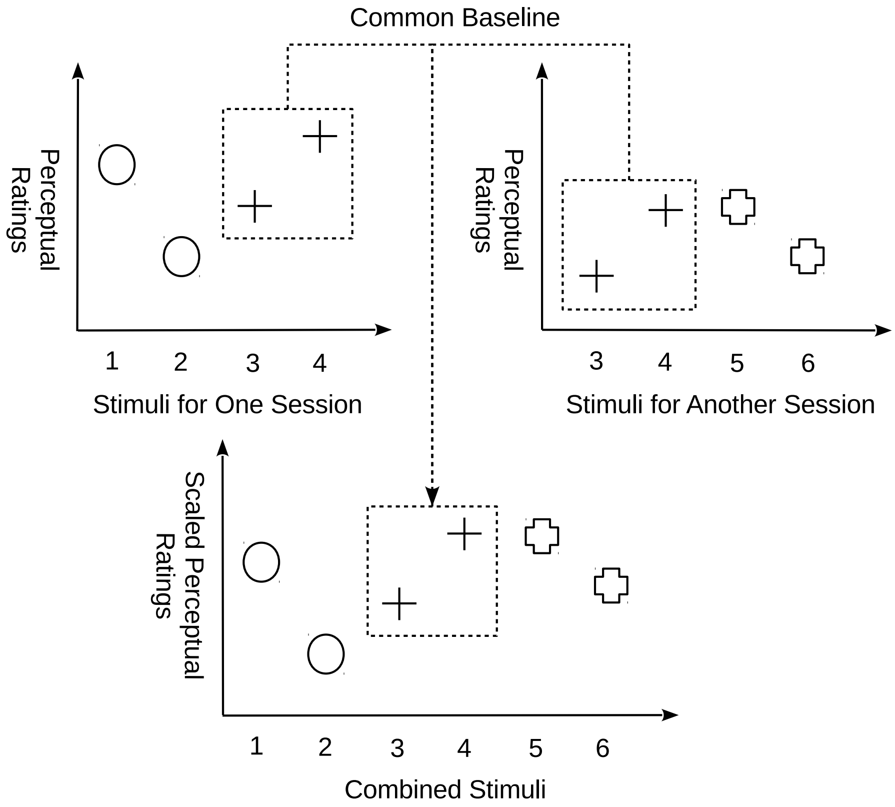


Figure 2.13: Depiction of the basic concept of baseline adjustment method

its definition and it eliminates the problems of origin shifting and unequal range of judgment criterion scales. Beyond this, the scores compute readily with computer programs. In this context, the raw ratings are scaled in the following way [21]:

$$BZ_{ij} = \left(R_{ij} - BMR_j\right) / BSDR_j$$

where BZ_{ij} stands for the baseline-adjusted Z-score of stimulus i for observer j, R_{ij} stands for the ratings assigned to stimulus i by observer j, and BMR_j stands for the mean of ratings assigned to the baseline stimuli by observer j, and $BSDR_j$ stands for the standard deviation of ratings of the baseline stimuli by observer j; then the BZ_{ij} are then averaged across all observers to generate one scale value per stimulus as BZ_i.

2.5 Image Quality Assessment Frameworks

In the past, much research have been conducted to develop various types of image quality assessment frameworks. For example, Zhang et al. [205] proposed an image quality assessment framework for printing. The printed image reproductions were assumed to be color patches and they were scanned four times in total and each time the scanning resolution was set to 1200 DPI. In the first time of scanning, the printing was scanned as it was; while it was scanned three more times with different color filters. Then the color information of pixels were converted into CIELAB color space, where the image quality metrics S-CIELAB [206] was applied. In their approach, there was no image registration or descreening process applied. Xu et al. [194] further extended the S-CIELAB metric based framework to adopt two control points located at the top-left and bottom-right corners of prints respectively to provide assistance to image registration. The prints were scanned at 300 DPI and the acquired images were descreened to eliminate halftone frequencies. Eerola et al. [181, 41, 42] proposed a full reference image quality metric based framework for halftone prints. The main idea was taking advantage of a high quality scanner to acquire the printed images, descreening the acquired images with Gaussian low pass filter to eliminate halftone frequencies, registering the descreened images at a sub-pixel level with their originals based on scale-invariant features and rigid 2D homography transformation, and then applying full reference image quality metrics to each pair of registered images. In this process, the image quality degradation due to scanning was assumed to be magnitudes smaller than the one of printing. In a similar fashion, Pedersen et al. [131, 132] also proposed a full reference image quality metrics based framework for printing. In their approach, control points at four corners of the printed images were used to produce appropriate affine transformation in the image registration process. Nuutinen [127] proposed an image quality assessment framework specifically for image sharpness measurement based on reduced reference metrics. Their method also employed scale-invariant features to determine the geometric distortion. However, only the correspondence areas with high local contrast (determined with Difference of Gaussian formula) were considered to be matched. In addition, an orientation histogram based point descriptor was used to calculate the match features for the key points in the images. For no reference image quality metrics, Moorthy et al. [121] proposed a two-stage image quality assessment framework to incorporate natural scene statistics. In their solution, support vector machine was used to classify image distortions into one of four distortion categories, and then support vector regression was employed to assess the distortion-specific image quality. Since no reference metrics are not dependent on the original image, then image registration is not required.

With respect to the discussions above, it is clear that most of the existing image quality assessment framework research were performed in the domain of either halftone printing or digital photography. The state-of-art image quality assessment framework commonly involves both control points for image registration and descreening process to eliminate halftone frequencies. Especially, the image registration typically assumed that the geometric distortions can be linearly corrected by inverting typical rigid spatial transformations, such as translation, scaling, rotation, and skew. Control point based image registration can be assumed to be more reliable than image feature based approach, since it is independent from the image content and the magnitude of actual image quality degradation. However, for projection displays, conventional control points based approach might not be suitable. Because the control points should be small enough in order to indicate precisely the locations of specific key points. However, for projection displays, the observers or cameras might be far away from the projection screen, so that the control points can be invisible or have a low appearance quality due to the limited visual acuity of observers or the spatial resolution of cameras. So the control points projected on the screen must have a certain size in order to be detectable. Thus, it is impossible to place these control points on the edge of the projection area. In addition, the projection appearance (e.g. position, size, and orientation) on the captured images is unknown in advance of image registration; depending on the relative position, viewing angle and orientations of projector, screen and camera, the projections are expected to have different spatial distortions on the captured images. Thus, appropriate distortion correction models shall be applied. A good image quality assessment framework should be adaptive and robust to the variations of viewing distance, viewing angle, and relative orientations with minimized assumptions. After all, the existing full reference, reduced reference, and no reference image quality metrics can be incorporated and evaluated under the framework without any modification, and they can focus on simply getting the registered image inputs and generating the corresponding image quality scores. Subsequently, these scores are correlated with perceptual evaluation results to evaluate their performance automatically. From this point view, a good image quality assessment framework for projection displays should be such a computational environment, which identifies the principal operational components of image quality assessment in a well-defined workflow, be adaptive and robust to various experimental setup and viewing conditions, incorporates existing image quality metrics without modifying them, and performs the image acquisition and quality assessment tasks in a fully automated fashion. Obviously, the existing image quality assessment frameworks primarily designed for printing and digital photography did not meet this requirement well. Thus, we need to propose a novel image quality assessment framework to confront the research challenges.

Chapter 3

Summary of Included Papers

> A research problem is not solved by apparatus; it is solved in a man's head.
>
> CHARLES F. KETTERING

In this chapter, we present a brief summary of the seven included papers (A-G). This summary outlines the scope, objective, methods used, results and principal conclusion related to each paper.

3.1 Paper A: Camera-based Measurement of Relative Image Contrast in Projection Displays

Ping Zhao, Marius Pedersen, Jon Yngve Hardeberg, and Jean-Baptiste Thomas

In *4th European Workshop on Visual Information Processing*
IEEE
Paris, France
pp. 112-117
June, 2013

3.1.1 Abstract

This research investigated the measured contrast of projection displays based on pictures taken by uncalibrated digital still cameras under typical viewing conditions. A spectroradiometer was employed as a reference to the physical response of projection luminance. Checkerboard, grayscale and color complex test images with a range of the projector's brightness and contrast settings were projected. Two local and two global contrast metrics were evaluated based on the acquired pictures. We used contrast surface plots and Pearson correlation to investigate the measured contrast versus the projector's brightness and contrast settings. The results suggested that, as expected, the projector contrast has a more significant impact on measured contrast than projector brightness, but the measured contrast based on either camera or spectroradiometer has a nonlinear relationship with projector settings. The results also suggested that simple statistics based metrics might produce a higher Pearson correlation value with both projector contrast and projector brightness than more complex contrast metrics. Our results demonstrated that the rank order of un-calibrated camera based measured contrast and spectroradiometer based measured contrast is preserved for large steps of projector setting differences.

3.1.2 Motivation

At the beginning of this research project, it was found that the image quality assessment for projection displays was not well studied in the existing literature. The existing research was largely carried out based on the measurements of spectroradiometer. The spectroradiometers were designed to give accurate measurement to the physical response of incidental light at a certain spot or over an area, but they were not suitable to measure a large number of samples. Specifically, it is impossible to measure the colors of individual pixels in a complex color image. In addition, in low ambient light conditions which are very typical for projection displays, it takes a long time for a spectroradiometer to have a single measurement due to the time cost of essential photon integration and dark current estimation. In contrast, using a digital still camera as the acquisition device can be an alternative approach to engage the study. The camera can record all pixels in the projection in one shot. The modern cameras are capable of capturing images in high resolutions, and the camera sensors are common to have high dynamic ranges. These features can be helpful for the acquisition of image projections, so it is worth to investigate the possibility of using the camera to acquire the projected images.

Then, it came to the problem of how we can use the camera to achieve the best acquisition for image quality assessment, and what research challenges might be encountered. However, to the best of our knowledge there was no such research conducted in the existing literature. Therefore, the best way to conduct the research was to set up an experimental environment and simulate the typical viewing conditions for the projection displays; so that the digital still camera can be used to capture the projected images in the field. In this process, we could observe the phenomena, identify the problems, and further come up with corresponding solutions. In order to evaluate the image quality of projection displays, we needed to characterize it with image quality attributes, such as brightness, contrast, colors, sharpness and noises. Based on these attributes, there were many attempts to characterize devices like CRT [57, 17] and LCD [57, 17, 45] displays. The goal of image quality assessment of projection displays can be achieved in a similar fashion. Previous characterizations of projection displays focused on black level estimation [10], display uniformity [108, 176, 109] and colorimetry [176, 61], but limited attention was paid to measured contrast of the displayed images. The measured contrast of a displayed image has been shown to be of a significant impact on visual experience [140]. In conventional approaches, the contrast of displays was largely determined based on the simple ratios between the highest and lowest measured luminance in the projections. The state-of-art image quality metrics simulated the human visual system to determine the contrast with much more complicated definitions. Nevertheless, the contrast predicted by these metrics had strong connection to the measured luminance, but has limited dependence on the variations of projection geometry and chromacity in the captured image. Thus, the measurement of relative contrast in projection displays by using digital still cameras can be a good starting point for the entire research project.

3.1.3 Methods

In this research, a low-end webcam Logitech QuickCam Pro 9000 (3264 x 2448 in pixels), a prosumer DSLR camera Nikon D200 (3872 x 2592 in pixels), a high-end DSLR camera Hasselblad H3D II (6490 x 4870 in pixels), and a spectroradiometer Minolta CS1000 were used for the image quality assessment. The webcam was mounted on a table, which was approximately three meters away from the projection screen. The other two cameras were mounted on a tripod respectively, and they were approximately four meters away from the projection screen. The pictures were always taken remotely with software installed on the controlling laptop without physically touching the cameras. We used a LCD projector SONY APL-AW15 (1280 x 768 in pixels). The projector was placed on a flat table in front of the screen, and it was approximately 3 meters away. The projection was approximately 2 x 1 in meters on the screen. All projector settings related to the brightness, contrast and color enhancements were switched off to make sure the input image was projected as it is. One checkerboard, one gray patch and one color complex test image (768 x 512 in pixels) were incorporated in the experiments. The color complex image was selected from Kodak Photo

CD PCD0992 [52] based, on which the gray scale version was generated by using Matlab function ”rgb2gray”: $L = 0.2989 \times R + 0.5870 \times G + 0.1140 \times B$. The test images were always projected at the original resolution on the screen. Fluorescent light was incorporated to simulate the ambient light for the typical projection environment at approximately 0 Lux (dark room), 30 Lux (meeting room) and 300 Lux (day-light office) respectively.

The projections in the captured pictures were surrounded by the dark background. In order to estimate the influence of surround on the measured contrast, all captured pictures were processed respectively to generate a cropped version with the image content only. However, both the cropped and noncropped picture versions were forwarded to the image quality metrics. Eventually, for each group of captured pictures with the same image content and under the identical viewing condition, we generated a surface plot for the measured contrast with respect to the metric scores. In our experiments, we evaluated four contrast metrics: Michelson [120], LAB Variance [139], RAMMG [150] and RSC [160]. Since the RAMMG and RSC shared various input parameter coefficients, we evaluated several combinations of them which involved: channel weightings: $(1, 0, 0)$, $(1/3, 1/3, 1/3)$, $(0.5, 0.25, 0.25)$, pyramid scales: linear and log based scales, radius of center and surround of receptive field: $(1, 2)$, $(2, 3)$, $(3, 4)$. These parameters were used and recommended by Simone et al. in their investigation of measuring perceptual contrast [159]. Because the selected image quality metrics had no parameter related to viewing distance, we placed the instruments at the same location to make sure that they share a constant distance to projection screen. In this case, the influence of viewing distance on measured contrast was equal to all metrics. In order to reduce computational complexity, the level weighting method was always set to variance and pictures were transformed into CIELAB space by the metrics themselves. We also evaluated the performance of metrics by determining their correlation between measured contrast and projector contrast, and the correlation between measured contrast and projector brightness with respect to Pearson correlation coefficient.

3.1.4 Results

Based on the observation of experimental results, a conclusion could be made that the webcam was not suitable for measuring contrast for projection displays, because the corresponding contrast surfaces were not consistent in different viewing conditions. In contrast, the Nikon D200 produced similar results as Hasselblad H3D II, however it was not the case for Michelson contrast in the high light condition. In that case, the contrast surfaces for Nikon D200 were smoother than the ones for Hasselblad H3D II. This was mainly because the Hasselblad H3D II camera was more sensitive to the luminance and it was capable of detecting the small spatial variance under both low and high light conditions. Although the acquisition was more accurate, the corresponding relative contrast predicted by the metrics was more sensitive to image noises. From this point of view, the camera Nikon D200 made a trade-off between the two extreme cases; so it was preferred as a measurement instrument for relative contrast of projection displays.

Under the low light condition, Michelson contrast had the maximum absolute value for all types of cameras and test images, despite the projector contrast and brightness settings. This result suggested that Michelson contrast was sensitive to measurement noises, and it became very unstable in the high light condition. LAB Variance, RAMMG and RSC metrics produced logistic-shape-like contrast surface. RAMMG and RSC metrics shared a general shape of normalized contrast surface despite the parametric coefficients for the radius of receptive center and surround. They were more sensitive to the increasing rate of projector brightness and contrast than the LAB Variance metric, since the contrast surface appears to be more bended. Pearson correlation was employed to determine the correspondence between measured contrast and projector settings. The spectroradiometer based Michelson contrast correlated well with camera based RSC metric, while keeping the projector brightness constant. The LAB Variance metric produced a higher correlation with low projector contrast, but the correlation decreased a lot while projector contrast increased. In the case projector contrast remained constant, the LAB Variance metric based contrast had a higher correlation to spectroradiometer based Michelson contrast. The measured contrast of RAMMG and RSC metrics correlated well with projector contrast and projector brightness, and RSC metric produced approximate 12% higher correlation than RAMMG metric. With respect to the experiments, we could see that the projector contrast has more significant impact than the projector brightness for the measured contrast with all metrics, as expected. It lead to an asymmetric contrast surface. The measured contrast for digital still cameras has a consensus with the one for the spectroradiometer.

3.1.5 Conclusion

In this research, several contrast metrics were evaluated based on pictures taken by uncalibrated digital still cameras in the typical viewing conditions of projection displays. The results showed that the projector settings have a great impact on the measured image contrast, and the impact of projector contrast setting is even stronger. Camera based Michelson contrast was proved not to be suitable for projection contrast measurement, while the global metric LAB Variance produces higher Pearson correlation values than the complicated local metric RAMMG and RSC on both brightness and contrast correlations. Thus, we demonstrated that the rank order of uncalibrated camera based measured contrast and spectroradiometer based measured contrast is preserved for large steps of projector setting differences, and the digital still camera based acquisition could be an alternative approach to the spectroradiometer based acquisition.

3.2 Paper B: Image Registration for Quality Assessment of Projection Displays

Ping Zhao, Marius Pedersen, Jon Yngve Hardeberg, and Jean-Baptiste Thomas

In *21st International Conference on Image Processing*
IEEE
Paris, France
pp. 3488-3492
October, 2014.

3.2.1 Abstract

In the full reference metric based image quality assessment of projection displays, it is critical to achieve accurate and fully automatic image registration between the captured projection and its reference image in order to establish a sub-pixel level mapping. The preservation of geometrical order as well as the intensity and chromaticity relationships between two consecutive pixels must be maximized. The existing camera based image registration methods do not meet this requirement well. In this paper, we propose a markerless and view independent method to use an un-calibrated camera to perform the task. The proposed method including three main components: feature extraction, feature expansion and geometric correction, and it can be implemented easily in a fully automatic fashion. The experimental results of both simulation and the one conducted in the field demonstrate that the proposed method is able to achieve image registration accuracy higher than 91% in a dark projection room and above 85% with ambient light lower than 30 Lux.

3.2.2 Motivation

In the full reference metric based image quality assessment [134, 137, 181], image quality was evaluated with respect to a series of selected attributes. For projection displays, the camera could be used to quickly acquire the projection and record all pixels in one shot. In addition, the measurement results were correlated well with spectroradiometer based measurements for large steps of image distortions. In order to apply existing full reference image quality metrics to the captured images without modifying the metrics, it was critical to achieve an accurate and fully automated image registration between the captured image and its original version. Subsequently, the image content in the captured image would share the dimension and resolution as the original one. The preservation of geometrical order as well as the intensity and chromaticity relationships between two consecutive pixels on the screen could be maximized. However, the existing camera based image registration methods did not meet this requirement well, because they either placed simple assumptions on the projections and cameras in order to reduce the problem complexity, or they tended to implicitly modify the captured image quality. The captured images were expected to have various spatial distortions with respect to the relative positions and orientations of the projector, screen, and camera. The camera lens introduced additional spatial distortions. Hence, establishing a robust, accurate and reliable image registration for image quality assessment for projection was required.

3.2.3 Methods

In this research, we proposed a marker-less and view independent method to use a camera to quickly capture the projection and correct its nonlinear spatial distortions without calibrating a camera in advance. It had three major components: feature extraction, feature expansion, and geometric correction.

3.2.3.1 Feature Extraction

One dot pattern and one cross pattern image was generated and projected in full screen size in order to estimate spatial distortions. The dot pattern incorporated round solid black dots evenly distributed in a $M_d \times N_d$ grid layout, where M_d and N_d represent the number of columns and rows respectively. The cross pattern included crosses that shared the center locations and radius with the dots in the dot pattern. Let's denote the captured dot pattern as I_d, then a contour map C could be generated as

$$C = M_c \left(G_a \left(I_d \right) - G_b \left(I_d \right) \right), \tag{3.1}$$

where the Gaussian filter G_a with kernel size a (empirically $a < 5$) is adopted to reduce the screen-door effect [7, 203]. The kernel size a should be kept as small as possible to preserve the details in the captured image. The Gaussian filter G_b with kernel size b (empirically $b >$ 41) was adopted to spread energy from highly illuminated pixels to their neighborhoods. The median filter M_c with kernel size c (empirially $c = 3$) was adopted to remove the salt-and-pepper like noises, and to smooth the detected object contours. False contours might be visible in the generated map C, and they could be eliminated by applying a binary threshold. The output binary image I_b could be expressed as

$$I_b = \begin{cases} 1 & C_i > (1-p) \cdot L_{min} + p \cdot L_{max} \\ 0 & otherwise \end{cases} \tag{3.2}$$

where C_i denoted the ith pixel in contour map C, L_{min} and L_{max} denoted the minimum and maximum gray values in the image respectively, and $p \in [0, 1]$ was a constant threshold. Smaller value of p forced the detected projection boundaries to be compressed and vice versa. The value of p had a contrary effect on the detected dot contours. Hence, the pixels corresponding to positive thresholding were kept and the rests were removed. The algorithm proposed by Suzuki et al. [168] was adopted to determine the contours and their hierarchy relationships in the binary image I_b. The outermost and longest object contour corresponded to the projection area, while the innermost and shortest contours corresponded to the dots. The remaining contours were therefore discarded. For each identified contour, a moving window (empirically size equals to 5) was placed along its pixels and a dynamic threshold with respect to local statistics in the corresponding area of the original captured image I_d was calculated. Then, the contour pixel at the window center was shifted toward its neighborhood either horizontally or vertically to achieve the goal of local optimization. The local threshold T was determined as

$$T = L \cdot \left(1 - k \cdot \frac{\sigma}{L/2}\right), \tag{3.3}$$

where σ denoted the standard deviation of gray values within the local window, L denoted the maximum gray scale level (256 for 8 bit image) and the constant k (empirically $k \in [0.1, 0.3]$) indicated the confidence of the image quality of the captured image I_d. In the case of good image quality, the value of k could be scaled down to 0. Otherwise, it should be scaled up. Eventually, we adopted the algorithm proposed by Fitzgibbon et al. [51] to fit dot contours into ellipses with respect to the least square error minimization, so the actual center, size, and orientation of each ellipse could be estimated simultaneously. The estimated ellipse centers would be slightly shifted according to the ”cornerSubPix” algorithm provided by the OpenCV library [16]. Such an algorithm incorporated the detected cross centers from the cross pattern image to locally optimize the ellipse centers, since the dots and crosses shared the center location and radius.

3.2.4 Feature Expansion

The detected dot grid needed to be expanded to cover the entire projection area. In this case, we fitted the coordinates of all dot centers in the same row or in the same column into a parametric natural cubic spline function as sample points, and estimate the parametric coefficients accordingly. Once the spline functions were determined, we could generate a smooth parametric cubic spline passing through each set of the feature points. In turn, each spline was extrapolated to intersect with the detected contour of the projection area to generate a pair of new feature points. Thus, in total, $2(M_d + N_d)$ new feature points were generated. The four extreme corners of projection contour could be determined by applying the split-and-merge algorithm proposed by Heckbert et al. [64] iteratively to eliminate redundant pixels until only four corners are left. These corners were used as the feature points as well. Eventually, we had $(M_d + 2) \cdot (N_d + 2)$ feature points covering the entire projection area. The reason to employ the natural cubic spline was to take the advantage of its unique mathematical properties. A typical parametric formulation could be presented as

$$x\left(p_x\right) = \sum_{k=0}^{3} \alpha_{ik} \left(p_x - c_i\right)^k, y\left(p_y\right) = \sum_{k=0}^{3} \alpha_{jk} \left(p_y - c_j\right)^k \tag{3.4}$$

where $p_x \in [0, M_d + 1]$ and $p_y \in [0, N_d + 1]$ were the two parametric coefficients for the spatial coordinates of a point on the ith and jth spline section respectively; α_{ik} and β_{jk} were the local polynomial regression of kth parametric coefficient of the ith and jth spline sections respectively; $c_i \in [0, M_d + 1]$ represented the parametric coordinate of ith sample point on spline $x\left(p_x\right)$, and $c_j \in [0, N_d + 1]$ represented the parametric coordinates of jth sample point on spline $y\left(p_y\right)$. Each feature point in the expanded dot grid represented one sample point for the corresponding spline. The coefficients were estimated to make sure that around each key point the two consecutive spline sections share the same first and second derivatives; so the whole spline curve was differentiated and continuous below the third polynomial order at all possible locations. In addition, the estimated splines were adapted to local variance within each spline section. Higher order spline may not be employed to avoid adaptation to the errors inherited from the capturing process or from the calculations above.

3.2.5 Geometric Correction

We created a sub-pixel level mapping between the captured image and its reference, and the captured image could be undistorted by down sampling with respect to a specified interpolation method. The basic idea was to register pixels between the Cartesian coordinate system in the camera space and a distortion independent coordinate system defined by the expanded feature grid. Suppose the reference image resolution was given as $N_x \times N_y$ in pixels and the capturing resolution as $M_x \times M_y$ in pixels. Any pixel $P_o = (x, y)$ where $x \in [0, N_x - 1]$ and $y \in [0, N_y - 1]$ in the original image corresponded to pixel $P_c = (u, v)$ where $u \in [0, M_x - 1]$ and $v \in [0, M_y - 1]$ in the captured image and the pixel p_u in the un-distorted image. Their coordinates were defined in the Cartesian coordinate systems and their pixel correspondences in the distortion independent space were

$$Q_o = \left(\frac{x \cdot (M_d + 1)}{(N_x - 1)}, \frac{y \cdot (N_d + 1)}{(N_y - 1)} \right) \tag{3.5}$$

$$Q_c = \left(\frac{u \cdot (M_d + 1)}{(M_x - 1)}, \frac{v \cdot (N_d + 1)}{(M_y - 1)} \right) \tag{3.6}$$

respectively. The correspondence between P_r and Q_o, as well as P_c and Q_c were established with respect to the expanded feature grid which was generated based on cubic splines. Since the undistorted image was expected to exactly register with the original image, then the coordinates of P_r were equal to the ones of P_u in the distortion independent space. Special attention must be paid to the screen-door effect [7, 203]. The geometric correction might introduce wave-like artifacts. A trade off had to be made between blurring the captured image to register the geometry with the reference image, or distorting the reference image to register it with the captured image. In this paper, we adopted the former approach to make sure that existing full reference image quality metrics can be incorporated without any modification.

3.2.6 Results

The experiment was performed in a controlled lab environment. A portable LCD projector SONY APL-AW15 (1280 x 768 in pixels) was placed in front of a planar screen, and a DSLR camera Nikon D200 (3872 x 2592 in pixels) was used for image acquisition. All 24 images from Kodak Photo CD PCD0992 [52] were adopted for the test. We evaluated the proposed method against the pictures either generated by simulation tools or the ones taken in the field. The experiments were separated into two parts: evaluations with artificial images and images captured in the field.

In the first part of the experiment, the reference images and pattern images were scaled, rotated, and translated respectively at a series of levels to simulate a specific type of spatial distortion. The output images had the same resolution as the captured images. Since the actual distortions were known prior to the simulation, the image registration accuracy could be evaluated with respect to the maximum absolute displacements of pixels from their ideal locations. In cases where the scaling factors were greater than 1, the maximum displacements are below 0.2 pixel. These small errors were largely negligible, if the capturing resolution was at least two times higher than the original image resolution. The lowest displacements were given for special rotation angles as expected. In other cases, the absolute displacements were between 0.5 and 2 pixels, and they correspond to the misadjustments of the contour fine-tune algorithm (Equation 3.3) due to blurred edges in the captured images. The proposed method was completely independent from spatial translations. We also scaled and rotated all test images in a similar fashion and applied SSIM image quality metric [185] (kernel size 5) to measure the structural similarity. This was largely ignored by the conventional image registration evaluations. This metric incorporated the visibility of structural errors; it concerned the displayed image content and was able to detect complicated image quality issues like artifacts. The mean of similarity increased rapidly and the variance becomes smaller and more stable. Image rotation had limited influence on the proposed method since the structure similarity were always above 0.98.

In the second part of the experiment, we used the camera to take pictures of each of the projected reference image at 25 random locations and orientations in the field under low light (0 Lux) and dimmed light (30 Lux) conditions respectively, since the light condition had a great impact on the visual experience [207]. Then we applied SSIM metric to the registered images and their references due to the lack of ground truth for the actual projections. The minimum structural similarity was higher than 0.91 in all cases under the low light condition, and it was above 0.85 in the dimmed light condition. The variance of structural similarity between random locations were small, so the proposed method produced similar results despite the changes of camera position and orientations. The mean and variance under the dimmed light condition were worse. This was because the ambient light reduced the contrast between projection boundary and its surroundings, and the adjustment accuracy of contour fine-tune algorithm was influenced.

3.2.7 Conclusion

In this research, we proposed a marker-less view independent method to use an uncalibrated camera to achieve a sub-pixel-level registration between the captured projections and their reference images. The preservation of geometrical order as well as the intensity and chromaticity relationships between two consecutive pixels on the display were maximized. The experimental results against distortion simulations and captured images proved that the registration accuracy was considerably high under typical light conditions for projection systems. By incorporating this method, we could apply existing full reference image quality metrics to captured projections without any modification to the metrics.

3.3 Paper C: Perceptual Spatial Uniformity Assessment of Projection Displays with a Calibrated Camera

Ping Zhao, Marius Pedersen, Jean-Baptiste Thomas, and Jon Yngve Hardeberg

In *22nd Color and Imaging Conference*
Society for Imaging Science and Technology
Boston, MA, USA
pp. 159-164
November, 2014.

3.3.1 Abstract

Spatial uniformity is one of the most important image quality attributes in visual experience of displays [177, 178, 114]. In conventional studies, spatial uniformity was mostly measured with a spectroradiometer and its quality was assessed with non-reference image quality metrics. Cameras are cheaper than radiometers and they can provide accurate relative measurements if they are carefully calibrated. In this paper, we propose and implement a work-flow to use a calibrated camera as a relative acquisition device of intensity to measure the spatial uniformity of projection displays. The camera intensity transfer functions for every projected pixels are recovered, so we can produce multiple levels of linearized non-uniformity on the screen in the purpose of image quality assessment. The experiment results suggest that our work-flow works well. Besides, none of the frequently referred uniformity metrics correlate well with the perceptual results for all types of test images. The spatial non-uniformity is largely masked by the high frequency components in the displayed image content, and we should simulate the human visual system to ignore the non-uniformity that cannot be discriminated by human observers. The simulation can

be implemented using models based on contrast sensitivity functions, contrast masking, etc.

3.3.2 Motivation

In the previous studies, it has been demonstrated that the digital still camera can be an effective and efficient alternative to the absolute optical instruments with respect to the correlations of measurement outcomes. With the proposed image registration method, it is easy to incorporate existing full reference image quality metrics without introducing any modification. The next research question to ask is how the proposed framework performs for evaluating image quality attributes for displays, and how well the metric outcomes correlate with perceptual results under this framework. For this purpose, we identified the most important image quality attributes for projection displays. With respect to the literature survey, it was found that spatial uniformity has long been regarded as one of the most important image quality attributes for displays [110, 112, 113, 119, 178]. Since the definition of spatial uniformity involves luminance aspect, it depends on accurate absolute measurements in a specific color space. For this reason, spectroradiometers were commonly incorporated to perform the task. In this research, we incorporated the camera as a relative acquisition device to record the intensity of projections, implemented the core components for projection displays, and used the framework to incorporate image quality metrics to evaluate the spatial uniformity, and finally the correlation between metric results and the perceptual results would suggest the reliability and efficiency of the proposed framework.

3.3.3 Methods

Spatial uniformity has a strong connection to the spatial variation of luminance across the entire imaging area, so the camera needs to be calibrated in advance to ensure that its optical and electronic systems introduce no additional unwanted influence to image quality of captured images. For this reason, a method for vignetting correction and a method for exposure optimization were proposed to quickly eliminate the vignetting effect and optimize linearity of camera sensor responses.

3.3.3.1 Experimental Setup

The experiments took place in a controlled lab environment where the only illuminant in the room was the projector. We used a portable three chip LCD projector SONY APL-AW15 (throw ratio: 1.5) to produce projections on a planar screen, which was naturally hanging from the ceiling. The projector was placed on a table in front of the projection screen, and the distance is approximately 3m with respect to the throw ratio of the projector. A remote controlling laptop was connected to the projector via a VGA cable in order to generate full screen projections (approximately 2×1.2 in meters, 1280×768 in pixels). We used a DLSR camera Nikon D610 (6048×4016 in pixels) with a Sigma VR 24-105mm f/4G (VR off) lens to capture the projections. The camera was fixed on a tripod and the pictures were taken remotely with a software control on the laptop without physically touching the camera. The pictures were saved in raw format and rendered with Aliasing Minimization and Zipper Elimination demosaicing algorithm [117] without automatic vignette correction, brightness adjustment, gamma correction, noise reduction, etc.

3.3.3.2 Vignetting Correction

Vignetting effect stands for an undesirable gradual intensity fall off from the image center to its external limits. It is caused by the non-uniform energy distribution of light on the camera sensor array after light's passing through the camera lens, assuming that the incident light is uniform. We corrected the camera vignetting based on the captured pictures of a hazy sky, which was closely uniform in gray [128]. In the lab, we took several trial shots of projections with either minimum or maximum projector input intensity. In this process, we adjusted the camera settings iteratively until all the captures are neither underexposure nor overexposure. Then we kept all camera settings except the exposure time, used a neutral light diffuser (white and semi-transparent) over the camera lens, and used the camera to take multiple pictures toward the same spot of the hazy sky. Each time we took a picture we rotated the camera a bit. Then we calculated the median of intensity response for each pixel in all the pictures we took, and used them to generate a vignetting mask, which was then applied to the camera RGB channels separately to correct the vignetting. In the experiment, it was found that empirically 10 pictures were sufficient to generate convergent median results, and the mask center was shifted upward and also a bit to the right. In order to maximize the validity and reliability of image quality assessment, we should offer the best effort to avoid assumptions. Our method placed no assumption about the camera or the light condition, and the whole procedure can be finished within a few minutes.

3.3.3.3 Exposure Optimization

By applying the vignetting mask generated in the daylight condition to the low light condition for projection displays, we implicitly assumed that the camera always produced linear responses. In order to verify this prerequisite, we separated the input intensity equally into 15 levels. For each level, we displayed a gray patch on the projection screen, and captured it under all possible camera exposure times ranging from 1/4000s to 30s. Meanwhile, a light meter was used to measure the luminance on the projection screen as a reference to the camera. Then, we constructed surfaces of camera intensity responses versus the projector luminance and exposure time. It was found that the response surface of one camera sensor in one certain channel can be separated into two regions. In one region, the responses were closely linear to all possible projector luminance with constant exposure time, and vice versa. However, in the other region, the camera sensor had a very large boost in the responses. This was obviously not due to the saturation protection. From this point of view, the camera produced linear responses only for limited combinations of projector luminance and camera's exposure time. For this reason, we determined the strongest responses over each camera intensity response for the maximum luminance under the two light conditions with a common exposure time, and we continued to decrease the exposure time until the ratios between such two sensor responses were approximately equal. A very small exposure time was always safe, but it did not take the full advantage of the dynamic range of camera sensors. Once this condition is met, the generated vignetting mask can be applied to the camera despite of light conditions.

3.3.3.4 Projector Calibration

In this research, we adopted and extended the method proposed by Brown et al. [20] in order to produce multiple levels of linearized non-uniformity on the screen. First, we equally separated the projector intensity into 15 levels, and for each level we displayed and captured a gray patch 10 times. Then the projector intensity transfer functions were recovered by polynomial regression upon the median responses over all gray patches in a color channel basis, in order to avoid temporary stability of both camera and projector. After this, we inverse the transform to compensate the non-linearity of camera responses in order to create flattened projections. Suppose that the scaling ratio of one pixel p_{ij} in an individual color channel on the ith row and jth column of the registered image is $r_{ij} \geq 1$, the corresponding regression function for the reference pixel is $f\left(x\right)$ and its inverse function is denoted as $f^{-1}\left(x\right)$. The x stands for the projector input intensity of the pixel p_{ij}. The camera response of pixel p_{ij} is denoted as $c_{ij} = f\left(x\right) \cdot r_{ij}$. In this context, the projector input intensity for the pixel p_{ij} at a certain non-uniformity level is defined as $g\left(x\right) = f^{-1}\left(f\left(x\right) \cdot s\left(r_{ij} - m\right)\right)$, where $m = \sum_{i=1}^{n_y} \sum_{j=1}^{n_x} r_{ij} / \left(n_x \cdot n_y\right)$, n_x and n_y stand for the width and height for the projection in pixels respectively, and s stands for a linear scaling factor of non-uniformity and it is under the constraint that $f(0) \cdot r_{ij} \leq g_{ij}(x) \leq f(x) \cdot r_{ij}$ assuming that the projector input intensities are normalized to between 0 and 1. The value of r_{ij} can be determined as $max(c_{ij}) / f(1)$, where the operator max stands for the maximum value of c_{ij}.

3.3.3.5 Experimental Procedure

We displayed seven types of test images: two natural color pictures (the 15th and 23th picture from Kodak Photo CD PCD0992 [52]), three uniform colored patches with opponent colors: yellow, magenta and cyan respectively, one gray patch (gray level: 0.5), and one presentation slide like image with dark texts on a gradient background. We linearly scaled each test image to produce multiple levels of non-uniformity. These scaling ratios were normalized into the range between -1 and 1, and then they were split into five levels: -0.6, -0.2, 0, 0.2 and 0.6. The level 0 corresponded to flattened projections. We also displayed one image as it was to preserve the natural projector nonuniformity; so 42 images in total were presented to each observer. The experiment was set up as a category judgment experiment with test images displayed in a randomized order. Ten observers were asked to use category numbers between 1 to 5 to indicate the perceptual uniformity. The numbers correspond to the ranks between ”not uniform at all” and ”perfectly uniform”. At the same time, the observers were also asked to use numbers between 1 to 5 to indicate how the non-uniformity affect their pleasantness. The numbers correspond to the rank between ”very disturbing” and ”not disturbing at all”. All ratings were scaled to generated Z-scores [44]. We evaluated the uniformity with the following image quality metrics: LR defined in VESA FPDM [183], LG based definition [104] (SFA), averaged standard deviation of RGB values (Stddev), coefficients of variation [149] (Coeff), averaged Euclidean distance $\overline{\Delta E^*_{ab}}$ in CIELAB color space (ΔE_{ab}), PSNR-M [142], SSIM [185], and S-CIELAB [86].

3.3.4 Results

The experimental results included two parts corresponding to the subjective and objective experiments.

3.3.4.1 Subjective Results

The first observation was that the rank order of non-uniformity was largely preserved for the seven types of test images, as expected. If we assumed that the general tendency of Z-scores was smooth, then they could be represented by parabolic curves. The curves might be more or less skewed depends on the projected image content. The flattened projections did not necessary correspond to the highest overall Z-scores, while small negative non-uniformity and natural projection images had similar or relative lower Z-scores in many cases, and either positive or negative large non-uniformity lead to the lowest Z-scores. This observation supports the fact that the human visual system was not sensitive to small variation of non-uniformity. The spatial non-uniformity was largely masked by the high frequency components in the displayed image content, and we should simulate the human visual system to ignore the nonuniformity that could not be discriminated by observers. The simulation could be implemented using models based on contrast sensitivity functions, contrast masking, etc. For the distorted slide like images, the Z-scores of flattened versions were clearly greater than others. This was because such reference image has dark texts on a large gradient background in a bright color, and the nonuniformity on a gradient background could be easier to be detected by human visual system than that on a flat background, which was the case of a gray patches. The general tendency of mean Z-scores of pleasantness were similar to the ones of perceived uniformity, and the Pearson correlation between them were all above 0.98 for all test images, except the absolute mean values of pleasantness were slightly larger in general. This observation suggested that the human visual system had a certain degree of but limited tolerance on average against nonuniformity on the displays. For the gray patches, the observers had a difficulty to distinguish the differences between the small minus non-uniform, flattened, natural projections. In a similar fashion, the pleasantness of small minus non-uniformity, flattened and natural projections for the two natural images had similar values, but their corresponding perceived uniformity had different mean values. This observation suggested that the non-uniformity was masked by the complex colors of natural pictures, and in such cases achieving a restrained uniform was not the only way to produce the best perceptual experience.

3.3.4.2 Objective Results

Based on the Pearson and Spearman correlations between the mean Z-scores of perceived uniformity and objective results for all metrics. Obviously, none of the metrics worked well for all types of test images, especially for natural color images. Simple metrics, such as LR and SFA worked surprisingly better than others in many cases. We thought that it was because in our experiment the non-uniformity for all pixels was globally scaled, so the rank order of intensities in each primary color channel was largely preserved; although we apply negative scalars to non-uniformity as well, the magnitude of scaled non-uniformity was still comparatively smaller than the reference intensity values. However, in real practice, the non-uniformity level of projections should be relatively small. The metric Coeff also gave high correlations for patches but negative values for natural pictures. However, no metric worked well for the natural color images and slide like images. In such cases, the correlations were largely below 0.6. The metric S-CIELAB also adopted contrast sensitivity functions, but it had slightly better correlation results than PSNR-M and SSIM metrics in all cases. It was also interesting to figure out the reason why metric LR did not work well in many cases, so we generated the plots of the subjective results versus the objective results for the LR metric. It was clear that for the non-patch test images, the variance of metric scores were largely compressed and a few outliers were visible. By examining the metric scores, we found that these outliers corresponded to the flattened projection and natural projection. Similar phenomena could be observed for other metrics. It suggested that either that the metrics gave lower values for the flattened projection, or higher values for the natural projection comparing to their expected values. In other words, the distance between the two consecutive levels of perceived uniformity was more compressed than the results of metrics.

3.3.5 Conclusion

In this research, we proposed a series of methods to calibrate a camera, and used it as a relative acquisition device of intensity in order to evaluate the spatial uniformity of projection displays by incorporating image quality metrics. The experimental results suggested that none of the frequently referred metrics worked well for all types of test images, especially for the flattened projections and natural projections. In such cases, the spatial non-uniformity was largely masked by the high frequency components, and we should simulate the human visual system to ignore the non-uniformity that cannot be discriminated by observers. The simulation could be implemented by using models based on contrast sensitivity functions, contrast masking, etc. In addition, the colors could be considered to be transformed into the frequency domain and analyzed at a smaller granularity in order to engage the issue of contrast masking.

3.4 Paper D: Measuring The Relative Image Contrast Of Projection Displays

Ping Zhao, Marius Pedersen, Jon Yngve Hardeberg, and Jean-Baptiste Thomas

In *Journal of Image Science and Technology*
Society for Imaging Science and Technology
Volume 59, Issue 3, pp. 030404-1-030404-13
April, 2015.

3.4.1 Abstract

Projection displays, compared to other modern display technologies, have many unique advantages. In this paper, we propose an objective approach to measure the relative contrast of projection displays based on the pictures taken with a calibrated digital camera in a dark room where the projector is the only light source. A set of carefully selected natural images is modified to generate multiple levels of image contrast. In order to enhance the validity, reliability, and robustness of our research, we performed the experiments in similar viewing conditions at two separate geographical locations with different projection displays. In each location, we had a group of observers to give perceptual ratings. Further, we adopted state-of-art contrast measures to evaluate the relative contrast of the acquired images. The experimental results suggest that the Michelson contrast measure performs the worst, as expected, while other global contrast measures perform relatively better, but they have less correlation with the perceptual ratings than local contrast measures. The local contrast measures perform better than global contrast measures for all test images, but all contrast measures failed on the test images with low lightness or dominant colors and without texture areas. In addition, the high correlations between the experimental results for the two projections displays indicate that our proposed assessment approach is valid, reliable, and consistent.

3.4.2 Motivation

The main motivation was similar to the research for evaluating the spatial uniformity of projection displays 3.4. In the existing literature, contrast has been proven to be an important image quality attribute for displays [111, 108, 8, 154, 57]. For projection displays, the contrast was largely evaluated with a spectroradiometer, but not with the digital still camera. In this research, we continued to use the implementation of the core components of the image quality assessment framework. The main purpose was to evaluate the state-of-art of image quality metrics predicting the relative contrast, and benchmark them against the perceptual results with respect to their correlations to the perceptions. The image quality metrics were incorporated under the proposed framework. The results of the evaluation can be used to improve the design of image quality measures, and they can also be extended in the development and enhancement of general image reproduction technologies.

3.4.3 Methods

3.4.3.1 Contrast Measures

The contrast measures for images could be broadly classified into two categories with respect to their measurements at either the global or local level. With respect to the survey of existing literature, we adopted the image quality measures, such as Michelson contrast [120], RMS [129], Lab variance [139], RAMMG [150], RSC [160], and GCF [118], for the evaluation. The Michelson contrast measure was selected because it was representative of global contrast measurement and it was typically used as a reference for contrast measurement in research. RMS and LAB variance measures were selected because they were representative of measurements, which relied on statistics; however the RMS measure worked only on luminance, while the LAB variance measure further took colors into account in the perceptual uniform CIELAB color space. RAMMG and RSC measures were representative of the measures incorporating low-level visual system models. The GCF measure addressed the problem from the spatial frequency perspective.

3.4.3.2 Experimental Setup

In order to enhance the validity, reliability, and robustness of this research, we performed the experiments under the same viewing conditions but at two separate geographical locations with two different projection displays and one group of observers at each location. In this case, we had two separate experimental sessions in total. The first session was conducted in France with 10 observers, and we used a portable three-chip LCD projector Mitsubishi XL9 (1024×768 in pixels) to display images on the screen. The second session was conducted in Norway with 17 observers, and we used another three-chip LCD projector, a SONY APL-AW15 (1280×768 in pixels). We used the same DSLR camera Nikon D610 (6016×4016 in pixels) with a VR 18-100 mm F/3.5-5.6G (VR off) lens to capture the images. We selected 10 test images from the Colourlab Image Database: Image Quality [106] with respect to their image content (800×800 in pixels). We normalized the RGB values of all pixels in the test images, and transformed them in each color channel simultaneously with the formula

$$S_i = (C_i - m) * (j + 6)/6 + m, \tag{3.7}$$

where S_i stand for the scaled RGB value for the ith pixel in the distorted image, j is an integer scaling factor for contrast distortion in the range [-3, 3], C_i stands for the normalized input RGB value for the ith pixel in the input image, and m stands for the mean of all C_i in the same color channel, so we obtained seven distortion levels for each test image. Overscaled values (either larger than 1 or smaller than 0) were clipped.

3.4.3.3 Experimental Procedure

The subjective experiment was conducted by using the software QuickEval [126], which is an interactive software running on the controlling laptop for psychometric scaling experiments. All observers operated directly on the laptop, and they were experiencing exactly the same stimulates in identical viewing conditions. Based on this system, each observer was required to perform two assignments. In the first assignment, we displayed each group of distorted images at the same time in randomized order. The observers ranked them in a descending order with respect to their perception of contrast. In the second assignment, the distorted images were ranked in a descending order with respect to the observers' preference of contrast. All subjective ratings are scaled to generate Z-scores [44]. For the objective experiment, we used a camera as the acquisition device, and applied contrast metrics to the registered captured images [208]. In this process, we set the camera up with ISO 100, and performed the MTF test [175] to acknowledge that the best aperture was f/7.1. We adjusted the shutter speed setting iteratively to make sure that no camera sensor was either underexposed or overexposed. We captured images in raw format, and applied the spot white balance algorithm to correct the captured colors. We incorporated the method proposed in our previous research to eliminate vignetting effect [209]. In the experiment, we incorporated the contrast measures, such as Michelson contrast [120], RMS [129], Lab variance [139], RAMMG [150], RSC [160], and GCF [118], into the proposed image quality assessment framework.

3.4.4 Results

3.4.4.1 Subjective Results for Ranked Perceived Contrast

Based on the observation of the Z-scores of ranked perceived contrast for all projectors, it was clear that the rank of perceived contrast has a closely linear relationship with the actual rank of modified contrast. Since the Z-score values in all plots were monotonically increasing, the relationship between perceived contrast and the actual image contrast was almost linear for all types of images.

3.4.4.2 Subjective Results for Preferred Perceived Contrast

The general tendency of the Z-scores of preferred contrast did not follow a linear relationship with the actual image contrast. This observation suggested that the observers tend to rank all distortions into two groups: either relatively less preferred (contrast level -3 to -1) or more preferred perceived contrast (contrast level 1 to 3). In the group of less preferred contrast, since the confidence intervals of Z-scores were largely overlapped, the perceived contrasts had no significant difference, while in the group of more preferred contrast, the confidence intervals were less overlapped. This suggested that the majority of observers prefer the enhanced contrast even though the luminance might be overscaled. In some cases, for both projectors at the contrast level 0 (original image) was neither preferred nor not preferred because it was very close to the center line for all test images. The preferred perceived contrast values for the two projectors were obviously different.

3.4.4.3 Objective Results for Ranked Perceived Contrast

We calculated the Pearson correlation coefficients between the metric scores and the mean Z-scores of ranked perceived contrast. It was clear that, for the Mitsubishi projection display, most contrast measures produce high correlation coefficients above 0.85 for most images, except that the RMS and GCF measures produced low coefficients on test image 6. However, the observation cannot be obtained from the SONY projector. For the SONY projection display, the Michelson contrast measure performed worst. Other contrast measures had similar performance for both projection displays on test images 2, 3, 4, 5, 7, 8, 9, and 10, but not on test images 1 and 6. For the Mitsubishi projection display, the contrast measure GCF performs badly with respect to its confidence interval. Although both projection displays were supposed to produce different contrast on the screens, the mean of correlation coefficients over all test images followed a very similar general tendency. Based on the observation on the variance of confidence intervals, the RSC contrast measure produced the most stable outcome regardless of the image content.

3.4.4.4 Objective Results for Preferred Perceived Contrast

It was clear that the Michelson contrast measure performs the worst. In addition, the RMS and GCF measures both performed relatively worse for test image 6 for the two projection displays as well. For the preferred contrast of both projection displays, the RAMMG and RSC had the highest correlations; however, the correlation from the RAMMG was slightly higher than that for the RSC contrast measure. This observation was different from the one for ranked perceived contrast. The rank order between RMS, LAB, GCF, RAMMG, and RSC was largely preserved for test images 2, 3, 4, 5, 7, 8, 9 and 10, but not for test images 1 and 6. This observation could be obtained from the ranked perceived contrast for both projection displays as well, but not from the preferred contrast for the SONY projection display. By looking at the average overall contrast measurement performance, the general tendency of the average Pearson correlation over all test images was almost the same as the one obtained from the preferred contrast.

3.4.4.5 Overall Results

We determined the average performance of the contrast measures over all test images for each projection display. Then, we calculated the Pearson correlation coefficients not on a per image basis but over all test images, so we could observe the metric performance regardless of the image content. We also calculated the Pearson correlations between the average performances over all contrast measures with respect to their types of contrast versus the types of projection displays. With respect to the results, it was concluded that the most preferred perceived contrast corresponds to the highest ranked perceived contrast: even for test images 1, 2, 4, and 5 the highest preferred perceived contrast corresponded to the second highest ranked perceived contrast. For related research in the future it is unnecessary to explicitly distinguish them and do the experiments twice.

3.4.5 Conclusion

In this research, we proposed an objective approach to measure the relative contrast of projection displays in a controlled environment by using a calibrated digital camera. To our best knowledge, it was the first of this type of research. The approach could be easily extended to measure other image quality attributes, such as sharpness and nonuniformity, for all types of displays. The experimental results based on two separate projection displays suggest that the Michelson contrast measure had very low performance over all test images, as expected. Other global contrast measures (RMS and LAB) performed relatively better than the Michelson contrast measure, but they had less correlation with the perceptual ratings compared to the local contrast measures. The local contrast measure GCF had similar performance to the RMS and LAB measures, but it performed worse than other local contrast measures (RAMMG and RSC). The contrast measures RAMMG and RSC performed the best overall, and have very close performance on contrast measurements for almost all test images. With respect to the 95% confidence interval of the average measurement performance over all test images, RAMMG had slightly improved correlations with the preferred contrast over other metrics. Many contrast measures did not perform well on the test images 1 and 6. These two images either had large area of low luminance component or dominant color component, and they did not have obvious texture area. We recommended local contrast measures incorporating low-level human visual system models, since they had better overall performance over global contrast measures in terms of both contrast prediction accuracy and stability regardless of the image content. Since the average correlations and stability of local contrast measures were good for many test images, we did not need to propose a new contrast measure, but rather to improve the models of human visual system to predict the image contrast better in future research.

3.5 Paper E: Measuring Perceived Sharpness of Projection Displays with a Calibrated Camera

Ping Zhao, and Marius Pedersen

Submitted to *Journal of Visual Communication and Image Representation*.

3.5.1 Abstract

Perceived sharpness is one of the most important image quality attributes for displays, because it determines how much details humans are able to perceive on the screen at certain distances. However, this attribute was not well studied for projection displays in the existing literature. In this paper, we conduct an experimental study on measuring perceived sharpness of projection displays based on the pictures taken with a calibrated camera in a darkroom, and evaluating the performance of state-of-art sharpness metrics accordingly. The basic idea is to apply Gaussian filtering to natural test images in order to simulate the optical blurring process of projection systems, so that we can generate multiple levels of image sharpness in a controlled manner without influencing the original properties of projection displays. We project these filtered images onto the screen and invite a group of human observers to give perceptual ratings on them. We calculate the correlation coefficients between perceptual sharpness and the one measured with state-of-art image quality metrics. We find out that the average performance of full reference metrics are comparatively better than the reduced and no reference metrics. Among the full reference metrics, SSIM, VIF and FSIM metrics perform well in terms of both accuracy and stability.

3.5.2 Motivation

The main motivation was similar to the research of evaluating the spatial uniformity and contrast of projection displays. In the existing literature, sharpness has been recognized as an important image quality attribute for displays [55, 190, 28]. For projection displays, to our best knowledge, there was no such research related to using digital still camera to do the image quality assessment. In this research, we continued to use the implementation of the core components of the image quality assessment framework. The main purpose was to evaluate the state-of-art of image quality metrics predicting the sharpness, and benchmark them with respect to their correlations with the perceptual results. The sharpness metrics could be broadly classified into three categories: full reference based, reduced reference based, and no reference based. However, the existing research largely focused on only one of them. In this context, one interesting research to do could be comparing the performance of all types of metrics by referring to the identical perceptual data, and making conclusion that which category of metrics has more advantages over others. In addition, we could also compare the metric performance within each category to see which one performs better and figured out the reason behind.

3.5.3 Methods

We used a portable three chip LCD projector SONY APL-AW15 (throw ratio 1.5) to produce projections (1920 × 1080 in pixels, approximately 2 × 1.2 in meters on screen) on a planar screen. The projector was put on a table placed in front of the projection screen about 3m away. We used a DLSR camera Nikon D610 (6048 × 4016 in pixels) with a Sigma VR 24-105mmf/4G (VR off) lens to capture projections. It was mounted on a tripod and placed in front of the screen about 4m away. Pictures were taken remotely in raw format with a software control on the laptop without physically touching the camera, and processed with aliasing minimization and zipper elimination algorithm [117]. We selected seven test images from the Colourlab Image Database: Image Quality [106] to generate six levels of Gaussian blur with kernel size 11 and standard deviation 0, 0.5, 1, 1.5, 2, and 3 respectively. We incorporated the method from our previous research to eliminate the vignetting effect [209], and optimize camera settings in order to ensure the linear response of camera sensors [209], and register the projections in captured images with their original image content [208].

We invited 15 observers to give perceptual ratings to the perceived sharpness of projected image distortions. Each of them sat on a chair, which was placed at the camera position. The viewing condition was similar to a home theater- like environment, where the room was completely dark and the visual angle was about 15 degrees. The blurred test images were displayed in a randomized order for every observer, and each time only one of them was displayed. The experiment is set up with category judgment method. For each displayed image, the observers were asked to indicate the overall perceptual sharpness with a category label, which stands for the rank between ”no blurring at all” and ”completely blurred” corresponding to the ratings numbers ranging from 1 to 9 respectively. We adopt eleven representative image quality metrics in all three categories: SSIM [185], VSNR [27], VIF [156], FSIM [204], RRIQA [186], RRED [163], LPC-SI [63], S-Index [96], CPBD [124], JNBM [50], and S3 [184]. The metric performance were evaluated with respect to the Pearson and Spearman correlation coefficients between the metric results and the mean Z-scores of perceptual ratings.

3.5.4 Results

3.5.4.1 Subjective Results

The perceptual ratings were collected from human observers, and they were scaled to generate Z-scores [44]. It was clear that the perceived sharpness decreased while the blur level increased. However, their relationship should not be simply interpreted with a linear regression model, since the Z-scores for test image 1, 4, 6 and 7 appeared to have a flat region between the first and second blur levels. This observation suggested that there was a lower bound threshold for observers to detect the sharpness changes. Another observation was that the general tendency of Z-scores for all test images were fairly similar, and their value ranges were almost identical. Investigation of the overall results incorporating all test images showed differences in the agreement between observers. For example, the variance for the blur level 4 was larger than others in test image 1, and also the variance for the blur level 1 in test image 5. However, the one or two outliers were minorities comparatively to all human observers in such cases. This observation suggested that the observers had agreements regarding perceptual sharpness despite of image content.

3.5.4.2 Objective Results

We calculated the Pearson correlation coefficients between the objective and subjective sharpness for each test image. The purpose was to understand how well the metrics perform with respect to specific image content. It is clear that in most cases the correlation coefficients are larger than 0.85; especially, for the SSIM, VIF, FSIM and LPCSI metrics, the correlation coefficients were above 0.9 for all test images. In addition, the top rank metrics had fairly close performance for most of the test images. From this perspective, the state-of-the-art image quality metrics had good correlations with perceptual sharpness in general. It was interesting to see which metric performs the best, and it could be more interesting to figure out the root causes. We generated the plots of objective sharpness versus perceptual sharpness for the VSNR, RRED, RRIQA, SIndex, CPBD and S3 metrics for specific test images. It was clear that the VSNR and RRED metrics had inconsistent rank orders on measured sharpness for test image 2 and 3. For the test images 4, the RRIQA, SIndex, CPBD and S3 metrics all reserve the rank orders well, but they had inconsistent derivative of curve between consecutive distortion levels. For the strongly blurred images, the derivatives were less than expected; for the slightly blurred images, the derivatives were larger than expected. In contrary, the VIF and FSIM metrics performed very well for test images 4 in both terms of rank order and derivatives.

3.5.5 Conclusion

In this research, we conducted an experimental study of perceived sharpness on projection displays in a home-theater like dark room. The perceptual results suggested that the perceived sharpness followed a nonlinear tendency pattern, but its rank order remained the same as the blur level increases. The correlations between the metrical and perceptual results indicated that SSIM, FSIM and VIF metrics give excellent prediction performance in most cases in terms of both correlation and its variance. According to the group comparison, full reference sharpness metrics had comparatively better prediction performance than reduced reference and no reference metrics. In the coming future, we should turn to focus on the design of a good sharpness metric based on the VIF metric for projection displays following the hints that we obtained from this research.

3.6 Paper F: Extending Subjective Experiments for Image Quality Assessment with Baseline Adjustments

Ping Zhao and Marius Pedersen

In Mohamed-Chaker Larabi & Sophie Triantaphillidou, eds. *Image Quality and System Performance XII, Proceedings of 27th Annual Symposium on Electronic Imaging*

The International Society for Optics and Photonics

San Francisco, CA, USA

pp. 93960R-1 - 9360R-13

February, 2015.

3.6.1 Abstract

In a typical working cycle of image quality assessment, it is common to have a number of human observers to give perceptual ratings on multiple levels of distortions of selected test images. If additional distortions need to be introduced into the experiment, the entire subjective experiment must be performed over again in order to incorporate the additional distortions. However, this would usually consume considerably more time and resources. Baseline adjustment is one method to extend an experiment with additional distortions without having to do a full experiment, reducing both the time and resources needed. In this paper, we conduct a study to verify and evaluate the baseline adjustment method regarding extending an existing subjective experimental session to another. Our experimental results suggest that the baseline adjustment method can be effective. We identify the optimal distortion levels to be included in the baselines should be the ones of which the stimulus combinations produce the minimum standard deviations in the mean adjusted Z-scores over all human observers in the existing rating session. We also demonstrate that it is possible to reduce the number of baseline stimuli, so the cost of extending subjective experiments can be optimized. In contrast to conventional research mainly focusing on case studies of hypothetical data sets, we perform this research based on the real perceptual ratings collected from a subjective experiment.

3.6.2 Motivation

Suppose that in an existing subjective experimental session, we have ratings of four levels of image distortions; later, two additional distortions need to be introduced to extend the existing subjective experiment. In this case, conventionally, we have to conduct a new experiment with all six levels of image distortions. The overall amount of workload can be demanding. In addition, in practice, the human observers involved in the new session are unlikely to be identical to those who participated in the existing session. However, if we have enough human observers, the averaged ratings regarding the existing image distortions should be statistically similar across the two sessions. In this context, a natural research question to ask would be is it possible to take this advantage without executing the entire experiment over again, especially when the amount of distortions is large. Baseline adjustment can be a potential answer to this research problem. This method introduces common stimuli (one or more distortions) to form a baseline in order to determine the comparability of ratings between different experiment sessions, and allows the computation of scale values expressed relative to responses for the baseline stimuli [23]. The baseline adjustment is carried out separately for each original stimulus. Many existing studies had introduced baseline adjustments into their scaling procedures [35, 66, 21, 23, 22, 65]. However, in these experiments, either the selection criteria of baseline was not discussed in depth [35, 66, 21, 23], or the baseline stimuli were simply selected randomly from existing candidates [22, 65]. A natural research question to ask is what types of stimuli should be included in order to form a representative baseline, and how many stimuli are essential? In this research, we conducted a study to verify and evaluate the baseline adjustment method for extending subjective experiments. The first goal was to verify that the baseline adjustment is an effective method, and the second was to identify the type and number of stimuli that we should use in the common baseline in order to minimize the experimental workload and complexity.

3.6.3 Methods

Compared to conventional research focusing on case studies of hypothetical data sets, we performed our research based on real perceptual data, which were collected from subjective experiment regarding perceptual spatial uniformity evaluation. We used the data to simulate real scenarios. The natural non-uniformity for every pixel was scaled into seven levels, and then we stacked them onto eight selected test images respectively; so 56 stimuli in total were shown to each observer in a completely randomized order. We conducted the experiment in a control lab environment, where we tried to simulate a home theater-like environment and avoid unwanted imaging artifacts. In this case, two calibrated SONY APL-AW15 LCD projectors (throw put: 1.5) were placed right in front of and about 3m away from a planar screen to produce two projections (both 1.5×0.9 in meters) in parallel. 20 human observers were invited to do the experiment; 13 of them had color science background, while the rest did not. 14 of them were male and the rest were female. All of them were required to have a mandatory visual acuity test. The observers were asked to sit in front and between the two projections displayed in parallel. The visual angles were around 20 degrees and the viewing distance was approximately 4 meter. Each observer was asked to use a natural number between 0 (corresponding to completely uniform) to 10 (corresponding to not uniform at all) to indicate his/her opinion regarding the overall magnitude of perceived spatial non-uniformity. All observers were required to do the experiment twice, which resulted in 2240 perceptual ratings in total.

In this research, we separated the image distortions into two groups (Figure 3.1). For example, the first group included distortion level 1 to 6, and the second group included distortion level 4 to 7. The ratings for distortion level 1 to 3 in the session 2 were ignored. In this case, we were simulating a scenario which extended existing subjective experiment with distortion level 1 to 6 in order to adopt additional distortion level 7; and the ratings for distortion level 4 to 6 were used to form the adjustment baseline for Z-score scaling in both rating sessions. Since the ratings were scaled on observer basis, we assumed that we were scaling the ratings for ″Image 1″ from ″Observer 1″. The ratings in ″Part 1″ were scaled with respect to the ″Baseline 1″ in order to generate scaled ratings in ″Adjusted Part 1″, the rating in ″Part 2″ was scaled with respect to the ″Baseline 2″ in order to generate scaled ratings in ″Adjusted Part 2″, and the ratings for distortion level 4 to 6 on ″Image 1″ from ″Observer 1″ in ″Session 1″ were scaled with respect to ″Baseline 1″ in order to generate ″Adjusted Ratings in Baseline 1″. Notice that the two baselines shared the same distortion levels, but they might have different rating values. Each baseline included only the ratings from ″Observer 1″ for all test images on the corresponding distortion levels. Eventually, all adjusted ratings in the table below were merged to generate a full set of adjusted Z-scores. Then the mean adjusted Z-scores over all observers were correlated with the non-adjusted Z-scores over all observers in original ″Session 1″ to determine the performance of the underlying baseline. Since the ratings for the two original sessions were collected from the identical observers in the same circumstance, the average correlations are expected to be high if the baseline was appropriately specified. In this context, we calculated both Pearson and Spearman correlations. Obviously, there were many possible combinations of distortion levels and unique distortion levels among the two sessions, so we wrote a computer program to permute all combination possibilities and calculate corresponding correlations accordingly.

3.6.4 Results

The experimental results were presented in two parts. In the first part, the average correlation results and analysis were presented, the results could be regarded as image content

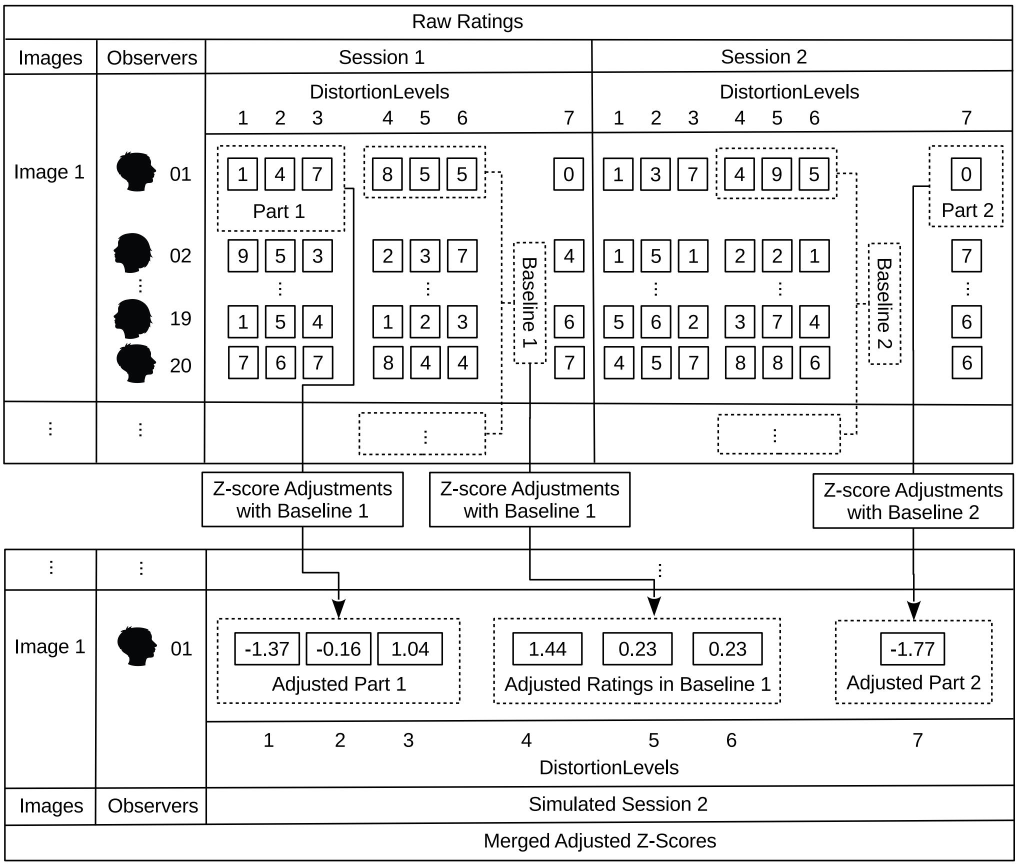


Figure 3.1: The approach we proposed to study the validity and reliability of baseline adjustment for scaling subjective ratings.

independent. In the second part, the correlation results regarding individual image content were presented. Regarding the overall correlation results, we elevated the observation to a higher level, where we focused on the average correlations over all human observers and all test images. The purpose was to identify the general tendency which enables the analysis on how the correlation values associated with the number of distortion levels included in baselines, despite of individual image content. So, we generated all possible distortion combinations and calculated the average over all test images. Then, it was found that both the mean Pearson and Spearman correlation values were monotonically increasing, while the standard deviations were shrinking, as the number of distortion levels included in the baseline increased. We concluded that, in general, despite the image content, the more distortion levels we adopted in the baseline, the better and more stable correlations we should have between the existing ratings and the expanded ratings; to a certain distortion level combination included in the baseline, no matter how the unique distortion levels in two session varied, the correlation values tend to less variant. However, meanwhile, the workload of repeating the subjective experiments increased as well. Notice that, in each category, the correlation values could rise up close to one. In other words, it was possible to achieve high correlations without adopting all distortion levels in the baseline. Then the question arises as to what were the most optimal distortion levels to be included in the baseline, so the correlation values were as high as possible despite the actual combinations of unique distortions levels involved for the existing and expanded sessions. Regarding the correlation results for individual images, we found that the most optimal distortion levels to be included in the baseline did not necessarily correspond to the highest possible correlation values, but they should produce the least correlation variance no matter how the unique distortion levels are combined. In other words, with the presence of an optimal baseline, the unique distortion level combinations had limited influence on the correlation variance. In some cases, with a non-optimal baseline, the correlation might happen to have the highest values, because the perceptual ratings were fuzzy in nature and the highest correlation might be caused by random rating noises. This theory could be also supported by the observations of mean and standard deviation of correlations. We came to the conclusion that the most optimal baseline gave the lowest possible standard deviation on the correlation results. These optimal distortion levels should be included as many as possible to improve the correlation results, and they were image content dependent.

In a typical situation of extending subjective experiments, we have only the raw ratings for existing image distortion levels but not the ones for additional distortion levels. In this case, we cannot relie on the correlations between different rating sessions to determine the optimal baselines. However, we can use each of the known image distortion levels as a baseline and calculate the adjusted Z-scores of the rest of ratings. Then, we determine the correlation between the adjusted Z-scores and the original non-adjusted Z-scores to find out which baseline is most optimal with respect to the method described in the previous paragraph. This is an approximation approach because in this context we implicitly make an assumption that the newly introduced image distortions have limited influence on scaling the existing ratings. In other words, mean and standard deviation of selected baseline in the new session are expected to be close enough to the ones in the existing session, and the number of additional distortion levels should be small. Ideally, we should introduce one additional image distortion at one time. If two or more are required, then the researcher should adopt them one by one in an iterative fashion. The validity of this constraint is supported by the fact the best correlations are always associated with the cases that only one additional image distortion is introduced at a time in the experiments.

3.6.5 Conclusion

In this research, we conducted a study to verify and evaluate the baseline adjustment method regarding extending subjective experiments from one existing session to a new session. The experimental results suggested that the baseline adjustment method works effectively because both Pearson and Spearman correlations gave high values once the optimal baseline was specified. We identified the most optimal baseline should include the combinations of image distortion levels that produce the minimum standard deviation of the mean adjusted Z-scores over all human observers in the existing rating session. We demonstrated that it was possible to reduce the number of image distortion levels included in the baseline, however the trade off is the lost confidence of Z-score correlations between the two rating sessions.

Chapter 4

Discussions of Papers

> Attempt the end and never stand to doubt; Nothing's so hard, but search will find it out.
>
> ROBERT HERRICK

In the previous chapter, the individual included paper has been summarized with respect to the corresponding discussions and conclusions. In this chapter, the contribution of the included papers will be discussed in context and the relationships between them will be demonstrated. Then, a conclusion of the research contributions and possible ideas for future works will be presented.

4.1 Paper A: Camera-based Measurement of Relative Image Contrast in Projection Displays

In this research, it was our first attempt to use a digital still camera as the acquisition device of static images displayed on the projection screen, and evaluate their image quality objectively with state-of-art image quality metrics. Since there was no standard procedure for using a camera to perform such a task that can be referenced in the literature, it was the best option for us to set up an experimental environment, simulate typical viewing conditions of projection displays, and try to establish an image quality assessment workflow. In this process we were able to identify and recognize the potential research challenges and come up with corresponding solutions. As a result, absolutely no uniformity correction, camera optimization, restricted light controlling, color correction, image registration, and subjective evaluation were involved in this image quality assessment process. Due to this fact, we decided to use the camera to evaluate the contrast attribute of projection displays. For one thing, the four image quality metrics we incorporated counted in only luminance information to predict image contrast at either a global or a local level. For another, the relative luminance measurement was a relatively easier task for a camera to perform. Although the patches displayed on the screen might not be fully uniform, the ratios of measured luminance at different spatial locations were expected to be preserved despite of the camera settings. In other words, the camera was uncalibrated, however its impact on the measured contrast was assumed to be limited at the current stage of research. We used typical test images for contrast evaluations, such as checkerboard patterns, gray scale images, and color complex images, in the experiments; it was because we had no prior information of how the image content influence the final assessment results. In this research, we displayed the projections naturally as they were without any correction or enhancement, but with a range of projector brightness and contrast settings. Although we had no prior information regarding the tone response characteristics of the projection display, the measured image contrast was expected to monotonically increase while either the value of projector brightness or contrast setting increases. In this manner, even without projector calibration, we could be able to generate multiple levels of contrast distortions for each test image. Thus, a contrast surface could be produced for each image quality metric. Based on the plots of these contrast surfaces, we could observe the interactions between the measured contrast based on the camera and the projector settings. One trade off had to be made in this context. Each image quality metric actually had its own independent contrast space, in which the range and scaling of predicted contrast might be significantly different. However, their relationships were not likely possible to be quantified precisely. Alternatively, we normalized the contrast values predicted by each image quality metric, and compared the general tendency of contrast surfaces extensively among different image quality metrics. In the experiment, the most important observation was that the rank orders of camera based measured contrast for all image quality metrics were preserved for large steps of projector setting differences. However, we still lack information on how the image quality metrics would perform for small steps of differences. This requires a greater number of image distortion levels, which were expected to be considered as future works. The experimental method in this research was naive, however the results provided evidence to support that the camera based image quality assessment approach actually worked. Then, the next stage of the research should be investigating how to improve the camera based method by introducing uniformity correction, camera optimization, image registration, etc. In this approach, the image quality assessment results were expected to be more accurate, more consistent, more reliable and more robust. All these factors were important to establish a good image quality assessment framework design, and the research upon them were conducted in the upcoming research presented in **Paper B**, **Paper C**, **Paper D**, and **Paper E** respectively.

4.2 Paper B: Image Registration for Quality Assessment of Projection Displays

There is a huge amount of image quality metrics in the existing literature, based on which either an extension can be created or a relatively new research can be conducted. Thus, it can be elegantly satisfying to have an unified image quality assessment framework to incorporate all the metrics without modifying them, and robustly evaluate the performance of existing or newly introduced ones against the state-of-art. It is known that the image quality metrics can be broadly classified into full reference, reduced reference and no reference categories. Among them, the full reference metrics require to establish an accurate correspondence of pixels between one acquired image reproduction and its original. The dimension and resolution of the two images must be exactly the same; meanwhile, the preservation of geometrical order as well as the intensity and chromaticity relationship between consecutive pixels on the displays must be maximized. Thus, an accurate image registration method is required. Although one projection display was typically shared by multiple observers in the field, the image quality of the display is only possible to be optimized for one observer at one location. However, this sweet spot is unknown in advance of image registration design and it may vary over time. Hence, the underlying image registration method must be fast, robust, and view independent. In order to achieve the goals, we proposed a computer vision based method for uncalibrated camera. The basic idea was to take the advantage of a series of predefined light patterns with known geometric information to detect the geometric distortion of captured images, and the distortions were encoded in a cubic spline based distortion free coordinate system; then the distortions of all subsequently captured pictures were corrected right after the pictures were taken. Correspondingly, the proposed image registration method could be decomposed into three major components: feature extraction, feature expansion and geometric correction. The feature extraction component corresponded to the computer vision based geometric distortion detection. The feature expansion component corresponded to the cubic spline based distortion free encoding. The geometric correction component essentially incorporated existing image interpolation algorithms to correct the spatial distortions with respect to the encoded distortion information, so it could be performed very efficiently. The experiments were conducted from two perspectives. In the first part, we generated many artificially distorted images by rotating, translating, and scaling the original test images independently, then applied the proposed image registration method. In this case, since all distortions were known, we could evaluate the image registration performance in terms of minimum, average and maximum absolute pixel shift errors. In the term of maximum shift errors, our proposed algorithm could achieve less than 0.2 pixel for scaling with factor larger than 2 times, and up to 2.5 pixels for all types of rotations. For translation, the proposed method was totally independent from this type of spatial distortion. These evaluation results were independent from the captured image content. We also determined the magnitude of structure changes with respect to SSIM metric in order to evaluate the dependence of registration errors against scaling and rotations. It was found that if the image resolution was at least two times higher than the projection resolution the average structure loss was less than 0.06 for all types of scaling and less than 0.02 for all types of rotations. In the second part of the experiment, we used the camera to take many pictures of the projection in the field, at completely random locations, with random viewing angles and orientations. Since we had no ground truth in this case, we evaluated the image registration method in terms of structural similarity. The experimental results of both parts suggested that our proposed method was able to achieve registration accuracy higher than 91% in a dark room and above 85% with ambient light lower than 30 Lux. From this point of view, the proposed method produced less errors in the lower light condition. After a careful examination, the main source of image registration errors was identified to be the projection contour detection algorithm. In some cases, the algorithm did not work well to discriminate the projection area from the background due to the color aberration along the edges. The root cause was that the demosaicing algorithm did not handle the optical transaction from the background to the foreground well. Although the proposed image registration method accounted for only the luminance information to perform, the color aberrations issues led to unstable luminance variations along the projection boundaries. From this perspective, one possible way to improve the method could be by taking multiple shots of one projection with the camera, modeling the noises of camera sensor responses under the current viewing conditions with respect to statistical analysis, and increasing the stability of measured data by post-processing the data. Another possible improvement can be incorporating one additional instrument (e.g. spectroradiometer) to characterize the background intensity, and use the information to discriminate the foreground from the background independent of the camera based background detection. The proposed image registration method was originally designed for projection displays, however it can also be applied to other types of flat panel displays with limited modifications.

4.3 Paper C: Perceptual Spatial Uniformity Assessment of Projection Displays with a Calibrated Camera

Spatial uniformity had long been recognized as one of the most important image quality attributes for projection displays. Compared to conventional research which largely reliess on spectroradiometers to measure the averaged intensity response over several spot areas at discrete spatial locations on the displays, using a camera as the acquisition device of displayed images has the advantage of quickly recording the intensity information of all pixels in one shot. The goal of this research was to evaluate the uniformity of projection displays, so the nonuniformity impact introduced by the camera's optical and electronic subsystems should be minimized. For this purpose, we proposed a method to take advantage of a hazy sky as a nearly uniform light source to quickly create a vignetting mask for the underlying camera lens, and applied them to all subsequently captured projection images in order to eliminate the vignetting effect. One additional benefit of using this method was that the dust shading effect could be eliminated as well, as long as the shadings of dust particulates in the captured images appeared to be semi-transparent to the incident light coming from all directions. With respect to the experimental data, it was found that empirically 10 pictures of the hazy sky were sufficient to create a good vignetting mask. In addition, we found that the camera sensors did not necessarily always give linear responses with different combinations of shutter speeds and luminance magnitudes of incident light. Based on the observation of our experimental data, we proposed an iterative method to quickly determine the best camera settings in order to optimize linearity of camera sensor responses. In one of our previous experiments, we had proposed an image registration method to eliminate the geometric distortion of captured images. By integrating this method into the current image quality assessment workflow, it was possible to compensate the intensity transfer functions of projection displays in order to produce multiple levels of linearized nonuniformity on the screen for the purpose of image quality assessment. In the experiments, we incorporated several state-of-the-art image quality metrics measuring the image uniformity, and benchmark them with respect to their correlations with perceptual results. It was found that none of these metrics worked well with all types of test images, as expected. The main issue was that the spatial non-uniformity was largely masked by the high frequency components in the displayed image content. In this case, the image quality should simulate the human visual system to ignore the minor non-uniformity that cannot be discriminated by observers. The underlying simulations could be implemented by using models based on contrast sensitivity functions, contrast masking, etc. Using a digital still camera to assess the spatial uniformity of projection displays has several advantages by incorporating our proposed methods, however this research also had several limitations. For example, modern cameras commonly have very high image resolutions (up to 50 million pixels in the state-of-art). In this case, calculating the medians for all pixels over many captured hazy sky pictures can be computationally inefficient, if the image processing software can not handle the parallel computing well. GPU based computing can be an alternative way to implement the proposed method. Another limitation of this research is that we have found the high frequency components are important to uniformity evaluation, but we did not really dig into the details. Especially those of determining the just-noticeable-difference of spatial uniformity based on perceptual data analysis. However, this part of the work requires a considerable amount of workload for greatly increasing the number of uniformity distortion levels and the resources to conduct a larger scale subjective experiments. For this reason, we may consider the related research as a part of the future work.

4.4 Paper D: Measuring The Relative Image Contrast Of Projection Displays

Contrast is among the most important image quality attributes of image reproductions in many different research domains, such as photography, printing, medical imaging, and display imaging etc. It defines the magnitude of presence of an object that can be recognized in a scene. In the past, due to the lack of knowledge of biological structure and information processing procedure of the human visual system, contrast was simply defined as a ratio related to the highest and lowest measured luminance in an image, known as Michelson contrast. Inspired by this formulation, several alternative formulations defined contrast in similar ways, such as Weber fraction, root-mean-square, and LAB variance. Since these formulations measured image contrast based on only two extreme pixels, in such cases the contrast was defined at a global level. Recently, more contrast formulations were proposed based on not only the luminance components but also the chrominance components, such as Weber-Fechner, RAMMG, and RSC. These image quality metrics increased the contrast measurement granularity and calculate the contrast at a local level. In such cases, it was common to simulate the capabilities and behaviors of human visual systems based on numerical calculations. In one of our previous studies (**Paper A**), we evaluated the relative contrast of projection displays based on a digital still camera. In this research, we incorporated our proposed image quality assessment framework to evaluate the performance of state-of-the-art contrast measures based on a set of carefully selected natural images. In order to enhance the validity, reliability, and robustness of our research, we performed the experiments in similar viewing conditions at two separate geographical locations with different projection displays. In each location, we had a group of observers give perceptual ratings. Further, we adopted state-of-the-art contrast measures to evaluate the relative contrast of the acquired images. The experimental results suggested that the Michelson contrast performs the worst, as expected, while other global contrast measures perform relatively better, but they had less correlation with the perceptual ratings than the local contrast measures. The local contrast measures performed better than global contrast measures for all test images, but all contrast measures failed on the test images with low luminance or dominant colors and without texture areas. In addition, the high correlations between the experimental results for the two projections displays indicated that our proposed assessment framework was valid, reliable, and consistent.

4.5 Paper E: Measuring Perceived Sharpness of Projection Displays with a Calibrated Camera

Perceived sharpness determines how much detail the observers are able to perceive on the displays at a certain distance. Assuming that the human visual system is consistent for sharpness perception over time, then the perceived sharpness is correlated with the sharpness reproduction capability of the displays. From this point of view, sharpness is important for assessing the image quality of displays with respect to their perception. However, this part of research was not engaged in the past, particularly for projection displays. In this research, we incorporated our proposed image quality assessment framework to perform the task and evaluate the performance of state-of-the-art sharpness metrics accordingly. In the experiments, the selected test images were blurred with a Gaussian filter to generate multiple levels of sharpness distortions. The purpose was to simulate the optical blurring process of a projection system without influencing its natural image quality properties. These blurred images were displayed in a home-theater like dark room. In the objective manner, several state-of-the-art image quality metrics measuring sharpness were used, while a group of observers were invited to give perceptual ratings. The correlations between the metric results and the perceptual results suggested that full reference metrics had comparatively better performance than the reduced reference and no reference metrics. Among the full reference metrics, SSIM, VIF and FSMI metrics performed well in both terms of accuracy and stability.

4.6 Paper F: Extending Subjective Experiments for Image Quality Assessment with Baseline Adjustments

In a typical working cycle of image quality assessment, it is common to have a number of observers give perceptual ratings on multiple levels of artificially distorted test images. In some cases, the number of distortion levels in an existing subjective experiment session may be found to be insufficient. For example, with no prior knowledge regarding a specific image quality attribute, additional distortion levels will be needed in order to increase the granularity of estimation of the just noticeable difference. In this case, a completely new subjective experiment including the existing and newly added levels of distortion need to be conducted. However, the new experiment may consume considerable time and resources to conduct. In this research, we investigated the possibility of using baseline adjustment method to extend the existing subjective experiment. One purpose was to first verify that the baseline adjustment method worked. Another purpose was to identify the most optimal distortion levels to be included in the baseline, and the number of them. For this purposes, we designed an experiment to incorporate the subjective data collected in an existing image quality assessment experiment, and divided to data into two groups. In this process, we made sure that there was an overlapping of distortion levels between the two groups of data. The subjective ratings for in overlapped region were used as the baseline, however the ratings for the baseline in the two groups of data might not necessarily be exactly the same. After this, we adjusted the subjective ratings for the unique distortion levels in the two groups of data with respect to their own baseline respectively. All subjective ratings were originally collected from one original experimental session. From this point of view, if the selected baseline was optimal, then the Z-scores of adjusted subjective ratings for the two groups of data were expected to have a high correlation. Compared to conventional research regarding baseline adjustment method, we used the real subjective experimental data rather than self-generated hypothetical data. With respect to the experimental results, it was found that the most optimal distortion levels to be included in the baseline should be the ones in which the stimulus combinations produce the minimum standard deviations in the mean adjusted Z-scores over all observers in the existing subjective experiment. Although it was the best for the baseline to include all distortion levels, it was possible to reduce the number of baseline stimuli. The trade off in these cases would be the reduced confidence interval of correlation values.

4.7 Discussion of Papers in Context

The main goals of this research were developing a digital still camera based image quality assessment framework for displays, and using it to evaluate the state-of-art image quality metrics regarding the most important image quality attributes for projection displays. Since the research project has a limited time frame, we decided to focus our research on projection displays. In order to achieve our goals, the research conducted in this thesis could be broadly divided into three major components: framework design, performance evaluation of state-of-the-art image quality metrics, and extending subjective experiments (Figure 4.1).

The core contribution was the digital still camera based display image quality assessment framework, which was not bound by theories but a practice oriented design. For this reason, at the very beginning of the PhD research, we evaluated the contrast of projection displays based on the acquisitions of a digital still camera (**Paper A**). It was our first attempt to set up a home theater-like dark room environment and perform the image quality assessment accordingly. In this process, there was no projector calibration, camera calibration, image registration, and subjective experiment involved at all. The main purpose was to observe what research challenges we might encounter, then we could come up with corresponding solutions. An important finding from this part of the research was that the camera based image quality assessment method worked, since the rank orders to measured contrast were preserved for large steps of projector settings. Inspired by this positive finding, we decided to extend the research further. Although it was not stated in the research paper, but we did establish a naive image quality assessment workflow based on a digital still camera. The next research question to ask was about how to improve this workflow and produce a concrete full framework.

In the existing literature, the image quality assessment frameworks incorporating reduced reference and no reference metrics have been addressed. However, to our best knowledge, there was no research related to such a framework incorporating full reference metrics. The main research challenge was that a flexible, robust and accurate image registration method was required in this context. For this reason, we proposed a novel image registration method, which took advantage of cubic spline to establish a distortion free coordinate system and correct the geometric distortions of captured images in the distortion free space (**Paper B**); so the type and magnitude of spatial distortions had very limited influence on the image registration results. From this point of view, compared to conventional image registration methods, the camera did not necessarily have to be calibrated in advance. This feature saved a lot of time and resources for the experiment. The image registration method could be divided into two parts. The first part, known as geometric detection (Figure 4.1), used projected light patterns to detect the geometric distortion in the captured images based on computer vision algorithms. This part of processing was computationally intensive, however it needed to proceed offline only once. In the second part, known as geometric correction (Figure 4.1), the distortions in all the subsequently captured images can be immediately corrected in real-time by applying a specified image interpolation algorithm. The proposed method was proved to work well in a dark room environment, however the registration errors were sensitive to ambient light due to the color aberration issues along the contour of the actual projection area. The image registration method was designed for projection displays. It is also possible to apply this method to other types of displays with limited modifications. With the proposed method, the full reference, reduced reference, and no reference image quality metrics can be incorporated into the current image quality assessment workflow (**Paper A**). This is an unique feature comparing with related research in the existing literature. An interesting phenomenon was noticed in the experiments. Since the captured images needed to be down- sampled in order to correct the geometric distortion, the screen door effect could be an issue in this context [18, 203]. It stood for the unwanted high frequency tiny grid line artifacts introduced by the projection system. Some of the digital still cameras with highly sensitive sensors and a high image resolution were capable of capturing these details (e.g. Hasselblad HD III). In such cases, with respect to specific implementations of image interpolation algorithms, the undistorted images would appear to have various banding like artifacts. Basically these unwanted high frequency components were magnified unexpectedly. However, perceptually, the majority of projector users were not be able to perceive these artifacts on the displayed images. In this context, we argued that the image quality assessment of projection displays was not necessarily to incorporate the most advanced cameras captured all the details on the displays, but use a relatively worse camera to make sure that the perceptually visible objects on the displayed images were captured. In fact, a similar conclusion had been made in **Paper A**, where we incorporated three types of cameras and compared them extensively with respect to the magnified noises.

In order to further extend the research, we implemented the proposed image registration method, integrated it into the image quality assessment work-flow, and used it to evaluate the most important image quality attributes for projection displays. In our research, we identified spatial uniformity (**Paper C**), contrast (**Paper D**) and sharpness (**Paper E**) as the most important image quality attributes, and we used the developed framework to evaluate the prediction performance of state-of-the-art image quality metrics regarding the three image quality attributes. Particularly, in this process of evaluating the spatial uniformity attribute, we proposed a hazy sky based method to eliminate the vignetting effect mainly caused by the camera lens, and we also proposed a method to quickly determine the common linear response region for all camera sensors with respect to the current combination of shutter speed setting and luminance level of incident light (**Paper C**). The spatial uniformity, contrast and sharpness were evaluated objectively with the state-of-the-art image quality metrics and subjectively with respect to the perceptual ratings. The relationships between the objective and subjective results were mainly explored based on Pearson and Spearman correlation coefficients, which were commonly used by many other researchers to determine the linear correspondence and rank orders of paired data samples. It was possible to apply other types of methods, such as Kendall Tau [125] and Gamma Statistic [60]; but in most cases the Pearson and Spearman correlation coefficients were simple, effective and sufficient for our research. In this process, we found that for a few specific types of test images all the image quality metrics had low correlations with the perceptual assessment results. In the failed cases of objective image quality assessments, new metrics should be proposed by either extending the existing ones or coming up with new ideas. In the former case, it is important to figure out the root causes. However, they strongly depend on the metric design details. Different image quality metrics may have totally different reasons to fail on the predictions. In this context, the proposed image quality assessment framework can be a very helpful tool to improve the design of metrics. Normally, the metrics has several potential components or parametric coefficients that can be improved. Once the modifications are made, the identical image distortions can be forwarded to the metrics to get a new set of results. These results can be compared to the existing metric results to see how significantly the metrics have been improved. In the cases that additional test images and their distortions need to be introduced in order to verify the improvements, the image quality assessment framework can incorporated to automatically acquire the image reproductions, register them with their originals, determine the metric scores, and correlate them automatically with the perceptual ratings. Thus, the whole process can proceed much more robustly and faster, meanwhile the validity and reliability of metric improvements are expected to be greater. For the newly proposed image quality metrics, it is a similar situation, but in addition to benchmark the new metric against the state-of-the-art.

In the evaluations of image quality attributes, the primary acquisition instrument of luminance and chrominance component was a digital still camera. The camera itself was not originally designed to give highly accurate measurement in all circumstances. It is only possible to incorporate cameras to perform assessment of relative image quality attributes. With different camera settings, such as ISO, shutter speed, and aperture etc, the values of acquisition results vary quite a lot. Subsequently, the final image quality assessment results regarding specific image quality metrics might be influenced. In this context, it is likely that the framework users will make common mistakes (e.g. applying non-optimal aperture setting for sharpness evaluation) without professional photography training and framework usage experience. In addition, the acquired colors by one camera are only available in the camera's own color space. In the cases that standardized color spaces, such as CIE LAB and CIE XYZ, are mandatory for applying certain image quality metrics, the camera needs to be calibrated in advance to make sure that the camera's color space is approximately the same as the sRGB space, for example. In this case, the camera calibration introduces additional measurement noises into the acquisition process. Besides, it is known that the camera is composed of optical, mechanical, electronic and software subsystems. Among them, the electronic subsystem is expected to the main source of imaging noises. For example, the photons hit the photon-well in a similar fashion like the rain drops to the ground. Even the incident light is ideally uniform, the number of photons dropping into the photons-wells are not identical. The numbers of received photon by an individual photon-well during the camera's exposure period approximately follows a Poisson distribution. If the number of photons received is large enough, it can be approximated with a normal distribution. From this point of view, the photon-well is the first source of light measurement errors. After this, the camera sensors need to convert light signal into analog electronic signal. The conversion results might be greatly influenced by the thermal energies due to the gradually increased camera's internal temperature. The modern digital still cameras typically are equipped with advanced mechanical system and digital signal processors. They generate a lot of heat while the pictures are being taken. These heats increase the working temperature of camera sensors and influences their dark current, sensitivities and response curves of camera sensors. In addition, the electronic signal as well as the signal noises are both magnified or re-scaled with respect to the current ISO setting. So, the signal conversion is another potential source of imaging noises. The analog signals need to be converted to digital signals, so that the captured signals can be stored or further processed. It is known that this conversion re-samples the continuous analog signals at fixed intervals to produce discrete digital signals. The losing of signal details can be a new potential source of imaging errors. Since in the research we used raw image signals to perform image processing and image quality assessment, the camera noises may have a great impact to the assessment results. From this point of view, the modeling of camera imaging noises should be considered as an essential part of the image quality assessment framework, especially when the modern cameras are common to have a large variance dynamic range for the result signals. However, it requires considerable time and resources to establish the noise models,

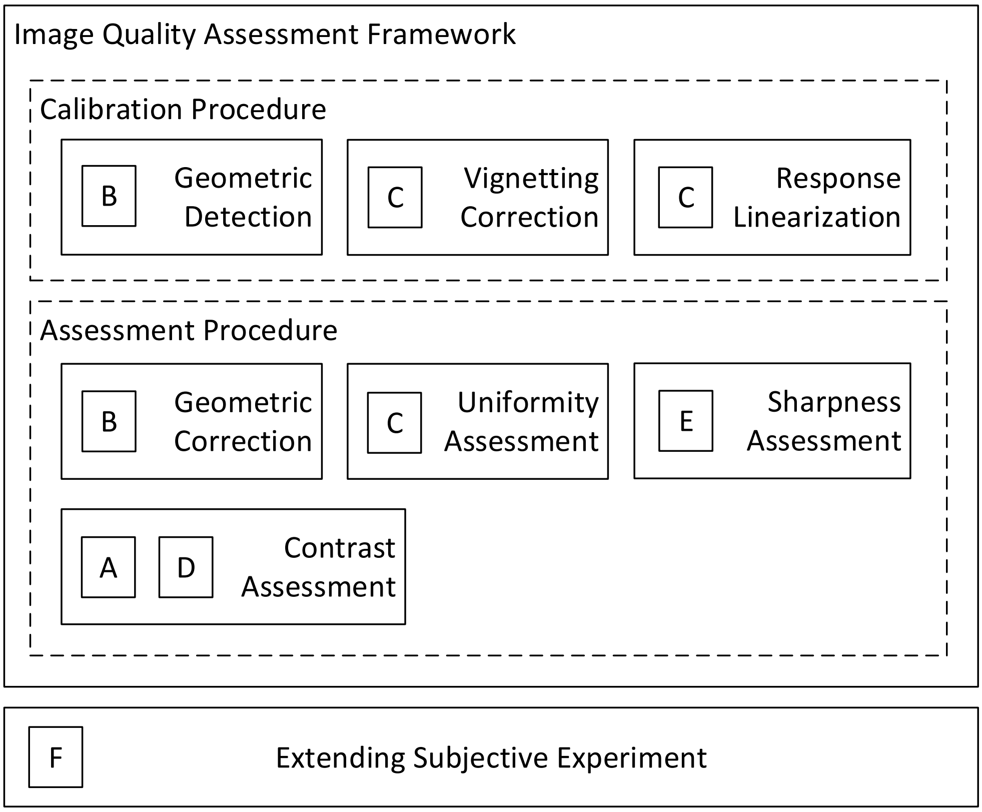


Figure 4.1: The novel contributions carried out in this research.

and such studies are beyond the current research scope of this PhD thesis. It can be very interesting for us to engage such research, but due to a limited time frame we leave them to future works.

In the subjective experiments, it was found in many scenarios that the subjective experiments needed to be conducted in multiple sessions, especially when additional distortion levels were required to be integrated into the existing subjective experiment sessions. In such cases, redoing a huge subjective experiment to include all distortion levels can be time and resource consuming. We designed an experiment to investigate the possibility of using baseline adjustment methods to extend the subjective experiments (**Paper F**). We found that the answer was positive, and pointed out the most optimal distortion levels to be included in the baseline and their number. The baseline adjustment method can be very helpful when many distortion levels are needed in the subjective experiment. For example, without prior knowledge the number of existing distortion levels of spatial uniformity (**Paper C**), contrast (**Paper D**), and sharpness (**Paper E**) might be insufficient to determine the just noticeable difference of these perceived image quality attributes. Avoiding redoing a full subjective experiment can be very beneficial to scientific research. It was our first attempt to address such a research challenge, so the experimental design did not cover all possible scenarios of extending subjective experiments. For example, we used the existing subjective data to engage the study in the paper, and these data were collected from the same group of observers. In real practice, the observers might not be identical between two experiment sessions. However, in such cases, the proposed method can also be applied but the current research needs to be further extended to take the variance of observers into consideration. In addition, the amount of data we used in the experiment is limited. In order to shrink the confidence intervals further more subjective data are mandatory. From these points of view, in order to extend the existing research, we will need to conduct a relatively larger experiment by taking all potential influence factors into consideration in advance. After this, it is possible to produce more valuable research outcomes.

Chapter 5

Conclusion and Perspectives

> Do what you can where you are with what you have.
>
> THEODORE ROOSEVELT

5.1 Conclusion

In this research, we proposed a camera based display image quality assessment framework, which in general was composed of a calibration procedure and an evaluation procedure. Although the calibration procedure can be computationally intensive with respect to the numerical calculations related to computer vision processing, vignetting mask generation, and searching for common linear response region of camera sensors, it was required to performed only once in advance of the evaluation procedure. As long as the relative positions, angles, and orientation of projector, screen and camera, as well as the camera settings remain constant, the evaluation procedure can be performed very efficiently in a fully automated fashion. One unique feature of our proposed framework was the capability of incorporating existing full reference image quality metrics without modifying them. In this research, we implemented the framework for projection displays, and used the framework to evaluate the prediction performance of state-of-the-art image quality metrics regarding the most important image quality attributes for projection displays. The evaluated image quality attributes were uniformity, sharpness, and contrast, however the proposed framework was not bound by the possibilities. All the metric evaluations were supported by the correlation of objective and subjective experimental results. In addition, we also investigated the strategies to extend subjective experiments with baseline adjustment method, which is expected to save quite a lot of time and resources for subjective experiments. In a broader point of view, the originally defined research scope have been fully covered by the research presented in this thesis, all research goals have been successfully achieved, and the corresponding research questions have been answered. The proposed image quality assessment frameworks were originally designed for projection displays, but could be easily adapted to other types of displays with limited modifications.

In conclusion, with the results we have obtained, we believe that the camera based acquisition approach can be a good complement to the conventional colorimeter and spectroradiometer based methods for image quality assessment. The trade offs to make in this context are mainly the accuracy and stability of measurements. However, with careful calibration and optimization these shortcomings can be compensated. In addition, the investment cost of time can be largely reduced, while the flexibility of image quality assessment is greatly increased. Therefore, we believe that our proposed methodology and procedures related to the proposed framework can not merely be helpful for conducting scientific research, but also potentially be helpful in the process of image quality enhancement for general displays in industrial applications. So, with the unique features and advancements introduced, we believe that our research outcome holds a positive position in the global competition as opposed to alternative solutions.

5.2 Perspectives

In this research, we have identified many research challenges in the process of evaluating the image quality attributes with the proposed image quality assessment framework. The methods corresponding to the challenges proved to work well. However, our thoughts about the potential future research areas were not limited by the existing research scope, many present methodology and procedures can be further improved.

- **A wider coverage of displays**: The proposed image quality assessment framework was originally designed for projection displays. However, it can be easily extended to other types of displays, such as LCD or LED desktop monitors, AMOLED smart phone screen, OLED television, etc. However, in such cases, more practical image quality factors should be taken into consideration. For example, these displays are typically used in a daylight environment or dimmer light environment, and the viewing conditions can be significantly different from the ones for projection displays. The viewing distance, viewing angle, light artifact (e.g. sun light reflected by the front glass of display), many other practical issues are influencing the image quality assessment results. It is worth to apply our proposed framework in such environment and observe what the potential research challenges can be found. These differences can be addressed by modifying our proposed methods, or introducing new procedure components into the existing work-flow. Since the framework is proceeded in a fully automated manner, the improvement process and corresponding performance evaluations can be greatly accelerated.

- **Camera modeling**: The central topic of this research was the assessment of captured displayed images, however the final outcomes have strong connections to the quality of image produced by the camera. Thus, the image quality influence coming from the camera itself should be minimized. In previous discussion (Section 4.7), we have addressed the image quality issues related to imaging noises. However, more factors need to be taken into consideration. For example, the international standards, such as ISO 9241 Part 304 [75], Part 305 [76], Part 307 [77], ISO 12232, ISO 12233 [175], ISO 14524 [78], ISO 15739 [79], ISO 19567 Part 1 [80], ISO 20462 [174], CIPA DC-003 [164], and CIPA DC-004 [165] defined the methods to determine the best exposure index, ISO speed, and aperture settings, as well as the ways to evaluate the camera's resolution, spatial frequency response, opto-electronic conversion functions, texture reproduction capability, noise level, camera sensor sensitivity, etc. These standards introduced the possibilities to improve the camera imaging from multiple perspectives. In this context, it could be advantageous to establish a camera model, decompose the camera imaging pipeline into several different subsystems and signal processing components, and improve the accuracy and stability of camera based image quality assessment. The detailed process can be complicated but it has a great potential to promote the proposed image quality assessment framework by incorporating the camera models as an essential part.

- **Developing a software toolbox**: In this research, the proposed image quality assessment framework is expected to be practice oriented. For this purpose, we expect to implement the framework as an open source software toolbox, and make it available to the public. Notably, image quality assessment can be processed in a fully automated fashion, and the proposed framework can be put into a more complicated real practice oriented environment for improvement and advanced design. In addition, the numeric results can be easily reproduced for validation and verification purposes. In this context, other image quality researchers are able to benefit from our research by incorporating the framework for their own benefit. Besides, it is also possible to engage in collaborative opportunities with industrial partners, such as display manufacturers, camera manufacturers, technical oriented medias, etc.

Bibliography

[1] Logitech QuickCam Pro 9000, 2007. Available from: http://www.cnet.com/products/logitech-quickcam-pro-9000/specs/. 149

[2] Nikon Digital SLR Camera D200 Specifications, 2014. Available from: http://imaging.nikon.com/lineup/dslr/d200/spec.htm. 151

[3] Nikon Digital SLR Camera D610 Specifications, 2014. Available from: http://imaging.nikon.com/lineup/dslr/d610/spec.htm. 151

[4] Optical Transfer Function, 2015. Available from: https://en.wikipedia.org/wiki/Optical_transfer_function. 14

[5] ADELSON, E. H. Lightness Perception and Lightness Illusions. In *The New Cognitive Neurosciences*, M. S. Gazzaniga, Ed., second ed., vol. 3. MIT Press, Cambridge, MA, USA, 2000, ch. 24, pp. 339–351. 10

[6] ALDRICH, J. E., AND RUTLEDGE, J. D. Assessment of PACS Display Systems. *Journal of Digital Imaging 18*, 4 (dec 2005), 287–295. 17

[7] ARORA, H., AND NAMBOODIRI, A. Projected Pixel Localization and Artifact Removal in Captured Images. In *IEEE Region 10 Conference* (Hyderabad, India, nov 2008), IEEE, pp. 1–5. 33, 35

[8] ASANO, T., TAKAGI, Y., KONDO, T., YAO, J., AND LIU, W. Image Quality Evaluation based on Contrast Sensitivity Function. In *IEEE International Conference on Mechatronics and Automation* (Beijing, China, aug 2011), IEEE, pp. 658–663. 15, 41

[9] BAE, S. H., PAPPAS, T. N., AND JUANG, B.-H. Subjective Evaluation of Spatial Resolution and Quantization Noise Tradeoffs. *IEEE Transactions on Image Processing 18*, 3 (mar 2009), 495–508. 15

[10] BAKKE, A. M., THOMAS, J.-B. B., AND GERHARDT, J. J. Common Assumptions in Color Characterization of Projectors. In *Gjøvik Color Imaging Symposium* (Gjøvik, 2009), vol. 4, pp. 45–53. 30

[11] BARTEN, P. G. J. *Contrast Sensitivity of The Human Eye and Its Effects on Image Quality*, first edit ed. Society for Imaging Science and Technology, 1999. 15

[12] BARTLESON, C. J. Combined Influence of Sharpness and Graininess on The Quality of Colour Prints. *Journal of Photographic Science 33*, 8 (1982). 12

[13] BEKKAT, N. Coded Image Quality Assessment based on A New Contrast Masking Model. *Journal of Electronic Imaging 13*, 2 (jan 2004), 341. 15

[14] Berger, K., Lipski, C., Linz, C., Sellent, A., and Magnor, M. A Ghosting Artifact Detector for Interpolated Image Quality Assessment. In *International Symposium on Consumer Electronics* (Braunschweig, Germany, jun 2010), ACM Press, pp. 128–128. 15

[15] Berns, R. S. *Billmeyer and Saltzman's Principles of Color Technology*, 3 ed. Wiley, New York, NY, USA, 2000. 10

[16] Bradski, G., and Kaehler, A. *Learning OpenCV*, 1 ed. OReilly Media, Inc., Sebastopol, CA, USA, 2008. 34

[17] Brainard, D. H., Pelli, D. G., and Robson, T. Display Characterization. In *Encyclopedia of Imaging Science and Technology*, D. H. Brainard, D. G. Pelli, and T. Robson, Eds. Wiley, New York, NY, USA, 2003, ch. Display Ch, pp. 172–188. 3, 10, 30

[18] Brennesholtz, M. S., and Stupp, E. H. *Projection Displays*, 2 ed. John Wiley & Sons, Ltd, 2008. 57

[19] Briggs, D. J. C. The Dimensions of Colour, 2007. Available from: http://www.huevaluechroma.com/. 11

[20] Brown, M. M., Majumder, A., and Yang, R. Camera-based Calibration Techniques for Seamless Multiprojector Displays. *IEEE Transactions on Visualization and Computer Graphics 11*, 2 (mar 2005), 193–206. 38

[21] Brown, T. C., and Daniel, T. C. Scaling of Ratings: Concepts and Methods. Tech. rep., Rocky Mountain Forest and Range Experiment Station, 1990. 21, 22, 24, 25, 47

[22] Brown, T. C., and Daniel, T. C. Landscape Aesthetics of Riparian Environments: Relationship of Flow Quantity to Scenic Quality Along a Wild and Scenic River. *Water resources research 27*, 8 (aug 1991), 1787–1795. 47

[23] Brown, T. C., Daniel, T. C., Schroeder, H. W., and Brink, G. E. Analysis of Ratings: A Guide to RMRATE. Tech. rep., Rocky Mountain Forest and Range Experiment Station, 1990. 24, 47

[24] Calabria, A. J., and Fairchild, M. D. Perceived Image Contrast and Observer Preference I. The Effects of Lightness, Chroma, and Sharpness Manipulations on Contrast Perception. *Journal of Imaging Science and Technology 47*, 6 (2003), 494–508. 15

[25] Calabria, A. J., and Fairchild, M. D. Perceived Image Contrast and Observer Preference II. Empirical Modeling of Perceived Image Contrast and Observer Preference Data. *Journal of Imaging Science and Technology 47*, 6 (2003), 494–508. 15

[26] Cbookes, W. Improvement in Apparatus for Indicating The Intensity of Radiation, sep 1876. Available from: http://www.google.no/patents/US182172. 17

[27] Chandler, D. M., and Hemami, S. S. VSNR: A Wavelet-Based Visual Signal-to-Noise Ratio for Natural Images. *IEEE Transactions on Image Processing 16*, 9 (sep 2007), 2284–2298. 45

[28] Choudhury, A., and Medioni, G. Perceptually Motivated Automatic Sharpness Enhancement using Hierarchy of Non-local Means. In *IEEE International Conference on Computer Vision* (2011), pp. 730–737. 45

[29] Chu, X.-q., Yang, C., and Li, Q. Contrast-sensitivity-function-based Clutter Metric. *Optical Engineering 51*, 6 (apr 2012), 067003. 13

[30] Chung, R., and Rees, M. A Survey of Digital and Offset Print Quality Issues. Tech. rep., Rochester Institute of Technology, Rochester, NY, USA, jul 2006. 15

[31] Ciocca, G., Corchs, S., Gasparini, F., and Schettini, R. Modeling Image Quality, 2015. 8

[32] Clark, J. H. The Ishihara Test for Color Blindness. *American Journal of Physiological Optics 5* (1924), 269–276. 155

[33] Daly, S. The Visible Differences Predictor: An Algorithm for The Assessment of Image Fidelity. In *Human Vision, Visual Processing, and Digital Display III* (San Jose, CA, USA, aug 1992), B. E. Rogowitz, Ed., vol. 1666, International Society for Optical Engineering, pp. 179–206. 12

[34] Dance, C., Willamowski, J., Csurka, G., and Bray, C. Categorizing Nine Visual Classes with Bags of Keypoints. In *European Conference on Computer Vision* (Prague, Czech Republic, may 2004), pp. 1–2. 13

[35] Daniel, T. C., and Boster, R. S. Measuring Landscape Ethetics: The Scenic Beauty Estimation Method. Tech. rep., Rocky Mountain Forest and Range Experiment Station, 1976. 47

[36] Datta, R., Joshi, D., Li, J., and Wang, J. Z. Studying Aesthetics in Photographic Images Using a Computational Approach. In *The 9th European Conference on Computer Vision* (Graz, Austria, may 2006), Springer Berlin Heidelberg, pp. 288–301. 13

[37] Datta, R., and Wang, J. Z. ACQUINE: Aesthetic Quality Inference Engine - Realtime Automatic Rating of Photo Aesthetics. In *International Conference on Multimedia Information Retrieval* (Philadelphia, PA, USA, mar 2010), ACM Press, pp. 421–424. 14

[38] de Ridder, H. Naturalness and Image Quality: Saturation and Lightness Variation in Color Images of Natural Scenes. *Journal of Imaging Science and Technology 40*, 6 (jan 1996), 487–493. 15

[39] de Ridder, H., and Endrikhovski, S. 33.1: Invited Paper: Image Quality is FUN: Reflections on Fidelity, Usefulness and Naturalness. *SID Symposium Digest of Technical Papers 33*, 1 (jul 2002), 986–989. 8

[40] DeValois, R. L., and DeValois, K. K. *Spatial Vision*. Oxford University Press, 1988. 11

[41] Eerola, T. *Computational Visual Quality of Digitally Printed Images*. Phd thesis, Lappeenranta University of Technology, 2010. 26

[42] Eerola, T., Lensu, L., Kálviáinen, H., Kamarainen, J.-K. K., Leisti, T., Nyman, G., Halonen, R., Oittinen, P., Kalviainen, H., Kamarainen, J.-K. K., Leisti, T., Nyman, G., Halonen, R., and Oittinen, P. Full Reference Printed Image Quality: Measurement Framework and Statistical Evaluation. *Journal of Imaging Science and Technology 54*, 1 (2010), 010201. 26

[43] Engeldrum, P. G. Image Quality Modeling: Where Are We? In *Image Processing, Image Quality, Image Capture Systems Conference* (Savannah, Georgia, apr 1999), Society for Imaging Science and Technology, pp. 251–255. 7, 9

[44] Engeldrum, P. G. *Psychometric Scaling: A Toolkit for Imaging Systems Development*. Imcotek Pr, 2000. 3, 21, 39, 42, 46

[45] Fairchild, M., and Wyble, D. R. Colorimetric Characterization of The Apple Studio Display (Flat Panel LCD). Tech. rep., Rochester Institute of Technology, New York, NY, USA, jul 1998. 3, 10, 30

[46] Fairchild, M. D. Image Quality Measurement and Modeling for Digital Photography, may 2002. 8

[47] Fairchild, M. D. *Color Appearance Models*, 2nd ed. John Wiley & Sons, Ltd, 2005. 10, 12, 15

[48] Fairchild, M. D., and Johnson, G. M. Meet iCAM: A next-generation Color Appearance Model. In *Color Imaging Conference* (nov 2002). 15

[49] Fei, X., Xiao, L., Sun, Y., and Wei, Z. Perceptual Image Quality Assessment based on Structural Similarity and Visual Masking. *Signal Processing: Image Communication 27*, 7 (aug 2012), 772–783. 12, 15

[50] Ferzli, R., and Karam, L. J. A No-reference Objective Image Sharpness Metric based on The Notion of Just Noticeable Blur (JNB). *IEEE Transactions on Image Processing 18*, 4 (apr 2009), 717–28. 45

[51] Fitzgibbon, A., and Fisher, R. A Buyer's Guide to Conic Fitting. In *British Machine Vision Conference* (Birmingham, UK, 1995), British Machine Vision Association, pp. 513–522. 34

[52] Franzen, R. PhotoCD PCD0992, 1999. Available from: `http://r0k.us/graphics/kodak/`. 31, 35, 39, 157

[53] Fry, E., Triantaphillidou, S., Jarvis, J., and Gupta, G. Image Quality Optimization, via Application of Contextual Contrast Sensitivity and Discrimination Functions. In *Image Quality and System Performance XII 93960K, Proceedings of Electronic Imaging* (San Francisco, CA, USA, jan 2015), M.-C. Larabi and S. Triantaphillidou, Eds., International Society for Optics and Photonics, pp. 93960K–93960K–12. 15

[54] Fu, W., Gu, X., and Wang, Y. Image Quality Assessment Using Edge and Contrast Similarity. In *International Joint Conference on Neural Networks* (Hong Kong, China, jun 2008), IEEE, pp. 852–855. 15

[55] Gao, S., Wang, Y., Jin, W., and Zhang, X. Perceptual Sharpness Metric based on Human Visual System. *Electronics Letters 50*, 23 (2014), 1695–1697. 45

[56] Gibson, J. E., and Fairchild, M. D. Colorimetric Characterization of Three Computer Displays (LCD and CRT). Tech. rep., Rochester Institute of Technology, New York, NY, USA, jan 2000. 3

[57] Gille, J., Arend, L., and Larimer, J. O. Display Characterization by Eye: Contrast Ratio and Discrimination Throughout the Grayscale. In *Human Vision and Electronic Imaging IX, 218* (San Jose, CA, USA, jul 2004), B. E. Rogowitz and T. N. Pappas, Eds., vol. 5292, International Society for Optics and Photonics, pp. 218–233. 3, 10, 30, 41

[58] Gong, M., and Pedersen, M. Spatial pooling for measuring color printing quality attributes. *Journal of Visual Communication and Image Representation 23*, 5 (jul 2012), 685–696. 10, 15

[59] Gong, R., Xu, H., Wang, Q., Wang, Z., and Li, H. Investigation of perceptual attributes for mobile display image quality. *Optical Engineering 52*, 8 (aug 2013), 083104. 16

[60] Goodman, L. A., and Kruskal, W. H. Measures of Association for Cross Classifications. *Journal of the American Statistical Association 49*, 268 (dec 1954), 732. 57

[61] Hardeberg, J. Y., Farup, I., and Stjernvang, G. Color Quality Analysis of A System for Digital Distribution and Projection of Cinema Commercials. *SMPTE Motion imaging 114*, 4 (2005), 146–151. 30

[62] Hartley, R. I. Self-Calibration from Multiple Views with a Rotating Camera. In *European Conference on Computer Vision* (Stockholm, Sweden, may 1994), J.-O. Eklundh, Ed., Springer Berlin Heidelberg, pp. 471–478. 3

[63] Hassen, R., Wang, Z., and Salama, M. M. A. Image Sharpness Assessment Based on Local Phase Cohernce. *IEEE Transactions on Image Processing 22*, 7 (mar 2013), 2798–2810. 12, 45

[64] Heckbert, P. S., and Garland, M. Survey of Polygonal Surface Simplification Algorithms. Tech. Rep. May, School of Computer Science, CarnegieMellon University, 1997. 34

[65] Hetherington, J., Daniel, T. C., and Brown, T. C. Is Motion More Important Than It Sounds?: The Medium of Presentation in Environment Perception Research. *Journal of environmental psychology 13*, 4 (dec 1993), 283–291. 47

[66] Hull, R. B., Buhyoff, G. J., and Daniel, T. C. Measurement of Scenic Beauty: The Law of Comparative Judgment and Scenic Beauty Estimation Procedures. *Forest Science 4*, 30 (1984), 1084–1096. 47

[67] Hung, P.-S., and Guan, S.-S. A Research on The Visual Assessment Methods for Evaluating The Quality of Motion Images Displayed at LCD. *Journal of Science and Technology 16*, 2 (2007), 153–164. 10

[68] Hunt, R. W. G. *Measuring Colour*, 4 ed. Wiley, 2011. 10

[69] International Committee for Display Metrology. Information Display Measurements Standard. Tech. rep., Society for Information Display, Campbell, CA, USA, jun 2012. 3

[70] International Electrotechnical Commission. IEC 61966 Multimedia Systems and Equipment - Colour Measurement and Management - Part 3: Equpment Using Cathode Ray Tubes, 2000. 3

[71] International Electrotechnical Commission. ISO 61966 Multimedia Systems and Equipment - Colour Measurement and Management - Part 4: Equpment Using Liquid Crystal Display Panels, 2000. 3

[72] International Electrotechnical Commission. CIE 179:2007 Methods for Characterising Tristimulus Colorimeters for Measuring The Colour of Light. Tech. rep., Vienna, Austria, jan 2007. 18

[73] International Electrotechnical Commission. IEC 61966 Multimedia Systems and Equipment - Colour Measurement and Management - Part 6: Front Projection Displays, 2007. 3

[74] International Electrotechnical Commission. IEC 61966 Multimedia Systems and Equipment - Colour Measurement and Management - Part 5: Equpment Using Plasma Display Panels, 2008. 3

[75] International Standard Organization. ISO 9241 Ergonomics of Human-system Interaction - Part 304: User Performance Test Methods for Electronic Visual Displays. Tech. rep., Geneva, Switzerland, nov 2008. 3, 62

[76] International Standard Organization. ISO 9241 Ergonomics of Human-system Interaction - Part 305: Optical Laboratory Test Methods for Electronic Visual Displays, 2008. 3, 62

[77] International Standard Organization. ISO 9241 Ergonomics of Human-system Interaction - Part 307: Analysis and Compliance Test Methods for Electronic Visual Displays. Tech. rep., Geneva, Switzerland, nov 2008. 3, 62

[78] International Standard Organization. ISO 14524:2009 Photography - Electronic still-picture cameras - Methods for measuring opto-electronic conversion functions (OECFs). Tech. rep., International Standard Organization, Geneva, Switzerland, 2009. 62

[79] International Standard Organization. ISO 15739:2013 Photography - Electronic still-picture imaging - Noise measurements. Tech. rep., International Standard Organization, Geneva, Switzerland, 2013. 62

[80] International Standard Organization. ISO/DTS 19567-1 Photography - Digital Cameras - Texture Reproduction Measurements - Part 1: Frequency Characteristics Measurements using Cyclic Pattern. Tech. rep., International Standard Organization, Geneva, Switzerland, 2015. 62

[81] Jaakkola, T., and Haussler, D. Exploiting Generative Models in Discriminative Classifiers. *Advances in neural information processing systems* (1999), 487–493. 13

[82] Jacobson, R. E. An Evaluation of Image Quality Metrics. *Journal of Photographic Science 43*, 1 (1995), 7–16. 7

[83] Janssen, R. *Computational Image Quality*. SPIE Publications, 2001. 7

[84] Jiang, W., Loui, A. C., and Cerosaletti, C. D. Automatic Aesthetic Value Assessment in Photographic Images. In *IEEE International Conference on Multimedia and Expo* (Suntec City, Singapore, jul 2010), IEEE, pp. 920–925. 13

[85] Johnson, G. M. *Measuring Images: Differences, Quality, and Appearance*. phdthesis, Rochester Institute of Technology, New York, NY, USA, mar 2003. 12, 16

[86] Johnson, G. M., and Fairchild, M. D. A Top Down Description of S-CIELAB and CIEDE2000. *Color Research & Application 28*, 6 (dec 2003), 425–435. 39

[87] Jonas, P. *Photographic Composition Simplified*. Watson-Guptill Pubns, 1976. 13

[88] Keelan, B. W. *Handbook of Image Quality: Characterization and Prediction*. Taylor & Francis, 2002. 3, 8

[89] Kerren, A., Ebert, A., and Meyer, J. Interacting with Visualizations. In *Human-centered Visualization*, A. Kerren, A. Ebert, and M. J., Eds., Volume 4417 of LNCS Tutorial. Springer, 2007, ch. 3, pp. 77–162. 20

[90] Kim, J. S., Westland, S., and Luo, M. R. Image Quality Assessment for Photographic Images. In *Congress of the International Colour Association* (Granada, Spain, may 2005), J. L. Nieves and J. Hernandez-Andres, Eds., pp. 1095–1098. 15

[91] Kim, Y. J., Luo, M. R., Choe, W., Kim, H. S., Park, S. O., Baek, Y., Rhodes, P., Lee, S., and Kim, C. Y. Factors Affecting The Psychophysical Image Quality Evaluation of Mobile Phone Displays: The Case of Transmissive Liquid-crystal Displays. *Journal of the Optical Society of America. A, Optics, image science, and vision 25*, 9 (sep 2008), 2215–2222. 16

[92] Kuang, J., Johnson, G. M., and Fairchild, M. D. iCAM06: A Refined Image Appearance Model for HDR Image Rendering. *Journal of Visual Communication and Image Representation 18*, 5 (jul 2007), 406–414. 15

[93] Kuo, C., and Ng, Y. Perceptual Color Contouring Detection and Quality Evaluation Using Scanners. *Journal of Imaging Science and Technology 49*, 1 (jan 2005), 41–46. 15

[94] Land, L., Juricevic, I., Wilkins, A., and Webster, M. Visual Discomfort and Natural Image Statistics. *Journal of Vision 9*, 8 (mar 2010), 1046–1046. 16

[95] Lau, S., Ng, K. H., and Abdullah, B. J. J. Viewing Conditions in Diagnostic Imaging : A Survey of Selected Malaysian Hospitals. *Hong Kong Journal of Radiology 4* (jan 2001), 264–267. 17

[96] Leclaire, A., and Moisan, L. No-reference Image Quality Assessment and Blind Deblurring with Sharpness Metrics Exploiting Fourier Phase Information. *Journal of Mathematical Imaging and Vision* (2014). 45

[97] Legge, G. E., and Foley, J. M. Contrast Masking in Human Vision. *Journal of the Optical Society of America 70*, 12 (dec 1980), 1458–1471. 12

[98] Lehtimaki, T. M., Saaskilahti, K., Pitkaaho, T., and Naughton, T. J. Evaluation of Perceived Quality Attributes of Digital Holograms Viewed with A Stereoscopic Display. In *Euro-American Workshop on Information Optics* (Helsinki, Finland, jul 2010), IEEE, pp. 1–3. 16

[99] Leisti, T., Radun, J., Leisti, T., Radun, J., Virtanen, T., Halonen, R., Virtanen, T., Nyman, G., and Halonen, R. Subjective Experience of Image Quality: Attributes, Definitions, and Decision Making of Subjective Image Quality. In *Electronic Imaging Conference* (San Jose, CA, USA, jan 2009), vol. 7242, International Society for Optics and Photonics, pp. 72420D–72420D–9. 10

[100] Li, C., Gallagher, A., Loui, A. C., and Chen, T. Aesthetic Quality Assessment of Consumer Photos with Faces. In *International Conference on Image Processing* (Hong Kong, China, sep 2010), IEEE, pp. 3221–3224. 14

[101] Li, Q., and Wang, Z. Reduced-Reference Image Quality Assessment Using Divisive Normalization-Based Image Representation. *IEEE Journal of Selected Topics in Signal Processing 3*, 2 (2009), 202–211. 14

[102] Liang, Z. Method for Reproducing Color Images Having One Color Gamut with a Device Having a Different Color Gamut, jun 1994. 18

[103] Lindberg, S. *Perceptual Determinants of Print Quality*. phdthesis, Stockholm University, Stockholm, Sweden, jun 2004. 15

[104] Ling, T. The Assessment of Ceiling Uniformity for Indirect Lighting Systems. Tech. rep., Rensselaer Polytechnic Institute, 1996. 39

[105] Liu, C., and Fairchild, M. d. Measuring The Relationship between Perceived Image Contrast and Surround Illumination. In *Color Imaging Conference* (Scottsdale, AZ, USA, nov 2004), vol. 2004, Society for Imaging Science and Technology, pp. 282–288. 10, 11

[106] Liu, X., Pedersen, M., and Hardeberg, J. Y. CID:IQ - A New Image Quaity Database. In *International Conference on Image and Signal Processing* (Cherbourg, Normandy, France, 2014), A. Elmoataz, O. Lezoray, F. Nouboud, and D. Mammass, Eds., vol. 8509 of *Lecture Notes in Computer Science*, Springer International Publishing, pp. 193–202. 42, 45, 157

[107] Liu, Z., and Wu, W. The Use of The Contrast Sensitivity Function in The Perceptual Quality Assessment of Fused Image. *International Journal of Image and Data Fusion 2*, 1 (feb 2011), 93–103. 15

[108] Majumder, A. Contrast Enhancement of Multi-displays Using Human Contrast Sensitivity. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Providence, 2005), vol. 2, pp. 377–382. 12, 16, 30, 41

[109] Majumder, A., and Gopi, M. Modeling Color Properties of Tiled Displays. *Computer Graphics Forum 24*, 2 (2005), 149–163. 10, 30

[110] Majumder, A., He, Z., Towles, H., and Welch, G. Achieving Color Uniformity Across Multi-projector Displays. In *Proceedings Visualization 2000. VIS 2000 (Cat. No.00CH37145)* (Salt Lake City, UT, USA, oct 2000), IEEE, pp. 117–124. 15, 16, 37

[111] Majumder, A., and Irani, S. Contrast Enhancement of Images using Human Contrast Sensitivity. In *The 3rd Symposium on Applied Perception in Graphics and Visualization* (New York, NY, USA, jun 2006), pp. 69–76. 41

[112] Majumder, A., and Stevens, R. LAM: Luminance Attenuation Map for Photometric Uniformity in Projection Based Displays. In *Proceedings of the ACM symposium on Virtual reality software and technology* (New York, 2002), VRST '02, ACM, pp. 147–154. 16, 37

[113] Majumder, A., and Stevens, R. Color Non-Uniformity in Projection Based Displays: Analysis and Solutions. *IEEE Transactions on Visualization and Computer Graphics 10* (2004), 177–188. 37

[114] Majumder, A., and Stevens, R. Perceptual Photometric Seamlessness in Projection-based Tiled Displays. *ACM Transactions on Graphics 24*, 1 (jan 2005), 118–139. 10, 36

[115] Males, M., Hedi, A., and Grgic, M. Compositional Rule of Thirds Detection. In *International Symposium ELMAR* (Zadar, Croatia, sep 2012), IEEE, pp. 41–44. 14

[116] Marchesotti, Luca Perronnin, F., Larlus, D., and Csurka, G. Assessing The Aesthetic Quality of Photographs Using Generic Image Descriptors. In *IEEE International Conference on Computer Vision* (Washington, DC, USA, nov 2011), IEEE, pp. 1784–1791. 13

[117] Martinec, E., and Lee, P. AMaZE Demosaicing Algorithm, 2010. Available from: http://www.rawtherapee.com/. 37, 45, 157

[118] Matkovic, K., Neumann, L., Neumann, A., Psik, T., and Purgatholer, W. Global Contrast Factor - a New Approach to Image Contrast. In *Eurographics Workshop on Computational Aesthetics in Graphics, Visualization and Imaging - Computational Aesthetics* (Girona, Spain, 2005), L. Neumann, M. S. Casasayas, B. Gooch, and W. Purgathofer, Eds., pp. 159–167. 41, 42

[119] Menu, G., Peigne, L., Hardeberg, J. Y., and Gouton, P. Correcting Projection Display Non-uniformity Using A Webcam. In *Color Imaging X: Processing, Hardcopy, and Applications* (San Jose, USA, jan 2005), R. Eschbach and G. G. Marcu, Eds., International Society for Optics and Photonics, pp. 364–373. 37

[120] Michelson, A. A. Studies in Optics. *Dover Publications* (1927). 11, 31, 41, 42

[121] Moorthy, A. K., and Bovik, A. C. A Two-stage Framework for Blind Image Quality Assessment. In *IEEE International Conference on Image Processing* (Hong Kong, China, 2010), IEEE, pp. 2481–2484. 26

[122] Morales, J. A. *Assessment of Iris Reflection Artifacts and Alignment in Fundus Images*. phdthesis, Saint Mary's University, San Antonio, TX, USA, 2011. 15

[123] Moroney, N., Fairchild, M. D., Hunt, R. W. G., Li, C., Luo, M. R., and Newman, T. The CIECAM02 Color Appearance Model. In *Color Imaging Conference* (Scottsdale, AZ, USA, nov 2002), Society for Imaging Science and Technology, pp. 23–27. 15

[124] Narvekar, N., and Karam, L. J. A No-reference Perceptual Image Sharpness Metric based on A Cumulative Probability of Blur Detection. In *International Workshop on Quality of Multimedia Experience* (San Diego, CA, USA, jul 2009), K. O. Egiazarian, S. S. Agaian, A. P. Gotchev, J. Recker, and G. Wang, Eds., IEEE, pp. 87–91. 45

[125] Newson, R. Parameters Behind nonparametric Statistics: Kendalls Tau, Somers D and Median Difference. *Stata Journal 2*, 1 (2002), 44–64. 57

[126] Ngo, K. V., Storvik, J. J., Dokkeberg, C. A., Farup, I., and Pedersen, M. {QuickEval}: A Web Application for Psychometric Scaling Experiments. In *Image Quality and System Performance XII, 93960R* (San Francisco, CA, USA, 2015), vol. 9396, SPIE, p. 93960O. 42

[127] Nuutinen, M., Orenius, O., Saamanen, T., and Oittinen, P. A Framework for Measuring Sharpness in Natural Images Captured by Digital Cameras based on Reference Image and Local Areas. *EURASIP Journal on Image and Video Processing 2012*, 1 (2012), 8. 26

[128] Pagani, A., and Stricker, D. Spatially Uniform Colors for Projectors and Tiled Displays. *Journal of the Society for Information Display 15*, 9 (2007), 679. 37

[129] Pavel, M., Sperling, G., Riedl, T., and Vanderbeek, A. Limits of Visual Communication: The Effect of Signal-to-noise Ratio on The Intelligibility of American Sign Language. *Optical Society of American 4*, 12 (dec 1987), 2355–2365. 41, 42

[130] Pedersen, M. *Image Quality Metrics for The Evaluation of Printing Workflows*. Phd thesis, University of Oslo, Gjøvik, aug 2011. 3, 9, 10, 15

[131] Pedersen, M., and Amirshahi, S. A. A modified framework for the application of image quality metrics for color prints. 26

[132] Pedersen, M., and Amirshahi, S. A. Framework for The Evaluation of Color Prints using Image Quality Metrics. In *5th European Conference on Colour in Graphics, Imaging, and Vision (CGIV)* (Joensuu, Finland, 2010), pp. 75–82. 26

[133] Pedersen, M., Bonnier, N., Hardeberg, J. Y., and Albregtsen, F. Attributes of a New Image Quality Model for Color Prints. In *17th Color Imaging Conference* (Albuquerque, NM, USA, nov 2009), Society for Imaging Science and Technology, pp. 204–209. 10

[134] Pedersen, M., Bonnier, N., Hardeberg, J. Y., and Albregtsen, F. Attributes of Image Quality for Color Prints. *Journal of Electronic Imaging 19*, 1 (jan 2010), 011016–01–1016–13. 4, 32

[135] Pedersen, M., Bonnier, N., Hardeberg, J. Y., and Albregtsen, F. Estimating Print Quality Attributes by Image Quality Metrics. In *Color and Imaging Conference* (San Antonio, TX, USA, nov 2010), Society for Imaging Science and Technology, pp. 68–73. 10

[136] Pedersen, M., Bonnier, N., Hardeberg, J. Y., and Albregtsen, F. Validation of Quality Attributes for Evaluation of Color Prints. In *Color and Imaging Conference* (San Antonio, TX, USA, nov 2010), Society for Imaging Science and Technology, pp. 74–79. 10

[137] Pedersen, M., Bonnier, N., Hardeberg, J. Y., and Albregtsen, F. Image Quality Metrics for The Evaluation of Print Quality. In *Proceedings of the SPIE* (San Francisco, USA, jan 2011), S. P. Farnand and F. Gaykema, Eds., vol. 7867, pp. 786702–786702–19. 32

[138] Pedersen, M., and Farup, I. Simulation of Image Detail Visibility using Contrast Sensitivity Functions and Wavelets. In *Color and Imaging Conference* (Los Angeles, CA, USA, nov 2012), vol. 1, Society for Imaging Science and Technology, pp. 70–75. 12

[139] Pedersen, M., Rizzi, A., Hardeberg, J. Y., and Simone, G. Evaluation of Contrast Measures in Relation to Observers Perceived Contrast. In *European Conference on Color in Graphics, Imaging and Vision* (Terrassa, Spain, 2008), pp. 253–256. 31, 41, 42

[140] Peli, E. Contrast in Complex Images. *Journal of the Optical Society of America 7*, 10 (oct 1990), 2032–40. 11, 30

[141] Poikonen, T. *Characterization of Light Emitting Diodes and Photometer Quality Factors*. phdthesis, Aalto University, Espoo, Finland, dec 2012. 17

[142] Ponomarenko, N., Silvestri, F., Egiazarian, K., Carli, M., Astola, J., and Lukin, V. On Between-coefficient Contrast Masking of DCT Basis Functions. In *Proceedings of the Third International Workshop on Video Processing and Quality* (Scottsdale, USA, 2007), pp. 1–4. 39

[143] ProjectorCentral.com. Mitsubishi XL9U Projector, 2015. Available from: `http://www.projectorcentral.com/Mitsubishi-XL9U.htm`. 149

[144] ProjectorCentral.com. Sony BRAVIA VPL-AW15 Projector, 2015. Available from: `http://www.projectorcentral.com/Sony-BRAVIA_VPL-AW15.htm`. 148

[145] Radiant Vision Systems LLC. Methods for Measuring Flat Panel Display Defects and Mura as Correlated to Human Visual Perception. Tech. rep., Redmond, WA, USA, 2014. 18

[146] Radiocommunication Sector of International Telecommunication Union. Methodology for The Subjective Assessment of The Quality of Television Pictures. Tech. rep., International Telecommunication Union, 2012. 158, 159

[147] Ramesh, R., Brown, M. S., Yang, R., Chen, W.-C., Brent, G. W., Towles, H., Scales, B., and Fuchs, H. Multi-projector Displays Using Camera-based Registration. In *Proceedings of IEEE Visualization* (San Francisco, CA, USA, oct 1999), IEEE, pp. 161–168. 3

[148] Rao, D. V., and Reddy, L. P. Contrast Weighted Perceptual Structural Similarity Index for Image Quality Assessment. In *IEEE India Council Conference* (Gujarat, India, dec 2009), IEEE, pp. 1–4. 15

[149] Rea, M. S., Ed. *The IESNA Lighting Handbook - Reference and Application*, 9th editio ed. Illuminating Engineering, 2000. 39

[150] Rizzi, A., Algeri, T., Medeghini, G., and Marini, D. A Proposal for Contrast Measure in Digital Images. In *European Conference on Colour in Graphics, Imaging, and Vision* (Aachen, Germany, 2004), pp. 187–192. 31, 41, 42

[151] Rizzi, A., Simone, G., and Cordone, R. A Modified Algorithm for Perceived Contrast Measure in Digital Images. In *Computer Graphics, Imaging and Visualization* (Barcelona, Spain, jan 2008), no. 1, Society for Imaging Science and Technology, pp. 249–252. 15

[152] Ross, W. D., and Pessoa, L. Lightness from Contrast: A Selective Integration Model. *Perception & psychophysics 62*, 6 (sep 2000), 1160–1181. 15

[153] Schubert, E. F. Human Eye Sensitivity and Photometric Quantities. In *Light-Emitting Diodes*, second edi ed. Cambridge University Press, Cambridge, UK, 2006, ch. 16, pp. 275–291. 17

[154] Seetzen, H., Li, H., Ye, L., Heidrich, W., Whitehead, L., and Ward, G. 25.3: Observations of Luminance, Contrast and Amplitude Resolution of Displays. *SID Symposium Digest of Technical Papers 37*, 1 (2006), 1229–1233. 41

[155] Sharma, A. Understanding Color Management. *IPA Bulletin 94*, 6 (mar 2005), 16–21. 10

[156] Sheikh, H. R., and Bovik, A. C. Image Information and Visual Quality. *IEEE Transactions on Image Processing 15*, 2 (feb 2008), 430–444. 45

[157] Shen, M.-Y., and Kuo, C.-C. Review of Postprocessing Techniques for Compression Artifact Removal. *Journal of Visual Communication and Image Representation 9*, 1 (mar 1998), 2–14. 15

[158] Silverstein, L. D. Color Display Technology: From Pixels to Perception. In *Color and Imaging Conference* (Scottsdale, AZ, USA, 2005), Society for Imaging Science and Technology, pp. 136–140. 10

[159] Simone, G., Pedersen, M., and Hardeberg, J. Y. Measuring perceptual contrast in uncontrolled environments. *2010 2nd European Workshop on Visual Information Processing (EUVIP)* (jul 2010), 102–107. 31

[160] Simone, G., Pedersen, M., and Hardeberg, J. Y. Measuring Perceptual Contrast in Digital Images. *Journal of Visual Communication and Image Representation 23*, 3 (apr 2012), 491–506. 31, 41, 42

[161] Sing, B. K. Automatic Removal of Chromatic Aberration from a Single Image. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Minneapolis, MN, USA, jun 2007), IEEE, pp. 1–8. 15

[162] Son, C.-H., and Ha, Y.-H. Color Correction of Images Projected on a Colored Screen for Mobile Beam Projector. *Journal of Imaging Science and Technology 52*, 3 (jun 2008), 1–11. 18

[163] Soundararajan, R., and Bovik, A. C. RRED Indices: Reduced Reference Entropic Differencing for Image Quality Assessment. *IEEE Transactions on Image Processing 21*, 2 (2012), 517–526. 45

[164] Standard of The Camera and Imaging Products Association. CIPA DC-003 -Translation -2003 Resolution Measurement Methods for Digital Cameras. Tech. rep., Tokyo, Japan, dec 2003. 62

[165] Standard of The Camera and Imaging Products Association. CIPA DC-004-Translation-2004 Sensitivity of Digital Cameras. Tech. rep., Tokyo, Japan, jul 2004. 62

[166] Standard Panel Working Group. SPWG Notebook Panel Specification. Tech. rep., Newark, NJ, USA, 2007. 3

[167] Strand, M., Hardeberg, J. Y., and Nussbaum, P. Color Image Quality in Projection Displays: A Case Study. In *Image Quality and System Performance II, Color Imaging Conference* (San Jose, CA, USA, jan 2005), R. Rasmussen and Y. Miyake, Eds., SPIE, pp. 185–195. 10, 16

[168] Suzuki, S., Be, K., and Abe, K. Topological Structural Analysis of Digitized Binary Images by Border Following. *Computer Vision, Graphics, and Image Processing 30*, 1 (apr 1985), 32–46. 33

[169] TCO Development AB. TCO Certified Displays. Tech. rep., Stockholm, Sweden, mar 2012. 3

[170] Teunissen, K. *Flat Panel Display Characterization: A Perceptual Approach*. Phd thesis, Eindhoven University of Technology, jan 2009. 10

[171] The International Commission on Illumination. CIE 156:2004 Guidelines for the Evaluation of Gamut Mapping Algorithms. Tech. rep., The International Commission on Illumination, Vienna, Austria, 2004. 158

[172] The International Commission on Illumination. ISO 11664-1 CIE Colorimetry - Part 1: Standard Colorimetric Observers. Tech. rep., Vienna, Austria, oct 2007. 18

[173] The International Commission on Illumination. CIE 053:1982 Methods of Characterizing The Performance of Radiometers and Photometers. Tech. rep., Vienna, Austria, 2014. 17

[174] The International Organization for Standarization. ISO 20462-1: Photography - Psychophysical Experimental Methods for Estimating Image Quality - Part 1: Overview of Psychophysical Elements. Tech. rep., The International Organization for Standarization, Geneva, Switzerland, 2005. 62

[175] The International Organization for Standarization. ISO 12233 Photography - Electronic Still Picture Imaging - Resolution and Spatial Frequency Responses. Tech. rep., The International Organization for Standarization, Geneva, Switzerland, feb 2014. 13, 42, 62, 155, 156

[176] Thomas, J.-B. *Colorimetric Characterization of Displays and Multi-display Systems*. Phd thesis, Université de Bourgogne, Dijon, France, oct 2009. 10, 16, 18, 30

[177] Thomas, J.-B., and Bakke, A. M. A Colorimetric Study of Spatial Uniformity in Projection Displays. In *Computational Color Imaging Workshop* (Etienne, Saint, France, 2009), A. Trémeau, R. Schettini, and S. Tominaga, Eds., vol. 5646, Springer Berlin Heidelberg, pp. 160–169. 10, 36

[178] Thomas, J.-B., Bakke, A. M., and Gerhardt, J. Spatial Nonuniformity of Color Features in Projection Displays: A Quantitative Analysis. *Journal of Imaging Science and Technology 54*, 3 (2010), 030403. 36, 37

[179] Trussell, H. J., and Vrhel, M. J. *Fundamentals of Digital Imaging*, first edit ed. Cambridge University Press, 2008. 20

[180] TSUSUROVÁ, D., SUROVÝ, P., AND YOSHIMOTO, A. Assessment of Aesthetic Quality of Forest Areas Using Aerial Photograph Image Data. *Forest Resources and Mathematical Modeling 12* (apr 2013), 55–73. 14

[181] TUOMAS, E., KÄMÄRÄINEN, J.-K., LENSU, L., AND KÄLVIÄINEN, H. Framework for Applying Full Reference Digital Image Quality Measures to Printed Images. In *Scandinavian Conference on Image Analysis* (Oslo, Norway, jun 2009), A.-B. Salberg, J. Y. Hardeberg, and R. Jenssen, Eds., vol. 5575, Springer Berlin Heidelberg, pp. 99–108. 4, 26, 32

[182] VAN HEESCH, F., KLOMPENHOUWER, M., AND DE HAAN, G. Masking Coding Artifacts on Large Displays. In *Digest of Technical Papers - IEEE International Conference on Consumer Electronics* (jan 2006), vol. 2006, IEEE, pp. 207–208. 15

[183] VIDEO ELECTRONICS STANDARDS ASSOCIATION. Flat Panel Display Measurements Standard 2.0, oct 2005. 39

[184] VU, C. T., AND CHANDLER, D. M. S3: A Spectral and Spatial Sharpness Measure. In *International Conference on Advances in Multimedia* (Colmar, France, jul 2009), IEEE, pp. 37–43. 45

[185] WANG, Z., BOVIK, A. C. C., SHEIKH, H. R., AND SIMONCELLI, E. P. Image Quality Assessment: from Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing 13*, 4 (apr 2004), 600–612. 35, 39, 45

[186] WANG, Z., AND SIMONCELLI, E. P. Reduced-Reference Image Quality Assessment Using A Wavelet-Domain Natural Image Statistic Model. In *Proceedings of SPIE Human Vision and Electronic Imaging X* (San Jose, USA, mar 2005), B. E. Rogowitz, T. N. Pappas, and S. J. Daly, Eds., SPIE, pp. 149–159. 45

[187] WEIBRECHT, M., SPEKOWIUS, G., QUADFLIEG, P., AND BLUME, H. R. Image Quality Assessment of Monochrome Monitors for Medical Soft Copy Display. In *Medical Imaging, Image Display* (Newport Beach, CA, USA, feb 1997), Y. Kim, Ed., vol. 3031, International Society for Optical Engineering, pp. 232–244. 17

[188] WENG, J., COHEN, P., AND HERNIOU, M. Camera Calibration with Distortion Models and Accuracy Evaluation. *IEEE Transactions on Pattern Analysis and Machine Intelligence 14*, 10 (oct 1992), 965–980. 15

[189] WHITE, M. The Early History of Whites Illusion. *Journal of the International Colour Association 5*, 7 (jul 2010), 1–7. 15

[190] WILLIAMS, D., AND BURNS, P. D. Measuring and Managing Digital Image Sharpening. In *Proceedings of IS&T Archiving Conference* (Bern, Swizerland, 2008), SPIE, pp. 89–93. 45

[191] WINKLER, S. Visual Fiedelity and Perceived Quality: Towards Comprehensive Metrics. In *Storage and Retrieval for Image and Video Databases* (San Jose, CA, USA, jun 2001), B. E. Rogowitz and T. N. Pappas, Eds., SPIE, pp. 114–125. 12

[192] WOODRUFF, A. E. The Radiometer and How It Does Not Work. *The Physics Teacher 6*, 7 (1968), 358. 17

[193] WU, C. Y., LO, M. T., TSAO, J., CHU, A., CHOU, Y. H., AND TIU, C. M. Factor Analysis in Both Spatial and Temporal Domains of Color Blooming Artifacts in Ultrasound Investigations Utilizing Contrast Agents. *Computerized Medical Imaging and Graphics 28*, 3 (apr 2004), 129–140. 15

[194] Xu, Y., Wang, Y., and Ming, Z. Color Reproduction Quality Metric on Printing Images based on The S-CIELAB Model. In *International Conference on Computer Science and Software Engineering* (Wuhan, Hubei, China, 2008), vol. 6, IEEE, pp. 294–297. 25

[195] Yamazaki, A., Liu, P., Cheng, W.-C., and Badano, A. Image Quality Characteristics of Handheld Display Devices for Medical Imaging. *PLOS ONE 8*, 11 (nov 2013), e79243. 17

[196] Yang, H. A Quality Assessment of Watermarked Color Image Based on Vision Masking. *International Journal of Information Technology and Computer Science* 2, 2 (dec 2010), 47–54. 15

[197] Yang, S., and Lee, B. Hue-preserving Gamut Mapping with High Saturation. *Electronics Letters 49*, 19 (sep 2013), 1221–1222. 10

[198] Yang, S. Y. S., Hu, Y.-H. H. Y.-H., Tull, D. L., and Nguyen, T. Q. Maximum Likelihood Parameter Estimation for Image Ringing Artifact Removal. In *International Conference on Image Processing* (Vancouver, BC, USA, sep 2000), vol. 1, IEEE, pp. 888–891. 15

[199] Yendrikhovskij, S. Image Quality and Colour Categorisation. *Colour Image Science: Exploiting Digital Media* (2002), 393–420. 8

[200] You, J., Xing, L., Perkis, A., and Wang, X. Perceptual Quality Assessment for Stereoscopic Images based on 2D Image Quality Metrics and Disparity Analysis. In *International Workshop on Video Processing and Quality Metrics for Consumer Electronics* (Scottsdale, AZ, USA, 2010), pp. 1–6. 16

[201] Zhang, F., Li, S., Ma, L., and Ngan, K. N. Limitation and Challenges of Image Quality Measurement. In *Visual Communications and Image Processing 2010* (jul 2010), P. Frossard, H. Li, F. Wu, B. Girod, S. Li, and G. Wei, Eds., International Society for Optics and Photonics, p. 8. 15

[202] Zhang, L., Zhang, L., Mou, X., and Zhang, D. A Comprehensive Evaluation of Full Reference Image Quality Assessment Algorithms. In *International Conference on Image Processing* (Orlando, FL, USA, 2012), IEEE, pp. 1477–1480. 15

[203] Zhang, L. L., and Nayar, S. Projection Defocus Analysis for Scene Capture and Image Display. *ACM Transactions on Graphics 25*, 3 (2006), 907–915. 33, 35, 57

[204] Zhang, L. L., Zhang, L. L., Mou, X., and Zhang, D. FSIM: A Feature Similarity Index for Image Quality Assessment. *IEEE Transactions on Image Processing 20*, 8 (aug 2011), 2378–2386. 45

[205] Zhang, X., Silverstein, D. A., Farrell, J. E., and Wandell, B. A. Color Image Quality Metric S-CIELAB and Its Application on Halftone Texture Visibility. In *IEEE COMPCON Digest of Papers* (San Jose, CA, USA, 1997), IEEE, pp. 44–48. 25

[206] Zhang, X., and Wandell, B. a. A Spatial Extension of CIELAB for Digital Color Image Reproduction. *Journal of the Society for Information Display 5*, 1 (1997), 61. 25

[207] Zhao, P., Pedersen, M., Hardeberg, J. Y., and Thomas, J.-B. Camera-based Measurement of Relative Image Contrast in Projection Displays. In *4th European Workshop on Visual Information Processing* (Paris, France, 2013), IEEE, pp. 112–117. 36

[208] Zhao, P., Pedersen, M., Hardeberg, J. Y., and Thomas, J.-B. Image Registration for Quality Assessment of Projection Displays. In *The 21st International Conference on Image Processing* (Paris, France, oct 2014), IEEE, pp. 3488–3492. 42, 45

[209] ZHAO, P., PEDERSEN, M., THOMAS, J.-B., AND HARDEBERG, J. Y. Perceptual Spatial Uniformity Assessment of Projection Displays with a Calibrated Camera. In *The 22nd Color and Imaging Conference* (Boston, MA, USA, nov 2014), Society for Imaging Science and Technology, pp. 159–164. 42, 45

[210] ZHOU, J., WANG, L., AKBARZADEH, A., AND YANG, R. Multi-projector Display with Continuous Self-calibration. In *Proceedings of the 5th ACM/IEEE International Workshop on Projector Camera Systems* (New York, NY, USA, oct 2008), ACM Press, p. 1. 15

Part II

Included Papers

Chapter 6

Paper A

Camera-based Measurement of Relative Image Contrast in Projection Displays

Ping Zhao, Marius Pedersen, Jon Yngve Hardeberg, and Jean-Baptiste Thomas

Published in

Proceedings of

4th European Workshop on Visual Information Processing

(EUVIP 2013),

Page 112-117,

Paris, France,

June, 2013.

CAMERA-BASED MEASUREMENT OF RELATIVE IMAGE CONTRAST IN PROJECTION DISPLAYS

Ping Zhao, Marius Pedersen, Jon Yngve Hardeberg

Gjøvik University College
Gjøvik, Norway

Jean-Baptiste Thomas

Université de Bourgogne
Dijon, France

ABSTRACT

This research investigated the measured contrast of projection displays based on pictures taken by un-calibrated digital cameras under typical viewing conditions. A high-end radiometer was employed as a reference to the physical response of projection luminance. Checkerboard, gray scale and color complex test images with a range of the projector's brightness and contrast settings were projected. Two local and two global contrast metrics were evaluated on the acquired pictures. We used contrast surface plots and Pearson correlation to investigate the measured contrast versus the projector's brightness and contrast settings. The results suggested, as expected, the projector contrast has a more significant impact on measured contrast than projector brightness, but the measured contrast based on either camera or radiometer has a nonlinear relationship with projector settings. The results also suggested that simple statistics based metrics might produce a higher Pearson correlation value with both projector contrast and projector brightness than more complex contrast metrics. Our results demonstrated that the rank order of un-calibrated camera based measured contrast and radiometer based measured contrast is preserved for large steps of projector setting differences.

Index Terms— measured contrast, projection display, digital camera, radiometer, metrics, Pearson correlation

1. INTRODUCTION

In the conventional market, LCD displays have been dominating the share for a long time. This is especially true in the desktop and mobile display market. However, nowadays, customers have an increased wish to own a display with higher resolution and larger area to visualize the information in a rich user experience. Due to many manufacturing limitations, it is difficult or not cost effective to produce a large scale display with LCD flat panels, while a projection display has a strong competency of high resolution, portability and flexibility on specifications. The application [1], for example, could be tiling multiple high brightness projectors to produce a large perceptual seamless image. One core research topic for the projection methodology is establishing a systematic approach to quantify and evaluate the quality attributes of projected images in an objective and automatic manner with carefully designed and selected image quality metrics.

In general, Display Image Quality (DIQ) is characterized based on the sets of device dependent and independent attributes. The latter set includes, but is not limited to, the attribute families of brightness, contrast, colors, sharpness and artifacts (such as noise). Based on these attributes, there were many attempts to characterize devices like CRT [2, 3] and LCD [2, 3, 4] displays. The characterization of a projection display shares a lot with these methods. Previous characterizations of projection display focused on black level estimation [5], display uniformity [1, 6, 7] and colorimetry [6, 8, 9], but limited attentions were paid to measured contrast in the actual quality of the displayed image. More specifically, the measured contrast of a displayed image has been shown to be of a significant impact [10] on visual experience. However, apart from subjective observer based evaluation and classic contrast measurement on the display itself there is no convenient method to evaluate this parameter on a displayed image.

This paper presents a study on the measured contrast of projected images. This study aims to understand the basic interactions between brightness and contrast under typical viewing conditions, evaluate the performance of contrast metrics, and correlate projector settings, camera based measured contrast and radiometer based measured contrast of the projected images. In this context, three types of cameras and a radiometer were employed as the DIQ measurement tools. The results of evaluation can be used to improve the design of projection DIQ assessment methods. They can also to be extended in the development and enhancement of general projected image reproduction technologies.

This paper is organized as follows: first, in Section 2, we introduce the background of image contrast and the metrics that were evaluated in this research. Then, in Section 3, a full description of the controlled environment, equipment setup and operation procedure for the contrast measurement is given. With respect to the acquired pictures, a series of discussions upon the interaction between projector settings and measured contrast is presented in Section 4. At last, conclusions are drawn based on the data observations.

ISBN: 978-82-93269-13-7 ©2013 University of Paris 13

2. BACKGROUND

Brightness can be defined as the perceived intensity of light coming from the image itself, rather than any property of the portrayed scene [11]; while we may say that contrast is a measure of the luminance and chromatic variations in one region relative to the average variance in the surrounding region in the same scene. So far, no one could give a convincible standard definition to contrast in a color complex scene. Somehow, contrast in either simple or complex scenes could be measured by metrics at local and/or global levels.

2.1. Global Contrast

One physical definition of contrast is given by Michelson formula [12] for luminance based global contrast:

$$C^M = (L_{max} - L_{min})/(L_{max} + L_{min}),$$

where L_{max} and L_{min} are the minimum and maximum luminance values respectively over all pixels of the entire image. This metric can be implemented easily and is widely incorporated in researches as a reference in the performance evaluation of other contrast metrics. In the family of global contrast metrics, Weber-Fechner [13], Root-Mean-Square (RMS) [14] and other definitions [15, 16, 17] were proposed as well. They share the similar concept with Michelson contrast by taking luminance information of extreme bright and dark pixels into account, but they have problem to deal with measurement noises (illustrated in contrast measurement section) and in many cases the contrast prediction is not appropriate [10].

The chrominance information in image should also contribute to the prediction of perceived contrast. In this context, an alternative metric referred as LAB Variance [18] was proposed:

$$C^{LAB} = \sqrt[1/3]{std^2(L) * std^2(a) * std^2(b)}$$

This metric was defined in the perceptually uniform CIELAB space and it takes simultaneously the luminance and chromatic channels into account. However, the equal weighting on channels does not reflect a well-known fact that luminance has stronger impact on the perceived contrast than chrominance [19, 20, 21].

2.2. Local Contrast

Many researchers realized that perceived contrast is highly local in nature. One of the local algorithms is RAMMG [22], that takes the advantage of multiple pyramid levels of local contrast information. The RAMMG metric subsamples the input image and generates pyramid images in the CIELAB space with nearest neighborhood algorithm. Then the local contrast is calculated by summing up the absolute differences between one pixel and its surrounding pixels in every channel and at every pyramid level. The local contrast values from the same channel would be normalized and finally be weighted. The mean of outputs from all levels would be the prediction of contrast over the entire image:

$$C^{RAMMG} = \frac{1}{N_L} \sum_{i=1}^{N_L} \sum_{j=1}^{3} \sum_{k=1}^{N_p} W_j C_k$$

where N_L stands for the number of pyramid levels, and N_p stands for number of pixels in each pyramid images, C_k stands for the local contrast for each pixel and its surroundings, and W_j stands for the weight of each CIELAB channel. Inspired by the design of RAMMG metric, other improved versions like RSC [23] employed Difference of Gaussians (DOG) formula to calculate the local contrast:

$$DOG(x, y) = \frac{R_c(x, y) - R_s(x, y)}{R_c(x, y) + R_s(x, y)}$$

where (x, y) stands for the spatial location of center point in the respective field, R_c and R_s stands for neuron responses of center and surround component respectively. The DOG formula was employed to take the spatial sensitivity in the center of the receptive field into consideration, and the purpose was to extend the edges and gradient information in the input images.

3. EXPERIMENTAL SETUP

3.1. Projector Setup

We used the SONY APL-AW15 which was a portable three chip LCD high brightness projector. The projector was placed on a flat table in front of the projection screen and was exactly 3m away as it is depicted in Figure 1. The projection principal axis was pointed at the screen and was perpendicular to it. The projection resolution was 1280x768 in pixels. On the screen, the projection size was approximately 2x1.2 meters. The projector was connected to a controlling laptop with a VGA cable. In order to minimize the influence of projector temporally stability, the projector lamp was warmed up at least one hour before each experiment section. All settings related to projector's brightness, contrast and color enhancement were switched off to make sure the input image was projected as it is.

3.2. Digital Camera

Three types of digital cameras were employed in the experiments as the DIQ measurement tools. They were a low-end webcam Logitech QuickCam Pro 9000, a prosumer DSLR Nikon D200 with VR 18-200mm f/3.5-5.6G (VR off) lens and a high-end DSLR Hasselblad H3D II with HC 80mm lens. The resolutions for the three cameras were 3264x2448, 3872x2592 and 6490x4870 in pixels respectively. The webcam was mounted on a table which was 3m away from the

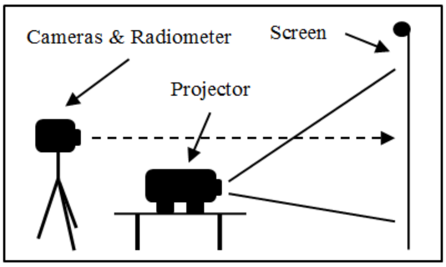


Fig. 1. Projector and camera placement

projection screen, while the other two cameras were both fixed on tripods that were 4m away. All cameras were manually focused with highest possible optical sharpness and all camera settings were fixed at certain values. The principal axes of all cameras were pointed at the center of the projected image and were quasi-perpendicular to it with a slightly varying angle due to the manual setup. The pictures were always taken remotely with software installed on the controlling laptop without physically touching the cameras. In order to confirm the physical responses of projected images on the screen, a radiometer Minolta CS1000 was mounted on a tripod beside the digital cameras as a measurement reference.

3.3. Test Images

Three test images which had a resolution 768x512 were included in the experiments as it is illustrated in Figure 2. They covered the image categories of simple checkerboard, gray complex and color complex. The use of checkerboard was recommended in the projection contrast measurement sections of international standard IDMS [24]. The color complex image was selected from Kodak Photo CD PCD0992 [25] based on which the grayscale version was generated by using Matlab function rgb2gray. The two complex images provides a smooth transition from a laboratory artificial stimuli to a natural scene in a way which people usually perceive and interpret objects in the real environment. The images were always projected at the original resolution on the screen.

3.4. Viewing Conditions

Three typical viewing conditions were considered in the experiments. The projection screen was illuminated by the fluorescent light with fixed luminance values (measured with a light meter at a fixed position) at approximately 0 (low light), 30 (dimmed light) and 300 lux (high light) respectively. A darkroom gives 0 lux luminance. In a dimmed meeting room like environment, the typical luminance is around 30 lux; while in the office like environment in daylight, the luminance could be around 300 lux. Whenever the viewing condition was switched, there was always at least 30 minutes idle period between two continuous experiment sections to maximize the illuminant temporal stability.

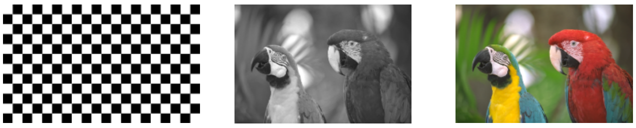


Fig. 2. Test images

3.5. Procedures

Since the resolution of original test image was smaller than the projection resolution in the grayscale and color complex cases, the actual image content of pictures taken by cameras were wrapped by the surround. In order to estimate the influence of surround on the measured contrast, in the post-processing step, all pictures taken by cameras were processed with Matlab scripts to generate a cropped version. The cropped version contains only the actual image content without surround or background. In the non-cropped version, the surrounds were simply left as how they were illuminated by the controlled light. Both the cropped and non-cropped picture versions would be input to the metrics.

Finally, for each set of pictures taken (project the same test images that under the same viewing condition), we generated the contrast surface plots with respect to the outcomes of metrics like it is demonstrated in Figure 3a. In our experiments, we evaluated four contrast metrics: Michelson [12], LAB Variance [18], RAMMG [22] and RSC [23]. Since the RAMMG and RSC shared various input parameters, we evaluated several combinations of them which involved: channel weightings: (1, 0, 0), (1/3, 1/3, 1/3), (0.5, 0.25, 0.25), pyramid scales: linear and log based scales, radius of center and surround of receptive field: (1, 2), (2, 3), (3, 4). These parameters were used and recommended by Simone et al. in their investigation of measuring perceptual contrast [26].

Because these metrics had no parameter related to viewing distance which might have influence on the metric selection, we fixed measurement devices at the same location to make sure that they share a constant distance to projection screen. In this case, the influence of viewing distance on measured contrast were equal to all metrics. In order to reduce computational complexity, the level weighting method was always set to variance and pictures were transformed into CIELAB space by the metrics themselves. We also evaluated the performance of metrics by determining the correlation between measured contrast and projector contrast, and the correlation between measured contrast and projector brightness with respect to Pearson correlation coefficient.

4. RESULTS AND DISCUSSION

4.1. Projector

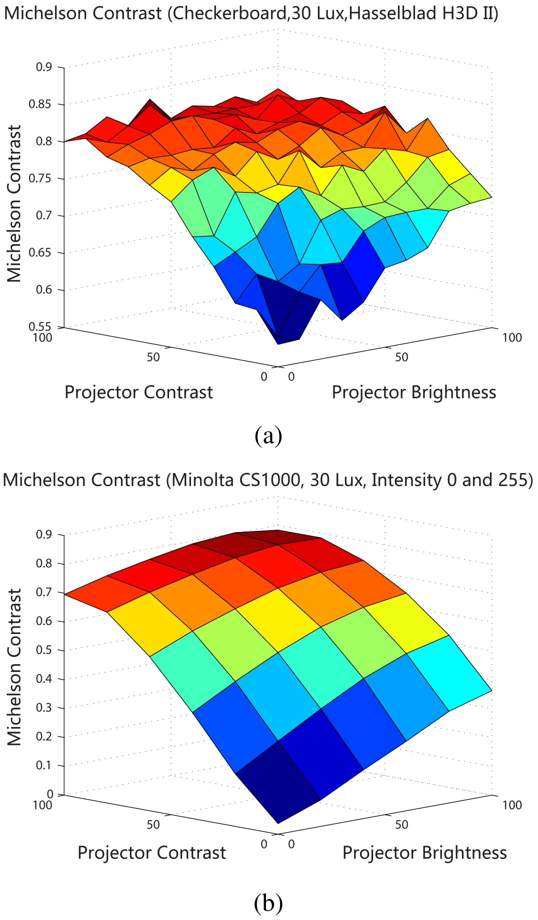
The projector SONY APL-AW15 had a light leak issue due to a fact that it is difficult to completely block the projector's backlight. The black level of the projection increased while the projector brightness increased, and it generated a halo around the image content. This halo should contribute to the perceived contrast, but it is beyond the scope of this article. The solution we adopted was cropping the image content from the pictures and accounting only the pixels in that image. In the case that the image resolution is smaller than projector resolution, both cropped and noncropped (the area other than the image content was black) picture versions were used in the post-analysis step. In the evaluation of the four metrics listed, the absolute values of measured contrast were reduced comparing with non-cropped version but the general shape of normalized contrast surfaces were almost identical. This observation suggests that it is not necessary to crop the image for either global or local contrast metric.

4.2. Digital Cameras

The webcam is not suitable for measuring contrast in the projection system, because the shape of contrast surfaces for the webcam were not consistent under varying viewing conditions. In analogy, the results for Nikon D200 are very similar to the ones for Hasselblad H3D II, except in the case of Michelson contrast under high light environment. In such case, the contrast surface for Nikon D200 is smoother than the one for Hasselblad H3D II. Based on the observations, the consumer DSLR camera Nikon D200 is preferred as a measurement tool for projection contrast. Hasselblad H3D is a little bit more sensitive to small luminance variation especially in the low and high light conditions, but it consumes much more time to take a large volume of pictures.

4.3. Contrast Measurement

Two global contrast metrics: Michelson's definition, LAB Variance, and two local contrast metrics: RAMMG and RSC were evaluated. Under the low light condition, Michelson's definition always gave contrast value 1 for three types of cameras and three types of test images despite the changes of projector contrast and brightness. This metric is very sensitive to measurement noises as it is depicted in Figure 3a and becomes very unstable under high light condition. LAB Variance, RAMMG and RSC metrics all produce logistic shape like contrast surface. The latter two metrics share a general shape of contrast surface despite their radius of receptive center and surround values, somehow the absolute values of measured contrast are different. They are more sensitive to increasing rate of projector brightness and contrast than the LAB Variance metric, since the contrast surface appears to be more bended.



(a)

(b)

Fig. 3. Michelson contrast surface for checkerboard pictures taken by Hasselblad H3D II (a) and Minolta CS1000 (b) under dim light condition

Pearson product-moment correlation coefficient was employed to determine the correspondence between measured contrast and projector settings. In Figure 5, we plotted Pearson correlation between radiometer based Michelson contrast and camera based contrast of various metrics. The general tendency of radiometer based contrast correlated well with camera based contrast of RSC metric while remain projector brightness constant. LAB Variance metric produced higher correlation value at low projector contrast but the correlation decreased a lot while projector contrast increased. In the case projector contrast remained constant, the LAB Variance metric gave a higher contrast correlation to radiometer based Michelson contrast. The measured contrast of RAMMG and RSC metrics correlated well with projector contrast and projector brightness, and RSC metric produced approximate 12% higher correlation than RAMMG metric.

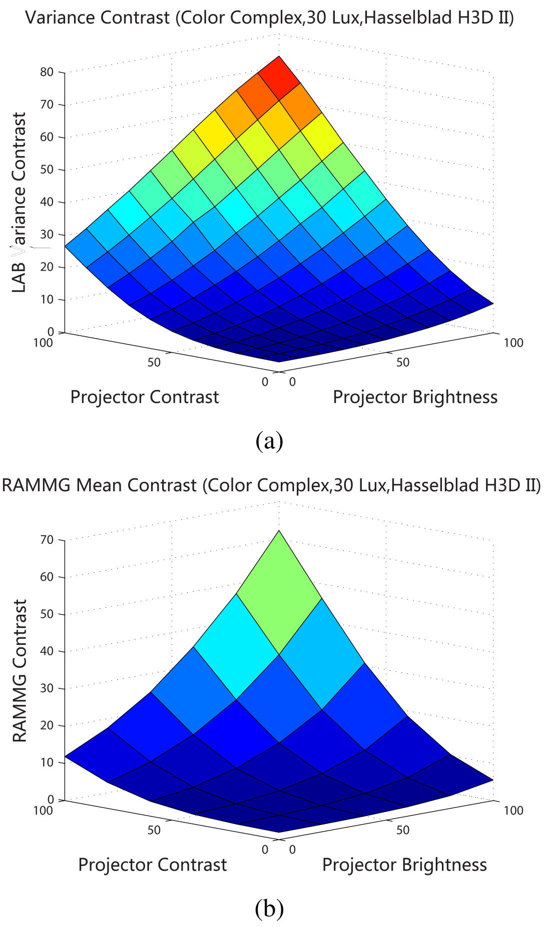


Fig. 4. LAB Variance (a) and RAMMG (b) contrast surface for color complex pictures taken by Hasselblad H3D II under 30 Lux viewing condition

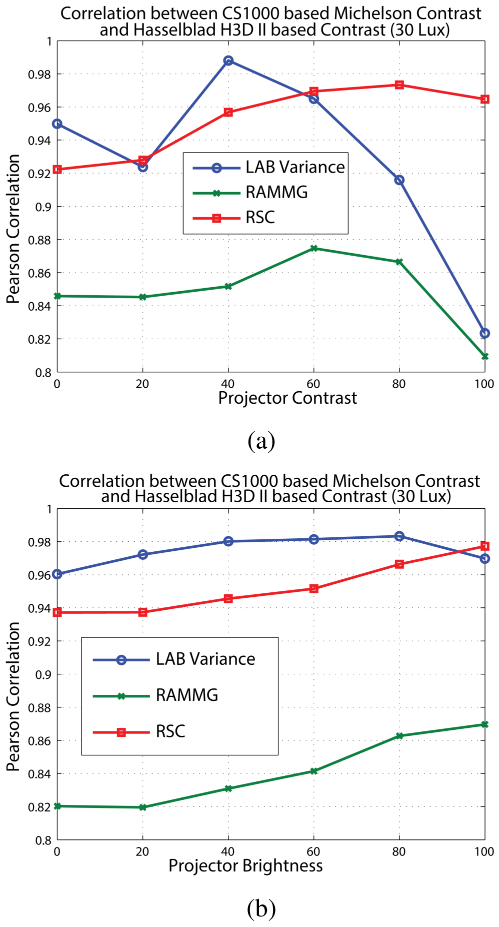


Fig. 5. Pearson correlation between measured contrast and projector settings for color complex pictures taken by Hasselblad H3D II under 30 lux viewing condition

4.4. Generals

Projector contrast has more significant impact than projector brightness on the measured contrast for all metrics as expected. The phenomenon can be observed from the height difference of left most and right most corner points of contrast surface in Figure 3 and Figure 4. It leads to an asymmetric contrast surface. The measured contrast for digital cameras has a consensus with the one for the radiometer. As it is depicted in Figure 3, the general tendency of the two Michelson contrast surfaces are similar to each other despite the contrast value range.

5. CONCLUSIONS AND FUTURE WORKS

In this research, several contrast metrics were evaluated based on pictures taken by un-calibrated digital cameras under typical viewing conditions. The results showed that the projector settings have a great impact on the measured image contrast, and the impact of projector contrast setting is even stronger. Camera based Michelson contrast was proved not to be suitable for projection contrast measurement, while the global metric LAB Variance produces higher Pearson correlation values than the complicated local metric RAMMG and RSC on both brightness and contrast correlations. Thus, we demonstrated that the rank order of un-calibrated camera based measured contrast and radiometer based measured contrast is preserved for large steps of projector setting differences. In the coming future, it will be important to incorporate psychophysical experiments to investigate the correspondence between the perceived contrast and measured contrast.

6. REFERENCES

[1] A. Majumder, “Contrast Enhancement of Multi-displays Using Human Contrast Sensitivity,” in *IEEE Computer*

Society Conference on Computer Vision and Pattern Recognition, Providence, 2005, vol. 2, pp. 377–382.

[2] J. Gille, L. Arend, and J. O. Larimer, "Display Characterization by Eye: Contrast Ratio and Discrimination Throughout the Grayscale," in *Human Vision and Electronic Imaging IX*, Bernice E. Rogowitz and Thrasyvoulos N. Pappas, Eds., San Jose, June 2004, vol. 5292, pp. 218–233, SPIE Proceedings Vol. 5292.

[3] D. H. Brainard, D. G. Pelli, and T. Robson, "Display Characterization," 2002.

[4] M. Fairchild and D. Wyble, "Colorimetric Characterization of The Apple Studio Display (flat panel LCD)," Tech. Rep., Munsell Color Science Laboratory, New York, 1998.

[5] A. M. Bakke, J. B. Thomas, and J. Gerhardt, "Common Assumptions in Color Characterization of Projectors," in *Gjøvik Color Imaging Symposium*, Gjøvik, 2009, pp. 45–53.

[6] J. B. Thomas, *Colorimetric Characterization of Displays and Multi-display Systems*, Ph.D. thesis, Universite de Bourgogne, 2009.

[7] A. Majumder and M. Gopi, "Modeling Color Properties of Tiled Displays," *Computer Graphics Forum*, vol. 24, no. 2, pp. 149–163, 2005.

[8] J. Y. Hardeberg, L. Seime, and T. Skogstad, "Colorimetric Characterization of Projection Displays Using a Digital Colorimetric Camera," in *Proceedings of the SPIE, Projection Displays IX*, 2003, vol. 5002, pp. 51–61.

[9] J. Y. Hardeberg, I. Farup, and G. Stjernvang, "Color Quality Analysis of A System for Digital Distribution and Projection of Cinema Commercials," *SMPTE Motion imaging*, vol. 114, no. 4, pp. 146–151, 2005.

[10] E. Peli, "Contrast in Complex Images," *Journal of the Optical Society of America. A, Optics and image science*, vol. 7, no. 10, pp. 2032–40, Oct. 1990.

[11] E. H. Adelson, "Lightness Perception and Lightness Illusions," in *The new cognitive neurosciences*, Gazzaniga M., Ed., vol. 3, chapter 24, pp. 339–351. MIT Press, Cambridge, MA, 2nd edition, 2000.

[12] A. A. Michelson, *Studies in Optics*, Dover Publications, 1995.

[13] P. Whittle, "The Psychophysics of Contrast Brightness," in *Lightness, Brightness, and Transparency*, pp. 35–110. Lawrence Erlbaum Associates, New Jersey, 1994.

[14] M. Pavel, G. Sperling, T. Riedl, and A. Vanderbeek, "Limits of Visual Communication: The Effect of Signal-to-noise Ratio on The Intelligibility of American Sign Language," *Optical Society of American*, vol. 4, no. 12, pp. 2355–2365, 1987.

[15] P. E. King-Smith and J. J. Kulikowski, "Pattern and Flicker Detection Analysed by Subthreshold Summation.," *The Journal of Physiology*, vol. 249, no. 3, pp. 519–548, 1975.

[16] D. A. Burkhardt, J. Gottesman, D. Kersten, and G. E. Legge, "Symmetry and Constancy in The Perception of Negative and Positive Luminance Contrast," *Journal of the Optical Society of America. A, Optics and image science*, vol. 1, no. 3, pp. 309–16, Mar. 1984.

[17] P. Whittle, "Increments and Decrements: Luminance Discrimination," *Vision Research*, vol. 26, no. 10, pp. 1677 – 1691, 1986.

[18] M. Pedersen, A. Rizzi, J. Y. Hardeberg, and G. Simone, "Evaluation of Contrast Measures in Relation to Observers Perceived Contrast," in *CGIV 2008 - Fourth European Conference on Color in Graphics, Imaging and Vision*, Terrassa, Spain, 2008, pp. 253–256.

[19] E. B. Goldstein, *Sensation and Perception*, Wadsworth Publishing, 8th edition, 2009.

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

[21] M. D. Fairchild, *Color Appearance Models*, John Wiley & Sons, Ltd, 2nd edition, 2005.

[22] A. Rizzi, T. Algeri, G. Medeghini, and D. Marini, "A Proposal for Contrast Measure in Digital Images," in *CGIV 2004 Second European Conference on Colour in Graphics, Imaging, and Vision*, Aachen, Germany, 2004, pp. 187–192.

[23] G. Simone, M. Pedersen, and J. Y. Hardeberg, "Measuring Perceptual Contrast in Digital Images," *Journal of Visual Communication and Image Representation*, vol. 23, no. 3, pp. 491–506, Apr. 2012.

[24] *Information Display Measurements Standard*, International Committee for Display Metrology, Society for Information Display, California, USA, 1.03 edition, 2012.

[25] R. Franzen, "Kodak Lossless True Color Image Suite: PhotoCD PCD0992," 2013.

[26] G. Simone, M. Pedersen, and J. Y. Hardeberg, "Measuring perceptual contrast in uncontrolled environments," *2010 2nd European Workshop on Visual Information Processing (EUVIP)*, pp. 102–107, July 2010.

Chapter 7

Paper B

Image Registration for Quality Assessment of Projection Displays

Ping Zhao, Marius Pedersen, Jon Yngve Hardeberg, and Jean-Baptiste Thomas

Published in

Proceedings of

21st International Conference on Image Processing

(ICIP 2014),

Page 3488-3492,

Paris, France,

October, 2014.

IMAGE REGISTRATION FOR QUALITY ASSESSMENT OF PROJECTION DISPLAYS

Ping Zhao, Marius Pedersen, Jon Yngve Hardeberg

Gjøvik University College
Gjøvik, Norway

Jean-Baptiste Thomas

Université de Bourgogne
Dijon, France

ABSTRACT

In the full reference metric based image quality assessment of projection displays, it is critical to achieve accurate and fully automatic image registration between the captured projection and its reference image in order to establish a subpixel level mapping. The preservation of geometrical order as well as the intensity and chromaticity relationships between two consecutive pixels must be maximized. The existing camera based image registration methods do not meet this requirement well. In this paper, we propose a marker-less and view independent method to use an un-calibrated camera to perform the task. The proposed method including three main components: feature extraction, feature expansion and geometric correction, and it can be implemented easily in a fully automatic fashion. The experimental results of both simulation and the one conducted in the field demonstrate that the proposed method is able to achieve image registration accuracy higher than 91% in a dark projection room and above 85% with ambient light lower than 30 Lux.

Index Terms— image quality, image registration, spatial distortion, projection display, full reference metric

1. INTRODUCTION

In the past decades, tremendous growth in the use of digital media indicates that our daily life has been greatly impacted by the rapid advancement of imaging technologies. Projection displays among various display technologies have unique advantages such as portability, high resolution, and flexibility. Recently, there has been an increased popularity of embedding projectors in smart phones and video cameras [1]. Furthermore, multiple projections can be tiled up to generate a large perceptually seamless picture with the help of a digital still camera [2]. It is cost effective for customers to visualize information in a very high resolution without issuing a customized manufacturing demand. Hence, projection display image quality assessment becomes an increasingly interesting and essential topic among both of scientific research and industrial communities. In full reference metric based image quality assessment [3, 4, 5], image quality is evaluated with respect to selected attributes. In a typical projection system, cameras are commonly used to acquire the scene, since they are capable to record all pixels in one shot. It is critical to achieve a highly accurate and fully automatic image registration between the captured projection and its reference image; then it is possible to apply metrics which require the distorted image and its reference image share the dimension and resolution. The preservation of geometrical order as well as the intensity and chromaticity relationships between two consecutive pixels on the screen must be maximized. However, existing camera based image registration methods do not meet this requirement well, because they either place simple assumptions on the projections and cameras in order to reduce the problem complexity, or they tend to modify the captured image quality implicitly. The captured images are expected to have various spatial distortions with respect to the relative positions and orientations of the projector, screen, and camera. The camera lens introduces additional spatial distortions. Hence, establishing a robust, accurate and reliable image registration for PDIQ assessment is a non-trivial task.

In this paper, we propose a marker-less and view independent method to use an uncalibrated camera to capture the projection scene and correct its nonlinear spatial distortions. The rest of this paper is organized as follows: first, in Section 2, we present the conventional image registration methods. Then, in Section 3, the proposed method is presented. In Section 4, experimental results are shown. At last, in Section 5, conclusions are drawn based on the data observations.

2. BACKGROUND

Cameras are conventionally calibrated off-line in order to eliminate lens distortion. The camera is typically modelled as a pinhole and the global homography [6, 7] are presented as a 4×3 extended nonsingular projective transformation matrix H. The intrinsic and extrinsic parametric coefficients [8, 9, 10] are estimated by a least-square-fitting technique with respect to a large number of pair-wise feature observations. The captured features are assumed to be distributed on a plane in the physical world. The pixel P_o in the original image corresponds to the pixel P_d on the screen and the pixel $P_c = H \cdot P_d$ in the captured image. H is defined as Equation 1 where s stands for the scaling factor for the homogeneous coordinates of pixel P_c and it correlates to the camera settings like capturing resolution, focal length and aperture. a_{ij} ($i, j \in \{0, 1, 2\}$) are collectively called intrinsic parameters of the camera, while r_{ij} and t_i defining the rotation and trans-

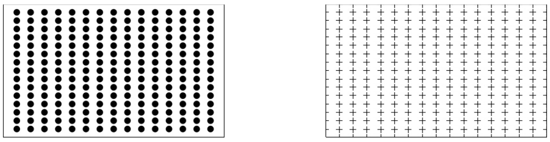


Fig. 1. Gray patterns with evenly distributed dots and crosses

lation of the view transformation respectively are collectively called extrinsic parameters.

$$H = s \cdot \begin{bmatrix} a_{00} & a_{01} & a_{02} \\ 0 & a_{11} & a_{12} \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} r_{00} & r_{01} & r_{02} & t_0 \\ r_{10} & r_{11} & r_{12} & t_1 \\ r_{20} & r_{21} & r_{22} & t_2 \end{bmatrix} . \quad (1)$$

The pincushion, barrel, and moustache distortions of the camera lens are corrected by estimating the coefficients of inward or outward displacements of feature points from their ideal locations and inversing the transform. The features are detected from a series of captured patterns whose physical dimension and appearance are known in prior. Many formulations [11, 12, 13] are introduced accordingly. The geometric distortions are then corrected by down sampling the captured image and register it with the reference image by inversing the perspective transformations like translation, scaling, rotation, skewing, and shearing. In a summary, the conventional image registrations require the camera lens distortion to be identified manually in prior, and the camera settings must be fixed for the use in the future. In real practice of projection displays, an identical projection may appear to have different types of distortions with slightly varied camera positions and/or orientations. The camera is likely to be relocated in the field in order to obtain a view dependent optimized image quality. In such cases, the camera settings will be adjusted and the camera must be recalibrated accordingly. The process of conventional camera calibration involving both offline and online procedures is non-trivial, so it makes camera based PDIQ assessment inflexible and impractical.

3. PROPOSED METHOD

The proposed method has three major components: feature extraction, feature expansion, and geometric correction.

3.1. Feature Extraction

Two gray patterns (Figure 1) are generated and projected in full screen size in order to estimate spatial distortions. The dot pattern contains round solid black dots evenly distributed in a $M_d \times N_d$ grid layout, where M_d and N_d represent the number of columns and rows respectively. The cross pattern includes crosses that share the center locations and radius with the dots in the dot pattern. Lets denote the captured dot pattern as I_d, then a contour map C can be generated as

$$C = M_c \left(G_a \left(I_d \right) - G_b \left(I_d \right) \right), \quad (2)$$

where the Gaussian filter G_a with kernel size a (empirically $a < 5$) is adopted to reduce the screen-door effect [14, 15]. The kernel size a should be kept as small as possible to preserve the details in the captured image. The Gaussian filter G_b with kernel size b (empirically $b > 41$) is adopted to spread energy from highly illuminated pixels to their neighborhoods. The median filter M_c with kernel size c (empirially $c = 3$) is adopted to remove the salt-and-pepper noises, and smooth the detected object contours. False contours might be visible in the generated map C. In this context, the false detections of object contours can be eliminated by applying a binary thresholding. The output binary image I_b can be expressed as

$$I_b = \begin{cases} 1 & C_i > (1-p) \cdot L_{min} + p \cdot L_{max} \\ 0 & otherwise \end{cases} \quad (3)$$

where C_i denotes the ith pixel in contour map C, L_{min} and L_{max} denote the minimum and maximum gray values in the image respectively, and $p \in [0, 1]$ is a constant threshold. Smaller value of p forces the detected projection boundaries to be compressed and vice versa. The value of p has an inverse effect on the detected dot contours. Hence, the pixels corresponding to positive thresholding are kept and the rests are removed. The algorithm proposed by Suzuki et al. [16] is adopted to determine the contours and their hierarchy relationships in the binary image I_b. The outermost and longest object contour corresponds to the projection area, while the innermost and shortest contours correspond to the dots. The rest of contours are therefore discarded. For each identified contour, a moving window (empirically size equals to 5) is placed along its pixels and a dynamic threshold with respect to local statistics in the corresponding area of the original captured image I_d is calculated. Then, the contour pixel at the window center is shifted toward its neighborhood either horizontally or vertically to achieve the goal of local optimization. The local threshold T is determined as

$$T = L \cdot \left(1 - k \cdot \frac{\sigma}{L/2} \right), \quad (4)$$

where σ denotes the standard deviation of gray values within the local window, L denotes the maximum gray scale level (256 for 8 bit image) and the constant k (empirically $k \in [0.1, 0.3]$) indicates the confidence of the image quality of the captured image I_d. In the case of good image quality, the value of k can be scaled down to 0. Otherwise, it should be scaled up. Eventually, we adopt the algorithm proposed by Fitzgibbon et al. [17] to fit dot contours into ellipses with respect to the least square error minimization, so the actual center, size, and orientation of each ellipse can be estimated simultaneously. The estimated ellipse centers will be slightly shifted according to the ”cornerSubPix” algorithm provided by the OpenCV library [18]. Such an algorithm incorporates the detected cross centers from the cross pattern image to locally optimize the ellipse centers since the dots and crosses share the same center location and radius.

3.2. Feature Expansion

The detected dot grid needs to be expanded to cover the entire projection area, so that the image registration can be independent from the geometry and content of projected images. In this case, we fit the coordinates of all dot centers in the same row or in the same column into a parametric natural cubic spline function as sample points, and estimate the parametric coefficients accordingly. Once the spline functions are determined, we can generate a smooth parametric cubic spline passing through each set of the feature points. In turn, each spline is extrapolated to intersect with the detected contour of the projection area to generate a pair of new feature points. Thus, in total, $2(M_d + N_d)$ new feature points are generated. The four extreme corners of projection contour can be determined by applying the split-and-merge algorithm proposed by Heckbert et al. [19] iteratively to eliminate pixels until only four corners are left. These corners are used as feature points as well. Eventually, we will have $(M_d + 2) \cdot (N_d + 2)$ feature points covering the entire projection area. The reason to employ the natural cubic spline is to take the advantage of its unique mathematical properties. A typical parametric formulation can be presented as

$$x\left(p_x\right) = \sum_{k=0}^{3} \alpha_{ik} \left(p_x - c_i\right)^k, y\left(p_y\right) = \sum_{k=0}^{3} \alpha_{jk} \left(p_y - c_j\right)^k \quad (5)$$

where $p_x \in [0, M_d + 1]$ and $p_y \in [0, N_d + 1]$ are the two parametric coefficients for the spatial coordinates of a point on the ith and jth spline section respectively; α_{ik} and β_{jk} are the local polynomial regression of kth parametric coefficient of the ith and jth spline sections respectively; $c_i \in [0, M_d + 1]$ represents the parametric coordinate of ith sample point on spline $x\left(p_x\right)$, and $c_j \in [0, N_d + 1]$ represents the parametric coordinates of jth sample point on spline $y\left(p_y\right)$. Each feature point in the expanded dot grid represents one sample point for the corresponding spline. The coefficients are estimated to make sure that around each key point the two consecutive spline sections share the same first and second derivatives; so the whole spline curve is differentiate and continuous below the third polynomial order at all possible locations. In addition, the estimated splines are adapted to local variance within each spline section. Higher order spline may not be employed to avoid adaptation to the errors inherited from the capturing process or from the calculations above.

3.3. Geometric Correction

We create a sub-pixel level mapping between the captured image and its reference, and the captured image can be undistorted by down sampling with respect to a specified interpolation method. The basic idea is to register pixels between the Cartesian coordinate system in the camera space and a distortion independent coordinate system defined by the expanded feature grid. Suppose the reference image resolution is given as $N_x \times N_y$ in pixels and the capturing resolution as $M_x \times M_y$ in pixels. Any pixel $P_o = (x, y)$ where $x \in [0, N_x - 1]$ and $y \in [0, N_y - 1]$ in the reference image corresponds to pixel $P_c = (u, v)$ where $u \in [0, M_x - 1]$ and $v \in [0, M_y - 1]$ in the captured image and the pixel p_u in the undistorted image. Their coordinates are defined in the Cartesian coordinate systems and their pixel correspondences in the distortion independent space are

$$Q_o = \left(\frac{x \cdot (M_d + 1)}{(N_x - 1)}, \frac{y \cdot (N_d + 1)}{(N_y - 1)} \right) \quad (6)$$

$$Q_c = \left(\frac{u \cdot (M_d + 1)}{(M_x - 1)}, \frac{v \cdot (N_d + 1)}{(M_y - 1)} \right) \quad (7)$$

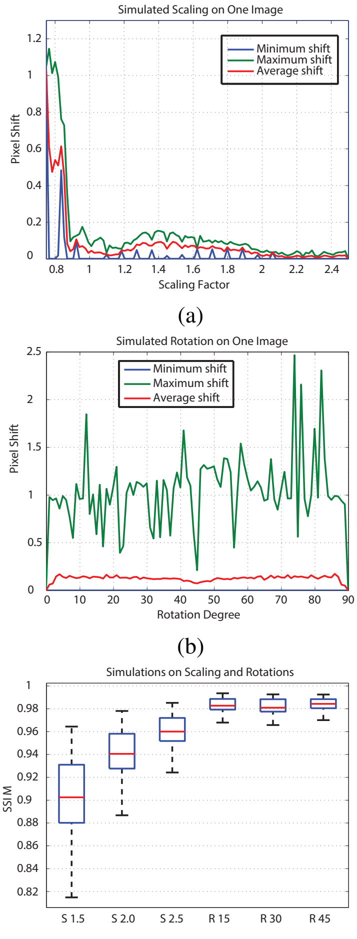
respectively. The correspondence between P_r and Q_o, as well as P_c and Q_c are established with respect to the expanded feature grid which is generated based on cubic splines. Since the undistorted image is expected to exactly registered with the reference image, then the coordinates of P_r is equal to P_u in the distortion independent space. Special attention must be paid to the screen-door effect [14, 15]. The geometric correction may introduce wave-like artifacts. A trade off has to be made between blurring the captured image to register the geometry with the reference image, or distorting the reference image to register it with the captured image. In this paper, we adopt the former approach to make sure that existing full reference image quality metrics can be incorporated without any modification.

4. EXPERIMENT

The experiment is performed in a controlled lab environment. A portable LCD projector SONY APL-AW15 (1280x768) is placed in front of a planar screen, and a DSLR Nikon D200 (3872x2592) is used for image acquisition. All 24 images from Kodak Photo CD PCD0992 [20] are adopted for the test and they are displayed as they are. We evaluate the proposed method against the pictures either generated by simulation tools or the ones taken in the field.

4.1. Simulation

The reference images and pattern images are scaled, rotated, and translated respectively at a series of levels to simulate a specific type of spatial distortion. The output images have the same resolution as the captured images. Since the actual distortions are known in prior for the simulation, the image registration accuracy can be evaluated with respect to the maximum absolute displacements of pixels from their ideal locations. In the cases scaling factors are greater than 1 (Figure 2a), the maximum displacements are below 0.2 pixel. These small errors are largely negligible if the capturing resolution is at least two times higher than the reference resolution (very typical). The lowest displacements are given for special rotation angles as expected (Figure 2b). In all other cases, the absolute displacements are between 0.5 and 2 pixels, and they correspond to the mis-adjustments of the contour fine-tune algorithm (Equation 4) due to blurring edges in the captured images. The proposed method is completely independent from spatial translations. We also scale and rotate all 24 testing images in a similar fashion and apply SSIM metric [21] (kernel size 5) to measure structural similarity. This is largely ignored by conventional image registration evaluations. The metric



(a)

(b)

(c)

Fig. 2. Simulated scaling on one image with max shift $<$ 1.2 pixels (a), simulated rotation on one image with max shift $<$ 2.5 pixels (b), simulated scaling and rotation on all images suggest the capturing resolution should be as large as possible and the image registration is insensitive to rotations (c)

incorporates the visibility of structural errors; it concerns the displayed image content and is able to detect complicated image quality issues like artefacts. Figure 2c (”S” for scaling, ”R” for rotation) and its corresponding statistics table (Table 1) illustrates that the mean of similarity increases rapidly and the variance becomes smaller and more stable. Image rotation has limited influence on the proposed method and since the structure similarity are always above 0.98.

4.2. Captured Images

In a controlled lab environment, we use the camera to take pictures of each of the projected reference image at 25 random locations and orientations in the field under low light (0 Lux) and dimmed light (30 Lux) conditions respectively, since the light condition has a great impact on the visual experience [22]. Then we apply SSIM metric to the registered images and their references due to the lack of ground truth for the projections. The minimum structural similarity is higher than 0.91 for all cases under the low light condition and it is still above 0.85 under the dimmed light condition. The variance of structural similarity between random locations are small, so the proposed method produces similar results despite the changes of camera position and orientations (Figure 3). The mean and variance under the dimmed light condition are worse (Table 2). This is because the ambient light reduces the contrast between projection boundary and its surroundings, and the adjustment accuracy of contour fine-tune algorithm is influenced.

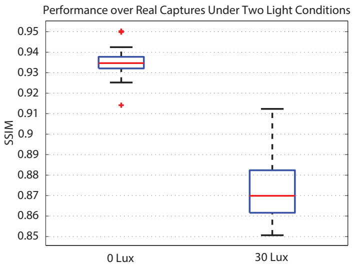


Fig. 3. SSIM for real captures under two light conditions, the image registration is more accurate and stable under the low light condition despite the changes of camera’s position and orientation

Table 1. Statistics of SSIM for simulation results

	S 1.5	S 2.0	S 2.5	R 15	R 30	R 45
min	0.815	0.887	0.924	0.968	0.966	0.970
max	0.964	0.978	0.985	0.994	0.993	0.993
95% int.	0.902 ± 0.017	0.940 ± 0.010	0.960 ± 0.007	0.983 ± 0.003	0.982 ± 0.003	0.984 ± 0.003

Table 2. Statistics of SSIM for real captures

	0 Lux	30 Lux
min	0.914	0.851
max	0.950	0.909
95% int.	0.935 ± 0.002	0.872 ± 0.004

5. CONCLUSION

In this paper, we propose a marker-less view independent method to use an un-calibrated camera to achieve a sub-pixel-level registration between the captured projections and their reference images. The preservation of geometrical order as well as the intensity and chromaticity relationships between two consecutive pixels on the display are maximized. The experimental results against distortion simulations and captured images prove that the registration accuracy is considerably high under typical light conditions for projection systems. By incorporating this method, we can apply existing full reference image quality metrics to captured projections without any modification to the metrics. In the future, we will integrate the method into an unified full reference metric based image quality assessment framework for projection displays, and study the perceptual properties of displayed images with respect to the correlations between human observations and metrics results despite of the actual projection geometries.

6. REFERENCES

[1] C. H. Son and Y. H. Ha, "Color Correction of Images Projected on a Colored Screen for Mobile Beam Projector," *Imaging Science and Technology*, vol. 52, no. 3, pp. 1–11, 2008.

[2] R. Raskar, M. S. Brown, R. Yang, W. C. Chen, G. Welch, H. Towles, B. Seales, and H. Fuchs, "Multi-projector Displays Using Camera-based Registration," in *IEEE Visualization*, San Francisco, USA, 1999, pp. 161–168.

[3] M. Pedersen, N. Bonnier, J. Y. Hardeberg, and F. Albregtsen, "Attributes of Image Quality for Color Prints," *Journal of Electronic Imaging*, vol. 19, no. 1, pp. 011016–1–011016–13, Jan. 2010.

[4] M. Pedersen, N. Bonnier, J. Y. Hardeberg, and F. Albregtsen, "Image Quality Metrics for The Evaluation of Print Quality," in *Proceedings of the SPIE*, S. P. Farnand and F. Gaykema, Eds., San Francisco, USA, Jan. 2011, vol. 7867, pp. 786702–786702–19.

[5] T. Eerola, J. K. Kamarainen, L. Lensu, and H. Kalviainen, "Framework for Applying Full Reference Digital Image Quality Measures to Printed Images," in *Scandinavian Conference on Image Analysis*, A. B. Salberg, J. Y. Hardeberg, and Robert. Jenssen, Eds., Oslo, Norway, 2009, vol. 5575, pp. 99–108, Springer Berlin Heidelberg.

[6] O. Bimber, D. Iwai, G. Wetzstein, and A. Grundhöfer, "The Visual Computing of Projector-Camera Systems," *Computer Graphics Forum*, vol. 27, no. 8, pp. 2219–2245, 2008.

[7] A. Agarwal, C. V. Jawahar, and P. J. Narayanan, "A Survey of Planar Homography Estimation Techniques," Tech. Rep., Centre for Visual Information Technology, 2005.

[8] Z. Y. Zhang, "Flexible Camera Calibration By Viewing a Plane From Unknown Orientations," in *International Conference on Computer Vision*, Kerkyra, Greece, 1999, pp. 666–673.

[9] P. F. Sturm and S. J. Maybank, "On Plane-Based Camera Calibration: A General Algorithm, Singularities, Applications," in *Computer Vision and Pattern Recognition*, Fort Collins, USA, 1999, pp. 1432–1437, IEEE.

[10] Z. Y. Zhang, "A Flexible New Technique for Camera Calibration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 11, pp. 1330–1334, 2000.

[11] R. Strand and E. Hayman, "Correcting Radial Distortion by Circle Fitting," in *Procedings of the British Machine Vision Conference*, Oxford, UK, 2005, pp. 9.1–9.10, British Machine Vision Association.

[12] T. Thormahlen, H. Broszio, and I. Wassermann, "Robust Line-Based Calibration of Lens Distortion From A Single View," in *International Conference on Computer Vision / Computer Graphics Collaboration Techniques and Applications*, Nice, France, 2003, pp. 105–112.

[13] F. Bukhari and M. N. Dailey, "Robust Radial Distortion from a Single Image," in *International Symposium on Visual Computing*, Las Vegas, USA, 2010, vol. 6454, pp. 11–20.

[14] H. Arora and A. Namboodiri, "Projected Pixel Localization and Artifact Removal in Captured Images," in *IEEE Region 10 International Conference*, Hyderabad, India, Nov. 2008, pp. 1–5, IEEE.

[15] L. Zhang and S. Nayar, "Projection Defocus Analysis for Scene Capture and Image Display," *ACM Transactions on Graphics*, vol. 25, no. 3, pp. 907–915, 2006.

[16] S. Suzuki and K. Abe, "Topological Structural Analysis of Digitized Binary Images by Border Following," *Computer Vision, Graphics, and Image Processing*, vol. 30, no. 1, pp. 32–46, 1985.

[17] A. Fitzgibbon and R. Fisher, "A Buyer's Guide to Conic Fitting," in *British Machine Vision Conference*, Birmingham, UK, 1995, pp. 513–522, British Machine Vision Association.

[18] G. Bradski and A. Kaehler, *Learning OpenCV*, OReilly Media, Inc., Sebastopol, CA, USA, first edit edition, 2008.

[19] P. S. Heckbert and M. Garland, "Survey of Polygonal Surface Simplification Algorithms," Tech. Rep. May, School of Computer Science, CarnegieMellon University, 1997.

[20] R. Franzen, "PhotoCD PCD0992," 1999.

[21] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image Quality Assessment: from Error Visibility to Structural Similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–12, Apr. 2004.

[22] P. Zhao, M. Pedersen, J. Y. Hardeberg, and J. B. Thomas, "Camera-based Measurement of Relative Image Contrast in Projection Displays," in *European Workshop on Visual Information Processing*, Paris, France, 2013, pp. 112–117, IEEE.

Chapter 8

Paper C

Perceptual Spatial Uniformity Assessment of Projection Displays with a Calibrated Camera

Ping Zhao, Marius Pedersen, Jean-Baptiste Thomas, and Jon Yngve Hardeberg

Published in

Proceedings of

22nd Color and Imaging Conference

(CIC 2014),

Volume 2014, Number 2014, Page 159-164,

Society for Imaging Science and Technology,

Boston, MA, USA,

November, 2014.

Perceptual Spatial Uniformity Assessment of Projection Displays with a Calibrated Camera

Ping Zhao, Marius Pedersen, Jean-Baptiste Thomas, and Jon Yngve Hardeberg; Gjøvik University College, Gjøvik, Norway; *Université de Bourgogne, Dijon, France.*

Abstract

Spatial uniformity is one of the most important image quality attributes in visual experience of displays. In conventional researches, spatial uniformity was mostly measured with a radiometer and its quality was assessed with non-reference image quality metrics. Cameras are cheaper than radiometers and they can provide accurate relative measurements if they are carefully calibrated. In this paper, we propose and implement a work-flow to use a calibrated camera as a relative acquisition device of intensity to measure the spatial uniformity of projection displays. The camera intensity transfer functions for every projected pixels are recovered, so we can produce multiple levels of linearized non-uniformity on the screen in the purpose of image quality assessment. The experiment results suggest that our work-flow works well. Besides, none of the frequently referred uniformity metrics correlate well with the perceptual results for all types of test images. The spatial non-uniformity is largely masked by the high frequency components in the displayed image content, and we should simulate the human visual system to ignore the non-uniformity that cannot be discriminated by human observers. The simulation can be implemented using models based on contrast sensitivity functions, contrast masking, etc.

Introduction

In the past decade, tremendous growth in the use of digital media implies that our daily life and work have been greatly impacted by the rapid advancement of display technologies. Hence, the image quality assessment of displays become an essential topic for both scientific research and industrial communities. Projection displays have advantages like high resolution, portability and flexibility. For example, multiple projectors can be tiled up to form a large perceptual photometric seamless image [1]. It is cost effective for users to visualize information in a very high resolution without issuing a customized manufacturing demand.

In general, the image quality of displays can be characterized by groups of image quality attributes. One group of them includes physical screen dimension, display resolution, refreshing rate, viewing distance, and viewing angle etc. These attributes are associated with a specific display and its viewing condition. They indeed have impacts on the perceptual image quality, but in most cases they are assumed to be constants within one working cycle of image quality assessment. The rest of the attributes include, but are not limited to, brightness, contrast, color gamut, sharpness and artifacts (including noises). Among these attributes, the spatial uniformity can be of a major importance for projection displays [2, 3]. Researchers tried to achieve objective spatial uniformity with mathematical modeling, but soon they realized that some restraints can be relaxed due to the limitation of perception of Human Visual System (HVS) [1]. In recent studies [4, 5, 6], radiometers were used as absolute acquisition devices to measure the luminance and chrominance of projection displays. Measuring with radiometers is time consuming. The devices are expensive and they are likely to be unavailable in real practice. Digital still cameras have been used to record projection pixels including its background and surrounding on the displays [7, 8, 9]. Cameras have the advantage that they can be placed at different locations in order to achieve a location- and view-dependent image quality assessment, and the acquisition process are much accelerated. However, cameras offer relative sensor responses upon the incoming lights, so they need to be carefully calibrated in advance.

In this paper, we propose and implement a work-flow to use a calibrated camera as a relative acquisition device to record the intensity of projections, and evaluate the spatial uniformity of projections against image quality metrics. The correlation between perceived and measured results suggest that the camera based image quality assessment can be reliable and accurate.

This paper is organized as follows: first, in the background section, we review the existing image quality metrics for spatial uniformity assessment. Then in uniformity assessment section, we describe the setup of our control lab environment, and demonstrate how to calibrate a camera and a projector to produce multiple levels of linearized non-uniformity on the projection screen. In addition, we also describe the experiment procedure and show the experiment results. In the last section, the conclusions and future works are presented.

Background

Many uniformity metrics have been proposed based on luminance measurements of gray patches. Among the international standards for image quality assessment of displays, FPDM [10] defines uniformity as $(100\% \cdot (L_{min}/L_{max}))$, where L_{min} and L_{max} stand for minimum and maximum measured luminance respectively. TCO 6.0 [11] defines a compliance threshold based on four luminance samples as (L_{max}/L_{min}), assuming that the minimum luminance is not even close to zero. SPWG [12] defines uniformity as $(100\% \cdot (L_{max} - L_{min})/L_{max})$ based on thirteen independent luminance measurements. These metrics associate uniformity with Luminance Ratio (LR) between two extreme pixels.

However, Tang [13] and Ngai [14] demonstrated that the LR based methods have inaccurate predictions of the non-linear HVS. Tang [13] incorporated the viewing distance d and spatial derivatives s of luminance to define the uniformity as $SFA = d^2 \left(L_{max} + L_{min} - 2\overline{L}\right)/s^2$, where $\overline{L}$ stands for the average of measured luminance. Further research from Samuelson et al. [21] quantify the image quality of an illuminated surface with a proposed spatial frequency analysis algorithm incorporating the difference of Gaussian function, and they suggested that the average magnitudes of luminance contrast within selected spatial frequency bands are related to the lighting quality of the scene represented by the image. Beyond these studies, Ashdown [15] investigated the influence of Luminance Gradient (LG) on the spatial uniformity and they indicated that their results were more correlated to the subjective perceptual ratings than LRs. However, these studies ignore the factor of viewing distance which is important to the uniformity assessment. Meanwhile, other metrics based on statistical analysis and/or color distances in specific color spaces were proposed. Poulin et al. [16] proposed a metric to determine the spatial uniformity as $(100\% - STDEV(L))$, where $STDEV(L)$ stands for the standard deviation (STDEV) of luminance. Thomas et al. [4] used color differences ΔL and ΔC measured with a spectroradiometer. The results suggested that the chromaticity shifts are significant and they should be accounted for. Another statistics based uniformity is defined as the variation of coefficient $\sqrt{\sum_{i=1}^{N}\left(L_i - \overline{L}\right)} / \left(\overline{L}(N-1)\right)$, where N stands for the number of sample points [17].

There are also existing works that are relevant to the uniformity assessment based on captured images. In the domain of printing, Green [18] proposed a metric for measuring smoothness of color transforms. The metric computes the second derivative from the vector of color differences. Besides, wavelet analysis [19] and standard deviation [20] are common methods to analyze non-uniformity (mottle). These methods were originally designed for printings, but they can be used for displays in a similar fashion.

Uniformity Assessment

In this section, we describe the experimental setup and how to use a camera and a projector to produce multiple levels of non-uniformity on the projection display.

Experimental Setup

The experiments take place in a controlled lab environment where the only illuminant in the room is the projector. We use a portable three chip LCD projector SONY APL-AW15 (throw ratio: 1.5) to produce projections on a planar screen which is naturally hanging on the ceiling. The projector is placed on a table in front of the projection screen, and the distance is approximately 3m with respect to the throw ratio of the projector. A remote controlling laptop is connected to the projector via a VGA cable in order to generate full screen projections which have resolution 1280 × 768 in pixels. On the screen, the dimension of projection area is approximately 2 × 1.2 in meters. We use a DLSR Nikon D610 which has an imaging resolution 6048 × 4016 in pixels and with a Sigma VR 24-105mm f/4G (VR off) lens to capture the projections. The camera is fixed on a tripod and the pictures are taken remotely with a software control on the laptop without physically touching the camera. The pictures are saved in raw format and rendered with Aliasing Minimization and Zipper Elimination demosaicing algorithm [22] without automatic vignette correction, brightness adjustment, gamma correction and noise reduction etc.

Vignetting Correction

Captured pictures are known to have vignetting effect which stands for an undesirable gradual intensity fall off from the image center to its external limits. In this paper, we correct camera vignetting based on the captures of a hazy sky which is closely uniform in gray [23]. In the lab, we take several trial shots of projections with either minimum or maximum projector input intensity. In this process, we adjust the camera settings iteratively until all the captures are neither underexposure nor overexposure. Then we keep all camera settings except exposure time, hook a neutral light diffuser (white and semi-transparent) over the camera lens, and use the camera to take multiple pictures toward the same spot of the hazy sky. Each time we take a picture we rotate the camera a bit. Then we calculate the intensity median response for each camera pixel over all pictures we have taken, and use them to generate a vignetting mask which is then applied to the camera RGB channels separately to correct the vignetting.

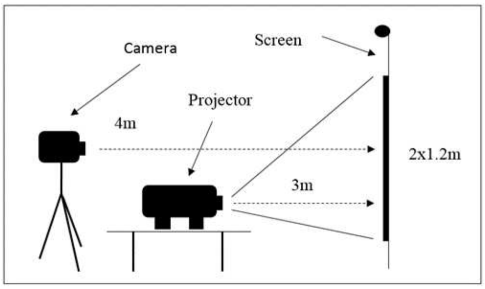


Figure 1. *Experiment setup (throw ratio: 1.5)*

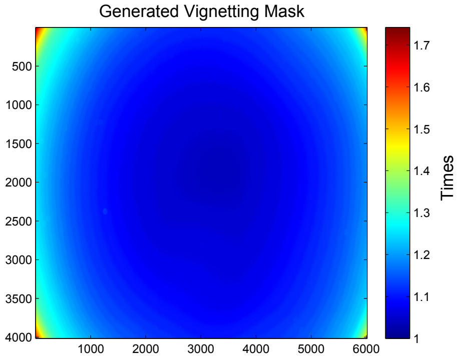


Figure 2. *The generated vignetting mask for our camera Nikon D610 with a Sigma VR 24-105mm f/4G (VR off) lens*

In the experiment, we take 60 pictures of the hazy sky and put 40 of them into a training set and the rest into a validation set. The median responses are obtained based on 5, 10, ..., and 60 pictures in the training set respectively. Then we apply corresponding masks to the pictures in the validation set. The minimum averaged standard deviation over all validation pictures indicate that empirically 10 pictures are sufficient to generate convergent median results. The mask we generate for our camera is shown in Figure 2. We can see that the vignetting is not even closely symmetric. The center has shifted upward and also a bit to the right. This observation is contrary to common assumptions about the vignetting symmetry in many literature (cos four law [24] for example). In order to maximize the validity and reliability of image quality assessment, we should offer the best effort to avoid assumptions. Our method places no assumption about the camera or the light condition, and the whole procedure can be finished within a few minutes.

Exposure Optimization

The daylight environment has a much higher luminance (normally above 1000 Lux) than the projection environment (around 10 Lux for example). In order to avoid either underexposure or overexposure of captures, the camera's exposure time varies between the two light conditions. The vignetting mask generated in a daylight condition might not be appropriate for the low light condition, since in this context we implicitly assume that all camera sensors have linear responses. In order to verify the linearity, we equally separate the range of projector input intensity into 15 levels. For each level, we display a gray patch and capture it under all possible camera exposure times ranging from 1/4000s to 30s. Meanwhile, we use a light meter to measure the physical luminance on the projection screen as a reference to the camera. Then we construct surfaces of camera intensity responses versus the projector luminance and camera's exposure time.

The first picture in Figure 3 depicts the intensity response of one camera sensor in the red channel. In the deep blue region, the responses are closely linear to all possible projector luminance while the exposure time is fixed, and vice versa. However, in the aqua regions, such sensor has a large boost in responses. It may be argued that this is because the camera sensor is closely saturated in these cases. Then we can have a look at the second picture in Figure 3 where the responses of another camera sensor in the green channel is obviously not saturated. In this case, the boost is still available at the areas where the blue and aqua regions intersect. In this context, we can see that the camera gives linear responses corresponding to limited combinations of projector luminance and camera's exposure time.

This conclusion seems to be trivial because the exposure time should be kept below 2s in most cases. However, in order to apply the vignetting mask generated in a high light condition to a low light condition, we have to make sure that the camera responses are all linear with respect to a common exposure time. For this reason, we determine the strongest responses over each camera intensity response for the maximum luminance under the two light conditions with a common exposure time, and we continue to decrease the exposure time until the ratios between such two sensor responses are equal. However, applying an exposure time which is too small would not take the full advantage of the dynamic range of the camera. Once this condition is met, the camera's exposure time is optimized, and the generated vignetting mask can be applied to the camera despite of light conditions.

Image Registration

In the context of image quality assessment incorporating full reference metrics, it is critical to achieve accurate and automatic image registration between the captured image and its reference image. Then we can apply existing full reference metrics with no modification to them. The preservation of geometrical order as well as the color relationships between two consecutive pixels on displays are maximized. In our previous research [25], we proposed a marker-less and view-independent method to use a camera to do the image registration. The maximum pixel shift error is below 0.2 pixel in the cases that camera resolution is higher than projection resolution. With respect to the performance evaluations with SSIM metric [28], the image registration accuracy is higher than 0.95 in a dark room environment and it is above 0.9 in a dimmed light condition where ambient light is below 30 Lux. In this paper, we adopt this method to extract projections from the captured images, and make them have exactly the same dimension and resolution as their reference images.

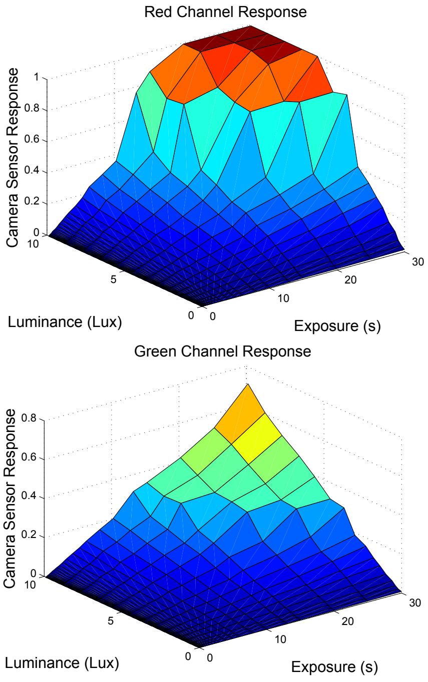


Figure 3. *Intensity responses of one camera sensor in the red channel (1st picture), and another sensor in the green channel (2nd picture)*

Projector Calibration

In order to assess the perceptual spatial uniformity, we produce multiple levels of linearized non-uniformity on the screen to be observed and captured. Brown et al. [9] located the minimum common achievable projector response for all pixels and generated a luminance attenuation map to correct the projection colors. The linearity of projector's intensity responses is assumed; otherwise the inverse projector intensity transfer function is applied to compensate for that. Pagani et al. [23] proposed a shading table based automatic uniformity correction. The colors of each shading point are corrected by iteratively refining the projector output intensities in order to avoid temporary stability problem of projectors, and the colors of other pixels are linearly interpolated based on its shading point neighbors.

In this research, we adopt and extend the method proposed by Brown et al. [9] for its simplicity and effectiveness. First, we equally separate the range of projector input intensity into 15 levels, and for each level we display and capture a gray patch 10 times. Then the projector intensity transfer functions can be recovered by polynomial regression upon the median responses over the gray patches at all intensity levels. In this way, we can avoid the temporary stability problems of both camera and projector. However, it is computational inefficient to determine the regression coefficients for all twenty million pixels of the camera Nikon D610. We calculate only the coefficients for the reference pixel which gives the lowest camera sensor response upon maximum projector luminance. The coefficients for other pixels can be obtained by linearly scaling the one of the reference pixel.

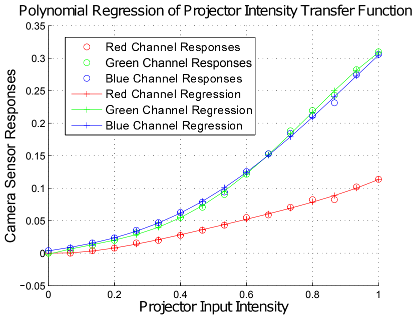


***Figure 4.** Polynomial regression of camera sensor responses*

The method is processed in a color channel basis. After this, we inverse the regression functions to compensate the non-linearity of camera responses in order to create flattened projections. In our experiment, 5th order polynomial regression is sufficient to achieve good approximation. The projectors intensity transfer functions for the reference pixel are depicted in Figure 4. Lower order regressions (2nd order for example) produces slightly different curves, but they may cause overcasting of dominated colors. The polynomial regression may produce negative values which are invalid. In such cases, we simply clip them because the absolute values are small ($< 1e-3$) to be negligible.

Suppose that the scaling ratio of one pixel p_{ij} in an individual color channel on the ith row and jth column of the registered image is $r_{ij} \geq 1$, the corresponding regression function for the reference pixel is $f(x)$ and its inverse function is denoted as $f^{-1}(x)$. The x stands for the projector input intensity of the pixel p_{ij}. The camera response of pixel p_{ij} is denoted as $c_{ij} = f(x) \cdot r_{ij}$. In this context, the projector input intensity for the pixel p_{ij} at a certain non-uniformity level is defined as $g(x) = f^{-1}\left(f(x) \cdot s\left(r_{ij} - m\right)\right)$, where $m = \sum_{i=1}^{n_y} \sum_{j=1}^{n_x} r_{ij} / \left(n_x \cdot n_y\right)$, n_x and n_y stand for the width and height for the projection in pixels respectively, and s stands for a linear scaling factor of non-uniformity and it is under the constraint that $f(0) \cdot r_{ij} \leq g_{ij}(x) \leq f(x) \cdot r_{ij}$ assuming that the projector input intensities are normalized to between 0 and 1. The value of r_{ij} can be determined as $max\left(c_{ij}\right)/f(1)$, where the operator max stands for the maximum value of c_{ij}.

Experimental Procedure

We incorporate human observers and full reference image quality metrics to assess the spatial uniformity of projection displays. We display seven types of test images (see Figure 5): two natural color pictures (the 15th and 23th picture from Kodak Photo CD PCD0992 [26]), three uniform colored patches with opponent colors: yellow, magenta and cyan respectively, one gray patch (the gray level equals to 0.5) and one slide like image with dark texts on a gradient background. For each test image, we linearly scale its natural projection's non-uniformity to produce multiple levels of non-uniformity. These scaling ratios can be normalized into the range between -1 and 1, and then we split it into five levels: -0.6, -0.2, 0, 0.2 and 0.6. The level 0 corresponds to flattened projections where the projector's natural non-uniformity is canceled. We also display one image reserving the projectors natural non-uniformity by displaying the image as it is; so 42 images in total are presented to each human observer.

The viewing condition is similar to a home theater environment where the room is dark, and the observers are located at the camera's position. The test images are displayed to observers in a randomized order. The experiment is set up as a category judgment experiment. So, for each displayed image, the observers are asked to indicate the perceptual uniformity with a category label which stands for the rank between not uniform at all and perfectly uniform corresponding to the ratings numbers from 1 to 5. At the same time, the observers are also asked to indicate how the non-uniformity affect their pleasantness with a category label which stands for the rank between very disturbing and not disturbing at all corresponding to the rating numbers from 1 to 5. The perceptual ratings are collected from 10 human observers and then they are scaled to generate Z-scores [28].

We also evaluate the uniformity with the following image quality metrics: LR defined in VESA FPDM [10], LG based definition [13] (SFA), averaged standard deviation of RGB values (Stddev), coefficients of variation [17] (Coeff), averaged Euclidean distance $\overline{\Delta E_{ab}^{*}}$ in CIELAB color space (ΔE_{ab}), PSNR-M [27], SSIM [28], and S-CIELAB [29]. The first four metrics are commonly referred uniformity metrics in literature, while the metrics $\overline{\Delta E_{ab}^{*}}$ is frequently referred to determine the perceptual distance between two colors. Since the non-uniformity changes the structure information in the images, we adopt SSIM as well.

Subjective Results

The first observation is that the rank order of non-uniformity is largely preserved for the seven types of test images as expected (see Figure 5). If we assume that the general tendency of Z-scores are smooth, then they can be represented by parabolic curves. The curves might be more or less skewed depends on the projected image content. The flattened projections do not necessary correspond to the highest overall Z-scores, while small negative non-uniformity and natural projection images have similar or relative lower Z-scores in many cases, and either positive or negative large non-uniformity leads to the lowest Z-scores. This observation supports the fact that HVS is not sensitive to small variation of non-uniformity. The spatial non-uniformity is largely masked by the high frequency components in the displayed image content, and we should simulate the human visual system to ignore the non-uniformity that cannot be discriminated by human observers. The simulation can be implemented using models based on contrast sensitivity functions, contrast masking, etc. For the distorted slide like images (correspond to the 7th test image), the Z scores of flattened versions are clearly greater than others (higher mean value and no overlapping of confidence intervals). This is because such reference image has dark texts on a large gradient background in a bright color, and the non-uniformity on a gradient background can be easier to be detected by HVS than that on a flat background which is the case of a gray patches (correspond to the 1st test image). The general tendency of mean Z-scores of pleasantness are very similar to the ones of perceived uniformity and the Pearson correlation between them are all above 0.98 for all test images, except the absolute mean values of pleasantness are slightly larger in general. This observation suggests that the HVS has a certain degree of but limited tolerance on average against non-uniformity on the display. For the gray patch test images, the observers have a difficulty to distinguish the differences between the small minus non-uniform, flattened, natural projections. In a similar fashion, the pleasantness of small minus non-uniformity, flattened and natural projections for the two natural images have

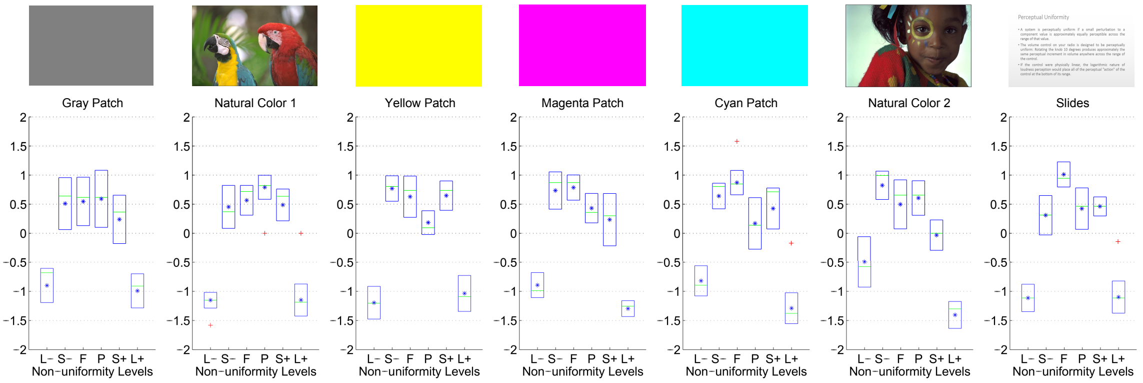


Figure 5. *Mean Z-scores of perceived uniformity, the blue box indicates the 95% confidence interval of Z-scores, the green bars stand for the median values of Z-scores, and the red crosses stand for the outliers with respect to quartile based statistic tests (against outer fences)*

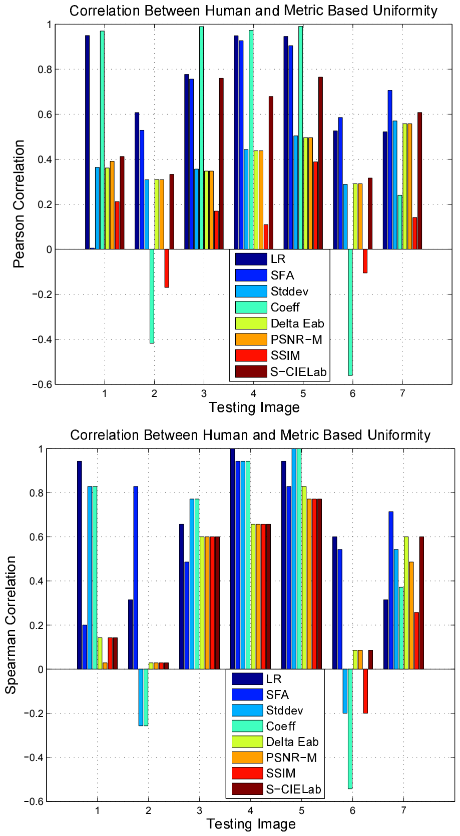


Figure 6. *Pearson (1st picture) and Spearman (2nd picture) correlations between the mean Z-scores of subjective ratings and objective metric results*

similar values but their corresponding perceived uniformity have different mean values. This observation suggests that the non-uniformity is masked by the complex colors of natural pictures and in such cases achieving a restrained uniform is not the only way to produce the best perceptual experience.

Objective Results

Figure 6 demonstrates the Pearson and Spearman correlations between the mean Z scores of perceived uniformity and objective results from all metrics. Obviously, none of these metrics works well for all types of images, especially for natural color images (the 2nd and 6th test images). Simple metrics like LR and SFA work surprisingly better than others in many cases. We think this might because in our experiment the non-uniformity for all pixels is globally scaled, so the rank order of intensities in each primary color channel is largely preserved; although we apply negative scalars to non-uniformity as well, the magnitude of scaled non-uniformity is still comparatively smaller than the reference intensity values in the reference images. However, in real practice, the non-uniformity level of projections should be relatively small, otherwise the optical components of such a projector should be replaced with new ones. The metric Coeff also gives high correlation scores for patches but negative values for natural pictures (the 2nd and 6th test images). However, no metric works well for the natural color images and slide like images (the 7th test images). In such cases, the correlation values are largely below 0.6. S-CIELAB also adopts CSF but it has slightly better correlation results than PSNR-M and SSIM metrics in all cases. It is also interesting to figure out the reason why metric LR does not work well in many cases, so we generate the plots of the subjective results versus the objective results for the LR metric (see Figure 7). It is clear that for the non-patch test images, the variance of metric scores are largely compressed and a few outliers are visible. By examining the metric score values, we find out that these outliers correspond to the flattened projection and natural projection. Similar phenomenon can be observed for other metrics. It suggests that either the metrics give lower values for the flattened projection, or higher values for the natural projection comparing to their expected values. In other words, the distance between the two consecutive levels of perceived uniformity is more compressed than the results of metrics.

Conclusion and Future Works

In this paper, we propose and implement a work-flow to use a calibrated camera as a relative acquisition device of intensity to measure the spatial uniformity of projection displays. The experimental results suggest that none of the frequently referred spatial uniformity metrics works well for all types of test images, especially for the flattened projections and natural projection of natural color images. In such cases, The spatial non-uniformity is largely masked by the high frequency components in the displayed image content, and we should simulate the human visual system to ignore the non-uniformity that cannot be discriminated by human observers. The simulation can be implemented using

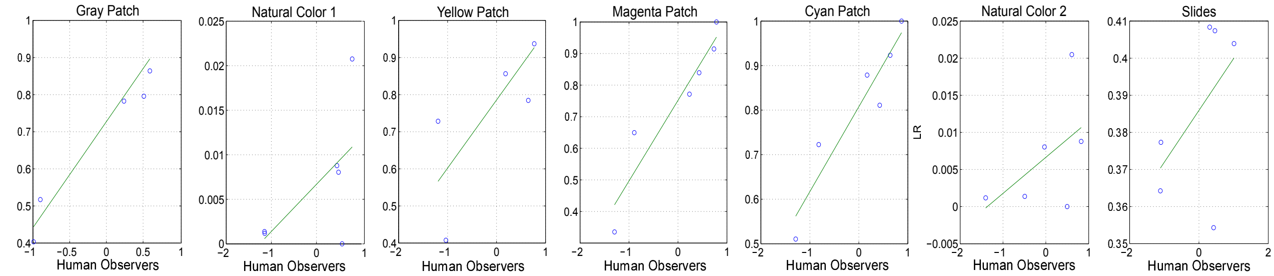


Figure 7. LR metric scores versus subjective mean Z-scores for all test images

models based on contrast sensitivity functions, contrast masking, etc. In addition, the colors can be considered to be transformed into the frequency domain and analyzed at a smaller granularity in order to engage the issue of contrast masking. In the coming future, we should either improve the existing metrics or design a new one to evaluate the spatial uniformity of projection displays. Such a metric should be incorporated into a unified image quality assessment framework for projection displays.

References

[1] A. Majumder and R. Stevens, Perceptual Photometric Seamlessness in Projection-based Tiled Displays, ACM Trans. Graph., vol. 24, no. 1, pp. 118-139, Jan. 2005.

[2] A. M. Bakke, J. B. Thomas, and J. Gerhardt, Common Assumptions in Color Characterization of Projectors, in Gjøvik Color Imaging Symposium, 2009, pp. 45-53.

[3] J. Thomas, A. M. Bakke, and J. Gerhardt, Spatial Nonuniformity of Color Features in Projection Displays: A Quantitative Analysis, J. Imaging Sci. Technol., vol. 54, no. 3, p. 030403, 2010.

[4] J. B. Thomas and A. M. Bakke, A Colorimetric Study of Spatial Uniformity in Projection Displays, in Computational Color Imaging, vol. 5646, A. Trémeau, R. Schettini, and S. Tominaga, Eds. Springer Berlin Heidelberg, 2009, pp. 160-169.

[5] G. Menu, L. Peigne, J. Y. Hardeberg, and P. Gouton, Correcting Projection Display Non-uniformity Using A Webcam, in Color Imaging X: Processing, Hardcopy, and Applications, 2005, pp. 364-373.

[6] J. B. Thomas, Colorimetric Characterization of Displays and Multi-display Systems, Université de Bourgogne, 2009.

[7] R. Raskar, M. S. Brown, R. Yang, W. C. Chen, G. Welch, H. Towles, B. Seales, and H. Fuchs, Multi-projector Displays Using Camera-based Registration, in IEEE Visualization, 1999, pp. 161-168.

[8] R. I. Hartley, Self-Calibration from Multiple Views with a Rotating Camera, in European Conference on Computer Vision, 1994, pp. 471-478.

[9] M. Brown, A. Majumder, and R. Yang, Camera-based Calibration Techniques for Seamless Multiprojector Displays, IEEE Trans. Vis. Comput. Graph., vol. 11, no. 2, pp. 193-206, 2005.

[10] Flat Panel Display Measurements Standard 2.0. Video Electronics Standards Association, Newark, USA, pp. 1-18, 27-Oct-2005.

[11] TCO Certified Displays 6.0. TCO Development AB, Stockholm, Sweden, pp. 1-125, 27-Oct-2012.

[12] SPWG Notebook Panel Specification 3.8. Standard Panel Working Group, Newark, USA, pp. 1-59, 2007.

[13] T. Ling, The Assessment of Ceiling Uniformity for Indirect Lighting Systems, 1996.

[14] P. Y. Ngai, The Relationship Between Luminance Uniformity and Brightness Perception, J. Illum. Eng. Soc., vol. 29, no. 1, pp. 41-50, Jan. 2000.

[15] I. Ashdown, Luminance Gradients: Photometric Analysis and Perceptual Reproduction, J. Illum. Eng. Soc., vol. 25, no. 1, pp. 69-82, Jan. 1996.

[16] F. Poulin and M. Caron, Display Measurement - A Simple Approach to Small-area Luminance Uniformity Testing, Television Displays, pp. 1-9, 2009.

[17] M. S. Rea, Ed., The IESNA Lighting Handbook - Reference and Application, 9th editio. Illuminating Engineering, 2000, pp. 1-1000.

[18] P. J. Green, A Smoothness Metric for Colour Transforms, in Color Imaging XIII: Processing, Hardcopy, and Applications, 2008, vol. 6807, p. 68070I-68070I-5.

[19] J. P. Bernie, H. Pande, and R. Gratton, A New Wavelet-based Instrumental Method for Measuring Print Mottle, Measuring Instruments, vol. 9, pp. 197-199, 2004.

[20] A. Sadovnikov, P. Salmela, L. Lensu, J. K. Kamarainen, and H. Klvi-inen, Mottling Assessment of Solid Printed Areas and Its Correlation to Perceived Uniformity, in Scandinavian Conference on Image Analysis, 2005, vol. 3540, pp. 409-418.

[21] C. E. Samuelson, I. Ashdown, P. Kan, A. Kotlicki, and L. A. Whitehead, A Proposed Lighting Quality Metric Based on Spatial Frequency Analysis, in Illuminating Engineering Society Annual Conference, 1999, pp. 51-61.

[22] E. Martinec and P. Lee, AMAZE Demosaicing Algorithm, 2010. [Online]. Available: http://www.rawtherapee.com/. [Accessed: 04-May-2014].

[23] A. Pagani and D. Stricker, Spatially Uniform Colors for Projectors and Tiled Displays, J. Soc. Inf. Disp., vol. 15, no. 9, p. 679, 2007.

[24] D. A. Kerr, Derivation of the Cosine Fourth Law for Falloff of Illuminance Across a Camera Image.pp. 1-12, 2007.

[25] P. Zhao, M. Pesersen, J. Y. Hardeberg, and J. B. Thomas, Image Registration for Image Quality Assessment of Projection Displays, in IEEE International Conference on Image Processing, 2014.

[26] R. Franzen, PhotoCD PCD0992, Kodak Lossless True Color Image Suite, 1999. [Online]. Available: http://r0k.us/graphics/kodak/. [Accessed: 11-May-2014].

[27] N. Ponomarenko, F. Silvestri, K. Egiazarian, M. Carli, J. Astola, and V. Lukin, On Between-coefficient Contrast Masking of DCT Basis Functions, in Proceedings of the Third International Workshop on Video Processing and Quality, 2007, pp. 1-4.

[28] Engeldrum, P. G. (2000). Psychometric Scaling: A Toolkit for Imaging Systems Development (pp. 1-200). Imcotek Pr.

[29] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, Image Quality Assessment: from Error Visibility to Structural Similarity, IEEE Trans. Image Process., vol. 13, no. 4, pp. 600-612, Apr. 2004.

[30] G. M. Johnson and M. D. Fairchild, A Top Down Description of S-CIELAB and CIEDE2000, Color Res. Appl., vol. 28, no. 6, pp. 425-435, Dec. 2003.

Chapter 9

Paper D

Measuring The Relative Image Contrast Of Projection Displays

Ping Zhao, Marius Pedersen, Jon Yngve Hardeberg, and Jean-Baptiste Thomas

Published in

Journal of Imaging Science and Technology (JIST),

Volume 59, Issue 3, Page 030404-1-030404-13,

Society for Imaging Science and Technology,

April, 2015.

Journal of Imaging Science and Technology® 59(3): 030404-1–030404-13, 2015.
© Society for Imaging Science and Technology 2015

Measuring the Relative Image Contrast of Projection Displays

Ping Zhao▲, Marius Pedersen▲, and Jon Yngve Hardeberg▲
Gjøvik University College, Gjøvik, Norway
E-mail: marius.pedersen@hig.no

Jean-Baptiste Thomas▲
Université de Bourgogne, Dijon, France

Abstract. *Projection displays, compared to other modern display technologies, have many unique advantages. However, the image quality assessment of projection displays has not been well studied so far. In this paper, we propose an objective approach to measure the relative contrast of projection displays based on the pictures taken with a calibrated digital camera in a dark room where the projector is the only light source. A set of carefully selected natural images is modified to generate multiple levels of image contrast. In order to enhance the validity, reliability, and robustness of our research, we performed the experiments in similar viewing conditions at two separate geographical locations with different projection displays. In each location, we had a group of observers to give perceptual ratings. Further, we adopted state-of-art contrast measures to evaluate the relative contrast of the acquired images. The experimental results suggest that the Michelson contrast measure performs the worst, as expected, while other global contrast measures perform relatively better, but they have less correlation with the perceptual ratings than local contrast measures. The local contrast measures perform better than global contrast measures for all test images, but all contrast measures failed on the test images with low luminance or dominant colors and without texture areas. In addition, the high correlations between the experimental results for the two projections displays indicate that our proposed assessment approach is valid, reliable, and consistent.* *© 2015 Society for Imaging Science and Technology.*
[DOI: 10.2352/J.ImagingSci.Technol.2015.59.3.030404]

INTRODUCTION

Flat-panel display technologies in the liquid crystal display (LCD) family have dominated the consumer market for many years, and this is especially true for desktop/laptop monitors, mobile phone screens, televisions, and many large outdoor information displays. The strongest appeal for consumers on displays is perhaps the ability to share information and collaborate with teammates quickly, easily, and conveniently. Projection displays, compared to other display technologies, have unique advantages in terms of portability, flexibility for deployment, and large screens to visualize information for a target audience. Recently, there has been an increasingly general interest to embed mini projectors into portable imaging devices such as smart phones or handheld video recorders, so that the pictures/videos can be reviewed and shared with a crowd of people in the field right after they are recorded.[1,2] In more advanced scenarios, multiple projections can be tiled up to produce a single large perceptual seamless image which visualizes information to target audiences, and they will enjoy a fully immersive visual experience.[3,4] In this context, image quality assessment of projection displays has gradually become an essential topic in both academic research and industrial commercial communities. The goal of image quality assessment is not limited to the establishment of a unified approach to evaluate the quality of image reproductions, but also in defining a systematic way to continuously improve the perceptual image quality within a closed work flow.

Image quality can be characterized and interpreted based on a set of image quality attributes which are terms of human perceptions of lightness, contrast, colors, sharpness, and artifacts (including noises).[5] Physical properties such as screen dimension, display resolution, and refreshing rate have impacts on the perceived image quality, but in a typical work flow of image quality assessment they can be assumed to be constants, since they are independent from image content and normally do not vary over time. In this paper, we only focus on the image quality attributes which are content independent. According to the existing literature, there have been many attempts to characterize displays such as cathode-ray tube (CRT)[6,7] and LCD[6–8] desktop/laptop monitors. The characterization of projection displays has a similar approach. Previous characterizations of projection displays primarily focused on black level estimation,[9] display uniformity,[10–12] and colorimetry,[11,13] but limited attention has been paid to measuring the contrast of projected images on the screen. More specifically, the measured contrast of an image has been shown to be of a significant impact on the visual experience.[14–16]

Experiments measuring the contrast of projection displays have largely been conducted based on absolute acquisitions with a radiometer or a spectrometer, etc.[11,17,18] These devices are well designed to produce accurate measurements, but they are expensive and require professional training; in a typical projection environment where it is common to have a low light condition, it takes a long time to collect a large number of measurements at discrete sample spots. Using a

▲ IS&T Members.

Received Mar. 22, 2015; accepted for publication June 1, 2015; published online June 22, 2015. Associate Editor: Maria V. Ortiz Segovia.
1062-3701/2015/59(3)/030404/13/$25.00

camera as a relative acquisition device has the advantage of recording all displayed pixels in one shot.[19] Once we have the captured images, we can process them with image quality measures to predict the actual image contrast and correlate these results with perceptual ratings. So, camera based acquisition can be a fast alternative approach to measure the relative contrast of projection displays at low cost.

This paper presents a study on the measurement of relative contrast of projection displays based on acquisitions with a digital single-lens reflex (DSLR) camera. The main goal of this work is to evaluate state-of-art contrast measures based on their correlations with subjective ratings. The results of the evaluation can be used to improve the design of image quality measures, and they can also to be extended in the development and enhancement of general image reproduction technologies.

This paper is organized as follows. First, in the next section, we introduce the background of image contrast and the state-of-art of contrast measurements. Then, in the third section, a full description of the experimental environment, setup, and experimental procedure is given. The results and discussions on the interaction between measured contrast and perceptual contrast are presented in the fourth section. Finally, in the fifth section, conclusions are drawn based on the data analysis.

CONTRAST MEASURES

The contrast measures for images can be broadly classified into two categories with respect to their measurements at either the global or local level. The global contrast measures determine the contrast at each pixel or a few representative pixels of the input image; so the contrast operator is applied individually to each component without involving its neighborhoods. However, at the local level of contrast measurement, the neighborhoods are involved possibly by following a hierarchical structure.

Global Contrast Measures

It is important to have a clear understanding of what image contrast is before we start to measure it. However, giving a comprehensive definition of perceptual contrast can be difficult, because it depends on how subjective the observers are, how the observers are related to the observation task, and how much experience the observers turn out to have.[20] Due to these difficulties, the early research on perceptual contrast confined itself to controlled viewing conditions with limited types of visual stimuli. The studies began with measuring the contrast of a periodic pattern such as a sinusoidal grating with a simple formula at a global level.

The most commonly used global contrast measure is defined with the Michelson formula

$$C^M = (L_{\max} - L_{\min})/(L_{\max} + L_{\min}),$$

where $L_{\max}$ and $L_{\min}$ stand for the maximum and minimum luminance values, respectively.[21] In a similar fashion, the contrast can also be defined with the Weber fraction

$$C^W = \Delta L/L_b,$$

where $\Delta L = (L_{\max} - L_{\min})/2$ and L_b stands for the luminance of a uniform background around the stimulus.[22] King-Smith and Kulikowski[23] defined contrast as

$$C^K = (L_{\max} - L_a)/L_a,$$

where L_a stands for the average luminance of the visual stimulus, while Burkhardt et al.[24] replaced L_a with the average luminance of the stimulus background. Among these measures, Michelson contrast is the most widely incorporated as a performance reference against others. It is obvious that the measures assume that extreme luminance values dominate the contrast of the whole image.

Pavel et al.[25] proposed a root-mean-square (RMS) measure

$$C^{\mathrm{RMS}} = \sqrt{\sum_{i=1}^{n} (x_i - x')^2/(n-1)},$$

where x_i stands for the normalized luminance value at the ith pixel, x' stands for the mean of x_i, and n stands for the number of pixels.[25] With respect to the formula definition, it is clear that this measure ignores the spatial frequency of image content and spatial distribution of contrast in that image. Pedersen et al.[26] proposed a LAB variance measure

$$C^{\mathrm{LAB}} = \sqrt[3]{\mathrm{std}^2(L) * \mathrm{std}^2(a) * \mathrm{std}^2(b)},$$

where L, a, and b define the coordinate of each pixel in the perceptually uniform CIELAB color space. This measure accounts both luminance and chromatic channels; however, the equal weighting for each channel is inconsistent with the known fact that luminance has stronger impact on the perceived contrast than that of chrominance.[27–29]

For image quality assessment of displays in industry, the international standards TCO Certified Display 6.0[30] and SPWG Notebook Panel Specification 3.8[31] both recommend contrast defined as

$$C^{\mathrm{TCO}} = L_{\max}/L_{\min}.$$

The Information Display Measurements Standard 1.03[32] follows a similar fashion, but further classifies contrast measurements into multiple categories as signal contrast, sequential contrast, starfield contrast, and corner-box contrast by taking the spatial information into account. Part 307 of ISO standard 241 defines contrast as

$$C^{\mathrm{ISO}} = (L_{\max} + L_D + L_S)/(L_{\min} + L_D + L_S),$$

where L_D and L_S stand for the luminance component reflected from diffuse illumination and the luminance component specularly reflected from large aperture sources of illumination, respectively.[33] The contrast definitions in the international standards mentioned above were originally designed to verify the display performance; however, the contrast of actual displayed images is not a part of their concerns.

In summary, the existing global contrast measures are largely inheritances or variations of Michelson contrast to determine the contrast of displays. The contrast definitions above account only the extreme or average luminance values, and they are confined to specific viewing conditions with gray patches or periodic patterns such as sinusoidal gratings; as a result, their use in natural images might be inappropriate.

Local Contrast Measures

For contrast measurement, the local nature of contrast changes across an image and spatial frequency content are related and should be considered together.[14] Local contrast measures divide the images into many subimages, possibly at multiple hierarchical levels, which may be overlapped with each other; the contrast is defined by taking pixel neighborhoods or specific local features into account. Depending on the division granularity, the contrast can be determined at a pixel level with respect to the luminance and/or chrominance information in a certain color space.

Boccignone et al.[34] followed the Weber–Fechner law to replace the subject luminance with the luminance $I(x, y, t)$ of pixel (x, y) at instant t and to replace the background luminance with the average luminance $I_b(x, y, t)$ in the surrounding area of pixel (x, y) at instant t. The instant t is changed by an iterative application of the anisotropic diffusion equation; so the most optimal local contrast for a pixel (x, y) is determined as

$$C^{\mathrm{MWF}}(x, y) = \max_{t\in[t_{\inf}, t_{\sup}]} \ln[I(x, y, t)/I_b(x, y, t)].$$

A. J. Ahumada[35] applied two rounds of low-pass filters F_a and F_b to the input image I and generated two filtered images:

$$I_a(x, y) = I(x, y) * F_a(x, y),$$
$$I_b(x, y) = I_a(x, y) * F_b(x, y).$$

Then the local contrast for each pixel (x, y) is defined as

$$C(x, y) = I_a(x, y)/I_b(x, y) - 1,$$

and eventually the final local contrast is calculated as

$$E(x, y) = C^2(x, y) * F_e(x, y),$$

where F_e is another low-pass filter. Despite the luminance information, the chrominance components in images contribute to the measured contrast as well. Matkovic et al.[36] introduced a global contrast factor (GCF) method to compute the local contrast by averaging the differences between spatially filtered super pixels, and then the global contrast is determined as the mean of local contrast with respect to weighting factors that are estimated based on a psychophysical experiment.

Peli et al.[14] proposed calculating the contrast separately at each pixel of an image to address the variation of contrast across the whole image. In this case, multiple band limited versions of the original image are obtained by applying a radically symmetric band-pass filter in the frequency domain. Then the contrast for each limited band is defined as the ratio between the filtered image and its local luminance mean image. Tadmor and Tolhurst[37] proposed a modified contrast measurement for natural scenes based on the conventional difference of Gaussian (DOG) receptive field model. They proposed a contrast measurement scheme as

$$C^{\mathrm{MDOG}} = [R_c(x, y) - R_s(x, y)]/[R_c(x, y) + R_s(x, y)]$$

at a pixel (x, y) in the image, where

$$R_c(x, y) = \sum_{i=x-3r_c}^{x+3r_c} \sum_{j=y-3r_c}^{y+3r_c} \mathrm{Center}\,(i - x, j - y)$$

and

$$R_s(x, y) = \sum_{i=x-3r_s}^{x+3r_s} \sum_{j=y-3r_s}^{y+3r_s} \mathrm{Surround}\,(i - x, j - y)$$

stand for the center and surrounding components of the receptive field, respectively, with

$$\mathrm{Center}\,(x, y) = \exp[-(x/r_c)^2 - (y/r_c)^2]$$
$$\mathrm{Surround}\,(x, y) = 0.85(r_c/r_s)^2 \exp[-(x/r_s)^2 - (y/r_s)^2],$$

where r_c and r_s stand for the radius of the center and the surroundings of the receptive field, respectively. Eventually, the global contrast is calculated as the mean of the local contrast measurements at many randomized locations in the image.

Rizzi et al.[38] proposed a contrast measure RAMMG that subsamples the input image in order to generate multiple pyramid images in the CIELAB space with a nearest neighborhood algorithm. Then the local contrast is calculated by summing up the absolute differences between one pixel and its surrounding pixels in every channel and at every pyramid level. The local contrast values from the same channel are normalized and finally weighted. The final global contrast is the mean of outputs from all levels:

$$C^{\mathrm{RAMMG}} = \frac{1}{N_L} \sum_{i=1}^{N_L} \sum_{j=1}^{3} \sum_{k=1}^{N_p} W_j C_k,$$

where N_L stands for the number of pyramid levels, N_p stands for number of pixels in each pyramid image, C_k stands for the local contrast for each pixel and its surroundings, and W_j stands for the weighting factor which needs to be determined for each CIELAB channel. Inspired by the RAMMG measure, Simone et al.[16] proposed a measure RSC which employs the DOG formula. They did not merely recombine the mean of averaged local contrast from each pyramid level in the lightness channel, but also in the chromatic channel.

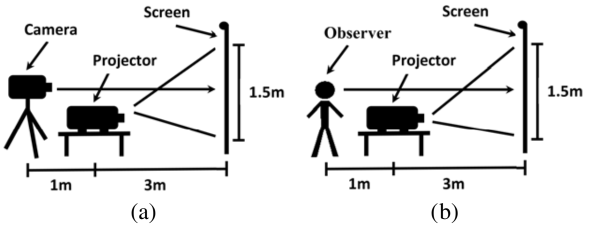
In summary, the existing local contrast measures account for luminance and chrominance components as well as the frequency component in the input images to determine the local contrast at various granularities. The global contrast is eventually determined by pooling local contrast values. In recent years, there has been a general increasing interest in incorporating low-level neuron science knowledge to improve contrast models further.

EXPERIMENTAL SETUP AND PROCEDURE

In order to enhance the validity, reliability, and robustness of our research, we performed the experiments under the same viewing conditions but at two separate geographical locations with two different projection displays and one group of observers at each location. In this case, we had two separate experimental sessions in total.

Experimental Setup

The first experimental session was conducted at the University of Burgundy in France with 10 observers (6 males and 4 females, age from 24 to 33), and we used a portable three-chip LCD projector, a Mitsubishi XL9 (1024 × 768) to display images on the screen. The second experimental session was conducted at the Gjøvik University College in Norway with 17 observers (14 males and 3 females, age from 25 to 53), and we used another three-chip LCD projector, a SONY APL-AW15 (1280 × 768). All observers were confirmed to have neither myopia vision nor color deficiency. Both the Mitsubishi and SONY projectors are three-chip LCD based, and they represent one dominant projector category in the current consumer market. However, the Mitsubishi projector has a more powerful bulb and it appears to be much brighter than the SONY projector in the default settings. Consequently, the Mitsubishi projector suffers from a stronger light leaking problem. The Mitsubishi projector appears to be optimized for document presentation automatically so that the color of the displayed pictures appear to be more bluish. The two projectors are widely used by people for meetings and presentations on a daily basis; their device status is "natural" so that the corresponding perceived contrast is close to what we expect to experience in real practice. Other aspects of the experimental sessions were exactly the same. The projector was placed on a flat table in front of the projection screen at a distance of 3 m away (Figure 1(a)). In our experiments, we were simulating a typical home-theater-like environment. All observers sat in a dark room at an equal distance from the screen. The viewing distance, projection area, and visual angles were all fixed. It is possible for observers to sit closer to or get further away from the screen in practice, but in that case the experimental environment is totally different and the underlying research should be extended to consider additional factors (non-uniform sunlight illumination in a daylight meeting room, for example). In this experiment, since all observers were confirmed to have no myopia or color deficiency difficulty, the visual acuity for them was approximately the same. In the objective experiments, the camera was replaced by the observers (Fig. 1(b)). The principal projection axis is pointed at and is perpendicular to the screen center. On the screen, the projection size was approximately 2 × 1.5 m. The projector was connected to a controlling laptop with a VGA cable. In order to minimize the influence of projector temporally stability, the projector lamp was warmed up at least one hour in advance. All settings related to projector brightness, contrast, and color enhancements were switched off to make sure that the input image was projected as it was. In this case, we assumed that the projection displays were uniform in terms of both their luminance and chromatic nature.



(a) (b)

Figure 1. The experimental setup for both experimental sessions. The projectors were placed at a distance of 3 m away from the screen. The camera and observers were located at the same position, which was about 4 m away from the screen. The experiments were conducted in a home-theater-like environment which was typical for projection systems. (a) The setup for the camera and the projectors. (b) The setup for the observers and the projectors.

For all experiments, we used the same camera to capture the projections. We used a Nikon D610 DSLR camera with imaging resolution of 6016 × 4016 and with a VR 18–100 mm $f/3.5$–5.6G (VR off) lens to capture the images. We set the camera on a tripod and placed the camera approximately at the height of the observers. The pictures were always taken remotely with software installed on the controlling laptop without physically touching the camera. We selected 10 test images (Figure 2) from the Colour Lab Image Database: Image Quality[39] with respect to their image content (800 × 800 in pixels), so we can cover as many features as we may have in the natural images. The features are, for example highlight/lowlight components, wide range/dominant colors, and large smooth/texture areas. We normalized the RGB values of all pixels in the test images and transformed them in each color channel simultaneously with the formula

$$S_i = (C_i - m) * (j + 6)/6 + m,$$

where S_i stands for the scaled RGB value for the ith pixel in the distorted image, j is an integer scaling factor for contrast distortion in the range $[-3, 3]$, C_i stands for the normalized input RGB value for the ith pixel in the input image, and m stands for the mean of all C_i in the same color channel, so we obtained seven distortion levels for each test image (Figure 3). Any overscaled RGB values (either larger than 1 or smaller than 0) were clipped.

Since the main goal of this research was to evaluate the performance of contrast measures, we only needed to produce multiple levels of contrast distortions, and certain perceptual contrast distances were expected between consecutive distortion levels. A linear contrast tuning is sufficient for achieving the goal without introducing brightness differences between the distorted images, so the variances of perceived contrast due to the perceptual adaptation of luminance are minimized. It is possible to tune the contrast with respect to other types of curves like sigmoid curves; however, the tuning is not expected to significantly

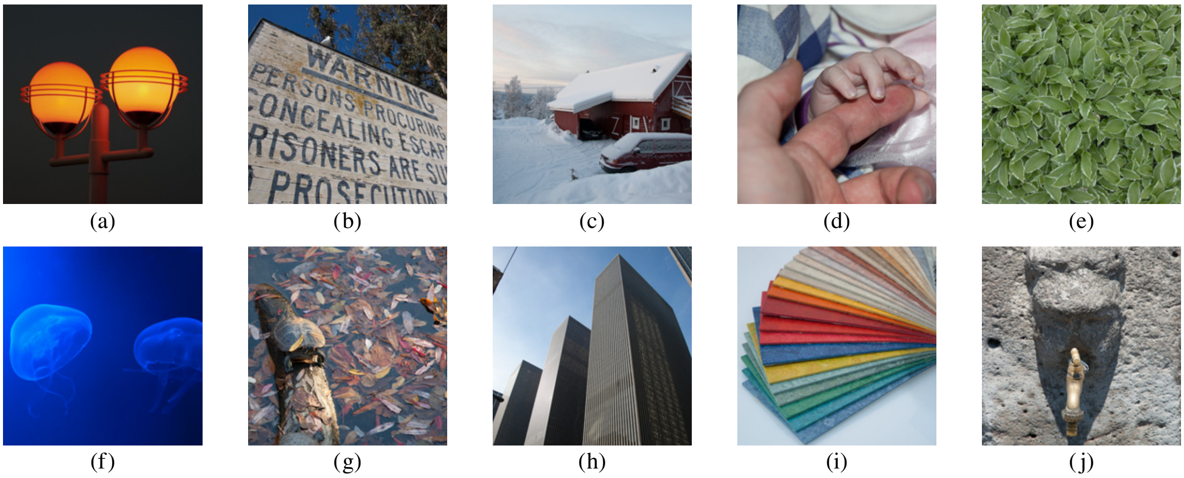


Figure 2. The thumbnails of the 10 test images. We generated 7 levels of contrast distortions of each test image, so there are 70 distorted images in total for each observers to evaluate. The test images were carefully selected to cover many features such as highlight/lowlight components, wide range/dominant colors, and large smooth/texture areas. (a) 1st test image, (b) 2nd test image, (c) 3rd test image, (d) 4th test image, (e) 5th test image, (f) 6th test image, (g) 7th test image, (h) 8th test image, (i) 9th test image, (j) 10th test image.

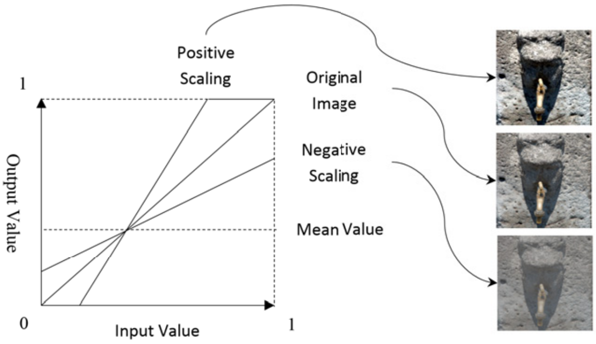


Figure 3. The linear transformation function. Positive scaling represents an enhancement of the actual image contrast, while negative scaling represents a decrement. After the scaling, the mean luminance remains the same. All overscaled values were clipped.

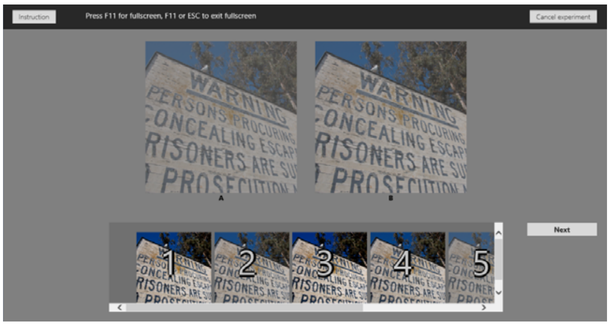

Press F11 for fullscreen, F11 or ESC to exit fullscreen

Figure 4. A screenshot of the software QuickEval, which is an interactive software running on the controlling laptop for psychometric scaling experiments. The software interacts with users within a standard web browser. Each observer is required to rank the displayed images with respect to either perceived or preferred contrast. The images at the bottom are distortion thumbnails, and the two larger windows on the top are used to display images of interest in their original size.

affect the rank order of measured contrast, which is an important aspect of determining correlation coefficients.

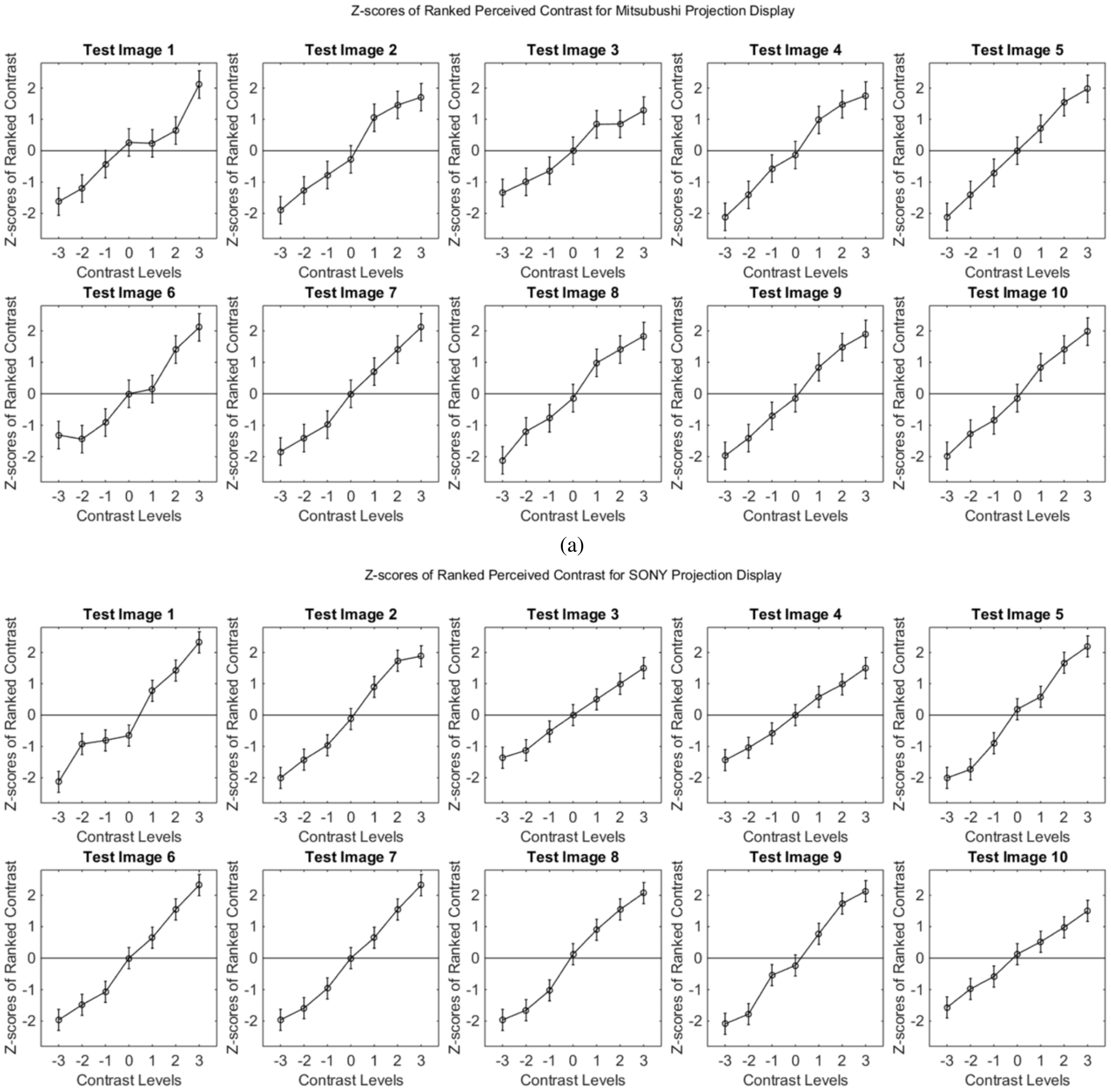
Experimental Procedure

Subjective Experiment

The subjective experiment was conducted by using the software QuickEval,[40] which is an interactive software running on the controlling laptop for psychometric scaling experiments. The software interacts with users within a standard web browser. All observers operate directly on the laptop and they are experiencing exactly the same stimulates in identical viewing conditions. In short, the experiment is performed locally in a controlled manner and it is very different from many typical web-based perceptual experiments.[41,42] Based on this system, each observer is required to perform two tasks. In the first assignment, we display each group of distorted images in randomized order (corresponding to the same input image) on the projection screen at the same time (Figure 4). The observers need to rank this group of distorted images in a descending order based on their perceived contrast; so the images with higher contrast should be ranked to the left, while the rest with lower contrast will be ranked to the right. In the second assignment, we display the images in the same way but we require the observers to rank each group of distorted images in descending order with respect to their own preference of contrast. The images with the preferred contrast should be ranked to the left, while the rest with less preferred contrast should be ranked to the right. The two windows on the top of the software are used to display observers' selected images at their original size. The software automatically records the ranking results and exports them as a rating matrix in the final report.

Objective Experiment

For the objective experiment part, we used a camera as an acquisition device and further processed these captured images with all types of contrast measures. We set the camera



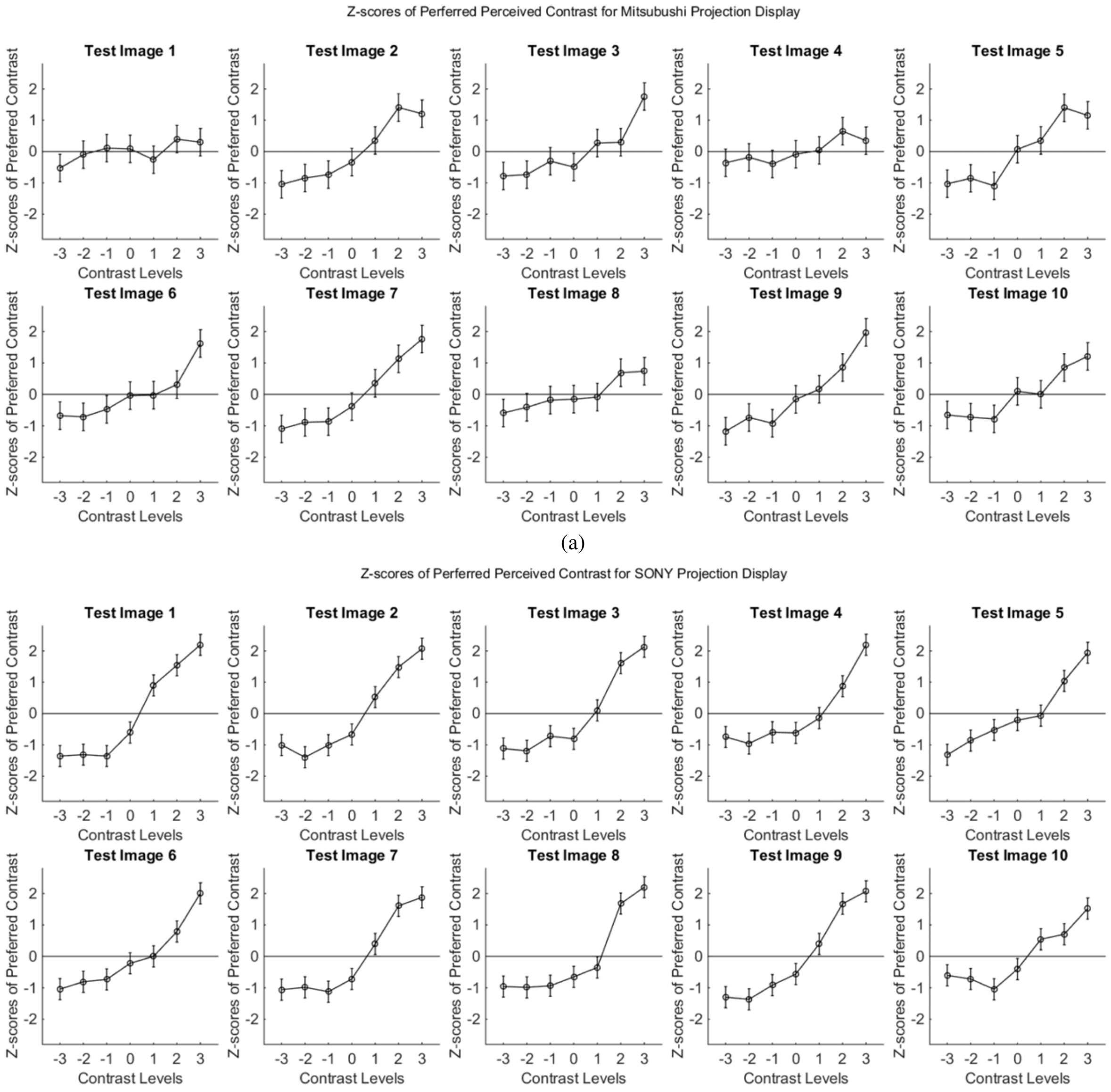
(a)

(b)

Figure 5. The *Z*-scores of ranked perceived contrast. The *X* labels stand for contrast distortion levels in the range [−3, 3]. The number 0 stands for the original image. All plots have identical *Y* value ranges. The circles stand for each *Z*-score of mean ratings for the distorted image, and the vertical bars indicate the 95% confidence interval[46] of *Z*-scores as $1.96 \times (1/\sqrt{N})$, where *N* stands for the number of observers. (a) Mitsubishi XL9 projection display. (b) SONY APL-AW15 projection display.

up with ISO 100 to minimize the camera sensor noise, and performed a standard MTF test[43] in order to acknowledge that the best aperture for the underlying camera and lens was $f/7.1$. We set the shutter speed at a certain value at the initial state and we took several pictures of the peak white projection and observed their histograms. We adjusted the shutter speed setting iteratively to make sure that no camera sensor was either underexposed or overexposed. Since we capture images in raw format, we can apply the spot white balance algorithm to determine the linear scaling factors of the RGB channels respectively. Then we apply these factors to linearly scale all subsequent pictures we take in order to correct the captured luminance. Captured pictures are known to have a vignetting effect, namely an undesirable gradual intensity fall off from the image center to its external limits. We incorporated the method proposed in our previous research[44] to correct camera vignetting based on the captures of a hazy sky which is closely uniform in gray.

In the experiment, we used the following image quality measures to evaluate the image contrast: Michelson contrast,[21] RMS,[25] Lab variance,[26] RAMMG,[38] RSC,[16] and GCF.[36] The Michelson contrast measure was selected because it is representative of global contrast measurement and it is typically used as a reference for contrast measurement in research. RMS and LAB variance measures are selected because they are representative of measurements which



(a)

(b)

Figure 6. The Z-scores of preferred perceived contrast. The X labels stand for contrast distortion levels in the range $[-3, 3]$. The number 0 stands for the original image. All plots have identical Y value ranges. The circles stand for each Z-score of mean ratings for the distorted image, and the vertical bars indicate the 95% confidence interval of Z-scores as $1.96 \times (1/\sqrt{N})$, where N stands for the number of observers. (a) Mitsubishi XL9 projection display. (b) SONY APL-AW15 projection display.

account on statistics; however the RMS measure works only on luminance, while the LAB variance measure further take colors into account in the perceptual uniform CIELAB color space. RAMMG and RSC measures are representative of the measures incorporating low-level visual system models. The GCF measure addresses the problem from the spatial frequency perspective.

EXPERIMENTAL RESULTS

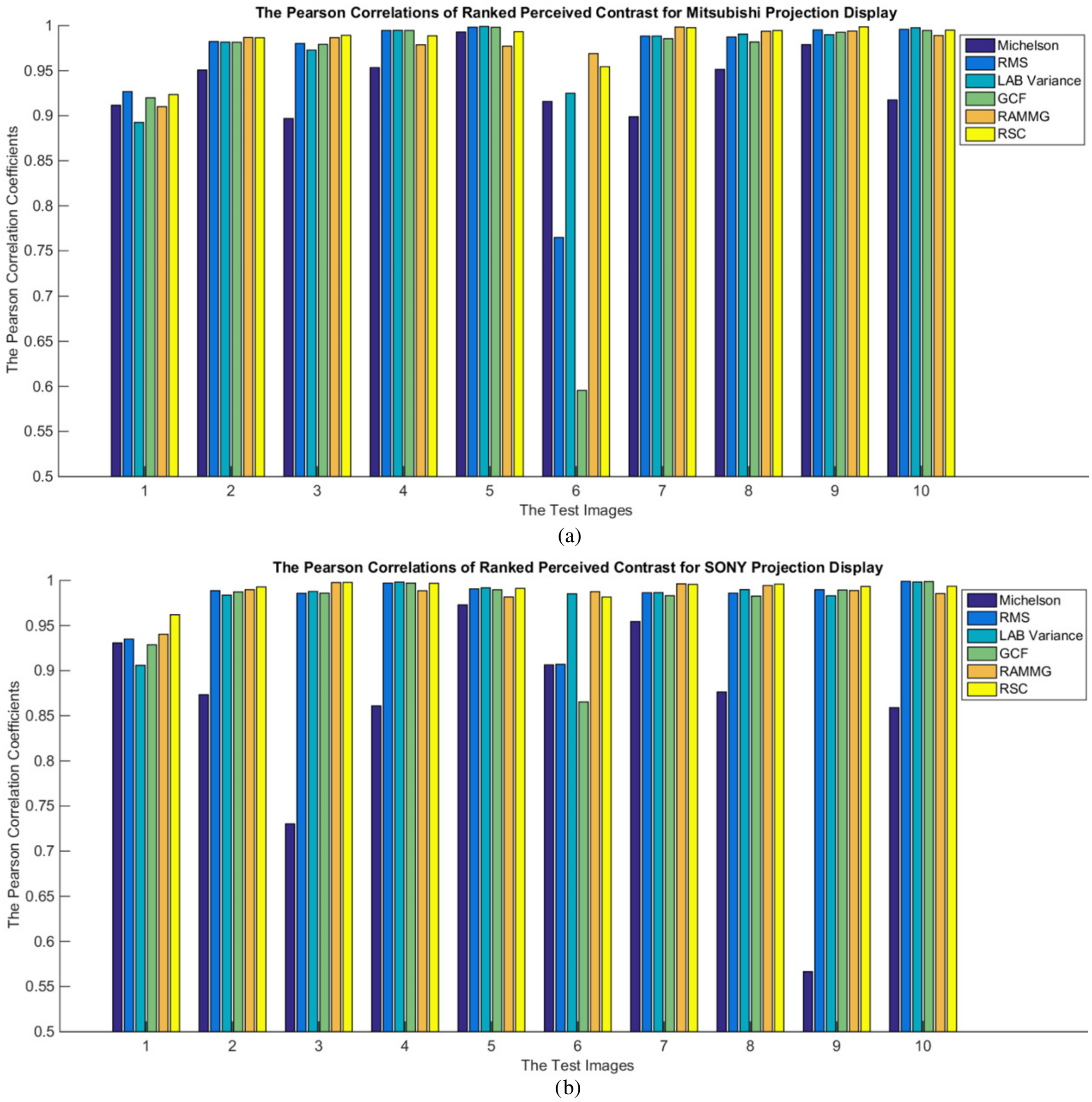
We collected raw subjective ratings, scaled them, and calculated the Z-scores;[4] meanwhile, we processed them with selected image quality measures in order to evaluate the image contrast.

Subjective Results

We collected the subjective ratings for the ranked perceived contrast and preferred perceived contrast, respectively. All collected raw ratings were scaled in order to calculate their Z-scores.

Ranked Perceived Contrast

The Z-scores of ranked perceived contrast for the projectors are shown in Figure 5. It is clear that the rank of perceived contrast has a closely linear relationship with the actual rank of modified contrast. Since the Z-score values in all plots are monotonically increasing, the relationship between perceived contrast and the actual image contrast is almost linear for all types of images.



(a)

(b)

Figure 7. The Pearson correlations between the mean Z-scores of ranked perceived contrast and the measurement scores for the ten selected test images for (a) the Mitsubishi projection display and (b) the SONY projection display. The Y values are limited to between 0.5 and 1.

Preferred Perceived Contrast

The Z-scores of preferred perceived contrast for the projection displays are shown in Figure 6. The general tendency of the Z-scores of preferred contrast no longer follows a linear relationship with the actual image contrast. This observation suggests that the observers tend to rank all distortions into two groups: either relatively less preferred (contrast level -3 to -1) or more preferred perceived contrast (contrast level 1 to 3). In the group of less preferred contrast, since the confidence intervals of Z-scores are largely overlapped, the perceived contrasts have no significant difference, while in the group of more preferred contrast, the confidence intervals are less overlapped. This suggests that the majority of observers prefer the enhanced contrast even though the luminance has been overscaled. In some cases, for both projectors, the contrast level 0, which stands for the original image, is neither preferred nor not preferred because it is very close to the center line for all test images. The preferred perceived contrast values for the two projectors are obviously different.

Objective Results

The evaluations of the objective contrast measures are presented first for the ranked contrast, and then for the preferred perceived contrast for the two projection displays.

Ranked Perceived Contrast

We applied the measures to all modified images to calculate the objective scores, and to determine the Pearson correlation

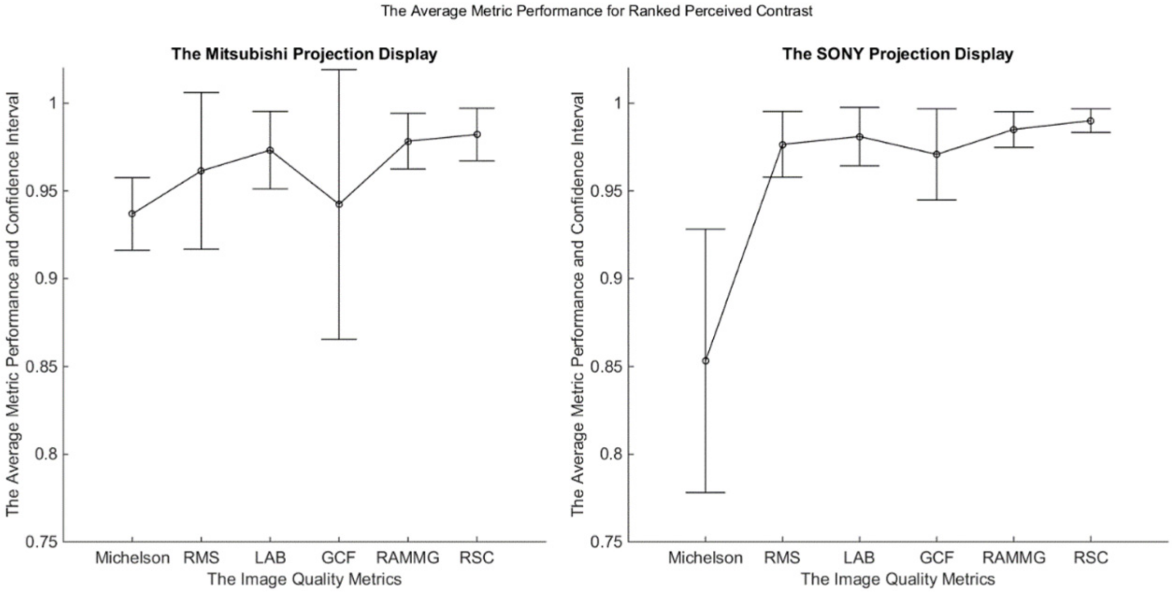


Figure 8. The average performance for ranked perceived contrast with respect to distribution of their Pearson correlation coefficients. The circles indicate the mean of correlation coefficients which are calculated on a per image basis. The bars stand for the 95% confidence interval.

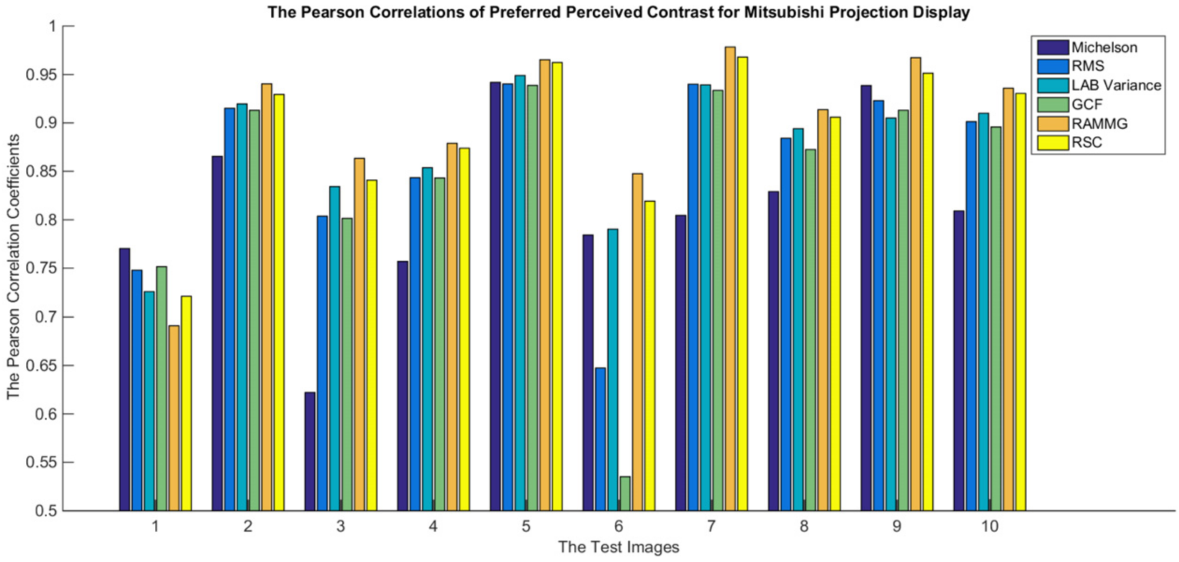
coefficients between the objective scores and the mean Z-scores of both ranked perceived contrast (Figure 7). Based on the observation, it is clear that, for the Mitsubishi projection display, most contrast measures produce high correlation coefficients above 0.85, except that the RMS and GCF measures produce low coefficients on test image 6. However, the observation cannot be obtained from the correlation results for the SONY projector. For the SONY projection display, the Michelson contrast measure performs relatively worse than the other contrast measures, and this is especially true for test image 9. Other contrast measures have very similar performance for both Mitsubishi and SONY projection displays on test images 2, 3, 4, 5, 7, 8, 9, and 10, but not on test images 1 and 6. It is not very clear which measure has the best performance in general. In this case, we generated the box plots of the Pearson correlation coefficients over all test images for both projection displays (Figure 8). It is clear that the Michelson contrast measure performs worse than other contrast measures, not merely because it has a low average correlation value around 0.85, but also its 95% confidence interval is much larger. For the Mitsubishi projection display, the contrast measure GCF performs badly with respect to its confidence interval as well. Although the Mitsubishi and SONY projection displays are supposed to produce different contrast on the screens, the mean of correlation coefficients over all test images follow a very similar general tendency. Based on the observation on the variance of confidence intervals, the RSC contrast measure produces the most stable outcome regardless of the actual image content.

Preferred Perceived Contrast

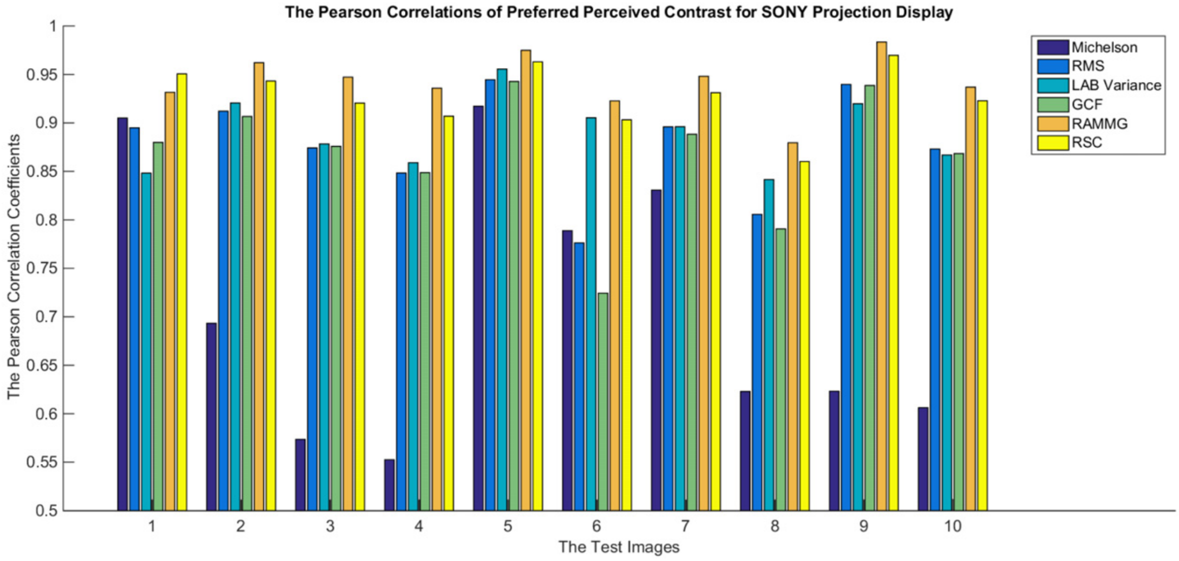
For the preferred perceived contrast, we followed a similar approach to calculate the Pearson correlation coefficients for all contrast measures on all test images (Figure 9). It is clear that the Michelson contrast measure performs the worst for both ranked and preferred contrast. In addition, the RMS and GCF measures both perform relatively worse for test image 6 for the two projection displays as well. For the preferred contrast of both Mitsubishi and SONY projection displays, the RAMMG and RSC still have the highest correlations; however, the correlation from the RAMMG contrast measure is slightly higher than that for the RSC contrast measure. This observation is different from the one for ranked perceived contrast. The rank order between RMS, LAB, GCF, RAMMG, and RSC contrast measures is largely preserved for test images 2, 3, 4, 5, 7, 8, 9 and 10, but not for test images 1 and 6. This observation can be obtained from the ranked perceived contrast for both projection displays as well, but not from the preferred contrast for the SONY projection display. By looking at the average overall contrast measurement performance shown in Figure 10, the general tendency of the average Pearson correlation over all test images is almost the same as the one obtained from the preferred contrast.

Overall Results

In Figs. 8 and 10, we showed the average performance of the contrast measures over all test images for each projection display. In this case, we calculated the Pearson correlation coefficients not on a per image basis but we did the calculation over all test images, so we could observe how the metrics performed regardless of the image content (Figure 11). In Fig. 11 we can see that, for the Mitsubishi projection display, the mean correlation coefficients are almost identical, and the 95% confidence intervals are largely overlapped for both ranked and preferred perceived contrast. This indicates that for the Mitsubishi projection display the contrast measurements have almost the same



(a)



(b)

Figure 9. The Pearson correlations between the mean *Z*-scores of preferred perceived contrast and the measurement scores for the ten selected test images for (a) the Mitsubishi projection display and (b) the SONY projection display.

performance. However, for the SONY projection display, the Michelson contrast measure performs relatively worse. For both projection displays, the ranked and preferred perceived contrasts share a similar general tendency.

We also calculated the Pearson correlations between the average performances over all contrast measurements with respect to their types of contrast versus the types of projection displays. The results are shown in Table I. Considering that the Michelson contrast measure produces low correlation coefficients and large variances for most test images for all projection displays, we removed the Michelson contrast measure and recalculated the data; the results are shown in Table II.

On looking at the data in Table I, it is clear that the average performances of ranked and preferred perceived contrast measurements have high correlations; they are all above 0.9. However, there is no evidence to indicate that there is any good relationship for ranked or preferred contrast between one projection display and another, since their correlation values are all below 0.85. After taking the Michelson contrast measure away, the low coefficients presented in

Table I. The Pearson correlations between the average correlation coefficients.

Contrast/Projector	Ranked, Mit.	Ranked, SONY	Preferred, Mit.	Preferred, SONY
Ranked, Mit.	1	0.7431	0.9629	0.8396
Ranked, SONY	0.7431	1	0.8155	0.9668
Preferred, Mit.	0.9629	0.8155	1	0.9248
Preferred, SONY	0.8396	0.9668	0.9248	1

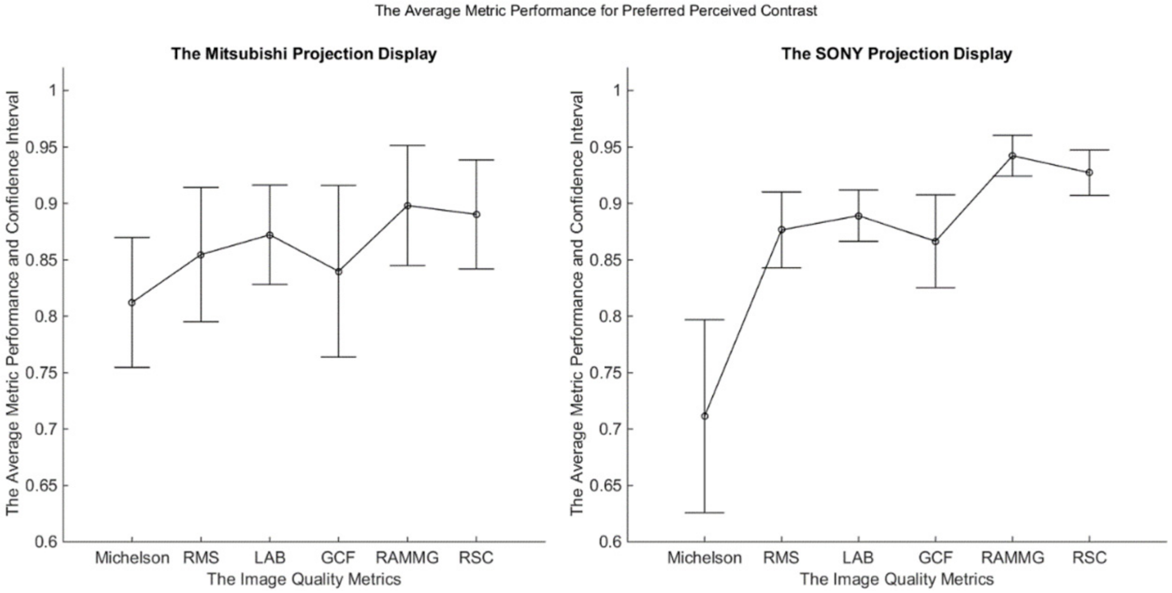


Figure 10. The average measurement performance for ranked perceived contrast with respect to the distribution of their Pearson correlation coefficients. The circles indicate the mean of the correlation coefficients which are calculated on a per image basis. The bars stand for the 95% confidence interval.

Table II. The Pearson correlations between the average correlation coefficients without the Michelson contrast measure.

Contrast/Projector	Ranked, Mit.	Ranked, SONY	Preferred, Mit.	Preferred, SONY
Ranked, Mit.	1	0.9608	0.9409	0.8424
Ranked, SONY	0.9608	1	0.9354	0.8829
Preferred, Mit.	0.9409	0.9354	1	0.9696
Preferred, SONY	0.8424	0.8829	0.9696	1

Table I increase by a certain amount, and their values are all above 0.9 in Table II. However, the correlation coefficients of both ranked and preferred contrast between different projection displays show no significant improvement. This observation suggests that human preference on the perceived contrast has a closely linear relationship with the ranked perceived contrast. In this circumstance, we conclude that the most preferred perceived contrast corresponds to the highest ranked perceived contrast; even for test images 1, 2, 4, and 5 in Fig. 6(a) the highest preferred perceived contrast corresponds to the second highest ranked perceived contrast. For related research in the future it is unnecessary to explicitly distinguish them and do the experiments twice.

CONCLUSION AND FUTURE WORKS

In this paper, we have proposed an objective approach to measure the relative contrast of projection displays based on pictures taken with a calibrated digital camera in a controlled environment. To the best knowledge we have, this is the first research regarding evaluating the perceived contrast on projection displays based on the images captured with a calibrated camera. This objective approach can be easily extended to measure other image quality attributes such as sharpness and non-uniformity for all types of information displays. The metric performance evaluation is based on two separate projection displays, so the validity and reproducibility of the research have been enhanced. The research feasibility is supported by the high correlations between subjective and objective experimental results, as well as the correlations between the two projection displays. We classified the contrast measures into local and global categories. For each category, we selected the representative contrast measures and evaluated their performance with respect to the Pearson correlations between subjective and objective assessment results. The experimental results based on two separate projection displays suggest that the Michelson contrast measure has very low performance over all test images, as expected. Other global contrast measures (RMS and LAB) also perform relatively better than the Michelson contrast measure, but they have less correlation with the perceptual ratings compared to the local contrast measures. The local contrast measure GCF has similar performance to the RMS and LAB measures, but it performs worse than other local contrast measures (RAMMG and RSC). The contrast measures RAMMG and RSC perform the best overall, and they have very close performance on contrast measurements for almost all test images. With respect to the 95% confidence interval of the average measurement performance over all test images, RAMMG has slightly improved correlations with the preferred contrast. It is interesting to see that many contrast measures do not perform well on the test images 1 and 6. These two images either have large area of low luminance component or dominant color component, and they do not have obvious texture area. We recommend local contrast measures incorporating low-level human visual system models since they have better overall

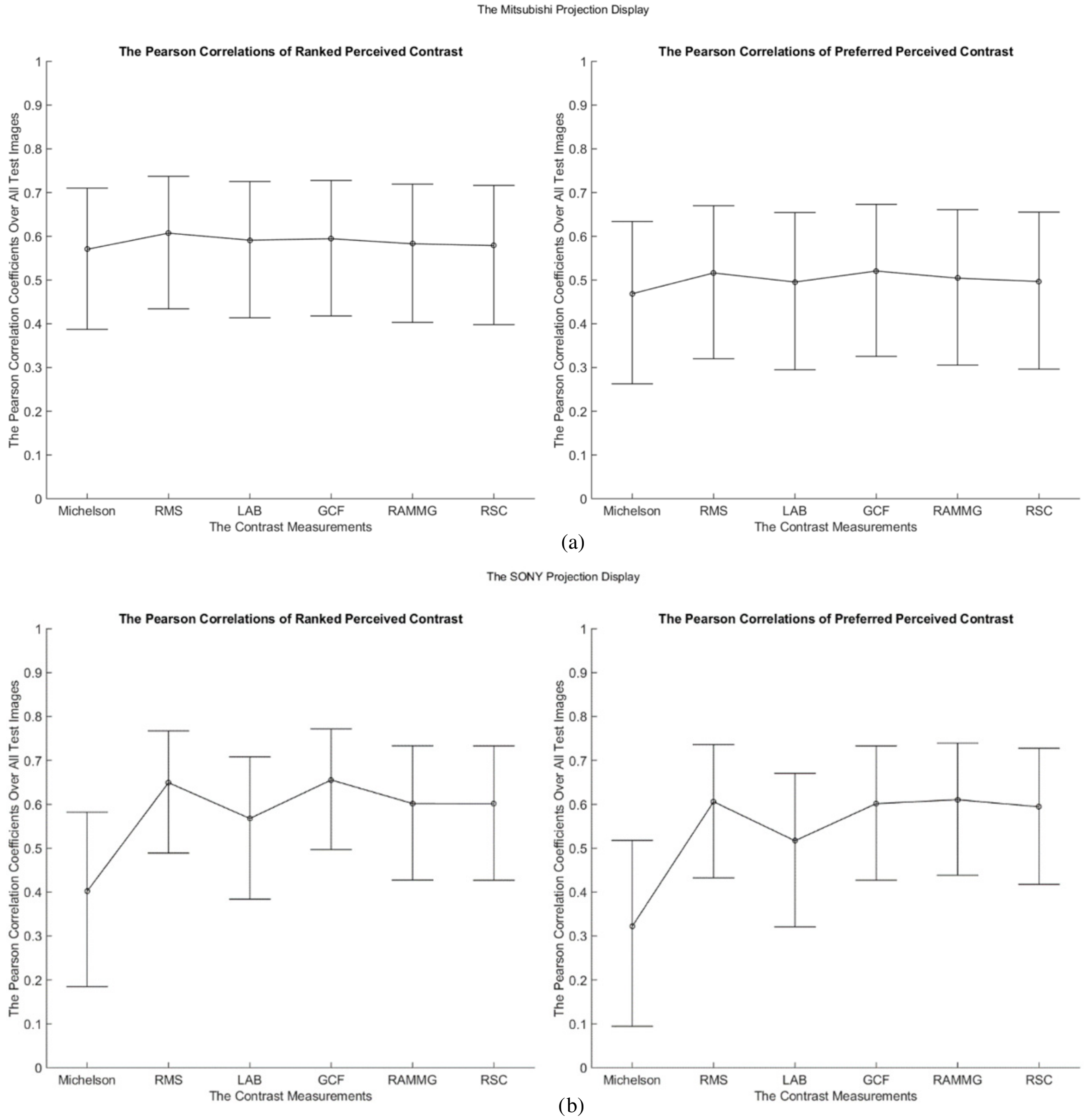


Figure 11. The average measurement performance for ranked perceived contrast with respect to the distribution of their Pearson correlation coefficients over all test images. The circles indicate the mean of correlation coefficients which are calculated on a per image basis. The bars stand for the 95% confidence interval calculated based on Fisher Z transformation.[47] (a) The Mitsubishi projection display. (b) The SONY projection display.

performance over global contrast measures in terms of both contrast prediction accuracy and stability regardless of the image content. Since the average correlations and stability of local contrast measures are good for many test images, we do not need to propose a new contrast measure, but rather to improve the models of human visual system to predict the image contrast better in future research.

REFERENCES

[1] O. Bimber, A. Emmerling, and T. Klemmer, "Embedded entertainment with smart projectors," IEEE Comput. **38**, 48–55 ACM Press, New York, USA (2005).

[2] W. Zou and H. Xu, "Colorimetric color reproduction framework for screen relaxation of projection display," Displays **32**, 313–319 (2011).

[3] O. Bimber and E. Andreas, "Multifocal projection: a multiprojector technique for increasing focal depth," IEEE Trans. Vis. Comput. Graphics **12**, 658–667 (2006).

[4] B. Sajadi, M. Lazarov, A. Majumder, and M. Gopi, "Color seamlessness in multi-projector displays using constrained gamut morphing," IEEE Trans. Vis. Comput. Graphics **15**, 1317–1325 (2009).

[5] M. Pedersen, N. Bonnier, J. Y. Hardeberg, and F. Albregtsen, "Attributes of image quality for color prints," J. Electron. Imaging **19**, 011016 (2010).

[6] J. Gille, L. Arend, and J. Larimer, "Display characterization by eye: contrast ratio and discrimination throughout the grayscale," Proc. SPIE **5292**, 218–233 (2004).

[7] D. H. Brainard, D. G. Pelli, and T. Robson, "Display characterization," *Imaging Science and Technology* (Wiley, 2002), pp. 172–188.

[8] M. Fairchild and D. R. Wyble, *Colorimetric Characterization of The Apple Studio Display (Flat panel LCD)* (New York, 1998).

[9] A. M. Bakke, J.-B. Thomas, and J. Gerhardt, "Common assumptions in color characterization of projectors," *Gjøvik Color Imaging Symposium* (Gjøvik, 2009), pp. 45–53.

[10] A. Majumder, "Contrast enhancement of multi-displays using human contrast sensitivity," *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition 2* (Providence, RI, 2005), pp. 377–382.

[11] J.-B. Thomas, *Colorimetric Characterization of Displays and Multi-display Systems* (Université de Bourgogne, 2009).

[12] A. Majumder, "Modeling color properties of tiled displays," Comput. Graph. Forum **24**, 149–163 (2005).

[13] J. Y. Hardeberg, I. Farup, and G. Stjernvang, "Color quality analysis of a system for digital distribution and projection of cinema commercials," SMPTE Motion Imaging **114**, 146–151 (2005).

[14] E. Peli, "Contrast in complex images," J. Opt. Soc. Am. A **7**, 2032–2040 (1990).

[15] M. Pedersen, N. Bonnier, J. Y. Hardeberg, and F. Albregtsen, "Attributes of a new image quality model for color prints," *Proc. IS&T/SID Seventeenth Color and Imaging Conf.* (IS&T, Springfield, VA, 2009), pp. 204–209.

[16] G. Simone, M. Pedersen, and J. Y. Hardeberg, "Measuring perceptual contrast in digital images," J. Vis. Commun. Image Represent. **23**, 491–506 (2012).

[17] A. Majumder and R. Stevens, "Color nonuniformity in projection-based displays: analysis and solutions," IEEE Trans. Vis. Comput. Graphics **10**, 177–188 (2003).

[18] M. Brown, A. Majumder, and R. Yang, "Camera-based calibration techniques for seamless multiprojector displays," IEEE Trans. Vis. Comput. Graphics **11**, 193–206 (2005).

[19] J.-B. Thomas, "Webcam based display calibration," *Proc. IS&T/SID Twentieth Color and Imaging Conf.* (IS&T, Springfield, VA, 2012), pp. 82–87.

[20] G. Simone, M. Pedersen, J. Y. Hardeberg, and A. Rizzi, "Measuring perceptual contrast in a multi-level framework," Proc. SPIE **7240**, 72400Q (2009).

[21] M. A. Abraham, *Studies in Optics* (Dover, 1927).

[22] P. Whittle, "The psychophysics of contrast brightness," *Lightness, Brightness, and Transparency* (Lawrence Erlbaum Associates, New Jersey, USA, 1994), pp. 35–110.

[23] P. E. King-Smith and J. J. Kulikowski, "Pattern and flicker detection analysed by subthreshold summation," J. Physiol. 519–548 (1975).

[24] D. A. Burkhardt, J. Gottesman, D. Kersten, and G. E. Legge, "Symmetry and constancy in the perception of negative and positive luminance contrast," J. Opt. Soc. Am. A **1**, 309–316 (1984).

[25] M. Pavel, G. Sperling, T. Riedl, and A. Vanderbeek, "Limits of visual communication: the effect of signal-to-noise ratio on the intelligibility of American sign language," J. Opt. Soc. Am. **4**, 2355–2365 (1987).

[26] M. Pedersen, A. Rizzi, J. Y. Hardeberg, and G. Simone, "Evaluation of contrast measures in relation to observers perceived contrast," *Proc. IS&T CGIV2008: Fourth European Conf. on Colour in Graphics, Imaging and Vision* (IS&T, Springfield, VA, 2008), pp. 253–256.

[27] E. B. Goldstein, *Sensation and Perception*, 8th ed. (Cengage Learning, Belmont, CA, USA, 2009).

[28] R. L. DeValois and K. K. DeValois, *Spatial Vision* (Oxford University Press, 1988).

[29] M. D. Fairchild, *Color Appearance Models*, 2nd ed. (John Wiley & Sons, Ltd, 2005).

[30] "TCO certified displays 6.0," pp. 1–125, TCO Development AB, Stockholm, Sweden (2012).

[31] "SPWG notebook panel specification 3.8," pp. 1–59, Standard Panel Working Group, Newark, USA (2007).

[32] "Information display measurements standard 1.03," Campbell, CA, USA (2012).

[33] "ISO9241 Ergonomics of human-system interaction - Part 307: Analysis and Compliance Test Methods for Electronic Visual Displays," pp. 1–217, International Standard Organization (2008).

[34] G. Boccignone, M. Ferrearo, and T. Caelli, "Encoding visual information using anisotropic transformations," IEEE Trans. Pattern Anal. Mach. Intell. **23**, 1–16 (2001).

[35] A. J. Ahumada Jr., "Simplified vision models for image quality assessment," *SID Int. Symp. Dig. Tech. Pap. 27* (1996), pp. 397–402.

[36] K. Matkovic, L. Neumann, A. Neumann, T. Psik, and W. Purgatholer, "Global contrast factor – a new approach to image contrast," *Eurographics Workshop on Computational Aesthetics in Graphics, Visualization and Imaging – Computational Aesthetics*, edited by L. Neumann, M. S. Casasayas, B. Gooch and W. Purgathofer (Girona, Spain, 2005), pp. 159–167.

[37] Y. Tadmor and D. J. Tolhurst, "Calculating the contrasts that retinal ganglion cells and LGN neurones encounter in natural scenes," Vision Res. **40**, 3145–3157 (2000).

[38] A. Rizzi, T. Algeri, G. Medeghini, and D. Marini, "A proposal for contrast measure in digital images," *Proc. IS&T CGIV2004: Second European Conf. on Colour in Graphics, Imaging, and Vision* (IS&T, Springfield, VA, 2004), pp. 187–192.

[39] X. Liu, M. Pedersen, and J. Y. Hardeberg, "CID:IQ – a new image quaity database," *Int'l Conf. on Image and Signal Processing 8509*, edited by A. Elmoataz, O. Lezoray, F. Nouboud and D. Mammass (Springer International Publishing, Cherbourg, Normandy, France, 2014), pp. 193–202.

[40] K. Van Ngo, J. Storvik Jr., C. Andre Dokkeberg, I. Farup, and M. Pedersen, "QuickEval: a web application for psychometric scaling experiments," Proc. SPIE **9396**, 93960Q (2015).

[41] I. Sprow, Z. Baranczuk, T. Stamm, and P. Zolliker, "Web-based psychometric evaluation of image quality," Proc. SPIE **7242**, 72420A (2009).

[42] C. Keimel, J. Habigt, C. Horch, and K. Diepold, "Qualitycrowd – a framework for crowd-based quality evaluation," *Picture Coding Symposium* (IEEE, Krakow, Poland, 2012), pp. 245–238.

[43] "ISO 12233:2014 photography – electronic still picture imaging – resolution and spatial frequency responses" (2014).

[44] P. Zhao, M. Pedersen, J.-B. Thomas, and J. Y. Hardeberg, "Perceptual spatial uniformity assessment of projection displays with a calibrated camera," *Proc. IS&T Twenty-second Color and Imaging Conf.* (IS&T, Springfield, VA, 2015), pp. 159–164.

[45] P. G. Engeldrum, *Psychometric Scaling: A Toolkit for Imaging Systems Development* (Imcotek Pr, 2000).

[46] J. Morovic, *To Develop a Universal Gamut Mapping Algorithm* (University of Derby, 1998).

[47] The Video Quality Experts Group, "Final Report from The Video Quality Experts Group on The Evaluation of Objective Models of Video Quality Assessment—Full Reference Television Phase I" (2000).

Chapter 10

Paper E

Measuring Perceived Sharpness of Projection Displays with a Calibrated Camera

Ping Zhao, and Marius Pedersen

Submitted to

Journal of Visual Communication and Image Representation (JVCIR), Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Journal of Visual Communication and Image Representation 00 (2015) 1–12

JVCIR

Measuring Perceived Sharpness of Projection Displays with A Calibrated Camera

Ping Zhao, Marius Pedersen

Gjøvik University College, Teknologivn. 22, 2821 Gjøvik, Norway

Abstract

Perceived sharpness is one of the most important image quality attributes for displays, because it determines how much details the humans are able to perceive on the screen at certain distances. However, this attribute was not well studied for projection displays in the existing literature. In this paper, we conduct an experimental study on measuring perceived sharpness of projection displays based on the pictures taken with a calibrated camera in a darkroom, and evaluating the performance of state-of-art sharpness metrics accordingly. The basic idea is to apply Gaussian filtering to natural test images in order to simulate the optical blurring process of projection systems, so that we can generate multiple levels of image sharpness in a controlled manner without influencing the original properties of projection displays. We project these filtered images onto the screen and invite a group of human observers to give perceptual ratings on them. We calculate the correlation coefficients between perceptual sharpness and the one measured with state-of-art image quality metrics. We find out that the average performance of full reference metrics are comparatively better than the reduced and no-reference metrics. Among the full reference metrics, SSIM, VIF and FSMI metrics perform well in terms of both accuracy and stability.

Keywords: sharpness, perception, image quality, projection display, image capture, camera calibration

1. Introduction

Nowadays, modern digital imaging with advanced media technologies composes an essential part of our daily life and works. It is easy for ordinary consumers to capture what they see and record what they experience with portable imaging devices even without professional training. Sharing stories with friends either in a face-to-face manner or over the networks can be achieved by simply clicking a few buttons. One common way to do this is via projection systems, which are typically configured with high brightness light sources and high definition displays to visualize image reproductions. Comparing to other types of display technologies, projection displays have many unique advantages like portability, flexibility for deployment, and large projection area for sharing information to a crowd of people. In some scenarios, it is required to tile up multiple projections in order to produce a large perceptual seamless images which visualize information to a crowd of target audience for immersive visual experience [1, 2]. In recent years, there is an increasing general interest to embed projection systems into portable devices to further enhance the continuity experiences between mobile imaging devices and socialization over the networks [3, 4]. Hence, display image quality

Email address: ping.zhao@hig.no (Ping Zhao, Marius Pedersen)
URL: http://www.colourlab.no (Ping Zhao, Marius Pedersen)

assessment of projection systems becomes an interesting topic for both scientific research and industrial commercial communities; because the underlying concept does not merely associate with a well defined systematic approach to measure the quality of image reproductions with respect to human perceptions, but take real industrial engineering practices into consideration as well.

Image quality is largely evaluated with respect to perceptual assessment results. The ultimate goal is to correlate the objective results with the subjective results, so that we are able to eventually eliminate the demand of human observers. In this context, image quality can be interpreted based on image quality attributes which are essentially terms of perception, such as, but not limited to, lightness, contrast, color accuracy, sharpness, artifacts (including noises), and physical properties of displays (screen dimension, display resolution and refreshing rate, etc.). Although there are many attributes can be used to depict image quality, for the researches in various domains the selection of the most important attributes has different priorities. For printing, Pedersen et al. [5] suggested that all the image quality attributes mentioned above are important; Johnson [6] specially remarked color accuracy, sharpness, and contrast for printing. For stereo displays, You et al. [7] and Lehtimaki et al. [8] pointed out noise, sharpness and perceived depth are priorities for stereoscopic imaging. However, specific for projection displays, limited works have been done so far. Thomas et al. [9] and Strand et al. [10] remarked lightness and color accuracy, while Majumder et al. [11, 12] indicated that lightness is more important than the color accuracy. With respect to the literature above, it is clear that despite of specific research domains and imaging technologies involved, sharpness is commonly recognized to be an important image quality attribute for perceptual evaluation, and it is closely associated with other attributes like lightness and contrast. Since sharpness defines the amount of details the human can observe in image reproductions at certain distances, it is commonly referred as the counterpart of blur which is another typical image quality distortions. Human Visual System (HVS) has a remarkable capability to detect image blur without seeing the original image, but unfortunately the underlying mechanisms are not well understood [13, 14].

In this paper, we conduct an experimental study on measuring perceived sharpness of projection displays based on the pictures taken with a calibrated camera in a controlled environment. The goal is to evaluate the performance of state-of-art image quality metrics for image sharpness with respect to their correlations with perceived contrast obtained from psychophysical experiments.

The rest of this paper is organized as follows. First, in Section 2, we present the state-of-art image quality metrics for image sharpness. Then, in Section 3, we present our methods for calibrating the camera, introduce the experimental environment. In Section 4, we demonstrate the subjective and objective results. Last, in Section 5, the conclusions and future works are presented.

2. Sharpness Metrics

2.1. Overview

Conventionally, in an objective manner, sharpness was largely measured with respect to the definition of edges in the image reproductions. The main idea is to detect edges in local regions, then compute the sharpness quality scores in these regions at the detected edges, and eventually pool these scores to generate a number representing the global sharpness quality. The edge information can be extracted based on Kurtosis [15, 16, 17], derivatives [18, 19], edge-width [20, 21], histogram [22], power spectrum [23], and wavelet [24] etc. In the cases that the edges features are largely visible, these features represent the quality of optical components in such an imaging system, so the measured edge responses can be used as an estimate of the modulation transfer function [25]. In the ISO 12233 standard, the modulation transfer function is determined with respect to the spatial frequency response of slanted edges [26]. However, it is insufficient to measure the sharpness by only focusing on a few strong edges, because images with very sharp edges may have only a small amount of details [27].

The clarity of fine details or textures is another important factor of sharpness measurement. The details based methods are slightly more advanced than the edge based methods, because they can be used to measure highly degraded images from where the edge information are difficult to be extracted. In the existing literature, the amount of details are related to the perceived contrast in the regions of interest; so, recently, there is an increasing general interest on developing perceptual models to simulate HVS. For example, Gao et al. [28] proposed a perceptual contrast model by introducing Weber's law into an isotropic local contrast model in order to account in human luminance masking effect, and then the sharpness is defined as an average contrast measured by the model in the regions of interest. Nuutinen et al. [29] determined sharpness based on the local energies which were calculated as the standard deviation of the wavelet coefficients from the correspondence blocks detected between the image reproductions and their corresponding reference images. They indicated that smooth regions cannot be used for sharpness measurements since their structure energy remain unaltered with low-pass operations. With respect to the existing literature, just like contrast, the definition of detail is fuzzy. In many cases, the researchers interpret sharpness in term of detail, but what they actually refer to is edge; so, sharpness measurement is somehow context dependent. It is known that, the image quality metrics can be generally classified into Full Reference (FR) based, Reduced Reference (RR) based, and No Reference (NR) based methods [30]. The existing studies regarding sharpness metrics mainly focus on one specific category of them, and the comparison between them is missing. In this paper, we propose to classify sharpness metrics based on how they refer to the original test images, and evaluate their performance after. The main focus is on the state-of-art sharpness metrics in each category.

2.2. Full Reference Metrics

The metric ”Structure Similarity Index Metric” (SSIM) [31] is commonly used to predict the degradation of structures in image reproductions, as well as in many performance benchmarks against sharpness metrics. SSIM was not specifically designed to measure sharpness, but it accounts in brightness, contrast and structure information to estimate the image quality. These factors are widely accepted to have great influence on the perceptual sharpness. Marziliano et al. [21] proposed two FR metrics to measure the sharpness of JPEG2000 compressed images. The proposed metrics measure the magnitude of image blurring which is an image quality attribute contrary to sharpness. Zhang et al. [32] proposed a metric ”Feature Similarity based Index Metric” (FSIM) to measure the overall quality of images. Firstly, they generated a local image quality map with phase congruence and image gradient magnitude as features, and then utilized the phase congruence information again as a weighting function to derive the final image quality score. The metric was not originally designed to measure sharpness, but the results might correlate with sharpness. Another commonly referred image quality metric ”Visual Information Fidelity” (VIF) proposed by Sheikh et al. [33]. This metric was derived from a statistical model for natural scenes, a model for image distortions, and a HVS model in an information-theoretic setting. The image quality metric ”Visual-Signal-to-Noise-Ratio” (VSNR) presented by Chandler et al. [34] takes advantage of low-level HVS properties on contrast sensitivity and visual masking via a wavelet-based model to determine if the distortions are below the threshold of visual detection. If the distortions are supra-threshold, the low-level HVS property of perceived contrast and the mid-level HVS property of global precedence are accounted as an alternative measure of structural degradation. These factors are known to have influence on measured sharpness.

2.3. Reduced Reference Metrics

Wang et al. [35] proposed a RR metric ”Reduced Reference Image Quality Assessment” (RRIQA) to decompose image reproductions into multiple scales and orientations with respect to a steerable pyramid framework. The wavelet coefficient histograms obtained from the reference image are fitted into a generalized Gaussian density model, while the histograms are determined from the image reproduction as well. The Kullback-Leibler distance between the probability distribution of the two sets of wavelet coefficients is adopted as a image quality predictor. Nuutinen et al. [29] measured sharpness as the average difference in standard deviation of the wavelet coefficients among the top rank correspondence blocks within the reference image and its reproductions. The blocks are detected by SIFT algorithm which is believed to be robust. The main author also also proposed a similar wavelet based RR metric to measure sharpness of digital printed natural images, but focusing only on the middle frequency energy in order to avoid high frequency noises in the image reproductions [36]. In a similar fashion regarding the wavelet transformations and sharpness determination, Cheng et al. [37] takes advantage Laplace distribution to determine the magnitude of gradient in images, while Xue et al. [38] uses Weibull distribution. From the entropy information perspective, Soundararajan et al. [39] proposed a method ”Reduced Reference Entropic Differencing” (RRED) to measure the difference between the entropy of wavelet coefficients of the reference image and its reproduction version. Although it was not originally designed for sharpness measurement, but it was reported to correlate with perceptual image quality well. Nevertheless, RR metrics largely rely on wavelet transformation, and they either takes advantage of edge or detail as features to determine the image distortions.

2.4. No Reference Metrics

Caviedes et al. [16] developed a content independent NR sharpness metric based on the local frequency spectrum around the image edges, however this method has problems to predict sharpness quality when the artifacts become dominant. Maalouf et al. [40] defined a sharpness metric based on the eigenvalues of the wavelet-based multi-scale structure tensor to accumulate multi-scale gradient information of local regions. The structure tensor has the advantage to identify edges in spite of the presence of noises, so the metric is suitable to be used to measure the sharpness of color edges. Cao et al. [41] introduced a sharpness metric which takes the advantage of anisotropic diffusion to build up a preliminary map of ringing artifacts and refined it by considering the property of ringing structure. Samira et al. [42] proposed a method to measure color differences to determine the sharpness in local regions. This method is good in the cases of which the color management is critical to the applications. Vu et al. [43] presented a block-based metric ”Spectral, Spatial, Sharpness” (S3) to quantify the local perceived sharpness within and across images. Both spectral and spatial properties of images are utilized to build up indexes for the standard deviation of the impulse response used in Gaussian blurring.

Hassen et al. [14] developed a metric ”Local Phase Coherence based Sharpness Index” (LPC-SI) to identify sharpness as strong local phase coherence in the complex wavelet transform domain. They incorporated this metric into a framework that allows for computation of local phrase coherence in arbitrary fractional scales. Leclaire et al. [44] introduced a metric ”Sharpness Index” (SIndex) which can be used to measure the sharpness in a probabilistic scene the surprisingly small variation of an image compared to that of certain associated random-phase fields. Narvekar et al. [45] presented an improved no-reference metric based on ”Cumulative Probability of Blur Detection” (CPBD). This metric splits the image reproduction into several regions, and for each region a distinct quality class or qualitative score is assigned, then a training base method was proposed to determine the centroid of the quality classes for the assigned scores, and finally the index of image quality class is assigned as the measured image quality. Narvekar et al. [46] proposed a NR metric based on a cumulative probability of blur detection. Comparing to the saliency-weighted foveal pooling based measure developed by Sadaka et al. [47], their metric require no additional visual attention or saliency maps. In the former case, the computational complexity can be largely reduced. Besides, Ferzli et al. [48] derived from the measured just-noticeable blurs to develop a perceptual-based sharpness metric which is applied to 8x8 blocks instead of the entire test image. The metric account in the response of the HVS to sharpness at different contrast levels. Wang et al. [49] proposed a metric to predict wavelet coefficients of local phase coherence structures across scale and space in a coarse-to-fine manner. Another no-reference metric sharpness metric ”Just Noticeable Blur Metric” (JNBM) proposed by Ferzli et al. [50]. It integrated the concept of just noticeable blur into a probability summation model. The metric was reported to be able to predict the relative amount of blurriness in images with different content.

3. Experimental Setup and Procedure

In our experiment, we use a calibrated camera to capture all pixels on the projection screen in one shot, and the performance of selected state-of-art sharpness metrics are evaluated with respect to the perceptual ratings collected from human observers. The experiments take place in a controlled lab environment where it is totally dark. We use a portable three chip LCD projector SONY APL-AW15 to produce projections on a planar screen which is naturally hanging on the ceiling. The projector is put on a table placed in front of the projection screen about 3m away with respect to the throw ratio (1.5) of the projector. A remote controlling laptop is connected to the projector via a HDMI cable in order to generate full screen projections which have resolution 1920 × 1080 in pixels. On the screen, the dimension of projection area is approximately 2 × 1.2 in meters. We use a DLSR Nikon D610 which has an imaging resolution 6048 × 4016 in pixels and with a Sigma VR 24-105mmf/4G (VR off) lens to capture the projections. The camera is fixed on a tripod which is placed right in front of the projections about 4m away. Pictures are taken remotely with a software control on the laptop without physically touching the camera. The pictures are saved in raw format and rendered with aliasing minimization and zipper elimination demosaicing algorithm [51] without automatic vignetting correction, brightness adjustment, gamma correction and noise reduction etc. We select 7 test images (see Figure 1) from the CID database [52] to generate 6 levels of Gaussian blur with kernel size 11 and standard deviation 0, 0.5, 1, 1.5, 2, and 3 respectively. The selection criteria of test images is established based on the coverage of different image features such as hue, saturation, lightness, contrast, skin colors, sky colors, grass colors, size of neutral gray areas, color transition, fine details and text presence etc.

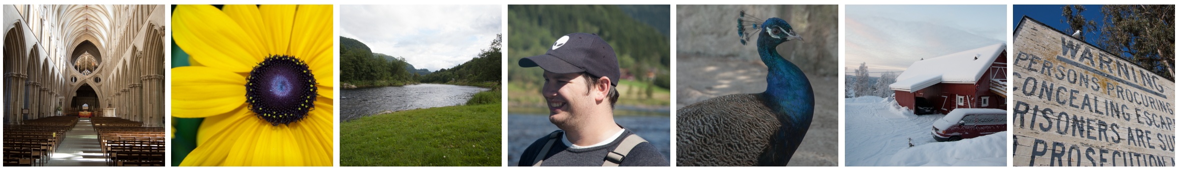


Figure 1. The selected 7 test images selected from the CID test image database, each of them is incorporated to generate 6 levels of Gaussian blurred images with fixed kernel size 11 and standard deviation 0, 0.5, 1, 1.5, 2, and 3 respectively. The selection criteria of the test images is established based on the coverage of image features such as hue, saturation, lightness, contrast, skin colors, sky colors, grass colors, size of neutral gray areas, color transition, fine details and text presence etc. The main purpose is to generalize the image features, so there will be no specific specific dominant.

A digital camera needs to be calibrated in advance to make sure that the pictures taken from this camera will not corrupted due to optical and electronic issues of the imaging system. For example, the vignetting effect is an optical phenomenon which stands for the undesirable gradual intensity fall off from the image center to its external limits. It corrupts all pixels in the captured pictures, and mask the luminance channel in a non-uniform manner. In this paper, we incorporate the method from our previous research to eliminate the vignetting effect [53]. The basic idea is to take the advantage of a hazy sky to use it as a closely uniform light source to illuminate on the camera. A luminance mask is generated and to be applied to all pictures that we take subsequently in order to eliminate the unwanted vignetting effect. Besides, we notice that some cameras do not always produce linear intensity response as expected with respect to the current camera settings. For example, the cameras may intend to suppress the sensor responses when the actual signal strength is reaching its upper bound limit. In this case, we adopt the method that we proposed in the same paper to optimize camera settings in order to ensure that all camera sensors do give linear responses in all circumstances.

The FR and RR image quality metrics are known to require accurate pixel-wise or feature-wise correspondence between the image reproductions and their originals respectively. Many existing researches, especially in the FR approach, place assumptions either on the measurement environment or how the camera is actually deployed and used in the field. We proposed an image registration method with minimized assumptions to engage the challenge [54]. Since the captured image contents are registered with their originals with confidence, we are able to apply FR, RR and NR sharpness metrics without worrying about the geometry and resolution correspondence issues. In order to make a fair comparison, in our experiments, we apply the registered images to all three types of image quality metrics.

We invite 15 human observers to give perceptual ratings to the projected image distortions. Each of them sits on a chair which is placed exactly where we are supposed to place a camera. The viewing condition is similar to a home theater like environment where the room is completely dark and the visual angle from the projection boundaries to the principal axis of observation is about 15 degrees. The blurred images are displayed in a randomized order for every observer, and each time only one image is displayed. The experiment is set up with category judgment method. For each displayed image, the observers are asked to indicate the overall perceptual sharpness with a category label which stands for the rank between no blurring at all and completely blurred corresponding to the ratings numbers ranging from 1 to 9 respectively.

In this paper, we want to evaluate and compare the sharpness prediction performance of state-of-art image quality metrics. For one thing, we can find out the metrics that correlate the best with perceptual sharpness of projection displays; for another, we can compare the FR, RR, and NR metrics to see which category of them has more advantages on the sharpness predictions in a group-wise comparison manner. To the best knowledge we have, researches from this perspective have not been well engaged in the past. So, we adopt eleven representative image quality metrics in all three categories: SSIM [31], VSNR [34], VIF [33], FSIM [32], RRIQA [35], RRED [39], LPC-SI [14], S-Index [44], CPBD [45], JNBM [50], and S3 [43]. There metrics were either designed to or can be potentially used to measure image sharpness. The selection criteria is established based on the citation frequency, as well as the sharpness features that the metrics rely on; so the typical sharpness metrics are covered in a certain degree, and the corresponding observations will produce hints to the design of a good sharpness metric for projection displays. The performance of these metrics are evaluated with respect to the Pearson and Spearman correlation coefficients between the metric results and the mean Z-scores of perceptual ratings.

4. Experimental Results

4.1. Subjective Sharpness

The perceptual ratings are collected from human observers and they are scaled to generate Z-scores [55] (see Figure 2). It is clear that the perceived sharpness decreases when the blur level increases. However, their relationship should not be simply interpreted with a linear regression model, since the Z-scores for test image 1, 4, 6 and 7 appear to have a flat region between the first and second blur levels. This observation suggests that there is a lower bound threshold for observers to detect the sharpness changes. Another observation is that the general tendency of Z-scores for all test images are fairly similar, and their value ranges are almost identical. An overall plot to incorporate all test images is presented in the last plot of Figure 2. Investigation of the results show differences in the agreement between observers. For example, the variance for the blur level 4 is larger than others in test image 1, and also the variance for the blur level 1 in test image 5. However, the one or two outliers are minorities comparatively to all human observers in such cases. This observation suggests that the observers have agreements regarding perceptual sharpness despite of image content.

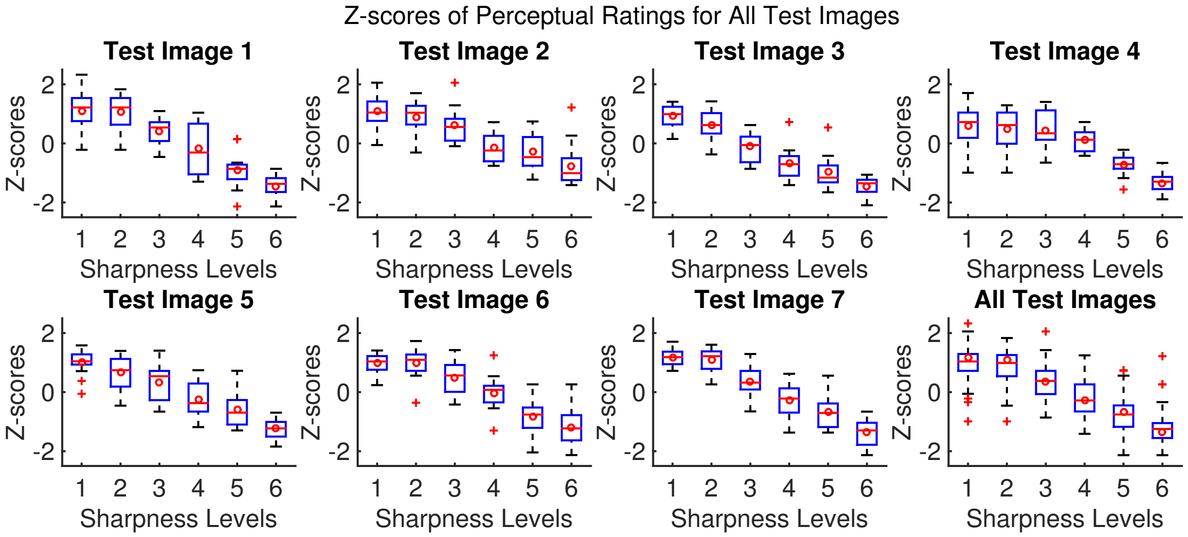


Figure 2. Z-scores of perceptual ratings collected from 15 human observers based on 6 blur levels of 7 selected test images. The blur is simulated with Gaussian filter with kernel size 11 and standard deviation 0.5. The red dots stand for mean Z-scores for each blur levels over all human observers, while the red bars stand for the median. The blue box stand for the 25% inner quantile of Z-scores, and the blue bars stand for 75% outer quantile of Z-scores. The red crosses stand for outliers with respect to the 75% outer quantile. All plots are scaled to have the identical value ranges for Z-scores from -2.5 to 2.5.

4.2. Objective Sharpness

4.2.1. Performance Comparison by Correlations

We calculate the Pearson correlation coefficients between the objective and subjective sharpness for each test image (see Figure 3). The purpose is to understand how well the metrics perform with respect to specific image content. It is clear that in most cases the correlation coefficients are larger than 0.85; especially, for the SSIM, VIF, FSIM and LPCSI metrics, the correlation coefficients are above 0.9 for all test images. In addition, the top rank metrics have fairly close performance for most of the test images. From this perspective, the state-of-art image quality metrics have good correlations with perceptual sharpness in general. It is interesting to see which metric performs the best, and it can be more interesting to figure out the cause of poor metric performance. For example, the VSNR metric for test image 2, RRED metric for test image 3, RRIQA, SIndex, CPBD and S3 metrics for test image 4, have correlation coefficients less than 0.8. In order to explore the common patterns, we generate the plots of objective sharpness versus perceptual sharpness for the VSNR, RRED, RRIQA, SIndex, CPBD and S3 metrics for specific test images (see Figure 4). The value ranges of perceptual sharpness are limited to between -2.5 and 2.5, and they are identical to the ones used in Figure 2. However, the value ranges of objective sharpness are not identical to all test images, since the scales of actual metric results are image dependent. The straight lines represent the linear regressions of objective sharpness versus perceptual sharpness based on least-square-fitting method. The circles stand for the measured sharpness, and each of them corresponds to one sharpness distortion level. By looking at the plots, it is clear that the VSNR and RRED metrics have inconsistent rank orders on measured sharpness for test image 2 and 3. For the test images 4, the RRIQA, SIndex, CPBD and S3 metrics all reserve the rank orders well, but they have inconsistent sharpness derivatives between consecutive distortion levels. For the strongly blurred images (the left most three distortion levels, their standard deviations for Gaussian blur are 3, 2 and 1.5 respectively) the sharpness derivatives are less than expected; for the slightly blurred images (the right most three distortion levels, their standard deviations are 1, 0.5 and 0 respectively), the sharpness derivatives are larger than expected. In contrary, the VIF and FSIM metrics perform very well for test images 4 in both terms of rank order and sharpness derivative.

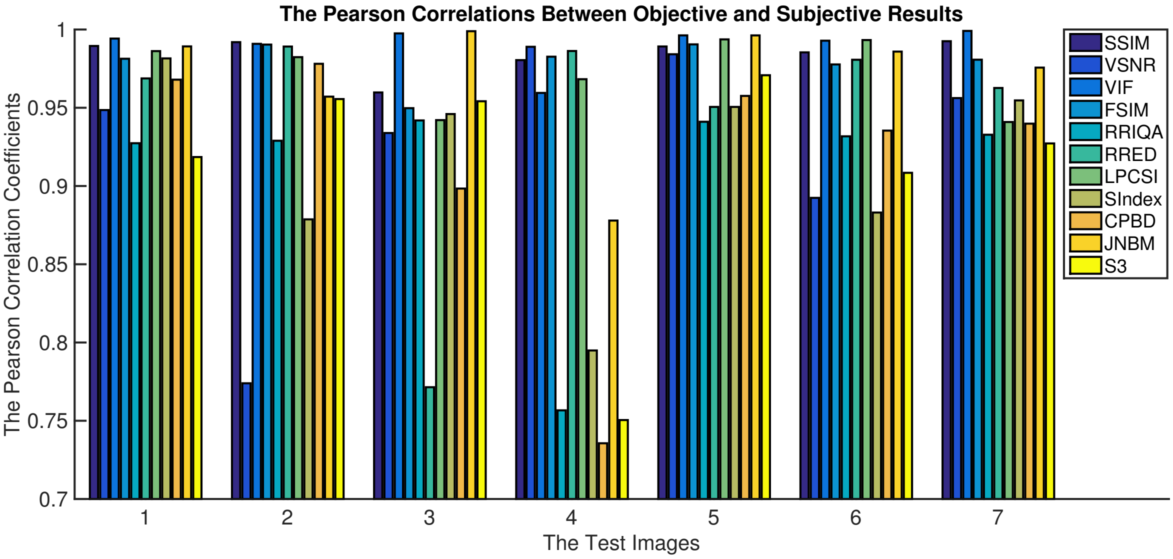


Figure 3. The Pearson correlation coefficient between objective and subjective sharpness for each test image over all distortion levels. In the experiment, we have 7 selected original test images, each image is blurred with Gaussian filter with kernel size 11 and standard deviation 0.5. We invite 15 human observers in the subjective experiments to give perceptual ratings.

4.2.2. *Performance Comparison by Groups*

It is not very convenient in Figure 3 to point out which metric performs the best overall with respect to the average prediction accuracy and stability. From this point of view, we generate box plots to depict the distributions of correlation coefficients for each metric despite of image content (see Figure 5), so each data column in the plots represents the correlation coefficients for all test images at all distortion levels for one specific metric. The red dots stand for the means of coefficients while the red bars stand for the medians. The box stand for the 25% inner quantiles and the bars stand for 75% outer quantiles of correlation values. The red crosses stand for outliers with respect to the 75% outer quantiles. Comparing to the typical analysis in existing literature, we generate box plots instead of introducing only means and confidence intervals. In this way, we are able to observe the mean, median and variance as well as the outliers. Since the captured images in the experiments are all registered with their original ones, all metrics in the experiments have the identical input images with the exactly the same dimension, content and optical degradation. For convenience, we group the metrics into FR, RR and NR categories. SSIM, VSNR, VIF and FSIM are FR metrics, RRIQA abd RRED are RR metrics, while the rest are NR metrics. Obviously, the average performance of FR metrics is higher than NR metrics, and the FR metrics tend to give more stable outcomes with respect to

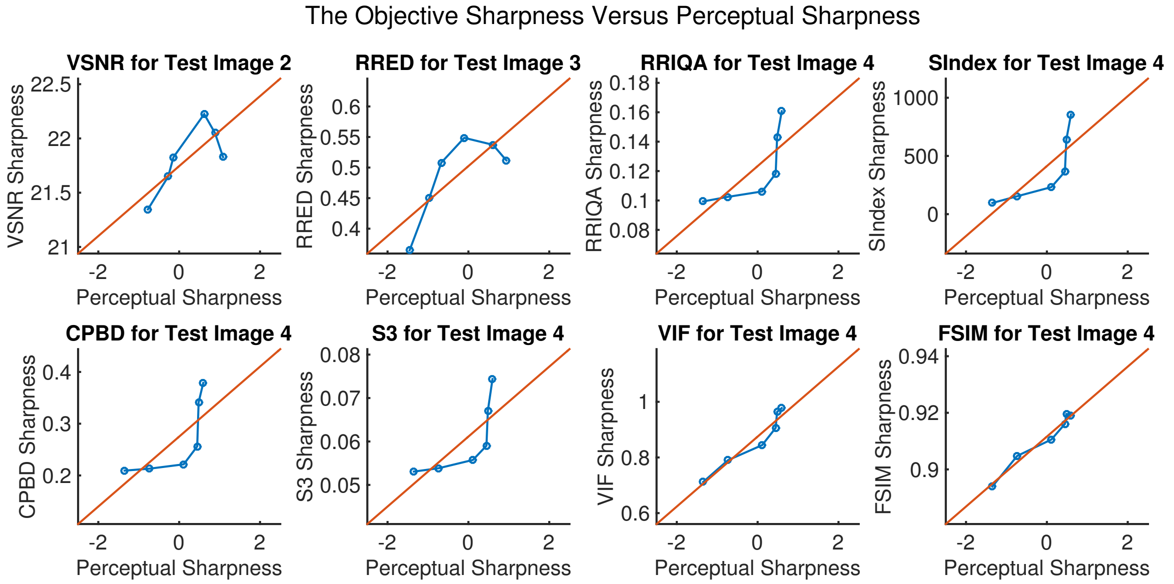


Figure 4. The plots of objective sharpness versus the perceptual sharpness for specific test images. The ranges of perceptual sharpness scaled to be between -2.5 and 2.5 which are identical to the ones used in Figure 2. However, the ranges of objective sharpness are not identical to all test images, since the scales of metric results are image dependent. The straight lines stand for the linear regressions for objective sharpness versus perceptual sharpness. The crosses stand for the sharpness predictions made by objective metric, and each of them corresponds to one sharpness distortion level.

the variance. Since one goal of image quality assessment is to maximize the prediction accuracy of image quality attributes, we should consider to incorporate FR metrics in priority.

4.2.3. Overall Performance Comparison

Notice that in the Figure 4 the scales of objective sharpness are not identical. This is mainly because each image quality metric has its own sharpness scale regarding specific image content. In other words, the rank order of sharpness predictions among different distortion levels might be well preserved but not between different test images for one metric (see Figure 6). In this figure, we can see that the rank order of objective sharpness is well preserved within each test image. However, the value range and variance of objective sharpness may vary a lot from one test image to another. Although the objective sharpness are not normalized for all metrics, the sharpness prediction curves for LPCSI metric are largely get across with each other, while for VSNR metric the curves have clearly distance from each other. This observation suggests that the LPCSI metric adapts to image content better than the other two metrics, and it should be applied in the circumstances that the generalization of sharpness prediction is a concern. In order to compare the generalization performance, we calculate the Pearson correlations between objective and subjective sharpness over all distortion levels for all test images for each metric (see Figure 7). It is clear that LPCSI metric delivers the best overall generalization performance, and the corresponding correlation coefficient is larger than 0.85 even in Figure 6 we have seen that its sharpness prediction curves for test image 1 and 3 are relatively away from others.

5. Conclusion

In this paper, we conduct an experimental study of perceived sharpness on projection displays in a home theater like dark room. The perceptual results suggest that the perceived sharpness follows a nonlinear tendency pattern but its rank order remain the same as the blur level increases. The correlations between the metrical and perceptual results indicate that SSIM, FSIM and VIF metrics give excellent prediction performance in most cases in terms of

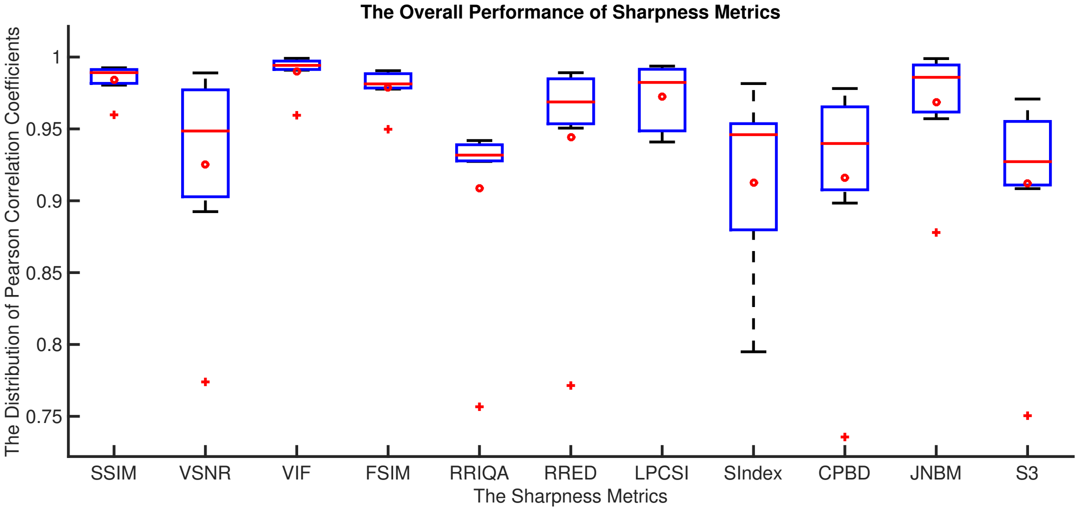


Figure 5. The prediction performance of sharpness metrics over all test images. The red dots stand for mean Pearson correlation coefficient for each sharpness metric over all test images, while the red bars stand for the median. The blue box stand for the 25% inner quantile of correlation values, and the blue bars stand for 75% outer quantile of correlation values. The red crosses stand for outliers with respect to the 75% outer quantile. The notation FR and NR beside the metric names are the indications for FR and NR metrics respectively.

both correlation values and their variances. According to the group comparison, FR metrics have comparatively better prediction performance than NR metrics. In the coming future, we should turn to focus on the design of a good sharpness metric based on the VIF metric for projection displays following the hints that we obtained from this research.

Acknowledgment

We would like to acknowledge the contributions of a large number of volunteers and Yao Cheng's hard works on data collection in the subjective experiments. This work is part of the HyPerCept research program funded by the Research Council of Norway.

References

[1] O. Bimber, E. Andreas, Multifocal Projection: A Multiprojector Technique for Increasing Focal Depth, IEEE Transactions on Visualization and Computer Graphics 12 (4) (2006) 658–67. doi:10.1109/TVCG.2006.75.

[2] B. Sajadi, M. Lazarov, A. Majumder, M. Gopi, Color Seamlessness in Multi-projector Displays Using Constrained Gamut Morphing, IEEE Transactions on Visualization and Computer Graphics 15 (6) (2009) 1317–25. doi:10.1109/TVCG.2009.124.

[3] O. Bimber, A. Emmerling, T. Klemmer, Embedded Entertainment with Smart Projectors, IEEE Computer 38 (1) (2005) 48–55. doi:10.1109/MC.2005.17.

[4] W. Zou, H. Xu, Colorimetric Color Reproduction Framework for Screen Relaxation of Projection Display, Displays 32 (5) (2011) 313–319. doi:10.1016/j.displa.2011.06.001.

[5] M. Pedersen, N. Bonnier, J. Y. Hardeberg, F. Albregtsen, Attributes of Image Quality for Color Prints, Journal of Electronic Imaging 19 (1) (2010) 011016–1–011016–13. doi:10.1117/1.3277145.

[6] G. M. Johnson, Measuring Images: Differences, Quality, and Appearance, Ph.D. thesis, Rochester Institute of Technology (2003).

[7] J. You, L. Xing, A. Perkis, X. Wang, Perceptual Quality Assessment for Stereoscopic Images based on 2D Image Quality Metrics and Disparity Analysis, in: International Workshop on Video Processing and Quality Metrics for Consumer Electronics, Scottsdale, AZ, USA, 2010, pp. 1–6.

[8] T. M. Lehtimaki, K. Saaskilahti, T. Pitkaaho, T. J. Naughton, Evaluation of Perceived Quality Attributes of Digital Holograms Viewed with A Stereoscopic Display, in: Euro-American Workshop on Information Optics, IEEE, Helsinki, Finland, 2010, pp. 1–3. doi:10.1109/WIO.2010.5582499.

[9] J.-B. Thomas, A. M. Bakke, A Colorimetric Study of Spatial Uniformity in Projection Displays, in: A. Trémeau, R. Schettini, S. Tominaga (Eds.), Computational Color Imaging Workshop, Vol. 5646, Springer Berlin Heidelberg, Etienne, Saint, France, 2009, pp. 160–169. doi:10.1007/978-3-642-03265-3_17.

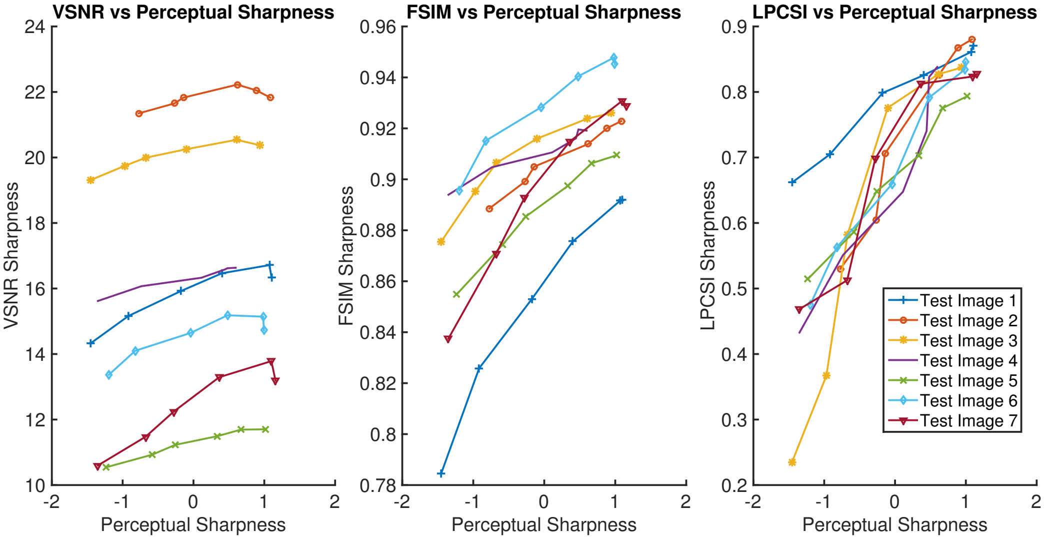


Figure 6. The prediction performance of sharpness metrics over all test images. The red dots stand for mean Pearson correlation coefficient for each sharpness metric over all test images, while the red bars stand for the median. The blue box stand for the 25% inner and 75% outer quantiles of correlation values. The red crosses stand for outliers with respect to the 75% outer quantile.

[10] M. Strand, J. Y. Hardeberg, P. Nussbaum, Color Image Quality in Projection Displays: A Case Study, in: R. Rasmussen, Y. Miyake (Eds.), Image Quality and System Performance II, Proceedings of SPIE, SPIE, San Jose, CA, USA, 2005, pp. 185–195. doi:10.1117/12.587281.

[11] A. Majumder, R. Stevens, Perceptual Photometric Seamlessness in Projection-based Tiled Displays, ACM Transactions on Graphics 24 (1) (2005) 118–139. doi:10.1145/1037957.1037964.

[12] L. Land, I. Juricevic, A. Wilkins, M. Webster, Visual Discomfort and Natural Image Statistics, Journal of Vision 9 (8) (2010) 1046–1046. doi:10.1167/9.8.1046.

[13] V. L. Jaya, R. Gopikakumari, IEM : A New Image Enhancement Metric for Contrast and Sharpness Measurements, International Journal of Computer Applications 79 (9) (2013) 1–9.

[14] R. Hassen, Z. Wang, M. M. A. Salama, Image Sharpness Assessment Based on Local Phase Cohernce, IEEE Transactions on Image Processing 22 (7) (2013) 2798–2810.

[15] N. F. Zhang, A. Vladar, M. T. Postek, R. D. Larrabee, A Kurtosis-based Statistical Measure for Two-dimensional Processes and Its Application to Image Sharpness, in: Proceedings of the 2003 Section on Physical and Engineering Sciences, Alexandria, VA, USA, 2003, pp. 4730–4736.

[16] J. Caviedes, F. Oberti, A New Sharpness Metric Based on Local Kurtosis, Edge and Energy Information, Signal Processing: Image Communication 19 (2) (2004) 147–161. doi:10.1016/j.image.2003.08.002.

[17] J. Caviedes, S. Gurbuz, No-reference Sharpness Metric based on Local Edge Kurtosis, in: Proceedings of International Conference on Image Processing, Vol. 3, IEEE, Rochester, NY, USA, 2002, pp. 2311–6. doi:10.1109/ICIP.2002.1038901.

[18] C. F. Batten, Autofocusing and Astigmatism Correction in The Scanning Electron Microscope, Phd thesis, University of Cambridge (2000).

[19] A. Choudhury, G. Medioni, Perceptually Motivated Automatic Sharpness Enhancement using Hierarchy of Non-local Means, in: IEEE International Conference on Computer Vision, 2011, pp. 730–737. doi:10.1109/ICCVW.2011.6130325.

[20] E. Ong, W. Lin, Z. Lu, X. Yang, S. Yao, F. Pan, L. Jiang, F. Moschetti, A No-reference Quality Metric for Measuring Image Blur, in: International Symposium on Signal Processing and Its Applications, Vol. 1, IEEE, Paris, France, 2003, pp. 469–472. doi:10.1109/ISSPA.2003.1224741.

[21] P. Marziliano, F. Dufaux, S. Winkler, T. Ebrahimi, Perceptual Blur and Ringing Metrics: Application to JPEG2000, Signal Processing: Image Communication 19 (2) (2004) 163–172. doi:10.1016/j.image.2003.08.003.

[22] X. Marichal, W.-Y. Ma, H. Zhang, Blur Determination in The Compressed Domain Using DCT Information, in: International Conference on Image Processing, Vol. 2, IEEE, Kobe, Japan, 1999, pp. 386–390. doi:10.1109/ICIP.1999.822923.

[23] N. B. Nill, B. H. Bouzas, Objective Image Quality Measure Derived from Digital Image Power Spectra, Optical Engineering 31 (4) (1992) 813. doi:10.1117/12.56114.

[24] R. Ferzli, L. J. Karam, No-reference Objective Wavelet Based Noise Immune Image Sharpness Metric, in: International Conference on Image Processing, Vol. 1, IEEE, Genova, Italy, 2005, pp. 405–408. doi:10.1109/ICIP.2005.1529773.

[25] D. Williams, P. D. Burns, Measuring and Managing Digital Image Sharpening, in: Proceedings of IS&T Archiving Conference, SPIE, Bern, Swizerland, 2008, pp. 89–93.

[26] ISO 12233:2014 Photography - Electronic Still Picture Imaging - Resolution and Spatial Frequency Responses, Tech. rep., International Organization for Standardization (2014).

[27] F. Cao, F. Guichard, H. Hornung, Measuring Texture Sharpness of A Digital Camera, in: IS&T/SPIE Electronic Imaging Symposium, Vol. 2009, SPIE, San Jose, CA, USA, 2009, pp. 1–8. doi:10.1117/12.805853.

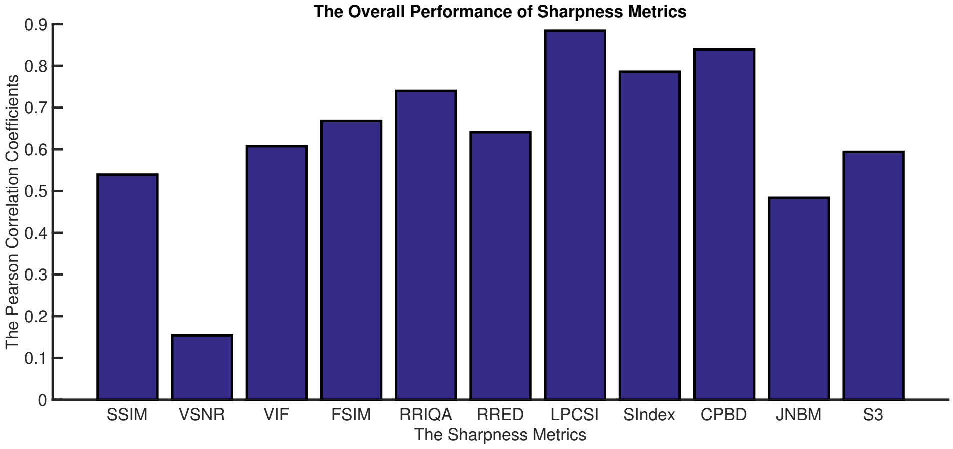


Figure 7. The plots of Pearson correlation between objective and perceptual sharpness over all test images for each image quality metric.

[28] S. Gao, Y. Wang, W. Jin, X. Zhang, Perceptual Sharpness Metric based on Human Visual System, Electronics Letters 50 (23) (2014) 1695–1697. doi:10.1049/el.2014.2844.

[29] M. Nuutinen, O. Orenius, T. Saamanen, P. Oittinen, A Framework for Measuring Sharpness in Natural Images Captured by Digital Cameras based on Reference Image and Local Areas, EURASIP Journal on Image and Video Processing 2012 (1) (2012) 8. doi:10.1186/1687-5281-2012-8.

[30] Z. Wang, A. C. Bovik, Modern Image Quality Assessment, 1st Edition, Morgan & Claypool, 2006. doi:10.2200/S00010ED1V01Y2005081VM003.

[31] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, Image Quality Assessment: from Error Visibility to Structural Similarity, IEEE Transactions on Image Processing 13 (4) (2004) 600–612. doi:10.1109/TIP.2003.819861.

[32] L. Zhang, L. Zhang, X. Mou, D. Zhang, FSIM: A Feature Similarity Index for Image Quality Assessment., IEEE Transactions on Image Processing 20 (8) (2011) 2378–2386. doi:10.1109/TIP.2011.2109730.

[33] H. R. Sheikh, A. C. Bovik, Image Information and Visual Quality, IEEE Transactions on Image Processing 15 (2) (2006) 430–444. doi:10.1109/TIP.2005.859378.

[34] D. M. Chandler, S. S. Hemami, VSNR: A Wavelet-Based Visual Signal-to-Noise Ratio for Natural Images, IEEE Transactions on Image Processing 16 (9) (2007) 2284–2298. doi:10.1109/TIP.2007.901820.

[35] Z. Wang, E. P. Simoncelli, Reduced-Reference Image Quality Assessment Using A Wavelet-Domain Natural Image Statistic Model, in: B. E. Rogowitz, T. N. Pappas, S. J. Daly (Eds.), Proceedings of SPIE Human Vision and Electronic Imaging X, SPIE, San Jose, USA, 2005, pp. 149–159. doi:10.1117/12.597306.

[36] M. Nuutinen, R. Halonen, T. Leisti, P. Oittinen, Reduced-reference Quality Metrics for Measuring The Image Quality of Digitally Printed Natural Images, in: Image Quality and System Performance VII, The Annual Symposium on Electronic Imaging, Vol. 7529, SPIE, San Jose, CA, USA, 2010, pp. 75290I–75290I–12. doi:10.1117/12.838883.

[37] G. Cheng, L. Cheng, Reduced Reference Image Quality Assessment based on Dual Derivative Priors, Electronics Letters 45 (18) (2009) 937. doi:10.1049/el.2009.1210.

[38] W. Xue, X. Mou, Reduced Reference Image Quality Assessment based on Weibull Statistics, in: Proceedings of International Workshop on Quality of Multimedia Experience, IEEE, Trondheim, Norway, 2010, pp. 1–6. doi:10.1109/QOMEX.2010.5518131.

[39] R. Soundararajan, A. C. Bovik, RRED Indices: Reduced Reference Entropic Differencing for Image Quality Assessment, IEEE Transactions on Image Processing 21 (2) (2012) 517–526. doi:10.1109/TIP.2011.2166082.

[40] A. Maalouf, M.-C. Larabi, A No Reference Objective Color Image Sharpness Metric, in: European Signal Processing Conference, Aalborg, Denmark, 2010, pp. 1019–1022.

[41] Z. Cao, Z. Wei, G. Zhang, A No-Reference Sharpness Metric Based on Structured Ringing for JPEG2000 Images, Advances in Optical Technologies 2014 (2014) 1–13. doi:10.1155/2014/295615.

[42] B. Samira, M. Lindsay, Colour Difference Metrics and Image Sharpness, in: Color and Imaging Conference Final Program and Proceedings, Society for Imaging Science and Technology, Scottsdale, AZ, USA, 2000, pp. 262–267.

[43] C. T. Vu, D. M. Chandler, S3: A Spectral and Spatial Sharpness Measure, in: International Conference on Advances in Multimedia, IEEE, Colmar, France, 2009, pp. 37–43. doi:10.1109/MMEDIA.2009.15.

[44] A. Leclaire, L. Moisan, No-reference Image Quality Assessment and Blind Deblurring with Sharpness Metrics Exploiting Fourier Phase Information, Journal of Mathematical Imaging and Vision.

[45] N. D. Narvekar, L. J. Karam, A No-Reference Image Blur Metric Based on the Cumulative Probability of Blur Detection, IEEE transactions on image processing : a publication of the IEEE Signal Processing Society 20 (1) (2011) 2678–2683. doi:10.1109/TIP.2011.2131660.

[46] N. Narvekar, L. J. Karam, A No-reference Perceptual Image Sharpness Metric based on A Cumulative Probability of Blur Detection, in: K. O. Egiazarian, S. S. Agaian, A. P. Gotchev, J. Recker, G. Wang (Eds.), International Workshop on Quality of Multimedia Experience, IEEE, San Diego, CA, USA, 2009, pp. 87–91. doi:10.1109/QOMEX.2009.5246972.

[47] N. G. Sadaka, L. J. Karam, R. Ferzli, G. P. Abousleman, A No-reference Perceptual Image Sharpness Metric based on Saliency-weighted Foveal Pooling, in: IEEE International Conference on Image Processing, IEEE, San Diego, CA, USA, 2008, pp. 369–372. doi:10.1109/ICIP.2008.4711768.

[48] R. Ferzli, J. J. Karam, Human Visual System Based No-Reference Objective Image Sharpness Metric, in: International Conference on Image Processing, IEEE, Atlanta, GA USA, 2006, pp. 2949–2952. doi:10.1109/ICIP.2006.312925.

[49] Z. Wang, E. P. Simoncelli, Local Phase Coherence and the Perception of Blur, in: Neural Information Processing Systems, no. December 2003, MIT Press, Whistler, Canada, 2003, pp. 9–11.

[50] R. Ferzli, 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) (2009) 717–28. doi:10.1109/TIP.2008.2011760.

[51] E. Martinec, P. Lee, AMAZE Demosaicing Algorithm (2010).

[52] X. Liu, M. Pedersen, J. Y. Hardeberg, CID:IQ - A New Image Quaity Database, in: A. Elmoataz, O. Lezoray, F. Nouboud, D. Mammass (Eds.), International Conference on Image and Signal Processing, Vol. 8509 of Lecture Notes in Computer Science, Springer International Publishing, Cherbourg, Normandy, France, 2014, pp. 193–202. doi:10.1007/978-3-319-07998-1.

[53] P. Zhao, M. Pedersen, J.-B. Thomas, J. Y. Hardeberg, Perceptual Spatial Uniformity Assessment of Projection Displays with a Calibrated Camera, in: Color and Imaging Conference, Society for Imaging Science and Technology, Boston, MA, USA, 2014, pp. 159–164.

[54] P. Zhao, M. Pedersen, J. Y. Hardeberg, J.-B. Thomas, Image Registration for Quality Assessment of Projection Displays, in: International Conference on Image Processing, IEEE, Paris, France, 2014, pp. 3488–3492. doi:10.1109/ICIP.2014.7025708.

[55] P. G. Engeldrum, Psychometric Scaling: A Toolkit for Imaging Systems Development, Imcotek Pr, 2000.

Chapter 11

Paper F

Extending Subjective Experiments for Image Quality Assessment with Baseline Adjustments

Ping Zhao, and Marius Pedersen

Published in

Mohamed-Chaker Larabi and Sophie Triantaphillidou, Ed.,

Image Quality and System Performance XII,

Proceedings of the 27th Annual Symposium on Electronic Imaging (EI 2015)

Page 93960R-1-93960R-13,

International Society for Optics and Photonics,

Boston, MA, USA,

February, 2015.

Extending Subjective Experiments for Image Quality Assessment with Baseline Adjustments

Ping Zhao and Marius Pedersen

Gjøvik University College, Teknologivn. 22, 2815 Gjøvik, Norway

ABSTRACT

In a typical working cycle of image quality assessment, it is common to have a number of human observers to give perceptual ratings on multiple levels of distortions of selected test images. If additional distortions need to be introduced into the experiment, the entire subjective experiment must be performed over again in order to incorporate the additional distortions. However, this would usually consume considerable more time and resources. Baseline adjustment is one method to extend an experiment with additional distortions without having to do a full experiment, reducing both the time and resources needed. In this paper, we conduct a study to verify and evaluate the baseline adjustment method regarding extending an existing subjective experimental session to another. Our experimental results suggest that the baseline adjustment method can be effective. We identify the optimal distortion levels to be included in the baselines should be the ones of which the stimulus combinations produce the minimum standard deviations in the mean adjusted Z-scores over all human observers in the existing rating session. We also demonstrate that it is possible to reduce the number of baseline stimuli, so the cost of extending subjective experiments can be optimized. Comparing to conventional researches mainly focusing on case studies of hypothetical data sets, we perform this research based on the real perceptual ratings collected from an existing subjective experiment.

Keywords: subjective experiment, baseline adjustments, image quality, psychometric scaling

1. INTRODUCTION

Image Quality Assessment (IQA) is a complicated task, because it associates with many systematic methodologies and one has to follow a well defined work-flow to engage his/her research problems. Many researches regarding IQA have been done in the domains like color printing,[1,2] flat-panel display,[3,4] image compression[5,6] and vision science.[7,8] Subjective IQA still remains the most precise way to quantify image quality,[9] despite the effort in finding an objective quality metric.[10] In order to extensively evaluate the quality of imaging metrics or systems, a large number of image stimuli and human observers are required, and the scaling method should be carefully specified as well. Experimental outcomes are usually constrained by the these factors. The number of image stimuli is proportional to the amount of time used by the human observers, and the length of the scaling study. There is constant trade off between the wish to have as many stimuli as possible and the acceptable resource (time, money, observers, etc.) consumption. Usually, an agreement between the number of stimuli and observers are found, which is a reasonable midpoint.

Perceptual ratings are commonly collected from multiple experimental sessions, where a session stands for a group of human observers and their ratings on a set of stimuli. The entire process consumes considerable time and resources. Many approaches have been proposed to address this research challenge. In ISO standard 20462-2,[11] triplet comparison was introduced to reduce the number of pairs to be evaluated by human observers or the use of incomplete data.[12–16] These methods have the advantage of decreasing the number of stimuli that can be evaluated. Engeldrum proposed a method to split the subjective experiment into different segments.[9] However it can be problematic since observers need to come back several times in order to finish the experiment. Time is always a limiting factor for subjective experiments, and therefore many guidelines can be found concerning the maximum amount of time required. One hour limitation is a common rule of thumb,[9,17,18] while the International Telecommunication Union recommends 30 minutes time limitation,[19] and Larabi recommends that the median

Further author information: Ping Zhao: E-mail: ping.zhao@hig.no, Telephone: +47-61135214, Marius Pedersen: E-mail: marius.pedersen@hig.no, Telephone: +47-61335246

time over the observers should not be more than 45 minutes.[20] Software have also been created to allow for large-scale experiments,[21] for both laboratory and on-line experiments. Regardless of the method used they can usually not be extended to include additional distortions or reproductions after the experiment is completed. Suppose that in an existing session, we have ratings of four levels of image distortions; later, two additional distortions need to be introduced to extend the existing subjective experiment. In this case, conventionally, we have to conduct a new experiment with all six levels of image distortions. The overall amount of workload can be demanding. In addition, in practice, the human observers involved in the new session are unlikely to be identical to those invited in the existing session. However, if we have enough human observers, the averaged ratings regarding the existing image distortions should be statistically similar across the two sessions. In this context, a natural research question to ask would be that is it possible to take this advantage without executing the entire experiment over again, especially when the amount of distortions is large.

Baseline adjustment can be a potential answer to this research problem. This method introduces common stimuli (one or more distortions) to form a baseline in order to determine the comparability of ratings between different experiment sessions, and allows the computation of scale values expressed relative to responses for the baseline stimuli.[22] The baseline adjustment is carried out separately for each original stimulus. Many existing researches have introduced baseline adjustments into their scaling procedures.[22–27] However, in these researches, either the selection criteria of baseline is not discussed in depth,[22–25] or the baseline stimuli were simply selected randomly from existing candidates.[26, 27] A natural research question to ask is about what types of stimuli should be included in order to form a representative baseline, and how many stimuli are essential.

In this paper, we conduct a study to verify and evaluate the baseline adjustment method for extending subjective experiments. The first goal is to verify that the baseline adjustment is an effective method, and the second goal is to identify the type and number of stimuli that we should use in the common baseline in order to minimize the experimental workload and complexity. Comparing to conventional researches focusing on case studies of hypothetical data sets, we perform this research based on real perceptual data collected from an existing subjective experiment. The rest of this paper is organized as follows: first, in Section 2, we study the existing psychometric models for subjective experiments and discuss the research challenges for psychometric ratings and scaling procedures. In Section 3, we present our experimental setup and results. At last, in Section 4, conclusions are drawn based on the data observations.

2. BACKGROUND

2.1 Psychometric Models

The goal of subjective IQA is obtaining the perceptual indications regarding a specific image quality attribute or overall image quality. A typical work-flow can be generalized as a conceptual psychometric model which is divided into two major procedures: rating and scaling.

In the rating procedure, the human visual system acquires the displayed images; then the brain interprets the information to generate opinions regarding the underlying image quality attribute. This implicit perceptual and cognitive processes vary largely from one observer to another, but they can be potentially influenced via the interactions with either the instructor or the environment in the field. In the case of IQA, the end product of the rating procedure is a matrix representing the numerical ratings of each level of image distortion from all human observers. Brown et al.[25] presented an excellent research regarding the challenges of interpreting the rating scales. They identified the research challenges and classified them into five major categories: unequal-interval judgment criterion scales, lack of inter-observer correspondence, linear difference between group average criterion scales, lack of intro-observer consistency, perceptual and criterion shifts. A good understanding of these problems is essential for advancing the design and improvements of psychometric model. Suppose that we have four human observers A, B, C, and D, and they are asked to rate three distortion levels of one test image with category judgment method; obviously, each human observer has his/her own judgment criterion scales. In this context, the challenges in the rating procedure can be briefly demonstrated in Figure 1. Typically, the judgment criterion scales have different origins, ranges, and intervals. The differences of origins and ranges are mainly due to the natural preference for perception and/or the temporary variations of judgment criterion scales; and they can be estimated with linear transfer functions which are formulated with psychometric scaling models. However,

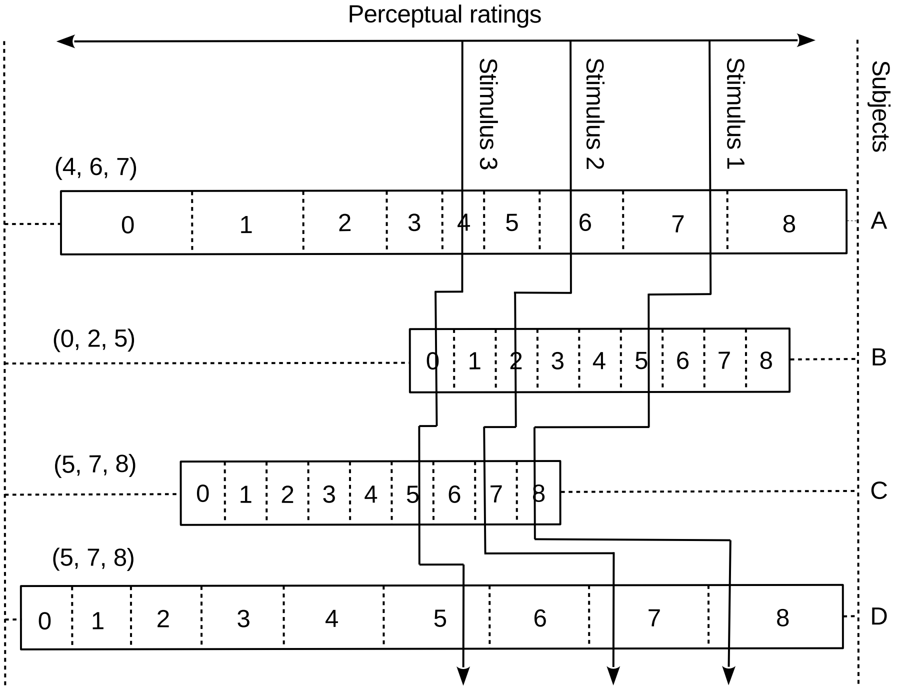


Figure 1. The judgment criterion scales of four human observers, A, B, C and D, and they are asked to rate three levels of distortions of one test image with category judgment method. Obviously, each human observer has his/her own judgment criterion scales. The ratings they have made are $(4, 6, 7)$, $(0, 2, 5)$, $(5, 7, 8)$, and $(5, 7, 8)$ respectively. Typically, the judgment criterion scales have different origins, ranges and intervals.

so far there is no effective way to quantify the interval differences, since the psychometric rating is completely an implicit perceptual and cognitive process. Empirically, well trained observers with color expertise are more likely to have above average equal intervals, while the non-experts are not. One may argue that it is possible to employ Monte Carlo like statistical analysis to estimate the judgment criterion scales, however the essential large amount of random tests are impractical to be applied to a large group of human observers. In many cases, the ratings from an individual observer can be inconsistent. To the same stimulus, regarding a specific image quality attribute, one observer may give completely different ratings in various rating sessions. If the variation can be assumed to be a random factor which follows a normal distribution around the true perceived value. The real perceived value can be estimated by statistical regression, but the trade off is that the regression requires a large number of random samples of which the collection is both time and resource consuming.

In the scaling procedure, the raw ratings are transformed in order to distinct the perception of a stimuli and the corresponding judgment criterion scale for assigning rating to that stimuli. The outcomes indicate the relative impression of the perceived image quality attribute or overall image quality. They are meaningless without the references to the observers' judgment criterion scales. Brown et al.[25] presented six typical scaling methods:

- Median rating: it uses median ratings over all human observers regarding a single stimuli as the scaled ratings. There is no assumption of equal intervals of judgment criterion scales. In contrast, it provides only the ordinal information of ratings.
- Mean rating: it uses mean rating as the scaled output, and it requires the interval of judgment criterion scales must be equal. However, this assumption does not hold in most cases.
- Origin-adjusted rating: it removes the rating mean in prior of aggregating them for each human observer and it cancels the differences of origins of judgment criterion scale, but not the differences of interval sizes.

- Z-score: it is similar to origin-adjusted rating in removing the differences of shift. In addition, it normalizes ratings with respect to their standard deviation, so the linear differences between observers are eliminated.
- Least square rating: it does not merely inherit the features of Z-score scaling, but also counts in the correlations between individual and all observers in the same group. Larger correlation indicates for larger contribution from individual observer to the same group of observers.
- Scenic beauty estimate: it was originally developed to scale ratings of scenic beauty of forest area, but the procedures are also appropriate for use with ratings of other types of stimuli. The differences in ratings are assessed by comparing an observers rating distribution (assumed to have a normal distribution) for one landscape area against each of several other landscape areas. It features with a relative operating characteristic, where a bi-variate graph of the cumulative probability of the ratings for the selected landscape, is compared against the cumulative probability of other ratings, respectively. The scaled outcomes are generated by calculating the distance of the standardized relative operating characteristic from a positive diagonal of difference matrix.

2.2 Baseline Adjustments

In a typical working cycle of subjective IQA, the case may occur, in the data post-processing phase, that the researcher realizes that it is mandatory to adopt observations on additional image distortions to draw the final conclusions. For example, a general tendency of human perception has been discovered; but the numerical distance between two consecutive distortion levels might be larger than they are expected. As a result, many perception details within these pre-defined intervals are not available. Conventionally, the researcher needs to conduct a large new subjective experiment incorporating all existing and additional image distortions. The purpose is to make all stimuli to be rated under the same circumstances, so the unwanted experimental artifacts between possibly two or more separate sessions can be largely avoided. However, the whole process is non-trivial and it may consume considerable time and resources.

Baseline adjustment can be a potential answer to this challenge. This method introduces common stimuli to form a baseline in order to determine the comparability of ratings between different experiment sessions, and allows the computation of scale values expressed relative to responses for the baseline stimuli.[22] The basic concept is depicted in Figure 2. The ratings for unique stimuli in both rating sessions are scaled respectively with respect to the selected common baseline in either session, and then they are merged to generate the final ratings. Suppose that we have one human observer, who is asked to rate four stimuli 1, 2, 3, and 4 in the first session and another two additional stimuli 5 and 6 in the second session. In this case, stimuli 3 and 4 are selected to form a common baseline. Notice that the ratings for them across the two rating sessions may have different values. The ratings of unique stimuli 1 and 2 in the first session are scaled with respect to the baseline in the rectangle on the left in Figure 2, and the ratings of unique stimuli 5 and 6 in the second session are scaled with respect to the baseline in the rectangle on the right in Figure 2. Since we are merging the scaled ratings from the second rating session to the first one, then the ratings for the baseline in the first session is scaled to generate the scaled baseline which has zero mean and normalized standard deviation. Finally, all scaled ratings are combined to be used as the final ratings. It is important that the stimuli for the two sessions are rated under the same circumstances. In order to achieve this goal, the following precautions should be followed:[25]

- the observers for each session should be randomly selected from the same observer population,
- the observer groups should be sufficiently large,
- the baseline stimuli should be representative of the full set of stimuli to be rated,
- the non-baseline stimuli should be randomly assigned to the different sessions,
- all other aspects of the sessions (e.g., time of day, experimenter) should remain constant.

Baseline adjustment is a higher level of abstraction on rating scaling, it must be integrated with specific scaling method which is well mathematically formulated. All scaling methods introduced in the previous Section 2.1 can be adopted as the candidates. In this paper, we choose to integrate Z-score scaling with baseline adjustment method. Z-score scaling is widely used in the psychometric modeling, mainly because of the simplicity on its definition and it eliminates the problems of origin shifting and unequal range of judgment criterion scales. Beyond this, the scores are straight forward to compute with computer programs. In this context, the raw ratings are scaled in the following way:[25]

$$BZ_{ij} = \left(R_{ij} - BMR_j\right) / BSDR_j$$

where BZ_{ij} stands for the baseline-adjusted Z-score of stimulus i for observer j, R_{ij} stands for the ratings assigned to stimulus i by observer j, and BMR_j stands for the mean of ratings assigned to the baseline stimuli by observer j, and $BSDR_j$ stands for the standard deviation of ratings of the baseline stimuli by observer j; then the BZ_{ij} are then averaged across all observers to generate one scale value per stimulus as BZ_i.

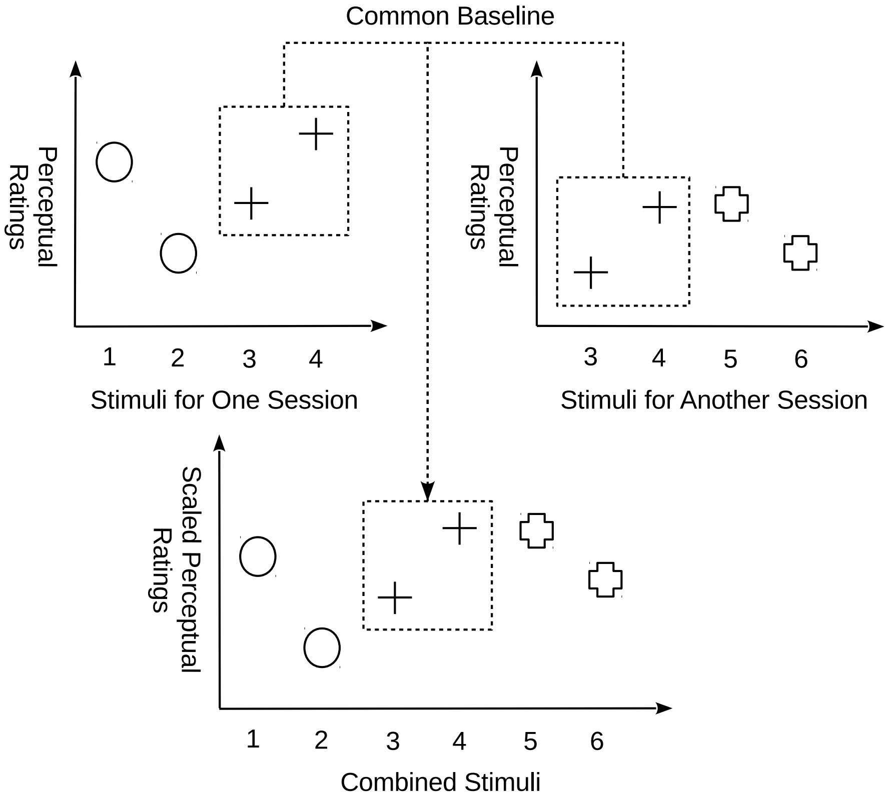


Figure 2. Depiction of the basic concept of baseline adjustment method. In this case, stimuli 3 and 4 are selected to form a common baseline. Notice that the ratings for them across the two rating sessions may have different values. The ratings of unique stimuli 1 and 2 in the first session are scaled with respect to the baseline in the rectangle on the left, and the ratings of unique stimuli 5 and 6 in the second session are scaled with respect to the baseline in the rectangle on the right. Since we are merging the scaled ratings from the second rating session to the first one, then the ratings for the baseline in the first session is scaled to generate the scaled baseline which has zero mean and normalized standard deviation. Finally, all scaled ratings are combined to be used as the final ratings.

3. EXPERIMENT

The first goal of this research is to verify that the baseline adjustment is an effective method for extending subjective experiments, and the second goal is to identify the type and number of image distortion levels that

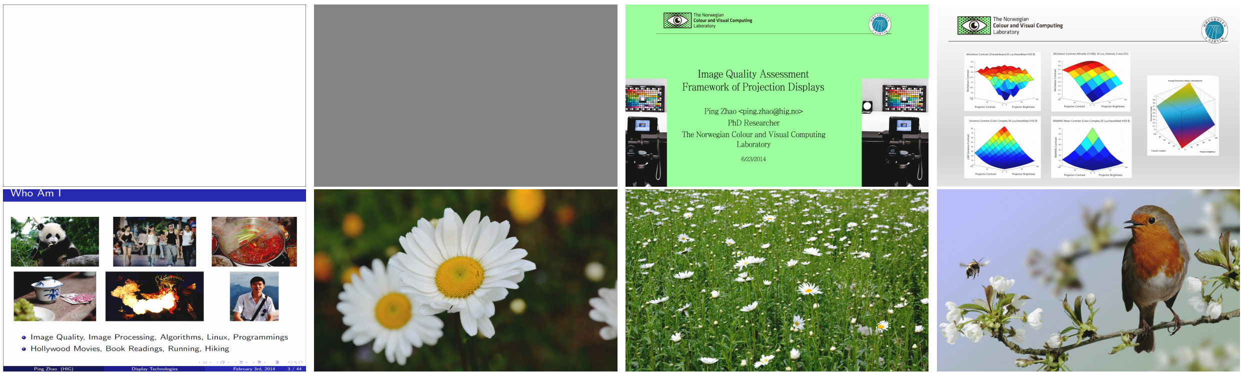

The Norwegian Colour and Visual Computing Laboratory
Image Quality Assessment Framework of Projection Displays
Ping Zhao <ping.zhao@hig.no>
PhD Researcher
The Norwegian Colour and Visual Computing Laboratory
6/23/2014
Who Am I
Image Quality, Image Processing, Algorithms, Linux, Programmings
Hollywood Movies, Book Readings, Running, Hiking
Ping Zhao (HIG)
Display Technologies
February 3rd, 2014

Figure 3. The 8 test images: one peak white patch, one neutral gray patch (normalized input gray value equals to 0.5), three natural color images, and three slide-like images.

we should include in the baseline. In this paper, we focus on integrating baseline adjustments with Z-scores to scale real perceptual ratings collected from an existing subjective experiment rather than hypothetical data.

3.1 Experimental Procedure

In this paper, we take the advantage of the data collected from an existing subjective experiment to simulate the real scenarios. The subjective experiment was designed to evaluate the perceptual spatial uniformity of projection displays. The natural spatial non-uniformity for every pixel was scaled into 7 levels, and then we stacked the multiple levels of non-uniformity onto 8 carefully selected test images (see Figure 3) respectively; so 56 stimuli in total are shown to each observer in a completely randomized order. Since projection displays are most commonly used in a dark room or a regular meeting room, then we decided to conduct the experiment in a control lab environment where we try to simulate a home-theater-like environment. In this environment, we can avoid wanted imaging artifacts (for example, non-uniform sun light passing though windows or varying ambient light disturbance). In this case, two calibrated SONY APL-AW15 LCD projectors (throw put: 1.5) were placed right in front of and about 3m away from a planar screen to produce two projections (both in 1.5x0.9m) in parallel simultaneously. 20 human observers were invited to join the experiment. 13 of them had color science educational background, while the rest did not. 14 of them are male and the rest are female. All of them were required to have a mandatory visual acuity test. The observers were asked to sit in front and between the two parallel projections, so visual angles (around 20 degrees from the center line to the left or right projection boundaries respectively) to the two projections were approximately the same (viewing distance was 4 meter). Each observer was asked to use a natural number between 0 (corresponding to completely uniform) to 10 (corresponding to not uniform at all) to indicate his/her opinion regarding the overall magnitude of perceived spatial non-uniformity. All observers were required to do the experiment twice (each session was 20 minutes on average), resulting in 2240 perceptual ratings in total.

For the experiment in this research, as it is demonstrated in Figure 4, we separate the image distortions into two groups. For example, the first group includes distortion level 1 to 6, and the second group includes distortion level 4 to 7. The ratings for distortion level 1 to 3 in the session 2 are ignored. In this case, we are simulating a scenario which extends existing subjective experiment with distortion level 1 to 6 in order to adopt additional distortion level 7; and the ratings for distortion level 4 to 6 are used to form the adjustment baseline for Z-score scaling in both rating sessions. Since the ratings are scaled on observer basis, in this case, we demonstrate how to scale the ratings for ”Image 1” from ”Observer 1”. The ratings in ”Part 1” are scaled with respect to the ”Baseline 1” in order to generate scaled ratings in ”Adjusted Part 1”, the rating in ”Part 2” is scaled with respect to the ”Baseline 2” in order to generate scaled ratings in ”Adjusted Part 2”, and the ratings for distortion level 4 to 6 on ”Image 1” from ”Observer 1” in ”Session 1” are scaled with respect to ”Baseline 1” in order to generate ”Adjusted Ratings in Baseline 1”. Notice that the two baselines share the same distortion levels, but they may have different rating values. Each baseline includes only the ratings from ”Observer 1” for all test images on the corresponding distortion levels. Eventually, all adjusted ratings in the table below are merged to generate a full set of adjusted Z-scores. Then the mean adjusted Z-scores over

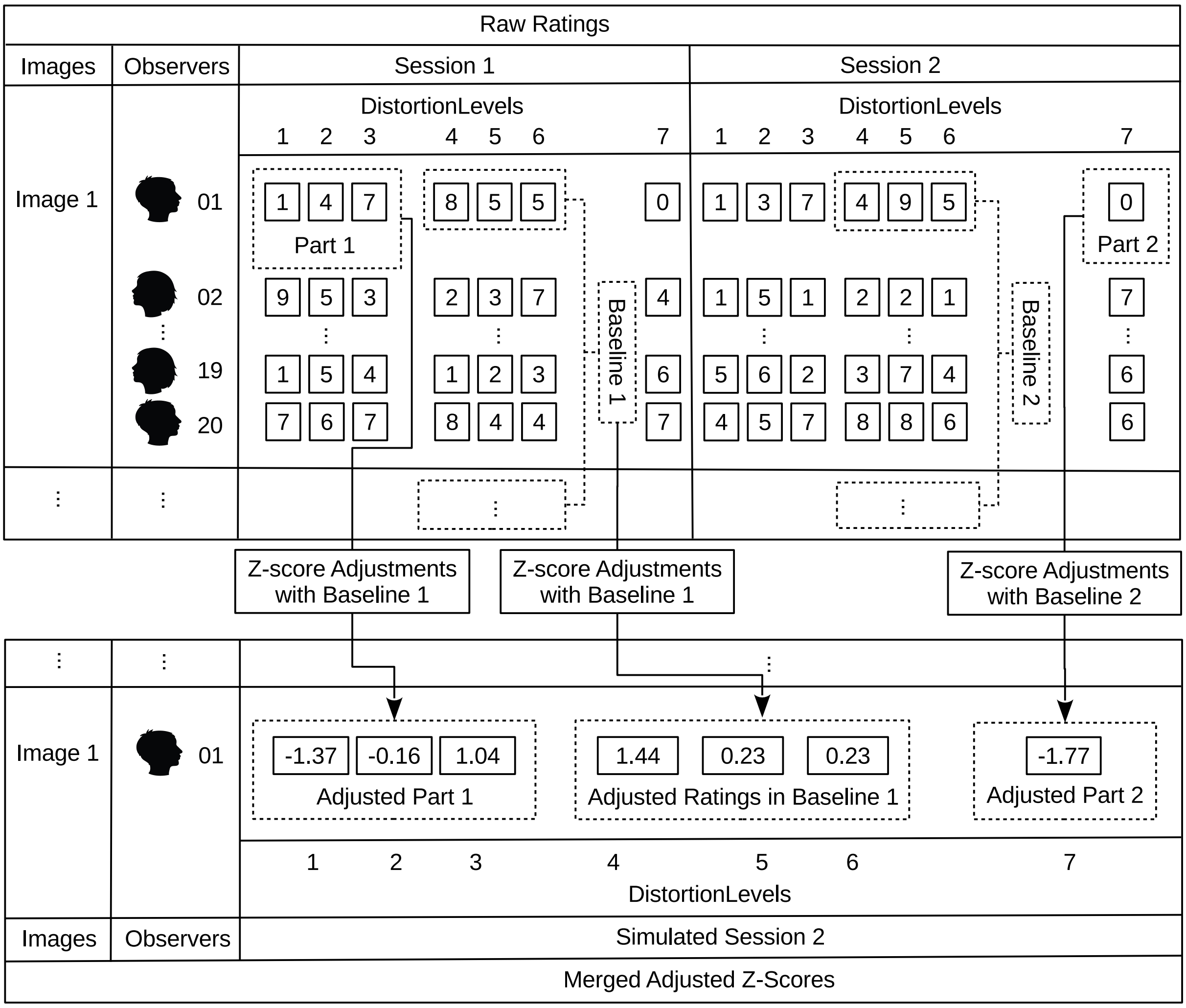


Figure 4. We separate the image distortions into two groups. For example, in this figure, the first group includes distortion level 1 to 6, and the second group includes distortion level 4 to 7. The ratings for distortion level 1 to 3 in the session 2 are ignored. In this case, we are simulating a scenario which extends existing subjective experiment with distortion level 1 to 6 in order to adopt additional distortion level 7; and the ratings for distortion level 4 to 6 are used to form the adjustment baseline for Z-score scaling in both rating sessions. Since the ratings are scaled on observer basis, in this case, we demonstrate how to scale the ratings for "Image 1" from "Observer 1". The ratings in "Part 1" are scaled with respect to the "Baseline 1" in order to generate scaled ratings in "Adjusted Part 1", the rating in "Part 2" is scaled with respect to the "Baseline 2" in order to generate scaled ratings in "Adjusted Part 2", and the ratings for distortion level 4 to 6 on "Image 1" from "Observer 1" in "Session 1" are scaled with respect to "Baseline 1" in order to generate "Adjusted Ratings in Baseline 1". Notice that the two baselines share the same distortion levels, but they may have different rating values. Each baseline includes only the ratings from "Observer 1" for all test images on the corresponding distortion levels. Eventually, all adjusted ratings in the table below are merged to generate a full set of adjusted Z-scores. Then the mean adjusted Z-scores over all observers are correlated with the non-adjusted Z-scores over all human observers in original "Session 1" to determine the performance of the underlying baseline. Since the ratings for the two original sessions are collected from the identical human observers in the same circumstance, the average correlations are expected to be high if the baseline is appropriately specified.

all observers are correlated with the non-adjusted Z-scores over all human observers in original "Session 1" to determine the performance of the underlying baseline. Since the ratings for the two original sessions are collected from the identical human observers in the same circumstance, the average correlations are expected to be high if the baseline is appropriately specified. In this context, we calculate both Pearson and Spearman correlations. Pearson correlation compares the general tendency of two groups of data, while Spearman correlation focuses on the rank order. Obviously, there are many possible combinations of distortion levels and unique distortion levels among the two session, so we write a computer program to permute all combination possibilities and calculate corresponding correlations accordingly. In this way, we have the flexibility to determine the optimal type and numbers of distortion levels included in the baseline with respect to the correlation values. The only constrain we apply here is that the merged full set of adjusted Z-scores must include all 7 levels of distortions, since the correlations require the two groups of numerical data must share identical length.

3.2 Experimental Results

The experimental results are presented in two parts. In the first part, the average correlation results and analysis are presented, the results can be regarded as image content independent. In the second part, the correlation results regarding individual image content are presented.

3.2.1 Overall Correlation Results

In this section, we elevate the observation to a higher level where we focus on the average correlations over all human observers and all test images. The purpose is to identify the general tendency which enables the analysis on how the correlation values associate with the number of distortion levels included in baselines, despite of individual image content. In order to achieve this goal, we generate all possible distortion combinations and calculate the average over all test images. Then, we get the plots depicted in Figure 5. In the plots, the horizontal axis indicates the number of distortion levels included in the baselines, and the vertical axis indicates the corresponding correlation values. Each box stands for a category in which the distortion combinations share the identical number of distortion levels in the baseline, even they might not have exactly the same distortion levels. The distortion combinations in the plots cover all combination possibilities, except the category 6 is not included because in that case there will be no distortion level selectable for simulated new rating session. The blue boxes indicate the inner 25% quartiles, while outer bars indicate outer 75% quartiles. The bars inside the inner boxes stand for the median values each category respecively, since they are very close to the mean values indicated by the dots, then we know that the correlation distributions are approximately non-skewed.

Among all categories, the category 0 indicates the case that we apply only Z-score scaling without baseline adjustment to the raw ratings of two sessions respectively, and calculate the correlations between them. In category 7, all distortion levels are included in the baseline, then we apply the baseline adjustment Z-score scaling to all ratings in both sessions. Since all distortions are adopted, in this case, the baseline adjusted Z-score scaling is equivalent to the non-adjusted Z-score scaling, then the correlation values are exactly ones. It is clear that both the mean Pearson and Spearman correlation values are monotonically increasing, while the standard deviations are shrinking, as the number of distortion levels included in the baseline increases. So we can make a conclusion that, in general, despite of the image content, the more distortion levels we adopted in the baseline, the better and more stable correlations we should have between the existing ratings and the expanded ratings; to a certain distortion level combination included in the baseline, no matter how the unique distortion levels in two session vary, the correlation values tend to less variant. However, meanwhile, the workload of repeating the subjective experiments increases as well. Notice that, in each category, the correlation values can rise up close to one. In other words, it is possible to achieve high correlations without adopting all distortion levels in the baseline. Then the question comes as what are the most optimal distortion levels to be included in the baseline, so the correlation values are as high as possible despite the actual combinations of unique distortions levels involved for the existing and expanded sessions.

3.2.2 Correlation Results for Individual Images

The correlation values do not merely vary with respect to the number of distortion levels included in the baseline, but also vary with individual image content. The research purpose in this section is to identify the connections between correlations and corresponding image content. We generate the plots of correlation values versus all

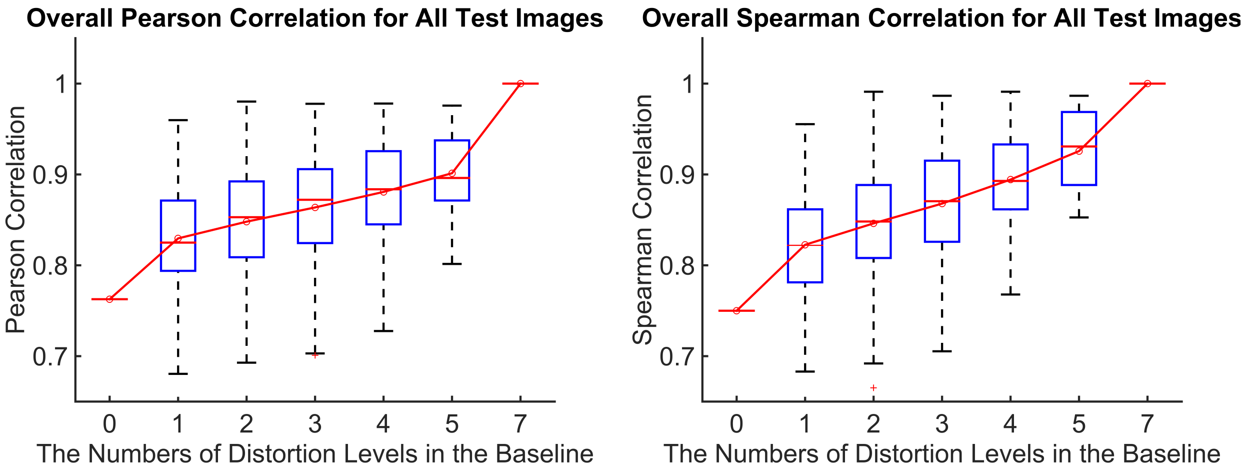


Figure 5. The box plots of Pearson (left) and Spearman (right) correlation values over all test images, the variance in each category is assumed to stem from the differences of judgement criterion scales between human observers but not between their perceptions. The mean values of correlations are depicted with dots around the bars which stands for the median values. Since mean values are calculated over all test images, they can be regarded as image content independent. In each category, all distortion combinations share the same number of levels included in the baseline.The category 0 indicates the case that we apply only Z-score scaling without baseline adjustment to the raw ratings in two sessions respectively and determine the correlations between them. In category 7, all distortion levels are included in the baseline, then we apply the baseline adjustment to all ratings in both sessions. Since all distortion levels are adopted, the baseline adjusted Z-score scaling is equivalent to the non-adjusted Z-score scaling, then the correlation values are exactly ones.

possible distortion level combinations. In this process, we find out that for a certain distortion level combination, the plot of Spearman correlation appears to be approximately a discrete version of Pearson correlation; so, in the following discussions, we focus only on Pearson correlations. We generate the plot of mean Pearson correlations over all human observers versus all possible distortion level combinations for each test image, and two of them are presented in Figure 6.

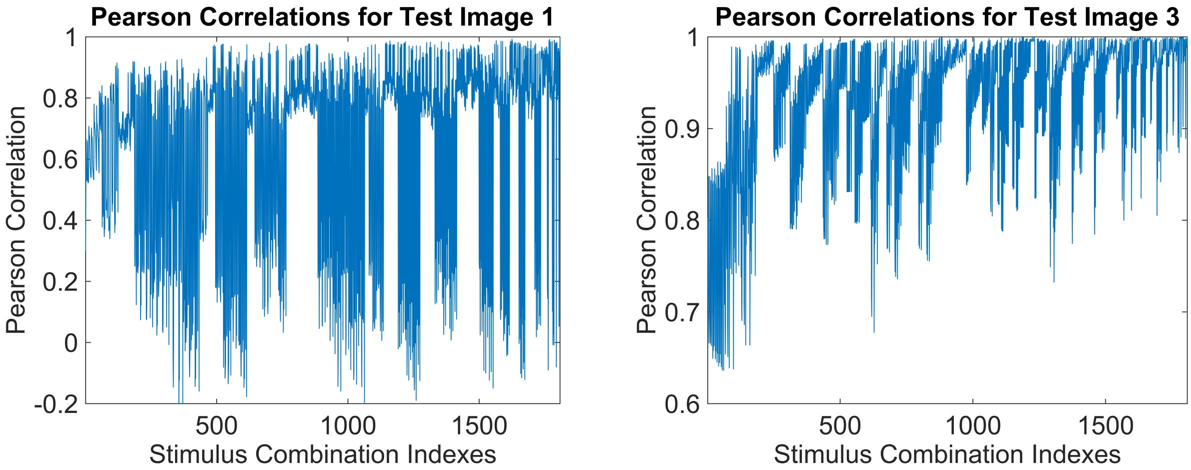


Figure 6. The plots of Pearson correlations versus all possible distortion level combination for test image 1 (left) and test image 2 (right). The X axis stand for the index of all possible distortion level combinations, it ranges from 1 to 1080 because we have 1808 possible combinations in total. The Y axis stand for the Pearson correlation values. For convenience to observe, we connect discrete correlation points with curves as they are illustrated in the figures. Consecutive correlation points for which the corresponding distortion level combinations may adopt similar distortion levels.

It is clear that the correlation values change dramatically depending on the actual choice of distortion combinations. For test image 1, one distortion combination (existing session: 1, 3, 5, 6, 7, simulated new session: 2, 4, 7, baseline: 7, the numbers stand for the indexes of distortion levels) gives Pearson correlation 0.856, while another distortion combination (existing session: 1, 4, 5, 6, 7, simulated new session: 2, 3, 7, baseline: 7) gives Pearson correlation -0.02. Both of them share the numbers of distortion levels in the baseline and two simulated rating sessions, and also they have exactly the same baseline, but the correlations are adversary. They only differ on the combinations of unique distortion levels. This observation gives us a hint that the most optimal distortion levels to be included in the baseline do not necessarily correspond to the highest possible correlation values, but they should produce the least correlation variance no matter how the unique distortion levels are combined. In other words, with the presence of an optimal baseline, the unique distortion level combinations have limited influence on the correlation variance. In some cases, with a non-optimal baseline, the correlation may happen to have the highest value among all, because the perceptual ratings are fuzzy in nature and the highest correlation might be caused by random rating noises.

This theory can be also supported by the mean and standard deviation of correlations represented in Table 5. Taking test image 3 into consideration, we represent the distortion combinations which have identical baselines with one column of data in the table. The mean and standard deviation are calculated with respect to their Pearson correlation values. First, we pay special attentions to the baselines which involve only one distortion level (first 7 columns on the first row). It is easy to see that the distortion levels can be ranked with respect to their corresponding standard deviations. In this case, the 4th baseline has the lowest standard deviation as well as the largest correlation mean. Among the baseline having two distortion levels, the one with distortion level 4 and 7 gives the minimum standard deviation as well as the highest correlation mean. Then we have similar observations on the cases (marked with bold texts) where three and more distortion levels are involved in the baselines. These baselines might not always give the best correlation values but they do give the ones close to the best. Meanwhile the standard deviations are always optimal which means with the presence of these baselines unique distortion levels have very limited influence on the correlation results. From this point of view, we can make a conclusion that the most optimal baselines must include the distortion levels which gives the lowest possible standard deviation on correlation results when only one of them is included in the baseline. These optimal distortion levels should be included as many as possible to improve the correlation results. The optimal baselines are always specified only to their corresponding image, since it is image content related.

In a typical situation of extending subjective experiments, we have only the raw ratings for existing image distortion levels but not the ones for additional distortion levels. In this case, we cannot reply on the correlations between different rating sessions to determine the optimal baselines. However, we can use each of the known image distortion levels as a baseline and calculate the adjusted Z-scores of the rest of ratings. Then, we determine the correlation between the adjusted Z-scores and the original non-adjusted Z-scores to find out which baseline is most optimal with respect to the method described in the previous paragraph. This is an approximation approach because in this context we implicitly make an assumption that the newly introduced image distortions have limited influence on scaling the existing ratings. In other words, mean and standard deviation of selected baseline in the new session are expected to be close enough to the ones in the existing session, and the number of additional distortion levels should be small. Ideally, we should introduce one additional image distortion at one time. If two or more are required, then the researcher should adopt them one by one in an iterative fashion. The validity of this constrain is supported by the fact the best correlations are always associated with the cases that only one additional image distortion is introduced at a time in the experiments of this paper.

4. CONCLUSION AND FUTURE WORKS

In this paper, we conduct a study to verify and evaluate the baseline adjustment method regarding extending subjective experiments from one existing session to a new session. The experimental results suggest that the baseline adjustment method works effectively because both Pearson and Spearman correlations give high values once the optiaml baseline is specified. We identify the most optimal baseline should includes the combinations of image distortion levels that produce the minimum standard deviation of the mean adjusted Z-scores over all human observers in the existing rating session. We demonstrate that it is possible to reduce the number of image distortion levels included in the baseline, however the trade off is to lose the confidence of Z-score correlations between the two rating sessions. This is our first tempt to address such a research problem. Since we are using the perceptual data collected from a small subjective experiment to simulate real scenarios, the type and number of test images, image distortions, and human observers are limited. In the coming future, we should extend this research with a larger experiment to claim statistic significance on the outcomes.

ACKNOWLEDGMENTS

We would like to acknowledge the contributions of a large number of volunteers in the subjective experiments. This work is part of the HyPerCept research program funded by the Research Council of Norway.

REFERENCES

[1] Pedersen, M., Bonnier, N., Hardeberg, J. Y., and Albregtsen, F., “Attributes of Image Quality for Color Prints,” *Journal of Electronic Imaging* **19**, 011016–1–011016–13 (Jan. 2010).

[2] Pedersen, M., Bonnier, N., Hardeberg, J. Y., and Albregtsen, F., “Attributes of a New Image Quality Model for Color Prints,” in [*17th Color Imaging Conference*], 204–209, IS&T and SID, Albuquerque, USA (2009).

[3] Teunissen, K., *Flat Panel Display Characterization - A Perceptual Approach*, phd thesis, Eindhoven University of Technology (2009).

[4] Hung, P. S. and Guan, S. S., “A Research on The Visual Assessment Methods for Evaluating The Quality of Motion Images Displayed at LCD,” *Journal of Science and Technology* **16**(2), 153–164 (2007).

[5] Osberger, W., *Perceptual Vision Models for Picture Quality Assessment and Compression Applications*, phd thesis, Queensland University of Technology (1999).

[6] Eckert, M. P. and Bradley, A. P., “Perceptual Quality Metrics Applied to Still Image Compression,” *Journal of Signal Processing* **70**(1998), 177–200 (2007).

[7] Brémond, R., Tarel, J. P., Dumont, E., and Hautiere, N., “Vision Models for Image Quality Assessment: One Is not Enough,” *Journal of Electronic Imaging* **19**, 043004 (Oct. 2010).

[8] Williams, D. and Burns, P. D., “Measuring and Managing Digital Image Sharpening,” in [*IS&T 2008 Archiving Conference*], 89–93 (2008).

[9] Engeldrum, P. G., [*Psychometric Scaling, a toolkit for imaging systems development*], Imcotek Press Winchester USA (2000).

[10] Pedersen, M. and Hardeberg, J. Y., “Full-reference image quality metrics: Classification and evaluation,” *Foundations and Trends in Computer Graphics and Vision* **7**(1), 1–80 (2012).

[11] ISO, “ISO 20462-2 photography - psychophysical experimental methods to estimate image quality - part 2: Triplet comparison method,” (jul 2004).

[12] Toriumi, N., Takayama, J., Ohyama, S., and Kobayashi, A., “New method for incomplete paired comparison using the bmpc method,” in [*Proceedings of the 41st SICE Annual Conference*], **4**, 2339 – 2341 (2002).

[13] Morrissey, J. H., “New method for the assignment of psychometric scale values from incomplete paired comparisons,” *Jounal of the optical society of America* **45**, 373–378 (1955).

[14] Gulliksen, H., “A least squares solution for paired comparisons with incomplete data,” *Psychometrika* **21**, 125–134 (1956).

[15] Silverstein, D. A. and Farrell, J. E., “An efficient method for paired comparison,” *Journal of Electronic Imaging* **10**, 394–398 (2001).

[16] Imrey, P. B., Johnson, W. D., and Koch, G. G., “An incomplete contingency table approach to paired-comparison experiments,” *Journal of the American Statistical Association* **71**, 614–623 (sep. 1976).

[17] Van Der Linde, I., Rajashekar, U., Bovik, A. C., and Cormack, L. K., “DOVES: a database of visual eye movements.,” *Spatial vision* **22**(2), 161–177 (2009).

[18] Keelan, B. W. and Urabe, H., “ISO 20462, a psychophysical image qualitt measurement standard,” in [*Image Quality and System Performance*], Miyake, Y. and Rasmussen, D. R., eds., *Proceedings of SPIE* **5294**, 181–189 (Jan 2004).

[19] International Telecommunication Union, “Recommendation ITU-R BT.500-11: Methodology for the subjective assessment of the quality of television pictures,” tech. rep., International Telecommunication Union/ITU Radiocommunication Sector (2009).

[20] Larabi, C., “Subjective quality assessment of color images,” in [*The CREATE 2010 Conference*], Simone, G., Hardeberg, J. Y., and Farup, I., eds., 373–377 (Jun 2010). ISBN: 978-82-91313-46-7.

[21] Ngo, K. V., Storvik, J. J., Dokkeberg, C. A., Farup, I., and Pedersen, M., “Quickeval: A web application for psychometric scalingexperiments,” in [*Image Quality and System Performance XI*], Larabi, M.-C. and Triantaphillidou, S., eds., **9396**, 9396–24 (Feb. 2015).

[22] Brown, T. C., Daniel, T. C., Schroeder, H. W., and Brink, G. E., “Analysis of Ratings: A Guide to RMRATE,” tech. rep., Rocky Mountain Forest and Range Experiment Station (1990).

[23] Daniel, T. C. and Boster, R. S., “Measuring Landscape Ethetics: The Scenic Beauty Estimation Method,” tech. rep., Rocky Mountain Forest and Range Experiment Station (1976).

[24] Hull, R. B., Buhyoff, G. J., and Daniel, T. C., “Measurement of Scenic Beauty: The Law of Comparative Judgment and Scenic Beauty Estimation Procedures,” *Forest Science* **4**(30), 1084–1096 (1984).

[25] Brown, T. C. and Daniel, T. C., “Scaling of Ratings: Concepts and Methods,” tech. rep., Rocky Mountain Forest and Range Experiment Station (1990).

[26] Brown, T. C. and Daniel, T. C., “Landscape Aesthetics of Riparian Environments: Relationship of Flow Quantity to Scenic Quality Along a Wild and Scenic River,” *Water Resources Research* **27**, 1787–1795 (Aug. 1991).

[27] Hetherington, J., Daniel, T. C., and Brown, T. C., “Is motion more important than it sounds?: The medium of presentation in environment perception research,” *Journal of Environmental Psychology* **13**, 283–291 (Dec. 1993).

5. APPENDIX

Table 1. The Mean and Standard Deviation of Pearson Correlations for All Groups of Distortion Combinations

Baseline	1	2	3	**4**	5	6	7	1, 2	1, 3	1, 4	1, 5	1, 6	1, 7	2, 3
Mean	.758	.852	.859	**.975**	.942	.892	.957	.894	.916	.969	.923	.898	.952	.864
Std	.086	.099	.088	**.012**	.042	.066	.026	.078	.066	.024	.059	.076	.042	.088

Baseline	2, 4	2, 5	2, 6	2, 7	3, 4	3, 5	3, 6	3, 7	4, 5	4, 6	**4, 7**	5, 6	5, 7	6, 7
Mean	.972	.901	.881	.928	.970	.903	.897	.943	.974	.979	**.984**	.926	.959	.948
Std	.020	.069	.095	.055	.018	.068	.085	.045	.014	.011	**.007**	.055	.027	.038

Baseline	1,2,3	1,2,4	1,2,5	1,2,6	1,2,7	1,3,4	1,3,5	1,3,6	1,3,7	1,4,5	1,4,6	1,4,7	1,5,6
Mean	.910	.973	.900	.897	.939	.982	.926	.913	.950	.976	.980	.996	.921
Std	.072	.021	.064	.084	.056	.015	.058	.075	.046	.021	.016	.001	.065

Baseline	1,5,7	1,6,7	2,3,4	2,3,5	2,3,6	2,3,7	2,4,5	2,4,6	2,4,7	2,5,6	2,5,7	2,6,7	3,4,5
Mean	.959	.945	.965	.895	.887	.922	.970	.981	.989	.906	.940	.927	.970
Std	.038	.052	.020	.070	.091	.058	.018	.012	.008	.077	.046	.064	.018

Baseline	3,4,6	3,4,7	3,5,6	3,5,7	3,6,7	4,5,6	**4,5,7**	4,6,7	5,6,7
Mean	.980	.990	.912	.949	.939	.977	**.987**	.987	.953
Std	.011	.008	.072	.045	.054	.013	**.006**	.006	.035

Baseline	1,2,3,4	1,2,3,5	1,2,3,6	1,2,3,7	1,2,4,5	1,2,4,6	1,2,4,7	1,2,5,6	1,2,5,7	1,2,6,7
Mean	.982	.922	.908	.938	.978	.982	.997	.917	.948	.935
Std	.014	.063	.081	.060	.020	.014	.001	.072	.051	.067

Baseline	1,3,4,5	1,3,4,6	1,3,4,7	1,3,5,6	1,3,5,7	1,3,6,7	1,4,5,6	1,4,5,7	1,4,6,7	1,5,6,7
Mean	.098	.985	.998	.924	.955	.947	.981	.996	.997	.951
Std	.018	.013	0	.066	.045	.055	.017	.001	.001	.048

Baseline	2,3,4,5	2,3,4,6	2,3,4,7	2,3,5,6	2,3,5,7	2,3,6,7	2,4,5,6	2,4,5,7	2,4,6,7	2,5,6,7
Mean	.968	.979	.986	.906	.923	.9238	.981	.990	.992	.938
Std	.014	.009	.012	.074	.051	.067	.012	.008	.006	.055

Baseline	3,4,5,6	3,4,5,7	3,4,6,7	3,5,6,7	**4,5,6,7**
Mean	.979	.991	.992	.946	**.988**
Std	.013	.008	.006	.048	**.005**

Baseline	1,2,3,4,5	1,2,3,4,6	1,2,3,4,7	1,2,3,5,6	1,2,3,5,7	1,2,4,5,6	1,2,4,5,7	1,2,5,6,7	1,3,4,5,6
Mean	.982	.922	.908	.938	.978	.982	.997	.917	.948
Std	.014	.063	.081	.060	.020	.014	.001	.072	.051

Baseline	1,2,3,4,5	1,2,3,4,6	1,2,3,4,7	1,2,3,5,6	1,2,3,5,7	1,2,3,6,7	1,2,4,5,6	1,2,4,5,7	1,2,4,6,7
Mean	.983	.987	.997	.927	.944	.9349	.9846	.998	.998
Std	.015	.011	.002	.068	.070	.087	.021	.001	.001

Baseline	1,2,5,6,7	1,3,4,5,6	1,3,4,5,7	1,3,4,6,7	1,3,5,6,7	**1,4,5,6,7**	2,3,4,5,6	2,3,4,5,7	2,3,4,6,7
Mean	.944	.985	.998	.999	.950	**.997**	.977	.987	.990
Std	.076	.020	0	0	.069	**0**	.007	.012	.010

Baseline	2,3,5,6,7	2,4,5,6,7	3,4,5,6,7
Mean	.934	.993	.993
Std	.064	.007	.005

Part III

Appendices

Appendix A

Specifications

In the chapter, we present the specification sheets of the projectors, cameras, spectroradiometers and test charts used in the research.

A.1 Projectors

The SONY BRAVIA APL-AW15 projector (Figure A.1) and Mitsubishi XL9U projector (Figure A.2) were used in the experiments. Their specifications are presented in Table A.1 and Table A.2 respectively.

A.1.1 SONY BRAVIA APL-AW15

SONY BRAVIA APL-AW15 (Figure A.1) is a portable three-LCD-chip projector, which was targeted at home-theater projection applications.

A.1.2 Mitsubishi XL9U

Mitsubishi XL9U (Figure A.2) is an affordable ultra-portable three-LCD-chip projector, which was targeted at the projection applications for business meetings or personal presentations.

A.2 Acquisition Devices

The Logitech QuickCam Pro 9000 webcam (Figure A.3), Nikon D200 DSLR camera (Figure A.4), Nikon D610 DSLR camera (Figure A.5), Hasselblad H3D II DSLR camera (Figure A.6), and Minolta CS1000 (Figure A.7) were used as the acquisition devices in the experiments.

Figure A.1: Sony BRAVIA VPL-AW15 Projector, reproduced from *www.crutchfield.com*

Table A.1: Specifications for Sony BRAVIA VPL-AW15 Projector [144]

Brightness	1,100 Lumens
Contrast	12,000:1
Auto Iris	Yes
Native Resolution	1280x720
Aspect Ratio	16:9 (HD)
Video Modes	720p, 1080i, 1080p/60, 1080p/24, 1080p/50, 480p, 480i
Data Modes	MAX 1920x1080
Digital Inputs	HDMI
Vertical Keystone Correction	Yes
HDBaseT	No
Max Power	265 Watts
Voltage	100V - 240V
Size	12 x 37 x 32
Weight	5.8 kg
Lamp Type	UHP
Lamp Wattage	165 Watts
Lamp Quantity	1
Display Type	2 cm 3 LCD
Standard Zoom Lens	1.60:1
Standard Lens Focus	Manual
Optional Lenses	No
Lens Shift	Vertical
Throw Dist (m)	1.9 - 5.9
Image Size (cm)	102 - 508
Throw Ratio (D:W)	1.36:1 - 2.19:1
Audible Noise	20.0 dB
Speakers	No
Digital Keystone	Vertical

The cameras' specifications are presented in Table A.3, Table A.4, Table A.5, and Table A.6 respectively.

A.2.1 Logitech QuickCam Pro 9000

Logitech QuickCam Pro 9000 (Figure A.3) was a very popular desktop webcam, which was designed to be an alternative to ordinary integrated laptop webcam. It provides relatively higher resolution and better image quality, and it incorporated a microphone and a speaker; so it is suitable for having online meetings. In our research, the camera was used in **Paper A**.

A.2.2 Nikon D200

Nikon D200 DSLR camera [**?**] is one of the classic DSLR cameras manufactured by Nikon. It has less imaging features comparing to the state-of-the-art Nikon DSLR cameras, but they share all the essential DSLR camera features. In our research, the camera was used in **Paper A** and **Paper B**.

A.2.3 Nikon 610

Nikon D610 DSLR camera [**?**] is one of the top rank DSLR cameras manufactured by Nikon. It has a lot of state-of-art imaging features. In our research, the camera was used in **Paper**

Table A.2: Specifications for Mitsubishi XL9U Projector [143]

Brightness	2,000 Lumens
Contrast	350:1
Auto Iris	No
Native Resolution	1024x768
Aspect Ratio	4:3 (XGA)
Video Modes	720p, 1080i, 525i, 525p, 576i, 576p, 625i, 625p, 1125i, 480p, 480i
Data Modes	MAX 1280x1024
Digital Inputs	No
Vertical Keystone Correction	No
HDBaseT	No
Max Power	280 Watts
Voltage	100V - 240V
Size	9 x 26 x 26
Weight	2.7 kg
Display Type	2 cm 3 LCD
Standard Zoom Lens	1.20:1
Standard Lens Focus	Manual
Optional Lenses	No
Lens Shift	No
Throw Dist (m)	1.4 - 7.6
Image Size (cm)	102 - 635
Throw Ratio (D:W)	1.49:1 - 1.81:1
Audible Noise	39.0 dB
Speakers	2.0 W Mono
Digital Zoom	Yes
Digital Keystone	Vertical

Table A.3: Logitech QuickCam Pro 9000 specifications [1]

Width	3.5 in
Depth	1.5 in
Height	1.6 in
Connectivity Technology	Wired
Total Pixels	2000000 pixels, 1920000 pixels
Optical Sensor Type	CMOS
Manufacturer	Logitech
Interfaces	1 x USB 2.0 - 4 pin USB Type A
Camera Type	Color, Color - fixed
Max Digital Video Resolution	1600 x 1200
Video Capture	640 x 480 @ 30 fps, 1600 x 1200 @ 30 fps
Microsoft Certifications	Certified for Windows Vista, Compatible with Windows 7
Image Sensor	2 MP CMOS, 2 MP, 1.9 MP
Battery Form Factor	none
Focal Length	3.7 mm
Lens Iris	F/2.0
Focus Adjustment	automatic
Min Focus Range	3.9 in
Computer Interface	USB 2.0

Figure A.2: Mitsubishi XL9U Projector, reproduced from *www.homecinesolutions.fr*

Figure A.3: Logitech QuickCam Pro 9000 webcam, reproduced from *www.techtail.org*

C, **Paper D**, **Paper E**, and **Paper F**.

A.2.4 Hasselblad H3D II

Hasselblad H3D II camera (Figure A.6) is a very high-end DSLR camera. It does not only have a very good linearity in the sensor response and a large imaging sensor array, but also it produces nice image quality. In our research, the camera was used in **Paper A**.

A.2.5 Minolta CS1000

Minolta CS1000 (Figure A.7) is a portable and well-designed high-end spectroradiometer to provide accurate measurement of the electromagnetic radiation at specific spatial locations in the visible spectrum. In our research, the camera was used in **Paper A**.

Table A.4: Nikon D200 DSLR camera specifications [2]

Effective pixels	10.2 million
Image sensor	RGB CCD, 23.6 x 15.8 mm, 10.92 million total pixels
Image size	L (3,872 x 2,592) / M (2,896 x 1,944) / S (1,936 x 1,296)
Sensitivity	ISO equivalency 100 to 1600 in steps of 1/3, 1/2, or 1 EV
Storage media	CompactFlash (CF) Card (Type I/II) and Microdrive
LCD monitor	2.5-in., 230,000-dot, low-temp. polysilicon TFT LCD
Exposure metering	Matrix, Center-Weighted and Spot
Exposure modes	P, S, A, and M
Interface	USB 2.0 (Hi-Speed): mass storage and PTP connectable
Power sources	Rechargeable Li-ion Battery EN-EL3e
Dimensions (W x H x D)	Approx. 147 x 113 x 74mm (5.8 x 4.4 x 2.9 in.)
Weight	Approx. 830g (1lb 13oz)

Table A.5: Nikon D610 DSLR camera specifications [3]

Effective pixels	24.3 million
Image sensor	35.9 x 24.0 mm CMOS sensor (Nikon FX format)
Image size	6,016 x 4,016 (L), 4,512 x 3,008 (M), 3,008 x 2,008 (S) 3,936 x 2,624 (L), 2,944 x 1,968 (M), 1,968 x 1,312 (S) 6,016 x 3,376 (L), 4,512 x 2,528 (M), 3,008 x 1,688 (S) 3,936 x 2,224 (L), 2,944 x 1,664 (M), 1,968 x 1,112 (S)
Sensitivity	ISO equivalency 100 to 1600 in steps of 1/3, 1/2, or 1 EV
Storage media	SD and UHS-I compliant SDHC and SDXC memory cards
LCD monitor	Monitor 8-cm (3.2-in.), approx. 921k-dot (VGA), TFT LCD
Exposure metering	Matrix, Center-weighted, Spot
Exposure modes	P, S, A, M, U1, U2
Interface	Hi-Speed USB
Power sources	One EN-EL15 Rechargeable Li-ion Battery
Dimensions (W x H x D)	Approx. 141 x 113 x 82 mm/ 5.6 x 4.4 x 3.2 in.
Weight	Qpprox. 760 g/1 lb 10.8 oz (camera body only)

Table A.6: Hasselblad H3D II DSLR camera specifications [?]

Field of View Crop Factor	0.8
Waterproof	No
Width	6 inch (153 mm)
Height	5.14 inch (131 mm)
Depth	8.35 inch (213 mm)
Weight	80.8 oz (2290 g)
Total Pixels	31 Megapixels
Optical Sensor Type	CCD
Color Depth	48 bit
Max Resolution	6496 x 4872
Sensor Dust Reduction	No
Viewfinder	Optical TTL (through the lens)
Supported Flash Memory	CompactFlash, CompactFlash Type II

Figure A.4: Nikon D200 DSLR camera, reproduced from *www.dpreview.com*

Figure A.5: Nikon D610 DSLR camera, reproduced from *www.dpreview.com*

Figure A.6: Hasselblad H3D II DSLR camera, reproduced from *www.camera-usermanual.com*

Figure A.7: Minolta CS1000 spectroradiometer, reproduced from *www.ueen.feec.vutbr.cz*

Appendix B

Test Charts

Several test charts were used for different purposes in our research. In advance of subjective experiments, Ishihara Color Plates - 38 Set [32] (Figure B.1) were shown to all invited observers in order to confirm that none of the observers had color deficiency difficulty. Meanwhile, the logMAR chart (Figure B.2) was also adopted to confirm that at a certain distance all observers were able to reach a certain visual acuity level. The camera resolution chart introduced by ISO 12233 standard [175] was incorporated into the camera calibration procedure. The purpose was to determine the best aperture setting for the current combination of camera body and its mounted lens. Such a test chart was used in both **Paper D** and **Paper E**.

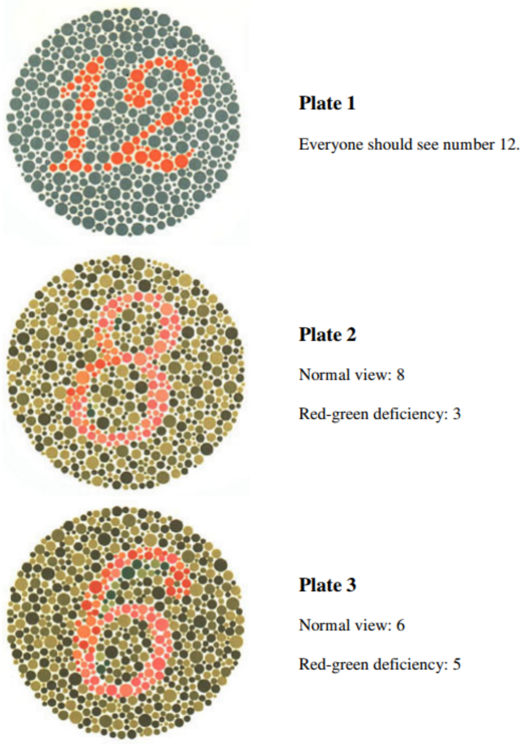


Figure B.1: Examples of Ishihara color plates for color deficiency test, reproduced from *unlimitedmemory.tripod.com*

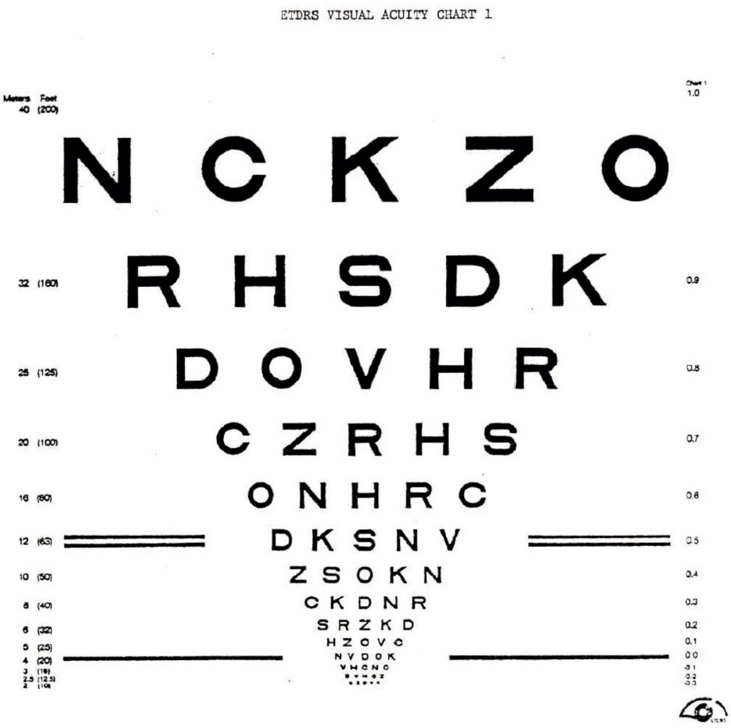


Figure B.2: The logMAR test chart for the visual acuity test, reproduced from *bjo.bmj.com*

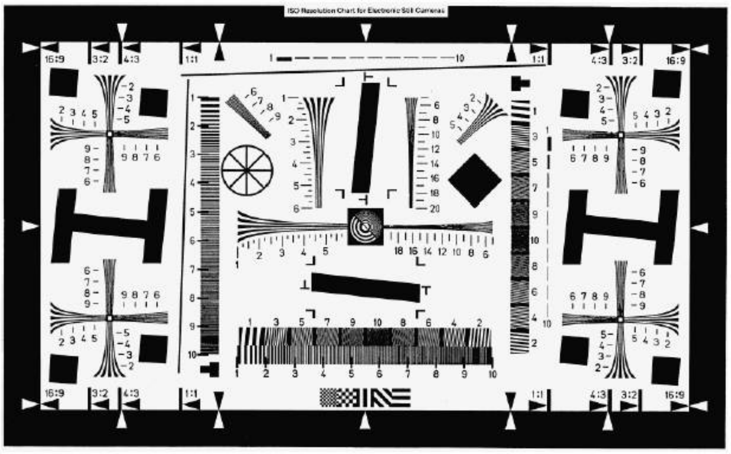


Figure B.3: The camera resolution test chart introduced by ISO 12233 standard [175]

Appendix C

Experimental Setup

In the experiments for **Paper C**, **Paper D**, and **Paper E**, we used a calibrated camera to capture all pixels on the projection screen in one shot, and the performance of state-of-art image quality metrics were evaluated with respect to the perceptual ratings collected from observers. We used the projectors to produce projections on a planar screen, which was naturally hanging on the ceiling. The projectors were put on a table placed in front of the projection screen about 3 m away with respect to the throw ratio (1.5) of the projector C.1. A remote controlling laptop was connected to the projector via either a VGA or HDMI cable in order to generate full screen projections. On the screen, the dimension of projection area was approximately 2×1.5 in meters. The camera was fixed on a tripod, which was placed right in front of the projections about 4 m away. Pictures were taken remotely with a software control on the laptop without physically touching the camera. The pictures were saved in raw format and rendered with aliasing minimization and zipper elimination demosaicing algorithm [117] without automatic vignetting correction, brightness adjustment, gamma correction and noise reduction etc. We selected test images from either The Colourlab Image Database: Image Quality database (CID:IQ) [106] (The full test image database CID:IQ is available for downloading from http://www.colourlab.no/cid.) or Kodak Photo CD PCD0992 [52] to generate multiple levels of image distortions. The selection criteria of test images was established based on the coverage of different image features such as hue, saturation, lightness, contrast, skin colors, sky colors, grass colors, size of neutral gray areas, color transition, fine details and text presence etc. The image selection criteria were established on expanding the coverage of these feature as much as possible in order to generalize the observation outcomes.

The projection displays were known to have a few typical viewing conditions, such as a home theater like dark room, a dimmed meeting room like environment with limited extra light, and an office like daylight environment with strong ambient light bouncing in the room. In this research, we setup the projector, screen and camera to simulate the first case in a controlled manner. All of the observers were confirmed to have normal visual acuity

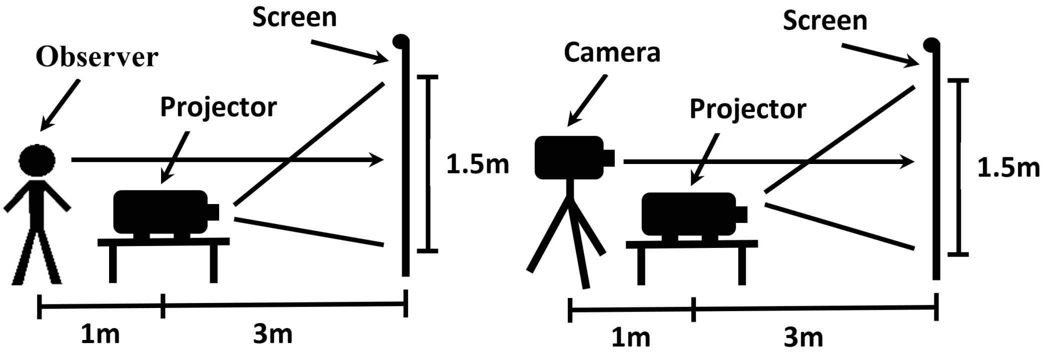


Figure C.1: The experimental setup for the research to simulate a home-theater-like dark room environment

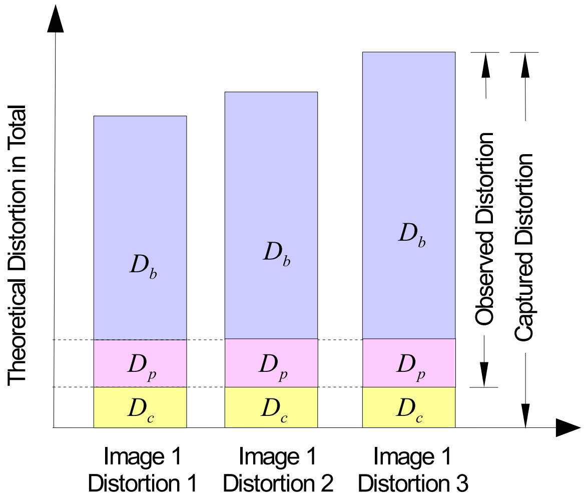


Figure C.2: Illustration of the distortions perceived by the observers and captued by the camera. In this figure, D_b, D_p, and D_c stand for the amount of image quality degradation introduced by the Gaussian blurring procedure, projection display, and the camera respectively. The values and summation of D_p and D_c are expected to be constant and they are much smaller than the value of D_b.

and color vision (the low room illumination and visual capability tests recommended in ITU-R BT.500-13 [146] and CIE 156:2004 [171]).

The digital still cameras were not originally designed to produce highly accurate absolute measurements, and they need to be calibrated in advance of the acquisition. It is known that a DSLR camera can be generally decomposed into optical subsystem, mechanical subsystem, electronic subsystem, and software subsystem. Each component in each subsystem has potential unwanted influence to the image quality of captured pictures. For example, the optical geometric distortion introduced by the misalignment in the lens array, the color aberration caused by the non-perfect demosaicing algorithm employed, and the analog-to-digital conversion noise magnification due to the thermal energy emitted from the internal imaging processor while pictures are taken etc. It is unlikely possible for us to eliminate all possible image quality degradation introduced by the projector and camera, however the majority of them, such as geometric distortions, vignetting effect, and non-linearity of camera sensor response. From another perspective, there are three sources of image quality degradation in our experiments: the Gaussian blurring procedure, projection display, and the camera. Let us denote the amount of image quality degradation introduced by them as D_b, D_p, and D_c, respectively (Figure C.2). Therefore, the values D_b should be image content dependent, and they are what the image quality metrics are expected to measure. Since we disabled all image enhancement features of the projector, then all blurred test images were displayed as they were. Then the values of D_p should be the equal for all image quality metrics. In a similar way of consideration, the camera was calibrated; so the values of D_c were also almost identical to all image quality metrics. So the image quality degradation introduced by the projector and camera had very limited influence on the performance of metrics.

We invited human observers (recommended in ITU-R BT.500-13 [146] and CIE 156:2004 [171]) to give perceptual ratings to the projected image distortions. The observers sit in the same location as the camera, which was placed approximately 4 m away from the screen. This was because we tried to avoid the potential image quality degradation due to the variation of observation position and viewing angle. The viewing condition was similar to a home theater like environment where the room is completely dark (recommended as low illuminant environment in ITU-R BT.500-13 [146]) and the visual angle from the projection boundaries to the principal axis of observation was about 15 degrees. The distorted images corresponding to the same test images were displayed in a randomized order to every observer.