Perceptual Quality Assessment of High Dynamic Range & Wide Colour Gamut Images and Video

by

Thilan Costa

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

Waterloo, Ontario, Canada, 2021

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner:	Michael S. Brown Professor, Dept. of Electrical Engineering and Computer Science, York University
Supervisor(s):	Vincent Gaudet Professor, Dept. of Electrical and Computer Engineering, Edward R. Vrscay Professor, Dept. of Applied Mathematics, University of Waterloo
Internal-External Member:	William Melek Professor, Dept. of Mechanical and Mechatronics Engineering, University of Waterloo
Internal Member:	Fakhri Karray Professor, Dept. of Electrical and Computer Engineering, University of Waterloo
Internal Member:	William Bishop Lecturer & Director of Admissions, Dept. of Elec. and Comp. Engineering, University of Waterloo

Author's Declaration

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

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

Abstract

While many successful algorithms have been proposed in the field of Image Quality Assessment (IQA) and Video Quality Assessment (VQA) through the years, High Dynamic Range (HDR) Wide Colour Gamut (WCG) technology and content have presented a unique set of challenges. Many IQA/VQA algorithms have ignored the colour in images and focused on performing quality assessment utilizing the luminance component of the image based on paradigms such structural similarity. However, HDR WCG technology emphasizes colour accuracy, creating a need for accurate assessment of colour quality in HDR content as well. In our work, we first study the performance of existing IQA/VQA methods in assessing the quality of HDR WCG content according to the existing IQA/VQA philosophy. We also check whether performance is affected by the use of constant luminance data as opposed to the commonly used non-constant luminance data in IQA/VQA applications. Then, we present the results of three colour difference measurement experiments performed using human subjects to study the performance of existing colour difference methods for HDR WCG content, as well as the effect of background luminance with respect to colour difference prediction. Our results show that existing methods are far from desirable in assessing colour difference in HDR WCG content. Therefore, we then utilize the data gathered from the colour difference tests to propose modifications to existing colour difference methods that significantly improve their performance for HDR WCG content. We then analyze the compatibility of the existing IQA/VQA philosophy with the requirements of IQA/VQA HDR WCG content. We demonstrate that the existing philosophy is inadequate in terms of the subjective testing methodology, as well as the operating paradigm of the IQA VQA algorithms. Finally, we propose a new IQA/VQA philosophy, and a new algorithm for HDR WCG IQA, followed by a UHD HDR WCG image database, subjective testing, and the performance results of the new proposed algorithm. The results indicate that the proposed algorithm performs successfully in HDR WCG IQA.

Acknowledgements

I would first and foremost like to give thanks to the most Holy Trinity, one God! I give thanks to Christ Jesus our Lord, especially under the title of the Infant Jesus of Prague. I thank thee, O Lord, for thy loving grace and blessings that made the completion of this thesis possible. I would also like to thank the Queen of Heaven, the Blessed Virgin Mary, for interceding with Christ on my behalf to obtain these graces, and also especially St. Joseph her most chaste spouse, St. Anthony of Padua, and St. Jude Thaddeus for joining her in praying for my intentions.

I would then like to thank my loving parents, my father Francis Costa and mother Subhadra Costa for all the love and care they have always given me. If not for their love and care, I would not be here and none of this would be possible. I would especially like to thank my mother who always encouraged me and offered guidance when things were difficult or unclear, and prayed many countless hours on my behalf to obtain the graces I needed to complete this thesis. Thank you always for tirelessly listening to my many problems and always providing guidance and advise. I must also thank my loving sister Sanduni Costa who was always generously willing to listen to any problems I had and also prayed on my behalf as well, and all this when she had very little time due to her own studies.

Last, but not least, I thank my supervisors Prof. Vincent Gaudet and Prof. Edward R. Vrscay, the PhD examination committee members Prof. William Bishop, Prof. Fakhri Karray, Prof. William Melek, and Prof. Michael S. Brown. My PhD journey was not an easy one, and these persons were patient and full of guidance as I completed my research work and thesis. I thank them all for taking the time to read my thesis, and all the suggestions and feedback they offered to produce the final version. I also thank Prof. Oleg Michailovich and Prof. Zhou Wang for being a part of my PhD at various times during my journey.

Dedication

I dedicate this thesis to Christ Jesus, our God and King, through his Blessed Mother Mary, the glorious and immaculate Queen of Heaven.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

With rapid development in display technologies, High Dynamic Range (HDR) Wide Colour Gamut (WCG) displays have started to become more available in the consumer market. A wide range of content providers are now already supporting or in the process of implementing support for content to take advantage of these display technologies. The downfall of the 3D TVs has also redirected the display technology market to invest heavily into HDR WCG content production, distribution and displays. HDR technology boasts a potential luminance range of 10,000 nits in peak brightness (though currently, most standard displays are around 1000 nits). The OLED technology has expanded the dynamic range due to its capabilities at displaying black levels as low as 0.0005 nits. All of these are in contrast to the standard dynamic range (SDR) displays that operate at only 100 to 200 nits peak luminance and much flexibility in regards to black levels.

Different quality assessment methods have been proposed specifically for HDR images and videos such as HDR-Visual Quality Metric (HDR-VQM) [29] and HDR-Visual Difference Prediction (HDR-VDP) [50], which a study in 2015 claimed performed well in comparison to other methods [34]. The study also found that HDR-VDP had a lengthy run-time as well. However, since this test was particularly performed for HDR/WCG images, their performance must be verified for video content. Furthermore, as indicated in the same study, these algorithms do not take colour information into consideration. Thus, there is much room for improvement for HDR image quality assessment.

The other additional factor driving HDR is WCG. HDR content producers claim to use a wider range of colours that were not available with SDR. In turn, the displays are also

required to support the wider range of colours. Traditionally, colour has not been at the forefront of image and video quality assessment. While there are several quality assessment methods (such as FSIMc [99]) that have considered colour, usually most quality assessment algorithms operate purely on the luma component. The algorithms that consider colour, usually do not consider how the colours interact with the background and merely compute differences by converting the colour data to a particular colour space in which colour difference computations can be performed. Since 1976, there have been several colour spaces that have been proposed to measure colour difference. CIELAB and ICtCp are examples of such colour spaces. These colour spaces also have different distance measures defined to compute the actual colour difference. The most recent study, performed for HDR WCG content by Dolby laboratories found that ICtCp performed best in comparison to existing colour spaces/measurements [69]. However, since ICtCp was proposed by Dolby laboratories, there has been no independent verification of the claims. Furthermore, these colour difference methods do not take into account the perception of colour under the context of different surrounding colours and lighting conditions. Therefore, there is much work that needs to be done in this regard.

1.2 Objectives

The goal of this research is to develop objective models that can predict the perceptual quality of HDR WCG content. To achieve the goal, this research attempts to understand the impact HDR and WCG has on the performance of state-of-the-art quality assessment algorithms. Since colour is also an important factor, this research will also try to understand the performance of colour difference methods and various colour spaces in regards to their accuracy/performance at measuring colour differences. We will then proceed to improve the existing colour spaces/distance measures for estimating colour differences. These methods will then be integrated into an image quality assessment algorithm designed for HDR WCG content.

1.3 Outline

The next section of the thesis describes the recent developments in the field of objective quality assessment and state-of-the-art HDR quality assessment models and algorithms. It also introduces the reader the background knowledge about colour as it pertains to display technology and discusses the related work in colour difference measurements. We study the

performance of the existing quality assessment algorithms on HDR WCG content according to the current paradigm followed by a study of the effect of using constant luminance as opposed non-constant luminance for improving the performance of HDR WCG.

We then study the performance of existing colour difference methods and colour spaces in predicting the colour difference in HDR WCG content, the effect of background luminance on colour perception, followed by a chapter on improvements to existing colour difference methods. Finally, we will demonstrate the incompatibility of the existing paradigm for proper and meaningful quality assessment of HDR WCG content, followed by our proposal for a new paradigm. We will then discuss the development of a novel HDR WCG image quality assessment (IQA) algorithm the formulation of a image database, subjective testing, followed by a performance evaluation of the proposed IQA algorithm.

1.4 Terminology

Unless specified otherwise, RGB and $YCbCr$ values in this thesis will be in UHD WCG.

1.5 List of Contributions

This thesis research makes the following contributions:

1. Detailed study of the performance of existing image/video quality assessment (IQA/VQA) algorithms for HDR WCG content, under the existing paradigm.
2. Study of the effect performance when using constant luminance data as input to existing IQA/VQA algorithms (as opposed to non-constant luminance, which is what is generally used).
3. Introduction of a novel and efficient test framework for performing colour difference testing for display based applications.
4. Comprehensive study of the performance of existing colour difference methods and colour spaces in evaluating colour difference in HDR WCG content.
5. Study of the effect of background luminance on colour difference perception.

6. Development of improvements based on existing colour difference methods and colour spaces that significantly improve the performance at HDR WCG colour difference prediction.

7. Discussion of the incompatibilities of the current paradigm with the requirements for HDR WCG quality assessment, and the proposal of a novel paradigm of IQA and VQA.

8. Development of requirements for a quality assessment algorithm and subjective testing process to satisfy this new paradigm.

9. Proposal of a novel subjective testing methodology to satisfy the proposed paradigm.

10. Development of a novel HDR WCG IQA algorithm that satisfies the novel paradigm, the construction of a HDR image database, and performance evaluation of the proposed algorithm using subjective testing performed according to the new paradigm.

Chapter 2

Literature Review

2.1 High Dynamic Range (HDR)

At any given moment, the average human visual system (HVS) can instantaneously adjust to perceive up to about five orders of magnitude in luminance levels [61] (the HVS can perceive a much larger range of luminance from 10^{-2} cd/m^2 up to 10^8 cd/m^2 given sufficient time to adapt [15]). Ideally then, one would desire to develop a display system that is capable of accurately reproducing five orders of magnitude in luminance levels to the human viewer. The display technologies that are widely available today are identified as low dynamic range (LDR) displays, and the best displays in this range are only capable of displaying about three orders of magnitude in luminance levels. In contrast, HDR displays can support up to $10,000$ cd/m^2, which accommodates the full range supported by the HVS.

The ability of HDR displays to display such a wide range of luminance levels provides the possibility for image and video content creators to capture and include more details throughout the luminance range when producing video or image content. In standard (low dynamic range (LDR)) content production, details in the extreme bright regions or extremely dark regions are not present. With HDR displays, the content providers have motivation to capture and preserve these details since the HDR displays are capable of displaying it to the viewer. In fact, lack of detail in highlights and shadow regions will cause those regions to appear much flatter when viewed with a HDR display in comparison to a LDR display.

One method that was widely used for creating HDR content was to capture the same scene under different exposure levels. The camera would therefore take multiple shots of

each frame at varying exposure levels and them combine them to a single HDR frame that spans a wide luminance range. However, newer cameras such as Arri Alexa [2] uses sensor technology that is more sensitive to light, which enables the camera to capture the HDR images in one go, without the need to compile multiple images into one frame. This also avoids the artifacts like ghosting that might occur when object motion exists.

2.2 HDR Content Production & Distribution

The captured data using a HDR capable camera is stored as floating point RGB data in file formats such as the Academy Colour Encoding Scheme (ACES) [1]. This data has a linear relationship to the luminance levels present in the scene that was captured. Due to this relationship, such data is labelled as *scene-referred* content. The ACES format supports 16-bit floating point values per colour channel, and can store any value from anywhere within the CIE 1931 color space.

The content creator will view this content through a reference monitor that confines the colour space to match that of the target display or the capabilities of the reference monitor. This data would now be *display-referred*. After colour grading and production completion, the content will then be mastered into the suitable formats to be distributed. During the mastering process, the colour space is confined to a standard space that has been defined for HDR displays by the International Telecommunication Union (ITU). The ITU standard for HDR content (ITU BT 2100) at the time of this writing, recommends the content to be in the Rec 2020 colour space. The HDR content also requires more bits to represent the extra detail in comparison to the LDR content which used 8-bits. The ITU standard for HDR content is defined for 10-bit data as well as 12-bit data. Most displays and content today are limited to 10-bit.

Even with the extra number of bits, storing the HDR scene referred data directly into 10-bit or 12-bits is inefficient. This is because the HVS does not have a linear response to the increase in luminance. The non-linear response of the HVS also means that if one were to directly quantize the full range of available luminance to the number of available bits, the difference between two adjacent bit code values will be more pronounced to the HVS. The effect of noise in the smaller values will also be more pronounced in a direct quantization. To avoid such issues and maintain a perceptually uniform increase in luminance, the data must therefore be mapped using a non-linear function, typically referred to as the Gamma function. This process of mapping the scene referred linear data using a Gamma function was performed in the case of LDR content as well. While for LDR content, it was generally referred to as a Gamma function, the ITU defines the function as the Opto-Electrical

Transfer Function (OETF). This is because the OETF resembles the inverse relationship followed by the pixels on the display when translating the 10-bit or 12-bit coded voltage value into light. The inverse OETF is defined as the Electro-Optical Transfer Function (EOTF).

The ITU standard defines the Electro-Optical Transfer Functions (EOTF) to which the display manufacturers conform their displays. In the case of LDR content, the defined EOTF is similar to the characteristics of the Cathode Ray Tube (CRT) monitor EOTF. For HDR, there are two different EOTF and OETF pairs defined in the present HDR display parameter standard by ITU (BT 2100) for program exchange. Those are Perceptual Quantizer (PQ) and Hybrid Log Gamma (HLG). The PQ EOTF is defined as

$$\begin{aligned} F_D = EOTF[E'] &= 10000Y \\ Y &= \left(\frac{\max[(E'^{1/m_2} - c_1), 0]}{c_2 - c_3 E'^{1/m_2}} \right)^{1/m_1} \end{aligned} \tag{2.1}$$

where $0 \leq E' \leq 1$, and the PQ OETF is simply the inverse EOTF function defined as

$$\begin{aligned} EOTF^{-1}[F_D] &= \left(\frac{c_1 + c_2 Y^{m1}}{1 + c_3 Y^{m_1}} \right)_2^m \\ Y &= F_D/10000 \end{aligned} \tag{2.2}$$

where $m_1 = 2610/16384$, $m_2 = 2523/4096$, $c_1 = 3424/4096$, $c_2 = 2413/4096$, $c_3 = 2392/4096$ and E' is the non-linear RGB values obtained by applying the OETF on the scene referred data. The PQ EOTF and PQ OETF are shown in Figure 2.1 and Figure 2.2 respectively.

The HLG OETF is given by the piecewise function

$$E' = OETF[E] = \begin{cases} \sqrt{3E} & 0 \leq E \leq 1/12 \\ a \cdot \ln(12E - b) + c & 1/12 < E \leq 1 \end{cases} \tag{2.3}$$

and the HLG EOTF is then given by

$$E = OETF^{-1}[E'] = \begin{cases} E'^2/3 & 0 \leq E' \leq 1/2 \\ [\exp((E' - c)/a) + b]/12 & 1/2 < E \leq 1 \end{cases} \tag{2.4}$$

where $a = 0.17883277$, $b = 1 - 4a$, $c = 0.5 - a \cdot \ln(4a)$.

After passing through the OETF, the scene referred HDR data is regarded as perceptually uniform.

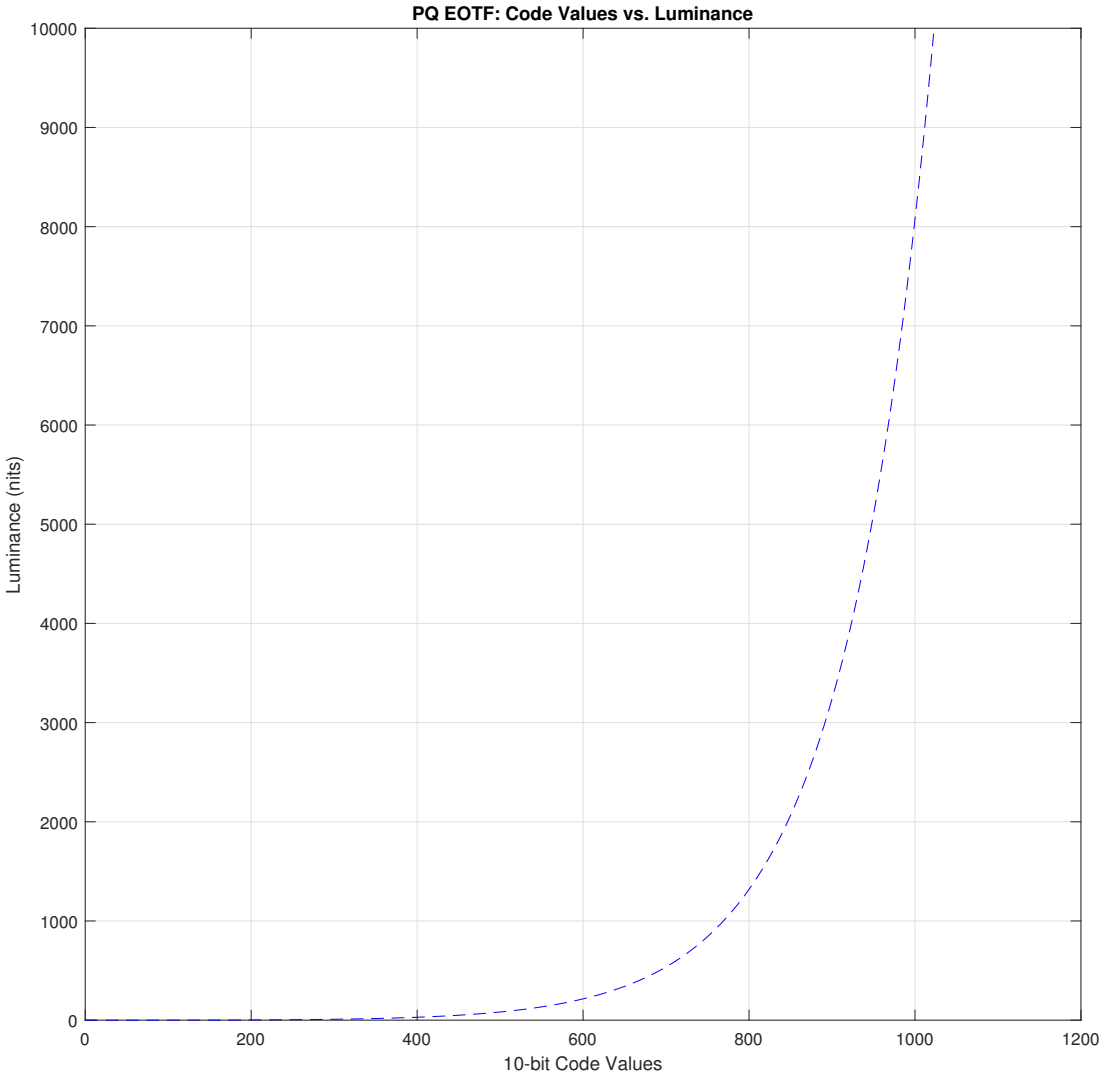


Figure 2.1: 10-bit PQ EOTF function

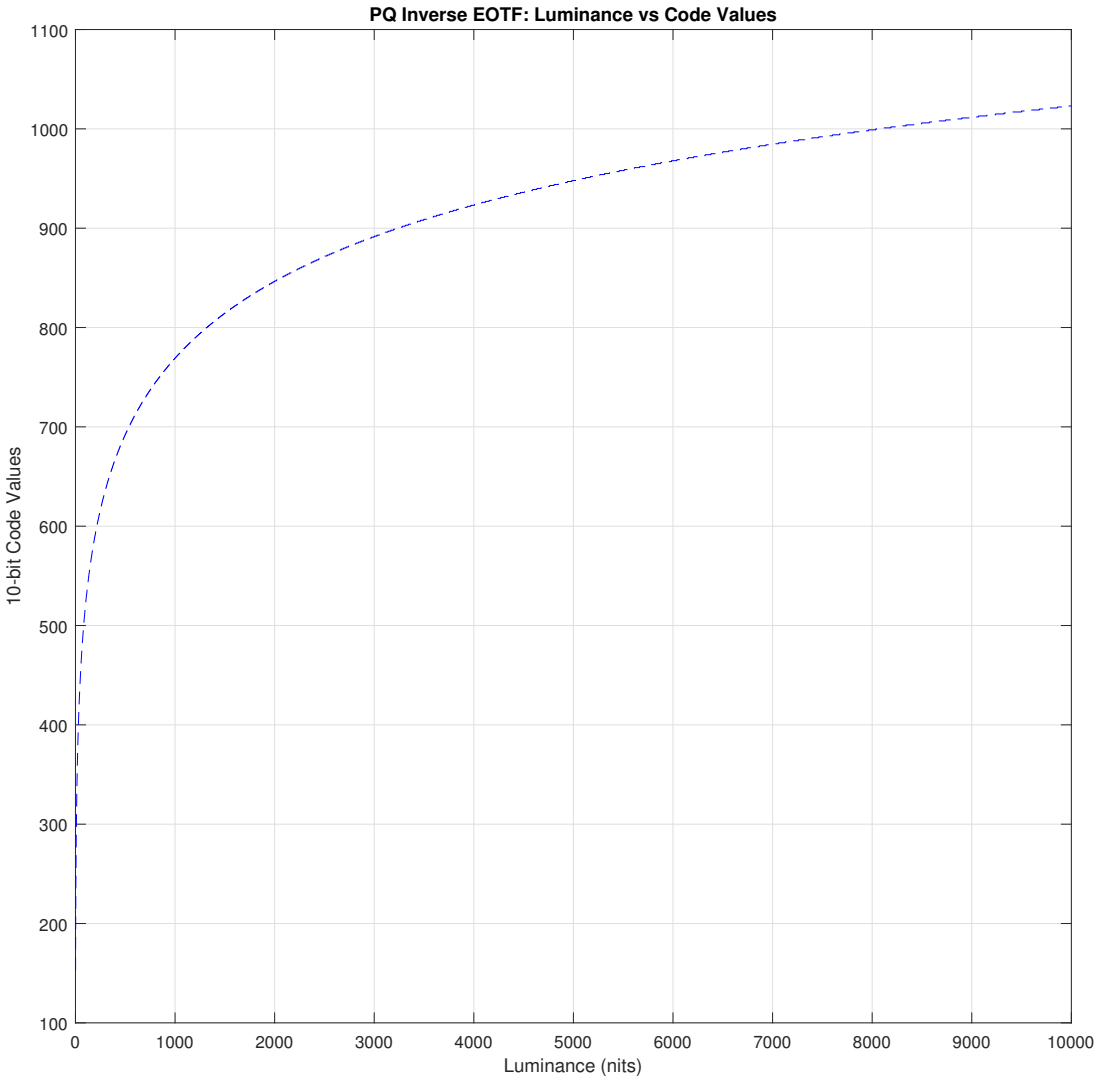


Figure 2.2: 10-bit PQ OETF function

A comparison of the PQ and HLG OETF is shown in Figure 2.3. It should be noted here that HLG operates under a relative luminance model compared to the PQ EOTF. The luminance input to both transfer functions are normalized. However, while PQ normalizes with a peak of 10,000 nits, HLG allows for a more flexible normalization depending on the brightness preference for the final scene. In this thesis, we will focus on PQ HDR WCG content, but the concepts and models developed will be equally applicable for HLG content.

2.3 Colour Images & Video

The HVS senses colour through three types of cone cell in the eye. Each type of cone cells has its unique spectral response which enables the eye to detect incident radiation wavelengths from 380 nm to 750 nm. This indicates that any colour can be represented as a weighted sum of the three spectral responses and is referred to as *tristimulus reproduction* [67]. As a result, in colour images and video, data for each pixel consists of three values such as $\{R, G, B\}$, $\{Y, Cb, Cr\}$, or $\{ICtCp\}$, which are sets of tristimulus values.

In 1931, the Commission Internationale de L'Eclairage (CIE) standardized a set of experimentally defined weighting curves for mapping the spectral power distribution to tristimulus values for describing colour. The curves are known as the CIE colour matching functions (CMFs). The CIE CMF functions denoted $\bar{x}(\lambda)$, $\bar{y}(\lambda)$, and $\bar{z}(\lambda)$ are shown in Figure 2.4. The spectral sensitivity of image/video capturing devices are required to conform to these curves identified as the *Maxwell-Ives-Luther criterion* [67]. Since the CMF's model the perception of the HVS, a capturing device that satisfies the *Maxwell-Ives-Luther criterion* is desired toward the end of obtaining a reproduction by the camera that matches the HVS reproduction of a scene.

The sample spectral power distribution of the output of the camera sensor is multiplied by the matrix consisting of the three sampled colour matching functions to produce the *X,Y,Z tristimulus* values. According to *Grassman's Third Law* [67] on additive colour mixtures, any other set of three colour components one can produce by a non-linear combination of X,Y, and Z (e.g., $\{R, G, B\}$, $\{Y, Cb, Cr\}$, and $\{I, Ct, Cp\}$) are also a set of tristimulus values. CIE also standardized a process for normalizing the XYZ tristimulus values to obtain chromaticity values x, y and z (though z is redundant from the definition). The x, y and z values are defined as

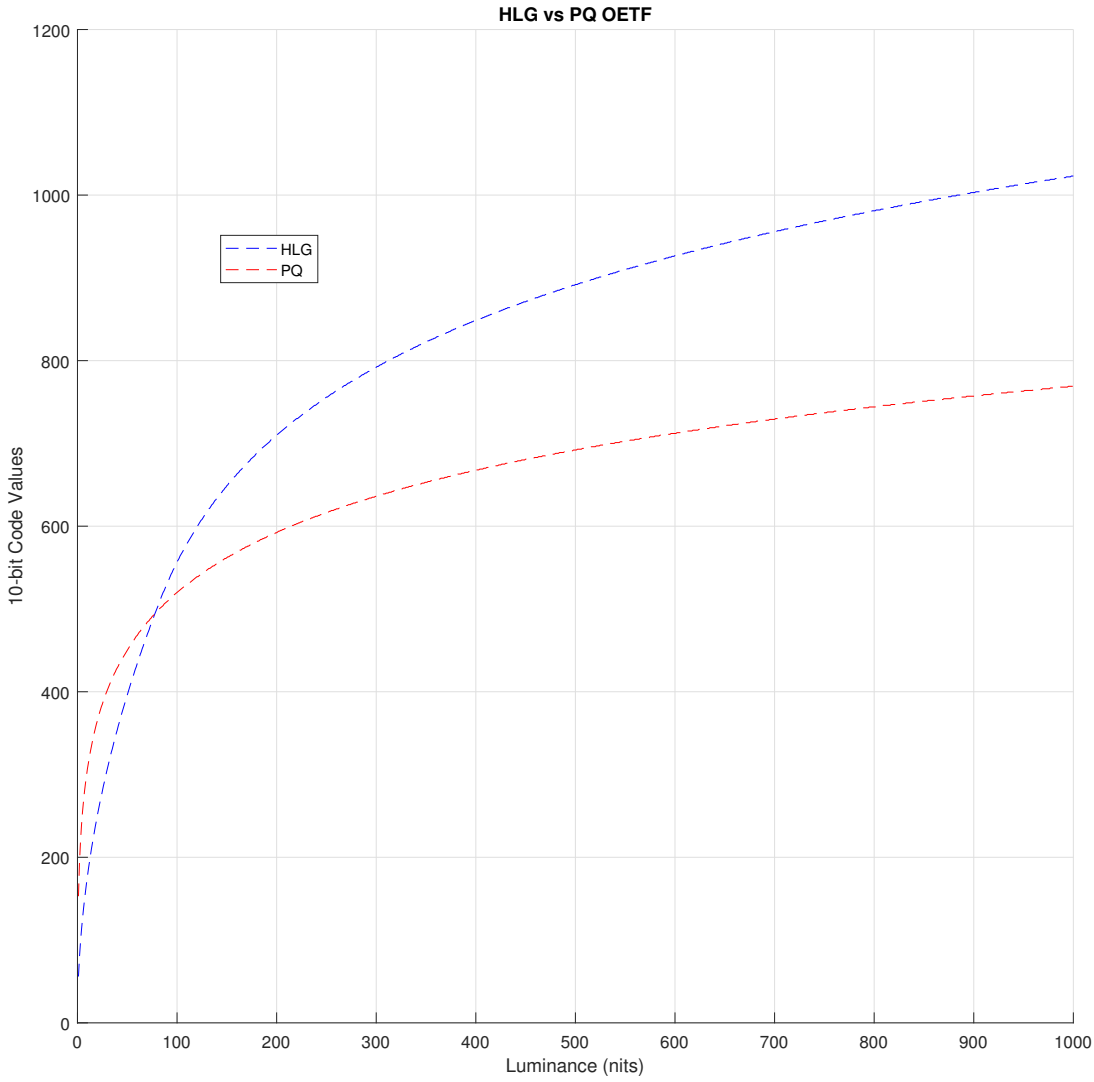


Figure 2.3: PQ vs HLG for 1000 nits peak

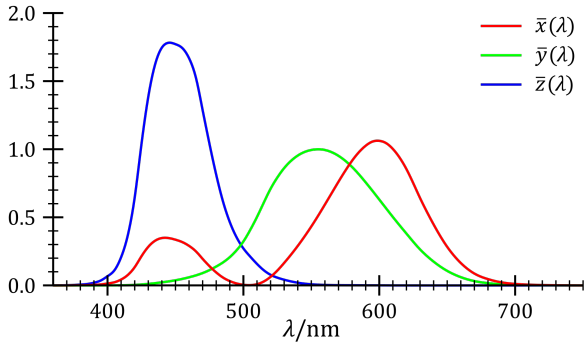


Figure 2.4: CIE colour matching functions [89]

$$\begin{aligned} x &= \frac{X}{X+Y+Z} \\ y &= \frac{Y}{X+Y+Z} \\ z &= 1-x-y. \end{aligned} \tag{2.5}$$

The $[x, y]$ values can be visualized as shown in Figure 2.5. Note that the $[x, y]$ figure shows a projection of the full gamut of colours along the luminance axis. Therefore to convert from $[x, y]$ coordinate to X, Y, Z coordinates, one can pick a luminance Y for the chosen colour. The combination of $[x, y]$ coordinates and the corresponding luminance value Y defines the xyY colour space. X, Z can be computed from xyY as follows:-

$$\begin{aligned} X &= \frac{x}{y}Y \\ Z &= \frac{(1-x-y)}{y}Y. \end{aligned} \tag{2.6}$$

If one were to have a monochromatic colour signal sweeping from 380 nm to 700 nm, the colours change across the boundary of the inverted-U shape in the xy chromaticity diagram (this is shown in Figure 2.5 and the wavelengths are indicated in micro-meters (μm).

As one may infer from Figure 2.5, there is no localized well defined co-ordinate for the colour white. If one were to define white as a signal with a uniform spectral power

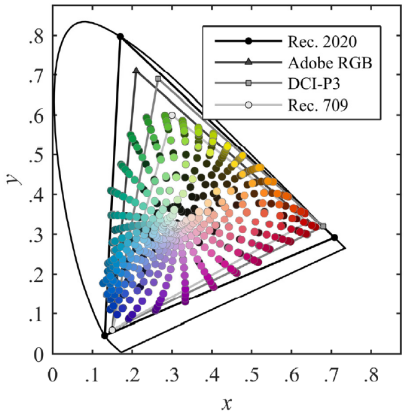


Figure 2.5: BT 2020, BT 709, DCI-P3, and Adobe RGB colour primaries in $[x, y]$ colour space [52]

distribution throughout the visible spectrum, then the white point in the xy chromaticity diagram will be located at $[1/3, 1/3]$. But in commercial and video and image production, the content producers usually prefer the spectral power distribution of white to approximate daylight, and is defined by CIE as the *illuminant* D_{65}. This white point is located in the xy chromaticity diagram at $[0.31271, 0.32902]$.

2.4 Colour Primaries & Wide Colour Gamut (WCG)

While the CIE X, Y, Z tristimulus format can describe any colour in the visible range, for content production and distribution, the colour range available is limited (due to display technology). The available colours can be denoted as a triangle contained within the coloured region of the $[x, y]$ chromaticity diagram as shown in Figure 2.5. Each triangle is uniquely defined by the coordinates of the three corners, and the white point. These three co-ordinates are referred to as the primaries. Without knowledge of the colour primaries and white point of given content, it is not possible to reproduce the content accurately on a display.

The ITU standard for program exchange parameters specify the acceptable colour primaries. For high definition (HD) content, BT 709 defined the colour primaries, while for ultra high definition (UHD) and HDR content, BT 2020 defined the colour primaries. As seen from Figure 2.5, the triangle formed by the BT 2020 colour primaries encapsulate a considerably larger area in comparison to the triangle formed by the BT 709 colour pri-

maries. For the content producers (and displays), this translates to being able to display content with a wider colour gamut to the viewer by using BT 2020 colour primaries. The phrase "Wide Colour Gamut (WCG) " content refers to the content that was graded using the BT 2020 colour primaries.

Each vertex of the colour triangle corresponds to the location of the colours Red, Green and Blue, hence defining the RGB colour primaries. Since any value within the triangle defined by the primaries must be expressible by the RGB primaries, we can define a conversion matrix M such that

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = M \begin{bmatrix} R \\ G \\ B \end{bmatrix}, \tag{2.7}$$

where R, G,B in the above equation refer to the linear RGB values, and M would be proportional to the colour primaries. Thus, for the BT 2020 colour primaries where the $[x, y]$ chromaticies for red, green and blue are $[0.708, 0.292]$,$[0.170, 0.797]$,$[0.131, 0.046]$ respectively [6], we have M of the form

$$M = \begin{bmatrix} 0.708 & 0.170 & 0.131 \\ 0.292 & 0.797 & 0.046 \\ 0 & 0.033 & 0.823 \end{bmatrix} \begin{bmatrix} J_1 & 0 & 0 \\ 0 & J_2 & 0 \\ 0 & 0 & J_3 \end{bmatrix}, \tag{2.8}$$

where the third row entries of the first matrix above are simply calculated as $1 - x - y$.

Since when $R = G = B = 1$, we should obtain the X, Y, Z values for the white point by summing the rows of M. The luminance Y should be 1. Hence, for $R = G = B = 1$ and the *illuminant* D_{65} white point defined at $[0.31271, 0.32902]$ [6], we have the following relation given by Equation 2.9 below

$$\frac{1}{0.32902} \begin{bmatrix} 0.31271 \\ 0.32902 \\ 0.35827 \end{bmatrix} = \begin{bmatrix} 0.708 & 0.170 & 0.131 \\ 0.292 & 0.797 & 0.046 \\ 0 & 0.033 & 0.823 \end{bmatrix} \begin{bmatrix} J_1 \\ J_2 \\ J_3 \end{bmatrix}. \tag{2.9}$$

Thus, we have

$$M = \begin{bmatrix} 0.6371 & 0.1446 & 0.1688 \\ 0.2628 & 0.6780 & 0.0593 \\ 0 & 0.0281 & 1.0602 \end{bmatrix}. \tag{2.10}$$

The matrix M and its inverse can be then used to convert colour from RGB colour space defined by BT 2020 to XYZ space and vice-versa.

2.5 Colour Spaces & Measuring Colour Difference

For various purposes, there have been many different colour spaces that have been proposed in the past decades besides RGB and XYZ/xyY colour spaces introduced above. The $CIELAB$ colour space was proposed in 1976. The conversion from XYZ space to $CIELAB$ is defined as [44]

$$\begin{aligned} L^* &= 116 f\left(\frac{Y}{Y_n}\right) - 16 \\ a^* &= 500\left(f\left(\frac{X}{X_n}\right) - f\left(\frac{Y}{Y_n}\right)\right) \\ b^* &= 500\left(f\left(\frac{Y}{Y_n}\right) - f\left(\frac{Z}{Z_n}\right)\right), \end{aligned} \tag{2.11}$$

where f is defined as

$$f(t) = \begin{cases} \sqrt[3]{t} & t > \delta^3 \\ \frac{t}{\delta^2} + \frac{4}{29} & \text{otherwise} \end{cases} \tag{2.12}$$

and $\delta = 6/29$. The X_n, Y_n, $Z_n$values of change depend on the chosen white point. If the white point is D65, then $X_n = 0.9505$, $Y_n = 0.1$, $Z_n = 0.1089$.

The motivation for the development of the colour space was the need to have a *perceptually uniform* colour space. Perceptually uniformity can be said to hold in a colour space if the smallest change to the coded colour value is approximately equally perceptible across the range of the fully set of colour values [67]. If a colour space is perceptually uniform, then one can simply measure the perceived difference between two colours located in the colour space by simply calculating the distance between them. For $CIELAB$, the colour difference method known as DeltaE calculates the Euclidean distance in the $L^*a^*b^*$ space. For two colours $L_1^*a_1^*b_1^*$ and $L_2^*a_2^*b_2^*$, the deltaE value can be calcuated as

$$\Delta E = \sqrt{(L_1^* - L_2^*)^2 + (a_1^* - a_2^*)^2 + (b_1^* - b_2^*)^2} \tag{2.13}$$

However, DeltaE is not 100% perceptually uniform. To correct for the distance measurements due to this discrepancy, additional distance measures were proposed, the most recent being CIEDE2000 (DeltaE2000) which was proposed in 1994 [76].

While the $CIELAB$ colour difference method was widely used, there was a question of how well it would perform in regards to HDR content. The analysis performed by Dolby laboratories [69] claimed that DeltaE2000 was not accurate in low luminance levels. Dolby has also proposed their own colour space $ICtCp$ [36][69] that they claim is better in terms of perceptual uniformity. After conversion from XYZ to linear RGB, the conversion to $ICtCp$ can be performed as [6].

$$\begin{aligned} L &= (1688R + 2146G + 262B)/4096 \\ M &= (683R + 2951G + 462B)/4096 \\ S &= (99R + 309G + 3688B)/4096 \end{aligned} \tag{2.14}$$

Then $L', M'.S'$ is computed by applying the PQ or HLG $OETF$ function described in Equation 2.2 and 2.4, to L, M, S computed above. Then I, Ct, Cp are computed as

$$\begin{aligned} I &= (2048L' + 2048M')/4096 \\ C_t &= (6610L' - 13613M'G + 7003S')/4096 \\ C_p &= (17933L' - 17390M' - 543S')/4096 \end{aligned} \tag{2.15}$$

The distance measure $\Delta ICtCp$ for two colours I_1, Ct_1, Cp_1 and I_2, Ct_2, Cp_2 is defined as [69]

$$\Delta ICtCp = \sqrt{(I_1 - I_2)^2 + 0.25(Ct_1 - Ct_2)^2 + (Cp_1 - Cp_2)^2} \tag{2.16}$$

The volume generated by mapping the entire 10-bit RGB space into $ICtCp$ is shown in Figure 2.6 below. Figure 2.7 shows the result from projection of the volume along the I-axis onto the Ct-Cp plane.

Another colour space that is defined for HDR is $Y'CbCr$ [6]. $Y'CbCr$ separates the video/image into luma (Y') and two colour components (Cb and Cr). This separation does lend itself to subsampling of the chroma components of video and image based on the assumption that human beings are more sensitive to the luminance component of an image/video. The conversion from non-linear RGB (obtained after applying OETF to linear RGB values) can be converted to the $Y'CbCr$ space as described in Equation 2.17 below.

$$\begin{aligned} Y' &= 0.267R' + 0.6780G' + 0.0593B' \\ Cb' &= \frac{B' - Y'}{1.8814} \\ Cr' &= \frac{R' - Y'}{1.4746} \end{aligned} \tag{2.17}$$

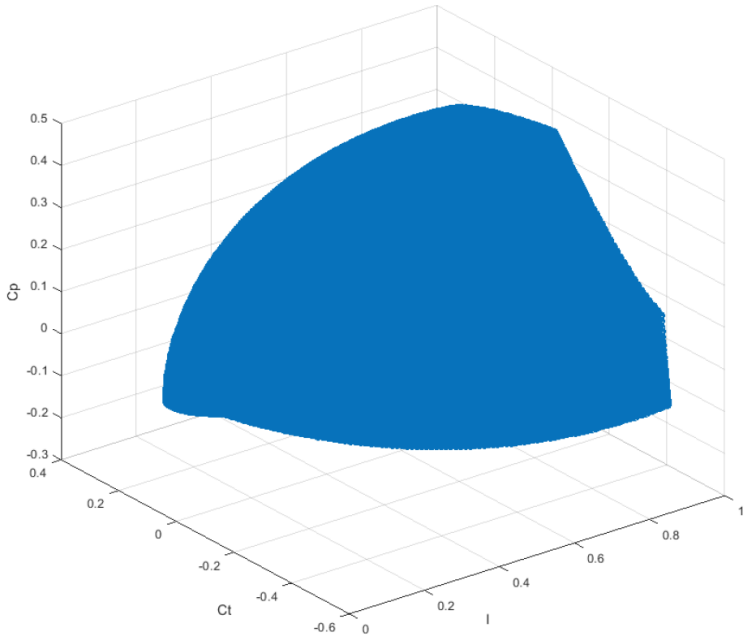


Figure 2.6: ICtCp colour volume containing the entire 10-bit RGB space

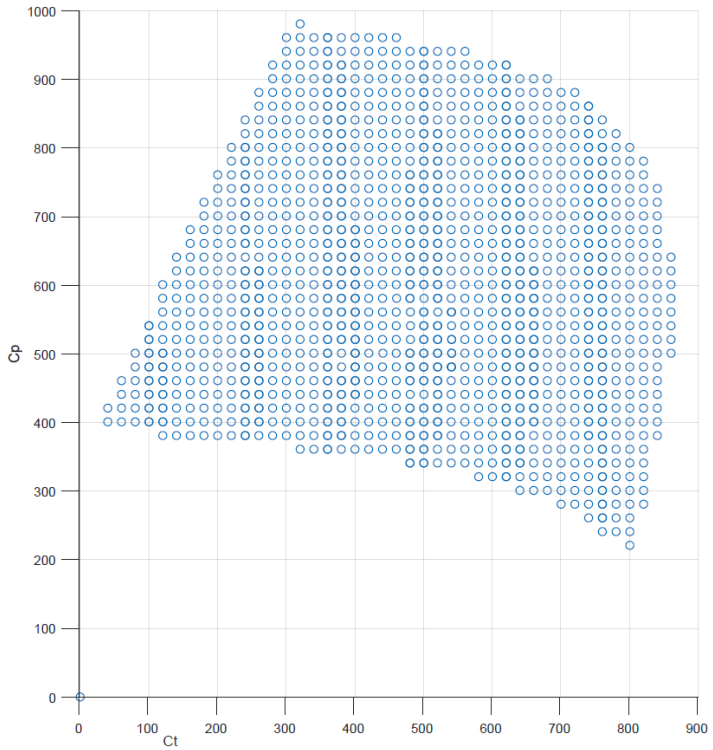


Figure 2.7: Projection of RGB volume along I onto the $Ct - Cp$ plane

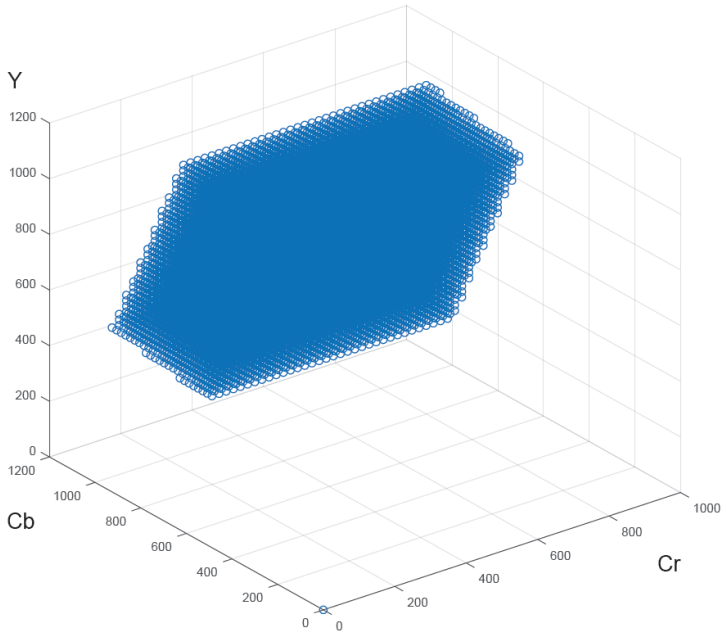


Figure 2.8: $Y'CbCr$ colour volume containing the entire 10-bit RGB space

where the prime indicates that it is the non-linear RGB and $Y'CbCr$ values. The $YCbCr$ volume containing the entire RGB space is shown in Figure 2.8.

In 2017, Pytlarz et al.[69] from Dolby laboratories performed a study on evaluating colour difference methods for HDR/WCG. Subjects were shown 4 squares placed in four quadrants of the display device, closer to the center. The background contained a noise pattern and the screen was periodically refreshed to prevent adaptation of the HVS. Three of the squares displayed a reference colour while one square contained a different colour. The subjects were asked to pick the square with the different colour. This was used to estimate the JND. The study found that CIEDE2000 performed considerably worse below luminances of 0.1 nits, but did perform quite well for colours above it and especially near neutral colours. $\Delta ICtCp$ was found to be the most accurate throughout the HDR WCG luminance range while $\Delta YCbCr$ (Euclidean distance between the two colours) and ΔRGB (Euclidean distance between the two colours) were found to perform worse than both CIEDE2000 and $\Delta ICtCp$ [69].

However, since $\Delta ICtCp$ was proposed by Dolby laboratories, the claim does require verification by an independent party. Moreover, the use of a relatively bright noise pattern

rather than a dark background under which colour differences are likely more noticeable, is also questionable. Another concern is that when designing the test colours for a given reference, one could unknowingly skew the result of the test by choosing a test colour that is within the JND, but not necessarily close to the JND. Such points can give an illusion of a uniform JND, while the colour space may not indeed be uniform.

2.6 Image & Video Quality Assessment Methods

When watching a video or observing an image on a display, the viewer can make a judgement on the quality of the content. Such form of quality assessment is referred to as subjective quality assessment. Subjective studies where participants evaluate the quality of images or video content have shown that humans tend to agree considerably on what constitutes a high quality image/video vs. a low quality image/video etc. These judgements of the viewer are important since content is produced to ultimately be enjoyed by the viewer. However, it is not feasible to have a subjective study for each new production and distribution of image or video content. Thus, there is a growing need for image & video quality assessment methods that can predict the subjective quality assessment results. Such methods are referred to as objective quality assessment methods. These methods provide a way for those who produce or distribute image & video content to quickly predict the subjective quality assessment and appropriately adjust the quality of the content.

Depending on the content that is required for the method to operate, the objective quality assessment methods can be divided into three categories as full reference (FR), reduced-reference (RR), and no-reference (NR) methods. FR objective quality assessment methods can be used when the reference image/video is available. The RR methods are used when the reference image/video is partly missing. The NR methods are used when the reference image/video is unavailable.

Over the years, there are many objective quality assessment methods that have been proposed. Naturally, the main desirable property of objective quality assessment methods is its high correlation with subjective scores. However, since content distributors and producers would desire to adjust the quality of the content based on the results of these methods, it is also desirable that these methods have a low computational complexity. Having certain mathematical properties such as convexity can also enable the content producers and distributors to easily optimize for quality. So when selecting an image/video quality assessment method, one would prefer the methods that satisfy all of the above qualities. Mean Square Error (MSE) and Structural SIMilarity (SSIM) index are two such

methods that satisfy the above qualities. This section will introduce these two methods briefly in the following subsections.

According to a comparison study of objective quality assessment metrics for HDR performed by Hanhart et al. [34], HDR-VDP-2 and HDR-VQM were the most reliable at predicting perceived quality for HDR content. While these objective quality assessment metrics do not satisfy the desirable qualities described previously, we have also dedicated subsections below to briefly describe these methods since they are the two methods that are said to best perform in evaluating the quality of HDR content (which is a main focus of the proposed research work).

2.6.1 Mean Square Error (MSE)

MSE is widely used in many engineering applications as a measure of distance between two signals. So it would seem reasonable to use MSE for measuring the difference between two images or videos. MSE for two images x and y of height L_1 and width L_2, where $x(i,j)$ and $y(i,j)$ represent the pixel value at location (i,j), can be computed as

$$MSE(x,y) = \frac{1}{L_1 L_2} \sum_{i=1}^{L1} \sum_{j=1}^{L_2} [x(i,j) - y(i,j)]^2 \tag{2.18}$$

As observable from the above equation, MSE is a differentiable convex function, and can be easily differentiated to derive an expression for finding the optimal. The function can also be shown to be energy preserving after an orthogonal transformation. The MSE is generally reported as the peak signal to noise ratio (PSNR) which can be computed as

$$\text{PSNR} = 20 \log_{10} \left(\frac{\text{L}}{\sqrt{\text{MSE}}} \right) \tag{2.19}$$

where L is the dynamic range of the pixel value. For an 8 bits/pixel image $L = 255$.

The problem with MSE is that its predicted scores have poor correlation with the scores given by human participants during image and video quality assessment subjective studies. Research has connected the failure of MSE to its inability to account for various aspects of the human visual system and human psychology [84] [85]. For an example, the human visual system does not merely consider an image pixel by pixel. The neighbouring pixels play an important role in how one perceives a pixel of an image. The general formulation of MSE cannot take such considerations into account. While MSE is still widely used, for the reasons mentioned above, it is not a good measure of image and video quality.

2.6.2 Structural Similarity (SSIM) Index

The Structual Similarity Index (SSIM) is based on the principle that the HVS exerts more weight toward distortions of structural components of an image when evaluating the quality of the image [85]. Given two images (or two frames of a video) x and y, SSIM consists of a product of three separate components that evaluate the luminance $l(x, y)$, contrast $c(x, y)$ and structural components $s(x, y)$ of two images as shown in Equation 2.20 given below [85].

$$\text{SSIM}(x, y) = l(x, y) \cdot c(x, y) \cdot s(x, y) \tag{2.20}$$

The luminance component $l(x, y)$, contrast $c(x, y)$ and structural components $s(x, y)$of the image is evaluated using 2.21, 2.22, and 2.23

$$l(x, y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \tag{2.21}$$

$$c(x, y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \tag{2.22}$$

$$s(x, y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \tag{2.23}$$

μ_x and μ_y are the mean intensity value of each image which can be computed by

$$\mu_x = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{2.24}$$

The standard deviation of each image, σ_x and σ_y, can be computed by

$$\sigma_x = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu_x)^2} \tag{2.25}$$

The cross-correlation of the two images σ_{xy} is computed as

$$\sigma_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y) \tag{2.26}$$

The constants C_1,C_2, and C_3 are included to avoid numerical instability when division by zero. Each coefficient can be computed as a simple function $(K_i \cdot L)^2$of the dynamic range L of the pixels in the image, where $K_i \ll 1$ where $i = 1, 2, 3$ corresponding to each coefficient C_1, C_2, C_3 respectively. Finally, by choosing $C_3 = C_2/2$, the combined product for the evaluating SSIM can be expressed as

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{2.27}$$

2.6.3 HDR-VDP-2

HDR-VDP-2 (HDR-Visual Difference Prediction-2) is a proposed image and video quality assessment method by Mantiuk et al. [50] in 2011, and was found to be the best performer in predicting quality score in a comparison study on objective quality assessment of HDR metrics performed by Hanhart et al. [34] in 2015.

The HDR-VDP-2 model operates by estimating the probability that the HVS would detect the difference between the reference and distorted images. The authors claim that their visual model takes into account the intra-occular light scatter, photo-receptor spectral sensitivities, separate rod and cone pathways, contrast sensitivity across the visible luminance range, intra-channel contrast masking, inter-channel contrast masking (masking here refers to a destructive interaction between two or more stimuli [38]) and spatial integration. The authors also state that no colour differences are taken into account in this version. A block diagram of the algorithm is shown in Figure 2.9.

To mimic the image decomposition that happens within the HVS, the algorithm first uses a steerable pyramid method proposed in [79] to decompose the image in terms of spatial frequency and orientation. Then the algorithm derives the threshold contrast in each of the spatial frequency and orientation channels which is then used to estimate the visual masking effect in each of the channels. A weighted sum of the effects of self-masking, masking across orientations, and masking from adjacent frequencies is used to compute the total masking effect. Then a psychometric function is used to translate this value into a probability value, and a steerable pyramid reconstruction transformation is used to generate an image where each pixel represents the probability that a difference in the reference and test image at that pixel, will be detected by the HVS. Under the assumption that a viewer will give equal attention to all regions of the image, a single probability for the entire image is considered to be the maximum probability value in the generated map. The authors also provide a weighted sum where the weights were fine

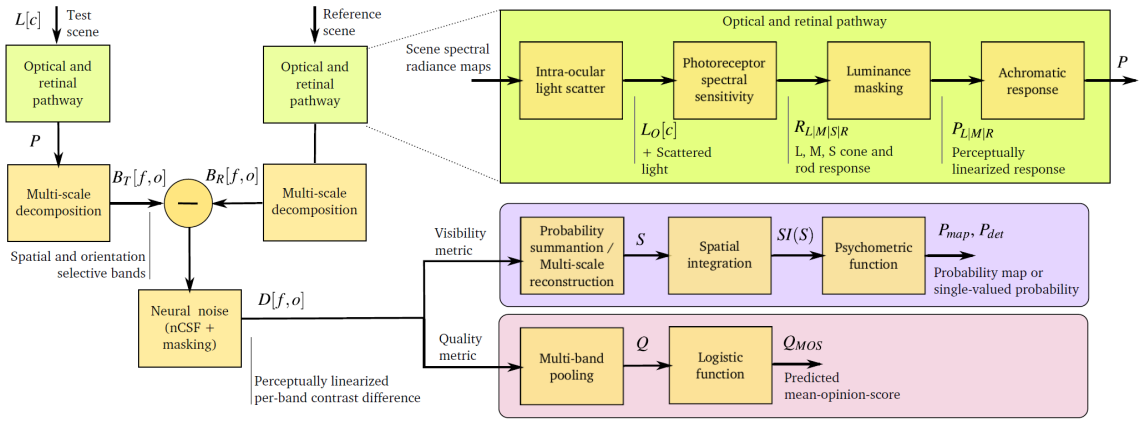


Figure 2.9: Block diagram of the HDR-VDP-2 algorithm [50]

tuned using available data to best match the desired outcome. The weighted sum of the probability values is then mapped to the MOS value using a logistical function that was trained to fit the LIVE database [35].

The advantage of this method is that the models used in the algorithm can take into account the viewing conditions and display capabilities under which the content is viewed, while also being able to handle the same dynamic range of luminance values that the HVS supports. Thus, it should in theory perform well in predicting the MOS values of HDR images and video.

On the other hand, the computation complexity of this method is heavily demanding. In addition, the use of experimental results to determine the best weights for the final weighting function in the algorithm, and the use of a trained logistical function at the end to map the resultant value to the MOS, make it hard to gauge the effectiveness of the algorithm. For these reasons, a better alternative is certainly desirable.

2.6.4 HDR-VQM

The HDR-VQM (HDR-Visual Quality Metric) method is based on decomposing the reference image (N_{src}) and distorted image (N_{hrc}) into subbands with the use of log-Gabor filters developed in [29]. A block diagram of of the HDR-VQM algorithm is given in Figure 2.10. The filter $H_{s,o}$ where s is the spatial scale index and o is the spatial orientation index of the filter, is given by

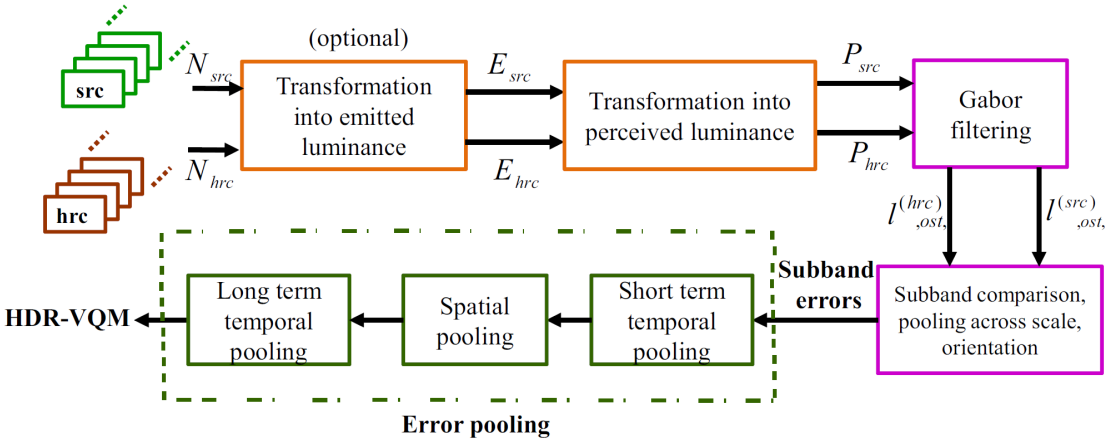


Figure 2.10: A block diagram of HDR VQM algorithm [62]

$$H_{s,o}(f,\theta) = \exp\left(-\frac{\log(f/f_s)^2}{2(\log(\sigma/f_s)^2}\right) \times \exp\left(-\frac{(\theta-\theta_0)^2}{2\sigma_0^2}\right), \tag{2.28}$$

where θ is the orientation, θ_0 is the centre orientation of the filter, f_s is the normalized centre frequency of the scale, and σ_s is the radial bandwidth and σ_0 the angular bandwidth. Using the filter in Equation 2.30, one obtains the subband values of the reference and distorted images, which are denoted as $l_{t,s,o}^{(src)}$ and $l_{t,s,o}^{(hrc)}$, respectively, where $s = 1, 2, \cdots, N_{scale}$ are the total number of scales, and $o = 1, 2, \cdots, N_{orient}$ are the total number of orientations, and $t = 1, 2, \cdots, F$ are the total number of frames (if it is a video sequence). The error in the subband values are then calculated using

$$Error_{t,s,o} = \frac{2 \cdot l_{t,s,o}^{(src)} \cdot l_{t,s,o}^{(hrc)} + k}{(l_{t,s,o}^{(src)})^2 + (l_{t,s,o}^{(hrc)})^2 + k}. \tag{2.29}$$

The total error between a reference image and distorted image can therefore be calculated as

$$Error_t = \frac{1}{N_{scale} \times N_{orient}} \sum_{s=1}^{N_{scale}} \sum_{o=1}^{N_{orient}} Error_{t,s,o}. \tag{2.30}$$

For the case of video, the total error can then be estimated by averaging the errors in each frame together.

The HDR-VQM algorithm does describe a methodology for transforming the reference image (denoted N_{src}) and distorted image (denoted N_{hrc}) into their respective emitted luminance values (E_{src} and E_{hrc} respectively). This transformation is performed to take into account the limitations of the display on which the content will be viewed. If the display peak luminance capability is below that of the peak luminance of the content, then this step will perform the necessary preprocessing to account for the disparity, and convert the data to a perceptually uniform space. This transformation is not necessary when evaluating content that is distributed since the present standard for distribution has already included the step to encode the data to a perceptually uniform space using PQ or HLG transfer functions.

Chapter 3

Perceptual Quality Assessment of UHD-HDR-WCG Videos[1]

There has been ample study of IQA and VQA algorithms and their performance in the past. However, the content eco-system at the time was mainly limited to resolutions of 1920×1080, with 8-bit dynamic range per colour pixel, and a smaller colour gamut defined by the ITU-R BT 709 colour primaries. While there has been recent work on constructing HDR databases [14, 63, 72, 54, 60, 12, 62] and subjective studies associated with these databases, the scope of the work is limited by one or more of the following problems.

1. The maximum spatial resolution of visual content is Full High Definition (FHD 1080p). UHD/4K content is lacking.

2. The colour gamut of the content and/or the displays used in these studies is usually BT.709, despite the growing popularity of WCG, such as DCI-P3 and BT.2020.

3. Most studies use content with a maximum temporal resolution of 30 frames per second (fps)

4. Fixed distortion levels such as a given bitrate in video encoding are used to create these databases, regardless of content complexity variations, leading to inadequate

[1]This is a paper that has been published in 2019 IEEE International Conference on Image Processing (ICIP) in 2019, available online: https://ieeexplore.ieee.org/document/8803179 [doi: 10.1109/ICIP.2019.8803179]. S. Athar, T. Costa, K. Zeng and Z. Wang, "Perceptual Quality Assessment of UHD-HDR-WCG Videos," 2019 IEEE International Conference on Image Processing (ICIP), 2019, pp. 1740-1744

perceptual separation across distortion levels and reduced overall effectiveness for objective benchmarking purposes

5. Only a limited number of Full Reference (FR) objective methods were evaluated, while state-of-the-art FR and No-Reference (NR) models were missing from these tests.

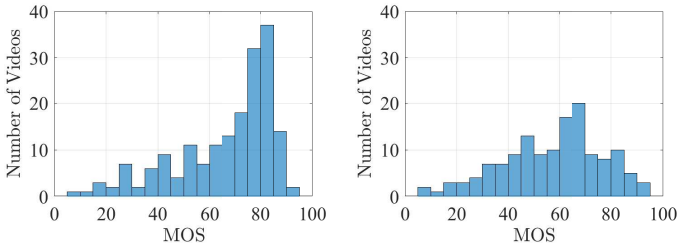
In this chapter, we present a new dataset namely the Waterloo UHD-HDR-WCG Database. It includes PQ-encoded 10-bit HDR content with UHD resolution, BT.2020 color gamut and two frame rates (24 fps and 59.94 fps). Adaptive bitrates are used to generate perceptually separated H.264 and HEVC compressed videos. We use a state-of-the-art professional 4K-HDR reference display, with a dedicated hardware pipeline, to construct the subjective experiment environment. We also present a novel data processing procedure to generate the Mean Opinion Scores (MOS). We then evaluate the performance of eleven FR and seven NR representative objective quality assessment methods and the results are provided. It should be noted that the IQA/VQA algorithms tested here do not consider colour components of the image/video data. These methods predict quality by analyzing the luminance channel of video/image data, while ignoring the chroma information.

3.1 Database and Hardware Setup

The Waterloo UHD-HDR-WCG database is created from 14 ten-second high-quality reference videos, all of which have Ultra High Definition (UHD) resolution (3840×2160), bit depth of 10 bits (Luma), YUV 4:2:0 chroma format, SMPTE ST 2084 (PQ) transfer function, and BT.2020 colour primaries to ensure Wide Colour Gamut (WCG) content. The frame rate is 59.94 fps and 24 fps for 9 and 5 reference videos, respectively. The focus of this work is to study the impact of compression on UHD-HDR-WCG content. Therefore, the reference videos are compressed by two encoders (H.264 and HEVC) at 5 bitrates each, resulting in 140 distorted videos. One way to construct VQA databases is to distort reference content at predefined distortion levels, that is, by using fixed bitrates for all contents. While this is a convenient approach, it does not lead to a uniform distribution of distorted content in the visual quality range. In order to uncover such issues, we first constructed a test HD-HDR database that had fixed distortion levels (bitrates) regardless of content and carried out a preliminary subjective test. The MOS histogram of this test database is shown in Fig. 3.1(a), where it can be seen that this database has a highly non-uniform

distribution of distorted content in the quality range. To address this, we selected the distortion levels for the Waterloo UHD-HDR-WCG database in a content-adaptive manner. Considering the visual quality range to be [0,100], where 100 is the highest quality, we first encoded the reference videos at multiple bitrates and used a state-of-the-art FR VQA method, SSIMplus [71], to select bitrates that led to distorted videos closest to predefined quality levels (94, 74, 54, 36, 18), followed by manual observation and bitrate adjustment to obtain perceptually separated distorted videos for each reference. It should be emphasized that SSIMplus was used to find a first approximation for the target bitrate for a quality level. The alternative would be to encode each video at a large number of bitrates and manually observe all of the encoded videos to arrive at the correct target bitrate for a desired quality level. This is infeasible due to the large amount of time that will be required to find target bitrates for a single video. The MOS histogram of the Waterloo UHD-HDR-WCG database is shown in Fig. 3.1(b), where it can be seen that the distorted content is more evenly distributed in the visual quality range for better perceptual quality separation.

Subjective experiments were carried out on a Canon DP-V2420 4K/UHD HDR Reference Display [3] which is a mastering monitor that is compatible with the Academy Colour Encoding System (ACES) [64] and supports both the SMPTE ST 2084 (PQ) and Hybrid Log Gamma (HLG) transfer functions. The display's peak luminance is 1000 cd/m^2, minimum black level is 0.005 cd/m^2, and screen size is 24 inch. To preserve the integrity of the content, we used the display's Quad 3G Serial Digital Interface (SDI) which supports a throughput of 12 Gbits/s. This fulfills the maximum throughput requirement of the high frame rate (59.94 fps) content of the database, which stands at around 11.12 Gbits/s. The workstation holding the database was connected through a Blackmagicdesign PCI Express Cable Kit to a Blackmagicdesign Ultrastudio 4K Extreme 3 [16] playback device. Here the single data stream is split into four streams that are connected to the Ultrastudio's SDI output interface, which is connected to the Reference Display. For smooth operation, the compressed videos were decoded and the entire database was stored in the YUV file format. It was ensured that all components in the video playback pipeline were capable of handling the high throughput requirements. Thus, the entire database (around 1.64 TBytes) was stored in a Samsung 2 TByte 960 Pro M.2 PCIe NVMe Solid State Drive (SSD) which is capable of sequential read speeds of up to 3.5 GBytes/s. The workstation is equipped with 32 GBytes of 3000 MHz DDR4 RAM to allow for storing an entire video in memory for quick transmission to the display. For optimal operation, customized video playback software was written by using the Blackmagicdesign Software Development Kit (SDK) which was invoked through MATLAB during subjective testing.



(a) Preliminary Database (b) Waterloo UHD-HDR-WCG Database

Figure 3.1: MOS distribution of Waterloo UHD-HDR-WCG database (b) in comparison with a preliminary database (a).

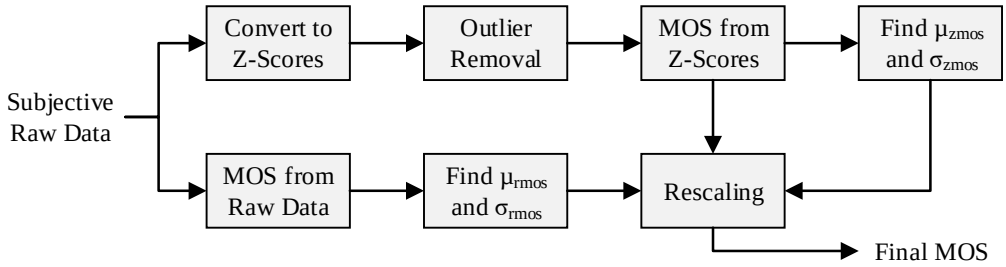


Figure 3.2: Process of Mean Opinion Scores (MOS) generation.

3.2 Subjective Study and Data Processing

The subjective study was conducted in the Image and Vision Computing (IVC) laboratory at the University of Waterloo in a dark room environment. A total of 51 subjects, including 29 males and 22 females aged between 18 and 35, took part in the study. Further, 8 subjects were regarded as experts since they worked in the area of VQA, while the remaining 43 were considered as naïve subjects. All the subjects had normal or corrected-to-normal vision, and were not colour-blind. The single-stimulus methodology with hidden reference [19] was used to carry out the study. The subjects were asked to evaluate the content at a viewing distance of approximately twice the screen height. The length of the study was around 80 minutes for each subject, which included two 30 minute rating sessions with a mandatory break in-between to reduce visual fatigue. After subjects viewed 10 s of content, the test display went black and a scoring GUI appeared on a secondary screen where subjects recorded their scores by using a sliding bar. The score range of 0 to 100 was

divided into intervals of 20 and labeled respectively as Bad, Poor, Fair, Good, Excellent, and subjects could select any integer value in this range. A higher score indicates better visual quality. To help orient the subjects with the test environment and to familiarize them with the quality range, a training session was carried out before the actual test which was composed of 5 distorted videos with varying distortion levels. The training videos had no overlap with the formal test videos and no instructions were given about which video should get what score.

Raw subjective scores are processed into final Mean Opinion Scores (MOS) by using the procedure shown in Fig. 3.2, where the goal is to take into account the variations in individual subject quality scales while maintaining the overall mean and variance of the raw scores. Since subjects may use the quality scale variably with respect to each other, raw scores per subject are first converted to Z-scores to account for these variations:

$$z_{ij} = \frac{s_{ij} - \mu_i}{\sigma_i} \tag{3.1}$$

where s_{ij} denotes the raw score assigned by subject i to video j, z_{ij} denotes the corresponding Z-score, μ_i and σ_i are respectively the mean and standard deviation of all the raw scores assigned by subject i in the test. Next, outlier detection and removal is performed as suggested in [19], which leads to the rejection of 9 subjects. The mean of the remaining Z-scores ($\hat{z_{ij}}$) for each video is computed which leads to the MOS in the Z domain (MOS_z), given as:

$$MOS_z = \frac{1}{N}\sum_{i=1}^{N} \hat{z_{ij}} \tag{3.2}$$

The MOS_z values lie in the range [-2.27, 1.43] and need to be rescaled. Although minmax normalization has been used to perform such rescaling [82, 97], we avoid using this technique since it can alter the distribution of data. Instead we use the following approach to generate the final MOS:

$$MOS = \sigma_{rmos}\left[\frac{MOS_z - \mu_{zmos}}{\sigma_{zmos}}\right] + \mu_{rmos} \tag{3.3}$$

where μ_{zmos} and σ_{zmos} are respectively the mean and standard deviation of MOS_z, whereas μ_{rmos} and σ_{rmos} are respectively the mean and standard deviation of the Mean Opinion Scores obtained from the raw subjective ratings. To evaluate the effectiveness of the

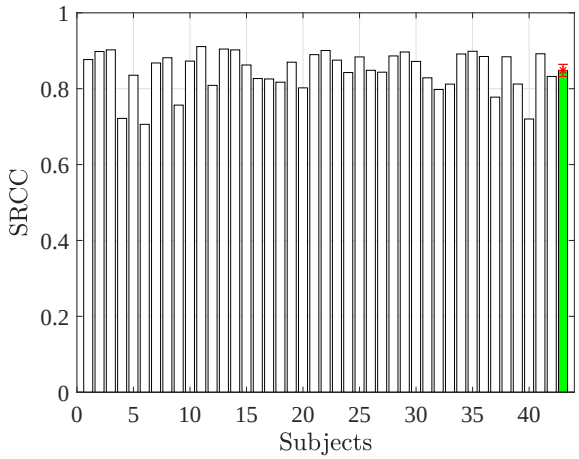


Figure 3.3: SRCC between MOS and individual subject scores. The right-most bar shows average subject performance with error bar.

final MOS, we compute its correlation with individual subjects' scores. Fig. 3.3 shows the Spearman Rank Correlation Coefficient (SRCC) of each valid subject with respect to MOS, where the right-most column shows the performance of an average subject with its corresponding error bar. It can be observed that there is a good degree of agreement between individual subjects and MOS.

3.3 Performance of Objective VQA Models

We tested the performance of representative VQA methods on the Waterloo UHD-HDR-WCG database. These include the FR methods: DSS [13], ESSIM [100], FSIM [99], GMSD [95], GSIM [39], HDRVDP2 [50], HDRVQM [62], IWSSIM [86], PSNR, SRSIM [98], and VIFDWT [73], and NR methods: BRISQUE [55], CORNIA [96], dipIQ [45], HOSA [94], LPSI [90], NIQE [56], and VMEON [41]. All but HDRVQM and VMEON are designed for objective Image Quality Assessment (IQA). These methods are applied to videos in a frame-by-frame manner and a final quality score is obtained by averaging across all frames. Among these methods, only HDRVDP2 and HDRVQM are designed specifically for HDR content, whereas all other methods have been designed and validated for Low Dynamic Range (LDR) content. PQ-encoding was substituted for the mapping used by HDRVQM (PU-encoding) to convert the linear light data into a perceptually uniform space. While PQ is one of the specifications recommended by ITU for mapping HDR data [6], PU is an older mapping that is not included in the recommendation. The

Table 3.1: Performance of Quality Assessment Algorithms. Best performing results in each category are in bold.

Category	Method	PLCC	SRCC	RMSE
FR	DSS [13]	0.7685	0.7456	12.3718
	ESSIM [100]	0.8512	0.8389	10.1485
	FSIM [99]	0.8693	0.8564	9.5568
	GMSD [95]	0.7366	0.7045	13.0781
	GSIM [39]	0.8596	0.8453	9.8812
	HDRVDP2 [50]	0.7035	0.6703	13.7423
	HDRVQM [62]	0.7783	0.7759	12.1428
	IWSSIM [86]	0.8088	0.7861	11.3730
	PSNR	0.5113	0.4615	16.6185
	SRSIM [98]	**0.8726**	**0.8630**	**9.4462**
	VIFDWT [73]	0.6809	0.6748	14.1612
NR	BRISQUE [55]	0.3622	0.3271	18.0241
	CORNIA [96]	**0.6497**	**0.6296**	**14.7003**
	dipIQ [45]	0.6192	0.5560	15.1845
	HOSA [94]	0.5379	0.5138	16.3015
	LPSI [90]	0.3941	0.3820	17.7718
	NIQE [56]	0.5286	0.4922	16.4152
	VMEON [41]	0.5776	0.5308	15.7845

performance of these methods was evaluated by using three evaluation metrics: Pearson Linear Correlation Coefficient (PLCC) and Root Mean Square Error (RMSE) to assess prediction accuracy and SRCC to assess prediction monotonicity [33]. A five-parameter logistic function [77] was used to perform non-linear mapping of objective scores to MOS before the computation of PLCC and RMSE. A better objective method should have higher PLCC and SRCC, and lower RMSE values. Table 3.1 shows the database-wide performance of the objective methods in terms of the three evaluation metrics. To draw statistically sound inferences about the performance of these methods, we carried out hypothesis testing on model prediction residuals (after non-linear mapping). Through the Jarque-Bera test [22] at the 5% significance level, we determined that the prediction residuals of all methods (except PSNR) are likely normally distributed. This enabled us to compare the model residuals through statistical significance testing by using the F-test [78]. The results are shown in Table 3.2, where "1", "–", or "0" mean that the method in the row is statistically (with 95% confidence) better, indistinguishable, or worse than the method in the column respectively.

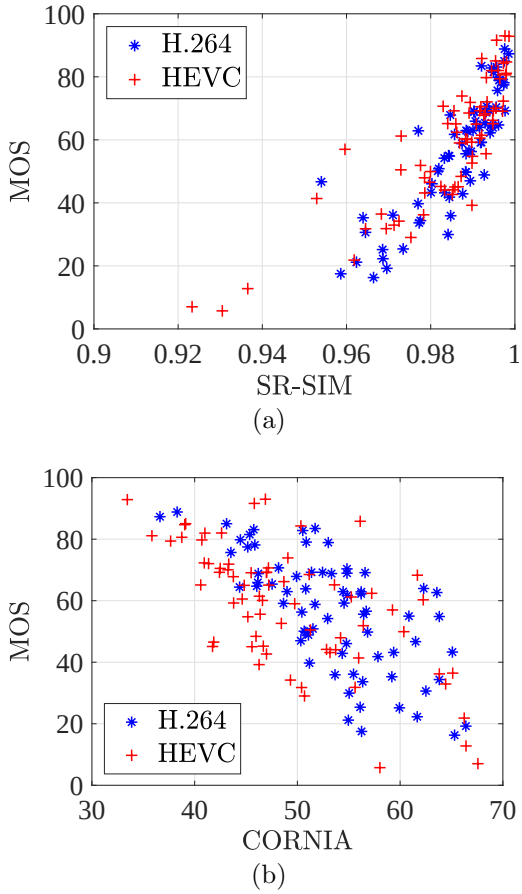


Figure 3.4: Scatter plots of best performing FR method SRSIM (a) and NR method CORNIA (b).

The LDR FR method SRSIM is found to be the top performer in terms of PLCC, SRCC and RMSE (Table 3.1). Other high performing FR methods include ESSIM, GSIM, and FSIM, where it can be seen from Table 3.2 that their performance is statistically indistinguishable from SRSIM. All of them inherit a similar formulation of signal fidelity measurement from SSIM [85]. Somewhat surprisingly, the HDR specific FR methods HDRVDP2 and HDRVQM do not offer superior performance on the Waterloo UHD-HDR-WCG database. This analysis suggests that LDR FR methods may be extended for HDR VQA, at least as far as compression is concerned, and potential further enhancement is possible by making HDR specific adjustments. Table 3.1 and Table 3.2 also indicate that all NR methods under testing perform rather inadequately. All these methods were designed for LDR content and most of them required some form of training, that was also done on LDR content. Clearly, there is significant room for improvement in HDR specific

design innovations. Similar performance evaluation results on FR- and NR-VQA models are observed when H.264 and HEVC compressed videos are evaluated separately. Fig. 3.4 shows the scatter plots for the top performing FR and NR methods, where the H.264 and HEVC compressed videos are separately identified.

Table 3.2: Statistical Significance Testing results for competing objective models on the Waterloo UHD-HDR-WCG database. A "1", "–", or "0" means that the method in the row is statistically (with 95% confidence) better, indistinguishable, or worse than the method in the column respectively. Legend: BRISQUE (m1), LPSI (m2), PSNR (m3), NIQE (m4), HOSA (m5), VMEON (m6), dipIQ (m7), CORNIA (m8), VIFDWT (m9), HDRVDP2 (m10), GMSD (m11), DSS (m12), HDRVQM (m13), IWSSIM (m14), ESSIM (m15), GSIM (m16), FSIM (m17), SRSIM (m18). FR methods are marked in **bold** while NR methods are marked in *italic*.

	m1	*m2*	**m3**	*m4*	*m5*	*m6*	*m7*	*m8*	**m9**	**m10**	**m11**	**m12**	**m13**	**m14**	**m15**	**m16**	**m17**	**m18**
m1	–	–	–	–	–	–	0	0	0	0	0	0	0	0	0	0	0	0
m2	–	–	–	–	–	–	0	0	0	0	0	0	0	0	0	0	0	0
m3	–	–	–	–	–	–	–	–	0	0	0	0	0	0	0	0	0	0
m4	–	–	–	–	–	–	–	–	0	0	0	0	0	0	0	0	0	0
m5	–	–	–	–	–	–	–	–	0	0	0	0	0	0	0	0	0	0
m6	–	–	–	–	–	–	–	–	–	–	0	0	0	0	0	0	0	0
m7	1	1	–	–	–	–	–	–	–	–	0	0	0	0	0	0	0	0
m8	1	1	–	–	–	–	–	–	–	–	–	0	0	0	0	0	0	0
m9	1	1	1	1	1	–	–	–	–	–	–	–	0	0	0	0	0	0
m10	1	1	1	1	1	–	–	–	–	–	–	–	–	0	0	0	0	0
m11	1	1	1	1	1	1	1	–	–	–	–	–	–	–	0	0	0	0
m12	1	1	1	1	1	1	1	1	–	–	–	–	–	–	0	0	0	0
m13	1	1	1	1	1	1	1	1	1	–	–	–	–	–	0	0	0	0
m14	1	1	1	1	1	1	1	1	1	1	–	–	–	–	–	0	0	0
m15	1	1	1	1	1	1	1	1	1	1	1	1	1	–	–	–	–	–
m16	1	1	1	1	1	1	1	1	1	1	1	1	1	1	–	–	–	–
m17	1	1	1	1	1	1	1	1	1	1	1	1	1	1	–	–	–	–
m18	1	1	1	1	1	1	1	1	1	1	1	1	1	1	–	–	–	–

3.4 Conclusion

This chapter presented a first-of-its-kind Waterloo UHD-HDR-WCG database composed of PQ-encoded UHD-HDR-WCG content compressed by H.264 and HEVC encoders in a content adaptive manner. A subjective study was carried out on a professional Canon DP-V2420 4K/UHD HDR Reference Display. To the best of our knowledge, such an endeavor has not been attempted before, and no database of its kind is available to the research community. We have also proposed a novel method to process subjective data into MOS that accounts for subject quality scale variations while keeping the overall mean and standard deviation of subjective scores unchanged. Finally, we have evaluated the performance of eleven FR and seven NR representative objective quality assessment methods on the new database. Our analysis indicates that FR methods developed for LDR content are promising to serve as the foundation for the development of highly effective FR-VQA models for UHD-HDR-WCG videos. On the other hand, there is substantial room for improvement when it comes to NR-VQA of UHD-HDR-WCG content.

Chapter 4

Constant luminance vs. Non-Constant luminance for HDR WCG VQA

It is known that the HVS possess a non-linear response to luminance [67]. Therefore, in the case of HDR, the linear light RGB data is transformed to conform to the human visual system response to luminance by applying Perceptual Quantizer (PQ) or Hybrid Log-Gamma (HLG) Opto-Electronic Transfer Functions (OETF) [6]. This allows for efficient quantization of the image or video data [67][53]. A display is said to support PQ or HLG when the display conversion of the electrical signal to an optical signal matches the PQ or HLG Electro-Optical Transfer Function (EOTF), which reverses the encoding that takes place when OETF is applied, reproducing the visual signal to the HVS. The PQ EOTF [6] and the OETF [6] are given in Equations (4.1) and (4.2) respectively. The parameters $m_1 = 2610/16384$, $m_2 = 2523/4096$, $c_1 = 3424/4096$, $c_2 = 32 \times 2413/4096$, $c_3 = 32 \times 2392/4096$ and $E >= 0$ (if $E < 0$, E would be set to 0 for the computation) are defined in BT 2100 [6].

$$E = EOTF(E') = 10,000 * \left[\frac{\max[(E'^{\frac{1}{m_2}} - c_1), 0]}{c_2 - c_3 E'^{\frac{1}{m_2}}} \right]^{\frac{1}{m_1}} \tag{4.1}$$

$$E' = OETF(E) = \left[\frac{c_1 + c_2 (E/10,000)^{m_1}}{1 + c_3 (E/10,000)^{m_1}} \right]^{m_2} \tag{4.2}$$

However, it should be noted that the EOTF function was derived to be applied to the linear light luminance component Y. However, in practice and due to the existing technological implementation ([67]), the EOTF function is applied to the R, G, and B components of a pixel. The relationship between the linear light luminance Y and linear light RGB components are given by the following relationship for BT 2020 [67],

$$Y = 0.2627R + 0.6780G + 0.0593B. \tag{4.3}$$

In the existing conversion process from linear light, the PQ encoded RGB data (usually denoted as $R'G'B'$) is then converted to $Y'C'_bC'_r$ following the procedure defined in BT 2100 [6]. This procedure computes the Y' component of the signal using the same formula given above in Equation 4.3. Thus, generally,

$$OETF(Y) \neq Y'.$$

Hence the terminology of *constant luminance* vs. *non-constant luminance*. An image is said to be encoded in *constant luminance* if the OETF function is applied to the linear light luminance Y directly [67][18]. In such a case, $OETF(Y) = Y'$. A signal is said to be *non-constant luminance* if the EOTF function is applied to each of the RGB components separately, and Y' is computed afterward using the non-linear $R'G'B'$ data [67].

In the case of applying full reference objective video quality assessment algorithms that are based on similarity measures (like SSIM, SR SIM, etc.), the objective quality assessment algorithm will operate on the Y' component that is extracted from image data. It is assumed that the Y' component that is used in this way, corresponds to the luminance information that is present in the image.

However, as explained above, Y' is what is denoted as the non-constant luminance, and in general, is not the same as $OETF(Y)$. Since $OETF(Y) \neq Y'$, this implies that Y' component does not necessarily contain all the luminance information that is actually observed by the human visual system. Some of the luminance information is leaked into the C_b and C_r data that is ignored by these quality assessment algorithms.

Thus, in the case of the UHD database presented in Chapter 3, the HDR video content is encoded using the PQ OETF function applied to each R,G and B component separately (i.e., non-constant luminance, as is the standard for content distribution). However, the PQ function was developed to quantize luminance information with sufficient bit depth to avoid quantization artifacts such as banding [53]. While applying PQ function to each linear RGB component is maximally efficient [31], the non-constant luminance computation process makes it even less likely to preserve the entirety of the luminance information within the Y' component.

In light of these points, this chapter investigates the impact of using the constant luminance as opposed to non-constant luminance for objective quality assessment methods. We confine our analysis to the similarity based methods since they rely on the luminance information more prominently, and have a much faster computation time.

4.1 Preprocessing Steps for Obtaining Constant Luminance

To obtain the actual luminance for the UHD HDR database videos presented in Chapter 3, we first upsample the $Y'C'_bC'_r$ 420 frames of a HDR video to obtain a $Y'C'_bC'_r$ 444 frames. This will provide a $Y'C'_bC'_r$ value for each pixel, which is missing in the subsampled $Y'C'_bC'_r$ 420 frames. Using these values, the corresponding $R'G'B'$ values for each pixel will then be computed using the following formulas,

$$R' = Y'/(2^n - 1) + 1.4746(C_r - 2^{n-1})/(2^n - 1), \tag{4.4}$$

$$B' = Y'/(2^n - 1) + 1.8814(C_b - 2^{n-1})/(2^n - 1), \tag{4.5}$$

$$G' = (Y'/(2^n - 1) - 0.2627R' - 0.0593B')/0.6780, \tag{4.6}$$

where n is the bit-depth of the content (10 bit to 12 bit for HDR WCG content). These equations are derived from the $R'G'B'$ to $Y'C'_bC'_r$ conversions for HDR BT. 2020 content that are provided in BT. 2100[6].

Then the linear light RGB components are obtained by applying the PQ EOTF function to the $R'G'B'$ components, and then the actual luminance is computed using Equation (4.3). Note that the actual luminance value computed using Equation (4.3) is in linear light. Since the luminance perception of the HVS does not correspond to a uniform scale in linear light, the use of the actual luminance Y in computation of differences/similarity becomes meaningless. To facilitate computation of similarity in a meaningful way, the Y component needs to be transformed to a space where the difference in the coded values correspond to equally perceptible changes in the visual stimulus by the HVS. Since the PQ transfer function was formulated with this very consideration (although it is applied to RGB components in practise due to the presently established content distribution pipeline[67][31]), one can achieve a meaningful actual luminance value, which we denote in this paper as Y_{PQ}, by applying the PQ OETF. The procedure to obtain Y_{PQ} and the quantization procedure used, are summarized in equation (4.7), where n corresponds to the bit-depth which can vary between 10-bit or 12-bit for HDR,

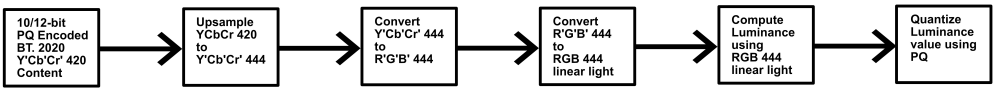


Figure 4.1: Preprocessing steps for obtaining constant luminance PQ (Y_{PQ})

$$Y_{PQ} = \text{round}[2^{n-1} \times PQ\_OETF(Y)]. \tag{4.7}$$

The above steps are summarized in Figure 4.1. Once these pre-processing steps are applied to the data, the objective quality assessment algorithms are applied to the Y_{PQ} component of the signal as usual, which is now a more accurate representation of the luminance content of a frame of video or an image than Y' as explained in the previous section.

4.2 Testing and Results

The ESSIM, FSIM, GSIM, and SRSIM algorithm performance with the added pre-processing steps were tested on the Waterloo UHD-HDR-WCG database described in the previous Chapter. These were the best performing algorithms as reported in Chapter 3[11] for the UHD HDR WCG database that is used here for testing, and PSNR was also tested with the additional pre-processing steps. The HDR-VDP and HDR-VQM algorithms were not tested since these algorithms operate by receiving the linear light data as inputs [50][62], so the difference in constant luminance vs non-constant luminance would have no impact. The results of running the algorithms without the pre-processing are shown in Table 4.1, and the results of running the algorithm with the pre-processing steps are shown in Table 4.2.

As can be seen from the results, the use of Y_{pq} does not seem to affect the performance of the objective quality assessment by a significant amount. The performance of SSIM becomes comparable to the best performing SRSIM quality assessment algorithm. This can be perhaps attributed to the fact that each of these algorithms are less sensitive to the absolute change in luminance values. The FSIM algorithm operates by creating a feature map based on phase congruence on which computes the similarity measure [99]. While the feature map is built by processing of the luminance data, and a gradient map is utilized to supply the contrast invariance of phase congruence, the absolute values of luminance may

Table 4.1: Performance of Quality Assessment Algorithms Without the Proposed Pre-processing [11].

Method	PLCC	SRCC	RMSE
ESSIM [100]	0.8512	0.8389	10.1485
FSIM [99]	0.8693	0.8564	9.5568
GSIM [39]	0.8596	0.8453	9.8812
PSNR	0.5113	0.4615	16.6185
SRSIM [98]	**0.8726**	**0.8630**	**9.4462**

Table 4.2: Performance of Quality Assessment Algorithms with the Proposed Pre-processing.

Method	PLCC	SRCC	RMSE
ESSIM [100]	0.8580	0.8468	9.9326
FSIM [99]	0.8682	0.8518	9.5942
GSIM [39]	0.8607	0.8465	9.8455
PSNR	0.5142	0.4653	16.5850
SRSIM [98]	0.8706	0.8578	9.5140

not be as important for these computations. Similarly, SRSIM operates on computing the similarity of salient maps[98], and the GSIM algorithm on the gradient maps [39]. It is also possible that compression type distortions (which are the only distortion type present in the UHD HDR database) do not generally lead to artifacts from leakage of the luminance components to the Cb and Cr components. It is possible that such artifacts are more noticeable when comparing a pristine image that is not sub-sampled ($Y'C'_bC'_r$ 444) to an up-sampled (which naturally takes place within a display) image from a heavily sub-sampled ($Y'C'_bC'_r$ 420) version of it.

4.3 Conclusion

In this chapter, we have studied the effect of operating the objective quality assessment algorithms on the constant luminance component as opposed to the lossy non-constant luminance on which it is usually designed to operate. In our evaluation of the performance with the pre-processing steps on the Waterloo UHD-HDR-WCG database, the performance results of similarity based methods on the Waterloo UHD-HDR-WCG database with the pre-processing steps did not show significant improvement. This could be, surprisingly,

due to the algorithms being less sensitive to variations in absolute luminance values of the image. It could also be the case that the potential artifacts are only visible when comparing the sub-sampled image to the pristine non-sub-sampled version of the image.

Chapter 5

Testing Methodology for Evaluating the Performance of a Colour Difference Method for HDR WCG[1]

Before existing colour difference measurement methods or new methods can be proposed to use in evaluating colour difference perception in HDR WCG images and video, one must first establish the performance of existing methods at detecting colour differences. Our work in regards to evaluating the performance of colour difference methodologies will be based on the notion of perceptual uniformity of a colour space. If a colour space is perceptually uniform, an approximately equal level of change in the colour values results in a perceptually equal difference across the entire colour space [67]. From an evaluation of colour accuracy point of view, a perceptually uniform colour space provides a fixed threshold over which the difference between two colour signals can be compared to determine if the HVS will notice a perceptual difference. If a colour space were not perceptually uniform, then one would be unable to determine a fixed threshold over which the HVS would detect a colour difference since the threshold would be subject to change depending on the location of the colour values within the colour space. This threshold over which a difference in colour is noticeable to the HVS is referred to as the just noticeable difference

[1]The content of this section include content from a paper that has been published in 2020 IEEE International Conference on Image Processing (ICIP) held in 2020, available online: https://ieeexplore.ieee.org/document/9191176 [doi: 10.1109/ICIP40778.2020.9191176]. T. Costa, V. Gaudet, E. R. Vrscay and Z. Wang, "Perceptual Colour Difference Uniformity in High Dynamic Range and Wide Colour Gamut," 2020 IEEE International Conference on Image Processing (ICIP), 2020, pp. 161-165

(JND) [67]. The JND denotes the smallest level of change in the colour values that the human visual system (HVS) can perceive.

Suppose we were in possession of information regarding a number of colour pairs where each colour pair consists of two colours within a JND of each other. If a colour space or colour difference method is perceptually uniform, then the evaluated colour difference for each of these colour pairs using the colour space or colour difference methodology must be a constant value. Thus, the approach of the testing methodology we propose in this chapter is to discover such colour difference pairs within a JND of each other. Once this information is gathered from the test, the colour spaces and colour difference methods can be evaluated. It is worth noting that a thorough test of JND colour uniformity would require large number of samples from the full colour space, followed by gauging the JND along multiple directions within the colour space for each sampling point. Unfortunately, the potential number of combinations prohibit such a test, resulting in very limited testing in reality [47][92][69][65]. However, to formulate strong conclusions regarding the performance of a colour space and to use the data for further analysis and development of a better colour space, is largely dependent on possessing information on larger number of colour pairs within a JND of each other. Therefore, it is desirable when formulating a testing methodology that one be able to efficiently gather information on as many colour points as possible.

Before introducing our testing methodology and test results, it must be said that the interest in finding the JND difference between colours does date back to the MacAdams experiments performed for measuring the JND in 1942 [47]. The experiment was conducted with a constant luminance for the test area of approximately 48 candela per square meter (cd/m^2 or nit). Another similar experiment by Wyszecki and Fielder [92] was conducted at a constant luminance level of 12 nits. Recent colour difference tests performed by Dolby Laboratories on a newly introduced colour space ICtCp are presented in [69][65]. The results suggested that the ICtCp space exhibits better perceptual uniformity than RGB, YCbCr, L*a*b* and ΔE2000, especially at low luminance levels.

However, apart from the recent colour difference tests performed by Dolby laboratories, the previous tests for determining JND were limited to very low luminance ranges compared to the HDR luminance range of 0 nits to $10,000$ nits. The particular colour primaries within which the tests were performed are also unclear. While the Dolby laboratories did gather the JND data for colour pairs within the range of HDR and WCG primaries, the number of colour pairs tested is still quite small. The experiments by McAdams [47], Wyszecki and Fielder [92], and two experiments by Dolby Laboratories [69][65] covered only 25, 30, 9 and 21 unique test colours. The latest experiment from Dolby laboratories was also limited to 3 luminance levels (0.1, 25, and 1000 cd/m^2) [65]. As shall be seen in the following section, our proposed method allows for much larger coverage to be tested within a shorter

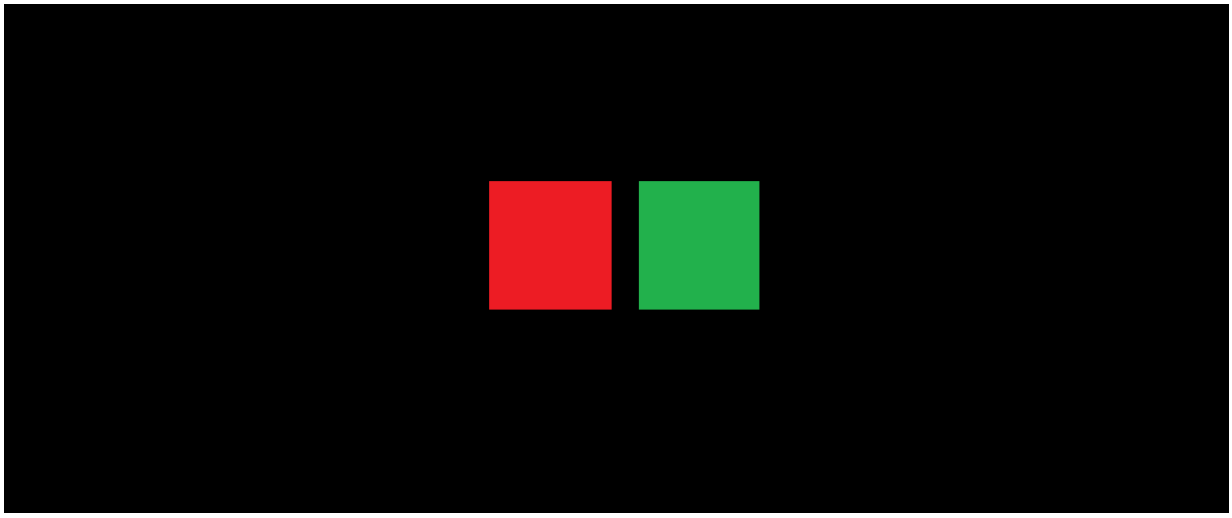


Figure 5.1: Test screen layout.

time span. The ability to gather data within as short time span as possible is extremely important to reduce fatigue of the HVS of the test subject and maintain reliability of the collected data.

In this chapter, we present a testing framework for evaluating the perceptual uniformity of colour spaces and colour difference measures. We also introduce methods for analyzing the collected data using the proposed framework and a means to evaluate the performance in such a way that the performance of various colour spaces and colour difference measures can be compared against each other.

5.1 Testing Methodology

The test consists of a collection of test cases where a subject is shown two squares of different colour located in a completely black background on a display as shown in Figure 5.1.

For each of these coloured pair of squares, one of the squares is chosen at random to be the reference colour, and the test subject is asked to adjust the colour of the other square (which we will refer to as the *test square*) to match it against the reference colour. The adjustment of the colour of the test square is facilitated by providing the subject with a slider displayed on a separate display placed adjacent to to the reference display with the colour squares. When the slider is at its leftmost limit, the test square will appear

the most different in colour. When the slider is at its rightmost point, the test will be designed such that the test square will be of the same colour as the reference square. The subject is instructed to find the leftmost point on the adjustment slider such that the two squares appear to be the same in colour. The subject will then press a button to record the position, and the next test case consisting of another pair of reference and test colour squares will begin.

The reference and test colours are chosen as follows.The reference colours for each test are randomly sampled from within the xy$Y$ colour space [67]. The luminance value Y is fixed at the required level for each test case. For each luminance level for which a reference colour is chosen, 1500 test colour points are then sampled in the xy$Y$ space along a straight line (of length 0.1 in the xy chromaticity diagram) from the reference, at a randomly chosen angle. The luminance Y for the test colour points is held constant at the same value as the corresponding reference colour. The reference and test colour points are then converted to PQ encoded [81] 10-bit RGB values [6] so that they can then be displayed to the test subject.

The colour of the test square at the beginning of each test case is set to the furthest sampled test point for that particular reference colour. This corresponds to the leftmost position of the slider accessed by the test subject on the adjacent display to adjust the colour of the test square. As the subject moves the slider, they navigate through the sampled 1500 test colour points, and the colour of the test square is set to the test colour point at which the slider is located. When a subject finds the leftmost point at which the two squares appear the same, this selection corresponds to the JND for that particular reference colour.

In terms of equipment and ambient lighting, the test is carried out on a Canon DP-2420 reference display in a dark environment with no ambient lighting. The display supports a maximum luminance of 1200 nits and the BT 2020 colour gamut. However, the Canon DP-2420 is not an OLED display, and is therefore incapable of displaying an absolute black level of 0 nits. Thus, black background of the two colour squares on the display is obtained by covering the entire display with black-out material, leaving only the two colour squares uncovered. A secondary display containing the slider bar is displayed on a HD LCD monitor with the display covered in black-out material leaving only the slider uncovered, to minimize any light from the display interfering with the subject perception of the colours. The secondary display is also positioned slightly behind the horizontal plane on which the reference monitor screen is positioned to further reduce any impact from the light emitted from this display. The slider is moved by the test subject using a mouse. When the subject has chosen the leftmost position on the slider where the two squares appear to be of same colour, the press of the keyboard ‘space’ key by the subject records

the slider position, and displays the new pair of squares corresponding to the next test case on the reference monitor, and resets the slider position to the left.

Since each person may have slight variations in colour perception between each other, and there may also be errors when performing the colour matching, more than 20 subjects will be used to determine the JND for a given reference colour.

Each test subject spent an hour on average to complete the test. A mandatory break is enforced at the halfway point of the test to reduce visual fatigue. Test subjects are also allowed additional breaks if required.

Several key choices described in the above mentioned test framework are explained as follows.

- **Sampling in the xy$Y$ space:** We aim for providing a common ground for testing, and sampling the colour points from the same colour space being tested could result in a systematic bias (*e.g., points could be sampled at sufficiently larger distance than the JND variance, thus giving the impression that it is perceptually uniform*). Therefore, we sample the colour points in the common xy$Y$ space where the standard colour gamuts such as ITU-R BT. 709 [21], BT. 2020 [17] and DCI P3 [80] are defined.

- **Conversion to RGB rather than YCbCr for display:** YCbCr space was designed mainly for industrial video distribution purposes with the intent of subsampling, which impacts the Cb and Cr components that are required for accurate colour representation.

- **Choice of luminance values:** The luminance range was chosen to cover a wide dynamic range with HDR applications in mind. It is also matched with the test carried out at the Dolby laboratories [69].

- **Difference in experimental setup:** In the subjective experiment carried out at Dolby laboratories [69], each test case consists of 4 colour squares placed in 4 quadrants of the screen, with a fixed test colour in 3 of the squares and the reference colour displayed in one of them. The same test and reference colour combination will repeat 4 times, with the reference colour contained in a different quadrant each time. As a consequence, the pairs of reference and test colour combinations that can be tested according to this framework is fairly limited considering the fact that the subjects would need to finish the experiment within a reasonable period of time to avoid visual fatigue. With the framework described in this paper, each test case covers a single reference colour and 1500 unique test colour points (and could be

simply scaled to even more points if necessary without affecting the length of the test), offering a far more efficient level of coverage.

- **Spatial separation of test colour squares:** While no spatial separation was used between the test and reference colours in the McAdams experiment [47], studies on visual perception indicate that there are optical illusions (*Cornsweet illusion* [26] and *Mach band effect* [30]) that can cause the HVS to incorrectly perceive differences when two colours are placed immediately next to each other. Therefore, separating the two colours (reference and test) with a neutral background colour can alleviate such effects.

- **Spatial versus temporal separation of reference and test colour squares:** One possibility that was considered was the use of a single square that can be toggled to display between reference and test colour samples, i.e. temporal separation between test and reference colours. However, earlier studies have shown that due to the successive contrast effect, the HVS may be affected by the previously viewed colour [93] [91].

5.2 Obtaining the JND from Testing Data

Given the JND J_k computed for each subject k for a particular test case i using the colour space or colour difference method under investigation, the final average JND value J_i for the test case i can then be computed as (similar computation has been performed in [69][65])

$$J_i = \frac{1}{N} \sum_{k=1}^{N} J_k. \tag{5.1}$$

Given the JND of J_i for each test case i, the JND for a particular luminance L, denoted J_L, can be computed as

$$J_L = \frac{1}{n(L)} \sum_{i \in L} J_i, \tag{5.2}$$

where $n(L)$ denotes the number of test cases corresponding to the luminance value.

To have a better comparison between the variation of the computed JND J_i for each test case i in comparison to the computed JND for a particular luminance based subsets

of the test cases J_L (or all luminance values), we compute the coefficient of variation (CV) as given by [10],

$$CV = \frac{standard\ deviation}{mean} = \frac{\sigma(\{J_{i \in L}\})}{J_L}. \tag{5.3}$$

If the value of CV is closer to zero, it follows that J_i for each of the test cases within a particular subset L are closer to the computed global JND J_L. We consider this to be a strong indicator that the space is perceptually uniform within that luminance range. In the same manner, if the CV value is very high, it then follows that the JND J_i of each test case within the particular subset L is largely varying. We consider such a large variation to indicate that the perceptual uniformity of the space is lacking. Thus, the CV is used as a measure of the perceptual uniformity of the colour spaces for a particular luminance range for which the test data is available. Since CV is normalized, it also allows us to compare the performance (the perceptual uniformity) across colour spaces and colour difference measures.

5.3 Conclusions

In this chapter, we introduce a testing framework that can be used for colour difference testing to efficiently measure a large number of colour points to measure the JND for display applications. We also introduced the use of the CV to compare the performance between colour spaces in regards to their perceptual uniformity. This framework and the performance measure will be used in subsequent chapters to study the performance of existing colour spaces in the subsequent chapters.

Chapter 6

Evaluating the Performance of Existing Colour Difference Methods for HDR WCG

As mentioned in the previous chapter, the most recent works evaluating the performance of existing colour difference methods as applicable for HDR WCG applications were from Dolby Laboratories [69][65]. This chapter presents the results of two subjective studies that we performed to analyze the performance of existing colour difference methods for HDR WCG. Both of these studies follow the testing methodology introduced in Chapter 5.

6.1 Colour Difference Subjective Study I

Our first colour difference study was performed in the March of 2018, and the main purpose of the study was to independently verify the results as indicated in the work by Dolby Laboratories [69]. Since the result of the 2017 study by Dolby laboratories concluded that the ICtCp space introduced by the same authors of the paper, was the most successful at predicting colour difference, it was deemed necessary to first confirm that this result was indeed accurate. If the results show that ICtCp is perceptually uniform, then one could potentially utilize ICtCp for detecting colour difference, which is a vital part of a colour quality assessment algorithm for HDR WCG content (also the focus of this thesis).

Since the purpose was to mainly verify the result provided by Dolby laboratories, 7 reference colours that were used by Dolby laboratories were used for our subjective study

as well. These colours consisted of Red, Green, Blue, Yellow, Magenta, Cyan and White. These colours were chosen at the primaries or boundaries of the gamut similar to [69]. However, we did incorporate a larger luminance range for the chosen reference colours ranging from 0.05 nits to 1000 nits, with points sampled for luminance values of 0.05 nits, 0.01 nits, 100 nits, 300 nits, 700 nits, and 1000 nits. The Dolby Laboratories experiment was limited to 0.05 nits to 500 nits. For each chosen colour, test colour points were sampled in 3 to 8 directions per reference colour (1500 test colour points per direction) for a total of 217 test cases.

The subjects were scheduled to complete the test over a time of 1 hour and 30 minutes. The subjects were given a break after each 25 minutes interval. They were asked to leave the room of the experiment setup and adapt their eyes to the natural light. The test cases were also ordered from the lower luminance range to the higher luminance range to prevent any errors or visual discomfort that may occur due to constant visual adaptation if the brightness levels changed with each test case.

6.1.1 Results & Analysis of Colour Difference Study I

The experiment consisted of data from 33 test subjects, aged between 22 and 35. All the subjects had normal or corrected-to-normal vision, and were not colour-blind. The data was processed as described in Chapter 5 to obtain the JND for each colour pair under the different colour spaces and colour difference measures. The colour spaces RGB, YCbCr, L*a*b*, ICtCp, xy$Y$, and the colour differencing method ΔE2000 were considered in this test.

First, the ΔRGB values were computed for each test case and each test subject using the following,

$$\Delta\text{RGB} = \sqrt{(R'_r - R'_t)^2 + (G'_r - G'_t)^2 + (B'_r - B'_t)^2} \tag{6.1}$$

where R'_r,G'_r,B'_r denote the reference RGB values and R'_t,G'_t,B'_t values denote the RGB values of the test colour sample chosen by the subject during the test. Outlier detection and removal were then performed on the collected data using these computed JND values as specified in [19].

Once we have removed the outliers, the RGB data of the reference colour and the subject chosen test colour point were then converted to the YCbCr, ICtCp, Lab and xy$Y$ colour spaces. To convert RGB to YCbCr, as per BT 2100, the Equations 2.17 were utilized [6]. Note that the R', G' and B' refer to the PQ mapped RGB values that were sent to the display during the test.

However, to compute ICtCp, La*b* and xy$Y$ (and the colour difference measure ΔE2000 that uses the L*a*b* values), we require the XYZ colour space values for each colour pair (the reference colour and the test colour point chosen by the test subject for that particular reference colour). To compute XYZ values, we computed the linear R, G, and B values by applying the PQ EOTF function given in 2.1 to each R', G' and B' component. Then the XYZ values were computed using the relationship described by Equation 2.7 and the matrix M given in Equation 2.10.

The xy$Y$ values were computed using the relationships given in Equation 2.5. Note that the Y values are already known once one possesses the XYZ values, and only the xy components need to be computed. The ICtCp values were then computed by converting the XYZ values to the LMS colour space, applying the PQ OETF, and multiplying by the conversion matrix to convert L',M' and S' values to ICtCp as defined in [36]. To convert to L*a*b*, the procedure described in Equations 2.11, 2.12 were used. These L*a*b* values were used to compute the ΔE2000 values as well.

The JND for each of the colour spaces and difference measures were then computed for each reference colour corresponding test colour chosen by each of the subjects as follows.

For YCbCr,

$$\Delta\text{YCbCr} = \sqrt{(Y_r - Y_t)^2 + (Cb_r - Cb_t)^2 + (Cr_r - Cr_t)^2} \tag{6.2}$$

was used.

L*a*b*, and ICtCp were calculated as follows [69].

$$\Delta\text{L*a*b*} = \sqrt{(L^*_r - L^*_t)^2 + (a^*_r - a^*_t)^2 + (b^*_r - b^*_t)^2} \tag{6.3}$$

$$\Delta\text{ICtCp} = \sqrt{(I_r - I_t)^2 + 0.25(Ct_r - Ct_t)^2 + (Cp_r - Cp_t)^2} \tag{6.4}$$

ΔE2000 and CIECAM02-UCS were also computed for each test case given in [32] and [58].

Once the JND as indicated by each test subject for each test case was computed, the average JND for each test case (each reference colour) was computed using the procedure described in Section 5.2 of the previous chapter, and the coefficient of variation was computed as listed in Table 6.1.

According to the results, it was difficult to conclude that the ICtCp colour space was more perceptually uniform than the other colour spaces. In fact, the results seemed to indicate that ΔRGB was perceptually uniform overall. The performance of ΔE2000 is also significantly worse compared to the other colour spaces under consideration. ΔICtCp is seen to perform well in the very low luminance regions, but even then, the performance of ΔRGB was not far behind, suggesting that the perceptual uniformity of the RGB colour space was not significantly worse than ICtCp.

Table 6.1: Colour Difference Study I, Coefficient of Variation for each colour space at each tested luminance level and the entire tested luminance range (denoted as "All").

Method	**0.05 nits**	**0.1 nits**	**100 nits**	**300 nits**	**700 nits**	**1000 nits**	**All**
ΔRGB	0.7441	0.7595	0.5655	0.3663	0.3436	0.3074	0.82867
ΔICtCp	0.6064	0.6574	0.5078	1.5496	1.6151	1.7805	1.6694
ΔYCbCr	0.8356	0.8461	0.6268	0.4771	0.4481	0.4242	0.8964
ΔL*a*b*	1.4919	1.6407	0.6722	1.6844	1.9269	2.1756	2.8641
ΔE2000	1.3338	1.5256	1.0218	1.7255	1.8580	2.0354	2.4883
Δxy$Y$	0.7235	0.6707	0.9116	1.7055	1.8776	2.1176	1.6054

6.1.2 Problems with the Colour Difference Subjective Study I

In discussion with industry partners involved in the manufacturing of reference displays, we discovered that the display capabilities can be unpredictable closer to the boundaries of the gamut. Although the Canon DP-2420 specifications indicated full support for the BT 2020 colour gamut, this made the results of a colour difference test performed with reference colours located on the boundary of the gamut, unreliable due to unpredictability of the display behaviour.

Another problem arose from the high luminance range test cases. For test cases that were in the range of 700 nits and higher, a significant potion of subjects found it difficult to evaluate colour difference due to the high level of perceived brightness. Such difficulty would have likely compromised the reliability of the test as well.

Furthermore, it was difficult to use the data to make a generalized conclusion regarding the entirety of the colour space using only the reference colours located in 7 different points within the gamut. A more uniform sampling of the gamut would provide a more generalized conclusion, and also provide useful insight to construct a better colour difference measure for the purpose of HDR WCG image and video quality assessment (in the event that the existing measures did not show satisfactory performance).

6.2 Colour Difference Subjective Study II[1]

To more uniformly sample the HDR WCG, we uniformly sampled the xy$Y$ colour space to obtain 245 reference colours contained within the ITU-R BT 2020 primaries. The luminance values for the reference colours were fixed at 0.01 nits, 0.1 nits, 1 nits, 10 nits, 100 nits, 300 nits and 500 nits. At each luminance level, 35 samples were chosen for a total of $7 \times 35 = 245$ reference colours. The location of these 245 reference colours are shown in the xy chromaticity diagram given in Figure 6.1. The testing methodology used for this subjective study is the same as the one described in Chapter 5 and the one used for the previously discussed colour difference subjective study.

Note that the previous study only contained 7 reference colours with 1500 test colour points per direction (and 3 to 8 directions per reference colour). Such sampling would allow one to determine the local uniformity of the colour space at the location of the 7 reference colours. However, for this colour difference study, we chose to focus on the global uniformity of the colour space. This was to reduce the time required to complete the subjective study without fatiguing the test subjects, obtaining sufficient information to evaluate the performance of the colour space, and also to potentially utilize the data to propose a new colour space. Moreover, the results of such an experiment can still produce insight into the local uniformity of the colour space (for a given luminance range), though not at the fine level of knowing the local uniformity as it holds for individual colours. At the same time, to determine the perceptual uniformity of a colour space, dense sampling of the colour space is required. While ideally, knowledge regarding both local and global perceptual uniformity is desirable, from the perspective of practicality (keeping the subjective test sufficiently short to reduce subject fatigue and collect reliable data) and utility (gaining sufficient knowledge to make a generalized conclusion), it would be of more benefit to determine global perceptual uniformity. Since the directions along which the test colours are sampled for each reference colour are chosen at random, global perceptual uniformity would indicate that the colour space is likely perceptually uniform locally as well. There is also the added benefit that one now possesses information regarding the behaviour of the space for a myriad of colours, which is more useful from the perspective of colour quality assessment of images and video. We considered the above reasons sufficient

[1]The content of this section include content from a paper that has been published in 2020 IEEE International Conference on Image Processing (ICIP) in 2020, available online: https://ieeexplore.ieee.org/document/9191176 [doi: 10.1109/ICIP40778.2020.9191176]. T. Costa, V. Gaudet, E. R. Vrscay and Z. Wang, "Perceptual Colour Difference Uniformity in High Dynamic Range and Wide Colour Gamut," 2020 IEEE International Conference on Image Processing (ICIP), 2020, pp. 161-165

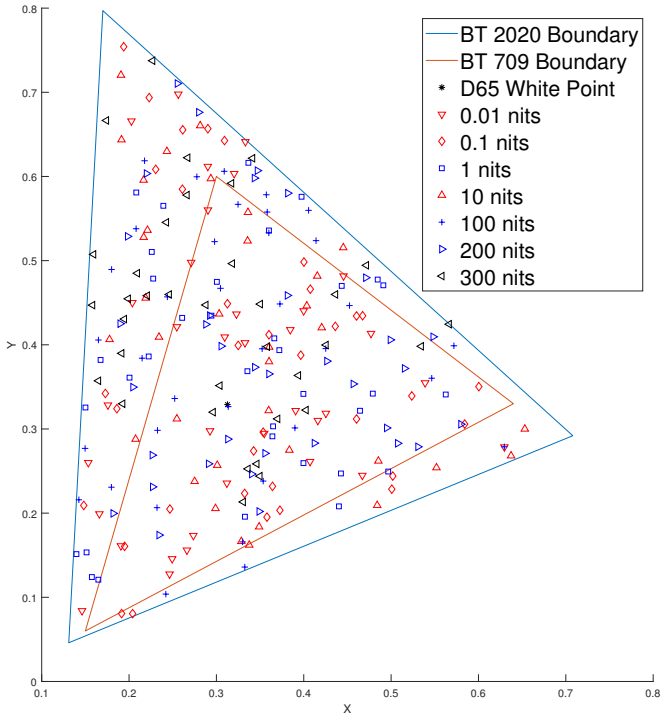


Figure 6.1: Location of chosen reference colours on the xy chromaticity diagram.

justification for choosing to only test the JND in a single direction for a reference colour, and maximize the number of reference colours.

We also excluded the higher luminance ranges of 700 nits and 1000 nits that were included in the previous colour difference study to avoid visual discomfort and possible sources of errors which was a defect of the previous study.

6.2.1 Results & Analysis of Colour Difference Study II

A total of 30 subjects, aged between 22 and 35, took part in the colour difference subjective study. All the subjects had normal or corrected-to-normal vision, and were not colour-blind. We then converted the study results which were in the RGB colour space to the YCbCr, ICtCp, L*a*b*, and xy$Y$ colour spaces as described in the analysis of the previous study discussed above. The JND values were then computed as described in Chapter 5 for the colour spaces and also for the colour difference measures ΔE2000 and CIECAM02-UCS. The results of the JND data is summarized in Table 6.2.

As seen from the JND results table, the JND does vary quite significantly for all of the colour spaces depending on the luminance. While ΔE2000 does appear to have a stable JND, the JND results table does not provide the necessary information in regards to the spread of the JND among the test cases for each of the colour spaces. So it is difficult to conclude on perceptual uniformity from the JND results table alone. Nevertheless, these results are valuable from the point of view of determining the threshold above which the colour difference evaluated using a colour space or colour difference measure signifies a perceptually noticeable difference (the most vital component of colour quality assessment of images and video content).

However, to better understand the perceptual uniformity and to compare the performance across colour spaces, the coefficient of variations were computed and the results are summarized in Table 6.3. The coefficient of variation (CV) can be considered as a measure of usefulness of the JND values provided in the JND results table. If the coefficient of variation is very high, the corresponding JND values in the JND results table become less useful. A high coefficient of variation indicates a larger standard deviation, indicating that when comparing two colours within the luminance region of interest, one may have an actual JND that is significantly higher or lower than the JND reported in the JND results table. So using the JND value in the JND results table, one is more likely to make an erroneous conclusion regarding the perceptual similarity of the two colours that are being compared. For the above reasons, it is desirable to have a low CV.

Table 6.2: Colour Difference Study II, Average JND in different colour spaces for each tested luminance level and the entire tested luminance range (denoted as "All")

Method	**0.01 nits**	**0.1 nits**	**1 nits**	**10 nits**	**100 nits**	**300 nits**	**500 nits**	**All**
ΔRGB	2.8930	4.6611	10.3475	15.6696	20.6305	19.8423	28.6965	14.6772
ΔICtCp	2.7888	3.9108	6.7452	10.4535	11.7651	11.8718	10.6578	8.3133
ΔYCbCr	0.0017	0.0027	0.0058	0.0091	0.0126	0.0107	0.0156	0.0083
ΔL*a*b*	11.0908	6.1569	11.3634	22.1332	42.1971	56.1250	66.8717	30.8483
ΔE2000	5.1055	2.5949	3.5564	3.4071	4.2322	5.0215	4.2041	4.0174
Δxy$Y$	0.0376	0.0206	0.0169	0.0148	0.0098	0.0090	0.0107	0.0171
CIECAM02-UCS	0.1629	0.0933	0.3203	0.2238	0.2090	0.1727	0.2125	1.6835

Table 6.3: Colour Difference Study II, Coefficient of Variation for each colour space at each tested luminance level and the entire tested luminance range (denoted as "All").

Method	**0.01 nits**	**0.1 nits**	**1 nits**	**10 nits**	**100 nits**	**300 nits**	**500 nits**	**All**
ΔRGB	0.4387	0.5285	0.8584	0.7705	1.0061	0.9764	1.3197	1.3943
ΔICtCp	**0.3536**	**0.3474**	**0.4111**	**0.4765**	0.5347	**0.3345**	**0.2861**	**0.6182**
ΔYCbCr	0.4307	0.5760	0.8034	0.7755	1.0506	0.8671	1.2565	1.3339
ΔL*a*b*	1.5192	0.6029	0.5277	0.4959	0.4463	0.4070	0.3925	0.9091
ΔE2000	1.4293	0.5206	0.4891	0.5054	0.6369	0.8487	0.5929	0.9186
Δxy$Y$	0.4264	0.6450	0.7956	0.8653	0.8180	1.1841	1.3462	0.9230
CIECAM02-UCS	1.4427	1.3239	0.6096	1.2456	**0.4418**	0.6302	0.6524	0.7801

From the results table, we see that ICtCp has the lowest CV over the entire luminance range that was tested. CIECAM02-UCS does also show better performance than ΔE2000 when one considers the entire luminance range. For colours in the 100 nits range, CIECAM02-UCS does outperform ICtCp by a small margin. Somewhat expectedly, both L*a*b* and ΔE2000 which is based on the same colour space, performs significantly worse in the very low luminance range, likely due to the fact that it was not formulated to measure colour differences in such low luminance ranges. These results are consistent with the findings in [69] as well.

Removing the low luminance range does improve the CV of L*a*b* to 0.8262 and the CV of ΔE2000 to 0.6900, much closer to the performance of ICtCp. Interestingly, CIECAM02-UCS was also seen to improve its CV to 0.5825 when luminance under 100 nits were excluded (thus, focusing on the higher end of the HDR luminance range). However, the ICtCp performance for higher luminance ranges is significantly better with a CV of 0.3345 and 0.2861 for the 300 and 500 nit ranges. Perhaps the most surprising result is that RGB, YCbCr and xy$Y$ show comparative performance to ICtCp in the lowest luminance range of 0.01 nits.

6.3 Conclusion

In this chapter, we provided two subjective studies to evaluate the performance of existing colour difference measures utilizing the testing framework and methodology introduced in Chapter 5. The first subjective study results were unfortunately made questionable in light of the display capabilities at the boundary of the gamut. However, the second subjective study covered 245 unique reference colours, in comparison to the 9 and 21 unique reference colours covered in the two experiments by Dolby laboratories [69][65]. This shows that our current proposed test framework is more efficient in obtaining the colour difference information. The agreement of our results with the results reported by Dolby Laboratories [69][65] further confirm that our testing framework and methodology is also reliable.

According to the analysis of the data from our colour difference subjective study, the results suggests that on average, the ICtCp colour space appears to be more perceptually uniform in comparison to the other colour spaces, confirming the claim by Dolby Laboratories. Meanwhile, the performance of ΔE2000 is comparable to the performance of ICtCp if the lowest luminance level of 0.01 nits is excluded, and CIECAM02-UCS performance is better than ICtCp for luminances greater than 100 nits. The analysis also indicates that the RGB and YCbCr colour spaces have a reasonable degree of perceptual uniformity at the lowest luminance range of 0.01 nits. This is especially interesting since these colour

spaces are widely used today for distribution of image and video content, and being able to utilize the image and video data directly (without any conversion to another colour space) is desirable from a IQA/VQA implementation perspective. Given that ICtCp, the most perceptually uniform colour space under test, still has only a significantly high overall CV of 0.6182, our work suggests that there is still a large room for improvement in the development of a perceptually uniform colour space and colour difference measures. We will investigate the possibility of constructing a colour space using the data gathered in this experiment in the coming chapters of this thesis.

Chapter 7

Colour Difference Study III: Effect of Background on Colour Difference Perception [1]

Various factors affecting the viewing conditions such as the surround luminance, surrounding colours and ambient luminance, can affect how a patch of colour would be perceived by the HVS. While the colour difference value may change for different background luminance values, ideally we would prefer the colour space to still be perceptually uniform i.e., for the JND of each colour observed under the same background luminance to remain similar. After all, images and video content do not have a single colour under a single luminance. Thus, studying the effect of background luminance on colour different perception is essential for developing an accurate colour difference measure for IQA. Such a study can confirm whether a particular colour space is more suitable due to the JND remaining constant under different background luminances (unlikely, given the colour difference experiment observations from Chapter 6 indicating that none of the existing colour spaces or colour difference methods were even close to being perceptually uniform), or provide information to determine colour difference by compensating for the change in the JND with background luminance.

This chapter focuses on studying the effect of background luminance on colour difference perception. Similar to the preceding chapters, we focus on the performance of $\Delta E2000$

[1] The content of this chapter include content from a paper that has been published in the 17th International Conference on Image Analysis and Recognition (ICIAR) held in 2020. T. Costa, V. Gaudet, E.R. Vrscay and Z. Wang (2020), Variation of Perceived Colour Difference Under Different Surround Luminance, in Image Analysis and Recognition, ICIAR 2020, LNCS 12131, 56-50 (Springer)

based on the L*a*b* [32], IC_tC_p [36], CIECAM02 colour spaces [58], and RGB, YC_bC_r colour spaces. CIECAM02 is claimed to be more perceptually uniform than $\Delta E2000$ [59], while is IC_tC_p claims to be more perceptually uniform than them both [69][65].

We first describe the experimental design, and present the results obtained from the experimental data. Since it is infeasible to consider different surrounding colours with different luminances, our experiment is limited to an achromatic stimulus as the surrounding luminance. The data and insight obtained from this experiment will be used to refine the objective colour quality assessment methods proposed in Chapter 8 for HDR WCG image and video content.

7.1 Experiment Design

For this experiment, we sampled colour points from the colours contained within BT 2020 primaries [17][6]. 12 reference colour samples were randomly chosen per luminance level, for a total of 72 reference colours at 6 luminance levels of 0.05 nits, 0.5 nits, 5 nits, 50 nits, 150 nits, and 300 nits. While HDR WCG content supports up to $10,000$ nits, the current HDR WCG displays are limited to the 1000 nits range, and the higher luminances are usually used for highlights in an image/frame of video. Thus, a maximum luminance of 300 nits is a realistic choice since a large colour patch would be less likely to be graded to be displayed at higher luminance levels.

The sampling of the colours were performed in the xyY colour space. For each reference colour, 1500 test colour points were sampled at fixed distances from the reference colour (by manually verifying the intra-distance between the points to be far smaller than the perceptual difference, but also such that the furthest point to clearly be perceptually different from the reference colour) along a randomly chosen direction on the xy plane of the xyY colour space, holding the luminance level constant. The colour for the surrounding luminance was chosen to be achromatic, and luminance levels of 0, 0.01, 1, 10, and 100 were chosen for the background.

The test was carried out on a Canon DP-2420 Reference monitor in a dark room with a backlight of approximately 5 nits to reduce eye fatigue. The test procedure is described in the following steps.

- Each reference colour and the furthest sampled test point from the 1500 sampled test points corresponding to the reference colour would be shown on two squares as shown in Figure 7.1. The region in which the surrounding luminance will be changed

is shown in white in the same Figure. Note that the entire monitor screen was not used to show the same surrounding luminance for this experiment since the peak luminance capability/stability of a display decreases as the pixel arrays on the entire screen is activated. Keeping the activated surround region smaller enables the display luminance to be reliably maintained at the desired value.

- The Canon Reference monitor used in this experiment is not an OLED display that can support black levels at 0 nits. Therefore, the display was covered using non-reflective blackout materials, exposing only the two squares to simulate the 0 nits surround luminance. When the other surround luminance values were displayed, the region outside of the chosen surround luminance region was covered using the blackout materials to keep the surrounding luminance at 0 nits.

- The left or right square will be randomly chosen at the start of each test pair to contain the reference colour, while the other contains the test colour. The test subject would be able to adjust the colour in the test square by navigating through the 1500 samples test points using a slider. When the slider is at the leftmost position, the two squares would look the most different in appearance while the two squares would contain the identical colour at the rightmost position. Subjects are instructed to choose the leftmost position at which the two squares look the same.

- Once the leftmost position is found, the subject would press a button and the next test case would be displayed.

- The test consists of 72 test pairs per set, which takes about 15 minutes to complete. The surround luminance was fixed for the entire set at one of the chosen six luminance levels. Once the set is complete, the test subject would have a five minute break, and the next set is started which would contain the same 72 pairs of colours, but a different surround luminance. Each subject required approximately 2 hours to complete the test with the breaks.

7.2 Results

There were 30 test subjects in total, and outlier detection was performed to filter the data as defined in BT.500 [19].

We then first computed the JND for each test case for each test subject. Note that for the computation of CIECAM02-UCS we do incorporate the surround information into the

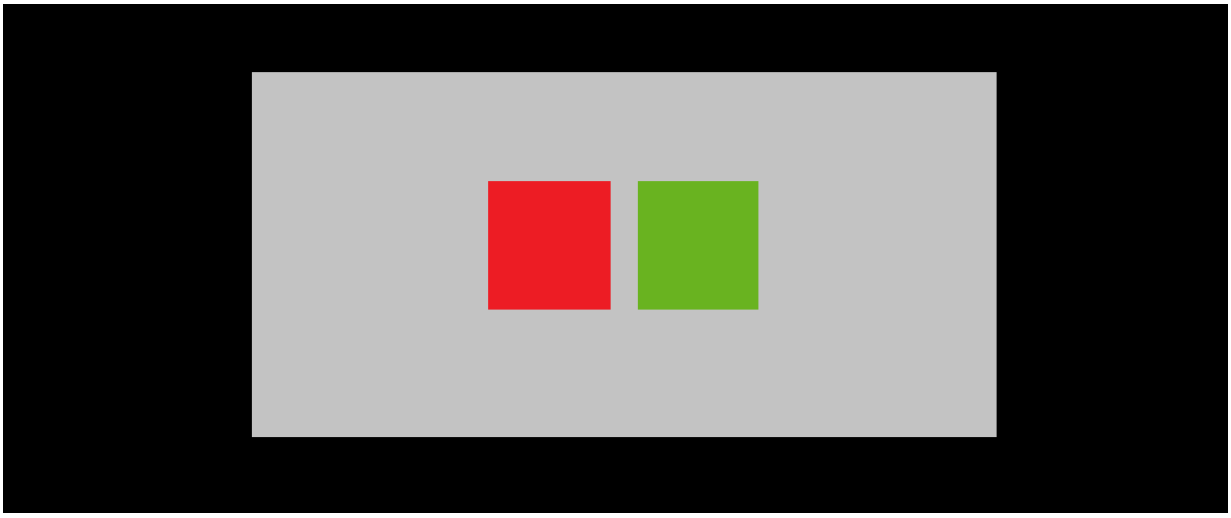


Figure 7.1: Test Setup for testing the impact of surround luminance on the perception of colour difference.

computation. Then we computed the final JND value for the test case as the average of the JND values for each subject. If the colour space or colour difference method is perceptually uniform, then the variation of the final JND value for each test case should be minimal. Therefore, we evaluate the perceptual uniformity of each colour space using the coefficient of variation (CV) given as follows [10],

$$CV = \frac{\textit{The standard deviation of the final JND values}}{\textit{The mean of the final JND values}}. \tag{7.1}$$

The CV was computed for each background luminance level as reported in Table 7.1.

Table 7.1: Coefficient of Variation (CV) measuring perceptual uniformity of the space (lower the CV, better the perceptual uniformity of the space).

Method	**0 nits**	**0.01 nits**	**1 nits**	**10 nits**	**100 nits**
ΔRGB	1.083	1.0917	1.0243	0.9969	0.8265
ΔICtCp	0.6057	0.6328	0.5840	0.5438	0.4884
ΔYCbCr	1.0215	1.0269	0.9649	09288	0.7827
ΔE2000	0.7956	0.8444	0.7739	0.7726	0.7555
CIECAM02-UCS	0.7191	0.7864	0.7249	0.6752	0.6768

It appears from the results that the perceptual uniformity of all the colour spaces

improve with increasing surrounding luminance. We also see confirmation of previous results [69][65][59] that indicated better performance by IC_tC_p over existing colour spaces, and also the better performance of CIECAM02-UCS over $\Delta E2000$.

However, now we turn to the question of whether the JND remains constant as the luminance changes. Table 7.2 contains the average JND computed for each background luminance. The average JND is computed for all the test colour points (72 colour points with 6 different foreground luminances) in each colour space or colour difference method.

Table 7.2: Variation of the global JND with changing background luminance.

Method	**100 nits**	**10 nits**	**1 nits**	**0.01 nits**	**0 nits**
ΔRGB	21.02	15.81	13.25	12.03	11.94
ΔICtCp	14.02	10.03	8.23	7.42	7.40
ΔYCbCr	12.09	8.88	7.41	6.76	0.07
ΔE2000	7.58	5.24	4.06	3.64	3.65
CIECAM02-UCS	3.09	2.24	1.87	1.71	1.67

As seen from the table, the JND does not remain constant in any of the colour spaces as the background luminance changes. Thus, some form of variable adjustment would need to be applied when estimating the JND under a particular background luminance. We shall revisit this problem in Chapter 10, as we discuss the development of an IQA algorithm for evaluating HDR WCG content.

7.3 Conclusions

In this Chapter, we presented an experiment performed to study the perceptual uniformity of prominent colour spaces as the perceived colour stimulus by the HVS is affected by a fixed surround luminance. The results show that the perceptual uniformity of the colour spaces that were tested increases as the surround luminance increases. We also observe that the JND fails to remain constant within any of the colour spaces or colour difference methods, requiring an additional adjustment to the JND value. Thus, algorithms proposed in the previous chapter for colour difference will ideally require an adjustment based on the background luminance.

Chapter 8

Constructing a Novel Colour Difference Measure for HDR WCG Content

Here we present our formulation of a novel colour difference measure based on the data obtained from the colour difference studies described in the previous chapters. It should be noted that the construction of a colour difference measure in this chapter is focused on utilizing the data that was obtained from the previously described colour difference studies. The aim is to produce a colour difference measure that is highly accurate at predicting whether a given pair of colours appear significantly different, and the extent of this perceived difference as well.

The work in this chapter is also limited to the situation where a given colour pair possess the same luminance. We do not focus on the case where the luminance is different since such a perceivable difference can be easily predicted by comparing the luminance values corresponding to the two colours. Thus, the scope of this chapter is to predict colour difference when luminance of the two colour pairs are of the same value.

8.1 Luminance Range Based Colour Difference Measure

In the analysis of results of colour difference studies described in the previous chapters, it was seen that the existing colour spaces and colour difference measures can be more

perceptually uniform at specific luminance ranges. In fact, Table 6.2 in Chapter 6 shows how the different colour difference methods and colour spaces demonstrate a different level of perceptual uniformity at the tested luminance levels in contrast to that of the overall perceptual uniformity (as indicated by the Coefficient of Variation (CV) values in the table). For colour spaces such as ICtCp, the CV values at each of the tested luminance ranges were significantly better than the CV value for the entire luminance range (0.6182), indicating that perceptual uniformity at a fixed luminance range is quite high for the ICtCp colour space. Such experimental evidence from our colour difference experiments supports the notion of constructing a better colour difference measure that utilizes these existing colour spaces while computing and determining the perceptible colour difference as some function of the luminance range.

At the same time, the CV values in Table 6.2 in Chapter 6 also indicate that some colour spaces perform better than others for a given luminance range. Therefore, one can potentially construct a better colour difference measure by computing the colour difference in different colour spaces (to maximize performance), depending on the luminance range as well. To better understand how to combine the existing colour difference measures and colour spaces in a way that maximizes perceptual uniformity, we need to determine the CV of each of them under various luminance ranges. Tables 8.1, 8.2, 8.3, 8.4, 8.5, 8.6 and 8.7 show the CV each colour space or colour difference method for various luminance range combinations. The tables are arranged such that the left-most column entries list the lower bound of the luminance range while the top row forms the list of upper bound of the luminance range. Each of the CV entries in the table correspond to the luminance range interval with the corresponding leftmost entry forming the lower bound and the corresponding top row entry forming the upper bound. The entry of the best performing colour space or colour difference method for each luminance range is bolded in the tables.

Table 8.1: Experiment 2, Coefficient of Variation for ΔRGB at each of the tested luminance level ranges.

ΔRGB	**0.01 nits**	**0.1 nits**	**1 nits**	**10 nits**	**100 nits**	**300 nits**	**500 nits**
0.01 nits	0.4387	0.5664	1.0400	1.0806	1.2168	1.1943	1.3943
0.1 nits	x	0.5285	0.9429	0.9573	1.0957	1.0865	1.2910
1 nits	x	x	0.8584	0.8346	0.9798	0.9866	1.1962
10 nits	x	x	x	0.7705	0.9388	0.9499	1.1607
100 nits	x	x	x	x	1.0061	0.9851	1.1867
300 nits	x	x	x	x	x	0.9764	1.2440
500 nits	x	x	x	x	x	x	1.3197

Table 8.2: Experiment 2, Coefficient of Variation for ΔICtCp at each of the tested luminance level ranges.

ΔICtCp	**0.01 nits**	**0.1 nits**	**1 nits**	**10 nits**	**100 nits**	**300 nits**	**500 nits**
0.01 nits	**0.3536**	**0.3902**	**0.5571**	**0.6996**	**0.7296**	**0.6709**	**0.6182**
0.1 nits	x	**0.3474**	**0.4871**	0.6109	0.6407	**0.5847**	**0.5384**
1 nits	x	x	**0.4111**	0.5136	**0.5493**	**0.4978**	**0.4604**
10 nits	x	x	x	**0.4765**	**0.5105**	**0.4541**	**0.4220**
100 nits	x	x	x	x	0.5347	**0.4419**	**0.4050**
300 nits	x	x	x	x	x	**0.3345**	**0.3167**
500 nits	x	x	x	x	x	x	**0.2861**

Table 8.3: Experiment 2, Coefficient of Variation for ΔYCbCr at each of the tested luminance level ranges.

ΔYCbCr	**0.01 nits**	**0.1 nits**	**1 nits**	**10 nits**	**100 nits**	**300 nits**	**500 nits**
0.01 nits	0.4307	0.5955	0.9748	1.0650	1.2621	1.1835	1.3339
0.1 nits	x	0.5760	0.8851	0.9464	1.1416	1.0781	1.2342
1 nits	x	x	0.8034	0.8273	1.0255	0.9801	1.1428
10 nits	x	x	x	0.7755	0.9838	0.9427	1.1060
100 nits	x	x	x	x	1.0506	0.9764	1.1313
300 nits	x	x	x	x	x	0.8671	1.1723
500 nits	x	x	x	x	x	x	1.2565

Table 8.4: Experiment 2, Coefficient of Variation for ΔL*a*b* at each of the tested luminance level ranges.

ΔL*a*b*	**0.01 nits**	**0.1 nits**	**1 nits**	**10 nits**	**100 nits**	**300 nits**	**500 nits**
0.01 nits	1.5192	1.4336	1.1237	0.9504	0.9707	0.9462	0.9091
0.1 nits	x	0.6029	0.6395	0.7578	0.8748	0.8595	0.8262
1 nits	x	x	0.5277	0.6162	0.7221	0.7157	0.6956
10 nits	x	x	x	0.4959	0.5701	0.5693	0.5626
100 nits	x	x	x	x	0.4463	0.4462	0.4503
300 nits	x	x	x	x	x	0.4070	0.4068
500 nits	x	x	x	x	x	x	0.3925

Table 8.5: Experiment 2, Coefficient of Variation for ΔE2000 at each of the tested luminance level ranges.

ΔE2000	**0.01 nits**	**0.1 nits**	**1 nits**	**10 nits**	**100 nits**	**300 nits**	**500 nits**
0.01 nits	1.4293	1.3924	1.1941	1.0832	0.9925	0.9678	0.9186
0.1 nits	x	0.5206	0.5267	**0.5191**	**0.5823**	0.7116	0.6900
1 nits	x	x	0.4891	**0.4939**	0.5661	0.7018	0.6788
10 nits	x	x	x	0.5054	0.5978	0.7388	0.7033
100 nits	x	x	x	x	0.6369	0.7699	0.7220
300 nits	x	x	x	x	x	0.8487	0.7566
500 nits	x	x	x	x	x	x	0.5929

Table 8.6: Experiment 2, Coefficient of Variation for ΔxyY at each of the tested luminance level ranges.

ΔxyY	**0.01 nits**	**0.1 nits**	**1 nits**	**10 nits**	**100 nits**	**300 nits**	**500 nits**
0.01 nits	0.4264	0.5816	0.6717	0.7328	0.8011	0.8687	0.9230
0.1 nits	x	0.6450	0.7148	0.7615	0.8113	0.8779	0.9428
1 nits	x	x	0.7956	0.8249	0.8635	0.9338	1.0078
10 nits	x	x	x	0.8653	0.8869	0.9721	1.0663
100 nits	x	x	x	x	0.8180	0.9964	1.1436
300 nits	x	x	x	x	x	1.1841	1.2802
500 nits	x	x	x	x	x	x	1.3462

Table 8.7: Experiment 2, Coefficient of Variation for ΔCIECAM02-UCS at each of the tested luminance level ranges.

ΔCIECAM02-UCS	**0.01 nits**	**0.1 nits**	**1 nits**	**10 nits**	**100 nits**	**300 nits**	**500 nits**
0.01 nits	1.4427	0.6877	0.7774	0.7663	0.8635	0.8945	0.8347
0.1 nits	x	1.3239	0.6889	0.6616	0.7583	0.7951	0.7439
1 nits	x	x	0.6096	0.5264	0.6357	0.6854	0.6466
10 nits	x	x	x	1.2456	0.5793	0.6281	0.5939
100 nits	x	x	x	x	**0.4418**	0.5976	0.5670
300 nits	x	x	x	x	x	0.6302	0.5724
500 nits	x	x	x	x	x	x	0.6524

When computing the CV for these luminance ranges, the new JND would be computed as the mean of all the individual JNDs for colours in Colour Difference Test II that fall within the region. As can be seen from the tables, the CVs of the colour spaces/difference methods are almost never better than the CV that was computed for the particular luminance value itself. In other words, the results from the above tables indicate that no benefit would be gained by using a JND value that is common to an entire luminance range for most cases. The perceptual uniformity would actually be lower when using a JND computed as the mean of all the colours for an entire luminance range. However, the tables do show one exception in that the CV does improve for the particular luminance range of 10 nits to 100 nits if one uses a common JND as computed using DeltaE2000. The results also show that while ICtCp dominates in its perceptual uniformity for individual luminance values, CIECAM02-UCS does outperform ICtCp in the 100 nits luminance. Table 8.8 lists the corresponding JND values for each of the best performing colour spaces in each luminance region. The italicized entries correspond to DeltaE2000 while the bolded entry corresponds to the CIECAM02-UCS. All the other entries correspond to the ICtCp colour space JNDs.

Table 8.8: Experiment 2, JNDs at each of the tested luminance level ranges for each best performing colour space.

Method	**0.01 nits**	**0.1 nits**	**1 nits**	**10 nits**	**100 nits**	**300 nits**	**500 nits**
0.01 nits	2.7888	3.3498	4.4816	5.9746	7.1327	7.9225	8.3133
0.1 nits	x	3.9108	5.3280	*3.1861*	*3.4477*	8.9493	9.2340
1 nits	x	x	6.7452	*3.4817*	9.6546	10.2089	10.2987
10 nits	x	x	x	10.4535	11.1093	11.3635	11.1870
100 nits	x	x	x	x	**0.2090**	11.8185	11.4316
300 nits	x	x	x	x	x	11.8718	11.2648
500 nits	x	x	x	x	x	x	10.6578

Using the above data, we propose several colour difference algorithms that incorporate the above principles in the following subsections.

8.1.1 ICtCp Colour Differencing with a Fixed JND per Luminance Range

As indicated from the colour difference experiment results described in Chapter 6, ICtCp was the best performing colour space for the entire luminance range with a CV of 0.6182. Also, as seen from Table 6.2, the ICtCp colour space has significantly better local perceptual

uniformity at each of the luminance ranges. Therefore, it seems very likely that a colour difference measure that computes the colour difference in the ICtCp colour space, but utlizes a JND dependent on the luminance range, will perform significantly better than using the ICtCp colour space with a fixed JND to determine colour difference.

Table 8.2 indicates the CV values as computed in the ITP colour space for various luminance ranges. As seen from the same table, the CV values is lowest in the ICtCp colour space when the JND is limited to the fixed luminance range. Thus, given two colour values, we propose the following algorithm that determines whether the two colours are perceptually different by performing colour difference in the colour space ICtCp with **F**ixed **JND** for an entire **L**uminance **R**ange (which we will denote as **ICtCp F-JND-LR**).

Algorithm 1: ICtCp F-JND-LR

Result: dif = 1 if perceptually different, or 0 otherwise. difm will store the computed colour difference as a factor of the JND.

difm := colour difference magnitude;
lum := luminance of reference/test colour pair in *nits*;
cdif := computed colour difference between the two colours in ICtCp space;
L_i := luminance range i;
JND_i := *JND* of luminance range i;
if *lum in* L_i **then**
    if $cdif \geq JND_i$ **then**
        $dif = 1$;
    else
        $dif = 0$;
    end
$difm = cdif \ / \ JND_i$;

In this algorithm, the JND_i corresponds to the computed JND entries for ICtCp colour space as given in Table 6.2 for each luminance value of 0.01, 0.1, 1 nits, 10 nits, 100 nits, 300 nits, 500 nits. Since the colour difference tests corresponding to the luminance range specified in Table 6.2 was performed with the corresponding luminance values of the tested colour pairs strictly limited to the luminance values just listed, the corresponding luminance ranges were defined as $L = \{[0, 0.01+\Delta_{0.01}], (0.01+\Delta_1, 0.1+\Delta_{0.1}], (0.1+\Delta_{0.1}, 1+\Delta_1], \cdots, (300+\Delta_{300}, 500+\Delta_{500}], (500+\Delta_{500}, 10,000]\}$. The terms $\Delta_{0.01}, \cdots, \Delta_{500}$ were used as a tolerance since it is naturally expected that the computed JND_i for a corresponding luminance, say 0.01 nits, must also hold for proximate luminance ranges that are close to it. We tested the proposed algorithm with the Δ values set to 50% of its corresponding JND_i luminance level, as well as a $L_i + \Delta_{Li}$ that corresponds to the midpoint between L_i

and L_{i+1}.

The *difm* in the algorithm is the computed colour difference as a factor of the JND, and is meant to be representative of the magnitude of perceived colour difference. A value less than 1 would indicate that there is no perceivable colour difference while values greater than or equal to 1 would indicate perceivable colour difference. A pair of colours that gives a *difm* value of 2 for example would be considered as perceptibly more different than a pair of colours that gives a *difm* of 1.

However, as indicated in the discussion before the start of this section in light of the results from Tables 8.1, 8.2, 8.3, 8.4, 8.5, 8.6 and 8.7, there is the possibility of improved performance by combining the ICtCp space based prediction with that of the DeltaE2000 colour space and CIECAM02-UCS. Thus, we propose the following two algorithms with minor changes to the ICtCp F-JND-LR proposed above.

The first, which we will denote ICtCp-CIECAM02 will simply evaluate the colour difference for the particular luminance range of 100 nits in the CIECAM02-UCS space, and determine whether the colour difference is perceptible and its magnitude according to the CIECAM02-UCS JND data. Algorithm 2 describes ICtCp-CIECAM02.

Algorithm 2: ICtCp-CIECAM02-UCS

Result: dif = 1 if perceptually different, or 0 otherwise. difm will store the computed colour difference as a factor of the JND.

difm := colour difference magnitude;
lum := luminance of reference/test colour pair in *nits*;
cdif := computed colour difference between the two colours in ICtCp space, except if the colours are within 10 nits to 100 nits;
L_i := luminance range i;
JND_i := *JND* of luminance range i;

```
if lum in L_i then
    if 10+ Δ10 ≥ lum ≤ 100 + Δ100 then
        cdif := computed colour difference in CIECAM02-UCS;
        if cdif ≥ JND_CIECAM then
            dif = 1;
        else
            dif = 0;
        end
        difm = cdif / JND_CIECAM;
    else
        cdif := computed colour difference in ICtCp;
        if cdif ≥ JND_i then
            dif = 1;
        else
            dif = 0;
        end
        difm = cdif / JND_i;
    end
```

The JND_{CIECAM} refers to the JND as shown in Table 8.8 corresponding to CIECAM02-UCS (the bolded entry in the table). The Δ_{10}, and Δ_{100} are the Δ_{Li} values that were described in the previous algorithm.

The second, which we will denote ICtCp-DeltaE2000 will simply evaluate the colour difference for the particular luminance range from 10-100 nits in the DeltaE2000 space, and determine whether the colour difference is perceptible and the magnitude of the colour

difference according the DeltaE2000 JND data. The ICtCp-DeltaE2000 is given below.

Algorithm 3: ICtCp-DeltaE2000

Result: dif = 1 if perceptually different, or 0 otherwise. difm will store the computed colour difference as a factor of the JND.

difm := colour difference magnitude;
lum := luminance of reference/test colour pair in *nits*;
cdif := computed colour difference between the two colours in ICtCp space, except if the colours are within 10 nits to 100 nits;
L_i := luminance range i;
JND_i := *JND* of luminance range i;
if *lum in* L_i **then**
  if $10 \geq lum \leq 100$ **then**
    cdif := computed colour difference in DeltaE2000;
    if $cdif \geq JND_{\Delta E2000}$ **then**
      $dif = 1$;
    else
      $dif = 0$;
    end
    $difm = cdif \ / \ JND_{\Delta E2000}$;
  else
    cdif := computed colour difference in ICtCp;
    if $cdif \geq JND_i$ **then**
      $dif = 1$;
    else
      $dif = 0$;
    end
    $difm = cdif \ / \ JND_i$;
  end

The $JND_{\Delta E2000}$ refers to the JND as shown in Table 8.8 corresponding to DeltaE2000 (the italicized entry in the table). As mentioned earlier, all of the above proposed algorithms will be tested with the Δ values set to 50% of its corresponding JND_i luminance levels, as well as a $L_i + \Delta_{Li}$ value that corresponds to the midpoint between L_i and L_{i+1}.

8.1.2 ICtCp Colour Differencing with a Varying JND per Luminance

Given the evidence that the JND varies with luminance in the ICtCp colour space (as seen from our colour difference experiment results in Table 6.2), one would expect a colour difference algorithm that functions based on the JND corresponding to the exact luminance of the colours would be more accurate (compared to one that utilizes a fixed JND for an entire luminance range). However, since the colour difference tests we described were limited to specific luminance values, such an algorithm would only be possible if there were a clear relationship between the JND and luminance that could be determined from the data obtained from the colour difference tests. Figure 8.1 is a plot of the ICtCp JND vs. $\log_{10}(Luminance)$ as determined from Colour Difference Study II .

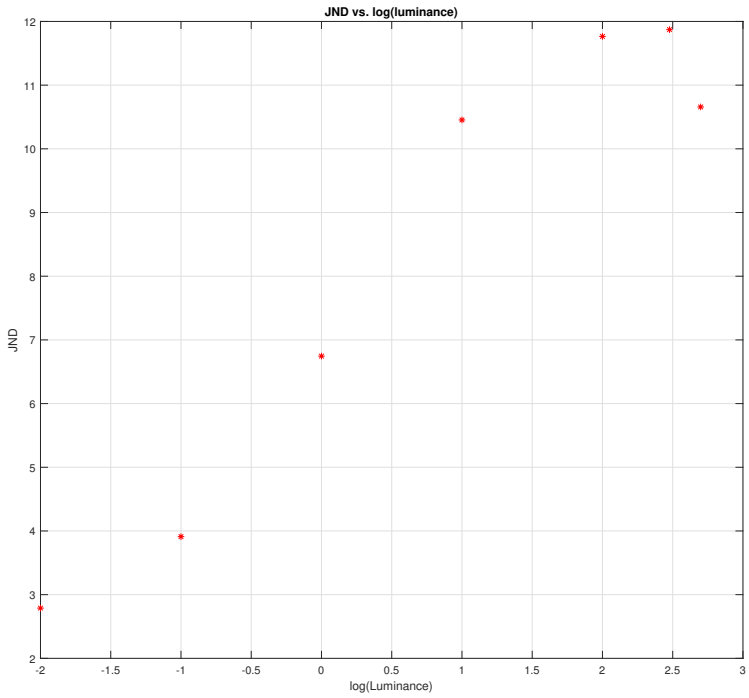


Figure 8.1: ICtCp JND from Colour Difference Study II vs. $\log_{10}(Luminance)$

It is clear from the distribution of points for the JND for 100 nits, 300 nits, and 500 nits that a simple linear relationship does not exist between the ICtCp JND and the luminance

for the entire HDR luminance range of 0 nits to 10,000 nits. Nevertheless, the distribution of the JND does seem capable of being approximated by a linear relationship over the range of 0 nits to 100 nits, although it is unclear from the data whether the linearity would hold right at a luminance of 100 nits. To determine a proper interval, we utilized the JND data collected from the surround colour difference study with the surround luminance set to 0 nits. The Figure 8.2 shows the JND for different luminance levels as determined from all the colour difference studies described in the previous chapters. From this distribution, it appears that the linear relationship between JND and luminance holds below 50 nits, and that a separate linear relationship exists between JND and luminance beyond 50 nits.

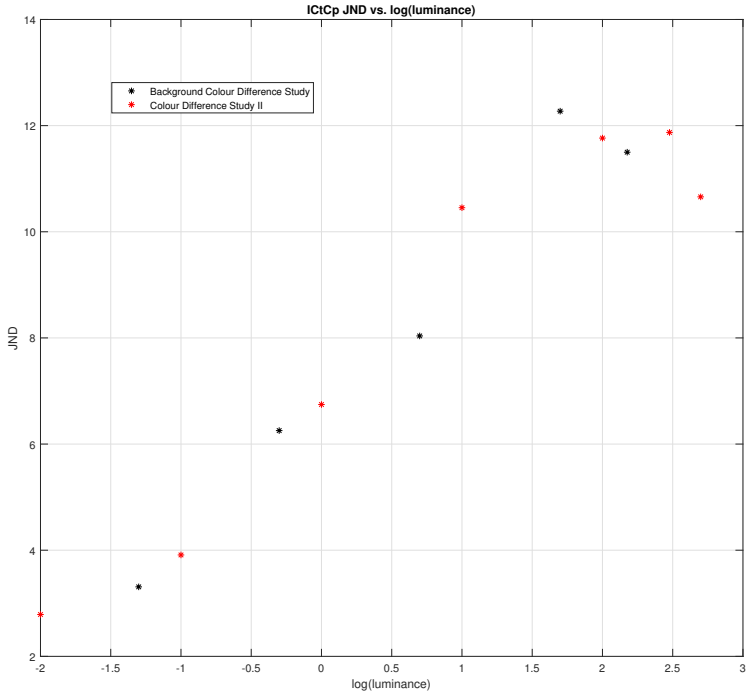


Figure 8.2: ICtCp JND from all Colour Difference Studies vs. Log(Luminance)

Thus, we propose an algorithm for determining colour difference based on a JND that directly varies with luminance by fitting a line of best fit for the JND points between 0 nits to 50 nits, and another line of best fit for the range of 50 nits to 500 nits. The algorithm operates in the ICtCp space with **V**arying **JND** using a **L**inear approximation based on

luminance (which we will denote as **ICtCp V-JND-L**).

Algorithm 4: ICtCp V-JND-L

Result: dif = 1 if perceptually different, or 0 otherwise. difm will store the computed colour difference as a factor of the JND.

difm := colour difference magnitude;
lum := luminance of reference/test colour pair in *nits*;
cdif := computed colour difference between the two colours in ICtCp space;
if $lum \leq 50$ **then**
    Estimate *JND* from 0 nits to 50 nits line of best fit;
    if $cdif \geq JND$ **then**
        $dif = 1$;
    else
        $dif = 0$;
    end
else if $lum > 50$ **then**
    Estimate *JND* from 50 nits to 10,000 nits line of best fit *JND*;
    if $cdif \geq JND$ **then**
        $dif = 1$;
    else
        $dif = 0$;
    end
$difm = cdif \ / \ JND$;

Figure 8.3 shows the lines of best fit used in the above proposed algorithm.

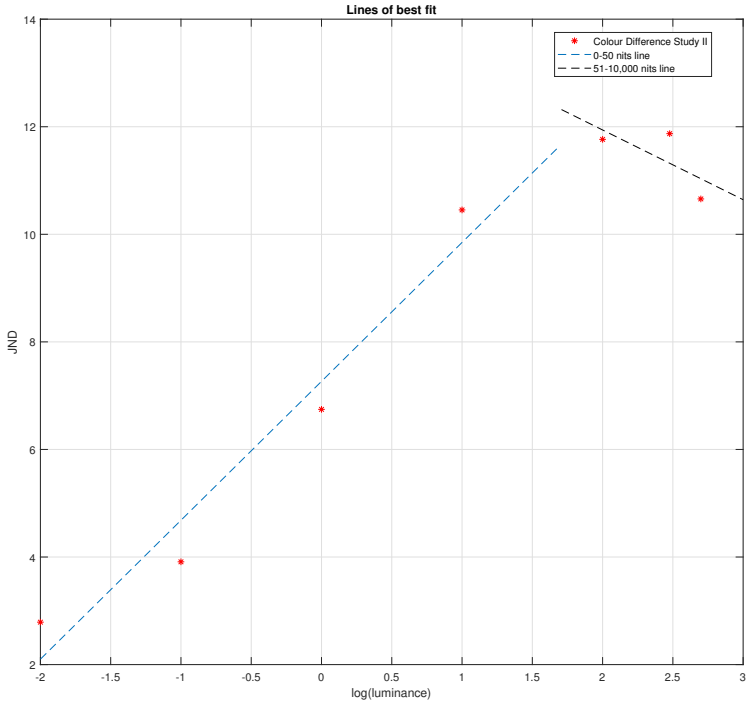


Figure 8.3: Lines of best fit for ICtCp JND data from Colour Difference Test II

As can be seen from this figure, there is a discontinuity in the JND at 50 nits, and depending on the line of best fit used to evaluate the JND, one would obtain a different value for the JND. It should be noted that the linear solution seen above was constructed by observing the data given in Figure 8.2. The data from the figure indicated a separate linear relationships for colour values with luminance under 50 nits and another for greater than 50 nits. For this reason, the discontinuity at 50 nits was allowed to remain rather than extend the line with the increasing slope to obtain a piecewise continuous function. The effect of this discontinuity will be investigated in the testing and analysis section. However, ideally we would like to have a continuous function that produces a JND for the entire luminance range (at least the range of 0 nits to 500 nits for which the experimental data is available). Thus, we propose a second algorithm that utilizes a second degree polynomial

of best fit (denoted **ICtCp V-JND-NL**).

Algorithm 5: ICtCp V-JND-NL

Result: dif = 1 if perceptually different, or 0 otherwise. difm will store the computed colour difference as a factor of the JND.

difm := colour difference magnitude;
lum := luminance of reference/test colour pair in *nits*;
cdif := computed colour difference between the two colours in ICtCp space;
Estimate *JND* for *lum* using 2nd degree Polynomial;
if $cdif \geq JND$ **then**
    $dif = 1$;
else
    $dif = 0$;
end
$difm = cdif \ / \ JND$;

The polynomial of best fit is shown in Figure 8.4. It should be mentioned that such a fit does seem inaccurate given the observable trend from the experiment results obtained for the 50 nits to 500 nits luminance range as seen from Figure 8.2. As seen from the figure, the JND does seem to show a decreasing trend as luminance continues to increase beyond 50 nits, while the best fit polynomial will not have such a rapid descent to match this trend. However, this is a limit of our available data and the observable trend does suggest that one would expect this algorithm to perform worse when determining colour difference for higher luminance colour pairs.

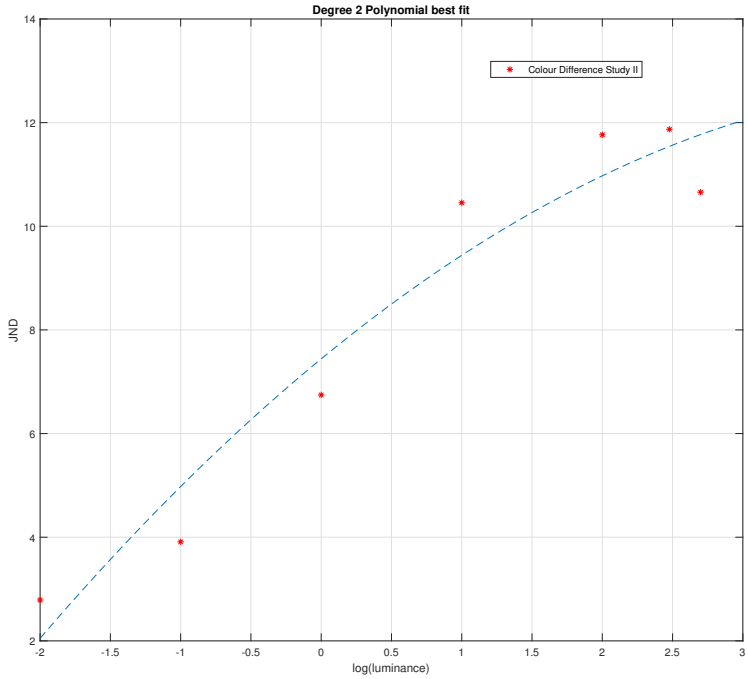


Figure 8.4: 2nd Degree Polynomial approximation for ICtCp JND data from Colour Difference Test II

8.2 Neural Network Based Colour Difference Measure

Another interesting way to employ the data obtained from the colour difference experiments described in previous chapters is to utilize a neural network approach. However, given the complexity of the problem, the available data points (245 from Colour Difference Test II and 72 points from the Colour Difference Test III to study the effect of surround luminance) for training, verification, and testing would seem limited. We overcome this limitation as follows.

We reserve the 72 data points from the colour difference study which focused on the effect of surround luminance for the testing and analysis of the performance of all the approaches proposed in this chapter. Taking the 245 data points from Colour Difference

Study II, we generate additional data points under the assumption that the computed JND for a particular reference colour holds true in its perceptual effect for at least up to a colour difference of 2 JND. Stated in another sense, implies that the JND of the test colour point located at a single JND distance from the reference colour would also possess the same JND. This would seem a reasonable assumption since the computed JND not only applies to the reference colour for which it was computed, but could also be viewed in reverse as the JND for the test colour that was located at a single JND distance from the reference colour.

Thus, we sample as many points as required along a straight line oriented in the same direction as that used to sample the test colour points for that particular reference colour in the experiment, limiting the interval to be within a distance of 2 JNDs from the reference colour point. For each of the sampled colour points, we compute the distance as a ratio of the computed JND for that particular reference. In this way, each of our data entries will consist of a reference colour, a test colour point, and the distance to the test colour point as a ratio of the JND (which would be a value in the interval between 0 and 2). We decided to train separate networks based on each of the colour spaces and colour difference methods, and as such, we created multiple sets of training data using the above approach for each colour space and difference method.

In this manner, we were able to obtain 14000 colour pairs with a corresponding distance (computed as a ratio of the JND for that particular reference colour). We then used 60% of the data for training, 20% for verification, and 20% for testing. It should be mentioned here that the division of the training, verification and testing data were performed in such a way that a reference or test colour would not appear in more than one set. This ensured that during training, the network was not exposed to the same reference colour or test colour, and thus the verification set would also guarantee some level of generalizability of the network performance. This is especially important since the neural network approach would be an attempt to predict colour difference between more than billions of colours using a relatively small subset of data (245 colour points).

We used a feed-forward neural network for our task, and experimented with various numbers of hidden layers and number of nodes per hidden layer using MATLAB. The inputs to the neural network were the reference colour tri-stimulus values for a colour space, the test colour tri-stimulus values for a colour space, and the computed distance between the two colours in the colour space. For the colour difference methods such as ΔE2000, the tri-stimulus values were the L*a*b* values, the space on which the difference method was based. It may appear at first to be questionable that the computed distance is used as an input to the neural network, when the distance information is already contained in the tri-stimulus values that are already passed in as inputs. However, the distance

measurements for certain colour spaces and colour difference methods (like ΔE2000) are complex. While a complex network could still approximate such a computation for distance, it was considered beneficial for the sake of network simplicity and being able to find an optimal network structure quickly, to simply pass in the computed distance as an input.

After much experimentation, we found success with a feed-forward neural network with a single hidden layer of 10 nodes. To better understand the relative performance of each network, we then used the original reference and test colour pairs that were used to form the 20% testing data set to determine the JND as computed by our trained neural networks using each colour space and colour difference method. The results are shown in Table 8.9.

Table 8.9: CV of each Neural Network (NN)

NN Colour Space	**CV**
ICtCp	0.4505
RGB	0.2599
L*a*b*	0.2790
E2000	0.2824
YCbCr	0.2678

As can be seen from Table 8.9, all of the neural networks do perform better than the CV we estimated from our Colour Difference Test II for the existing colour spaces. ICtCp which performed the best had a CV of 0.6182. In contrast, the best performing neural network trained using data from the RGB colour space shows a CV of 0.2599. The neural network based on ICtCp does have the highest CV value of 0.4505, but it should be noted that it is quite possible that a certain random set of the data for training would have just as easily produced a network similar in performance to the RGB network. However, the RGB neural network does prove attractive since most of the video and image data today is stored as RGB or YCbCr data. The conversion from YCbCr to RGB is quite inexpensive, and therefore the RGB neural network (and for the same reasons, the YCbCr network as well) can be easily applied without much pre-processing of the data.

All of this being said, it must be noted that the training, verification and testing of the neural network has been done in this section using the data from the Colour Difference Test II. All of the 14000 test points used in the preparation of the network were limited to luminances of 0.01 nits, 0.1 nits, 1 nits, 10 nits, 100 nits, 300 nits, and 500 nits. Thus, the low CV of the RGB network above may not hold for other colours that fall within other luminance values apart from that which were used for the training, verification and testing in this section. This question of how well this trained network will generalize at predicting

colour differences for a larger potion of the entire RGB colour space will be studied and addressed in the testing and analysis section of this chapter.

8.3 Testing & Analysis

As indicated in the previous subsections, the colour data from the Colour Difference Study III (performed to study the effect of surround luminance) will be used for the testing purposes of the above proposed algorithms and approaches. We will specifically be using the data collected with the 0 nits surround luminance from the surround luminance study (Colour Difference Study III), since this is the same condition under which Colour Difference Studies I and II were performed. The data from the surround luminance study is of further interest since the luminance values of the colours are different from that of the other colour difference study data that was used to formulate the proposed algorithms or train the neural networks. While the colours in the Colour Difference Study II had luminance values of 0.01 nits, 0.1 nits, 1 nits, 10 nits, 100 nits, 300 nits, and 500 nits, the surround luminance study luminance values are 0.05 nits, 0.5 nits, 5 nits, 50 nits, 150 nits, 300 nits. Apart from the 300 nits luminance, the other values would be outside of the data that was used to formulate the approaches so far mentioned in this chapter. Therefore, the data set can be vital to understanding the generalizability of the proposed algorithms.

From the perspective of testing, there are two types of functionality that must be tested in regards to the proposed algorithms and approaches. Given a pair of colours, the algorithm or approach should

1. have the ability to identify whether the two colours are perceptually different, and
2. provide a meaningful value that quantifies the *extent* of the perceivable difference between the two colour values.

Both of these qualities are of importance from the perspective of image and video quality assessment as explained in the introduction of this chapter. We first concentrate on testing the ability to identify whether the two colours are perceptually different. To do so, we utilized the 72 reference colours and JND data that we possess from the surround luminance study in the following manner. Knowing the JND of the reference colours, we sampled 1300 points from within 1 JND, and another 1300 points from within the interval of 1 JND to 2 JND. The sampling was performed along a straight line oriented in the same direction as the location of the test colour points in the original surround luminance test. The

sampled colour pairs within 1 JND are the colour pairs that appear perceptually to be the same. The colour pairs with a JND more than 1 are considered to appear as perceptually different. The goal of this particular test then is to see how well the proposed algorithm or approach correctly determines whether a given colour pair is perceptually different or the same.

Table 8.10 summarizes the performance of each of the proposed methods and existing colour spaces and colour differencing methods when it comes to correctly determining whether a colour pair appears perceptually different or the same.

Table 8.10: Performance at correctly identifying colour pairs as similar or different

Method	% correct no difference	% correct is different	Overall Performance
ICtCp	77.57	97.09	87.33
ΔE2000	85.64	92.28	88.96
La*b*	74.71	88.31	81.51
CIECAM2000	72.56	89.36	86.41
RGB	76.56	92.02	84.29
YCbCr	75.56	92.21	83.89
ICtCp F-JND-LR (0.5)	88.47	96.70	92.58
ICtCp F-JND-LR (mp)	80.02	98.48	89.25
ICtCp-CIECAM02 (0.5)	84.86	97.36	91.11
ICtCp-CIECAM02 (mp)	77.56	98.74	88.15
ICtCp-DeltaE2000 (0.5)	85.17	97.60	91.39
ICtCp-DeltaE2000 (mp)	78.17	98.30	88.23
ICtCp V-JND-L	87.86	97.34	92.60
ICtCp V-JND-NL	86.55	97.47	92.58
Neural Net (JND 0.9)	79.17	96.72	87.95
Neural Net (JND 0.95)	82.55	96.07	89.31
Neural Net (JND 1)	85.56	95.31	90.43
Neural Net (JND 1.05)	92.24	93.24	92.74
Neural Net (JND 1.1)	90.16	92.67	91.42

In the neural network approach, we make a distinction of the network based on the threshold above which two colours are considered different. Recall that the neural network was trained to output the distance as a fraction of the JND. Ideally, the neural network would output a distance of 1 for a colour pair located exactly within a JND. But, to accommodate for slight error, we considered the case where the 1 JND corresponds to 0.9, 0.95, 1, 1.05 and 1.1 of the neural network output. These are labeled as Neural Net (JND 0.9), (JND 0.95), (JND 1), (JND 1.05), and (JND 1.1) respectively in the results table. We

also have two entries for the ICtCp F-JND-LR algorithm postfixed as *'(0.5)'* and *'(mp)'*, and the same for ICtCp-CIECAM02 and ICtCp-DeltaE2000. These post fix labels refer to the different notions of Δ_{Li} introduced in for the luminance range algorithms in this chapter. The Δ_{Li} acts as the tolerance for the specified luminance range. The *'(0.5)'* version refers to the case where the algorithm uses an upper-bound that is the value in the luminance range plus an additional tolerance of 0.5 times the luminance value of the upper-bound of the range. The *'mp'* version refers to the case of using the midpoint between the next luminance regions luminance value and the luminance value of luminance region of interests as the upper-bound for the luminance region.

According to the results, almost all of the proposed algorithms and approaches outperform the existing colour spaces and colour difference methods at the task of detecting whether two colour values appear similar or different to the human visual system. The neural network (JND 1.05) approach performs best overall, correctly predicting 92.74% of the 2600 tested colour pairs. However, it does have a lower performance with correctly identifying colour pairs that are perceptually different (93.24%) compared to the ICtCp V-JND-L which correctly predicted 97.34% of the 1300 colour pairs as perceptually the same. However, the ICtCp V-JND-L with 87.86% predictions correct when it came to similar test colour pairs, does under-perform compared to the Neural Net (JND 1.05) and Neural Net (1.1).

We also see from the results for ICtCp F-JND-LR (0.5) and ICtCp F-JND-LR (mp), that ICtCp F-JND-LR (0.5) performs much better. Surprisingly, this method performs as well (92.58% overall performance compared to 92.60%) as ICtCp V-JND-L which operates by using a variable JND based on luminance rather than a fixed JND for an entire range. This may at first glance indicate that the JND enjoys some level of stability for a given luminance region that is within one order of magnitude in the luminance scale. However, referring to the Figure 8.5 which shows the plot of all the JND data we have from all colour tests (including the testing data) together with the lines of best fit (that were constructed for ICtCp V-JND-L), we see that the JND certainly does not seem to remain stable even within one order of magnitude in the luminance range. At the same time, the luminance range algorithms that use a hybrid of colour spaces such as ICtCp-CIECAM02 or ICtCp-DeltaE2000 do not perform as well as the other proposed algorithms, but still do outperform the existing colour spaces.

Next we turn to the ability to measure the magnitude of perceivable colour difference. Here, we compute the Pearson Correlation Coefficient (PLCC), Spearman Correlation Coefficient (SRCC) between the predicted distance by each algorithm as a fraction of the estimated JND vs. the distance as a fraction of the actual JND from the experimental results for the colour pairs. We also computed the RMSE between these two quantities.

The results are shown in Table 8.11.

Table 8.11: Performance at correctly identifying the magnitude of perceivable colour difference

Method	PLCC	SRCC	RMSE
ICtCp	0.7671	0.8577	5.4082
ΔE2000	0.6303	0.7993	14.61
La*b*	0.6622	0.7296	10.16
CIECAM2000	0.6602	0.7466	33.79
RGB	0.5324	0.6882	8.10
YCbCr	0.5502	0.6996	8.15
ICtCp F-JND-LR (0.5)	0.8027	0.8967	2.46
ICtCp F-JND-LR (mp)	0.7913	0.8907	2.96
ICtCp-CIECAM02 (0.5)	0.7645	0.8722	3.62
ICtCp-CIECAM02 (mp)	0.7470	0.8644	4.17
ICtCp-DeltaE2000 (0.5)	0.7463	0.8858	3.3847
ICtCp-DeltaE2000 (mp)	0.7170	0.8715	5.7865
ICtCp V-JND-L	0.8091	0.9001	2.40
ICtCp V-JND-NL	0.7979	0.8946	2.58
Neural Net (JND 0.9)	0.5591	0.7254	6.5775
Neural Net (JND 0.95)	0.5591	0.7254	6.9275
Neural Net (JND 1)	0.5591	0.7254	7.2685
Neural Net (JND 1.05)	0.5591	0.7254	7.5983
Neural Net (JND 1.1)	0.5591	0.7254	7.9156

As seen from the results, the neural network methods do show a weak correlation with the actual distance measurement as a fraction of the JND, indicating weaker performance. The same holds true for the existing colour spaces and methods. In contrast, the proposed algorithms that utilize a JND based on variable luminance and luminance ranges do show considerably higher PLCC and SRCC values. As seen from the previous table on colour difference prediction performance, the ICtCp F-JND-LR methods do also perform comparable to the variable luminance methods. The best performing from the group is ICtCp V-JND-L with a PLCC of 0.8091 and a SRCC of 0.9001. The RMSE for this algorithm is 2.4 as well.

Next, we also compute the correlation of coefficient (CV) to obtain insight on the perceptual uniformity of these proposed algorithms (and existing colour spaces and methods) with respect to the test data set. The results are shown in Table 8.12

Table 8.12: CV of existing colour spaces and proposed methods for the testing data set

Method	**CV**
ICtCp	0.5138
RGB	0.8916
L*a*b*	1.1366
E2000	0.7918
YCbCr	0.8474
CIECAM02-UCS	0.7637
ICtCp F-JND-LR (0.5)	0.3402
ICtCp F-JND-LR (mp)	0.3552
ICtCp-CIECAM02 (0.5)	0.6835
ICtCp-CIECAM02 (mp)	0.5690
ICtCp-DeltaE2000 (0.5)	0.3503
ICtCp-DeltaE2000 (mp)	0.3998
ICtCp V-JND-L	0.3354
ICtCp V-JND-NL	0.3461
Neural Net (JND 1)	0.2973

We do see from the results that the performance of existing colour spaces are significantly better for the testing data set compared to the CV values from Colour Difference Study II (given in Table 6 in Chapter 6). All the proposed methods, apart from the ICtCp-CIECAM02 algorithms, show a better CV than the CV of the existing colour spaces and colour difference methods for the same data. The Neural Net (JND 1) shows the best CV of 0.2973. (Note that this CV remains unchanged for the other 4 networks with JND 0.9, 0.95, 1.05, and 1.1, and hence separate entries are not included in the Table.) The CV for the proposed luminance range based and variable luminance based algorithms are also improved (CV of ~ 0.34) compared to the existing methods and colour spaces.

Overall, ICtCp V-JND-L seems to perform best at predicting colour difference (92.60%), as well as estimating the colour difference magnitude (PLCC 0.8091, SRCC 0.9001, RMSE 2.40). The lines of best fit used in the ICtCp V-JND-L algorithm are shown in Figure 8.5 and the points from both the Colour Difference Test II JND data that were utilized to determine the lines of best fit and the test data set JND are also shown.

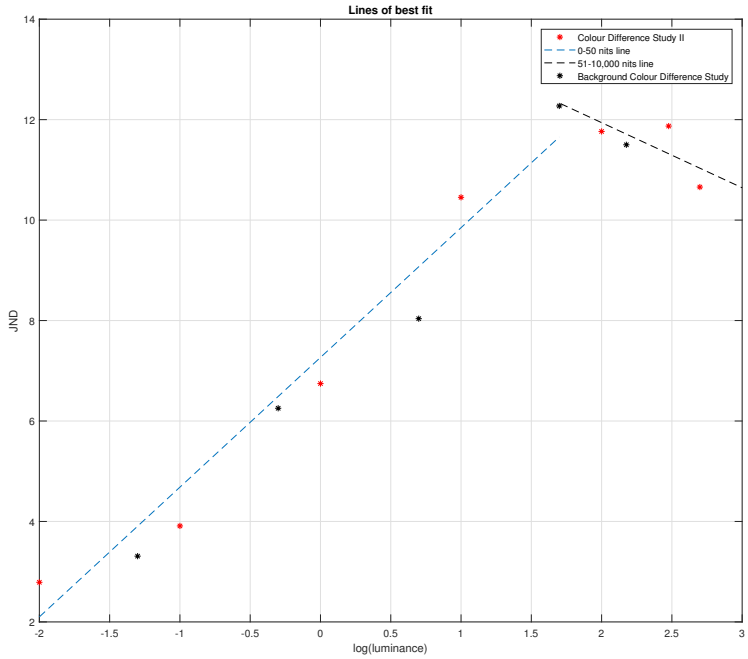


Figure 8.5: JND vs Log(Luminance) for all tests

As seen from the plot, the JNDs for the test data (from Colour Difference Test III) do fit the test data reasonably well. In fact, recomputing the lines of best fit using all the data, and then running the algorithm resulted in an overall performance of 92.49% (slightly less than its current performance), and PLCC 0.7979, SRCC 0.8946, RMSE 2.58, (compared to PLCC 0.8091, SRCC 0.9002, RMSE 2.41 as implemented), and a CV of 0.3332 which are hardly any different from the existing performance. This does suggest that the best performance that could be obtained by the approach of this algorithm is reached, and adding more JND data points will not increase it significantly. We postulate two reasons for this behavior.

1. The actual variability of the JND with luminance does not follow a simple linear function in the ICtCp colour space.

2. The JND varies significantly depending on the particular colour in the ICtCp colour space, even within the same luminance.

The above aspects do not come as a surprise since it was seen from the CV results summarized in previous chapters for the Colour Difference Tests, that even for a particular luminance, the CV was still considerably higher. In fact, the smallest CV value found was 0.3536, and this was for the case of 0.01 nits luminance colour pairs in the ICtCp colour space in Colour Difference Study II. This CV is still worse than that of the CV of the ICtCp V-JND-L algorithm (0.3354). From this fact, it also follows that the JND within the ICtCp colour space (or all the existing colour spaces considered) is not simply a function of just luminance. The JND variation with the colours within a luminance is still significantly large enough that it appears that one cannot simply approximate the JND as a function only of luminance, so that some error in predicting colour using a method that relies on luminance alone would be unavoidable. This seems to confirm the now obvious fact that the existing colour difference and colour spaces are still quite limited in their success at mapping colours to a perceptually uniform space. The deficiencies in this regard cannot be completely overcome by correcting the JND based on luminance.

It should also be pointed out that ICtCp V-JND-NL, ICtCp F-JND-LR offer comparably good performance to ICtCp V-JND-L as well. ICtCp V-JND-NL can be a more attractive algorithm in comparison to ICtCp V-JND-L due to the absence of the discontinuity at 50 nits. However, it does have its own shortcoming in that from the data, the JND trend for luminances above 50 nits seems to follow a linearly descending pattern. The 2nd degree polynomial of best fit that is used as shown in Figure 8.4 clearly does not follow this trend. So its effectiveness at predicting colour difference and difference magnitudes at higher luminances might be weaker. This would go undetected in our tests since the test data points were limited to 300 nits and below.

8.4 Summary & Conclusions

In this Chapter, we have presented our work on several new colour difference algorithms aimed toward colour difference prediction. The goals of the algorithms were to predict whether two colours appear different to our human visual system, and also to quantify the magnitude of the colour difference as well.

The proposed algorithms ICtCp V-JND-L, ICtCp V-JND-NL, and ICtCp V-JND-NR showed a 92.6% performance at correctly predicting whether a colour difference is perceptible, and a 0.8 PLCC, 0.9 SRCC with a RMSE of 2.4 when it came to predicting the magnitude difference. The same algorithms showed an improved CV (a CV of ~ 0.34) compared to existing methods. The Neural Net (JND 1.05) approach offered the best

performance at prediction (92.74%), but was considerably weaker at predicting the magnitude of colour difference with respect to even some of the existing colour spaces and colour difference methods. However, the Neural Net approach did show the best CV of 0.2973.

It was also argued in this chapter that the performance of the proposed algorithms could not be likely improved by collecting more JND data points to obtain a better line (or polynomial) of best fit. This was clear from the results obtained by modifying the algorithms to take into account the testing data, and retesting against the testing data itself. Ideally, the JND prediction for a given colour by the algorithm would have to not only take into account the luminance, but the colour as well. It does appear to be the case that the existing colour difference and colour spaces are still far limited in their success at mapping colours to a perceptually uniform space, and a simple JND correction based on luminance is insufficient to fully compensate for the shortcomings.

However, since the proposed algorithms do offer a significant improvement over the existing colour spaces and methods, any colour quality assessment algorithm would be more accurate when using one of the proposed algorithms in this chapter for computing and quantifying the colour difference.

Chapter 9

Incompatibility of Modern IQA Philosophy for HDR WCG Content, and Subjective Testing

For HDR WCG content, the preservation of the colour quality of the image is just as vital as that of the structural information. The HDR WCG eco-system is intended to reproduce the content as prepared by the content producer as accurately as possible [7][8] [9]. The shades of colours that are carefully chosen and the selected highlights and the details emphasized in the darker regions of the image by the colorists (the content producer in this case) should all be left untouched during the distribution of the content. We shall denote the colour choices and other details of the image as produced by the colorist as the **creative intent**. Now, it is also true that the colours within and image can be changed in a manner that a the HVS would still recognize the image as a pristine image. Similarly, it is not self-evident by looking at an image whether it is lack of highlights or shadow detail is intentional or the result of a loss during the distribution process. Therefore, it is true to say that the assessment of the quality of a HDR WCG images must necessarily take into account the creative intent of the content producer.

This chapter will first investigate the compatibility of the existing IQA models with requirement for HDR WCG IQA. This will be followed by the subject of colour quality assessment which is vital for HDR WCG IQA. Finally, we shall discuss the difficulty present in existing IQA subjective testing process under the current IQA paradigm.

9.1 Problem with Structural Similarity Models for HDR WCG IQA

Before we propose a novel algorithm for IQA of HDR WCG content, it is first important to clarify the deficiencies present in the popular structural similarity-based operating philosophy of many existing methods. This position that the current IQA models are deficient may come as a surprise to the reader since in Chapter 3, the subjective study presented indicates an acceptable level of performance with respect to VQA of HDR WCG content. Thus, it should be clearly stated here that the IQA performance measured in Chapter 3 is within the popular philosophical paradigm introduced in [85], and one that has become the foundation of many structural similarity based IQA methods since then.

Before the introduction of structural similarity-based IQA, the focus of the IQA methods were based on modeling the HVS for detecting the visible error in an image with respect to an original pristine image [88] [24][25] [42] [43] [49] [75] [83]. This philosophy can be said to weigh the fidelity of the distorted image to the pristine image. The latest proposed method at the time of writing to follow this paradigm for IQA can be said to be the Normalized Laplacian Pyramid method [37] from 2016. We shall address the problem with this method, and its predecessors shortly in the next section, while focusing on the more popular paradigm of structural similarity for this section.

In the introduction of structural similarity in [85], five reasons were given for departing from the error visibility model of IQA. These five reasons are listed as the "Quality Definition Problem", the "Suprathreshold Problem", the "Natural Image Complexity Problem", the "Decorrelation Problem", and The "Cognitive Interaction Problem" [85]. Here, we shall address each of these concerns as they pertain to HDR WCG IQA and IQA in general.

- *Quality Definition Problem*: Here it is argued in [85] that error visibility is not necessarily indicative of image quality. An example is provided that if the luminance of the entire image is increased by some fixed factor, the image would not be considered "objectionable". However, as explained above in regards to HDR WCG content, the goal of IQA is to verify that the **creative intent** of the content producers is preserved. Thus, increasing the luminance of an image by a noticeable scale factor is indeed to be considered objectionable under a philosophy of preserving creative intent.

- *Suprathreshold Problem*: It is argued in [85] that error sensitivity models used to measure the thresholds for just noticeable differences may not potentially apply to

large differences. However, this objection is fatal to not just the error visibility models, but to structural similarity models as well. If one is unable to discover a just noticeable difference that holds for large enough differences, then it would, for example, be incorrect to perform a numerical operation such as a subtraction operation with image data. Neither would it be reasonable to compute statistical quantities such as a mean, or standard deviation as is done in the computation of the structural similarity method proposed in [85] and subsequent developments under this philosophy. Performing such operations necessarily assume a uniform scale to exist for the image luminance values used in the computation. Or in other words, that the numerical differences between the coded image values are separated by a fixed just noticeable difference.

- *Natural Image Complexity Problem*: It is argued in [85] that the experiments performed to determine masking phenomena that may occur in natural images is modeled using studies performed on much simpler models. We once again point out that this objection is fatal for the structural similarity based IQA models as well. The masking can affect the visibility of structural components of an image as well. So as long as one incorporates masking effects into ones IQA model, this problem will be always present. Whether one utilizes a complex pattern or a simple one for experimentally accounting for the effects of masking (as in [87]), one would still be generalizing from a smaller subset of natural images or simple patterns to that of all possible natural images.

- *Decorrelation Problem*: It is argued in [85] that the use of aggregations such as the Minkowski metric for error pooling assumes wrongly that the pixels and image are statistically independent. This is most certainly a valid problem, and one which must be taken into account when devising a good IQA model. However, as we shall see in the subsequent discussion presented in this chapter and the IQA model we propose, one does not necessarily have to commit to error pooling in this manner under a error visibility model.

- *Cognitive Interaction Problem*: It is argued in [85] that a human subject may evaluate images differently depending on the situation. It is pointed out that different persons may fixate on different points of the image, different types of errors, and the information they possess in regards to the image. This is a very valid claim when one considers that the average human person may not necessarily perceive a difference of change in colour unless they are trained to spot such differences in an image. However, the entire driving force for HDR WCG IQA is the need to preserve **creative intent**. Thus, the colorist or content producer, who has spent the time developing

the content to have a certain appearance, is the human subject in this IQA model. We are no longer concerned with how the average human person will evaluate the image, but a professional who is well versed in operations such as colour correction and grading. Thus, this problem is irrelevant for the problem of HDR WCG IQA as it pertains to the fidelity of **creative intent**.

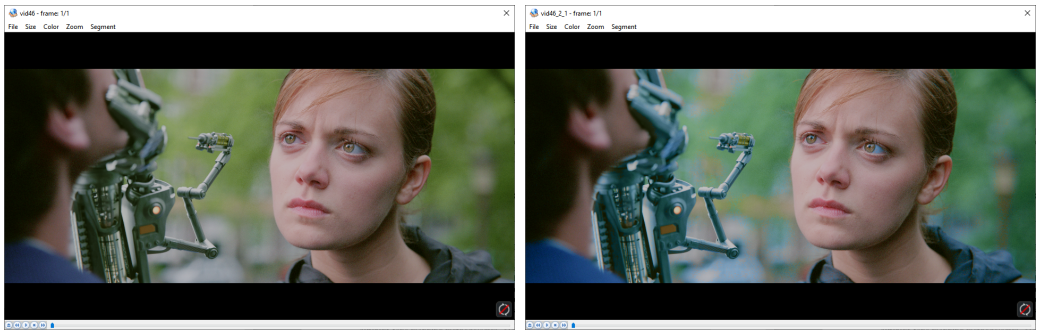
From the above comments addressing of the *Quality Definition Problem* and *Cognitive Interaction Problem*, it should be clear as to why the structural similarity methods are incompatible with the notion of evaluating the fidelity toward, or preservation of the creative intent, as required for HDR WCG IQA. Structural similarity methods focus on predicting the score a person would award a distorted image by drawing upon their experience of pristine natural images. This form of evaluation does not take into account the notion of creative intent.

9.2 Colour Quality Assessment

Another aspect that makes structural similarity methods, and also the error visibility methods of old and new like the normalized Laplacian Pyramid method [37] mentioned above, incompatible for HDR WCG IQA is the focus on the luminance component of images. This focus is reasonable from the perspective of detecting structural distortions or distortion that will be present in the luminance component. However, the colour components of images are completely ignored in such a process. A further challenge faced by structural similarity methods is that is impossible to construct an IQA model for colour distortions that may occur from gamut mapping for example, as a structural distortion problem. For these reasons, the structural similarity philosophy is incompatible with the goals of a IQA model for HDR WCG. Distortions that occur from tone mapping or gamut mapping do not effect any structural distortions, and can only be evaluated in terms of their fidelity to the pristine image that contains the creative intent.

To better understand this constraint of having to rely on the creative intent, consider the following two images given in Figure 9.1. It should be noted here that the colour distorted image in (b) evaluates to a 100% quality score using similarity based methods and existing error visibility methods for IQA since the luminance component of the pristine image is left untouched in the distorted image. However, more problematically, subjects will not only be unlikely to object to the image shown in (b), but they may even prefer the colour grading in it. Thus, the only way in which a meaningful and valuable quality assessment can be performed between these two images is to judge the distorted image with respect

to its fidelity to the reference image i.e., **creative intent**. Therefore, any consideration of colour quality with respect to IQA must necessarily operate on the paradigm that considers creative intent, rather than the currently used notion of what distortions would be considered objectionable by an average human observer. Any deviation from the creative intent in a visible manner would by this paradigm, be objectionable.



(a) (b)

Figure 9.1: Examples of a Pristine Image (a) Distorted Image (b).

9.3 Deficiencies in Present IQA (and VQA) Subjective Testing

In the paper describing the TID 2008 Image database [66], the authors express their concern that subjects have expressed difficulty in providing a score/grade for the quality of an image within the standard subjective testing setting. The authors point out that the subject would rate a particular image with the worst score, and then upon discovering a worse quality image, be faced with the problem of being unable to correct the prior choice. This form of correction is generally never allowed in a subjective experiment. To overcome this issue, the authors performed a prior training session where the subjects would be able to familiarize themselves with possible levels of distortions. However, it should be pointed out that this does not completely mitigate the problem. The training phase assumes that the subject would be able to commit the different levels of expected distortions to memory, and also formulate a consistent rating system within that short span of time that can be accurately applied to a myriad of different images observed over a lengthy period of time (typically ranging from 30 minutes to more than a hour in some cases). In our own experience, this can be quite difficult.

It should also be mentioned that while the TID 2008 database presented the distorted image simultaneously together with the reference image for the subject for rating, subjective experiments can take place with only one image content displayed at a time, especially in the case of video. While the reference image will also be presented to the subject to rate at some point during the testing process, the image order of the test is randomly set. Thus, there is no guarantee that the subject will view the reference image before they view the distorted image. As such, the subject would then need to rate the distorted image from ones experience of prior observed natural images. This method, as may be clear from the description, is unsuitable for determining whether the creative intent is maintained.

Supposing that we choose the option of displaying the reference image together with the distorted image to the subject, there is also a more serious problem, which to the best of our knowledge, has not been raised before. To understand the problem, consider a simple hypothetical subjective test where there are only two distorted images A and B of a single reference image, where A contains a slightly noticeable distortion that is present throughout the entire image, and B contains a severely noticeable distortion that is present in only a small part of the image. Now, the subject viewing these images must first decide whether they would be awarding both A and B with the same rating score or category. Once this is decided, the subject must mentally construct a mapping that quantifies the severity of the distortion, quantifies the percentage of the image that the distortion occupies, and then combine these quantities in some manner to determine an appropriate score. To make things more problematic, the subject must determine this mapping before they may have seen image A or image B (due to the ordering within the test). Thus, the task is an impossible one. Now, in an average subjective test, to accurately determine a single score that remains consistent throughout a typical subjective test, the subject would have to have first observed all the images within the subjective test, and then constructed a consistent mapping. This has never taken place in any subjective test, and is frankly infeasible given the limited amount of time that a subject is willing to invest in the task, and the cost of associated with the participation time. If one ignores the problem, then the mean opinion scores (MOS) computed using such subjective testing are likely not reproducible, and hence of no real value.

We propose that this difficulty can be easily overcome by relaxing the condition that causes the need for the subject to map between competing criterion (extent of the image that is distorted vs severity of the distortion). We shall provide our solution in the next chapter, but for now it must be noted that this problem exists within the current subjective testing methodology, a grave problem that makes it unsuitable for the sound development and testing of IQA methods for HDR WCG content.

9.4 Conclusion

In this chapter, we presented reasons for considering the existing IQA philosophy as well as subjective testing methodology problematic for adaptation to HDR WCG IQA. The existing IQA algorithms developed on top of the structural similarity philosophy are fundamentally incompatible with the notion of preserving creative intent. Both the structural similarity-based methods and error visibility philosophy-based methods are problematic since they consider only the luminance component of images and ignore the colour components. Thus, it is impossible to measure the colour fidelity of an image with respect to its reference. Finally, the subjective quality assessment methodology is problematic since the current scoring system demands a non-trivial level of effort toward awarding a score or category for the images and videos. The difficulty arises from the need to quantify different facets of distortions (area affected by a distortion, and the severity level of the distortion) into a single category or score. Thus, we propose that it will be more beneficial to remove the constraint of awarding a single score, and will expand upon this thinking in the following chapters.

Chapter 10

HDR WCG Image Quality Assessment Algorithm

From our experience in carrying out subjective tests such as the one described in Chapter 3 and taking part in numerous subjective experiments conducted to obtain the mean opinion scores reflecting the quality of images and video, it seems clear that there is a dilemma faced by the subject during the evaluation of the quality of an image (and by extension, the evaluation of video content which consists of a collection of images). We discussed this problem in the preceding chapter, and argued that it can be alleviated by evaluating competing facets of a distortion present in an image separately, rather than combining them into a single numerical score or qualitative label (worst, bad, moderate, good, excellent, etc.). The existence of a reliable and accurate method for a human subject to evaluate the quality of an image is an apriori requirement for creating a quality assessment algorithm. A good algorithm can accurately predict the quality score given by a human subject to an image or video content. Thus, without a reliable method of evaluation of images and video content for human subjects, any quality assessment algorithm cannot be reliably judged as superior or accurate in its operation.

Thus, we begin this chapter by introducing a system of subjective test evaluation that takes into consideration the percentage of the area of the image that is distorted, as well as the extent to which the distortion is perceptually visible. We propose that the subject evaluate these two criterion by selecting one of the following choices that would be presented to the subject.

1. Significant portion of the image is distorted by a significant degree.

2. Significant portion of the image is distorted by a slightly noticeable degree.

3. Significant portion of the image is distorted by a slightly noticeable degree, while a small part is distorted by a significant degree.

4. Small portion of the image is distorted by a significant degree.

5. Small portion of the image is distorted by a slightly noticeable degree.

6. No visible change.

At first glance, the above rating scheme may seem to be reducing subjective scores to a single category (an issue we argued was defective in the previous chapter). However, it should be noted that each category corresponds to a combination of the severity of the distortion and the area of the image that is distorted. We believe that the above scheme offers the advantage in that the subject can evaluate an image independently of the quality of the other images within a subjective test. The subject can gauge the difference as being just noticeable or significantly noticeable, and the level of distortion of other images that are present within the test has no bearing on this matter. Neither does the percentage of the area that is distorted in the other images within a subjective test have a bearing on deciding whether the distortion area is small or widespread throughout a particular image. A training session or instruction dealing with what constitutes a small percentage of the image vs. a significant portion of the image should be sufficient to keep the scores consistent and reproducible. Therefore, the above evaluation criterion can be an accurate and precise.

Still, there is a problem with the above categories in that they are qualitative. This qualitative aspect, while acceptable and even appropriate for certain applications, can be problematic in others that require even the smallest distinctions be made between the quality of an image. To better understand this problem and an application, consider the use of an IQA algorithm for the task of evaluating the performance of two different tone mapping algorithms. With tone mapping algorithms, the better performing tone mapping algorithm would be one that is perceptually close to the original reference image as possible. Now, an IQA algorithm that is only capable of classifying an image into the above categories may classify that the output of both tone mapping algorithms consist of a *'significant portion of the image that is distorted by a significant degree'*. However, it may very well be true that one tone mapping algorithm has actually distorted a lesser amount of the image (quantitatively) than the other, though they both would be qualitatively assessed as affecting a significant portion of the image. Thus, such an IQA algorithm would seem to be limited in its usefulness for such applications.

However, it is important to realize that an IQA algorithm that classifies images according to the above proposed set of categories would arrive at its classification by first performing a quantitative computation. In fact, we propose that any successful quality assessment for HDR IQA must necessarily be able to determine the following quantities.

- Percentage of image with significant change in luminance/colour
- Percentage of image with slightly noticeable change in luminance/colour
- Percentage of image with no noticeable change

It is through the evaluation of the above quantities that an IQA method will be able to classify as to which category the image may be assigned as described by a human subject. Additionally, if required, the percentage of the image undergoing pure colour change or luminance change can also be computed for required applications. Thus, the above quantitatively precise quantities can be utilized for evaluating different compression, tone mapping, or gamut mapping methods and differentiating between the performance of one method over another in a precise manner. In such an algorithm, an additional step will exist to map these multiple quantities representing image quality to one of the categories that may be chosen by a human subject as proposed.

A natural question then would be as to how the accuracy of the IQA algorithm in determining these quantities is evaluated. As mentioned at the start of this chapter, the IQA algorithm performance will always have to be measured by how well it reflects the quality perceived by a human subject. However, if the quality assessment information provided by the subject is limited to the above categories, then there is no possible way to verify the accuracy of the IQA algorithm in its quantitative determinations. At the same time, it is impossible for the human subject to provide the accurate quantitative values that we listed above. We propose to solve this problem by displaying the perceived error signal as seen by the IQA algorithm, to the subjective testing process. Such a visualization of the predicted perceived error signal is referred to in IQA as *quality maps*. The quality map can consist of a grey value coded or colour coded (according to the severity of the error) image that highlights the regions the algorithm detects as changed to the test subject. According to the paradigm we propose, we would require the subject be able to see the regions of the distorted image that the algorithm identifies as having a significant change in luminance/colour and the regions the algorithm identifies as slightly noticeable change in colour/luminance through its quality map. Thus, in addition to classifying an image into a category as we suggested above during the subjective testing process, the subject must also evaluate the quality map correctness. A successful IQA method is one that

not only predicts the subject choice reliably, but one that also captures all the errors and is able to display them accurately on its quality map. This methodology eradicates any imprecision or ambiguity that would be present in comparing performance against subject evaluated categories. It should be noted here that to the best of our knowledge, no published subjective test has been carried out that has displayed the quality maps for the subject to evaluate.

In summary, as per the above discussion, we then require the following elements from a subjective test for HDR WCG content under this new paradigm.

1. The subject be able to compare the reference image and distorted image before awarding a score.
2. The subject choose a category from those listed that best represents by comparing the two images.
3. The subject evaluate the quality map accuracy.

In light of the above, the IQA algorithm should also satisfy the following the requirements.

1. Given the reference and distorted image, evaluate the percentages of the distorted image that contains a significant/slightly noticeable change in colour/luminance.
2. Output a quality map displaying the detected error according to severity, that captures the actual changed in the image.
3. Using the percentages of the image that is distorted and the severity, predict the category that will be chosen by the subject.

It should be emphasized that the above requirements are novel recommendations in the field of IQA and VQA based on the paradigm we introduced in the previous chapter. Due to its novelty, these above listed requirements may certainly be developed further with ongoing research work to improve. But, we propose that the above requirements will sufficiently meet the needs of HDR WCG quality assessment by overcoming the shortcomings that we discussed in the previous chapter. Thus, we will now turn our attention to proposing a HDR IQA algorithm, and subsequently the subjective testing procedure, that satisfies the above requirements.

10.1 Obtaining the Perceivable Colour Difference

Given a reference image *img1* and distorted image *img2*, the proposed IQA algorithm operates as follows. We first begin by converting the image data into the ICtCp colour space, and evaluate the ΔIC_tC_p for each pixel as [69],

$$\Delta ICtCp = \sqrt{(I_1 - I_2)^2 + 0.25(Ct_1 - Ct_2)^2 + (Cp_1 - Cp_2)^2}.$$

Then, to determine the JND for each colour pixel, we utilize the *ICtCp V-JND-L* proposed in Chapter 8 to estimate the JND given the luminance of each colour pixel. Recall that the luminance-JND relationship was approximated using a piece-wise linear relationship as shown in Figure 10.1 below. This allows us to account for the luminance of the colour pixel, and thus obtain a more accurate estimate of the JND as demonstrated in Chapter 8. However, as discussed in Chapter 7, the JND of colours are also affected by the background luminance.

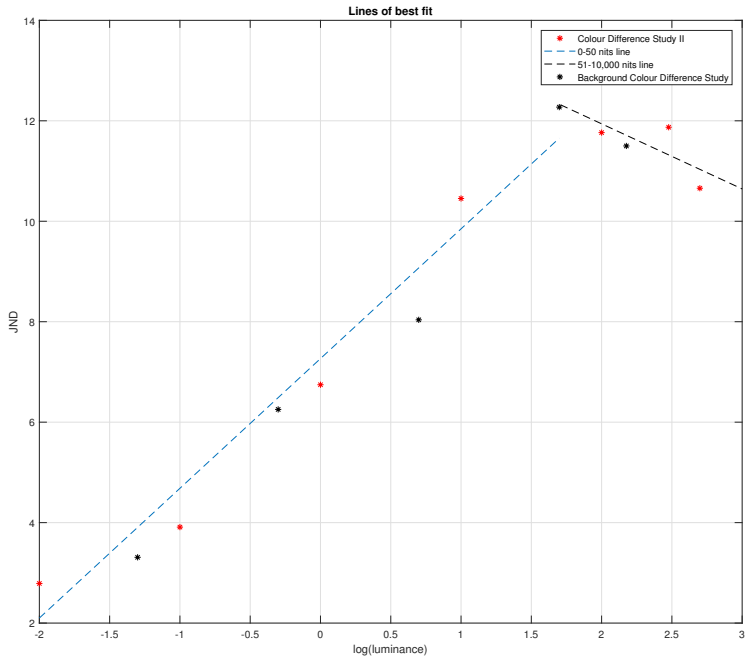


Figure 10.1: JND vs Log(Luminance) for all tests

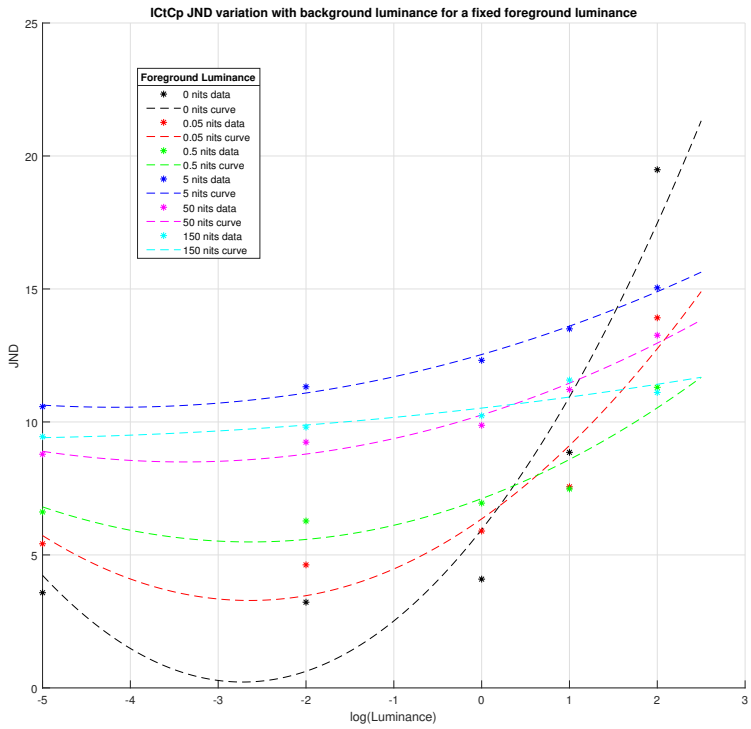


Figure 10.2: JND vs Log(Luminance) for all tests

Figure 10.2 shows the distribution of the JND data (in the IC_tC_p colour space) gathered from the colour difference test studying the effect of background luminance described in Chapter 7. The JND data points are colour coded according to each foreground luminance that was tested. All the points of a given colour represent the average JND for the same set of colours under each different background luminance for which the test was carried out (the x-axis of the plot corresponds to the $\log_{10}$ of the background luminance). Note, the lowest background luminance value that was used in the colour difference test was 0 nits. However, this background luminance was approximated as 0.00001 nits to allow it to be considered in approximating a relationship between the background luminance, foreground luminance, and JND. The value of 0.00001 nits was chosen since this is representative of the black level of the reference display on which the colour difference test was carried out. The best fitting 2nd degree polynomial curves for each foreground luminance are also shown in the same figure.

The best fitting curves for each foreground that is shown in the above figure can be used to adjust the JND to account for the effect of background luminance. However, the first problem with utilizing these curves is that given a natural image, it is highly unlikely to contain a single patch of colour with a uniform background luminance as was present during the colour difference test. We propose using the average luminance of a given image/scene as the background luminance of the particular image as a comparable substitute. Given a complex natural image, as the HVS focuses on a particular colour within the image, the HVS will be affected by the overall luminance of the surrounding image rather than the luminance of the immediately surrounding pixels. Thus, the average luminance of an image would appear to be a reasonable estimate of the *background luminance* in a complex image.

A second problem arises from the fact that the data we possess does not capture every possible foreground luminance, but only a few foreground luminances (0.5 nits, 0.05 nits, 5 nits, 50 nits, 150 nits, 300 nits). Ideally, one would like to sample the colour space densely under a large range of foreground luminances to obtain JND data as we have done through our colour difference experiments, and then produce a set of curves (as we have done in Figure 10.2) over which one could find a best fitting surface that could then be used to estimate the JND adjustment. Since we do not have a sufficiently large number of curves to construct an entire surface with good degree of accuracy, we opt for computing the JND adjustment based on the curves corresponding to the nearest foreground luminances (nearest luminance greater than the foreground luminance of interest, and nearest foreground luminance lesser than the foreground luminance of interest) which we have tested. We then linearly interpolate the JND adjustment for the particular foreground luminance of interest.

A third problem occurs with regard to how one should compute the JND adjustment itself. Each curve presented in Figure 10.2 represents the JND for a fixed foreground luminance across varying background luminance. We estimate the JND adjustment from the curves as given below:

$$\textit{JND adjustment} = \frac{\textit{JND for a foreground luminance } y \textit{ \& background luminance } x}{\textit{JND for a foreground luminance } y \textit{ \& background luminance of } \boldsymbol{0 \textit{ nits}}}.$$

When the foreground luminance x is one that directly corresponds to the curves, the JND under a background luminance of 0 nits is already known (the leftmost point of each of the curves given in Figure 10.2), and the curve provides the numerator of the formula above (JND for a particular background). When the foreground luminance x is one that does not correspond directly to the curves, the JND adjustment can be computed individually

for each of the nearest curves using the above formula, and then interpolated as explained before. Once the adjusted JND's are computed according to the manner described, we compute the perceived colour difference as a ratio of the JND for each pixel as follows.

$$\textit{Colour difference as a ratio of JND} = \frac{\Delta IC_tC_p}{\textit{Adjusted JND}}$$

If the perceived colour difference of a pixel as computed above is greater than 1, but less than 2, we classify the pixel as having a slightly noticeable colour change. If the perceived colour difference of a pixel is above 2, we classify it as having a significantly noticeable degree of colour change. These thresholds, which we will designate as *lower threshold* and *upper threshold*, respectively, can be adjusted to match the creator preference and sensitivity to colour difference. For example, if the creator is willing to tolerate a 2 JND colour difference, the upper threshold can be adjusted to 3. Similarly, if the creator is slightly more sensitive to colour difference, and wishes to preserve ones creative intent to meet that degree of sensitivity, the lower threshold can be set to 0.9, for example.

We have summarized the computation of the perceived colour difference as discussed above in the following algorithm.

Algorithm: IQA: Computing the perceived colour difference

img1 := Reference image;
img2 := Distorted image;
Step 1: Convert *img1* and *img2* to the IC_tC_p colour space;
Step 2: Compute the ΔIC_tC_p between *img1* and *img2*;
Step 3: Compute the corresponding JND for each pixel of *img1* using *ICtCp V-JND-L* proposed in Chapter 8;
Step 4: Compute JND adjustment for each pixel to account for the background luminance using the average luminance of *img1*;
Step 5: Perform JND adjustment for each of the JND's computed in *Step 3* ;
Step 6: Obtain the pixel by pixel perceived colour difference as a ratio of JND using $\frac{\Delta IC_tC_p}{\textit{Adjusted JND}}$;
Step 7: For each pixel:
if *lower threshold* $\leq$ *pixel by pixel difference* $\leq$ *upper threshold* **then**
  Changed by a slightly noticeable degree;
else if *pixelwise difference* $\geq$ *upper threshold* **then**
  Changed by a significant degree;
else
  No visible change;

10.2 Obtaining the Perceivable Difference in Luminance and Structure

We assess the luminance component separately since the structural elements are present in the luminance component. It is possible that changes in the luminance levels of large parts of an image (or the entire scene, which would be perceived as change in brightness) may still be captured from colour differencing. However, it would be difficult to quantify this change in the same manner in which colour difference is quantified. The threshold beyond which the HVS perceives the difference as a very noticeable difference, may not apply to a similar difference in luminance. More importantly, recall that the colour difference methods proposed in Chapter 8 (including *ICtCp V-JND-L* used in the above proposed algorithm) were devised using colour difference data gathered from the experiments where colour difference was measured when the luminance of the compared colours was held constant. Thus, there is no information about how well the luminance differences in ΔIC_tC_p would reflect perceived luminance differences. In contrast, the PQ function was developed as a non-linear mapping that quantizes the linear light luminance component Y to Y' such that a unit difference in coded values is just below the just noticeable difference of the HVS (to avoid banding artifacts) [53]. So we compute the structural/luminance differences, and determine the perceptual difference as follows.

Algorithm: IQA: Structural/Luminance differences

img1 := Reference image;
img2 := Distorted image;
Step 1: Pixelwise difference is computed using luminance components Y' by performing $Y'_{img1} - Y'_{img2}$;
Step 2: For each pixel:
if *lower threshold* $\leq$ *pixelwise difference* $\leq$ *upper threshold* **then**
  Changed by a slightly noticeable degree;
else if *pixelwise difference* $\geq$ *upper threshold* **then**
  Changed by a significant degree;
else
  No visible change;

For the lower threshold of determining whether a pixel luminance difference is noticeable to the human visual system, we check if the coded value difference is greater than or equal to 2. The reason for choosing a value of 2 as opposed to a value of 1 is that, as mentioned above, the PQ function is constructed in such a way that the quantization

of the luminance component produced using PQ will have coded values separated by a difference that is less than a JND (approximately 0.9 JND, according to [53]). Thus, a value of approximately 2 would occur when the HVS determines that the perceived luminance difference is slightly noticeable. The upper threshold was set to 5, and this value was determined by experimenting with various images to determine an acceptable threshold at which the luminance changes are classified in an acceptable manner as slightly noticeable vs. significantly noticeable according to our preference. However, the upper-threshold could always be adjusted depending on the preferences of the creator to how sensitive one wishes the algorithm to be toward the degree of change in an image that one would classify as a significant degree of change. In fact, in the event that the creator is more sensitive to the change in luminance, the lower threshold could be adjusted to cater to the creator's preference as well. In this manner, our proposed algorithm can be intuitively fine tuned using the just noticeable differences to match the 'sensitivity to change' of the creator. This capability is a vital feature that is required for checking the preservation of creative intent. It should be mentioned here that present IQA algorithms do not lend themselves to such straightforward adjustment.

10.3 Combining Perceived Error in Structure, Colour & Generation of Quality Maps

In the previous sections, we describe the proposed methods for computation of the perceived error, and the classification of the error as significant, slightly noticeable, or no visible change. Next remains the matter of determining the extent of the image that would be considered as being contaminated by the error. As explained in Chapter 9 and the discussion at the start of this chapter, we hold that combining the perceived error to a single score or a single qualitative judgment (excellent, good, moderate, bad, worse) is an unsuitable approach for IQA and VQA. Thus, we proposed the computation of multiple "scores" that capture the image quality as a percentage of the area of the image where the colour/luminance is distorted/changed. Given the computation of the perceived error using the algorithm steps proposed above, these percentages can be evaluated simply as

$$\textit{Percentage of particular change} = \frac{\textit{Total number of pixels with the particular change}}{\textit{Total number of pixels in the image}}.$$

Once the above quantities are computed, the decision making to map the above percentages to the choice made by the subject/creator can be performed simply by checking

whether the percentage of a certain type of error falls within thresholds for a small or significant portion of the image. Two thresholds in terms of area must be determined to perform this mapping. One needs to determine the area threshold above which a person would consider the distortion as affecting a significant portion of the image, and the area threshold below which one may consider the image as not distorted. This lower threshold is required since one may reasonably choose to ignore a single pixel change, for example. Furthermore, these thresholds can be determined for persons in general, and also be set according to the preference of a single subject (like in the case of a creator who judges the preservation of creative intent). For the subsequent testing of this method, we fixed thresholds according to the preferences of the author. Carrying out tests using a subset of images that were reserved for this process (apart from the image database that will be introduced shortly that will be used for testing the performance of the proposed method), we set the thresholds value at an acceptable level for the authors.

Now, evaluating the above percentages and subsequent decision making for the entire image based on these pixel counts and above thresholds can be misleading. It is possible, for example, that a 25% of the pixels of an image that are distorted are spread throughout the image. This can indicate to the subject that significant portion of the image is distorted, while the percentage suggest that only a small portion of the image is distorted. To overcome this problem, we segment the image into 3×3 smaller images of equal area. An example of this 3×3 division is shown in Figure 10.3. The motivation for using a 3×3 subdivision comes from the "rule of thirds" used for scene composition in photography [48]. It is quite possible that a more optimal segmentation exists that also considers the display size and subject viewing distance, but we use this simple segmentation as a first iteration of the proposed algorithm. These segments enable one to localize the spread of the error pixels within the image. By computing the percentages for each of the segments, one can then determine whether each segment can be classified as containing a significant degree or slightly noticeable degree of change. Then the decision on whether a small or large portion of the image is distorted can be made by looking at the level of distortion for the 9 segments. In our algorithm we consider more than four segments being distorted as comprising a large portion of the image, while four or less is considered to comprise a small portion of the image.

For the generation of the quality maps, the maps were output as 8-bit greyscale images. The regions that have slightly noticeable changes in the distorted image were indicated in the quality map using the colour grey (127, 127, 127). The regions with significant degree of change were indicated by black (0, 0, 0), while the unaffected regions were shown as white (255, 255, 255). An example of a slightly distorted image is given in Figure 10.4 and an example of a significantly noticeable degree of distortion is given in Figure 10.5. As can

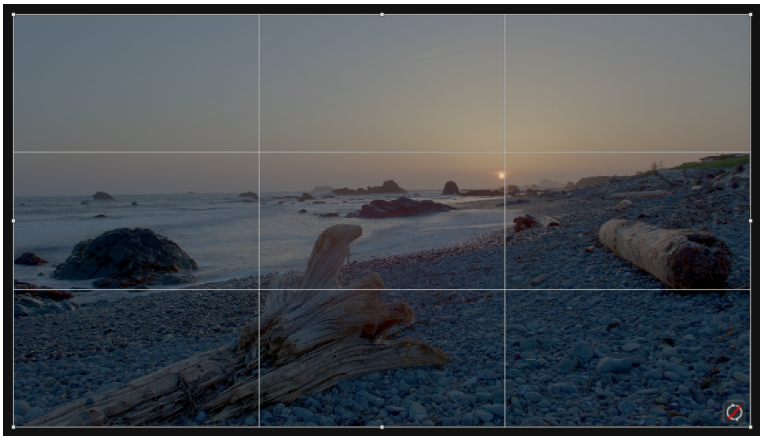
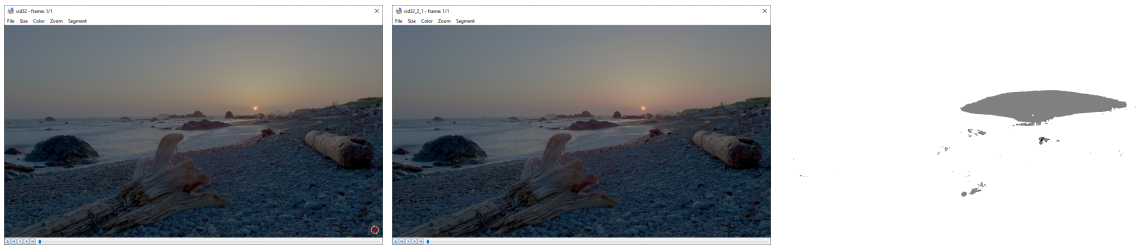


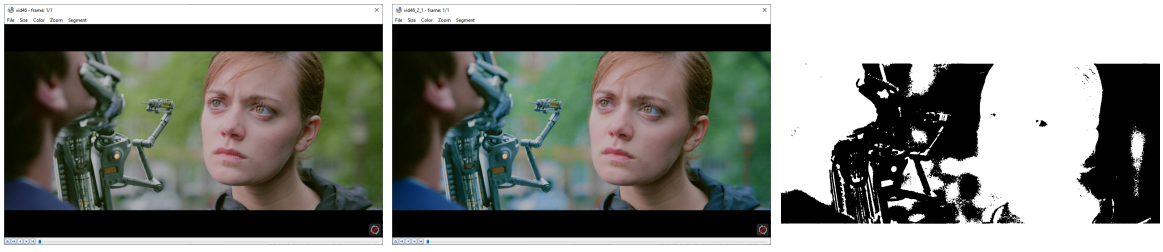
Figure 10.3: An example of splitting of the image to 3×3 parts.

be seen from the two figures, the slight distortion is highlighted as grey in the quality map. The larger distortion is highlighted in black. The unaffected regions are shown in white.



(a) (b) (c)

Figure 10.4: Example of a slightly distorted image and its Quality Map. Pristine Image (a) Distorted Image (b) and (c) Quality Map.



(a) (b) (c)

Figure 10.5: Example of a significantly distorted image and its Quality Map. Pristine Image (a) Distorted Image (b) and (c) Quality Map.

10.4 Testing the Performance of the Proposed HDR WCG IQA Algorithms

As mentioned in the beginning of this chapter, the performance of an IQA or VQA algorithm is measured by its ability to closely align with the quality as perceived by the HVS. Therefore, the traditional approach in IQA/VQA algorithm development has been to use an existing database or create a new database, with content that are affected by the interested types of distortions, and then compare the human evaluation of the content to that of the algorithm prediction. This is the process we followed in Chapter 3 where we evaluated the performance of existing quality assessment algorithms in predicting the quality of HDR WCG content suffering from compression based distortions. However, as we argued in this and the previous chapter, the philosophy required to satisfy the requirements of HDR WCG IQA is different from that of the existing methods. The fidelity to creative intent is the foundation upon which our new quality assessment algorithm has been devised. While the construction of a database of content affected by the interested distortions is still essential, there are additional requirements for the subjective testing process. We listed the following in the introduction of the chapter, but present them here again below.

1. The subject be able to compare the reference image and distorted image before awarding a score.
2. The subject choose a category from the above listed that best represents by comparing the two images.
3. The subject evaluate the quality map accuracy.

In addition to the above requirements, there can be said to be one additional preferential requirement. It is preferred under the proposed paradigm of evaluating fidelity to creative intent, that the subject who evaluates the quality of the distorted content be the same person who was behind the creative intent. In the traditional subjective testing methodology, one is focused on computing a mean opinion score by collecting the opinion of many subjects (as seen in Chapter 3). This interest is reasonable according to the traditional paradigm where the subject is expected to evaluate the quality of content by drawing upon their prior experience of natural images. However, for evaluating the fidelity to creative intent, the creator behind the intent would be best positioned for evaluating the quality. This is especially true when one considers the average time within which a subject must evaluate the quality of content within a traditional subjective test. For timely completion

of subjective tests, they are designed with less than one minute of time allocated for evaluation of a single distorted piece of content. Even with a subjective test that is designed to satisfy the main three requirements listed above, it can be difficult for a subject to accurately evaluate the preservation of creative intent in such a short period of time. The person behind the creative intent can likely spot the shortcomings of distorted content more readily since they were behind the engineering of the intent of the scene in the first place. It is also possible to allow for more time per evaluation of a single piece of content since there is only one subject that takes part in the study (in comparison to 51 subjects in the test described in Chapter 3, for example). For these reasons, the subjective scoring performed by a single subject who is the creator behind the creative intent would be a more accurate representation of the quality than the traditional mean opinion score.

The subjective study we propose and carry out in the following sections was designed to satisfy all of the four requirements laid out above.

10.4.1 Image Selection and Testing Methodology

We obtained the reference images for the database from two sources, the *Mark Fairchild's HDR Photographic Survey* [4] made available in raw *.exr*, and the frames of the open source film *Tears of Steel* [5] made available in *.tiff* files. As discussed above, we require the test to be evaluated by the person behind the creative intent. Thus, we began by color grading the selected raw images from these two sources using the display that will be used for testing. This process guarantees that the images will be displayed as intended during the subjective test (since the color grading has been performed on the exact display in which it will be viewed), and also makes it possible for the author to take part in the subjective test as the person behind the creative intent.

Our testing and grading setup consisted of a *BenQ PD 2700U* HDR display, calibrated for 100% Rec. 709 coverage, with support for PQ encoding and a peak luminance of 350 nits. This monitor was driven using a PC with *Blackmagic Decklink Mini Monitor 4K* device connecting to the monitor via a HDMI interface. The *Mini Monitor 4k* allows for fine control of the output video signal to the monitor when colour grading, as well as during the testing process when the image is displayed on the monitor. A secondary display was used for running the grading application and displaying the subjective testing user interface during the subjective testing phase. This secondary display was a regular display connected to the PC via a standard video output. This setup is illustrated in Figure 10.6.

The test consisted of 46 reference images chosen from the two sources mentioned. First,

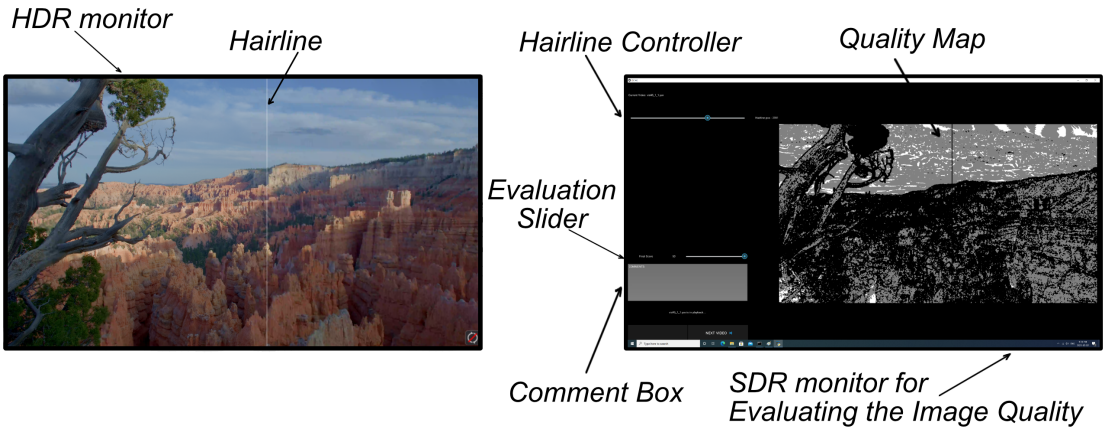


Figure 10.6: Subjective Testing Setup

we performed colour grading to match the capabilities of the HDR display used for the test (PQ, 10-bit, Rec 709). Then, 230 distorted images were produced using the graded reference images via *x264* compression that consisted of a mixture of structural and colour distortions. Since compression does not cause pure colour distortions on its own, we also created 20 more distorted images consisting of pure colour distortions using techniques such as mapping to a smaller gamut, selective skewing of colours in skin tones and other areas of interest in the images. The final database consists of 250 images in total.

10.4.2 Subjective Testing Procedure

Each distorted image was displayed on the *BenQ PD 2700U* HDR display, while a user interface on the adjacent standard display allowed the subjects to control a hairline on the HDR display such that to the right of the hairline would be displayed the corresponding reference image, while the left contained the distorted image. This allowed the subject to compare the distorted image to the reference image without having to rely on prior memory of the reference image. The user interface on the standard display also allowed the subject to choose the rating for the image from the categories discussed at the start of this chapter, and navigate to the next test once the scoring is complete.

The test was carried out in a room with very low ambient lighting (approximately 12 nits). The ambient lighting was produced by a backlight placed behind the displays to reduce eye fatigue of the subject. As indicated in the previous subsection, the subjective

test was administered to the author who performed the colour grading of the reference images. To reduce human error, the subjective test was administered three times, and the majority decision from all three tests was taken as the final "score" for each distorted image. The image sequence within the test was also randomized for each attempt. Sufficient breaks were taken to reduce any fatigue during testing.

The subject evaluated the quality of content and classifies each image into one of the following categories

1. Significant portion of the image is distorted by a significant degree.
2. Significant portion of the image is distorted by a slightly noticeable degree.
3. Significant portion of the image is distorted by a slightly noticeable degree, while a small part is distorted by a significant degree.
4. Small portion of the image is distorted by a significant degree.
5. Small portion of the image is distorted by a slightly noticeable degree.
6. No visible change.

and also evaluated the accuracy of the quality map as displayed on the user-interface for the subjective test. For the assessment of the quality maps, we utilized a binary scheme of 'correct' or 'incorrect'. If the quality map highlighted any region as incorrectly being distorted, it was considered an 'incorrect' quality map. This process of evaluating quality maps could likely be refined further, but we utilized this scheme for this particular subjective test.

10.4.3 Performance of the Proposed HDR WCG IQA Algorithm

As mentioned in the previous section on image selection, the database of images used for testing consisted of 250 images in total. Of these, 230 were generated from compression, and 20 were generated from pure colour distortions. Of the 230 images, 202 images consisted of significantly noticeable degree of distortions, while 28 images consisted of slightly noticeable degree of distortions. Within the 20 pure colour distortions, there were 12 images of significantly noticeable colour change, 8 images of slightly noticeable colour change. In both the colour and compression based distortions, a larger number of images were allocated for the category of significantly noticeable degree of distortion. The reason for the larger

number of images under this distortion category is that it is more vital that an IQA algorithm be able to detect such distortions. For example, a creator may choose to tolerate some degree of slightly noticeable changes within the distribution process of content while significantly noticeable changes may always be considered as unacceptable. Thus, most of the images were crafted to test this aspect of the IQA algorithm more heavily.

In comparing the classification of the images, the proposed IQA algorithm correctly classified 96.43% (183 of 202 images) of the images with compression distortions as having significant degree of distortion, and 90.59% (27 of 28 images) as having a slightly noticeable degree of distortion. The performance does drop to 61.54% (8 of 13) in correctly classifying images that contain a significant portion of the image with slightly noticeable distortion together with a small part that is distorted significantly noticeable degree. However, it should be noted that 10 of the 13 images (77 %) in this category were correctly identified as having a large distortion. For the images with pure colour distortions, 95% of the images (19 of the 20 images) were classified correctly by the proposed IQA algorithm. One image was incorrectly classified as having no visible change. For the total set of images, both the pure colour and compression based distorted images, the proposed algorithm correctly predicted the portion of the image that was distorted for in all the cases.

With regard to the case of classifying images with a significant degree of distortion in small parts together with a slightly noticeable degree of distortion in a large portion of the image, some more investigation was necessary to isolate the possible source of the error. The question in this case is whether the classification portion of the algorithm was to blame, or the part that computed the visible error. Thus, we turn to the accuracy of the quality maps. For the entire subjective test consisting of 250 distorted images, the algorithm generated quality map was classified as incorrect for only 5 of the images. This indicates that the algorithm is indeed very effective at correctly computing whether a pixel is distorted in terms of luminance or colour. But, the results do indicate that the classification portion of the algorithm that operates by computing the percentage of pixels distorted and the segmentation of the image does require further work. As suggested in the discussion of the algorithm, it is possible that a more optimal segmentation exists for an image based on viewing distance and display size. Our current proposed approach checks whether at least one segment is distorted by a significant degree when an image contains only a large portion of the image that is slightly distorted. This check determines whether to classify the image as having a small portion that is also largely distorted. However, the effectiveness of the threshold used for the pixel percentage above which a segment is classified as being significantly distorted can depend on the segment size. Thus, the classification aspect is an area in which future work would be necessary for improving the performance of the algorithm.

On the matter of incorrect quality maps, the low number of incorrect maps does indicate that the algorithm is certainly good in its performance. The incorrect maps are from images with very complex texture detail that seem to mask the visibility of the errors. Some improvement of this proposed algorithm is possible in future works by detecting these complex texture regions of an image and evaluating the colour and luminance difference accordingly.

Furthermore, we also re-ran the proposed algorithm with the JND based on background luminance disabled. The results of this re-run showed that without the background luminance adjustment, the IQA algorithm incorrectly classified close to 90% of all the images with slightly noticeable differences as having a significant degree of distortion. This result is consistent with what one would expect if background luminance is not considered. As seen in Figure 10.2, the JND does increase as the background luminance increases. Therefore, using the original JND would overestimate the level of distortion as observed during the re-run of the IQA algorithm. Conversely, this also shows that the JND adjustment we introduced in this chapter to the IQA algorithm is effective and necessary for correct prediction of the degree of distortion.

10.4.4 Conclusions

In this chapter, we introduced three necessary features for an effective HDR WCG IQA algorithm according to our proposed IQA paradigm (in Chapter 9) of preserving creative intent. We also introduced three necessary features for subjective testing according to this paradigm, with an added preferential requirement stressing the importance of having the creator as the subject who evaluates the quality. We then introduced our HDR WCG IQA algorithm that separately evaluates the colour difference, luminance difference, and then combines the result to predict the quality according to the proposed paradigm. Under this proposed paradigm, the subject evaluates the degree of distortion present in an image, as well as the extent (the area of the image) of the image that is distorted. Furthermore, the importance of evaluating the accuracy of the Quality Map generated by the IQA algorithm was also explained. We then detailed the construction of a HDR image database and the subsequent testing procedure. The images for the database were colour graded by the author, thus making the author the creator behind the creative intent who would be the preferred subject for the study. We then used the results of this evaluation of the content by the author to compare the IQA algorithm performance. The results indicated that the IQA algorithm was extremely accurate in predicting the quality degradation present in the distorted images as seen by its generation of accurate Quality Maps for 245 images out of 250. In terms predicting the categories selected by the subject, the IQA algorithm was

97% effective in correctly identifying images with significant degree of distortions as well as 91% effective in identifying images with slightly noticeable distortions.

The proposed HDR WCG IQA algorithm did show weakness in correctly classifying the images according to the extent of the distortion (area of the affected image). We proposed that this deficiency was likely due to the nature of the segmentation of the image that was used to aggregate the detected errors into a category. It is quite possible that a more optimal segmentation exists, and future research can be performed to better understand how best to segment the image to match the result of the algorithm with that of the human perception.

Chapter 11

Conclusions & Further Research

In this thesis, we studied the performance of existing Image Quality Assessment (IQA) and Video Quality Assessment (VQA) algorithms in evaluating the quality of High Dynamic Range (HDR) Wide Colour Gamut (WCG) content in accordance with the existing quality assessment paradigm. We devised a subjective experiment consisting of a PQ encoded 10-bit HDR WCG content database with a focus on image compression based distortions to determine the mean opinion of score (MOS) as perceived by human subjects. Then, the MOS was compared against the predicted quality by existing quality assessment algorithms, and these algorithms were seen to have acceptable levels of performance. We then studied the effect of using constant luminance vs. non-constant luminance data for evaluating the quality using the quality assessment algorithms. Since the constant luminance is the more correct representation of the luminance component of the image, it was hypothesized that use of the constant luminance component may achieve an improvement of the performance of existing IQA/VQA methods. After all, the existing methods were built on the principle that they operate on the luminance information of the image. Our analysis showed that there was no improvement in using constant luminance and that the performance remained almost identical to the case of non-constant luminance.

Next, we presented a testing framework for studying the performance of existing colour difference methods for HDR WCG applications. We then evaluated the performance of several colour difference methods and found them to be severely lacking. However the results did confirm that the claims by Dolby Laboratories regarding ICtCp being more perceptually uniform than existing colour spaces was accurate. We also studied the effect of background luminance on colour difference perception in a separate study as well. The data gathered from these colour difference experiments were used to construct new colour

difference methods. Some of these methods performed significantly better than existing methods.

We turned our attention to the problem of constructing an IQA method that met the requirements for quality assessment of HDR WCG content. We demonstrated that the existing paradigm of quality assessment was incompatible with the requirements for HDR WCG IQA. This led us to propose a more suitable paradigm and a quality evaluation scheme for human subjects that is highly reproducible between subjects. We also proposed requirements that must be satisfied by an IQA algorithm and the subjective testing process to evaluate the performance of an algorithm according to the new paradigm. Finally, we proposed a new HDR WCG IQA algorithm, performed a subjective study, and evaluated the performance of this new algorithm. The proposed HDR WCG IQA algorithm was seen to be highly effective in detecting distortions within an image. The section of algorithm that classified images according to the proposed evaluation scheme was also effective, but did show some weakness in the case of images where a slightly noticeable difference affects a large portion of the image with a significantly noticeable degree of distortion present in small parts of the image. This suggests that the classification part of the algorithm can be further improved, and we proposed a possible method of improving the performance as well. It was also seen that the adjustments introduced in the algorithm to take into account the effect of background luminance were vital for correct prediction of the severity of distortion. Removal of this adjustment process resulted in significantly worse performance at delineating slightly noticeable distortions from significantly noticeable distortions.

The work in our thesis has provided a foundation based on a novel paradigm for HDR WCG IQA. Our work has also integrated colour quality into the quality assessment process under this new paradigm. An extensive part of this thesis was also dedicated to the study of the performance of existing colour difference methodologies and colour spaces to better understand how they could be integrated to the IQA/VQA setting. Improvements built upon these existing methods were proposed in this thesis that are far more suitable for HDR WCG quality assessment applications.

11.1 Further Research

We list the following natural courses in which the research work in this thesis could be extended.

1. **Extend the proposed paradigm and HDR WCG IQA algorithm to Video content**: While the proposed IQA algorithm was extremely effective for IQA, ex-

tending it to Video Qualtiy Assessment (VQA) does pose its own challenges. If one defines VQA as every frame of the video content being faithful to the creative intent, then one can directly extend the proposed IQA algorithm to VQA by treating each frame as an image. Then the question becomes one of formulating a method through which one aggregates the evaluation for each frame to then describe the quality of the entire video. A more interesting challenge may also be found if one relaxes the condition of requiring every frame having to preserve creative intent. In a video, not every frame is observed by the subject. In a 24 frames per second cinematic production, the subject only perceives the frame for 1/24th of a second. Therefore, not all the detail within the frame is visible to the subject. At the same time, if the particular scene is static, then the subject is more likely to perceive lack of details within the scene. Therefore, to construct a quality assessment algorithm that satisfies the relaxed condition can be quite challenging. There is potential for research in carefully understanding the subject perception vs. time required for a particular level of detail within an image to be visible.

2. **Performance testing of the proposed IQA algorithm using WCG content**: Due to the lack of access to a display with significant coverage of the Rec 2020 primaries (WCG), we tested the performance of the proposed IQA algorithm using a display with 100% Rec. 709 coverage. While the algorithm operated by preprocessing the Rec. 709 content and converting it to the Rec. 2020 primaries, the image database has not, strictly speaking, included a coverage of the WCG. So it would be of interest to test the performance using a image database constructed using the Rec. 2020 primaries.

3. **Improving the classification part of the proposed IQA algorithm**: The performance of the proposed algorithm was excellent in distinguishing between images with only slightly noticeable differences or significantly noticeable differences. However, the performance was weaker for classifying images where the significant distortion was smaller and a large portion of the image contained slightly noticeable differences. The results from the Quality Maps proved that this was not the result of an inability of the IQA algorithm to detect the significant differences, but one that was the result of its classification. Therefore, research work should be done to improve the classification part of the algorithm. Apart from the methods we proposed in the thesis, perhaps a fuzzy-logic based approach may also be a potentially promising direction.

4. **Denser sampling of the colour space for construction of more effective JND adjustments for colour**: As seen in this thesis, the performance of exist-

ing colour difference methods was successfully improved by building upon the data gathered through colour difference experiments. The colour difference experiments we performed were limited to colours with a small subset of luminance values. Thus, having JND data for more densely sampled set of luminances can help make more accurate adjustments to the JND as proposed by existing colour difference methods and colour spaces. Furthermore, given the construction of potential displays supporting up to 10,000 nits peak luminances, there is a need for more colours with these higher luminance ranges to be sampled for JND data.

References

[1] ACES Central. https://acescentral.com. Accessed on 19.05.2021.

[2] Alexa Exposure Latitude Illustration. https://www.arri.com/en/2014-11-11-alexa-exposure-latitude-01-83718. Accessed on 19.05.2021.

[3] Canon DP-V2420 Reference Display product page. https://www.usa.canon.com/internet/portal/us/home/products/details/reference-displays/4k-uhd-reference-displays/dp-v2420. Accessed on 19.05.2021.

[4] Mark Fairchild's HDR Photographic Survey. http://markfairchild.org/HDR.html. Accessed on 19.05.2021.

[5] Tears of Steel: ALL 4K FRAMES NOW AVAILABLE ON XIPH.ORG. https://mango.blender.org/production/all-4k-frames-now-available-on-xiph-org/. Accessed on 19.05.2021.

[6] BT.2100: Image parameter values for high dynamic range television for use in production and international programme exchange. Standard, Radiocommunication Sector of International Telecommunication Union, 2017.

[7] SMPTE ST 2094-1. Dynamic Metadata for Color Volume Transform–Core Components. 2016.

[8] SMPTE ST 2094-10. Dynamic Metadata for Color Volume Transform–Application #1. 2016.

[9] SMPTE ST 2094-40. Dynamic Metadata for Color Volume Transform–Application #4. 2016.

[10] H. Abdi. Coefficient of variation. *Encyclopedia of research design*, 1:169–171, 2010.

[11] S. Athar, T. Costa, K. Zeng, and Z. Wang. Perceptual quality assessment of uhd-hdr-wcg videos. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 1740–1744. IEEE, 2019.

[12] M. Azimi, A. Banitalebi-Dehkordi, Y. Dong, M. T. Pourazad, and P. Nasiopoulos. Evaluating the Performance of Existing Full-Reference Quality Metrics on High Dynamic Range (HDR) Video Content. *arXiv preprint arXiv:1803.04815*, 2018.

[13] A. Balanov, A. Schwartz, Y. Moshe, and N. Peleg. Image quality assessment based on DCT subband similarity. In *Proc. IEEE Int. Conf. Image Process. (ICIP)*, pages 2105–2109, Sept. 2015.

[14] A. Banitalebi-Dehkordi, M. Azimi, M. T. Pourazad, and P. Nasiopoulos. Compression of High Dynamic Range Video using the HEVC and H.264/AVC Standards. In *Int. Conf. Het. Netw. Quality, Rel., Security, Robustness*, pages 8–12, Aug. 2014.

[15] F. Banterle, A. Artusi, K. Debattista, et al. Advanced high dynamic range imaging: theory and practice. *CRC Press*, 2011:33–113, 2015.

[16] Blackmagicdesign. Product Technical Specifications: UltraStudio 4K Extreme 3. Accessed on 19.05.2021.

[17] ITU-R Recommendation BT.2020-2. Parameter values for ultra-high definition television systems for production and international programme exchange. October 2015.

[18] ITU-R Recommendation BT.2087-0. Colour conversion from Recommendation ITU-R BT.709 to Recommendation ITU-R BT.2020. October 2015.

[19] ITU-R Recommendation BT.500-13. Methodology for the subjective assessment of the quality of television pictures. January 2012.

[20] Rec. ITU-R BT.709-5. Parameter values for the hdtv standards for production and international programme exchange. June 2015.

[21] ITU-R Recommendation BT.709-6. Parameter values for the HDTV standards for production and international programme exchange. June 2015.

[22] C. M. Jarque and A. K. Bera. A Test for Normality of Observations and Regression Residuals. *Int. Statist. Rev.*, 55(2):163–172, Aug. 1987.

[23] C. Chinnock. Dolby Vision and HDR10. *White Paper of Insight Media*, 2016.

[24] CARLSON C.R. and Cohen R.W. A simple psychophysical model for predicting the visibility of displayed information. 1980.

[25] S.J. Daly. Visible differences predictor: an algorithm for the assessment of image fidelity. In *Human Vision, Visual Processing, and Digital Display III*, volume 1666, pages 2–15. International Society for Optics and Photonics, 1992.

[26] R. P. Dooley and M. I. Greenfield. Measurements of edge-induced visual contrast and a spatial-frequency interaction of the cornsweet illusion. *JOSA*, 67(6):761–765, 1977.

[27] Z. Duanmu, W. Liu, Z. Wang, and Z. Wang. Quantifying visual image quality: A bayesian view. *arXiv preprint arXiv:2102.00195*, 2021.

[28] H. S. Fairman, M. H. Brill, and H. Hemmendinger. How the cie 1931 color-matching functions were derived from wright-guild data. *Color Research & Application: Endorsed by Inter-Society Color Council, The Colour Group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch Society for the Study of Color, The Swedish Colour Centre Foundation, Colour Society of Australia, Centre Français de la Couleur*, 22(1):11–23, 1997.

[29] D. Field. Relations between the statistics of natural images and the response properties of cortical cells. *Josa a*, 4(12):2379–2394, 1987.

[30] A. Fiorentini. Mach band phenomena. In *Visual psychophysics*, pages 188–201. Springer, 1972.

[31] J. Froehlich, T. Kunkel, R. Atkins, J. Pytlarz, S. Daly, A. Schilling, and B. Eberhardt. Encoding color difference signals for high dynamic range and wide gamut imagery. In *Color and Imaging Conference*, volume 2015, pages 240–247. Society for Imaging Science and Technology, 2015.

[32] P. Goldstein. Non-macadam color discrimination ellipses. In *Novel optical systems design and optimization XV*, volume 8487, page 84870A. International Society for Optics and Photonics, 2012.

[33] Video Quality Experts Group. Final Report from the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment. Mar. 2000.

[34] P. Hanhart, M. Bernardo, M. Pereira, A. Pinheiro, and T. Ebrahimi. Benchmarking of objective quality metrics for HDR image quality assessment. *EURASIP Journal on Image and Video Processing*, 2015(1):39, 2015.

[35] D. Kundu, D. Ghadiyaram, A. Bovik, and B. Evans. Espl-live HDR image quality database.

[36] Dolby Laboratories. ICtCp white paper. Version 7.2.

[37] V. Laparra, J. Ballé, A. Berardino, and E.P. Simoncelli. Perceptual image quality assessment using a normalized laplacian pyramid. *Electronic Imaging*, 2016(16):1–6, 2016.

[38] G. E. Legge and J. M. Foley. *Contrast masking in human vision*, volume 70. Optical Society of America, 1980.

[39] A. Liu, W. Lin, and M. Narwaria. Image Quality Assessment Based on Gradient Similarity. *IEEE Trans. Image Process.*, 21(4):1500–1512, Apr. 2012.

[40] A. Liu, W. Lin, and M. Narwaria. Image quality assessment based on gradient similarity. *IEEE Transactions on Image Processing*, 21(4):1500–1512, 2012.

[41] W. Liu, Z. Duanmu, and Z. Wang. End-to-End Blind Quality Assessment of Compressed Videos Using Deep Neural Networks. In *ACM Int. Conf. Multimedia*, pages 546–554, 2018.

[42] J. Lubin. The use of psychophysical data and models in the analysis of display system performance. In *Digital images and human vision*, pages 163–178. MIT Press, 1993.

[43] J. Lubin. A visual discrimination model for imaging system design and evaluation. In *Vision Models for Target Detection and Recognition: In Memory of Arthur Menendez*, pages 245–283. World Scientific, 1995.

[44] M. R. Luo, G. Cui, and B. Rigg. The development of the CIE 2000 colour-difference formula: CIEDE2000. *Color Research & Application*, 26(5):340–350, 2001.

[45] K. Ma, W. Liu, T. Liu, Z. Wang, and D. Tao. dipIQ: Blind Image Quality Assessment by Learning-to-Rank Discriminable Image Pairs. *IEEE Trans. Image Process.*, 26(8):3951–3964, Aug. 2017.

[46] K. Ma, T. Zhao, K. Zeng, and Z. Wang. Objective quality assessment for color-to-gray image conversion. *IEEE Transactions on Image Processing*, 24(12):4673–4685, 2015.

[47] D. L. MacAdam. Visual sensitivities to color differences in daylight. *Josa*, 32(5):247–274, 1942.

[48] L. Mai, H. Le, Y. Niu, and F. Liu. Rule of thirds detection from photograph. In *2011 IEEE International Symposium on Multimedia*, pages 91–96, 2011.

[49] J. Mannos and D. Sakrison. The effects of a visual fidelity criterion of the encoding of images. *IEEE transactions on Information Theory*, 20(4):525–536, 1974.

[50] R. Mantiuk, K. J. Kim, A. G. Rempel, and W. Heidrich. HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM Trans. Graphics*, 30(4):40:1–40:14, 2011.

[51] D. Marini and A. Rizzi. A computational approach to color adaptation effects. *Image and Vision Computing*, 18(13):1005–1014, 2000.

[52] K. Masaoka and Y. Nishida. Metric of color-space coverage for wide-gamut displays. *Optics Express*, 23(6):7802–7808, 2015.

[53] S. Miller, M. Nezamabadi, and S. Daly. Perceptual signal coding for more efficient usage of bit codes. In *The 2012 Annual Technical Conference Exhibition*, pages 1–9, Oct 2012.

[54] K. Minoo, Z. Gu, D. Baylon, and A. Luthra. On metrics for objective and subjective evaluation of high dynamic range video. In *Proc. SPIE Opt. Eng. Appl.*, volume 9599, pages 95990F:1–14, Sept. 2015.

[55] A. Mittal, A. K. Moorthy, and A. C. Bovik. No-Reference Image Quality Assessment in the Spatial Domain. *IEEE Trans. Image Process.*, 21(12):4695–4708, Dec. 2012.

[56] A. Mittal, R. Soundararajan, and A. C. Bovik. Making a Completely Blind Image Quality Analyzer. *IEEE Signal Process. Lett.*, 20(3):209–212, Mar. 2013.

[57] M. Miyahara, K. Kotani, and V. R. Algazi. Objective picture quality scale (PQS) for image coding. *IEEE Transactions on Communications*, 46(9):1215–1226, 1998.

[58] N. Moroney, M. D. Fairchild, R. W. G. Hunt, C. Li, M R. Luo, and T. Newman. The CIECAM02 color appearance model. In *Color and Imaging Conference*, volume 2002, pages 23–27. Society for Imaging Science and Technology, 2002.

[59] N. Moroney and Z. Huan. Field trials of the CIECAM02 color appearance. *CIE 25th Quadrennium*, 2003.

[60] R. Mukherjee, K. Debattista, T. Bashford-Rogers, P. Vangorp, R. Mantiuk, M. Bessa, B. Waterfield, and A. Chalmers. Objective and subjective evaluation of High Dynamic Range video compression. *Signal Process.: Image Commun.*, 47:426–437, 2016.

[61] M. Narwaria, M. Da Silva, and P. Le Callet. *High Dynamic Range Visual Quality of Experience Measurement: Challenges and Perspectives*, pages 129–155. Springer International Publishing, Cham, 2015.

[62] M. Narwaria, M. P. Da Silva, and P. Le Callet. HDR-VQM: An objective quality measure for high dynamic range video. *Signal Process.: Image Commun.*, 35:46–60, 2015.

[63] M. Narwaria, M. P. Da Silva, and P. Le Callet. Study of High Dynamic Range Video Quality Assessment. In *Proc. SPIE Opt. Eng. Appl.*, volume 9599, pages 95990V:1–13, Sept. 2015.

[64] The Academy of Motion Picture Arts and Sciences. Academy Color Encoding System (ACES). Accessed on 19.05.2021.

[65] E. Pieri and J. Pytlarz. Hitting the mark-A new color difference metric for HDR and WCG imagery. *SMPTE Motion Imaging Journal*, 127(3):18–25, 2018.

[66] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti. Tid2008-a database for evaluation of full-reference visual quality assessment metrics. *Advances of Modern Radioelectronics*, 10(4):30–45, 2009.

[67] C. Poynton. *Digital video and HD: Algorithms and Interfaces*. Elsevier, 2012.

[68] P. H. Putman. SMPTE Progress Report–Displays 2017. *SMPTE Motion Imaging Journal*, 126(7):83–87, 2017.

[69] J. Pytlarz, E. Pieri, and R. Atkins. Objectively evaluating high dynamic range and wide color gamut color accuracy. *SMPTE Motion Imaging Journal*, 126(2):27–32, 2017.

[70] U. Rajashekar, Z. Wang, and E. P. Simoncelli. Perceptual quality assessment of color images using adaptive signal representation. In *Human Vision and Electronic Imaging XV*, volume 7527. International Society for Optics and Photonics, 2010.

[71] A. Rehman, K. Zeng, and Z. Wang. Display Device-Adapted Video Quality-of-Experience Assessment. In *Proc. SPIE Electron. Imag.*, volume 9394, page 939406:111, Mar. 2015.

[72] M. Rerabek, P. Hanhart, P. Korshunov, and T. Ebrahimi. Subjective and Objective Evaluation of HDR Video Compression. In *Int. Workshop Video Process., Quality Metrics Consum. Electron. (VPQM)*, 2015.

[73] S. Rezazadeh and S. Coulombe. A novel discrete wavelet transform framework for full reference image quality assessment. *Signal, Image, Video Process. (SIViP)*, 7(3):559–573, 2013.

[74] A. M. Rohaly, J. Libert, P. Corriveau, A. Webster, et al. Final report from the video quality experts group on the validation of objective models of video quality assessment. *ITU-T Standards Contribution COM*, 1:9–80, 2000.

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

[76] G. Sharma, W. Wu, and E. N. Dalal. The CIEDE2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations. *Color Research & Application*, 30(1):21–30, 2005.

[77] H. R. Sheikh, M. F. Sabir, and A. C. Bovik. A Statistical Evaluation of Recent Full Reference Image Quality Assessment Algorithms. *IEEE Trans. Image Process.*, 15(11):3440–3451, Nov. 2006.

[78] D. J. Sheskin. *Handbook of Parametric and Nonparametric Statistical Procedures*. Chapman & Hall/CRC, 2011.

[79] E. Simoncelli and W. Freeman. The steerable pyramid: A flexible architecture for multi-scale derivative computation. In *Image Processing, 1995. Proceedings., International Conference on*, volume 3, pages 444–447. IEEE, 1995.

[80] EG SMPTE. 432-1:2010. *SMPTE Engineering Guideline - Digital Source Processing Color Processing for D-Cinema*, 2010.

[81] ST SMPTE. 2084: 2014. *High Dynamic Range Electro-Optical transfer function of mastering reference displays*, pages 1–14, 2014.

[82] W. Sun, F. Zhou, and Q. Liao. MDID: A Multiply Distorted Image Database for Image Quality Assessment. *Pattern Recognit.*, 61:153–168, 2017.

[83] P. C. Teo and D. J. Heeger. Perceptual image distortion. In *Proceedings of 1st International Conference on Image Processing*, volume 2, pages 982–986. IEEE, 1994.

[84] Z. Wang and A. C. Bovik. Modern image quality assessment. *Synthesis Lectures on Image, Video, and Multimedia Processing*, 2(1):1–156, 2006.

[85] 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.*, 13(4):600–612, Apr. 2004.

[86] Z. Wang and Q. Li. Information Content Weighting for Perceptual Image Quality Assessment. *IEEE Trans. Image Process.*, 20(5):1185–1198, May 2011.

[87] Z. Wang, E.P. Simoncelli, and A.C. Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. Ieee, 2003.

[88] A. B. Watson. The cortex transform- rapid computation of simulated neural images. *Computer vision, graphics, and image processing*, 39(3):311–327, 1987.

[89] Wikimedia Commons. CIE 1931 XYZ color matching functions, 2009. [Online; accessed March 20, 2018].

[90] Q. Wu, Z. Wang, and H. Li. A highly efficient method for blind image quality assessment. In *Proc. IEEE Int. Conf. Image Process. (ICIP)*, pages 339–343, Sept. 2015.

[91] G. Wyszecki, K. R. Boff, L. Kaufman, and J. P. Thomas. *Handbook of Perception and Human Performance: Sensory Processes and Perception*. 1986.

[92] G. Wyszecki and G. H. Fielder. New color-matching ellipses. *JOSA*, 61(9):1135–1152, 1971.

[93] G. Wyszecki and W. S. Stiles. *Color science*, volume 8. Wiley New York, 1982.

[94] J. Xu, P. Ye, Q. Li, H. Du, Y. Liu, and D. Doermann. Blind Image Quality Assessment Based on High Order Statistics Aggregation. *IEEE Trans. Image Process.*, 25(9):4444–4457, Sept. 2016.

[95] W. Xue, L. Zhang, X. Mou, and A. C. Bovik. Gradient Magnitude Similarity Deviation: A Highly Efficient Perceptual Image Quality Index. *IEEE Trans. Image Process.*, 23(2):684–695, Feb. 2014.

[96] P. Ye, J. Kumar, L. Kang, and D. Doermann. Unsupervised feature learning framework for no-reference image quality assessment. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 1098–1105, June 2012.

[97] A. Zarić, N. Tatalović, N. Brajković, H. Hlevnjak, M. Lončarić, E. Dumić, and S. Grgić. VCL@ FER Image Quality Assessment Database. *Automatika*, 53(4):344–354, 2012.

[98] L. Zhang and H. Li. SR-SIM: A fast and high performance IQA index based on spectral residual. In *Proc. IEEE Int. Conf. Image Process. (ICIP)*, pages 1473–1476, Sept. 2012.

[99] L. Zhang, L. Zhang, X. Mou, and D. Zhang. FSIM: A Feature Similarity Index for Image Quality Assessment. *IEEE Trans. Image Process.*, 20(8):2378–2386, Aug. 2011.

[100] X. Zhang, X. Feng, W. Wang, and W. Xue. Edge Strength Similarity for Image Quality Assessment. *IEEE Signal Process. Lett.*, 20(4):319–322, Apr. 2013.