Image quality in context

Proefschrift

ter verkrijging van de graad van doctor

aan de Technische Universiteit Delft,

op gezag van de Rector Magnificus prof. dr. ir. J. T. Fokkema,

voorzitter van het College voor Promoties,

in het openbaar te verdedigen op maandag 18 februari 2008 om 15.00 uur

door Jacobus BESUIJEN

elektrotechnisch ingenieur

geboren te de Noordoostelijke Polder

Dit proefschrift is goedgekeurd door de promotoren:
Prof. dr. G.C. van der Veer
Prof. dr. I. Heynderickx

Samenstelling promotiecommissie:

Rector Magnificus, voorzitter

Prof. dr. G.C. van der Veer,	Open Universiteit Heerlen, promotor
Prof. dr. I. Heynderickx,	Technische Universiteit Delft, promotor
Prof. dr. H. de Ridder,	Technische Universiteit Delft
Univ.-Prof. Dr.-Ing. Dipl.-Wirt.-Ing. C. Schlick,	Rheinisch-Westfälische Technische Hochschule Aachen, Duitsland
Prof. dr. ir. G. de Haan,	Technische Universiteit Eindhoven
Prof. dr. ir. J. Biemond,	Technische Universiteit Delft
Dr. G.P.J. Spenkelink,	AE-Rotor Techniek BV Hengelo (Ov)

Het onderzoek beschreven in dit proefschrift is mede mogelijk gemaakt door:
Universiteit Twente in Enschede

ISBN/EAN: 978-90-9022716-0

Samenvatting

Beeldkwaliteit in context

In het onderzoek is eerst een analyse van de ergonomische kwaliteit van normen voor de visuele kwaliteit van beeldschermen uitgevoerd. Hieruit volgen een aantal aanbevelingen voor de opzet van nieuwe internationale standaarden:

- Er zouden afzonderlijke standaarden voor verschillende soorten gebruikers, m.n. beeldschermontwerpers, inkopers en eindgebruikers, moeten worden opgezet.
- De standaarden zouden onafhankelijk moeten zijn van beeldschermtechnologie, om vergelijking tussen beeldschermtechnologieën mogelijk te maken.
- De standaarden zouden modulair moeten worden opgezet met verschillende kwaliteitsgradaties om toetsing voor verschillende toepassingen mogelijk te maken.
- De standaard zou een kwaliteitstoets moeten bevatten, die op de werkplek kan worden uitgevoerd, om rekening te kunnen houden met slijtage en verloop van onderdelen, en correctie van suboptimale instellingen.

De afzonderlijke variabelen, die de beeldkwaliteit van een brede categorie van beelden binnen de context van het gebruik beïnvloeden, en hun onderlinge samenhang binnen de cyclus van beoordeling en aanpassing van beeldkwaliteit, worden in het "Image Quality in Context Cycle"-model gepresenteerd. Tevens is een schema geconstrueerd om de afzonderlijke eigenschappen, die tezamen de beeldkwaliteit vormen, in te delen. In dit schema kunnen de beeldkwaliteitsfactoren onafhankelijk van de technologie, waarmee het beeld tot stand gekomen is, beschouwd worden. De gestructureerde opbouw in vijf verschillende domeinen en acht typen van factoren kan benut worden voor de systematische opzet van normen en beeldkwaliteitsonderzoek.

De praktische problemen bij de instandhouding van de optimale beeldkwaliteit worden onderzocht met betrekking tot de correcte en consistente weergave van kleuren op het beeldscherm. Dit aspect is ook voor de doorsnee computergebruiker van belang, bijv. bij het bewerken van digitale foto's, het voorbereiden van presentaties en het beoordelen van mogelijk online aankopen. Voor eenvoudige PC's is hiervoor de sRGB-standaard van toepassing. Afwijkingen van de correcte kleurweergave kunnen gemakkelijk ontstaan door verkeerde instellingen van apparatuur of programmatuur, verloop van componenten en invloed van omgevingsverlichting.

Bij dit onderzoek zijn eerst een aantal substantiële problemen met de modellen voor de overdrachtskarakteristieken van de kathodestraalbuis en de methoden om deze vast te stellen aan de hand van deze metingen aan het licht gekomen. De gangbare modellen gaan uit van een machtsfunctie met constante exponent gamma en mogelijke nulafwijkingen in elektrische spanning en/of luminantie (hoeveelheid licht richting gebruiker). De parameters van deze modellen worden geschat op basis van het minimaliseren van de fout in de fysische grootheid luminantie ten opzicht van een reeks metingen over het gehele luminantiebereik. Deze methoden leveren aanzienlijke fouten op in de bepaling van de nulafwijkingen. Een nieuwe methode wordt geïntroduceerd waarbij de optimalisatie plaats vind op basis van de fout in de gestandaardiseerde perceptuele grootheid 'lightness', die overeenkomt met de waargenomen helderheid. Uit diverse metingen blijkt, dat deze methode betere modellen met een nauwkeuriger schatting van de nulafwijkingen oplevert. Tevens blijkt, dat voor sommige beeldschermen betere resultaten worden bereikt als de aanname van de constante gamma wordt losgelaten.

De praktische mogelijkheden van een gebruiker om de kleurweergave van zijn/haar beeldscherm te optimaliseren door het correct instellen van het zwartniveau met behulp van de "contrast"- en "brightness"-instelmogelijkheden, en de visuele meting van de monitorgamma('s) zijn onderzocht in een experiment met 32 proefpersonen. De helft van de proefpersonen heeft het experiment uitgevoerd in vrij donker schemerlicht, de andere helft in normale kantoorverlichting. Alle personen hebben in het experiment twee beeldschermen ingesteld: één van het type kathodestraalbuis (beeldbuis) en één Liquid Crystal Display (LCD). Als uitgangspunten bij de instellingen zijn de ingebouwde sRGB-instellingen gebruikt.

Bij de instelling van het zwartniveau werd gebruik gemaakt van twee verschillende typen van instructie: een korte instructie bestaande uit één dialoogscherm met stimuli en instructie van Adobe Gamma Wizard en een uitvoerige instructie bestaande uit drie stimulischermen van Sonera DisplayMate met instructies op papier.

De meeste proefpersonen waren niet in staat om het zwartniveau in de buurt van optimaal in te stellen, onafhankelijk van de verlichtingsomstandigheden, het gebruikte beeldscherm of de uitgebreidheid van de aangeboden instructie. Hierbij is het mogelijk dat de technische beperkingen van de beeldschermen een rol hebben gespeeld. Bij het LCD-scherm kon de instelling van het zwartniveau niet beduidend worden aangepast en werd de kwaliteit van de instelling afgemeten aan het bereikte contrast. Bij de beeldbuis-monitor bleek de optimale zwartniveau- instelling een grotere afwijking van de sRGB-standaard op te leveren.

Voor de visuele meting van de exponent gamma zijn in de literatuur verschillende methoden voorgesteld. De theoretische merites en praktische problemen van deze methoden worden besproken en vergeleken, en verbeteringen worden voorgesteld. Een nieuwe methode om gammamodellen met een verschillend aantal parameters (wel of geen mogelijke nulafwijkingen) te vergelijken wordt geïntroduceerd.

In het experiment zijn verschillende methoden van helderheidsvergelijking getest: spatiële methoden (streepjespatronen met egaal vlak, zie blz. 96) met drie verschillende resoluties en een temporele methode (flikkerend vlak met egaal vlak, zie blz. 96). De verlichtingsomstandigheden hadden geen effect op de schatting van gamma. De proefpersonen hadden grote problemen met spatiële helderheidsvergelijking bij lage resoluties. Temporele en spatiële methoden leverden voor ongetrainde proefpersonen een hogere gamma op, dan voor de expert en de berekeningen op grond van fotometrische metingen.

Acknowledgements

First of all I have to thank my supervisors: Gerrit van der Veer for picking up an orphaned Ph.D.-candidate and providing support and advice to fulfill this thesis. Ingrid Heynderickx for her expert comments on image quality matters. And both for their patience with a self-willed and sometimes despondent candidate.

A number of minor but essential technical contributions by Johan Jonker and Robert Lodewijks to the light measurement equipment and by John Anemaet to the experimental computing environment made the handiwork possible. The basic fitting algorithm and the first steps in its application on the gamma model were supplied by dr. Ruud van Damme.

In their ultimately successful efforts to dismiss me, the Faculty of Behavioral Sciences of the University of Twente provided me with the sabbatical period needed to finish my research and write this thesis. The faculty director dr. Jan Verberne mediated the opportunities to conduct the experiments, while specifying the reasons for my dismissal. Professor Willem Verweij supplied the extra boost to finish this project by denouncing my capabilities. Antagonists are often essential for movements to reach their target.

Thanks to family, friends and colleagues, esp. roommate Ria Verleur, who stood by me in troubling times, and encouraged me to finish this effort. An extra pat for Bella, whose daily needs gave me time to contemplate away from the display.

Special thanks for my long time colleague Gerd Spenkelink, who introduced me to the world of ergonomic experimenting, and applied the same level of standards for valuable research.

Für Sabine,

die mich gelegentlich daran erinnert hat, dass sie dachte sie hätte einen intelligenten Mann geheiratet.

Contents

1 Introduction

1.1 The meaning of image quality

Image quality is one of the deciding factors for customers buying a display, a camera or a printer. All promotional material for these products advertises some image quality features, esp. the size, the number of picture elements (pixels), the sharpness and the vividness of the colors. Image quality is also an important feature in comparative tests in consumer magazines. Display showrooms in shops are set up as beauty contests, presumably to entice the customer to spend more money on a display with a better-looking image. Quality images also boost the sales of the products they advertise on online shopping sites. A Greenfield Online research study conducted in December 2001 [AIM02] showed that people are more likely to purchase products from Web sites that allow them to see the product in more detail.

The quality of images is not one static property: image quality appreciation varies with the context in which the image is seen. Ambient illumination, the condition of the human visual system, cultural differences, the viewer's experience and the purpose, for which the image is viewed, can affect the level of image quality appreciation. Expectations for the quality of images are raised over the years. Originally people were happy to be able to see any image of something they could not see directly at that moment: a small faint monochrome picture on photo (1826) or television (1926). Since then size, contrast and sharpness increased, and movement, color and depth were added.

The introduction of the personal computer with its enormous possibilities for retrieving, exchanging, storing and visualizing information has given an enormous boost to the quality of electronic displays. In 1981 the IBM PC was introduced with the Color Graphics Adapter (CGA) video card, which could show either an image with 320x200 picture elements (pixels) in four colors or a 640x200 pixel image in monochrome on the 11.5 inch (29.2 cm) diagonal display. In 2007, 26 years later, common portable computers on sale are advertised to show 1440x900 pixels in 16 million colors on a 17 inch (43.9 cm) display, more than doubling the area of the display with a ten to twentyfold increase in the number of pixels: Together with the millions of added colors amounting to an enormous increase in image detail.

Consequently, people now expect more quality from images, although they are temporarily satisfied with lower quality if the images show scenes they could not see before (medical, microscopy) or in places they could not see them before (mobile phones) or if the images are presented directly on demand (internet). But improved image quality sells, as the market proofs, although the transfer to a higher level of image quality for TV is slowed down by a lack of content (programs)[Atw07b], the phletora of choices presented to the consumer concerning display technology, cables, tuners and providers, and by the competition between the high definition DVD formats, Blu-ray disc and HD DVD [Atw07a; Burn06; Kim07; Swa06]. A lot of consumers have got lost in the technology jungle: A recent (2007) survey by the Leichtman Research group showed that 50% of HD-Television owners could not receive HD-programs, and about half of these don't even realize they are not watching HDTV [Abo07].

Due to the rapid developments in display technology, it looks like display image quality is approaching that of the original, and steps in the improvement of the quality of the image are getting ever smaller. With an equal pace of improvement the average PC display of the year 2033 for example would have a 24 inch (61 cm) diagonal with about 5680x3195 pixels and 64 trillion (10^{12}) colors. With 270 pixels per inch (ppi) such a display would probably exceed the resolution needed to display text (with anti-aliasing) with a quality at least equal to high quality print (1200dpi) [YK03]. The number of colors that the eye can distinguish is estimated to be about two million colors [McC98; PA98], which means there is really no need to improve upon the number of 16 million that are available now.

Research and development is now shifting its attention to enhancing the viewing experience [Hey07] e.g. by adaptation of the background illumination to the image content and/or user preference (Philips Ambilight) and the improvement of 3D-displays. For the research regarding the viewer this means a shift of focus from basic visual perception to more emotional and cognitive factors:

- The Ambilight concept enlarges the visual field of the display thereby increasing the immersion of the viewer.
- In viewing 3D-displays cognitive dissonances can occur if head movements don't provide the expected views of the scene or if the perceived distance does not agree with the angular size of the objects [MIJS04].
- Presence, the feeling of being part of the world in the image, is an important aspect of the quality of 3D-displays.

For 3D-displays the meaning of the concept image quality does not necessarily cover all aspects of the perceptual quality of the image: two subjective rating experiments reported by Seuntiëns [SHIJ07] showed that the stereoscopic depth of an image had no effect on the image quality rating, but did have an effect on the rating of the naturalness of the image. Basically this is a semantic problem also signaled by Janssen en Blommaert [JB98] for color images: What do viewers/users evaluate when they rate image quality: beauty, the perceived distance from the original, the usefulness, or the naturalness?

In a working situation the quality of images on a display is not just a matter of beauty and visual experience: it also has an ergonomic aspect. Ergonomic visual display quality is defined by the way it can affect the visual comfort and the performance of the worker, which can induce stress and health problems. The International Standards Organization's (ISO) new draft standard 9241-300: "Ergonomics of human-system interaction -- Introduction to electronic visual display requirements", states that it is: “aimed at ensuring effective and comfortable viewing conditions for users”. Furthermore a user performance and comfort test is specified as a route to compliance with the standard.

The problem is that health problems, stress, performance and visual comfort (e.g. eye strain) are indirect measures concerning phenomena that need considerable time to manifest itself, and depend on many factors not related to visual display quality (e.g. environmental, competence level, amount of training, psychosocial, and physical). It is very difficult therefore to create experimental circumstances in which these measures are distinctive for comparatively small differences in image quality [SB99].

1.2 Context of the research

The ergonomic visual display quality of simulated Flat Panel Displays [Spe90] was the starting point of our research in 1988. In that project the Display Evaluation Scale (DES) [SBB93], a direct subjective rating scale for visual display quality factors, was developed. After that project the focus of research broadened to the measurement of visual display quality on all displays, culminating in the PhD. dissertation on visual display quality of Gerd Spenkelink in 1994 [Spe94]. Inevitably the Display Evaluation Scale developed in our research had to be compared with the user tests suggested for the international standards for visual display quality [ISO92; ISO00]: the starting point of the current research project, reported in chapter 2.

The comparison shows that the role of standards needs reconsideration in the fast changing field of display technology and applications with different demands. An important conclusion reached in this comparison is that there should be separate standards for engineers and display users. Visual display quality should also be assessed in the field in the context of use. In a separate line of research a tool for the on site measurement of visual display quality has been developed [SBR00].

Because digital images are easily exchangeable over different systems, the range of possible contexts of use is enlarged. There is an ever increasing need to be able to compare the quality of images on different display systems and hardcopy. The distinction between the concepts of video display quality (computer work) and image quality (television and photography) is fading. A comprehensive model showing the way in which contextual factors can influence the quality of all kinds of images is presented in chapter 3.

A second conclusion from the analysis of the ergonomic quality of the video display quality standard in chapter 2 is that there should be a higher degree of independence of the standard from the display technology. The difference between standards for images of different sources and displays of different technologies inhibits the comparison of image quality between systems. Quality is in the eye of the beholder, and when the light distribution that forms the image reaches the eye, the source of the images and the technology of the displays no longer matter for the appreciation of the image quality. The construction of a set of features for image quality, independent of imaging and display technology is discussed in chapter 4.

An important aspect in the ergonomic context of working with image quality is service and maintenance. Over time the quality of the image itself is not static: The ambient light changes, hardcopy photos bleach and fade away, display and graphics card components age and alter the balance of the colors and the contrasts, and users or system administrators may change hardware controls and software settings. In many cases the defects in image quality found for a display can be serviced by changing hardware or software settings. The ability of the average user or system administrator to correct these settings and the type of instructions and instruments that could be provided to support them in their endeavor is the next focus of our research. The characterization and correction of the tone reproduction curve (TRC) is singled out for the research and the results are reported in chapters 5, 6, and 7.

The tone reproduction curve determines the relationship between the digital primary color values from dark to light of the digitized image and the luminance output of the primary colors on the display. The TRC is chosen for its impact on true-to-life, color display and color management. Both the technological and the human vision aspects of the TRC are thought to be well researched and standardized. The ideal shape of the display's TRC is laid down in standards [DIC03; IEC1998], and several models exist to describe it. The TRC of a display type or an individual display can also be measured and recorded in a color profile, which can be used in color management software to produce more accurate images on the display. The measurement and maintenance of color profiles is a task unsuited for the resources of the average computer user.

The visual perception of brightness, contrast and color differences is well researched and generally acknowledged models, CIELAB and CIELUV [CIE06], exist to describe it.

A great number of problems concerning the context of image quality discussed in chapters 2-4 can affect the shape of the TRC: hardware configuration, software settings, wear, environment, user control settings, differences between technologies and applications.

For the traditional Cathode Ray Tube (CRT) display the TRC is generally indicated by a single number: the parameter gamma of the power function describing the curve for the ideal CRT. Display system gammas usually lie in the range between 1.8 and 3.0, with most common values between 2.0 and 2.5. For a true-to-life display of the image, the system that acquired the image should transfer light to digital values with the inverse gamma function, e.g for a gamma 2.0 display system the camera system should have an inverse gamma of 0.5 = 1/2.0

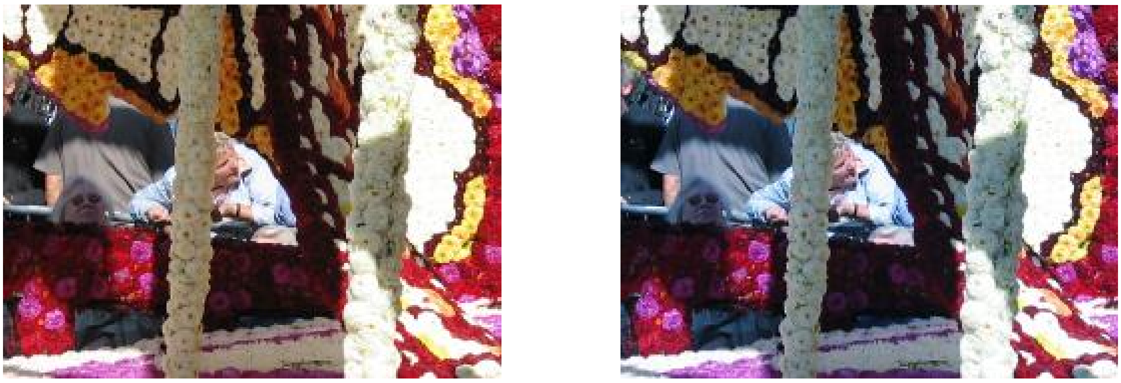
The inverse gamma function conversion produces very effective transmission and storage of images, where the steps between digital values are nearly equal to the just noticeable differences in luminance that humans can perceive. Therefore and for reasons of standardization all other display technologies have tried to adapt the gamma function shaped TRC of the CRT. Unfortunately many computer users have no idea what gamma is, or what the effects of a change in gamma on the displayed image could be. In practice the gamma of a display is not always what it should be, and there is no simple gamma control attached to correct it.

With a few examples we will try to show what the effect of an incorrect display gamma is on the displayed image. Figure 1.1a shows a print of the original image shot at the 2004 Lichtenvoorde Flower Parade. First a word of caution for the reader/viewer: Every printer or display system has its own TRC with its own shape and will present its own version of the original image. So the original could be too dark or look a bit washed out. To make sure that the examples of the effects are not submerged by the peculiarities of the specific printer or display system, the effects in the images are exaggerated. It is therefore not useful to present gamma values or black level settings with the pictures. From the author’s experience the effects will show up stronger on a CRT display than on print or projection.

If the gamma of the display is lower than expected (fig 1.1c), the TRC ascends too steep for the shadows and too slow for the highlights, the image will appear washed out, and detail in bright parts of the image may be lost. These effects can best be noted in the colors of the face and the shirt of the leaning man in the middle, and in the loss of detail in the white patch of flowers on the right. If the gamma is higher than expected (fig 1.1d), the TRC ascends too slow for the shadows and too steep for the highlights, the contrasts in the image will be hard, and detail in dark parts of the image may be lost. If the forms (gammas) of the TRCs for the primary colors differ, in this case a lower gamma for blue and a higher gamma for red (fig 1.1b), then the colors for the midtones will shift (to blue): best seen on the white flowers in the shadow and the grey T-shirt of the (headless) man on the left.

A wrong setting of the display black level, which can usually be corrected with the display brightness control, has a similar effect on the image as a wrong gamma. If the digital code value 0 is not black but dark gray (fig 1.1e), then this has a similar effect as a gamma that is too low. If not only the digital code value 0 but many lower code values display to black (fig 1f), then this has the same effect as a gamma too high.

The research project is set out to analyze the abilities of users to visually inspect and optimize the display settings concerning the shape of the TRC with different types of instructions and inspection methods. Measuring and fitting the TRC of a Cathode Ray Tube (CRT) to the physical model turns out to be more than a standard procedure. Current gamma models and fitting procedures are found to be inaccurate in modeling the darker part of the



a. "original"

b. blue gamma too low, red gamma too high

c. gamma too low

d. gamma too high

e. Black level too high

f. Black level too low, clipping

Figure 1.1. Examples of the effects of gamma and black level on color images

tone reproduction curve: they tend to clip the luminance to zero for digital values were luminance can be seen and measured (compare fig. 1.1f).

1.3 Contributions of this thesis

- We produce an analysis of the existing ergonomic visual display quality standard, and suggest measures to improve the usability and effectiveness. Some of these measures have already been incorporated in the drafts of the new standard.
- We present a more comprehensive model for the evaluation of image quality in the context of use.

- We present a classification model for image quality factors that allows a comparison of image quality from different technologies.
- We develop a perceptually more accurate method for fitting gamma functions to the tone reproduction curve of a cathode ray tube (CRT).
- We provide evidence that a fixed-gamma model cannot accurately describe the luminance output of all CRTs over the whole input voltage range.
- We hypothesize that reflected ambient illumination affects the ability to correctly adjust the black level of a CRT, and obtain statistical evidence through our experiments to support this theory.
- We develop a generally applicable and accurate method for fitting gamma curves to visual gamma measurement data.
- We develop a method for comparing gammas from models with and without offset parameters.
- We show a number of practical problems with the spatial as well as the temporal visual gamma matching method for CRTs and LCDs.

1.4 Thesis outline

The subsequent chapters are divided as follows:

Chapter 2: Standardizing visual displays. In chapter 2 the role of standards in the preservation of image quality on displays is discussed. The chapter is based on an article published in 1998 [BS98] on the at that time actual ISO 9241-3 [ISO92] standard for visual display quality and the proposed user performance tests. The standard is found to be more useful for defining design requirements for engineers than for ergonomic analysis in the use environment. Problems with system configuration, software applications, display settings, user behavior, wear and physical environment are addressed. Different methods of visual display quality measurement are compared and direct subjective methods are considered to be the most valid and sensitive. Results with the Display Evaluation Scale, a tool for rating visual display quality [Spe94], are evaluated. Three user groups of the standard are distinguished and a new configuration and content of standards is proposed to suit these user groups. The concluding paragraph calls for a validated image quality model based on a broader perspective combining display technology and human vision research.

Chapter 3: The image quality in context cycle. To address the problem of different standards for different applications and user groups a new model of image quality assessment is presented in chapter 3 that puts image quality in the ergonomic context: the user, the task, the environment, the interaction, the different (display) technologies, the physical image, and visual perception.

Chapter 4: A Classification framework for visual display and image quality. A framework for image quality features is presented based on five image signal domains and eight types of quality factors. Four of these quality factors are specific to each of the domains; the other four depend on several domains. The signal domains can be classified in two ways. The first distinction is made between the basic spatial and luminance domains, necessary to reproduce an image, and the extra domains chromaticity, temporal and depth (3D), which add extra qualities and naturalness to the images. The second distinction is made between the coordinate domains spatial and temporal, which define the position and time of the information and the value domains luminance and chromaticity, which define the value of the information. The 3D-domain is implicitly a coordinate domain, but the position in depth has to be reconstructed and cannot be directly pointed to in the data.

Chapter 5: The shifting gamma perception. A special study [Bes05] reported in chapter 5 is concerned with the analysis and improvement of the Gamma-Offset-Gain-(GOG) and Gamma-Offset-Gain-Offset- (GOGO) models and fitting methods used in color management standards. Two improvements are suggested: the models should be fitted by optimizing the root mean square error (RMSE) of the CIE lightness instead of the luminance. Secondly, based on reports in literature that extensive voltage measurements support refinements of the CRT physical model, an alternative gamma model is adopted, with a steeper TRC for lower input levels. Results show that the adapted models correspond better with the luminance measurements for the two CRTs in the study and for one reported in literature. The correct setting of the black level is an important step in the (re)calibration of an electronic display.

Chapter 6: Black-level offset: Characterization and correction. The study [Bes06] reported in chapter 6 looks at the consequences of black-level offset, the possibilities for display characterization with offset, offset correction, and the ability of average untrained users to visually correct the black-level setting with the contrast and brightness controls on the display. In an experiment, 32 subjects were asked to optimally set the black level according to two types of instructions (short and extensive, between subjects) under two levels of illumination (low and office, between subjects) for two types of displays (CRTs and LCDs, within subjects). Most subjects were not able to set the black level near optimal for either display, with any combination of instruction and illumination level. The LCD did not have an optimal black level. For the CRT, optimal black level did not provide minimal differences with the sRGB standard tone reproduction curve.

Chapter 7: Visual gamma measurement and methods to compare gamma models. The second part of this experiment [Bes07] is reported in chapter 7 and compares several methods for visual gamma estimation that have been proposed. The correct estimation of the gamma exponent describing the tone reproduction curve of a display is an important step in color management. In this study the theoretical merits and practical problems of a number of visual gamma estimation methods, allowing the user to calibrate a display system without photometric measurement instruments, are discussed and compared, and improvements are suggested. A new method to compare gamma models with different numbers of parameters is introduced. In the experiment spatial and temporal brightness matching methods were tested with 32 untrained subjects working on a CRT and an LCD with different resolutions under office and low illumination conditions. Illumination had no effect on gamma estimations. Subjects had great difficulties with spatial brightness matching at low spatial resolutions. Temporal and spatial visual brightness matching for untrained subjects showed a larger gamma than photometric fits.

Chapter 8: Conclusions and Recommendations. In the last chapter the conclusions and recommendations of the previous chapters are summarized and expanded.

2 Standardizing visual display quality

Abstract

The current ISO 9241-3 standard for visual display quality and the proposed user performance tests are reviewed. The standard is found to be more engineering than ergonomic and problems with system configuration, software applications, display settings, user behavior, wear and physical environment are addressed. Different methods of visual display quality measurement are compared, direct subjective methods are considered to be the most valid and sensitive. Results with the Display Evaluation Scale, a tool for rating visual display quality, are evaluated. Three user groups of the standard are distinguished and a new configuration and content of standards is proposed to suit these user groups.

2.1 Introduction

A visual display unit is a technical device that serves its master, the display user, by outputting information that the user needs to process. The visual display quality pertains to the efficiency with which information is transferred from the display to the visual system of the user. Visual display quality or image quality, therefore, is primarily defined by the degree to which images are adapted to the capacities of the user's visual information processing system. Visual display quality criteria should be user centered.

This is exactly what one would expect from part 3 of the ISO 9241 series [ISO92], which gives the ergonomic requirements for the visual display in office work situations. Lately there has been a lot of debate about this standard for two reasons. First, proposals for the user performance test, only applicable to novel display technologies, have met with a lot of criticism. Second, the standard is due for revision and it is generally acknowledged that this will have to be a major revision, because the display technology has changed enormously since the publication of the standard in 1992. Another point for discussion, however, should be high on the agenda of the parties involved with this standard. This point concerns the ergonomic relevance of the standard.

The ergonomic contents and validity of the standard depend on a number of factors: the relevance and validity of the criteria given by the standard, the validity and reliability of the compliance test and the usability or practical value of the standard. In this article we shall analyze the current part 3 of the display standard in the light of these factors, and formulate proposals for improvements. It will be shown that in order to deal effectively with the quality of visual human-display interaction, fundamental changes to the standard are essential. We shall devote much attention to user performance testing, as this turns out to be a critical factor in the usability and practical validity of the standard.

2.2 The current standard

The standard ISO 9241: 'Ergonomic requirements for office work with visual display terminals (VDTs)', is organized into 17 different parts, each dealing with a different aspect of VDT use. Four parts provide extensive general guidance, five parts deal with equipment, two with the work environment and six parts relate to the software used. Although the title of part 3: 'Visual display requirements' suggests that it contains all requirements for displays, additional requirements are set in part 7 for display with reflections and part 8 for displayed colors. The eventual quality of the displayed image could also depend on the requirements set in part 6 on environment and part 5 on workstation layout and posture.

In part 3 visual display quality is defined as sufficient if 25 requirements are met. The way in which the display requirements are defined is based on physical measurements of the display. Therefore, it is not surprising that the primary compliance route consists of a large number of physical measurements in the luminance and spatial domain. It requires complicated measurement instruments and special technical knowledge to perform the test.

ISO protocols demand that ergonomic standards are accompanied by a user performance test as an alternative compliance route. In an informative annex of the standard a comparative user performance test method was proposed. This test would become an addendum to the standard when it had been proven. Other parts of the 9241 standard, part 7 [ISO98] on reflections and part 8 [ISO97c] on color, as well as the draft flat panel standard ISO/DIS 13406-2 [ISO01] refer to the user performance test method in part 3 as an alternative.

Practice with the originally published test method showed several problems and based on this experience a new method is now proposed [ISO00]. The proposed test consists of a character searching task and a single visual comfort rating. A display under test will conform to the standard if both the search velocity and the visual comfort rating are not significantly lower than those for a benchmark display tested with the same subjects. The benchmark shall be a display known to meet or exceed the mandatory requirements based on the physical measurements defined in ISO 9241-3.

In this most recent proposal the validity of the user performance test method has been limited to displays that cannot be tested with the physical measurement protocol. This limitation seems to be caused by resistance from engineers to "a simple performance task and comfort rating", applied in a comparative set up, replacing all 25 items of the design requirements and recommendations. In view of the title of the standard and the way in which part 1 [ISO97b] focuses on the user performance approach, we can conclude that an ergonomic standard is threatened to be taken over by technical engineers.

2.2.1 Problems with ISO 9241-3

There are a lot of known (and as yet unrecognized) problems with the criteria and physical measurement part of ISO 9241-3. First of all, there is a modeling problem. The primary compliance route consists of a large number of physical measurements, but there is no model specified that weighs or integrates the different measurements. Therefore, for example, a meaningful overall quality index cannot be calculated. Many of the measured physical characteristics interact in perception, e.g. contrast sensitivity depends on the level of retinal illuminance and thus interacts with the perceived brightness. The perceptual quality characteristic sharpness has no equivalent physical characteristic in the standard. The nearest related measure, the inner-character contrast, depends on character size and shape.

The implicit model behind the primary compliance test consists of a collection of limits for physical characteristics that are treated as if they were independent and in combination define image quality. Imaginative readers might be able to experience that this model is incorrect, by trying to envision a display that would just, and only just, meet all the individual criteria of the compliance test. The result is a display that would pass the test, but as the mind's eye will reveal, also would be an awkward display no user would be willing work with.

The independent factors model behind the primary compliance test stems from the fact that the criteria are mostly based on (laboratory) research in which quality aspects often are studied in isolation. The interaction between brightness and contrast is well known and has consequences for the image quality [Spe94], chap. 3. Interactions also exist between e.g.

character height and leading (= line spacing) and between line width and leading [Spe94], chap. 7.

The results of extensive research on readability/legibility [Dil92; Tin63] indicate that a low level of discomfort on a single aspect of visual display quality does not slow performance, but a combination of low discomfort levels from several aspects does. Since a lot of the physical requirements in the standard are based on research with performance measures it is very likely that displays that just meet those requirements produce visual discomfort and even loss of performance with the users.

As it is difficult to design, carry out and analyze experiments with (multiple) interacting factors, it will take considerable time before a coherent, widely accepted and validated model of image quality is reached.

Second, there are a number of practical problems. VDTs have evolved rapidly in the last decade and the requirements and protocols of the standard are still based on CRTs with single scan rate and resolution, and a built-in character generator. Already another international standard ISO 13406: "Ergonomic requirements for visual display units based on flat panels" has been drawn up to be applicable to "flat panel display screens when used to perform office tasks." On an other competing level, standards are being devised for electronic image management systems and CAD/CAM work, which will also contain sections about visual display quality. The overlapping of these standards does not serve the interest of the users or designers of visual display units and a reorganization in more modular units seems pressing.

In countries that have ISO 9241 as a national standard, as the countries of the EU, only ISO 9241-3 approved video display units will be bought. Hence practically all new models of visual display units are tested by an accredited test house before they are put on the market. It is very likely that the display manufacturer will have conducted identical tests early in the design process, to make sure that the costs for testing are being well spent. The buyer of a new display for his or her office will be certain that the display is able to meet the requirements of ISO 9241-3 under certain conditions. What the buyer doesn't know is whether these conditions are met in the operational situation, which is the daily use in the office. The conditions are related to system configuration, software applications, display settings, user behavior, display wear and physical environment.

2.2.2 Hardware configuration and software characteristics

A major problem embedded in part 3 is that many of the requirements have a quality-determining source outside the VDT. In modern computer work stations, resolution and scan rate are largely determined by settings of the graphics card, and character properties by a software font generator. It might be possible to provide separate test reports for all combinations of possible resolutions and respective maximum scan rates, but it is unfeasible to measure character properties for all available fonts. Moreover it is useless information, because the user is practically unable to determine the exact font displayed: In MS Word on a Windows'95 PC the appearance of a 10 pt. Arial font depends on the resolution (800x600, 1024x768, etc.) and font size settings (small fonts, large fonts, etc.) in the display properties menu on the desktop, the zoom factor (percentage, page or page width) specified in the application's view menu and the amount of dots per inch (dpi) specified in the display driver.

At the time the physical requirements of the standard are measured the quality of the displayed image will very likely be different from the quality at the time of actual use.

2.2.3 Display settings / user behavior

Apart from the well-known brightness and contrast controls the user is now provided with an abundance of controls or menus with which vertical and horizontal convergence, size and position, trapezoid and pincushion distortion, rotation, pixel clock, etc. can be changed. (A separate standard could be drafted for this user interface alone.) Considering how users do sometimes deteriorate image quality by the adjustment of just brightness and contrast controls, it is doubtful if these new features contribute to an improved image quality for all users.

In many applications the user can exert a large influence on the image quality on the operational level, for instance by selecting a zoom factor. Often, users are not aware of their options or refrain from using them and never change the installed defaults. In an evaluation of display workplaces in one of the faculties of our university [EDUT96], where text processing is the main application, it was observed that 18% of the employees using a display for two hours daily or more, worked with displayed characters that were too small (<16') for acceptable legibility. Another 26% of the users worked with character heights of 16'-18', which is qualified as sub-optimal in the standard. Most of the users in these categories worked with an application in which they could have easily changed the zoom factor to improve the character size. In some cases the users could have reduced the (large) viewing distance. It could be that some of the users preferred to have more information on the screen, but it is alarming to see that 44% of the users spent two or more hours daily staring at sub-optimal images, or images that did not meet the requirements of the standard.

2.2.4 Display wear

If one examines the visual display quality in an office, the suspect cases will probably be older visual display units. It is not unusual for a CRT to loose 30%-50% of its luminance in the first year of use. An increase in geometrical distortions, misconvergence and non-uniformity can also be expected. Some effects of wearing might be temporarily compensated by adjusting the settings, but this is not an easy task for the average user and it might evoke other quality problems. In the end the quality will deteriorate to a questionable level and an employer will be faced with the question: "Do I have to replace these three year old displays or can I wait another year?". ISO 9241-3 does not provide any methods to solve this problem: It is not feasible to send all suspect displays to a test house.

2.2.5 Physical environment

The visual environment in the office and the work station lay-out are important factors for the quality of the displayed image. Several design requirements in the standard can only be assessed at the actual office work station. Viewing distance, line-of-sight angle and character height depend on the actual placement of the display with regard to the user. The visual environment determines luminance balance and glare, and has a large influence on angle of view, luminance contrast, temporal instability (flicker interference with fluorescent lights) and screen image color. One might argue that some of these influences are accounted for in measurement procedures and other parts of the standard, but it is impossible to account for all office situations. The use of an anti-reflection screen in front of the display for instance, requires at least new display luminance, luminance contrast and luminance balance measurements.

So far, we can conclude that part 3 of the standard is outdated, because it was based on a now outdated terminal concept and was drawn up in very technology dependent fashion. Further, the operational circumstances have a large influence on the actually achieved image quality.

These circumstances are difficult to deal with, because they are so variable. The image quality that counts, however, is the one achieved in practice. A physical measurements approach towards compliance testing does not appear to be fruitful. Apart from the modeling problem, the operational circumstances cannot be incorporated in physical measurements. The problems might be solved by taking a multi-level approach. A starting point for such an approach can be found in the diversity of interests between user groups of the standard.

2.3 Users of the standard

The standard covers the display design process from technical engineering to implementation and use. Three groups of users can be identified that have specific interests in the standard.

First there are the display designers. As it stands, part 3 of the standard serves to communicate the ergonomic interests of end users to designers. For this goal, visual ergonomic issues have been translated into display hardware related variables, for which performance objectives are specified. Unfortunately, as we have already discussed, some of these variables are now outdated and the control over other variables is now placed outside the display. Furthermore, modern displays generally are capable of better performance than is required by the standard. In practice, the standard is a passed station for most display technologies, although this does not imply that the image quality is okay.

Second, there are the system designers, managers or others that are responsible for purchase decisions. The decision which display is going to be used not only has economical consequences, but also has its impact on the ergonomics of the workplace. Here too, the users' interest should be considered, as stated in Part 1: "ISO 9241 is also relevant to purchasers who wish to specify VDT systems for use in their own organizations, and those who wish to evaluate the suitability of existing equipment, working environments and work practice." Currently, the factors that define the quality of a CRT in the eyes of the buyer are probably: screen size, bandwidth and dot pitch (it is what the sellers tell them is quality). The buyer probably will be interested to hear that a display complies with ISO 9241-3: (s)he doesn't want inferior displays in the office. Compliance with the standard, however, is of marginal interest when comparing the quality of available displays for a specific task. The standard only defines two categories of visual display quality: Bad and Not Bad. This might give the false impression that there is no distinction in quality for the majority of the displays that do comply with the standard.

Part 1 of ISO 9241 seems to allow a less absolute view of compliance than the one adopted in part 3: "What is appropriate in one set of circumstances may be inappropriate in a different context; when using VDT ergonomics standards it is important to recognize that the potential range of application is very broad. Therefore ergonomics standards often take the form of recommendations, or requirements, which are conditional upon certain defined circumstances."

The user is no longer content with 24 lines of 80 characters on a 12" display screen, even if it complies with the standard. The current standard (in 1998) is 'WYSIWYG' on a 15" or 17" screen and because 'What You See' on your screen is still different from 'What You Get' on your printer, further quality improvement of displays is inevitable. The requirements for visual display quality will have to be raised until reading from VDTs is as easy as reading from paper. If the desktop metaphor is further carried through, display size will tend to increase. A larger screen size provides operational comfort and reduces the amount of visual uncomfortable scrolling needed.

The relevant information for the buyer is what level of device independent visual quality can be reached without negative interference from the display technology. E.g. if a CRT of a certain size, while complying with the requirements on orthogonality, uniformity, spatial instability, etc., can show a certain graphics card resolution with a sufficient scan-rate and acceptable sharpness. The required resolution depends on the application. The present standard is not helpful to the potential buyer, because it was not set up for comparative evaluation and does not provide a quality index, let alone a composite quality index. Furthermore, the relations between display hardware and configuration, task environment, physical environment and workplace design are not accounted for. Therefore, in the present situation, a decision maker who wants to make an informed choice will have to go through a rather complex design process and will probably be in need of a comparative evaluation method valid for the intended application environment.

Third, there are the end users. Although in general the end users do not take the design decisions because manufacturers and display purchasers have done that for them, users certainly have a lot of control over the ergonomics of the human-display interaction. One prerequisite is that the user has the knowledge and tools to utilize this control in a positive manner. A standard could be very helpful in that respect. The user should be supported in establishing e.g. that the system is badly configured; that the interaction between display, character generator, graphics card, set up and control adjustment produces sub-optimal visual display quality. There might be interactions with work station layout and visual environment that deteriorate the image quality. The display may become worn out.

In the general introduction to ISO 9241 [ISO97b] these three user groups are also named as the intended users of the standard, if we assume the end user is represented by 'those who wish to evaluate the suitability of existing equipment, working environment and work practices.' At the moment the parts that deal with visual display quality seem to be mainly intended for display designers and manufacturers. And even the alternative compliance route through the user-performance test does not supply any support to the end user in evaluating the visual quality of the displayed information.

2.4 The proposed comparative user performance test methods

"ISO 9241 emphasizes the need to specify the factors affecting the performance of the users, and the need to adopt a user-performance approach to evaluate systems.", according to part 1, the general introduction of the standard. This approach is conveyed in a proposed comparative user performance test method, that is under consideration as an alternative route to compliance to ISO 9241-3. Since its publication a number of amendments have been made to this proposal [ISO00; TSM92]. All proposals consist of a comparison of the speed (and accuracy) with which a simple task is performed for a short period of time on the display under test and a benchmark display. In addition the subjects are asked to give some sort of visual comfort rating for the displays.

2.4.1 Performance measures

The relation between image quality and the outcome of this kind of performance of display tasks is not as straightforward as is it may seem. Performance measures are integral measures of the workload and are relatively insensitive to variations in image quality because many other factors codetermine the performance level. To produce quantifiable output any task in a performance test needs elements that are not directly related to image quality. These task elements require skills in motor behavior and cognitive information processing that are subject to learning effects during the task. At the same time the resulting performance tasks can differ so much from real world applications that their validity becomes questionable.

The practice effects where found by Travis, Stewart & Mackay [TSM92]for the proposed task in informative annex C of the standard. Their adjusted task was found to be suffering from the same effects [BS97a]. The latest proposal by Boschman & Travis [ISO00] has an even simpler performance task: pushing a key for every detected occurrence of a given character in a piece of text. To overcome the problem of initial learning effects the participants start with a training period of 10 screens of text, and the stimulus order will be counter-balanced. But a similar task, used in a validation of a simulator of flat panel displays, showed an increasing performance speed with counter-balanced stimulus order and an initial training period of 78 screens of text [SSW90].

Complicating factors are inter-individual and intra-individual variation in performance capacity. Variation in performance between individuals is often larger than variation caused by differences between displays. During the individual task performance there is a trade-off between accuracy and speed. The joint quality and quantity of task performance of an individual over the short test period will only vary with image quality, if either maximum effort is invested or the amount of effort invested in the task is somehow kept at a constant level. Therefore, both the quality and quantity of performance must be incorporated in a performance measure, and the amount of effort should either be measured or controlled for [Dru95].

Provisions against practice effects will extend the length of a performance measuring task. Drury & Forsman [DF96] observed that on a character search task subjects typically needed a practice time of 2 hours to reach a stable performance level. But even then performance measures will be relatively insensitive to differences in display quality and provide no information about the cause of these differences. Reviews of decades of reading research on typeset materials by Tinker [Tin63] show that most physical variables, when studied individually, have only a modest effect (10% or less), even when varied over a large range. Combination of several sub-optimal print conditions were found to reduce the reading rate by 20% or more. Gould et al [GAB+87] concluded the same for reading rates on VDTs.

2.4.2 Eye measurement

Fatigue, asthenopia, eye strain, visual stress, the effects of task performance on the visual system are measured in many studies. Eye movements and visual measures such as accommodation, vergence, and acuity are applied as measures of the load that is imposed on (parts of) the visual system. Blink rate, the amount of lacrimation, and subjective rating of the amount of visual discomfort (dry eyes, fatigue, headache...) are used as indicators of the stress on the visual system. Problems with the sensitivity and consistency of this type of measures have been reported [GK83; HB84; Nis90]. In a survey of literature Watten [Wat93] concluded that the progress in scientific explanations of the basic issues of fatigue of accomodation and vergence in the last century has been rather limited. It still has not been established that eye musculature can be fatigued. Just as performance measures, the visual measures are indirect measures of image quality, in the present case mediated by muscle or neural activity. In a recapitulation of their experimental work [BR97; RB97] in support of the latest user performance test proposal, Boschman and Roufs conclude that eye movement measurement seems to provide little more information than performance measures. Saccade lengths increase as fixation times decrease for better legible text, and all fixation times roughly add up to the total time needed for the task. Some measures require an intrusive measurement process and complex measurement instruments are often needed. And again, the cause of the possible performance difference cannot be specified.

2.4.3 Subjective methods

The subjective rating of the amount of visual discomfort was incorporated in the first version of the user performance test. These visual stress symptoms are subject to a large number of factors not related to visual display quality [Pic95]. This makes visual fatigue measures unsuited for a comparative user performance test.

The user himself could be asked to quantify the image quality, i.e. direct subjective measures can be used. Although subjective measures are widely used in studies into visual display quality, mostly they play the second violin. They are often used to provide additional support for other, 'objective' measurements. For the purpose of ergonomic evaluation, however, subjective measures are potentially very useful and valid measures. The literature confirms this [BPW89; BR97; LTT89; RB97; ShB95]. In many experiments where performance measures and ratings of visual display quality are used, the two methods provide similar results and often subjective ratings are more sensitive to image quality impairments [BR97; PSB83; Pas90; ShB95; Tin63; Wat93]. The objections that are commonly given against such measures seem to be based on an objectivity fallacy [MS92]. In measuring visual display quality, subjective measures are likely to be less biased then 'objective' measures. The quality of an image is inherently a subjective feature and some of the 'objective' requirements are probably derived from subjective measurements. (Unfortunately, ISO 9241-3 does not provide references to the sources of knowledge about the requirements.)

Subjective image quality measures can be obtained by a number of methods, such as rating, matching, paired comparisons and questionnaires. With all methods global, overall judgments can be obtained, but also judgments of specific quality aspects. Although every method has its specific advantages and drawbacks, a shared advantage is that the psychological evaluation of the sensations by the respondent is an integral part of the measurement process. Although individual variation may be inherent to subjective measurement, for visual perception the conformity in information processing is much larger than the variation. Thus, if response variability of observers can be controlled in some way, a reliable measurement process can be achieved. This is quite different from the situation in social-psychological research, where much knowledge about variability and errors in, for example, ratings has been generated. Here observers judge perceived attributes of other people or attitudes towards people or objects. These judgments are far more susceptible to subjective variation. For the measurement of display characteristics it is probably more accurate to make a comparison between subjective techniques and psycho-physical techniques, as was demonstrated by e.g. [BZ81].

2.4.4. The Display Evaluation Scale

In a number of experiments, e.g. [Spe89; SB94], in which we evaluated a flat panel display simulator, respectively investigated flat panel display characteristics, rating measurements showed to be more powerful than performance data.

- In most cases performance did not result in significant differences between conditions, while the ratings did. Some possible explanations for this have been given above.
- We consistently found a positive correlation between both data sets: the pattern of results was identical.
- Ratings of visual display quality aspects such as contrast, brightness and font size, closely agreed with the physical attributes, and exhibited meaningful and mostly simple (linear or log-linear) relations.

The theoretical considerations and the practical evidence and experience, as well as the relative ease of use, speak in favor of subjective techniques.

Visual perception pertains to phenomena in the temporal, the spatial, the spectral and the luminance domain. There are a number of different (perceivable) display properties that relate to the spatial, spectral, temporal and luminance characteristics of images, of which the perceptual correlates such as brightness, contrast and color are of specific interest. Display quality thus necessarily has a multi-dimensional nature.

The design of the scales was based on a mix of theoretical and practical considerations e.g. with respect to the number of response categories [HWH+02; Har87].

The DES has been applied in a large number of experiments. The inter-rater reliability [Gui54], repeated measures reliability and validity have been studied. Inter-rater reliability coefficients were determined from five different experiments [SBB93]. To give an indication: the mean correlation within subjects, between objects and collapsed over the DES items, was 0.809. In three experiments, in which different subjects participated, three monochrome flat panel displays were rated (an ELD, a reflective LCD and a back-lit LCD). The mean repeated measurements correlation of the DES items over these experiments was 0.848 [SBB93].

For the majority of individual DES items to be valid, the scores should correspond to another, credible source of knowledge about the rated objects. In a number of experiments the ratings were compared with physical measurements of the properties that the items were supposed to measure [Spe94], chaps 3 and 7 [SB92; SB94; SB95; SB96].

In the luminance domain, brightness scores highly correlated with the logarithm of display luminance [Spe94], chap. 3 [SB92], and could be predicted accurately with a linear regression equation including a term for luminance contrast. Also, the brightness appeared to be determined predominantly by the highest luminance (independent of display polarity) [SB92].

Over the range of contrasts that is normally found on displays, contrast ratings were found to be linearly related to the luminance modulation [Spe94], chap. 3 [SB92]. The spatial integration of luminances by the human eye was investigated in a study of gaps between display elements [SB94], in which we could discriminate between local and global physical measures of contrast on the basis of contrast ratings. Display luminance was found to contribute to perceived contrast [Spe94]. Both luminance and contrast have a subjective optimum, the exact value of which may vary due to other factors. The contrast item of the DES not only measures luminance contrast, but also color contrast [Lip86; SB96]. We obtained a very good linear fit between rated contrast and the effective contrast scale delta E [Lip86].

In the spatial domain, the text parameters character size, line spacing and line width were studied [Spe94] chap. 7. The optimal character height of about 20', which is generally agreed upon, was reproduced. The rated leading was shown to depend on both character height and line width and had an inverted u-shaped relation with these factors.

In an evaluation study of four different TFT displays and one CRT [BS97b] subjects performed a proof reading task in a word processing application or a search task in a CAD application on all screens. Each time after finishing the task on one of the displays, this display was rated on the DES. The mean rating scores provided a meaningful and useful characterization of the displays. The rated screen sizes maintained the order of the actual screen sizes; even the small difference of 0.7 cm between the CRT and one of the flat panels was reflected in the ratings. Further, as can be expected in a situation where large drawings are used, the CAD users in general would have preferred larger screens (the ones used ranged from 13.4 to 16.1 inch effective image diagonal).

In the experimental situation the diffuse reflection of the CRT was 5.4 cd/m2, as opposed to 0.8-1.7 cd/m2 for the flat panels, consequently, the mean CRT rating was 4.5 to 5.5 points lower than for the flat panels. The mean ratings of reflection also differed between polarity: In the negative polarity, (CAD application with dark background) the mean was 3.2 point below the mean rating in positive polarity (Word processing application with light background). The mean luminance uniformity scores correctly identified one positive and one negative outlier with a difference in score of about 2.5 and 4.5 point respectively. The difference between the remaining three displays was only 0.5.

The impact of the task on preferences was reflected in mean overall impression ratings. Whereas the word processors showed the highest preference for the flat panel display with the best rated legibility and character shape, the CAD-workers preferred the display that had received relatively good scores for most aspects, exhibited the largest viewing angle and the lowest reflection.

2.4.5 The latest proposal

The latest proposal for a visual performance and comfort test [ISO00] consists of a comparison between the display under test and a benchmark display for the search velocity in a letter search task and an overall subjective rating of visual comfort using a category scale. This proposal has met with a lot of criticism [ISO97a] mainly for the wrong reasons. The critics had not noticed the limited scope of the test and had a bias against any human involvement in the judgment of image quality. Nevertheless there are problems with the proposed test that limit its value as a normative instrument:

- The performance task is not ecologically valid, i.e. not a real task in a real environment.
- Performance measures are not sensitive enough.
- Only a low under limit is set to the quality of the benchmark display.
- The test provides no information about the nature of the shortcomings.

Display designers would be better served with a detailed rating of the visual display quality aspects from subjects working with a real world application, especially as the test is meant to evaluate novel display technologies. Exactly for this purpose we developed and tested the Display Evaluation Scale.

2.5 Discussion

The subjective rating of specific image quality aspects has been proven to be a powerful instrument in the assessment of visual display quality. In combination with a short task with a typical office application a slightly adapted version of the Display Evaluation Scale can be applied as a comparative user performance test in the current standard.

This solves all but one of the problems with the latest user performance test proposal mentioned above: The task is ecologically valid, the subjective measure is sensitive, and specific information about visual display quality aspects is provided. The remaining problem is the realization of the benchmark display that needs to be supplied or nominated by the supplier of the test display. The ideal benchmark display would just fulfill the physical requirements compliance test. In practice, benchmark displays on average will be considerably better, but individually, for the suppliers of test displays, they could be very differing in quality. So the benchmark display is not a standard display. In general a user performance test would benefit from additional requirements to the benchmark display like equal an approximately equal screen size and resolution to the test display.

In all, the user performance test within its current limited scope will not be of much use to the three user groups of the standard described above. To really improve the comfort and performance of people working with VDTs, more radical changes will have to be made to the standard. It will be difficult, if not impossible, to meet the needs of the three user groups with one standard. In the design stage, the actual application, application environment and users are unknown. Manufacturers will be inclined to market products that are capable of achieving a high level of visual quality for a broad range of users in a broad range of environments using unknown applications of different types. As the flat panel evaluation study [BS97b] indicated, text processing may put other demands on the display than CAD work.

In the purchase stage, knowledge of the display application situation will be available to some extent, but variability of users, software applications and environment will make it impossible to fix a design solution. It should be possible, however, to choose for a display that will not disappoint buyer and user once installed and in use. In the implementation phase, final decisions about the configuration, settings, workplace lay out and the control of light can be taken.

How can standardization aid these three user groups? The conclusion from the discussion, so far, is that there is enough reason to redesign the standard from scratch. We shall consider this from the point of view expressed in the first paragraph of the introduction. In particular, we shall address the topic of user performance testing as a means to re-ergonomize the standard.

Evidently, a designer needs information that can be applied in the engineering process, information that can be translated into design parameters. An overall performance or comfort score is too global to be of use. More specific subjective information, such as judgments of contrast or sharpness are useable if the designer can relate this information to display hardware parameters. The type of physical measurements that are part of the primary compliance route of the present standard are also useable, because they can be linked to beam spot sizes, bandwidths, reflectivity of layers, etc.

An engineers' standard could be an updated version of the current one, preferably formulated in a more technology independent fashion. Compliance criteria that can not be determined from the display alone will have to be written out of the standard (e.g. the criteria for the characters), they can be dealt with in a standard for the display buyer and display user.

Designers and consumers would also be helped with a form of gradation in visual display quality. One of the most important aspects of this gradation should be some kind of resolution or sharpness index, like the ISO number for photographic material or the number of dots per inch (dpi) for printers. At the moment there is no consistent relation between the dot pitch of a display of any technology and the perceived sharpness.

Because of the separation of display system and displayed information, a user performance test for the engineers' standard will be more difficult to define. But when suitable stimuli have been agreed upon, comparison with benchmark display(s) of a certain quality index can provide a better scaled quality assessment.

A display buyer could be helped with a method or tool, for a selection within the Not Bad class of 'approved displays'. Just as the (amateur) photographer can make his choice of film for sunny or dark conditions based on the ISO number specified, a display buyer should be able to select the appropriate display for the application and circumstances. If the decision makers are experts in the field of display ergonomics and are frequently involved in purchase decision making, theoretically there would be no problem to require that physical measurements are made (for which specialized instrumentation is needed). In practice the decision makers are often facility managers, system administrators, practicing ergonomists with limited knowledge of displays and human display interaction, and others that do not possess adequate knowledge or experience. System administrators for instance, will often be inclined to choose displays for their technical specifications only. Facility managers will put a great deal of emphasis on the cost aspect, creating a bad trade-off value for ergonomics. Many practicing ergonomists will evaluate displays very strictly according to the available standard. Generally they are not experts on displays, but on workplace adjustment and physical (bodily) factors. Indeed, a helpful standard is needed. It should, for the sake of standardization, at least prescribe the evaluation method and factors to be evaluated. Preferably it would also contain an evaluation instrument. This instrument should at least provide information on the basis of which displays can be compared. A global quality index is not sufficient, because displays with identical scores may be quite different. A subjective instrument that measures and expresses visual display quality in the (perceptual) dimensions that are experienced by users will combine detailed measurement and interpretability.

A typical user cannot be expected to install, configure, and fine tune a display. But who can? The adjustments cannot be made without reference to the operational situation, including the software application(s) and the nature of the visual tasks. System designers, hardware and software suppliers, company automation experts, users, ergonomists: each party has a partial knowledge of human display interaction. In order to achieve an optimal visual display quality, this knowledge has to be integrated. A standard can help by giving the end user an instrument with which positive and negative characteristics of the displayed images can be located. In order to assist in improving or optimizing the visual display quality, the user should be given practical guidance. An instrument like this can also be used for periodical quality assurance. In the operational environment, physical measurements are out of the question. A combination of subjective techniques (e.g. matching and rating) and simple measurements (e.g. screen diagonal) may be the answer.

A standard for the buyer and end user, can achieve a relatively high degree of completeness, because the operational circumstances will be known, at least to some extent. As we stated before, the relevant information for the buyer, apart from compliance with the engineers' standards, is if the display can produce images with a sufficient resolution, scan-rate and sharpness for the intended applications. So the users' standard should contain a method to calculate the needed resolution, sharpness and scan-rate (as long as high scan-rates are not a matter of course). Further, this standard should provide the criteria for the information that is displayed as well as offer help in configuring and optimizing the system. Attention will have to be given to hardware configuration (e.g. the relation with graphics cards), software (configuration and settings), workplace (where and how to put the display), (control of) lighting, and the user (e.g. viewing distance, optical corrections). Partly this can be done by referring to other standards, compliance to which here acts as a prerequisite.

2.6 Concluding

Three problems with the current part 3 of ISO 9241 have been discussed: the ergonomic quality, the user performance test, and the technological relevance (aging). Based on our analysis of the standard and its application we argued that the ergonomic quality of the standard can only be guaranteed by a) formulating a valid, reliable and useable method for testing the visual display quality; b) formulating a separate standard for engineers and display users; and c) a higher degree of independence of the standard from the display technology.

We realize that these three points will make the revision a major operation in which new problems will be encountered, such as the gearing to other standards that bear a relevance to display work situations. Although visual display quality is an intensely studied subject, we still lack a validated model of image quality. One reason for this, as we perceive it, is that most of the research in the area is conducted from different and often narrow perspectives.

2.6 Concluding

Due to the division of science we encounter work that ranges from technical studies that fully neglect the human to studies originating from the humanities that show a lack of understanding of the technical factors. From the point of view of ergonomics one could say that in most research there is an imbalance between understanding of the independent (technology) factors and the dependent (human) factors. Because engineers are the dominant party in display engineering and design, we are still in the situation that the technology on the desktop acts as the starting point for the design of the display work (situation).

As the visual display quality in practice continues to show, there is work to be done.

3 The image quality in context cycle

In this chapter a new model for a structured approach to image quality is presented. This framework is meant to integrate different fields of imaging and display technology and different applications of imaging. Image quality is modeled in the context of use: the user, the tasks and activities, the environment, the life cycle and standardization. Optimal image quality can only be achieved in optimal circumstances, in practice circumstances are rarely optimal. Modeling the context can bring the practical image quality closer to the optimal image quality.

Models build on the analysis of previous models and will be corrected and extended by future models. Building on the analysis of and experience with previous image quality models, each researcher has different approaches to image quality and visual display quality research and will zoom in on and refine the model in his or her own region of interest. Crucial to the usability of a model is the design rationale, which explains the structure and the contents, provides the scientific background and in future can facilitate a redesign process. The first step in the rationale should be a presentation of some results and considerations in image quality research and modeling.

3.1 Image quality models

The way a model of image quality takes shape highly depends on the definitions of "image" and "quality". In the most limited definition an image is a photo film or print of a natural scene and the quality is the beauty of the picture. Early research on image quality in the 1920s and 1930s [Sny88] uses this definition for the calculation and evaluation of photographic film image quality. As images are used more and more for all kinds of detection and recognition tasks, e.g. in aerial photography and medical imaging, the imaging application became a determining factor in the definition of quality. The need to adapt the image to the abilities of the human visual system is becoming apparent in the optimization of image quality. There has also been an enormous evolution in the way images can be displayed and processed. This means that image quality is no longer exclusively linked to densities and grain-size of chemical emulsions on celluloid or paper. The visual quality of electronic displays has become an important new field of research.

In their effort to link perception research and displayed image quality Roufs and Bouma [RB80] in 1980 observe the friction between the dedicated research effort needed for the understanding of human visual information processing fundamental for the proper understanding of image quality and the fast decisions required in display design, which are also based on many technical and economical considerations. They discuss a number of methods and the problems involved, but do not provide a structured approach or framework to come to a better understanding of image quality.

Snyder 1985 [Sny85] states that image quality has been used generally in two contexts: (1) that dealing with physical measures of the image itself and with little or no regard for the ability of the observer to obtain information from the image; and (2) that dealing with the perceived or measured quality from the human observer (i.e. behavioral measurements), sometimes with little regard for the physical characteristics of the image. From this observation he then draws the apparent conclusion that useful measures of image quality must contain both repeatable and design-relevant physical descriptors of the image as well as suitable behavioral measures by which to assess the validity of the metric. Snyder sees great possibilities for image-quality metrics based on the combination of the function describing the relative attenuation of the luminance modulation to the spatial frequency input of the system: the Modulation Transfer Function (MTF), and the function describing the contrast threshold of the human visual processing system for spatial frequencies: the Contrast Threshold Function (CTF).

As a vision researcher Klein (1993) [Kle93] is surprised at the use of complex images of natural scenes for image processing and judgment by the applied vision community and at the use of one-dimensional image-quality metrics. He argues, quite rightly, that image quality is multi-dimensional and simple stimuli more clearly show the consequences of image processing methods and make it easier to entangle the different image quality dimensions.

Spenkelink and Besuijen, 1997 [SB97] allege that image quality covers only the quality of the displayed image itself, while visual display quality also pertains to the consequences for visual information processing and the degree of adjustment to the task requirements in an operational environment. The measurement and specification of visual display quality should be user centered, but in order to generate applicable design knowledge the link with technology should not be lost.

In hindsight the distinction between image quality and visual display quality is mainly a matter of the dominating terminology in the research fields of photography and visual display ergonomics, as these fields of research become more and more overlapping this distinction disappears. Janssen and Blommaert [JB98] e.g. incorporate the use of the image within their description of image quality. They see image quality in terms of a compromise between partially conflicting demands maximizing either “naturalness”, the degree of correspondence to knowledge and reality, or “usefulness”, the internal signal-to-noise ratio.

3.1.1 The Image Quality Circle

Recently some important contributions to modeling image quality and the structured approach of the psychometric scaling of image quality factors have originated from the field of photography. To simplify the understanding of image quality Engeldrum, 1995,2004 [Eng95; Eng04b] has developed a step-by-step approach called the "Image Quality Circle" (IQC). This three-step approach (see fig. 3.1) first converts technology variables to physical image parameters through system models. In the next step visual algorithms define how the physical image parameters are related to customer perceptions. Finally, image quality models condense the customer perceptions to customer quality preference. The emphasis of the research effort lies on the psychometric scaling of image quality factors, in Engeldrum's terminology: "nesses" (e.g. sharpness and graininess), involved in the last step of the image quality circle. He defines image quality as "the integrated perception of the overall degree of excellence of an image".

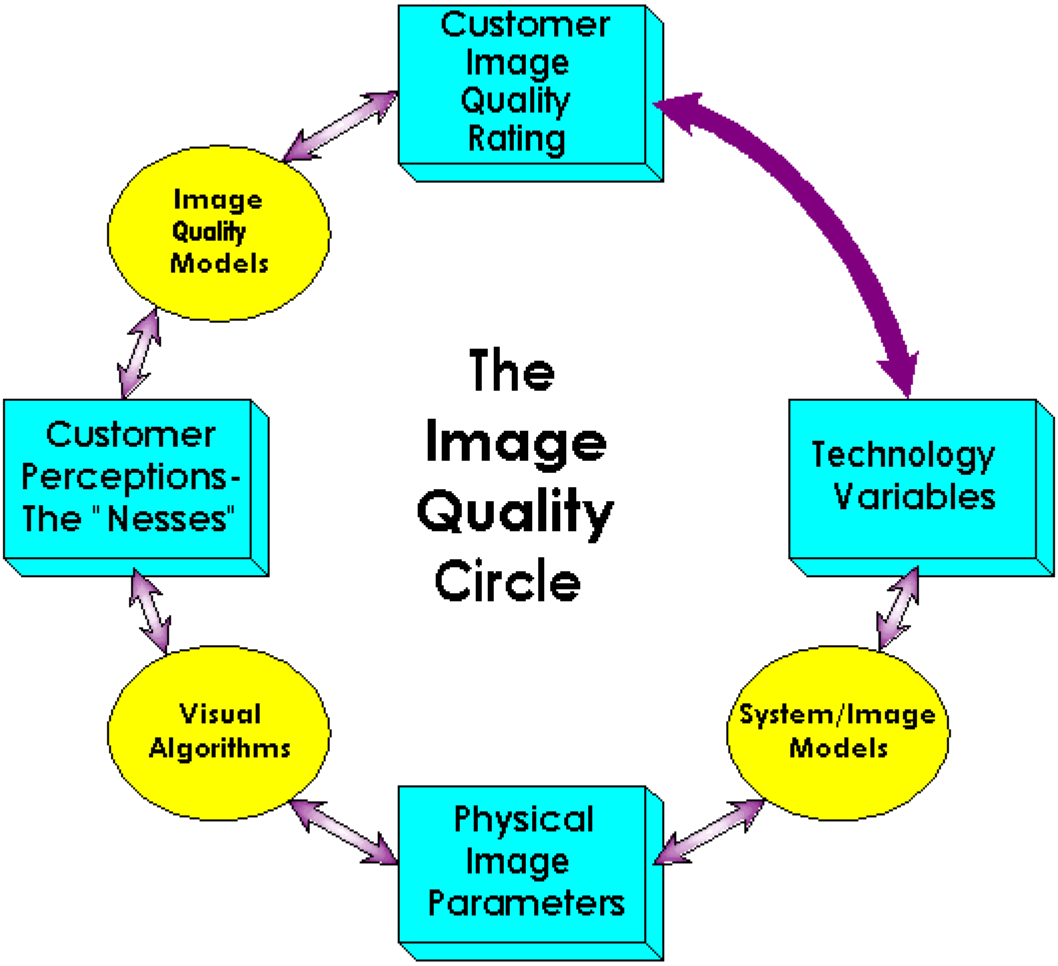


Figure 3.1. The Image Quality Circle by Engeldrum [Eng04b].

Although Engeldrum claims that his idea of image is very broad, the implementation of his image quality definition bears all the trademarks of a restriction to still photography. In conformance with the observation cited from Spenkelink and Besuijen above, Engeldrum limits image quality to the beauty aspects: “how good does the image look in the sense of a beauty contest” and has no intent to describe the utility, the application, the observers, the context or the component attributes. He rejects the reservations made by Janssen and Blommaert [JB98] and Keelan (2002) [Kee02] that certain aspects of image quality have different importance depending on the context in which the quality is judged. The view taken with the Image Quality Circle is that factors that affect the judgment and preference of image quality can be controlled or understood in any image quality judgment situation. The influence of the application on the image and subjective preference on the judgment of image quality is assessed by Engeldrum [Eng04b] (p. 450), to be easily accounted for by a partial compression or extension of the psychometric scale. This assessment does not seem to do justice to the complexity of image quality judgment and the possible application dependent change in relative importance of quality attributes to overall preference.

Engeldrum [Eng04b] also overestimates the role of the psychometric scaling step in his model: *"What can be particularly valuable and efficient about image quality models, in the Image Quality Circle framework, is that most of the model construction relies entirely on psychometric scaling studies. Thus, investments in infrastructure are minimal, compared to the considerable resources required for the measurement of Physical Image Parameters and*

Technology Variables. (p. 454)". Of course all three steps in the image quality circle are equally important if a complete model needs to be assessed. Psychometric scaling is only possible if the necessary customer perceptions can be produced. Therefore you need visual research to produce the necessary visual algorithms and you need the measurement of Physical Image Parameters and Technology Variables to produce the necessary system models.

3.1.2 Quality rating with a scaled reference

Keelan's [Kee02] definition of image quality is aimed at setting up a research program to develop a structured framework of psychometrically scaled image quality factors. He distinguishes four types of image quality attributes: personal, aesthetic, artifactual and preferential, and wants to develop comparable psychometric scales for the last two types of attributes. In this effort he wants to eliminate all unwanted sources of variation like personal involvement and aesthetical opinions from this experimental design: "*The quality of an image is defined to be an impression of its merit or excellence, as perceived by an observer neither associated with the act of photography, nor closely involved with the subject matter depicted.*"

The main instrument in Keelan's quality rating procedure is the image quality ruler. The hardcopy version consists of a series of photos depicting the same scene but linearly degraded in Just Noticeable Differences (JNDs) of unsharpness. The rating subject has to place the photo to be rated, which is subject to another kind of degradation, in a position where the quality equals that of a photo on the quality ruler. The softcopy version consists of a paired comparison on two electronic displays of the rated picture with sample pictures from the quality ruler combined with a binary search algorithm to optimize the number of comparisons.

The psychometric scaling research is mainly focused on consumer photography with film or digital camera, but the methods are also used to compare compression algorithms. Keelan and his team at Eastman Kodak have developed a coherent system of psychometric scales for evaluating image quality subject to all possible degradations in the photo acquisition process and optimized for the most common combinations of lighting and distances in consumer photography. The image quality evaluation system has been successfully used to design and select components for photo cameras.

Keelan's scope of image quality is still very limited in a world of rapidly expanding possibilities:

- Digital photo cameras are now far more important than photo film cameras.
- Most digital photo cameras also offer the possibility to make short video shootings.
- Most cameras have more options than an average user is able to remember and control.
- The effect of a change of configuration settings on the camera is hard to judge on the tiny viewer screens.
- Hardcopies are no longer exclusively made by professionals or automated professional equipment: they are printed at home or replaced by softcopies which can be placed on web sites or stored on CDs or DVDs and shown on PC displays, television screens or mobile phones.

3.2 The new image quality model

Here an ergonomic view on the assessment of image quality and the development of image quality models is taken. This leads to a comprehensive model, where the quality of an extensive range of images is accommodated. The model is not limited to photographic images, but it can also include moving images, 3D-images, medical images, synthetic images, and images of text and graphics on all possible kind of image carriers: displays of all technologies, hardcopies and projections. Because images for all kinds of technologies and applications are digitally recorded, the boundaries between different kinds of images fade away and it is more and more desirable to come to one comprehensive image quality model. The inclusion of the quality of display of the third dimension in the concept of image quality is not straightforward. Several studies have shown that the rating of 3D-quality is not covered by the term 'image quality' [SMIJ05; STMV00].

The image quality model is also extended ergonomically. It does not only comprise the quality of a separate image at a certain moment in the design stage, but also explicitly treats the quality of the image at the time of use, the context of the image quality: the quality in relation to other images (standardization) and the conservation of this quality (configuration, calibration, maintenance and repair).

Although Keelan’s work [Kee02] might seem limited to the link between technology variables and psychometric scaling, in his Advanced Photo System model (p451) the context of use is represented with eleven factors that are primarily related to the user interaction with the photo camera, with two factors that are related to the environment and with one factor about the content of the image.

Standardization is one of the main problem points indicated by Hunt [Hun05] in his evaluation of imaging performance of different display technologies: *“Properties where improvements are needed include greater consistency between pictures displayed from the same signals on different devices, greater color gamut, less impairment by ambient illumination, and better resolution.”*

In the newly proposed model image quality is assessed as a multidimensional entity, with separate scales and criteria in the different quality dimensions for separate applications. (The dimensions of image quality and a classification scheme for image quality factors are treated in chapter 4.) Engeldrum’s assumption that psychometric scales can be fitted from one application to another by compression or extension will probably hold for most applications, but the occurrence of strong interactions is expected to be significantly higher with more complex images (esp. moving and 3D) than the rare cases in Keelan’s experience [Kee02] (p252). Roberts [Rob02] e.g. reports that for certain imaging systems a higher degree of sharpness of moving objects can lead to an aggravation of movement artifacts.

The image quality model should be modular in structure with the accompanying advantages of reusability, ease of maintenance and reduced chances of errors. It allows the international standardizing organizations to adapt the existing standards for different applications (medical, office, telecommunication, broadcasting, etc.), technologies (photo, flat panels, CRT) and platforms (Windows PC, Apple, Silicon Graphics) to each other and manage them. It should also be easier for the producers of imaging and display devices and applications, because several standards could be benchmarked in one test.

In figure 3.2 the "Image Quality in Context" cycle is outlined. The cycle has six stages: Technology Variables, Physical Parameters, Image, Visual Responses, Cognitive Perception, and User Evaluation, and six models defining the conversion from one stage to the next:

Physical, Light Distribution, Visual Response, Feature Extraction, Psychometric Scaling, and Adjustment model.

On the right side is the image: a 2D or even 3D light distribution, on the left side is the user evaluation of the image quality. The image is determined by technology variables, which can be translated into physical parameters and these determine what the image looks like on print or display. The image that is projected on the human eye also depends on the environment and the viewing position. The projected image triggers visual responses in the human visual system. These responses lead to a cognitive perception of the contents of the image and the quality of the image. The user evaluation then is the result of a comparison of the user expectations of the image quality and the perceived image quality. The outcome of the user evaluation could be a reason to adjust the technology parameters of the image forming system.

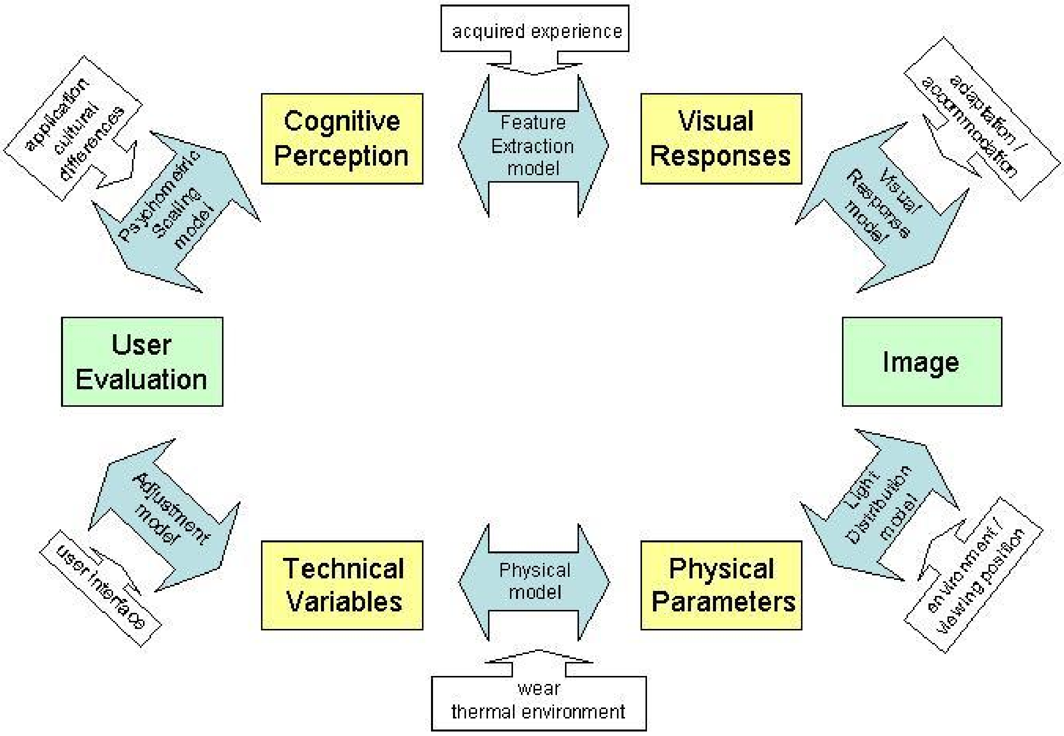


Figure 3.2. The cycle of image quality evaluation in context.

The context is represented by external factors acting upon each model: wear or aging and the thermal environment on the display or hardcopy, the environment and the viewing position on the light distribution, adaptation and accommodation on the visual responses, the experience on the feature extraction, the application on the psychometric scaling and the user interface on the adjustment of the technical parameters. Presenting the external factors separately from the models they act upon simplifies these models by separating these often complex parameters, and at the same time stresses their role in the image quality maintenance process.

All steps in the cycle should be carefully modeled; in the next part a closer look is taken at all the steps in the cycle of models.

Technology Variables

Technology is not constrained to the actual display system: the image acquisition system (camera, scanner) as well as compression and transmission variables determine the way an image looks. The main purpose of image quality rating is the optimization of technology variable settings. This optimization is important in the design cycle of the image forming system as well as in the life cycle. In the design phase care will be taken to minimize the amount of variation in all other stages of the cycle to provide ideal circumstances for optimizing the technology variable settings. In the life cycle of the system the variation in the other stages is expected to be considerably larger: a wide range of applications, cultural differences between users, sub-optimal viewing conditions and configuration settings will result in a lower average image quality with larger variation.

On a home computer system e.g. it is very likely that images come from different sources all having their own characteristics: they can be acquired by webcams, scanners, digital photo cameras, mobile phones, or television tuners, composed with some kind of drawing program, or downloaded from an unknown source via the internet.

This implies that image quality is also a matter of standardization of source formats. Therefore color profiles, which store information about variables relevant for color management purposes attached to the image data are essential to image quality in the life cycle of a system.

Wear and thermal environment

A number of properties of display systems and image carriers is not static, but changes over time. Light sources e.g. are generally not constant in luminous output. The output depends on the thermal environment and there usually is a warming up time before the maximum output is reached, and the light output over the lifetime of the display system usually decreases. This could become a serious problem if the light output of all components of the system does not decrease at the same rate: as a consequence, non-uniformities or color shifts could develop.

Physical model

The conversion from technology variables to physical parameters is determined by the physical model. This model differs from Engeldrum's System/Image model in that it does not comprise the whole system: the interface for adaptation of the technical parameters is excluded. Furthermore the physical model does not completely define the image: the effects of environment and viewing position also effect the distribution of light.

The physical model defines the limits of the capacities of a specific display or imaging technology, and is determined by physical measurements. Methods for measurement need to be standardized to allow a fair comparison between systems and technologies. At the moment at least four standards exist for Flat-Panel-Display measurement [Dow06], with different methods of measurement, e.g. for luminance uniformity. And the recent discussions in the display society over the capacities of Liquid Crystal Displays versus Plasma Display Panels show that parties disagree over the way measurements are made and presented.

It is recommended that researchers in image quality perform their own measurements in order to obtain reliable results: Individual products are often outperformed by their product specifications. Generic physical models for a certain technology are not always up to date with developments in imaging and display systems and measurement technology, as is illustrated for the CRT Tone Reproduction Curve model in chapters 5, 6 and 7.

Physical parameters

The physical parameters should describe the spatial and temporal distribution of the luminance and chromaticity in the image. The spatial domain includes the range of viewing angles with respect to possible artifacts, 3D-viewing or multiple image channels. The description should be free of any relation to the specific technology used to display the image.

Environment / Viewing Position

The viewing of the image can also be disturbed by classical ergonomic environmental factors like vibrations, blinding glare from specular reflections or light sources close to the display, or a diminished contrast caused by diffuse reflection. The viewing position determines what the user actually sees: the retinal size of the image, the contrast (esp. LCD), the depth (3D-displays) and the reflections.

Light Distribution Model

Although the physical model completely describes the output of visible radiation of the display system, it does not completely describe the image presented to the user / observer. The actual image presented also depends on the position of the viewer in relation to the displayed image and the ambient illumination, which can interfere with the image. The viewing angle has always been an important image quality issue in LCDs: off-normal viewing angles could lead to color shifts and even contrast inversions. Enormous progress has been made to resolve this issue, but a lot of LCDs currently in use still exhibit this impairment (see chapter 7), and a closer comparison of LCDs demonstrated in retail shops today will probably reveal that this issue is not yet completely settled.

Recently displays have been developed, which give completely different pictures depending on the viewing point. These displays could be used e.g. in cars to display the car navigation system to the driver and at same time provide entertainment for the passengers. For the display of 3D-images the difference in viewing position of the left and right eye is used to show different images and create the illusion of depth.

Image

The image is neither the image on the surface of the display screen or hardcopy, nor the image on the retina of the viewer. The image in this model is somewhere in between: at the precise moment that the image enters the optical system of the human eye, including optical aids like glasses or contact lenses. It does include the effects of the environment and viewing position on appearance of the displayed image, but it does not include the effects that the state of the eyes (adaptation, accommodation, vergence) has on the reception of the image on the retina.

Adaptation / Accommodation / Vergence / Individual differences

The eye is adapted to a certain level of luminance and accommodated to see sharp at a certain viewing distance. The adaptation level of luminance should correspond with the level of luminance of the image and the accommodated viewing distance should correspond with the distance between viewer and display. For 3D displays the discrepancy between the distance at which the eyes are converged and the distance at which they are accommodated can cause fusing or focusing problems [WDK93].

Differences in visual responses between individuals can also play a role in the image quality evaluation process. If the application needs the image to convey information to all possible users, the visual characteristics of special groups, such as color deficient and aging people become important in the evaluation cycle.

The human visual system has many special cases in the handling of stimuli: all kinds of interactions can occur depending on the size, the duration, the frequency, the chromaticity, the luminance level, the surroundings and the history of the stimuli. One of these special cases had an adverse effect on the performance in the visual gamma matching experiment with multiple brightness matches described in chapter 7. The prolonged fixation of the eye gaze on an image with saturated colors lead to chromatic adaptation. The sensitivity for strongly saturated colors at the locations where they are projected on the retina is decreased. Subsequent stimuli of the same color at a lower luminance seem washed out, and a subsequent look at a white field could show an afterimage in opposite colors.

Visual Response Model

This model could also be named the Standard Observer and describes how the image as a visual stimulus affects the state of the visual cortex in the brain given the current state of the eye with respect to adaptation and accommodation. The Commission International d’Eclairage (CIE) has drawn up Standard Observers for human vision of luminance and chromaticity and has constructed several 3D-spaces CIE-xyY, CIELAB and CIELUV for the representation of color stimuli[CIE06]. Currently vision researchers [WA05] are working on the development of a Spatial Standard Observer (SSO) based on the collected contrast thresholds for a set of 43 standard stimuli from a total of 16 human observers.

As yet there seem to have been no recorded attempts to set up a Temporal Standard Observer.

Visual Responses

A description of visual responses is difficult because of the endless number of possible stimuli. The types of stimuli and responses are often classified in three or four categories: spectral or color (luminance and chromaticity), spatial and temporal. For image quality purposes the description of just noticeable differences between responses is even more important than describing the exact response to a certain stimulus. For the color dimension the JNDs between simultaneously displayed colors are represented reasonably well in the CIELAB and CIELUV system, and color researchers are still suggesting new color spaces to improve the congruence between JNDs and the distance in color space.

Visual responses to spatial stimuli are usually described by the Contrast Sensitivity Function (CSF) or the Contrast Threshold Function (CTF), which specify the sensitivity of the human visual system for spatial frequencies at different orientations. Integrating formulas are used to derive the difference in JNDs. A major problem with the CSF and the SSO [WA05] is that they do not account for hyperacuity, a ten- to twenty-fold increase in spatial sensitivity for certain stimuli. Hyperacuity means e.g. that the human visual system is more sensitive to slight horizontal shifts in the middle of a vertical straight line, than to a vertical grating with the same spatial frequency. Artifacts caused by hyperacuity regularly occur with line drawings and texts on rectangular matrix displays.

The research in visual responses to temporal stimuli until recently was concentrated on the Critical Fusion Frequency (CFF), the frequency were flicker (CRT or film projection) becomes invisible. The rise of the LCD technology has shifted the focus of research on temporal visual responses to motion related artifacts.

Acquired experience / cultural differences

Experience plays an important role in the ability to recognize patterns. The symbols of the Roman alphabet are easily recognized by anyone who learned reading with texts written in Roman script, but the same person will have considerable difficulties in recognizing individual characters from Arabic, Chinese, Hebrew or Indian scripts.

It is less clear to what extent experience can influence image quality feature extraction and assessment. Brindsley [Bri70] distinguishes the Class A and Class B observations. A Class A observation simply asserts if the sensation produced by stimulus α and stimulus β under certain conditions is the same or not the same. Any other observation is a Class B observation, these "include all those in which the subject must describe the quality or intensity of his sensations, or abstract from two sensations some aspect in which they are alike." Experience should not play a role in Class A observations, but most observations regarding image quality are Class B observations where the influence of experience can not be excluded. Slight differences between stimuli might easily degrade class A observations to class B observations. In the experiment reported in chapter 7 slight temporal artifacts impeded an equal brightness match, which caused biased gamma matching results for the untrained user, while the expert was able to filter out these artifacts and provide more accurate results.

Another problem can arise from the incongruency of concepts for image quality features between languages and cultures. Most people from western cultures would have trouble defining the Japanese image quality aspect of "mura". This does not necessarily mean that they would have trouble rating the general image quality of displays with mura. JNDs of sharpness, one of the most important image quality features, appear to be uniform over different cultures: Liu et al. [LXHY04] found no difference between the assessment of sharpness by Chinese and European people for Chinese and Roman character sets.

Feature Extraction Model

This model describes the lower level processes with which quality features are extracted from the image. These processes are comparable to the way alphanumerical information is automatically recognized by a trained reader. For images with considerable differences in one quality aspect, immediate image quality judgments can be made, barely involving any cognitive processes, based on the outcome of the feature extraction. If the quality differences between images are in the order of one or two JNDs, cognitive intervention is needed to guide attention to parts of the image were the differences are most visible. If the images differ in more than one quality aspect in opposite ways, cognitive efforts will be needed to weigh the impairments.

Application

The application is an important context factor influencing the psychometric scaling. Image quality factors essential to one application can be non-significant for another. If color is the essential factor in applications where colors have to match, e.g. in color correction of photography or video shootings, or for large multi-panel displays, color JNDs will have more weight on the rating scale than JNDs for other image quality factors. If color is only used to enhance data from non-visible sources like satellite or medical imaging by coding with a limited set of colors on a single display, the significant factor is not so much what the precise colors are, as long as the difference between the colors remains large enough.

For the application 'text editing' different image quality aspects are important than for the application 'television viewing'. The first needs a stable high-resolution image for a short viewing distance; the latter needs a medium resolution fast switching image for a medium range viewing distance.

Psychometric Scaling Model

Psychophysics aims to relate discriminable differences in sensation to continuous physical properties. The term psychometrics is used if the stimuli vary along more than one objectively measurable physical dimension, i.e. when comparative judgments have to take place. Paired comparisons, rating and ranking experiments can be used to construct psychometric scales, but it is difficult to compare the results from different sources unless the scales are calibrated to some common standard. Keelan [Kee02] and his group at Eastman Kodak have accomplished a tremendous effort in developing a comprehensive set of psychometric scaling methods and results for a variety of image quality factors. The real test of their model and methods however lies in the (at least partial) reproduction of their results by other research groups and the extension to images with more complex quality attributes.

User Evaluation

A distinction has to be made here between the user evaluation as a result of an image quality rating experiment or as result of an image quality model evaluation of the image, and the individual evaluation of a user evaluating the quality of the system he or she is working on. In the first case the evaluation is often part of a series of evaluations of different systems or controlled system settings, and the user is asked to give an opinion on specific set of image quality aspects defined and presented in the experiment. This in contrast to the case of individual quality evaluation, where a single system is evaluated by a user, who wants to optimize the system settings based on a personal set of image quality aspects. Evaluations of image quality in practical work situations will also reveal more problems related to contextual aspects, e.g. the viewing environment, which are usually controlled in experimental settings.

User Interface

The user interface should provide a conversion from the adjustment of technological variables preferably to JNDs in a meaningful visual-perception dimension. For hardware controls this is usually not the case, which can lead to considerable problems for the average user to improve the quality of the image with help of these controls. Examples of these problems for the brightness and contrast controls of a CRT and an LCD are shown in chapter 6.

Adjustment Model

This is the most complex model because it should comprehend all other models. It describes how the technological variables can be adjusted with user controls in hardware or software that operate only in the desired perceptual dimension with an adjustment rate that is linearly proportional to a perceptual scale in JNDs. This is complex because there usually is no one-on-one relationship between perceptual dimensions and technological variables. The adjustment of the perceptual color dimensions hue, saturation and value (lightness) (HSV) is a clear example. Originally user interfaces allowed only the adjustment of the gain of the red, green and blue (RGB) primary colors. If only the lightness of a color needed to be adjusted, the gain of all primary colors had to be changed. Later different user interfaces provided controls to directly adapt the separate perceptual HSV color dimensions.

Of course it is impossible to convert the adjustment model in an optimal user interface that will encompass all contextual effects operating upon the image quality apprehension. It should be possible though to design a user interface that is robust to the effects to be expected in the specific context of use. In many cases this would involve the presentation of different levels of user interfaces for different types of users.

Another problem can be that different users make use of the same system and have different preferences regarding image quality.

3.3 Conclusion

A model is presented that describes image quality in the context of use. At first sight this model may appear fairly complex, as it consists of six sub-models with numerous external factors that influence these sub-models. In the practice of use many of these external factors will play no substantial role and not all of the sub-models will have to be worked out in detail. In many cases assuming standard models will suffice to provide acceptable results. It is however always useful to know the broader context of a problem, making deliberate decisions in the design process of image quality experiments and the development of the separate models. Not making shortcuts out of ignorance or lack of information, but because the effects of external factors are known to be insignificant in the given circumstances.

This model is not complete and might give disproportionate attention to some of the aspects involved in image quality assessment. It should not be seen as a rule of law, but as an instrument, part of the process of image quality assessment in practice and a stage in the development of better methods to comprehend image quality research.

4 A classification framework for visual display or image quality

Abstract

A new classification framework is proposed for visual display quality factors. It enables the multidimensional assessment of image quality for a variety of applications at a technology independent level. It can be used as design for standards and checklist for image quality experiments.

4.1 Introduction

The great variety in display technologies and the growing number of applications for electronic displays brings about an accumulation of display quality standards. Cathode Ray Tubes (CRTs), Flat Panel Displays (FPDs), and projection displays have their own figures of merit (FOMs) while medical imaging, office work, aviation, and television each have specific requirements for displays. At the same time the traditional tie between an application and its specific display is loosened; because of the growing number of connections between systems, images are available on the nearest display and Visual Display Terminal (VDT) work can be done anywhere. Most standards are based on a series of technology specific threshold requirements, that tell more about the state of the particular technology than about the level of visual comfort to be expected viewing the display (Chapter 2, [BS98]). Setting a specific standard for each technology and each application would lead to numerous tests on the production side and incomparable test results on the buyers' side.

This calls for a universal display standard that comprises all display technologies and allows the comparison of figures of merit. Application standards can be based on this display standard by defining application specific performance levels for display quality factors and adding requirements for the image content (characters, medical images, video, etc.).

For this purpose the quality factors of the displayed image need to be identified free of any technological denotations or connotations. Several proposals have been made for a taxonomy of the visual quality space. Shurtleff [Shu80], summarizing his research in the sixties and seventies, proposed a classification restricted to the then main technology of the CRT and the application of data display. Light, geometric, display surface, optical, temporal, electronic, and symbol generation factors were classified with regard to their importance to symbol identification and the technical knowledge available on the subject. Another summary of CRT figures of merit was made by Task [Tas79] distinguishing geometric, electronic and photometric factors. Based on this summary Rash et al [RKHM00] made an extensive list of FOMs for CRT and FPD, classified in four domains: spatial, spectral (chromaticity), luminance and temporal (see table 4.1). It provides a mix of technology specific FOMs, like: spot size and shape, focus, and bandwidth for CRT, and pixel size, number of defective

Table 4.1. Common CRT and FPD Figures Of Merit Classified by Domain (according to Rash et al [RKHM00])

Spatial	Spectral	Luminance	Temporal
CRT			
Viewing distance Resolution Spot size and shape Modulation transfer function Luminance uniformity Signal/noise ratio Display size Aspect ratio Number of scan lines Interlace ratio Scan line spacing Linearity Focus	Spectral distribution Color gamut Color purity Chromaticity	Peak luminance Luminance range Gray shades Contrast ratio Halation Ambient illuminance Gamma Dynamic range	Frame rate Field rate Bandwidth
FPD			
Pixel resolution (H x V) Pixel size Pixel shape Pixel pitch Subpixel configuration Number of defective (sub)pixels	Spectral distribution Color gamut Chromaticity	Peak luminance Luminance range Gray levels Contrast (ratio) Uniformity Viewing angle Reflectance ratio Halation	Refresh rate Update rate Pixel rise/fall times

(sub)pixels, viewing angle and pixel rise/fall times for FPD, and common human vision and viewing environment aspects, like spectral distribution, color gamut, luminance range and halation. Many of the common aspects: viewing distance, modulation transfer function, display size, color purity, reflectance ratio and ambient illuminance are linked oddly enough to only one of the two technologies. Glasser [Gla97], whose FOMs are more exemplary, combines the luminance and chromaticity domain and adds two external domains: viewing direction and lighting. He also makes a distinction between fidelity, the quality of the transfer of the input source, strongly linked to the space and time domain, and the legibility, the quality of the output image in the eyes of the viewer, strongly linked to the viewing direction and lighting domain.

Both Glasser and Rash classify their FOMs directly in one of the domains, and this poses problems as these FOMs are seldom confined to one domain only. This is illustrated by the different classification of the quality factor ‘contrast’: in the space domain by Glasser and in the luminance domain by Rash. Assigning FOMs to domains needs careful consideration. Take flicker for example: the visibility of flicker primarily depends on temporal and luminance factors, without a variation of luminance in time it would not be flicker. Some spatial extent is needed and a certain chromaticity value (including (0, 0) or neutral) is implied for the flickering stimulus to be perceived and the visibility of flicker is affected by the size and very slightly by the color of the stimulus, but the spatial and chromaticity domains are secondary to the temporal and luminance domains. The unclear definitions of FOMs are another obstacle for straightforward classification, e.g.: should it be called flicker if there is a chromaticity component, but no luminance component, or should it be color flicker? And what if there is a modulation in the size or position of the object? Any visual display quality classification scheme will need a stricter definition of the FOMs.

In his short image quality model taxonomy Engeldrum [Eng04a] distinguishes Detection/Recognition models and “Beauty Contest” models with or without Reference Image. He also introduces the term “Ness”-based image quality for what has been called subjective image quality in contrast to “Physically” based image quality for what is also known as objective image quality. Objective image quality is indeed a paradoxical term. As the saying goes: “Quality is in the eye of the beholder”, which is certainly true for image quality. Objective image quality can therefore only exist as the greatest common factor of a number of individual subjective image qualities. The term “Physically based” is not really an improvement: ‘objective’ image quality is usually computed as a function of the difference between two image signals, e.g. the signal before and after compression and decompression. These signals are detached from the physics of image acquisition and display; in fact the image does not need to be visible to compute its ‘objective’ image quality. The term image signal quality would be more appropriate. Because the ultimate destination of images usually is the human eye, the degradation of image signal quality should also be measured in perceptual attributes with psychometric scales. In his quest for a clear taxonomy it looks like Engeldrum [Eng04a] has neglected the importance to image quality research of measures combining physical and perceptual attributes, e.g. the Modulation Transfer Function (MTF) of the image and the Contrast Sensitivity Function (CSF) of the human visual system.

The denomination “Ness-based” is derived from the principal photographic image quality factors sharpness, noisiness / graininess, brightness, lightness and colorfulness. Engeldrum [Eng04b] uses “ness” as “a shorthand notation to mean some perceptual attribute, to emphasize the connection to human perception, and to distinguish a Customer Perception from a Physical Parameter.” This leads to rather artificial terms like contrast-ness, hue-ness, defect-ness and color accuracy-ness. On the one hand, terms like usefulness and naturalness are discarded for being functions of other perceptual dimensions, and on the other hand sharpness and blurriness are set as examples together. The term "Ness-based" appears to have no added value and can therefore better be replaced by perceptual.

Keelan [Kee02] distinguishes four types of image quality attributes: personal, aesthetic, artifactual and preferential. He wants to develop comparable psychometric scales for the last two types of attributes. Artifactual factors like sharpness, oversharpening, streaking, and contouring can only degrade an image, they have a rather sharp optimum on one side of the scale and every deviation from this optimum beyond the visibility threshold means degradation. Preferential factors like brightness and contrast have a broader optimum with degradation on both sides. This distinction between factors was also made in the work on the Display Evaluation Scale (DES), a multidimensional quality-scaling instrument for text displays, by Spenkelink et al [SBB93], who also found that the brightness rating distribution had a broad peak.

The question may be asked if this kind of distinction is not arbitrary: the artifactual factors sharpness and oversharpening could be combined in one preferential factor, and the preferential factor brightness might be split in two artifactual factors: dimness and blinding. There is e.g. a difference in sharpness between images from film and video camera, and some professionals make an effort to adjust the settings of their video cameras to obtain the softer look of film on HDTV [Rob02]. Not all broadcasters share this preference for "The Film Look" as can be seen by comparing coverage of international sporting events on channels from different European countries, esp. UK and Germany.

The aim here is to develop a framework for image quality factors at a level that is independent of the technology used to display the image and of the condition of the environment and the visual system of the viewer. The framework should provide a clearer procedure for classification of the image quality factors.

4.2 Dimensionality

Perceptual image quality factors should preferably be unidimensional. A set of unidimensional orthogonal image quality measures would provide the optimal solution for describing the perceptual image quality space. A unidimensional factor also has the advantage that shortcomings in image quality can easily be traced back to the physical parameters and technological variables of the image forming system. Sharpness and contrast as natural language terms are not unidimensional: unsharpness in moving images can be caused by simple spatial blur, motion blur, color misconvergence, flicker, and a lack of contrast; in fact it is not easy to find an image quality factor not affecting sharpness. Contrast is often confused with sharpness and brightness. If image quality is rated based on a name or definition of the quality factor, then ambiguity will always be a source of variance. The use of paired comparison techniques is a way to avert this problem. The most elegant solution is Keelan's [Kee02] image quality ruler, which links the comparison result to a scale of reference images calibrated in JNDs of unsharpness (caused by simple spatial blur).

The multidimensionality of sharpness might be the blessing of this method: because nearly all quality factors affect sharpness, it might be easier to compare them on a scale of unsharpness. It is probably not so easy to quantify quality factors on the image quality ruler that have very little in common with sharpness like color fidelity or non-linear update in picture frames (lost frames). For preference to colorfulness this is confirmed by Keelan [Kee02] p.299, who found a pattern deviating from most other image quality factors with two groups of observers: a sensitive group with smaller JNDs and a preference for less colorful images and a group of less sensitive observers with larger JNDs and a preference for more colorful images. Mapping both groups consistently on one JND calibrated quality ruler is impossible.

Engeldrum [Eng04b] recommends Multi-Dimensional Scaling techniques (MDS) to extract orthogonal image quality factor dimensions from experimental data. In our experience with the Display Evaluation Scale for text displays MDS-results depend very much on the experimental design: the dimensions found with MDS and the weight of these dimensions depend strongly on the physical parameters that are varied and the range over which they are varied. Reliable and reproducible MDS results need large amounts of experimental data for relevant variations of physical parameters.

4.3 Internal and external domains

In agreement with Glasser’s division in fidelity and legibility [Gla97] the visual quality factor space is here divided in internal and external domains. The image quality or fidelity is determined by the five internal domains: temporal, spatial, luminance, chromaticity and depth (3D). The external domains are involved in the viewing of the image: the surface of the display or hardcopy, the ambient illumination and the position of the viewer relative to the image; they define the viewing quality. The term 'legibility' should not be coupled to these external domains, because the size and sharpness of the pictures or characters, i.e. the content of the image, also affects the legibility. The internal and external domains are closely related and cannot be completely separated: size, luminance level, white point and natural depth are properties of the image that are mainly relevant in relation to the external properties of illuminance and viewing position. And a (small) aspect of viewing position is always present if the image is viewed with both eyes.

Table 4.2. Visual quality factors for images on displays or hardcopy classified by quality aspect and physical domain(s). Cells within rows with the same content, accentuated by the background color, belong together. Dark gray cells are non-existent, light gray cells not active in the row.

<table>
<tr><td colspan="2"></td><td colspan="2">value</td><td colspan="3">coordinate</td></tr>
<tr><td colspan="2">domain</td><td>chromaticity</td><td>luminance</td><td>spatial</td><td>temporal</td><td>3D</td></tr>
<tr><td colspan="2">quality aspects</td><td>extra (2D)</td><td>basic</td><td>basic (2D)</td><td>extra</td><td>extra</td></tr>
<tr><td>1</td><td>level</td><td>white point</td><td>lum. level</td><td>size</td><td></td><td>depth level</td></tr>
<tr><td>2</td><td>range</td><td>color gamut</td><td>dynamic range</td><td>angular size</td><td></td><td>depth range</td></tr>
<tr><td>3</td><td>addressability</td><td>chromaticity addressability</td><td>gray scale addressability</td><td>addressability</td><td>update addressability</td><td>depth addressability</td></tr>
<tr><td>4</td><td>equidistancy</td><td>non-uniform sampling</td><td>non-linearity</td><td>distortion</td><td>non-linear update</td><td>distortion</td></tr>
<tr><td rowspan="4">5</td><td rowspan="4">smearing</td><td>color blur</td><td>temporal blur</td><td>color blur</td><td>temporal blur</td><td></td></tr>
<tr><td>temporal color blur</td><td colspan="2">blur</td><td>temporal color blur</td><td></td></tr>
<tr><td></td><td colspan="4">temporal crossstalk</td></tr>
<tr><td></td><td colspan="2">crosstalk</td><td></td><td>crosstalk</td></tr>
<tr><td rowspan="2">6</td><td rowspan="2">discontinuity</td><td>color flicker</td><td colspan="2">raster separation</td><td>color flicker</td><td></td></tr>
<tr><td>color separation</td><td>flicker</td><td>color separation</td><td>flicker</td><td></td></tr>
<tr><td rowspan="3">7</td><td rowspan="3">instability</td><td colspan="3">color non-uniformity</td><td></td><td></td></tr>
<tr><td></td><td colspan="3">jitter, ripple</td><td></td></tr>
<tr><td colspan="2">color drift</td><td></td><td>color drift</td><td></td></tr>
<tr><td>8</td><td>noise</td><td colspan="5">noise</td></tr>
</table>

The internal/external division makes it possible to also incorporate standards for image quality or image fidelity used in telecommunications and image processing, the image signal quality, in the new framework. In these fields quality is determined as the amount of (perceptual) degradation from the original that is not necessarily perfect itself. Here is an important parallel with visual display quality standards, where the requirements are also set to levels, which appear to be more adapted to the technological state of the art than to the desired visual comfort (See chapter 2). In both cases there is no room or drive for technical improvement. As put forward by Klein [Kle93], the only approach to image quality determination that is immune to future technical improvements is the worst-case approach: comparison with a perfect image. For the moment at least perfect displays or images are an illusion, but if a perceptual distance from perfection could be determined it would be useful as a quality grade.

For the VESA Flat Panel Display Measurement standard (FPDM) a distinction is made between temporal and motion artifacts [Mis04]. For temporal artifacts changes of a display's characteristics are assessed with the stimulus at a fixed position over time. For motion artifacts the display's characteristics as a function of position change over time are evaluated. This is no ground to add a motion domain to the framework. As Teunissen et al. [TZLH06] have shown, good predictions of the consequences for image quality of motion artifacts can be made with fixed point temporal measurements combined with conversions from the temporal to the spatial domain and models of visual motion perception. Such a measurement system is easy to build and can be used for a broad field of motion artifacts.

4.4 The five internal domains

A closer look at the intrinsic properties of the domains and the current FOMs and a more fundamental approach to classification can produce a more solid framework for a universal image and display quality standard, and a guide for visual display quality research. In agreement with the position of the “Physical Parameters” and "Image" blocks in the image quality in context cycle model of chapter 3 the domains should be placed somewhere between the physical and the perceived world, just before it enters the human visual system including optical appliances. Between "Physical Parameters" and "Image" the description of the image is not fundamentally changed, only the influence of external factors is added.

The internal domains should provide a suitable framework to describe the image signal quality in terms that are also relevant for perception. So a chromaticity domain is more appropriate than a spectral domain, because different spectral distributions can produce the same perceived color. The domains should not be affected too much by the present momentary condition of the human eye, so luminance is more appropriate than brightness.

Based on their properties there are several ways to divide the five domains into groups (see table 4.2). The luminance and spatial domains can be seen as the basic domains, these domains are necessary to display any image. The temporal, chromaticity and depth (3D) domain are extra: they are not needed to display a still monochrome image. This is an important reason to keep these domains apart.

A division can also be made between the domains that determine the value of the stimulus, luminance and chromaticity, and the domains that determine where and when it is shown: the temporal, spatial and 3D are coordinate domains. The number of dimensions is another point of difference between the domains: the temporal and luminance domain have one dimension, the chromaticity domain has two and the spatial domain has two (or three). At the introduction of this framework in 2002 [BS02] the inclusion of quality aspects of 3D-images was deemed overambitious. Since then research and development in 3D-displays has thrived and public and commercial interest for this next frontier in displaying reality have been enormous. In this process the specific problems of 3D-image acquisition and display technologies have become apparent and laid down in overviews of 3D-image quality factors [MIJS04; WDK93]. This larger image quality scope can always be cropped to the domains of interest in the type of images under evaluation. The third dimension of the image puts extra constraints on the external domain of the viewing position, where small changes can have far reaching consequences for the perceived image.

For this framework the correspondence between the domains is far more interesting than the difference. The most important similarity is that they are all domains, which are sampled, i.e. they are mainly digital array mappings from real domains. These mappings have limitations, which are the same for all domains: limits in size and step size, linearity and stability. If we take a closer look at the relation the figures of merit have to the domains, common properties can be derived. Ideally these properties should be independent of a particular display technology or the state of the human eye and could be seen as intrinsic quality aspects of the representation of the specific domain(s) scaled to human perception. E.g. if we look at contrast and color gamut, they both represent a range within their domains, resp. luminance and chromaticity, that can be displayed. With the range comes a midpoint or level and the number of steps into which the range can be divided.

If we work this out for ‘all’ FOMs and domains eight quality aspects can be distinguished: the first four pertaining to single domains and the last four to multiple domains (table 4.2). The columns in table 4.2 represent the five internal domains of the displayed images. The basic luminance and spatial (2D) domains are in the middle; the value domains are on the left, the coordinate domains on the right. The rows give eight quality aspects that indicate how well the displayed image is represented for this aspect in the specific domain(s).

4.5 The eight quality aspects

The naming of these aspects poses some problems; the first four can be described as positive aspects or qualities, for the others negative aspects or impairments are chosen, because most quality factors (cells) are usually identified as impairments. The first two: level and range, are closely related. Together they determine if the image is suited for the viewing conditions (ambient illuminance, reflections, viewing position) and the image content (luminance and chromaticity of the scene, natural depth). The level gives an indication how suitable the image is for given conditions and the range denotes, which values and coordinates are available within the perceptual range for the specific domains.

The quality aspect addressability was named resolution in the previous version. Resolution is related to what human perception can dissolve in the picture. Addressability merely indicates how many values (gray levels) or coordinates (pixels) can be addressed by the system. The visibility of the separate addressed values and coordinates depends on the perceptual distance between these values and coordinates, which is determined by the relation with range and the non-linearity, and by the amount of smearing and impairment by other artifacts in the image. Resolution in one domain can also be relinquished to improve the resolution in another domain, e.g. dithering improves the gray scale resolution at the cost of a loss in spatial resolution.

Equidistancy is the accuracy with which the differences in value or coordinates in the domains are mapped onto the JND scales of perception. This includes non-linearity but also non-uniformity of the interval: e.g. a projection on a slanted screen is linear but is non-uniform with respect to the aspect ratio of the picture elements.

Smearing denotes the way in which the value that belongs to one coordinate in the original is distributed over that coordinate and the neighboring coordinates in the image.

Discontinuity refers to the visibility of inactive image parts between samples.

Instability is the non-random variation of the value of a part of the image that should have constant value for different coordinates in time or space.

Noise is the random variation of the value of picture elements.

4.6 The image quality factor cells

The cells in table 4.2 can be used to define the interface between image on display and hardcopy. Technology variables and physical parameters can be linked to visual response parameters to find the optimal match for environment, experience and application. A short review of the variables, parameters, applications and problems connected with the different cells is presented in Appendix A.

The naming of the quality factors in the cells is problematic and bound to produce some ambiguities. Commonly used terms such as ‘sharpness’ and ‘resolution’ have such a wide range of meanings and physical parameters involved that an attempt to restrict them to one cell or cell group in this framework would be infeasible. Some visual quality concepts can easily be mapped to cells in the table, such as: white point, color gamut, and geometric distortion. In some cases more concepts map to one cell (jitter, ripple), while in other cases, esp. in the temporal domain, there are no traditional image quality FOMs to fill the cells.

Newer projection technologies, such as sequential color projection, and applications like streaming video demand requirements for quality in these areas.

In the overview presented in Appendix A is chosen to name the cells with the neutral combination of column and row names. For ease of use the cells in the table can also be coded by a combination of letters for the domains(s) and the number in the left most column of the table. The domains involved in the cells are denoted by the first letters: L for Luminance, S for Spatial, C for Chromaticity, T for Temporal and D for Depth (3D). The kind of quality factor is denoted by the number of the row: 1 for level, 2 for range, 3 for addressability, 4 for equidistancy, 5 for smearing, 6 for discontinuity, 7 for stability and 8 for noise or the rest.

Traditional FOMs that can be linked to specific cells are listed as associated terms. The list is not exhaustive for the last four rows but presents the combinations that are represented by artifacts commonly encountered and described in literature.

4.7 Validity

The three main criteria for the validity of this classification scheme are completeness, orthogonality and applicability. For the moment the model looks complete for the internal domains: the known artifacts can be categorized in the framework. The viewing quality determined by the three external domains, viewing direction, lighting and display surface properties, mentioned above, are not included. These domains are always present when viewing an image, and in many circumstances they are of predominant importance to the image quality. Regrettably the classification of the external image quality factors is not straightforward.

Lighting can cause disturbance by reflections on the image and blinding by glare from light sources or strong reflections around the image. For emissive and transmissive displays reflections are not directly involved in the image formation and could be viewed as disturbing superimposed images and should therefore have additional quality factors. For reflective and transflective (transmissive and reflective) images the reflections are part of the image formation process and determine quality factors for the internal luminance and chromaticity domains. The effect of lighting on the image is strongly affected by the viewing direction and the image surface properties. A change of viewing direction can obscure glaring light sources or introduce new specular reflections on the image. The description of image quality factors concerned with lighting and reflections with all the degrees of freedom involved deserves a separate framework.

Aspects of viewing direction that act on quality aspects represented in internal domains of the framework like brightness, contrast, color and depth should be dealt with in the application of the framework.

The independence from display technology and the structure of the framework with the division in domains and abstract quality aspects provides an excellent base for orthogonality. Technical measurements can be defined that measure each separate quality factor. In human vision all kinds of trade offs, e.g. between spatial and gray-scale resolution, and interactions, e.g. between brightness and flicker perception, exist between quality factors. But impairments of individual quality factors should be identifiable if the quality is low enough.

The elaborate list of image quality factors in Appendix A shows that a large number of image quality aspects can be represented in the cells of the classification framework. The compound perceptual quality factors such as sharpness, resolution, and colorfulness are not represented in the framework. To assess the effect of the characteristics of the image quality aspects on the perceived quality a layer is needed that combines the relevant cells of the framework and characteristics of visual perception. This first layer is described in the Image Quality in Context Cycle (chapter 3) as the Visual Response model, which is about equivalent to the Visual Algorithms step in Engeldrum's Image Quality Circle [Eng04b].

This framework is meant to be a useful guideline for image quality research. The definition of the cells can be used to develop a set of simple stimuli as recommended by Klein [Kle93] to test the effect of image acquisition and display parameters on image quality. First the effects for individual cells can be worked out. Fortunately an enormous amount of recognized research results on visual responses concerning stimuli related to the cells in the framework already exists. It is not always easy though to find the results that exactly match the specific conditions needed to fit into this framework, so extra vision research will be needed, especially for the further stages where interactions between two or more image quality characteristics are examined. The new framework is a concrete step towards a more complete and orthogonal classification scheme that provides a more structured approach to the disentangling of image quality factors.

4.8 Discussion

The framework presented provides a structure for the definition of image quality and visual display quality independent of display or system technology. The main problem in developing the framework to a standard may lie in the definition of the more complex quality aspect smearing and the inclusion of the effect of viewing direction on certain quality aspects.

The framework should provide guidance in display quality research. A great deal of the documented experiments is concerned with comparisons using complex images between samples of display technologies or variations of technological parameters with effects on multiple quality factors. Understanding the effect that single quality factors and their interactions have on the overall quality can provide fundamental knowledge for camera and display designers in all technologies.

5 The shifting gamma perception[♦]

Abstract

True to life, color display and color management depend on a proper technical model of the display used. Current gamma models and fitting procedures are not accurate in modeling the lower part of the tone reproduction curve. The GOG- and GOGO-model used in color management standards tend to clip the luminance to zero for digital values were luminance can be seen and measured. Two improvements to the models are suggested. First, the models should be fitted by optimizing the root mean square error (RMSE) of the CIE lightness instead of the luminance.

Second, a shifting gamma model is adopted, with gamma increasing in value for lower voltages. Results show that the adapted models correspond better with the luminance measurements. The clipping values are nearer to the measured zero luminance threshold, and the average RMSE and ΔE^*_{ab} over the whole scale are smaller.

5.1 Introduction

Electronic displays are rapidly becoming part of nearly all human-machine interfaces. A substantial amount of visual inspection tasks, which used to be done on paper, on photo, or in real life, are now done on electronic displays. Some of these tasks set high demands to the reliability of the pictures shown on the displays. The visible impression caused by the pictures should be consistent for the viewers over different systems, with different graphics cards, different displays, and even different display technologies. Exchange of graphical designs, (remote) medical diagnostics, preparation and projection of presentations, and online shopping are examples of applications where visible display quality is a critical criterion. Various display standards and color management systems have been developed to preserve the correct representation of pictures. Many of these use gamma, the exponent of the power function describing the relation between input voltage and light output of a cathode ray tube (CRT) display, as the main parameter in the display's tone reproduction curve (TRC). The merits of gamma are not common knowledge: ***"Owing to poor understanding of TRC, and to misconceptions about nonlinear coding, gamma has acquired a terrible reputation in computer graphics and image processing,"*** Poynton stated in one of his persistent efforts to explain gamma and clear its reputation [Poy98]. The nonlinearity of the CRT is not a defect,

---

[♦] This chapter first appeared as article: Besuijen, J. (2005). "The shifting gamma perception", Color Research and Application **30**(5): 332-340. After publication a printing error in one of the formulas was discovered and a programming error in the CIE Lightness RMSE fitting method. The correction of the programming error resulted in considerably lower values and lower reduction for CIE Lightness RMSE, especially for the red and blue primary. The corrected formula and tables were published in: Besuijen, J. (2006), "Errata: The shifting gamma perception", Color Research and Application **31**(5): 438-439. This chapter contains the corrected tables and formula.

but a highly desirable feature, because it is very nearly the inverse of the lightness sensitivity of human vision. The nonlinearity causes a CRT's response to be roughly perceptually uniform.

In a video camera, digital still camera, or scanner, a nonlinear transfer function—gamma correction or inverse gamma coding, similar to the CIE L^* function of color science—is imposed. This provides an optimal perceptual performance with a limited number of bits throughout recording, storage, processing, compression, transmission, decompression, and presentation [Poy98].

The use of gamma has several other merits:

1. It characterizes the nonlinearity of the TRC with a single meaningful parameter, unlike the parameters of cubic splines or polynomial fittings.
2. It relates to the technical characteristics of the main display technology CRT.

But with the advances in measurement methods and technologies, the number and relevance of the disadvantages that have come to light are growing:

1. The single gamma model for CRT technology is disputed. Olson [Ols95] showed that gamma could be much higher (9.5) for low voltages than for high voltages (1.5).
2. The CRT has lost its supremacy in the market, and the liquid crystal display, the projected main technology for at least the next decennium, has a dissimilar TRC. [GF00; Gla97] And so have the main other contending display technologies PDP and OLED. The non-CRT display technologies therefore need a lookup table correction to display the images encoded for display on CRTs, which is the actual image encoding standard.
3. Several platforms, Macintosh and SGI, use a default partial gamma correction of 1/1.4 and 1/1.7, respectively, resulting in display system gammas, that is, the gamma from the stored digital image to the image on display, of about 1.8 and 1.47. For an extensive review of the consequences see Poynton [Poy98].
4. Recent models for gamma fitting (GOG, GOGO) are no longer single parameter models but have two or three parameters [BMG93; KDB01b].
5. The gamma figure is not an absolute standard; the sRGB and IEC standard [IEC98] has a nominal gamma of 2.2 (single parameter) for the characterization of the standard display, but uses a power of 2.4 in its formulas (with an extra gain/offset parameter).
6. The meaning of the value of the gamma is still confusing to the average user. It is unclear in which cases gamma correction is needed, and the term gamma is often used for inverse gamma as well.
7. Other curve fitting techniques like cubic splines give better fitting results.
8. The human brightness perception curve differs substantially from the gamma curve, especially for lower light levels.
9. Different standard curves based on human brightness perception are already developed (e.g. the Grayscale Standard Display Function for medical applications) [DIC03].

These in turn raise questions whether gamma is still the best way to characterize a display and whether users are able to handle the gamma concept to control the brightness and contrast reproduction on their displays.

5.2 The simple gamma model

The basic gamma model relates the electron beam current I of a CRT gun to the input voltage E with a simple power function with exponent γ:

$$I = E^{g} \tag{5.1}$$

The relation between the beam current and the phosphor luminance is nearly linear, [KDB01b; Ols95] leading to the one-parameter gamma function for the relation between digital code value ***d*** (or gray level) and the display luminance *Y*:

$$Y = \left(E_{\max} \frac{d}{d_{\max}} \right)^{g} = Y_{\max} \left(\frac{d}{d_{\max}} \right)^{g} \qquad (5.2)$$

if the digital to analog conversion form (digital code value d to voltage) is linear. The simplest solution to determine the exponent γ is by fitting a straight line through the measurement points in the log-luminance–log-voltage domain [VES98]. Roberts [Rob93] and Berns [BMG93] showed that small offsets in voltage or luminance result in curves at the ends of the straight line that have a severe impact on gamma found with this fitting method.

A positive voltage offset results in a luminance offset, a luminance for digital code value 0 and a low estimate for gamma in the log–log domain. A negative voltage offset results for low gray levels in too low voltages for the CRT to produce electrons and therefore luminance, with the result that all digital code values below the threshold are mapped to luminance 0. If this offset is not compensated the gamma fitted will be too high. In practice optimal black level adjustment is nearly impossible and small offsets will remain. Even small offsets can have considerable impact on the fitted gamma [BMG93; Rob93].

Not all luminance offsets result from voltage offsets; they can also be caused by internal flare or veiling glare also called external flare (see below). Berns [BK97] also argued that differences for unreliable low luminance measurements are overrated in the log–log domain.

5.2.1 The Dark Side of the Scale

Correctly measuring the low light levels linked with the lower digital levels on the graphics card has proven to be a problem for much research on display characterization. In color measurement with spectroradiometer, the incoming light is divided into many spectral bands reducing the sensitivity for low light levels considerably. In colorimeters, where the light is divided over three or more colored filters on as many sensors, the light level per sensor is substantially higher, but deviations in the spectral transmission curves can introduce considerable errors, especially for narrow band primaries, like the red primary of most CRTs. If luminance is measured with a photometer, measurements for the green primary will generally be accurate and sensitive, but the blue and red primaries fall on the sides of the spectral photometric band and their light is considerably attenuated by the photometric filter. The original objective of the research project was to develop a fast low cost display characterization maintenance method based on a combination of the merits of a contact photometer and a spectroradiometer. The spectroradiometer, a Photo Research PR-650, would be used to measure color coordinates and peak luminances. The contact photometer, a Macam L203, would be used to measure primary luminance profiles with a precision of 0.001 cd/m^2. The reliable spectroradiometer peak luminance measurements should be used to correct the photometer measurements, especially for the red and blue primary, with a scale factor.

5.3 The Gain-Offset-Gamma model (GOG)

Together with accurate luminance measurements for low light levels, the inclusion of the voltage offset in the gamma model should provide a better display characterization.

Berns [BMG93] developed a model for the amount of spectral radiant exitance $M_{\lambda,R}$ generated by a CRT gun, which incorporates an offset voltage and a gain parameter. The model derivation is limited to the red channel denoted by the subscript **R**.

$$M_{\lambda,R} = k_{\lambda,R}\left\{a_R\left[(v_{\max} - v_{\min})\left(\frac{d}{d_{\max}}\right) + v_{\min}\right] + b_R - v_{C,R}\right\}^{\gamma_R} \tag{5.3}$$

The amount of spectral radiant exitance of a computer-controlled display depends on the digital counts in the digital to analog converter (DAC) (d, d_{max}), the video generator (v_{min}, v_{max}), the video amplifier (a_R, b_R), the CRT ($v_{C,R}$, γ_R), and the properties of the faceplate and phosphor materials ($k_{\lambda,R}$). Note that a_R and b_R are supposed to correspond with the "contrast" and "brightness" settings of the monitor, and that γ_R is assumed constant under all conditions. By normalization the number of parameters can be reduced, leaving a normalized system gain term $k_{g,R}$ and offset term $k_{0,R}$, and producing a normalized tristimulus output:

$$\begin{aligned} R &= \left(k_{g,R}\frac{d}{d_{\max}} + k_{0,R}\right)^{\gamma_R}, && \left(k_{g,R}\frac{d}{d_{\max}} + k_{0,R}\right) \geq 0 \\ R &= 0, && \left(k_{g,R}\frac{d}{d_{\max}} + k_{0,R}\right) < 0 \end{aligned} \tag{5.4}$$

$$k_{g,R} = \frac{a_R(v_{\max} - v_{\min})}{a_R v_{\max} + b_R - v_{C,R}} \tag{5.5}$$

$$k_{0,R} = \frac{a_R v_{\min} + b_R - v_{C,R}}{a_R v_{\max} + b_R - v_{C,R}} \qquad k_{g,R} = 1 - k_{0,R} \tag{5.6}$$

Notice that the gain is dependent on the monitor offset setting ("brightness"), and the offset is dependent on the monitor gain setting ("contrast"), which could explain the problems faced by the users to correctly set their display configuration.

This GOG model was expanded by Katoh [Kat02] to the GOGO-model with a luminance offset parameter $Y_{0,R}$ for extra luminance caused by flare or glare.

$$Y_R = (Y_{\max,R} - Y_{0,R})\left(k_{g,R}\frac{d}{d_{\max}} + 1 - k_{g,R}\right)^{\gamma_R} + Y_{0,R} \tag{5.7}$$

5.3.1 Veiling Glare and Flare

Veiling glare and internal flare can disturb the light measurements. Veiling glare or external flare is caused by ambient illumination reflecting in the front glass plate of the display. When spot measurements are used, external flare can be removed by darkening the room or it can be neutralized by subtraction, but only if the source is constant and at the expense of an extra error source. With contact measurements, covering the immediate surround of the measuring probe on the glass plate will suffice to eliminate external flare. Internal flare is harder to counter, but because the sources are internal they should be included in the monitor model. Internal flare can have three sources as follows [BK97]:

1. Reflections in the glass face plate from light generated elsewhere on the display surface, also called neighboring pixel interreflections.

2. Reflections from light generated by electrons deflected from their original destiny, also called “secondary emissions.”
3. Black level luminance caused by one of the other primaries, also called cross-channel emissions or unwanted emissions [DK98].

If the background around the measurement stimulus is (nearly) black the internally reflected light will have the same color for a primary (R, G, or B) measurement stimulus. The amount of reflected light does depend on the size of the stimulus: the highest contribution will come from inside and just outside the measurement field of the light meter. In practice, it is impossible to match exactly the measurement stimulus to the measurement field of the instrument used. The risk of a stimulus being too small is far greater than the accuracy gained by an exact match.

The secondary emissions could add light from other primaries to the stimulus, but if these emissions are linearly proportional to the amount of electrons reaching the shadow mask, the effect on the measured color of a primary should be the same for all input levels. As the flare from these sources is presumably small and proportional to the electron current of the gun under measurement, for the purpose of display characterization, these forms of internal flare should be regarded as an inseparable part of the model.

Neither of the above sources will have any effect on the black level luminance ((R, G, B) = (0, 0, 0)). The luminance offset at black level is obviously the same for all the three primaries. And as most instruments are not able to measure color at these low luminances, it will be difficult if not impossible to determine the source. Most monitors do not support external black level (“brightness”) adjustment for separate primaries. Just subtracting the measured luminance at (0, 0, 0) from all measurements is not the solution, because the offset could be partly resulting from the primary under measurement, that is, if $k_{0,R}$ is positive. If displays are used in brightly illuminated places, a higher black level setting might be desired to be able to distinguish dark colors.

In the GOGO-model, Y_{max} is a fixed value that can be measured and the luminance for the input $d = d_{max}$. Y_0 is a model parameter, and for a characterization with separate primaries it typically does not equal the luminance for $d = 0$, because part of the luminance is caused by the voltage offset (if $k_g < 1$) of the target primary.

The model parameters are typically estimated by measuring the luminance for 9 (0, 32, 64, . . . , 224, 255) or 17 equi-stepped code values[*] (e.g., 0, 16, 32, . . . , 240, 255) and performing a nonlinear optimization on the mean square error (MSE) in the luminance domain. The resulting display characterization has an adequate color reproduction with average errors reported of about $\Delta E^*_{ab} = 0.4$ [BMG93] and $\Delta E^*_{ab} = 0.8$ [BK97] depending on the color set used for testing.

5.3.2 Measurements

The R, G, and B primary luminance profiles of a 21” Eizo Flexscan 780i-W delta gun and several Iiyama 17” Trinitron CRTs were measured with the MACAM L203 photometer. Ambient illumination was excluded by a dark grey foam cover. The linearity of the

---

[*] The last steps, 31 and 15, in both these code value sequences are of course smaller than the others, 32 and 16. A more elegant solution for an 8-bit system, suggested by Charles Poynton, would be to use 18 really equi-stepped code values with all steps equal to 15. However, this is not a general solution as it cannot be expanded to 10- and 12-bit display systems. For the measurements in this article, 17 steps starting with the smaller step: 0, 15, 31, . . . , 239, 255, were used. This way more bits are involved in the measurements, thereby minimizing the influence of calibration errors in individual bit conversions of the DAC.

Table 5.1. Results for different model fits to the measured data of the R, G, and B primaries of the delta gun CRT.

	γ	Range	k_g	L_0	Lum RMSE	CIEL RMSE	Clip value
Delta gun R method							
GOGO Lum	1.938		1.151	0.036	0.0181	1.802	33.45
GOGO CIEL	2.091		1.095	0.011	0.0243	0.891	22.12
Glid Lum	2.491	0.693	0.978	0.000	0.0072	0.711	-5.74
Glid CIEL	2.293	0.404	1.031	0.000	0.0099	0.358	7.67
Delta gun G method							
GOGO Lum	2.015		1.140	0.052	0.0074	2.111	31.32
GOGO CIEL	2.119		1.104	0.014	0.0131	0.860	24.02
Glid Lum	2.240	0.276	1.066	0.000	0.0021	0.348	15.79
Glid CIEL	2.205	0.228	1.077	0.005	0.0024	0.208	18.23
Delta gun B method							
GOGO Lum	2.088		1.157	0.007	0.0084	0.353	34.60
GOGO CIEL	2.147		1.136	0.005	0.0102	0.272	30.53
Glid Lum	2.378	0.345	1.065	0.001	0.0031	0.125	15.56
Glid CIEL	2.312	0.259	1.084	0.002	0.0036	0.097	19.76

Table 5.2. Results for different model fits to the measured data of the R, G, and B primaries of the CRT measured by Berns[Berns 1996].

	γ	Range	k_g	L_0	Lum RMSE	CIEL RMSE	Clip value
Delta gun R method							
GOGO Lum	1.618		1.097	0.000	0.0140	1.240	22.55
GOGO CIEL	1.648		1.091	0.002	0.0155	0.664	21.27
Glid Lum	1.671	0.116	1.080	0.002	0.0117	0.457	18.89
Glid CIEL	1.688	0.148	1.073	0.002	0.0118	0.432	17.35
Delta gun G method							
GOGO Lum	1.546		1.136	0.005	0.0120	3.872	30.53
GOGO CIEL	1.652		1.094	0.002	0.0229	1.562	21.91
Glid Lum	1.688	0.239	1.075	0.002	0.0032	0.307	17.79
Glid CIEL	1.697	0.260	1.072	0.002	0.0034	0.267	17.13
Delta gun B method							
GOGO Lum	1.539		1.099	0.005	0.0123	0.563	22.97
GOGO CIEL	1.583		1.082	0.004	0.0153	0.446	19.33
Glid Lum	1.675	0.231	1.042	0.002	0.0043	0.132	10.28
Glid CIEL	1.673	0.226	1.043	0.002	0.0043	0.131	10.51

Table 5.3. Results for different model fits to the measured data of the R, G, and B primaries of the Trinitron CRT.

	γ	Range	k_g	L_0	Lum RMSE	CIEL RMSE	Clip value
Delta gun R method							
GOGO Lum	2.161		1.094	0.022	0.0054	0.511	21.91
GOGO CIEL	2.218		1.074	0.008	0.0074	0.279	17.57
Glid Lum	2.344	0.207	1.037	0.000	0.0020	0.178	9.10
Glid CIEL	2.294	0.138	1.051	0.002	0.0027	0.115	12.37
Delta gun G method							
GOGO Lum	2.034		1.104	0.000	0.0074	2.208	24.02
GOGO CIEL	2.125		1.076	0.000	0.0131	0.906	18.02
Glid Lum	2.220	0.246	1.047	0.000	0.0018	0.310	11.45
Glid CIEL	2.201	0.223	1.052	0.000	0.0019	0.223	12.60
Delta gun B method							
GOGO Lum	2.078		1.071	0.000	0.0080	0.514	16.90
GOGO CIEL	2.140		1.052	0.000	0.0107	0.365	12.60
Glid Lum	2.293	0.274	1.006	0.000	0.0015	0.078	1.52
Glid CIEL	2.270	0.240	1.013	0.000	0.0018	0.059	3.27

photometer was checked by comparisons with measurements through neutral density filters: no trend was visible. The displays had at least 1 hour warm-up time. The time between stimulus change and luminance measurement was empirically optimized to 5 seconds.

5.3.3 Again the Dark Side

In using the GOG(O)-model for display characterization, the results are not completely satisfying: the model seems to work well for the high end of the digital code scale, but has serious defects on the low end. Even if luminance and luminance change can be perceived and measured in the lowest tenth of the digital code scale for a particular CRT gun, the fitted model tends to map these densities to zero luminance. Especially for displays were the black level is set correctly, with minimal luminance at digital code 0 and measurable luminance increase near digital code 10, the fitted solution could map about one tenth of the digital code scale to zero.

Measurements on a delta-gun CRT display with these characteristics were used to produce the model fitting results in Table 5.1: the rows marked “GOGO lum” show the results for the GOGO-model, with clipping thresholds above 30. For this display and graphics card, this would mean a reduction of the amount of possible colors from 16.7 million to 11 million. This is not an isolated case; results for other displays showed the same discrepancy between observed luminance and fitted clipping values. The effect can also be found in the data reported by Berns [Ber96] where luminance was measured for low digital counts (<10), but clipping values, d_c, of 24, 29, and 18 for R, G, and B are fitted by the GOGO model (see Table 5.2). This could lead to considerable color differences, especially if one or more of the primary densities is near threshold. These densities are rarely used in tests. Berns tested his display characterization either on color sets with factorial designs with digital counts of 0, 559, 755, 903, and 1023 (10 bits) [BMG93] or with digital counts of 0, 85, and 255 (8 bits) [DK98].

The same pattern can be observed for the Trinitron display tested (Table 5.3): the black level setting is nearly optimal, but nearly one tenth of the values is mapped to zero.

5.3.4 Minimizing the Error in Perceptual Space

An obvious cause for the failure of the model fitting could be that the space in which the error is minimized does not represent the perceptual reality. For luminance, only the relative weight of the spectral components is defined according to perception, but the luminance scale is not linearly proportional to brightness perception. The log–log domain might overrate the importance of small luminance values; the luminance domain underrates it. The small luminance values are at least equally important in modeling the gamma function, and maybe even more because the maximum digital value with zero luminance determines the size of the voltage offset. In the GOG(O)-model an error of 0.1 cd/m^2 is equally important for luminances of 100 cd/m^2 and 0.1 cd/m^2. If the available measurement equipment does not provide reliable low luminance measurements, then the measurements should not be used in the model-fitting procedure or if possible the reliability should be improved by averaging repeated measurements.

Minimization of the error in the CIE lightness domain would be preferable as the eventual goodness of fit is determined by color differences ΔE_{uv} or ΔE_{ab} in the CIELUV or CIELAB space, where lightness is one of the three dimensions. Another option is the Digital Imaging and Communications in Medicine (DICOM) Grayscale Standard Display Function (GSDF)

[DIC03] based on a cubic spline fit of a vision model developed by Barten [Bar92] (see Appendix B). This vision model was made to fit data of just-noticeable differences in luminance modulation. The GSDF does not pass through the origin of the grayscale-luminance plane, which is not a helpful feature for an error function in an optimization procedure. Moreover the ten eight-digit constants in this standard, and the eight/nine polynomial terms of log luminance in the inverse formula make this function too complex for an iterative procedure.

The CIE lightness L^* is a relative measure of the luminance of a color Y to the luminance of a reference white Y_n. For a display the obvious choice for Y_n is the luminance of white on the display.

$$L^* = 116\left(\frac{Y}{Y_n}\right)^{1/3} - 16 \qquad \text{if } Y/Y_n > 0.008856 \text{ and}$$

$$L^* = 903.3\left(\frac{Y}{Y_n}\right) \qquad \text{otherwise.} \tag{5.8}$$

As shown in the rows marked “GOGO CIEL” in Tables 5.1-5.3, an optimization in the lightness domain lowers the clipping threshold, whereas the decrease in lightness RMSE outweighs the increase in luminance RMSE. Because the voltage offset is lowered, the gain is decreased and gamma is increased. The fit for the luminance offset is more compatible over the three primaries.

5.3.5 Correcting the Physical Model

So the fitting in the lightness domain yields an improvement in the quality of display characterization, but on closer inspection several effects remain unexplained.

For simulated GOGO-model data with different voltage or luminance offsets and a little added noise, the fitting procedure never failed to find the model parameters. This leads to the conclusion that the GOGO-model might not correctly describe the relation between input voltage and output luminance of a CRT. The addition of gain and offset parameters has undoubtedly improved the gamma model, but it cannot correctly predict the luminance output over the whole input voltage range.

A possible explanation is that the century old single-gamma model for CRTs is too simple to model the physical reality, and not all the differences in gamma measurement results can be attributed to differences in voltage offsets. This is exactly what Olson [Ols95] established in extensive tests on CRTs over the whole input voltage and electron beam current domain. He found that gamma varied over the voltage domain from values as high as 9.5 on the low end to a minimum of 1.5 on the high end of the scale, with a range of about linear decrease in the middle. His measurements far exceeded the range of normal display operation and in his paper the normal range of voltage operation of the measured displays is not specified, and luminance was not measured, but it seems reasonable to assume the normal display operating range to be in the nearly linear decreasing gamma region. Berns [BK97] argues that changes in gamma only occur at very low voltages and luminances, and the effects could hardly be measured and would be too small to be perceived. But Olson’s findings place the normal gamma of about 2.2 in an area where gamma is rapidly decreasing with increasing voltage, and Olson concludes that a simple gamma does not suffice for the purpose of high-resolution film recording.

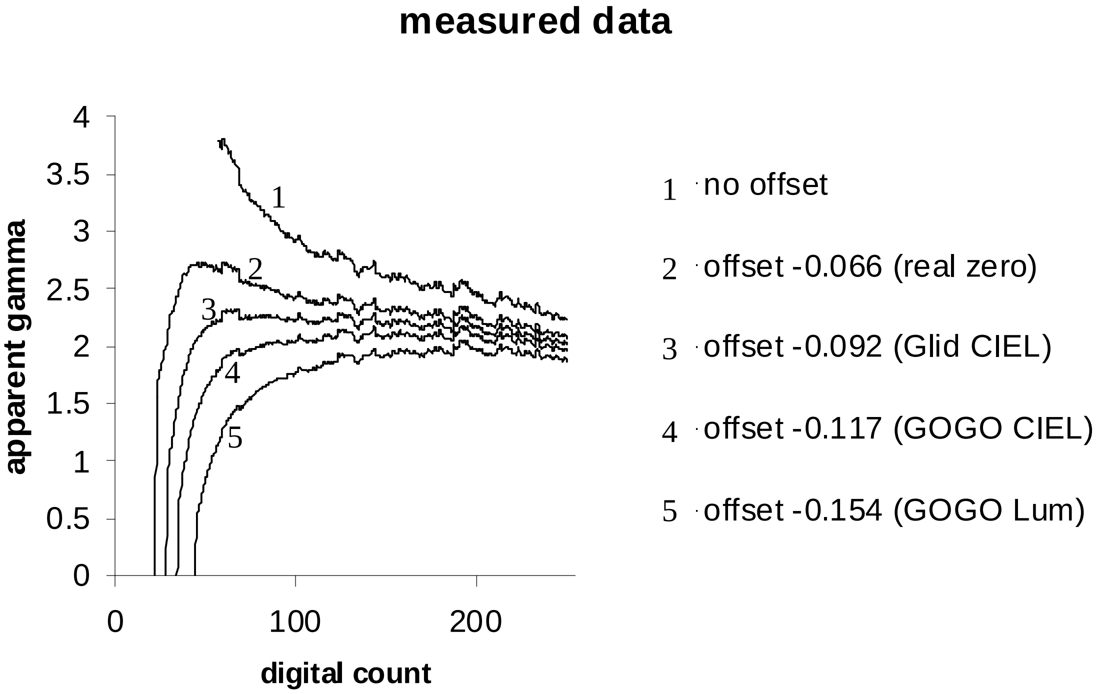


Figure 5.1. Apparent gamma for measured data with different values for offset and gamma range as provided by the different model fits. The descriptions in the legend have the same top-to-bottom order as the curves at digital count 100.

An indicative test for the changing gamma hypothesis can be performed by splitting the data samples in an upper 128–255 and a lower 0–128 half. For all the tested displays the gamma fitted in a GOGO-model for the lower half was invariably higher than the gamma fitted for the upper half.

The change in gamma can also be visualized by plotting the apparent gamma, a differential measure characterizing the slope at one point of the curve.

$$g_{app} = \frac{\log(Y_{i+1} - Y_0) - \log(Y_i - Y_0)}{\log(d_{i+1} - d_c) - \log(d_i - d_c)} \tag{5.9}$$

The apparent gamma is extremely sensitive to noise in the measurements and to a correct estimation of the voltage offset d_C. In Fig. 5.1, a moving average over 9 samples of a full scale (0, 1, 2, . . . , 255) measurement of a delta-gun CRT is plotted. Even then the small peaks caused by slight misalignments in the graphics card DAC (at 64, 128, and 192) and between different luminance meter scales (at 197) are clearly visible. It also shows the influence of different voltage offset estimations on the apparent gamma. If the GOG(O)-model is correct, then with the right voltage offset the apparent gamma would be a straight horizontal line at gamma level. In the figure, the lines are either curved or show a decreasing gamma for increasing digital counts in the range 60–255. These curves can be compared with the apparent gamma for simulated data shown in Fig. 5.2. Comparison shows that the measured and offset corrected curves in Fig. 5.1 have shapes similar to the curves with a descending gamma in Fig. 5.2. Instead of using Olson’s complex formula, with voltage constants that should be measured inside the display housing, these curves are based on the GOGO-model function with a linearly decreasing gamma over the luminance range of the CRT. This linear

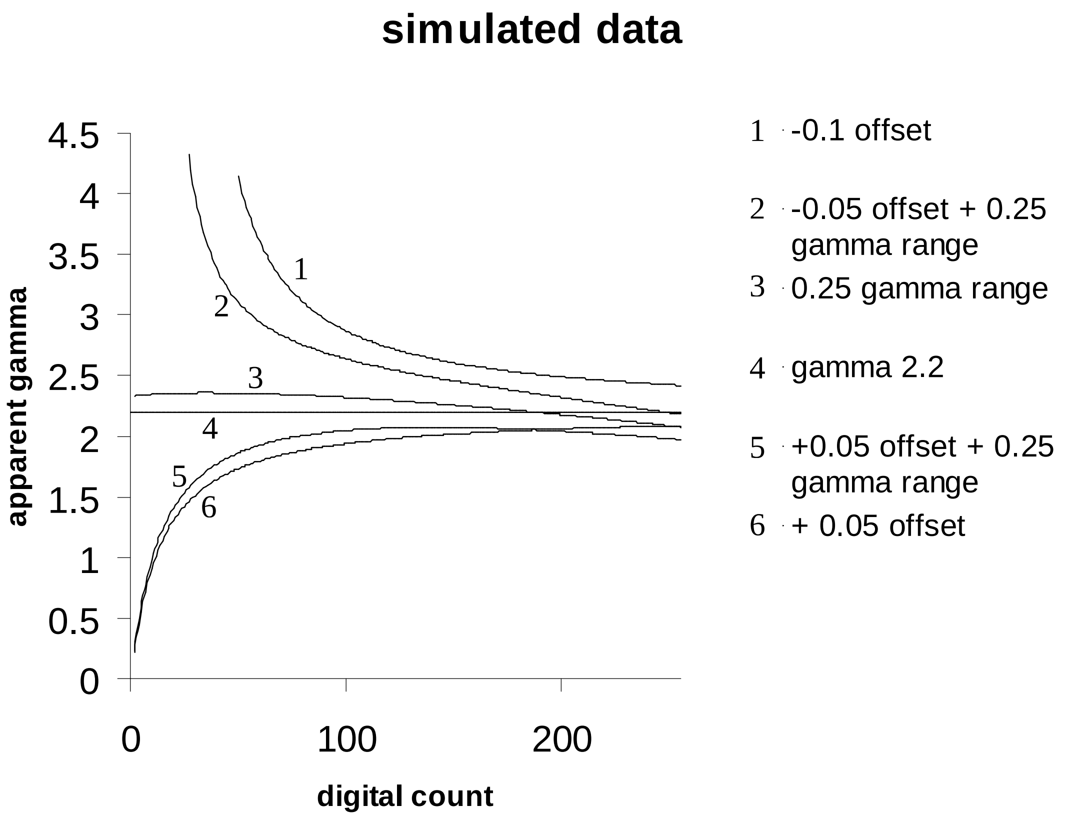


Figure 5.2. Apparent gamma for simulated data with gamma 2.2 and different values for offset and gamma range. The descriptions in the legend have the same top-to-bottom order as the curves at digital count 100.

shift can be characterized by the parameter A_R denoting the range of gamma variation in the gliding gamma model.

$$Y_{i,R} = Y_{0,R} \qquad \text{if } k_{g,R}\frac{d_i}{d_{\max}} + 1 - k_{g,R} \le 0,$$

$$Y_{i,R} = \left(k_{g,R}\frac{d_i}{d_{\max}} + 1 - k_{g,R}\right)^{g + A_R\left(0.5 - \frac{d_i}{d_{\max}}\right)} + Y_{0,R} \qquad \text{otherwise.} \qquad (5.10)$$

As shown in Tables 5.1-5.3, the gliding gamma model is better fitted to explain the variation in the luminance measurements. The total root mean square error is smaller in the CIE lightness domain as well as in the luminance domain. The clipping value is nearer to the apparent threshold in the measurements. For the delta gun display that would mean an increase in the amount of possible colors from 11 to 13.5 out of the maximum 16.7 million.

5.3.6 Computational Complexity

The introduction of a fourth parameter in the optimization procedure raises the chances of divergence or finding a suboptimal solution. We used an LU-decomposition method with the following constraint: $Y0 \ge 0$ and $Y(d_{max}) = Y_{max}$. The first is a physical constraint: luminances cannot be negative. The latter is a matter of computational efficiency in color generation:

there is no need to check if the generated color exceeds the maximum luminance for a primary.

Estimating the starting values for the parameters in the minimization can speed up the procedure and raises the chances of finding the right solution, with more accurate results. If there is any offset luminance, the most likely source is a voltage offset for the other primaries. It is reasonable to assume that the voltage offset for each primary is proportional to its maximum luminance.

$$Y_{0,R,start} = Y_0 \frac{Y_{\max} - Y_0 - Y_{\max,R}}{Y_{\max} - Y_0} \tag{5.11}$$

The clipping value and therefore the gain factor can be estimated by extrapolating the power function from the digital value d_{z1}, with the lowest luminance measurement higher than Y_0, Y_{z1}.

$$d_{c,R,start} = d_{z1,R} - 255\left(\frac{Y_{z1,R} - Y_{0,R}}{Y_{\max,R} - Y_{0,R}}\right)^{1/g} \tag{5.12}$$

$$k_{g,R,start} = \frac{d_{\max}}{d_{\max} - d_{c,R}} \tag{5.13}$$

with γ the nominal system gamma: 2.2 for PC, 1.8 for Macintosh, etc. A starting value for γ could be computed from the measured values at about 40% and 60% of the digital code scale. For a graphics card with 256 values and 17-step sampling this would be given as follows:

$$g_{R,start} = \frac{\log\left(Y_{144,R} - Y_{0,R}\right) - \log\left(Y_{112,R} - Y_{0,R}\right)}{\log\left(144 - d_{c,R}\right) - \log\left(112 - d_{c,R}\right)} \tag{5.14}$$

In practice, this gamma estimation is not reliable enough due to the cumulative errors of Y_0 and $d_{c,R}$ and the noise enhancing effect of the log-difference measure. Nominal gamma values, if available, generally provide better starting values. For the gamma range a value of 0.25 appeared to be a good starting point.

These starting values provide suitable approximations of the final parameters in the optimization; they cannot guarantee an optimal solution. However, the recognition of an optimal solution is not difficult. For a 17 sample fit, the total CIE Lightness RMSE should be smaller than 1. Correct solutions from other fitting methods can be used as starting values.

5.4 Discussion

The gliding gamma model with optimization in the CIE lightness domain provides a better characterization of the CRT display. Figures 5.3 and 5.4 show the error of the fits for different models in the luminance and the CIE-lightness domain, respectively. The GOGO model fit optimized on luminance error performs reasonably well in the luminance domain, but has substantial errors for low digital counts in the lightness domain. Adapting the optimization to the lightness domain provides a smaller average error, which is more spread over the whole domain. The error for the gliding gamma model with optimization in the lightness domain is hardly noticeable in the lightness as well as in the luminance domain. The results in the tables and in a great number of measurements and fits not reported here, show the model is consistent with the physical reality of the CRT configuration and gives good estimations of voltage offsets and luminance offsets caused by flare.

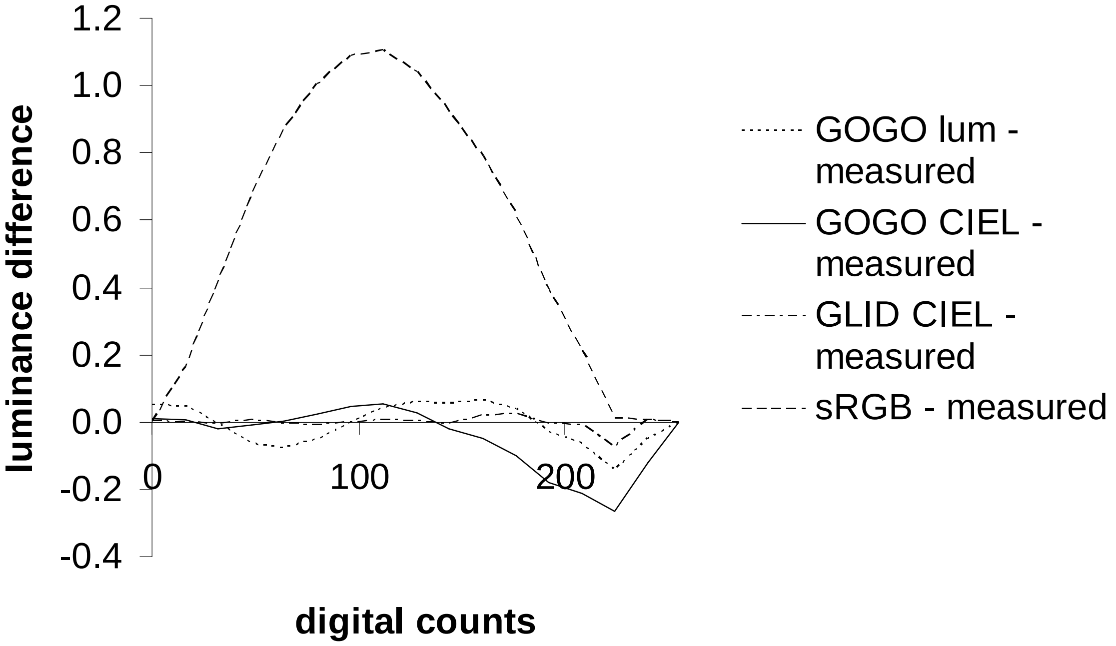


Figure 5.3. Luminance differences between measured data and fits for different models for the green channel of the delta gun CRT and the sRGB standard. The GLID lum data is very close to the GLID CIEL data and therefore left out.

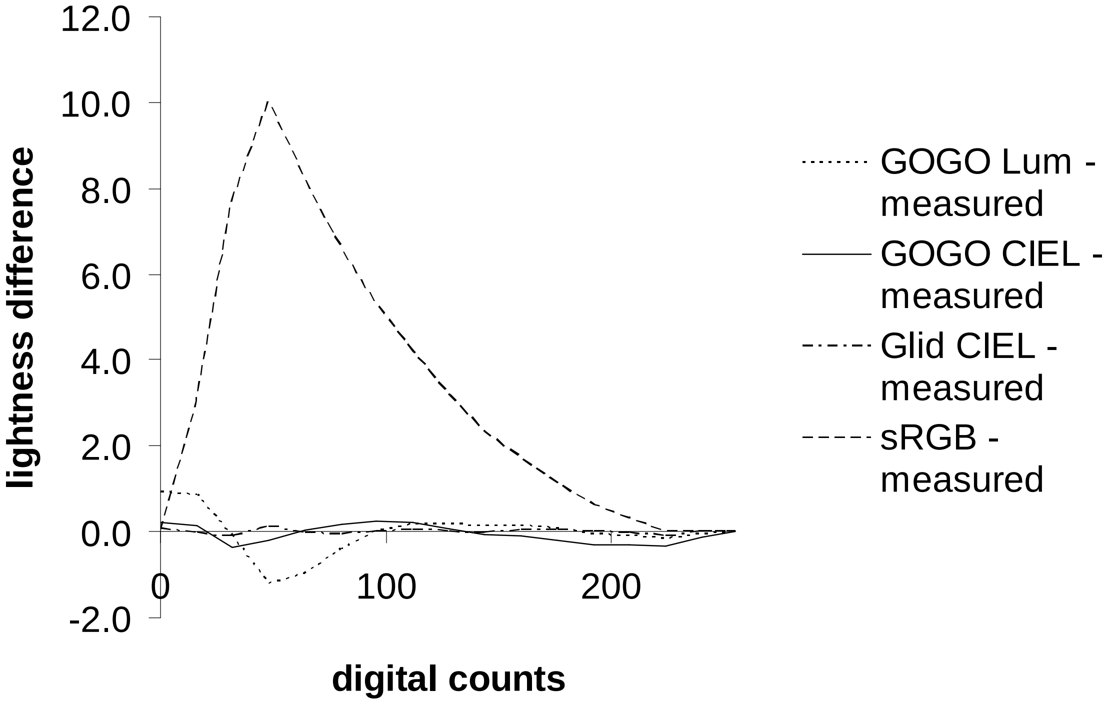


Figure 5.4. CIE lightness differences between measured data and fits for different models for the green channel of the delta gun CRT and the sRGB standard. The GLID lum data is very close to the GLID CIEL data and therefore left out.

The improved model comes at the cost of an extra parameter describing the linear range over which gamma varies. The gamma shift poses problems in determining the inverse function needed to compute the digital counts for each primary (R, G, B) from the CIE-coordinates of the desired color. This can be solved by using a short iterative procedure, but a more common approach is working with lookup tables. Both methods add to the computational complexity of correct color generation.

On the whole the shifting gamma seems to make display characterization more complex, and take us farther away from the most important advantage of the gamma model methods: the single parameter display characterization. Four parameters are needed to describe the model and gamma is no longer fixed. But a closer look at the results in Tables 5.1-5.3 shows that the central gamma value in the gliding gamma model fits has less variation over the primaries and is closer to the nominal gamma values than the fixed gamma in the GOGO fits. So the concept of gamma as a single figure describing the TRC of a CRT still stands.

5.5 Conclusion

The two methods suggested clearly improve the models fitted to the CRT’s TRC. The optimization in the lightness domain enhances the perceptual validity of the models. These could be further refined by making allowances for the viewing intent, that is, the intended view-surround contrast conditions (e.g., office, living room television, or cinema lighting conditions).

The voltage dependency of CRT gamma, detected and physically explained by Olson, [Ols95] can have a measurable effect on the TRCs within the luminance operating range of office CRTs. The gliding gamma model provides more accurate fits for these TRCs in the lightness and luminance domain.

The gliding gamma model appears too complex to impose it on other display technologies. A new standard should be developed rather than trying to fit the TRCs of the new technologies to the technical oddities of the CRT. As shown in Figs. 5.3 and 5.4 the sRGB standard does not do a great job of describing the TRC of at least one sample of the CRT technology it is meant to represent. The “physical properties” of the display technologies should not be the base for color management standards, and these should be based on human lightness perception, with CIE lightness the most likely candidate. New technologies already have lookup tables to correct their technical TRCs and these could easily be filled with lightness correction curves. And with the acceptance of the digital display interface there should be no problem to integrate lightness correction LUTs in the CRT. The gliding gamma model can then be used to compute the correction tables.

6 Black-level offset: Characterization and correction[♦]

Abstract

The correct setting of the black level is an important step in the (re)calibration of an electronic display. This study looks at the consequences of black-level offset, the possibilities for display characterization with offset, offset correction, and the ability of average untrained users to visually correct the black-level setting with the contrast and brightness controls on the display. In an experiment, 32 subjects were asked to optimally set the black level according to two types of instructions (short and extensive, between subjects) under two levels of illumination (low and office, between subjects) for two types of displays (CRTs and LCDs, within subjects). Most subjects were not able to set the black level near optimal for either display, with any combination of instruction and illumination level. The LCD did not have an optimal black level. For the CRT, optimal black level did not provide minimal differences with the sRGB standard tone reproduction curve.

6.1 Introduction

For the reliable reproduction of colors over the Internet, several color-management standards are proposed or already in use. The critical factor in these standards is the characterization of the user display. For an accurate characterization luminance and chromaticity measurement instruments are required. Special display characterization kits combining measurement instruments and software for stimulus generation and color management are now available, but for the general computer user the additional cost does not outweigh the gains. PC users can fall back on the simple and less-reliable default RGB color space (sRGB) for color management. The sRGB standard was initiated by the International Color Consortium and adopted by the International Electrotechnical Commission as standard 61966 and is primarily designed for cathode-ray tubes (CRTs). It assumes the computer display adheres to the sRGB definition or at least does not deviate much from this definition.

The standard sets reference values for

- phosphor chromaticities (primary-color coordinates),
- white point or correlated color temperature (CCT): $x = 0.3127$, $y = 0.3290$ (D65),
- transfer function, tone reproduction curve (TRC), or gamma: $\gamma = 2.2$,
- display-model offset or black-level offset: 0.0 cd/m^2,
- display luminance level: 80 cd/m^2,
- veiling glare (reflection level): 1 %.

Indeed, most computer displays currently on the market have an sRGB setting, which is supposed to set the right voltage levels to produce the correct TRC, white point, and luminance levels. To check or recalibrate the computer-display settings, a user without measurement instruments needs to fall back on the usually minimal support provided by the

---

[♦] This chapter first appeared as article: Besuijen, J. (2006). "Black level offset: Characterization and correction", Journal of the Society of Information Display **14**(6): 895-903.

manual or apply visual calibration methods provided as part of software applications or freely available on the Internet.

In this study, we look at the consequences of black-level offset for color fidelity and the effect it has on the characterization of other display parameters. We also look at the potential of visual calibration methods to correct the black-level offset, the usability issues concerned with the use of these methods, and the capabilities of the average untrained user to visually calibrate his/her computer monitor with these methods. We look especially at the influence of the black-level offset on the tone-reproduction curve (TRC), which defines the relationship between the code value d on the graphics card, with the maximum code value d_{max} (usually 255) and the resulting luminance Y on the display screen. The sRGB standard assumes a reference display characterization of a CRT based on a simple gamma function with $\gamma = 2.2$:

$$Y(d) = Y_{\max}\left(\frac{d}{d_{\max}}\right)^{2.2} \qquad (6.1)$$

A linear portion is integrated into the encoding specification of the transfer function of the dark end signal to optimize encoding implementations. According to the standard, the equations below closely fit Eq. (6.1). This should maintain consistency with the legacy of desktop and video images.

$$Y(d) = Y_{\max}\frac{d}{d_{\max}}/12.92 \qquad for\frac{d}{d_{\max}} \le 0.04045 \qquad (6.2a)$$

$$Y(d) = Y_{\max}\left(\frac{(\frac{d}{d_{\max}} + 0.055)}{1.055}\right)^{2.4} \qquad for\frac{d}{d_{\max}} > 0.04045. \qquad (6.2b)$$

There are several potential problems with the TRC characterization in this standard:

- some monitors have no sRGB settings,
- the sRGB monitor settings are not tested for conformation to the standard,
- the user can change the monitor's brightness and contrast settings,
- the monitor's luminous output changes over time,
- the TRC of the standard is based on the gamma model for the CRT and other display technologies have different TRCs that do not conform to this model.
- recent studies have shown that gamma might not be constant over the range of code values, but higher for lower code values [Bes05; Ols95].

There are two ways in which these problems show up in the TRCs for the primary-color channels of a CRT. The first is a black-level offset; if this offset is positive, it means $Y(0)$ does not equal 0, black looks gray, and the image might look washed out. If the offset is negative, clipping will occur: several code values are mapped to black and the image will look dark. The black-level offset can be corrected by adjusting the contrast and brightness controls on the display.

The second is a difference in gamma parameters between the primary-color channels, which means white and neutral grays do not have the same color. The gamma parameters can be measured by a visual matching task and then be corrected by setting a look-up table (LUT) on the graphics card.

6.2 Characterization and correction of black-level offset

For color-management purposes, there are two problems related to the black-level offset. If the display has to be used in an illuminated environment where veiling glare cannot be avoided, a black-level offset might be desired to preserve the color differences within the darker parts of the image. In that case, a correct characterization is needed to provide an optimal color fidelity for the circumstances. If the level of veiling glare is low, then the black-level offset should be corrected. A correct characterization of the display can be helpful in determining the brightness and contrast control settings that produce the lowest black-level offset. The direct measurement of the black-level offset requires sensitive measurement equipment and a controlled measurement environment. It is possible to estimate the black level by model fitting from measurements at higher luminance levels, but the estimations become more reliable with accurate measurements at lower luminance levels. Some combinations of models and estimation methods are extremely sensitive to measurement errors at low luminance levels.

Display characterization by fitting gamma to the slope in the log-luminance – log-digital-code domain is especially sensitive to black-level offsets. Roberts [Rob93] demonstrated the influence of luminance and voltage offsets on gamma fitting in the log–log domain and developed a method based on differential slope analysis to characterize these offsets. The combination of differentials and the log–log domain is bound to enlarge possible measurement errors. Berns [BMG93] developed the Gain-Offset-Gamma (GOG) model, of which Eq. (6.2b) is an example, later refined by Katoh, Deguchi, and Berns [KDB01a; KDB01b] in the gain-offset–gamma-offset (GOGO) model, which relates the amount of light Y_P generated by the CRT gun of primary P (which could be R,G, or B) to CRT parameters:

$$Y_P(d) = k_P\left\{a_P\left[(v_{\max}-v_{\min})\left(\frac{d}{d_{\max}}\right)+v_{\min}\right]+b_P-v_{C,P}\right\}^{\gamma_P}+Y_{0,P} \tag{6.3}$$

The amount of light of a computer-controlled display depends on the digital counts or code values in the DAC (d, d_{max}), the video generator voltages (v_{min}, v_{max}), the video amplifier parameters (a_P, b_P), the CRT gun ($v_{C,P}$, γ_P), and the properties of the faceplate and phosphor materials (k_P). $Y_{0,P}$ is the black-level offset, originating from sources other than the primary channel P.

Note that a_P and b_P are supposed to correspond to the contrast and brightness settings of the monitor, and that γ_P is assumed constant under all conditions. By normalization, the number of model parameters can be reduced to three: gamma (γ_P), the normalized gain/offset parameter $k_{g,P}/k_{0,P}$ and the black-level offset $Y_{0,P}$:

$$\begin{aligned} Y_P(d) &= \left(Y_{\max,P}-Y_{0,P}\right)\left(k_{g,P}\frac{d}{d_{\max}}+k_{0,P}\right)^{\gamma_P}+Y_{0,P}, & \left(k_{g,P}\frac{d}{d_{\max}}+k_{0,P}\right) &\geq 0 \\ Y_P(d) &= Y_{0,P}, & \left(k_{g,P}\frac{d}{d_{\max}}+k_{0,P}\right) &< 0 \end{aligned} \tag{6.4}$$

with:

$$k_{g,P} = \frac{a_P(v_{\max}-v_{\min})}{a_P v_{\max}+b_P-v_{C,P}} \quad (6.5) \qquad k_{0,P} = \frac{a_P v_{\min}+b_P-v_{C,P}}{a_P v_{\max}+b_P-v_{C,P}} \quad \text{and} \quad k_{g,P} = 1-k_{0,P} \quad (6.6)$$

Notice that the gain is also dependent on the monitor offset (brightness) setting, and the offset is also dependent on the monitor gain (contrast) setting, which could explain possible problems users have to correctly set their display configuration.

According to Poynton [Poy02], the names used for these controls for a CRT are misleading with respect to their functions: “The control called brightness mainly affects reproduced contrast, and the control called contrast ideally affects only brightness!.” He argues that “brightness” should be named “black level.” This corresponds with Eq. (6.6), which shows that b_P can be used to level out the cut-off voltage, $v_{C,P}$, in the CRT gun to produce the desired offset, assuming v_{min} is nearly 0 and therefore $a_P v_{min}$ is small compared to $v_{C,P}$. If the total offset is 0, then black will be truly black, contrast will be maximal, and changing a_P will only affect the brightness of the picture.

In practice, CRTs will not adhere to the ideal model and some black-level offset will always be present. On most computer monitors, there are no controls available to individually adjust the gain and offset of the primaries R, G, and B, only brightness and contrast controls are available to adjust all primaries at once. In practice, the parameters of the electronic circuits for the three primaries will differ, resulting in at least offset in one primary for any brightness/contrast setting. Any positive offset $k_{0,P}$ for one primary will manifest itself as (part of) a luminance offset $Y_{0,P}$ in another primary.

In their search for a more robust model fitting method, Deguchi and Katoh [DK97] looked at the effect on color for 12 different combinations of contrast and brightness control settings for a CRT. Their main purpose was to find the gamma model with the best fitting performance for the different settings. They found the gamma-offset–gain-offset (GOGO) model performed best with the largest variation in the offset and gain parameters and the least variation for the gamma parameter. This latter conclusion was based on the constant gamma assumption for CRTs. Recent research suggests that this assumption at least does not hold for all CRTs.

Through extensive measurements of the electrical characteristics of a CRT gun, Olson [Ols95] showed that gamma is not constant over the working range of voltages. In an earlier study [Bes05], we showed that a GOGO-like model optimized in the CIE lightness domain with an increasing gamma for lower digital counts better fitted the TRCs and black-level offsets for three CRTs.

6.3 Minimizing the black level

Park et al. [KPK03; PKB00] devised two methods based on polynomial fittings of part of the TRC for different settings of contrast and brightness to calculate the brightness setting with minimal black-level offset. One method is based on the GOG-model, the other approach is very similar to that of Roberts [Rob93] discussed above. The second-order-polynomial fitting parameters can be expected to be rather sensitive to measurement errors and deviations from the model, but the use of more measurement series might make this method more robust.

Based on a multiple measurements series, the optimum brightness can also be estimated with non-linear fitting methods based on Eqs. (6.4)–(6.6). With the estimates of $k_{g,P}/k_{0,P}$ for several brightness and contrast settings and assuming linear relationships b_P = brightness setting/100 and a_P = contrast setting/50, estimates can be made for v_{min}, v_{max}, and $v_{C,P}$. And the optimum brightness setting for a given contrast setting can be found for $a_P\, v_{min} + b_P - v_{C,P} = 0$.

Models are rarely perfect, and with the circuitry and firmware in displays getting more and more complex, it is unlikely that one model can be used to describe the behavior of all available CRTs. In providing a choice between preset correlated color temperatures (CCT) in CRTs, *e.g.*, 9300°K, 6500°K, and 5000°K, designers have to make concessions regarding optimal settings. The CCT can only be changed by changing the ratio of the brightness and contrast parameters of the individual primaries: this is bound to lead to black-level offset for some primary.

For LCDs, there is always some black-level offset: the liquid crystal is unable to achieve zero transmittance. The brightness control of an LCD typically alters the backlight luminance and therefore the black-level offset as well. The natural TRC of liquid crystal does not have the shape of the gamma function but more of an S-shape, with saturation effects for high code values. Conversion to a gamma-shaped TRC can be reached by a LUT or conversion circuit in the display. If an LCD has a contrast control, it probably changes the gain at the output of this LUT. Setting the contrast to maximum for an LCD might therefore lead to saturation problems. For the LCD used in the experiment, setting the contrast to maximum for CCT 6500°K clearly saturated the red primary at about code value 223.

6.4 Effect of black-level offset on color

A positive black-level offset adds an amount of light in a usually neutral color to the intended color, shifting the color coordinate towards the neutral point. This shift will be larger for dark colors. Based on this fact, Berns *et al.* [BFT03] devised a method to estimate the black-level offset by minimizing the variation in the estimated primary-color coordinates from color measurements of the primaries at series of digital counts. They reached good results with measurement series 16, 32, ... 255. The advantage of this method is that it can be used for any display technology.

In practice, users can expect considerable color differences from the sRGB standard. Bodrogi *et al.* [BSB$^+$02] looked at the differences between measured tone reproduction curves and the sRGB TRC for the primary-color channels for 11 different CRT monitors with nine different settings, including three brightness/contrast control settings. They found considerable color differences in CIELAB space, mean $\Delta E_{ab}^* > 20$ for a large set of colors, which seem to be mainly stemming from black-level offsets and differences between the TRCs of the primary-color channels.

To compare displays to the sRGB standard, the ideal gamma function Eq. (6.1) is more helpful than the optimized encoding implementation with the linear portion Eq. (6.2). The difference in luminance between these two functions is indeed small, but the difference on a perceptual scale with a reference white luminance, the CIE lightness domain from the CIELAB, and CIELUV color spaces is noticeable with color differences $\Delta E_{ab}^* > 1.0$ for many code values, as shown in Fig. 6.1 for the primary sRGB colors. This lightness difference can also lead to chromaticity differences; for about the worst case, sRGB-designed magenta color RGB = (176, 32, 176) on a display with a TRC according to Eq. (6.1), this will mean a significant loss of saturation. A CRT without a built-in LUT will have a continuous gamma function such as Eq. (6.1) over the entire scale, and in a comparison with Eq. (6.2) the discontinuities would at the least distract.

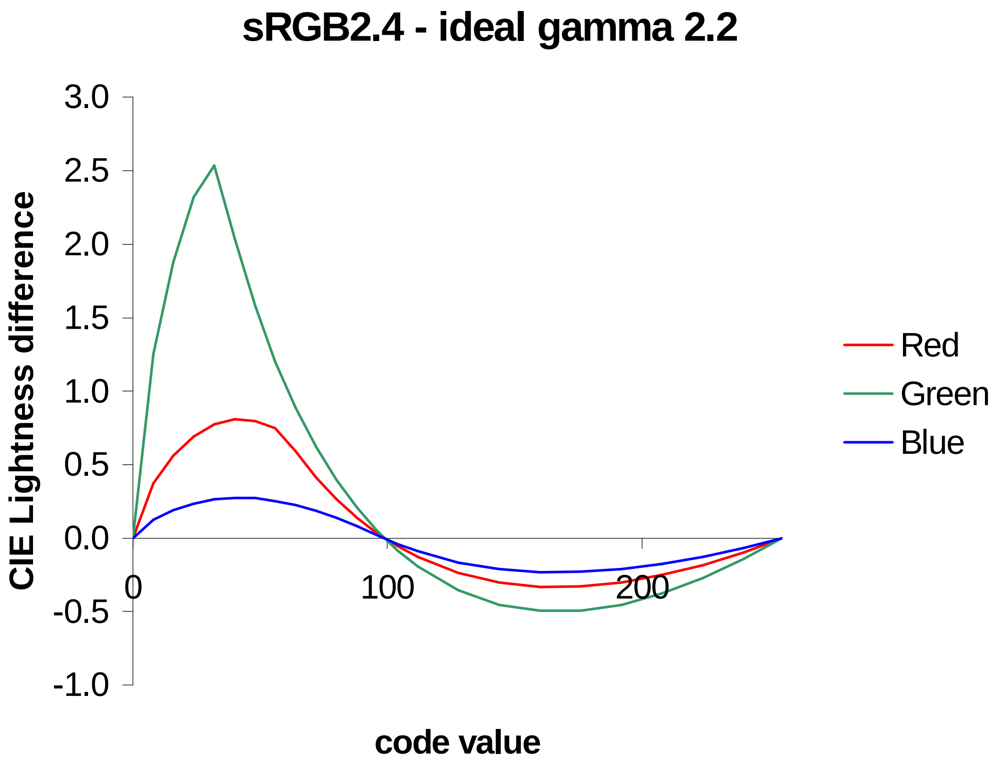


Figure 6.1. CIE Lightness differences between sRGB TRC and an ideal gamma = 2.2 function

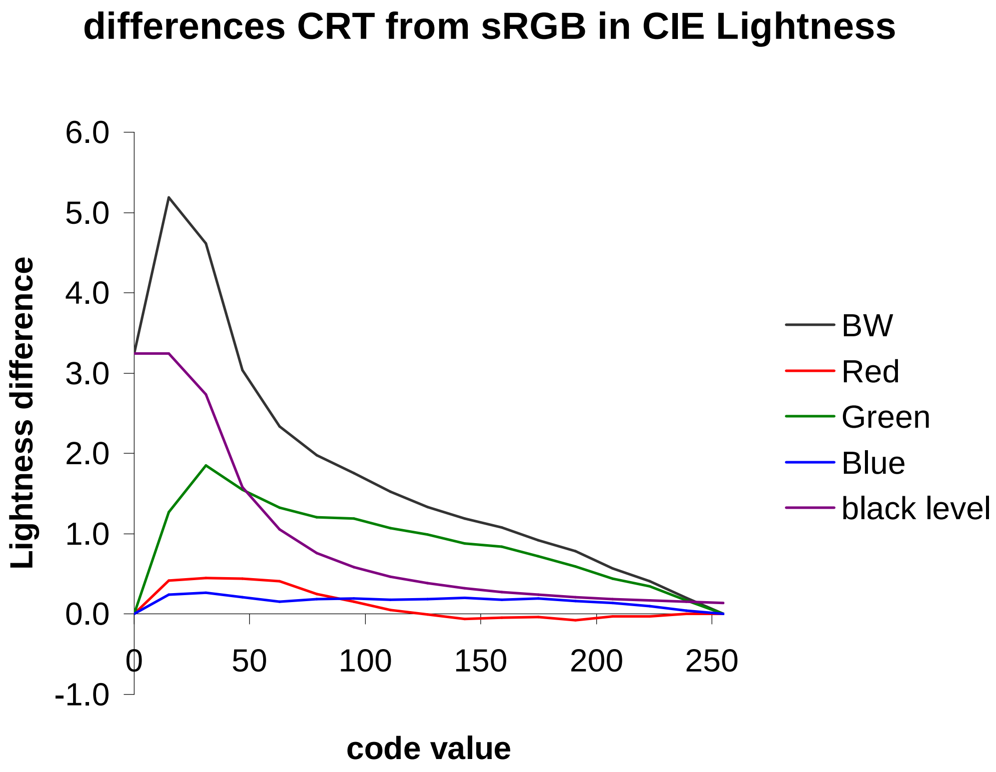


Figure 6.2. Differences in CIE Lightness from sRGB for the CRT in the experiment. BW gives the overall difference for neutral stimuli. Red, Green, and Blue give the differences for the separate primaries corrected for black level. 'black level' shows the difference if only black level offset were present.

Table 6.1. Experimental design

Instruction	Simple		Elaborate	
Illumination	Low	Office	Low	Office
Display	CRT and LCD	CRT and LCD	CRT and LCD	CRT and LCD
Number of subjects	8	8	8	8

6.5 Visual offset-correction experiment

Calibrating the display by adjusting the brightness and contrast settings to set only code value 0 to be precisely black while keeping a bright enough image is an essential step in maintaining compliance to the sRGB standard. In an experiment, 32 participants were asked to optimally adjust the brightness and contrast settings of a CRT and LCD (within subjects) according to types of stimuli with instructions, short and elaborate (between subjects), under two levels of illumination, low and office (between subjects). The design (see Table 6.1) was balanced over the participants to eliminate sequence effects.

After finishing the brightness and contrast adjustment, the values from the on-screen display were put on record. Then the participants performed a control visual detection task on the display to determine the threshold code value that provided a recognizable stimulus. This test was added to check if the subject's visual capabilities could be the cause of non-optimal settings. The experiment was performed in combination with a visual gamma matching experiment are reported in chapter 7.

6.5.1 Participants

The participants were psychology students from the University of Twente, participating as part of their first year curriculum. Thirty-two Students participated, six students were Dutch and 26 were German with a sufficient knowledge of the Dutch language, 24 students were female and 8 were male, ages ranged from 17 to 26. All subjects had corrected or uncorrected 20/20 or better visual acuity for stereoscopic vision measured by the TopCon Screenoscope-II, Tokyo Optical Co., Ltd. All subjects were tested for color-vision deficiencies with the Ishihara test for color blindness; one male subject had a weak red–green color deficiency. This was judged to have no consequences for the monochromatic brightness matching task in this experiment.

6.5.2 Apparatus

A CRT and an LCD connected to two separate Windows-operated personal computers were used in the experiment. The CRT was a 2-year-old Philips 17-in. model 107T5 controlled by an Intel Extreme Graphics 2 graphics card at a 800 × 600 resolution at a frame rate of 100 Hz. The LCD was a Philips 17-in. model 170B4MG02, also 2 years old, controlled by an NVIDIA GeForce2 MX 100/200 graphics card at the native 1280 × 1024 resolution at 60 Hz. Both displays were connected to the graphics card with a D-SUB cable. Care was taken that no color-management was used.

Brightness and contrast settings could be adjusted by an on-screen dialog displaying a blue bar on a white background with a percentage between 0 and 100 just above the middle of the screen, and minus and plus push buttons on the bottom of the screen. The effect of contrast and brightness settings on the light output of the displays for rectangular stimulus patches of 1/25 of the active area on a black background of white, RGB = (255, 255, 255), and black, RGB = (0, 0, 0), measured with a L203 photometer with a 6° field-of-view luminance probe

Table 6.2. (a-e) Luminance data of displays used in the experiment.

(a) LCD black level in cd/m^2

%brightness %contrast	0	50	100
0	0.18	0.35	0.50
50	0.21	0.37	0.53
100	0.20	0.38	0.51

(b) LCD white level in cd/m^2

%brightness %contrast	0	50	100
0	42.6	85.4	118.5
50	55.8	108.8	146.0
100	65.4	120.8	164.0

(c) CRT black level in cd/m^2

%brightness %contrast	0	50	100
0	0.01	0.33	6.13
50	0.02	0.23	5.47
100	0.04	0.16	4.47

(d) CRT white level in cd/m^2

%brightness %contrast	0	50	100
0	0.01	0.91	8.84
50	11.00	30.50	59.50
100	65.60	99.50	109.90

(e) Luminance for sRGB settings in cd/m^2

	CRT	LCD
black	0.333	0.459
white	92.770	154.440

excluded by a dark gray foam cover and care was taken that no pressure was exerted on the LCD screen.

The displays were measured with the default sRGB settings. The differences in CIE lightness with the sRGB TRC, assuming equal white luminance, are shown in Fig. 6.2 for the CRT and in Fig. 6.3 for the LCD. The figures show that the effect of black-level offset is relatively large for the CRT compared to the deviations from the ideal gamma curve; for the LCD, the deviations from the gamma curve have a larger effect on luminance and color than the black level offset.

6.5.3 Lighting conditions

Half the subjects performed the experiment in office lighting conditions provided by the standard experiment room fluorescent tube lighting. The other half worked under low illumination conditions provided by a small incandescent lamp reflected from the back wall. Under office conditions, the illuminance on the desktop measured with a Macam L203 photometer with an illuminance probe was 530 lux; on the CRT, 266 lux; and on the LCD, 296 lux. For low light conditions, the illuminance on the desktop was 38.1 lux, on the CRT, 12.6 lux; and on the LCD, 13.1 lux.

6.5.4 Stimuli and instructions

Several applications are available to help the user in calibrating the display. These applications provide visual stimuli and instructions on the sequence of operations and optimal state of the visual stimuli. In this experiment, two types of black-level setting tools were compared. The simple tool chosen for the experiment was the Adobe™ Gamma Wizard, part

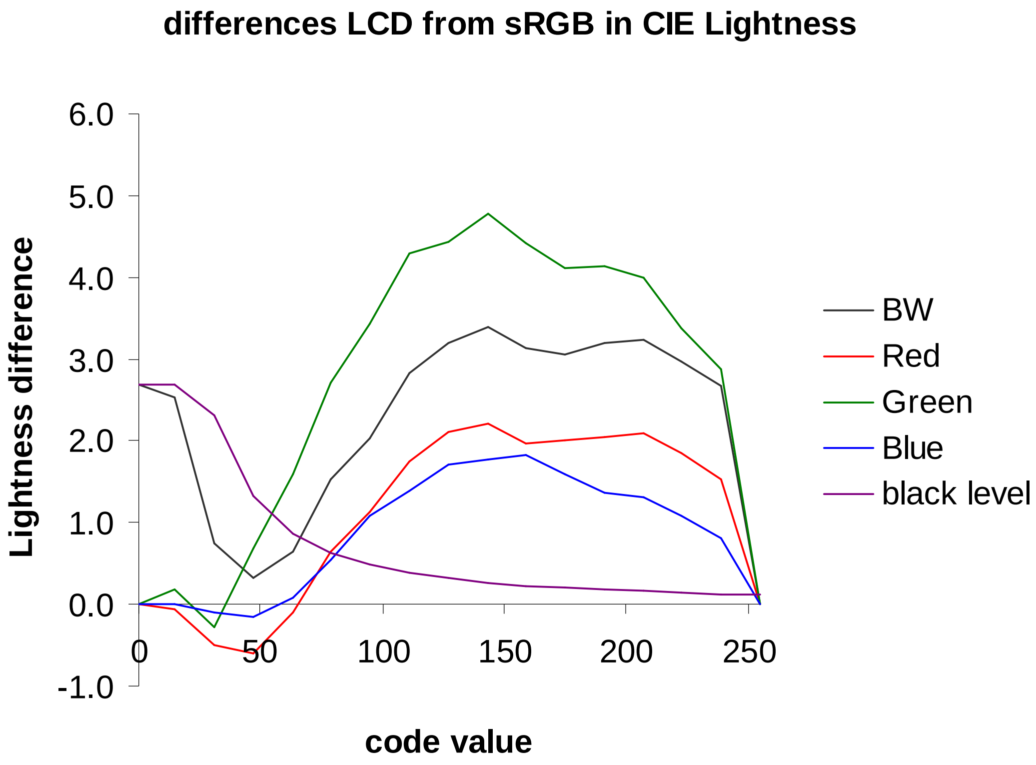
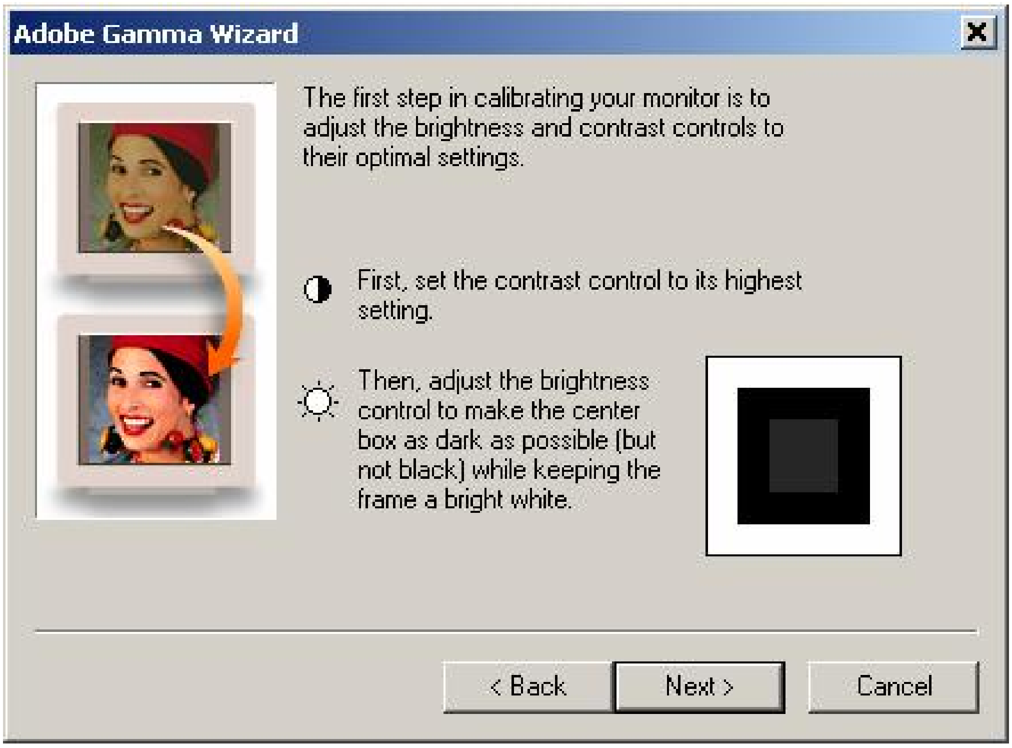


Figure 6.3. Differences in CIE Lightness from sRGB for the LCD in the experiment. BW gives the overall difference for neutral stimuli. Red, Green, and Blue give the differences for the separate primaries corrected for black level. 'black level' shows the difference if only black level offset were present.



The first step in calibrating your monitor is to adjust the brightness and contrast controls to their optimal settings.

First, set the contrast control to its highest setting.

Then, adjust the brightness control to make the center box as dark as possible (but not black) while keeping the frame a bright white.

Figure 6.4. Adobe Gamma Wizard instruction dialogue box for optimal brightness and contrast settings.

of Photoshop™ CS from Adobe Systems, Inc., consisting of one dialog box with a visual stimulus and a two-step instruction, as shown in Fig. 6.4. The visual stimulus consists of a neutral gray square of value 38 [RGB = (38, 38, 38)] in a black square of value 0 in a white square of value 255.The more elaborate tool was part of the calibration program DisplayMate for Windows video edition 1.21 from Sonera Technologies™, consisting of three pages of screen-wide visual stimuli with instructions provided on paper. The first page consists of lines of texts at different contrasts in gray and white (values 128, 192, and 255) to the black background (0) and is meant to be a coarse setting for brightness and contrast. These settings are refined with the second page showing two groups of overlaid rectangles, one with dark gray tones (values 0, 24, and 48) and one with bright gray tones (values 208, 232, and 255). The third screen consists of 32 rectangles of dark gray tones (values 1–64) on a black background and is meant for fine-tuning only the brightness setting.

The participants were handed written general instructions about the design and purpose of the experiment and a manual for handling the on-screen dialog and display pushbuttons.

Depending on the type of instruction used in the experiment, the participants were provided with written instructions in Dutch, comprising the relevant instructions from the DisplayMate manual pages or a written translation in Dutch of the Adobe dialog box. Participants were asked to read all instructions carefully before starting the adjustments. The adjustments were started with the displays in the default sRGB settings.

The experiment leader was present during the experiment to clear possible questions, demonstrate the working of the display controls, and register the resulting brightness and contrast settings.

The visual threshold recognition task was based on the stimuli in the Screenoscope visual acuity task. The character “E” was displayed in one of four possible orientations, the participants had to respond by pressing the arrow key on the keyboard with the arrow pointing towards the open side of the “E”. The first “E” was displayed at code value 76, for every correct answer the code value was decreased and for every false answer the code value was increased until the threshold was found. This process was repeated starting at code value 0. If the resulting threshold code value was no more than one apart, the lowest value was taken as threshold, otherwise the result for the subject was discarded.

The results can be compared in different domains. First, there are the brightness and contrast settings, scales from 0 to 100, which the user actually controls, but stand in a complex relation to the ultimate result: the black level. The perceptual scale for the black level is its CIE lightness with display white as the most obvious choice for the reference white luminance. With this specific task focusing on the black and dark grays and given the absence of white in the last DisplayMate instruction screen, it is necessary to evaluate the luminance of the black level as well.

The experiment was setup to test the following hypotheses for the three independent variables:

1. Display: black-level setting will be nearer to the general optimum for the CRT than for the LCD. The minimum luminance and CIE lightness for display black that a CRT can produce is much lower than that of an LCD.
2. Type of instruction: Black-level settings will be more accurate for the elaborate type of instruction. For the short type of instruction, the comparison between black and nearly black is between code values 0 and 38. For the elaborate type of instruction, comparison is made between code values 0 and 1.
3. Illumination: black levels will be higher for office-lighting conditions than for low-lighting conditions. Due to veiling glare, the minimum luminance of the display will

increase. It is more difficult to perceive small brightness differences if the average brightness is higher.

6.6 Results

To be able to test the hypotheses, an optimal black-level setting must be defined for CRTs and LCDs. With the results from expert visual inspection before the experiment and the luminance measurement conducted after the experiment, the optimal black-level setting for the CRT was determined for a contrast setting at 100% and a brightness setting at 31%. It should be noted that the optimal black-level setting does not guarantee an optimal TRC, which is close to the sRGB standard. For the CRT in the experiment, the optimal black-level setting turned out to increase the differences with the TRC of the sRGB standard. A brightness setting of about 40% with a small black-level offset produces an improved fit to the standard.

For the LCD, it was not possible to define an optimal setting. The lowest possible value for the black level of the LCD appeared to be 0.18 cd/m^2 for brightness and contrast settings of 0%. If we look at the brightness of black in relation to the brightness of white with the CIE lightness value, we see that for 100% contrast the lightness varies negligibly from 2.762 over 2.842 to 2.809. It is no wonder that a number of subjects expressed difficulties in setting the black level: changing brightness and contrast settings seemed to make little difference. Comparison of the tables 6.2(a-e) shows that the black level is determined by the brightness setting and the contrast settings determines the contrast ratio between black and white; for 0% contrast, the contrast ratio is about 237 and for 100% about 321. These changes may seem considerably large, but are negligible on a perceptual scale like the CIE lightness for black or, for instance, Michelson’s contrast (0.9916 and 0.9938, respectively, on a scale from 0 to 1). The changes are also small compared to the changes between settings for the CRT. Here, the contrast ratio pattern is more complex with a minimum contrast ratio of 1 (no contrast) for contrast and brightness settings of 0% and a maximum contrast ratio of 1640 for a brightness setting of 0% and a contrast setting of 100%; Michelson’s contrasts of 0 and 0.9988, respectively. The CIE lightness of black varies substantially more than that for the LCD; for 100% contrast setting, it varies from 0.551 at 0% brightness setting to over 1.453 at 50% to 23.89 at 100%; indicating that black for 50% brightness setting and higher has a clearly visible luminance.

6.6.1 Differences between displays

Due to the small contrast range of the LCD, it seems there is little use in testing the display hypothesis with the experiment because the LCD had no optimal black-level setting. The results of the experiment show as might be expected that the variance in the brightness settings for the LCD was high, with ranges near 80 on a scale of 100 for the LCD within groups with equal conditions. But for the CRT, the range of brightness settings was large as well, about 50. For the LCD, the variance in contrast settings was high for the elaborate instruction, CRT contrast settings showed a smaller range.

6.6.2 Differences between type of instruction

Effects of the type of instruction and lightning level were tested with an analysis of variance (ANOVA) test including both between-subject factors and their interaction. Contrast settings for the CRT differed significantly for the type of instruction $F(1, 28) = 16.481$, $p < 0.001$.

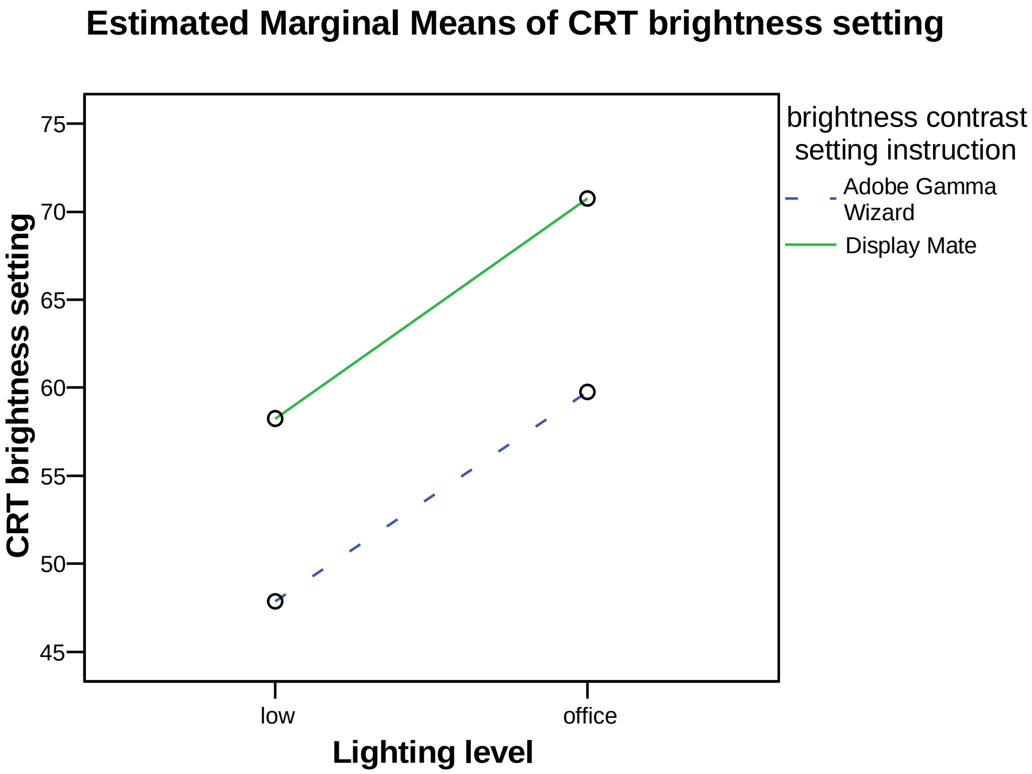


Figure 6.5. Effects of Lighting level and type of instruction on the user CRT brightness setting

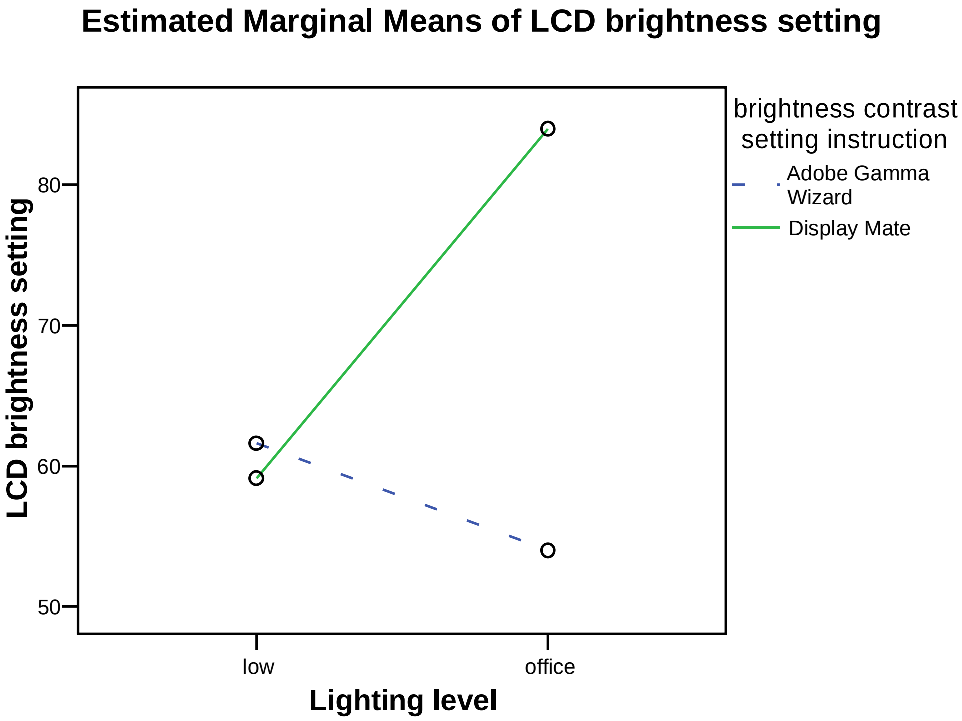


Figure 6.6. Effects of Lighting level and type of instruction on the user LCD brightness setting

Partial eta squared = 0.371, representing a large effect: the short type of instruction leading to higher contrast settings (see Fig. 6.5). It should be noted that the variances of the settings for the type of instruction were significantly different for Levene’s test [$F(3, 28) = 4.185$, $p = 0.014$].

Contrast settings for the LCD differed significantly for the type of instruction $F(1, 28) = 18.630$, $p < 0.001$. Partial eta squared = 0.400, representing a large effect. It should be noted that the hypothesis that the distribution differs from a normal distribution cannot be rejected. These results are not surprising as Adobe Gamma first instructs the user to set the contrast control to the highest setting. Subjects clearly failed to complete this instruction in two cases for the CRT (82 and $\underline{92}$%) and in four cases for the LCD ($\underline{67}$, 53, 64, and 92%) (underlined scores for same subject).

There were no large effects on the brightness settings, but there was a trend for the CRT for the type of instruction; $F(1, 28) = 2.680$, $p = 0.113$, partial eta squared = 0.087 representing a middling effect: Adobe Gamma results in lower brightness settings. These results oppose the hypothesis for this type of instruction. There were no significant effects on the luminance and the CIE lightness of CRT black for this type of instruction.

6.6.3 Differences between lighting conditions

There were no effects on the contrast settings and no large effects on the brightness settings, but there was a trend for the CRT for lighting level $F(1, 28) = 3.486$, $p = 0.072$, partial eta squared = 0.111, representing a large effect: low lighting results in lower brightness settings. There was an interaction trend for the LCD [$F(1, 28) = 2.480$, $p = 0.127$, partial eta squared = 0.081, representing a middling effect (DisplayMate) resulting in higher settings for office lighting (see Fig. 6.6). For CRT black luminance and CIE lightness, the results were somewhat clearer [$F(1, 28) = 3.766$, $p = 0.062$, partial eta squared = 0.119 and $F(1, 28) = 4.301$, $p = 0.047$, partial eta squared = 0.133]. These results (see Figs. 6.7 and 6.8) are in agreement with the hypothesis on illumination.

6.6.4 The recognition test

The results shown in Fig. 6.9 show that only four from the 16 subjects working with the Adobe Gamma instruction on the CRT have threshold detection values that lie somewhere near the nearly black value of 38 in the picture from the Adobe Gamma dialog, all other values were ≤ 5. The CRT brightness settings [$F(24, 3) = 196.358$, $p = 0.001$, lighting level $F(1, 3) = 144.000$, $p = 0.001$ and the interaction $F(2, 3) = 57.000$, $p = 0.004$] had a significant effect on the threshold values of the recognition test for the CRT. For values below 40% of the brightness setting scale, threshold values increased with decreasing brightness. The threshold level for office-lighting conditions is higher and steeper.

6.6.5 General performance

In general the average user performed worse than the default sRGB settings with higher mean values for CRT black luminance: $F(1, 28) = 15.439$, $p = 0.001$, partial eta squared = 0.355 and CRT black CIE Lightness: $F(1, 28) = 13.568$, $p = 0.001$, partial eta squared = 0.326, with only 10 subjects out of 32 producing lower black level settings. For low illumination levels the performance did not differ significantly from the default settings, with 7 subjects out of 16 producing lower black level settings.

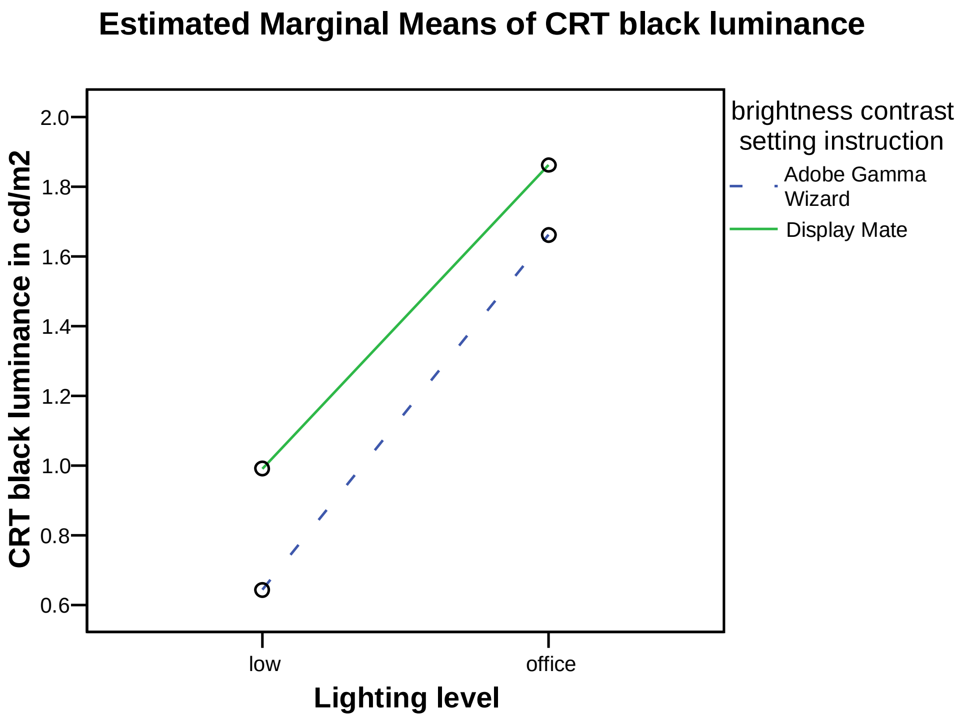


Figure 6.7. Effect of Lighting level and type of instruction on luminance of CRT black level

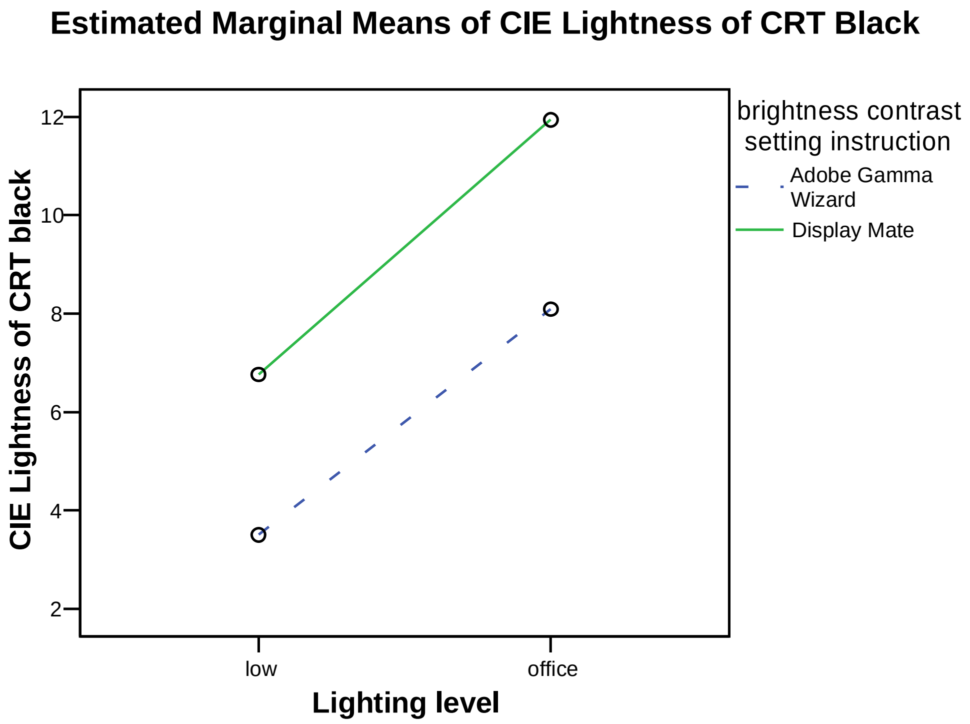


Figure 6.8. Effect of Lighting level and type of instruction on CIE Lightness of CRT black level

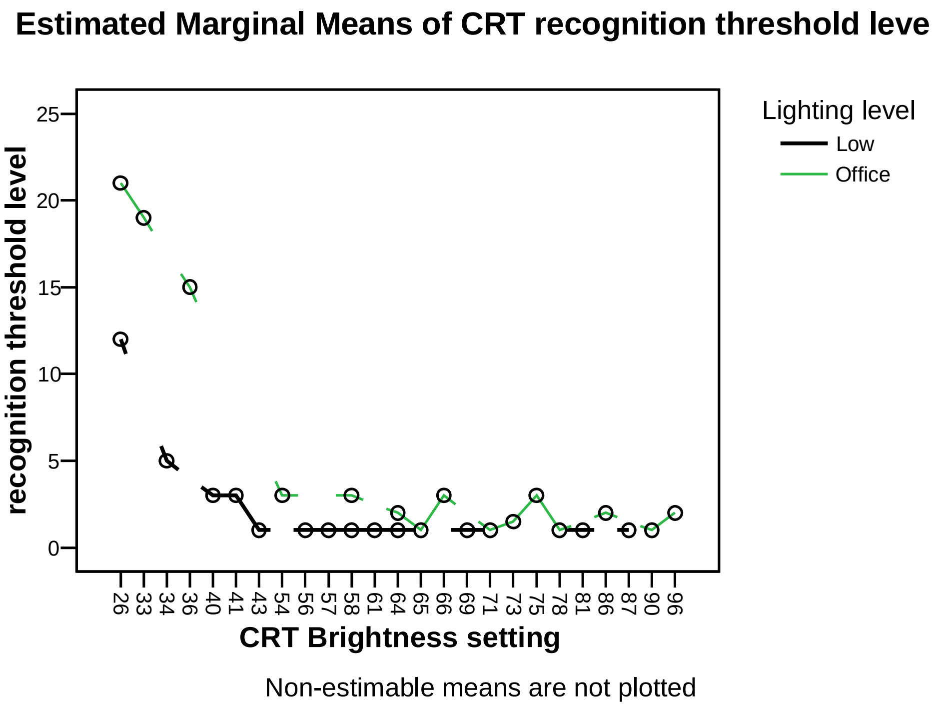


Figure 6.9. User CRT "Brightness" settings with related recognition threshold levels.

6.7 Conclusions

From the results of this experiment, it can be concluded that most of the average untrained users are not able to find the optimal black-level setting with the provided instructions by adjusting the contrast and brightness controls. This is not surprising for the LCD used in the experiment because it did not have a clear optimal setting. The contrast control on the LCD had no real purpose: the range of adjustment of perceived contrast was small and the red primary could be driven in saturation for high settings. For the CRT though both controls clearly have a function and an optimal setting would have significantly improved the color-management properties of the display. Black-level settings were clearly closer to optimal with low ambient illumination: black-level adjustments should be made under low ambient illumination. There are several improvements that can be made to the black-level optimizing procedure. More or more-detailed brightness and contrast controls setting instructions do not seem to work for the average user. During the experiment it could be observed that settings made with the first page of stimuli and instructions of the elaborate type of instruction were completely undone in the next page. The common user is served better by simple instructions presented one at a time. The Adobe Gamma instruction dialog seems very simple, but still has three instructions for the user. The dialog could best be split in two steps: First, set the contrast control to its highest setting. Then click OK, and in the next dialog adjust the brightness setting with the test stimulus. It is possible that the performance could be improved if the gray in the test stimulus had a lower value. This might even make the additional requirement “while keeping the frame a bright white,” redundant.

The black-level matching instruction to make a gray patch nearly but not quite black is not working for the average untrained user. Often there is no reference black available for the display screen and the user is asked to evaluate brightness differences, which is more difficult than matching brightness. The procedure used by the European Broadcasting Union (EBU) for black-level setting of television sets [EBU97] seems to have a better matching procedure. The picture line-up generating equipment (PLUGE) signal has fields with a +2% and –2% voltage value, and the black level can then be positioned directly in between. Roberts [Rob93] assumes the margin might not be not accurate enough though.

The on-screen settings dialogs should be adapted: the white background is very disturbing when setting the black level, the position in the middle of the screen takes the view from the relevant stimuli, and for the CRT in the test the allowed period of inactivity of just 3 sec before the dialog disappears leads to many errors.

7 Visual gamma measurement and methods to compare gamma models

Abstract

The correct estimation of the gamma exponent describing the tone-reproduction curve of a display is an important step in color management. Several methods for visual gamma estimation have been proposed. In this study, the theoretical merits and practical problems of a number of these methods are discussed and compared, and improvements are suggested. A new method to compare gamma models with different numbers of parameters is introduced. In an experiment, spatial and temporal brightness-matching methods were tested with 32 untrained subjects working on a CRT and an LCD with different resolutions under office and low-illumination conditions. Illumination had no effect on gamma estimations. Subjects had great difficulties with spatial brightness matching at low resolutions. Temporal and spatial visual brightness matching for untrained subjects showed a larger gamma than photometric fits.

7.1 Introduction

In display-standards and color-management systems, the relation between the input voltage and the light output of a primary-color channel is described by the tone reproduction curve (TRC). The main parameter in the function describing the TRC usually is gamma, the exponent of the power function defining the transmission for the cathode-ray tube (CRT). Through its years of display-market domination, the CRT has defined the standard and the TRC with power function form has its merits. The non-linearity of the CRT is not a deficiency but a highly desirable feature: it is very nearly the inverse of the lightness sensitivity of human vision. The non-linear TRC causes a CRT’s response to be roughly perceptually uniform, this means the coding of the images during storage and transmission is nearly optimal and therefore TRCs of other display technologies are adjusted to mimic this behavior. Another advantage of the power function is that the transmission is defined by a single parameter: the exponent gamma. A higher gamma means the image has relatively darker midtones and looks to have more contrast. A lower gamma means relatively brighter midtones, more details might be visible in the darker regions of the image, but the image may appear washed out.

In practice, the single parameter advantage is undermined by confusion with inverse gamma functions, different ideal gamma settings for different systems or system parts (Windows PC 2.2, Apple 1.8, TV broadcasting 2.5), different viewing circumstances (higher ambient illuminance means lower ideal gamma), and black-level offsets introducing extra parameters to the function.

CRT gamma is not a constant, and even if it would be a constant, then a small voltage offset would have about the same effect as a change in gamma. For an LCD, the native TRC must be transformed to behave like a gamma function. Thus, for an average computer display some deviation from the intended gamma in the default sRGB settings should be expected. Measurement of the TRC or estimation of the display gamma then is the first essential step to allow color-management applications to correct the TRC. For the average user without photometric measurement equipment, gamma brightness matching is the only method available to obtain an estimate of the TRC.

It is unlikely that the general public will take the time to try to understand all the problems related with the gamma function. Most of these problems can be solved and hidden for the user with appropriate software. But in order to display images the way they are meant to be displayed, a characterization of the actual TRC of the display system is needed regularly. If the user does not have a photometric device, there are visual measurement methods to attain at least a rough characterization. These methods usually consist of two stages: the correct setting of the display black level and a brightness-matching stage to determine a gamma estimate. In the first part of an experiment described in an earlier paper, [Bes06] the ability of users to correctly adjust the black level was tested. The results with a simple as well as with an elaborate off-the-shelf method were rather poor. In the second part of the experiment, reported here, the capabilities of the user and the available visual-matching methods to provide a correct estimation of the TRC based on a gamma function model are evaluated.

7.2 Visual gamma-matching theory

In a gamma brightness-matching task, subjects were asked to adjust the brightness of a uniform color field to that of an adjoining field composed of two differing values of the same color. In the simplest form, this test is performed for white (equal R, G, and B) only. If the gammas of the primary colors differ, better results can be obtained by gamma matching for the separate primary colors (red, green, and blue).

In the temporal gamma brightness-matching task, the differing values are displayed alternating in the field at half the frame rate of the display, intended to produce an image of the average luminance, which should be matched by the subject by adjusting the brightness of the uniform color field. Flicker may be perceived in the alternating field and jitter on the edge of the alternating and uniform fields, especially when the luminance difference in the alternating field is high.

In the spatial gamma brightness-matching task, the differing values are displayed next to each other, usually as alternating horizontal lines at half the resolution of the display to avoid the effects of inter-pixel dependency. For a better matching result, the viewing distance should be extended to enable the user to diffuse the pattern of horizontal lines into the average luminance. Some alignment artifacts may result on the divide of the fields.

Table 7.1. Gamma estimates by code value from single gamma brightness match.

code	182	183	184	185	186	187	188	189	190	191	192	193	194
gamma	2.055	2.089	2.124	2.160	2.197	2.235	2.274	2.314	2.356	2.399	2.443	2.488	2.535

An alternative approach is to display several fixed combinations of uniform and alternating fields representing different gamma matches and let the user choose the best match. In Table 7.1, the relation between the matched code value from a match with an alternating field with minimum (0) and maximum (255) code values for an ordinary PC graphics card and the resulting gamma is shown.

It is difficult to implement the temporal brightness-matching task so that it works on all combinations of graphic cards, displays, and operating system installations for Windows PCs. Most commercially available applications therefore use forms of spatial brightness matching.

The simple gamma-matching applications, such as Adobe Gamma, provide a single brightness match to determine a single gamma parameter for white or the separate red, green, and blue channels. With the likeliness of substantial black level offset, it is doubtful if a single parameter is sufficient to accurately describe the TRC. The best-performing gamma models used in display characterization with luminance measurements have 3–4 parameters.

Berns [BMG93] developed the Gain-Offset-Gamma (GOG) model, later refined by Katoh, Deguchi, and Berns [KDB01a; KDB01b] in the Gain-Offset-Gamma-Offset (GOGO) model, which characterizes the amount of light Y_P generated by the CRT gun of primary P (which could be R, G, or B) as a function of the digital counts or code values in the DAC (d, d_{max}) in three parameters: gamma (γ_P), the normalized gain/offset parameter $k_{g,P}/k_{0,P}$, and the luminance offset $Y_{0,P}$.

$$\begin{array}{ll} Y_P(d) = \left(Y_{\max,P} - Y_{0,P}\right)\left(k_{g,P}\dfrac{d}{d_{\max}} + k_{0,P}\right)^{g_P} + Y_{0,P}, & \left(k_{g,P}\dfrac{d}{d_{\max}} + k_{0,P}\right) \geq 0 \\ Y_P(d) = Y_{0,P}, & \left(k_{g,P}\dfrac{d}{d_{\max}} + k_{0,P}\right) < 0 \end{array} \qquad (7.1)$$

with:

$$k_{g,P} = 1 - k_{0,P} \qquad (7.2)$$

Gamma (γ_P) is assumed constant under all conditions. Clipping occurs when the normalized offset parameter $k_{0,P} < -k_{g,P}.d/d_{max}$ and several code values are mapped to the same luminance output $Y_{0,P}$. The luminance offset $Y_{0,P}$, if there is one, will originate from sources other than the primary channel P, usually from a positive k_0 of another primary channel.

The biggest problem of visual gamma matching is that it produces luminance ratios of code value triplets as a result and not absolute luminance values for a single code value as in photometric measurements. This makes it impossible to directly relate output results to each other, which makes it difficult to apply the standard photometric measurement fitting procedures to the results. The luminance offset parameter Y_0 cannot be fitted with this procedure. To adequately fit the remaining two or three parameters, two, three, or more different matching results are needed.

7.2.1 Previous research

In literature, several methods for TRC characterization with visual brightness matching are proposed. Colombo and Derrington [CD01] used fields alternating spatially as well as temporally to visually calibrate a CRT monitor with a specially tuned graphics card providing 1024 monochrome gray levels. Their procedure provides nine gamma brightness matches, which do not cover the lower third of the code value range. They also see luminance saturation effects as a major problem in characterizing the CRT, and therefore reject the gamma function for fitting purposes. This makes their gamma brightness-matching method unsuited for recalibrating computer displays to sRGB.

Daly and Lee [DL98] use a model similar to the GOGO-model:

$$Y_P(d) = (a_P d + b_P)^{g_P} + f_P = a_P{}^{g_P}(d + \frac{b_P}{a_P})^{g_P} + f_P \qquad (7.3)$$

with P denoting the primary color, and claim that all parameters in this equation can be visually measured, starting with

$$\frac{b_P}{a_P},$$

then γ_P, α_P, and finally φ_P.

They claim the offset β_P to be usually negative and in that case the offset can be visually measured by detecting the maximum code value that does not differ from code value 0. In reality, chances are equal so that the offset is positive and in that case β_P cannot be measured and neither can all the subsequent parameters.

If β_P is known, a series of gamma-matching patterns can be made from which the optimum gamma match can be chosen. The external flare parameter φ_P can then be determined by repeating the first visual measurement in a dark room. Finally, the authors admit that an absolute value for α_P cannot be determined by simple visual-matching techniques, but it is possible to estimate the ratios of α_P for the three primary colors from a visual white-point measurement.

In PsychToolBox, [BPI02] a special method was devised to generate stimuli and fit the gamma function to the matched code values. It is based on a recursive method starting with the anchors of code values 0 and 255 (normalized to 1.0) and the value matching the averaged luminance (brightness) of these two. The value of the first match V(0.5) is combined in the next two matches with the known values 0.0 and 1.0 to produce V(0.25) and V(0.75), respectively. These two values can then be used to produce matches with combinations with the other known values and so forth. The emphasis in this procedure should be lying on the lower region of the scale because this provides a better distribution of the resulting code values over the range for a gamma near the sRGB standard value of 2.2.

The advantage of this method is that it links input (code values) with output (fraction of the total luminance) and the results can be easily applied in standard photometric fitting procedures minimizing the error in the CIE lightness domain taking the maximum luminance as reference white.

A disadvantage is the possibility of a propagating error in the procedure. The error in the first match caused by either the user or the quantization error has its impact on the next matches and could accumulate with each match. This can easily be illustrated with the case where gamma = 1.0. If the user would make no matching error but would consequently select the lower value for equally distant choices, the matching sequence could look like this: V(0.5) = 127, V(0.25) = 63, V(0.125) = 31, and V(0.0625) = 15. And 15/255 = 0.0588 which constitutes an error of nearly 6%.

A different approach is taken by Miller and Yang [MY03], they generate a series of unrelated code value couples for matches. The results are therefore not directly related to the scale of possible output values. This calls for another type of optimization function. Miller and Yang use the following gamma model, comparable to the GOGO-model [Eq. (7.1)] and the model of Daly and Lee [Eq. (7.3)]:

$$Y(d) = a(d+b)^g + f \tag{7.4}$$

with:

$$b = d_{\max}\left(1-k_g\right)/k_g = d_{\max}k_0/\left(1-k_0\right) \tag{7.5}$$

where f represents only the ambient light that is reflected from the display and not any internal luminance offset. They view γ as the slope of a linear function, relating code value to luminance when the function is plotted in log–log space.

For the virtual match, the following equation is valid:

$$(d_M+b)^g = 0.5(d_H+b)^g + 0.5(d_L+b)^g\ , \tag{7.6}$$

with d_H and d_L the high and low alternating values and d_M the matched value, α and *f* falling out of the equation. Miller and Yang argue that by applying two or more matches with different d_H and d_L the unknown β and γ can be solved by minimizing the error function:

$$E = \sum_{i=0}^{n} E(i)^2$$

with

$$E(i) = \log[\frac{(d_H(i) + b)^g + (d_L(i) + b)^g}{2(d_M(i) + b)^g}] \qquad (7.7)$$

However, in an experiment where the actual β obtained from photometric data ranged from 3.3 to 24.9, they found values with this optimizing method for the visual β obtained by temporal brightness matching, which range from 1.2 to 1.4.

It is also remarkable that all the values they obtained for β based on photometric data were positive, this means that there was luminance for code value 0 in all conditions and that clipping never occurred, not even for brightness settings of only 15%. (It also contradicts the assumption of Daly and Lee mentioned above that β is usually negative). It is very likely that the error function used in Miller and Yang’s fitting procedure was not accommodated to handle negative values of β. The discrepancies between β-values from fits on photometric and matching data in their results do not occur with the implementation of their proposed optimization in the log code value domain used in this paper, with negative β-values resulting almost as likely as positive values.

In an experimental procedure aimed at visually estimating the white point, Wen and Wu [WW04] used a single visual match to determine γ with:

$$g = -\frac{\ln 2}{\ln(d_M / d_{\max})} \qquad (7.8)$$

Strange enough, their results show only four gamma values: 2.24, 2.32, 2.40, and 2.48, which are relatively far apart if compared with the possible results from a normal PC graphics system with 0–255 as possible values for each color (see Table 7.1).

7.2.2 Improved method for visual gamma-measurement model fitting

Based on these findings and the need to incorporate a parameter to describe black-level offset in the model, a new simple error function is introduced for the fitting of the gamma model parameters γ and β (and therefore k_0) to the code values resulting from multiple visual brightness matching results. The error function operates in the code value domain and can produce reliable estimations, without accumulative errors and the hazards of the error prone log–log domain. The difference between the matched code value d_M/d_{max} and the fitted code value is used as error function:

$$E(i) = \frac{d_M(i)}{d_{\max}} - \{[(\frac{d_H(i)}{d_{\max}} + b)^g + (\frac{d_L(i)}{d_{\max}} + b)^g]/2\}^{1/g} - b \qquad (9)$$

Optimizing in the code-value domain has certain advantages in the case of gamma brightness matching. All results from the matches are code values, so domain conversions with the inevitable errors are not needed, and the scaling of the code-value domain for the expected gamma values is close to the scaling in the perceptual CIE lightness domain.

The three two-parameter model estimators: Miller and Yang’s optimization in the logarithmic code value domain, the Psychtoolbox method optimized in the CIE lightness domain, and the

Table 7.2. Comparison between code value error and logarithmic code value error optimization of gamma brightness matching data for simulated data and photometric measurement data with optimization in CIELAB Lightness domain according to the GOG-gamma model with normalized voltage offset k_0 [Eq. (7.1)].

simulated		code value		Psychtool		Miller & Yang	
gamma	k_0	gamma	k_0	gamma	k_0	gamma	k_0
2.2	0.000	2.187	-0.004	2.184	-0.004	2.167	-0.007
2.2	0.020	2.228	0.026	2.222	0.023	2.245	0.027
2.2	0.100	2.219	0.107	2.103	0.063	2.239	0.112
2.2	-0.053	2.226	-0.047	2.221	-0.049	2.246	-0.043
2.4	0.052	2.431	0.060	2.397	0.050	2.436	0.061
1.959	0.000	1.951	-0.003	1.952	-0.003	1.950	-0.003
measured							
2.267	0.020	2.262	0.021	2.259	0.020	2.280	0.025
2.171	0.011	2.157	0.010	2.162	0.010	2.160	0.010
2.172	-0.003	2.209	0.011	2.208	0.009	2.319	0.032
2.184	0.014	2.183	0.017	2.185	0.017	2.233	0.027

straight optimization in the code value domain, were compared for six cases of simulated parameter data and four cases of photometric measured and fitted parameter data. Each data set was produced by eight matches on luminance data according to the scheme in Table 7.2, for the Psychtoolbox method the data point (255, 1.0) was added to the series.

The results represented in Table 7.2 show that all three methods deliver reasonably good parameter estimations. Comparison reveals that the Psychtoolbox method and the straight code-value optimization have a smaller error in the γ and k_0 parameter estimation than the logarithmic optimization of Miller and Yang in nine out of ten cases. The Psychtoolbox method performs slightly better in most cases than the code-value optimization but has a significantly larger error for one case of simulated data ($\gamma = 2.2$, $k_0 = 0.1$).

A closer look at the data revealed that this was a typical case of the accumulating quantization error signaled above: six of the eight luminance matches were rounded downwards. Because of its consistent performance, the straight code-value optimization was used for all two-parameter model estimations from multiple brightness matches in the experiment, described later in this paper.

7.2.3 A method for comparing gamma models with different parameters

There are gamma models with one (gamma), two (gamma and offset/gain), three (gamma, offset/gain, luminance offset), and four (gamma, offset/gain, luminance offset, gamma range) parameters. How to compare these models? The sRGB standard has two different gamma model definitions. The ideal TRC is a one parameter model with $\gamma = 2.2$:

$$Y(d) = Y_{\max}\left(\frac{d}{d_{\max}}\right)^{2.2} \tag{7.10}$$

For the encoding specification of the transfer function a two-parameter gamma model with a linear portion integrated at the dark end of the signal is supplied to optimize encoding implementations. According to the standard the equations below closely fit Eq. (7.10). This should maintain consistency with the legacy of desktop and video images.

$$Y(d) = Y_{\max} \frac{d}{d_{\max}} / 12.92 \qquad for \frac{d}{d_{\max}} \leq 0.04045 \tag{7.11a}$$

$$Y(d) = Y_{\max} \left(\frac{(\frac{d}{d_{\max}} + 0.055)}{1.055} \right)^{2.4} \qquad for \frac{d}{d_{\max}} > 0.04045. \tag{7.11b}$$

The offset value of 0.055 corresponds with k_0 = 0.052 and k_g = 0.948. In the sRGB standard it is assumed that the gamma models with γ = 2.2 and k_0 = 0.0 and with γ = 2.4 and k_0 = 0.052 are in practice equivalent. To test this assumption a method was developed to reduce gamma models with more than one parameter to a single-parameter gamma model. By making a fit for a two-parameter gamma model to data generated with a single parameter gamma function, the optimal k_0 with alternative gammas can be found. Based on a number of these fits shown in Fig. 7.1, a general comparison between different models can be made.

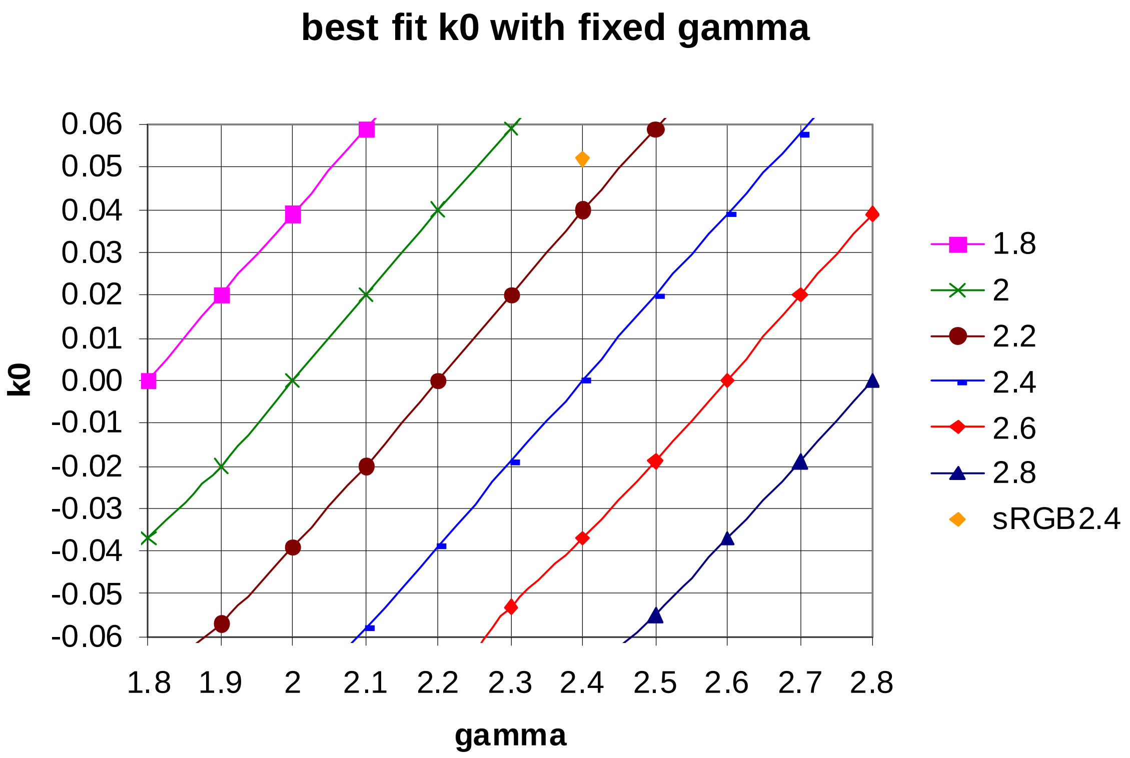


Figure 7.1. Best alternative fit in the CIELAB lightness domain for ideal gamma functions with two parameter fixed gamma optimization. The colored lines represent the results obtained with the ideal single gamma model data with the gamma value shown in the legend. The markers show the optimal normalized offset (and therefore gain) parameter k_0 [from Eq. (7.1)] setting on the vertical axis for a fit to the data with the fixed gamma on the horizontal axis.

The colored lines in the figure represent the results obtained with the ideal single-gamma model data with the gamma value shown in the legend. The markers show the optimal normalized offset (and therefore gain) parameter setting on the vertical axis for a fit to the data with the fixed gamma on the horizontal axis. The differences between the data and the fits on the lines are considerably smaller than for any other point in the plane lying on the same distance from the ideal point (where $k_0 = 0.0$) but not in the proximity of the ideal line, *e.g.*, the CIELAB lightness RMSE for a 17-points curve:

$$CIEL\ RMSE = \sqrt{\frac{\sum_{d=0,15,31,47,..}^{d<255} \left(L^*(d)_{est} - L^*(d)_{ideal}\right)^2}{16}} \qquad (7.12)$$

$$\text{with } L^*(255)_{est} = L^*(255)_{ideal}$$

For $\gamma = 2.4$ and $k_0 = 0.040$ (on the line), CIEL RMSE is 0.88, and for $\gamma = 2.4$ and $k_0 = -0.030$ it is 4.30, and for $\gamma = 2.3$ and $k_0 = 0.000$ the error is 1.34. A comparison of the normalized TRCs for the related ideal single-parameter gamma and the two-parameter equivalents on the line show that for a positive offset the luminance will be lower for low code values and higher for high code values. For a negative offset the luminance will be higher for low code values and lower for high code values. The relation between the fixed gamma γ_2 (*x* axis) and the offset k_0 (*y* axis) from the two parameter model and the gamma γ_1 (lines) from the one parameter model for $1.8 < \gamma_1 < 2.8$ and $\gamma_1 - 0.4 < \gamma_2 < \gamma_1 + 0.4$ can be estimated by:

$$g_1 \approx g_2 - c\,k_0 \qquad (7.13)$$

with $c \approx 5$ if no luminance offset Y_0 (not related to the voltage offset k_0) is present in the measured data. If there is a luminance offset Y_0 present, then *c* will be smaller, the lines in the figure will be steeper, and a larger absolute k_0 will be needed to compensate for the error in the fixed γ_2, especially if γ_1 is nearer 1.8.

7.3 Gamma brightness-matching experiment

In the gamma brightness-matching experiment, spatial stimuli with three resolutions and temporal stimuli were compared for the primary colors and white on a CRT and LCD. The experiment was setup to evaluate practical application with untrained users. The subjects were allowed to adjust the tilt of the display and no form of head fixation was used. Half the subjects performed the brightness matching tasks in low ambient illuminance, the other half in office lighting conditions. The results of the experiment allowed a comparison of single and multiple brightness match gamma estimation methods.

7.3.1 Participants

The participants were psychology students from the University of Twente, participating as part of their first-year curriculum. Thirty-two students participated, six students had the Dutch nationality and 26 had German nationality with a sufficient knowledge of the Dutch language, 24 students were female and 8 were male, ages ranged from 17 to 26. All subjects had corrected or uncorrected 20/20 or better visual acuity for stereoscopic vision measured with the TopCon Screenoscope-II, Tokyo Optical Co. Ltd. All subjects were tested for color-vision deficiencies with the Ishihara test for color blindness; one male subject had a weak red-green color deficiency. This was judged to have no consequences for the monochromatic brightness-matching tasks in this experiment.

Table 7.3. sRGB standard and experiment display characteristics for default sRGB settings.

	sRGB standard			CRT			LCD		
black level	0			0.340			0.459		
max luminance	80			92.8			154.4		
Correl. Color Temp	6500K			7217K			6319K		
gamma	2.2 (2.4)			2.171-2.267			1.608-1.816		
	CIE x	CIE y	rel. lum.	CIE x	CIE y	rel. lum.	CIE x	CIE y	rel. lum.
W	.313	.329	1.000	.301	.325	1.000	.315	.339	1.000
R	.640	.330	.2127	.623	.332	.2371	.638	.342	.2378
G	.300	.600	.7152	.295	.601	.6681	.295	.614	.6764
B	.150	.060	.0722	.149	.077	.0948	.147	.072	.0858

7.3.2 Apparatus

A CRT and an LCD connected to two separate Windows-operated personal computers were used in the experiment. The CRT was a 2-year-old Philips 17-in. model 107T5 controlled by an Intel Extreme Graphics 2 graphics card at a 800 × 600 resolution at a frame rate of 100 Hz. The LCD was a Philips 17-in. model 170B4MG02, also 2 years old, controlled by an NVIDIA GeForce2 MX 100/200 graphics card at the native 1280 × 1024 resolution at 60 Hz. Both displays were connected to the graphics card with a D-SUB cable. Care was taken that no color management was used.

For both displays, the luminances of all 256 code values for the separate (red, green, and blue) and combined (white) primary channels were measured in the sRGB-setting for rectangular stimulus patches of 1/25 of the active area on a black background with a L203 photometer with a 6° field-of-view luminance probe from Macam Photometrics, Ltd. Ambient illumination was excluded by a dark gray foam cover and care was taken that no pressure was exerted on the LCD screen.

A series of 17 samples (0, 15, 31, ..., 239, 255) from the measured data were fitted to the GOGO-model [Eq. (7.1)], minimizing the mean square error in the CIE-lightness domain with maximum white luminance as reference white. The three-parameter GOGO model was chosen because of the apparent luminance offset in both CRT and LCD. The four-parameter gliding gamma model which gave better fits for a number of CRTs in an earlier study [Bes05] did not produce significantly better results for the sRGB setting of the CRT in this study. Therefore the model with the smaller number of parameters was preferred. The results are shown in comparison with the sRGB standard values in Table 7.3.

Table 7.4. Best fits for GOGO-model by minimizing CIE lightness error for 17 equidistant samples for sRGB display settings for CRT and LCD in experiment. CLRMSE is the CIELAB Lightness Root Mean Square Error per sample. Note that the fitted luminance offsets are not equal to the black levels in table 7.3, mainly because the displays in this study did not fully adhere to the GOGO-model.

	CRT				LCD			
	gamma	k_0	Y_0	*CLRMSE*	gamma	k_0	Y_0	*CLRMSE*
red	2.267	0.020	.344	.0440	1.745	-0.101	.595	.5619
green	2.171	0.011	.360	.1462	1.675	-0.090	.672	1.1339
blue	2.172	-0.003	.354	.0657	1.608	-0.143	.531	.2305
white	2.184	0.014	.362	.6880	1.816	-0.078	.656	1.0749

The tone-reproduction curve and the values for maximum luminance, black level, and white point of the default sRGB setting of the displays used in the experiment deviated significantly from those prescribed in the standard. There is no measure defined in the standard to evaluate the difference from the standard.

The main problem with the default sRGB settings of the displays lies in the black-level offset. Both the CRT and LCD have black levels deviating about 3 ΔE (CIELAB) from absolute black. The CRT has some additional problems with the color setting: the color saturation of the red and blue phosphors seems to be insufficient. The LCD blue is also lacking in saturation, but the LCD's color temperature is much closer to the target.

The CRT's tone-reproduction curves can be adequately described by the GOGO model (Table 7.4), although the source of the luminance offset cannot be explained from the model. The gamma function seems not fit to describe the LCD's TRC due to the luminance saturation effects for high values. The differences in CIELAB lightness of the sRGB TRC assuming equal white luminance and the fitted model with the measured data are shown in Fig. 7.2 for the CRT and in Fig. 7.3 for the LCD. The figures show that the effect of black-level offset is relatively large for the CRT compared to the deviations from the ideal gamma curve because for the LCD the deviations from the gamma curve have a larger effect on luminance and color.

A better result for the gamma fitting on the LCD could be reached if the curve is not fitted through the maximum luminance value. In that case, the fit would project a higher maximum luminance, but the overall error of the fit would decrease, *i.e.*, the solid-line magenta curve in Fig. 7.3 would be flatter and closer to the x-axis, but with a sharp negative dip at 255. This might explain that the impact on the perceived image quality of the display is not so severe as might be expected from the error curve in Fig. 7.3. In practice, however, the use of a model with an error in the maximum luminance could lead to practical problems such as an error in the position of the white point.

7.3.3 Lighting conditions

Half the subjects performed the experiment in office-lighting conditions provided by the standard experiment room fluorescent-tube lighting. The other half worked under low-illumination conditions provided by a small incandescent lamp reflected from the back wall. For the office condition, the illuminance on the desktop measured with a Macam L203 photometer with illuminance probe was 530 lux; on the CRT, 266 lux; and on the LCD, 296 lux. For the low light conditions, the illuminance on the desktop was 38.1 lux; on the CRT, 12.6 lux; and on the LCD, 13.1 lux.

7.3.4 Stimuli and instructions

All subjects performed two gamma brightness-matching tasks on each display with the built-in sRGB contrast and brightness settings, and with the settings resulting from the brightness and contrast setting task, where the subjects tried to set the optimum black level. Because of the large variation in optimum black-level settings, only the gamma brightness-matching results for the sRGB settings are evaluated. In the gamma brightness-matching task, subjects were asked to adjust the brightness of a uniform color field to that of a field composed of two differing values of the same color. In the temporal brightness-matching task, the differing values were displayed alternating in the top half of a square at a 100-Hz frame rate intended to produce a 50-Hz image of the average value (Fig. 7.4). The square was placed in the middle

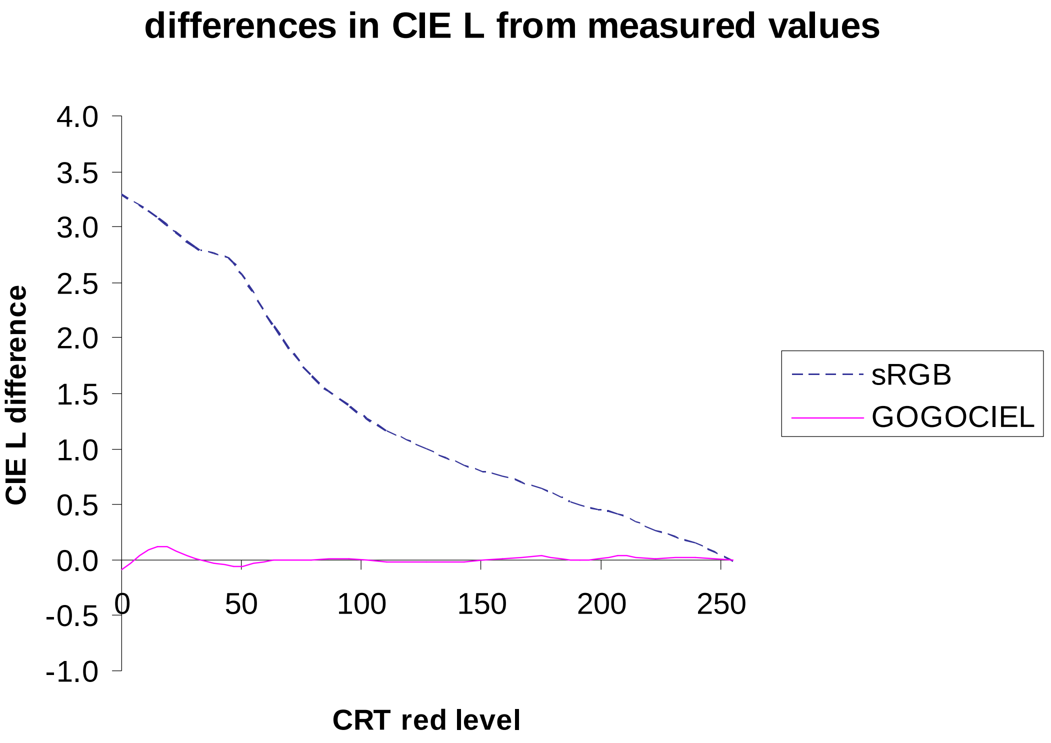


Figure 7.2. CIELAB Lightness differences from measured CRT Red values with sRGB and GOGO CIELAB Lightness optimized model fits.

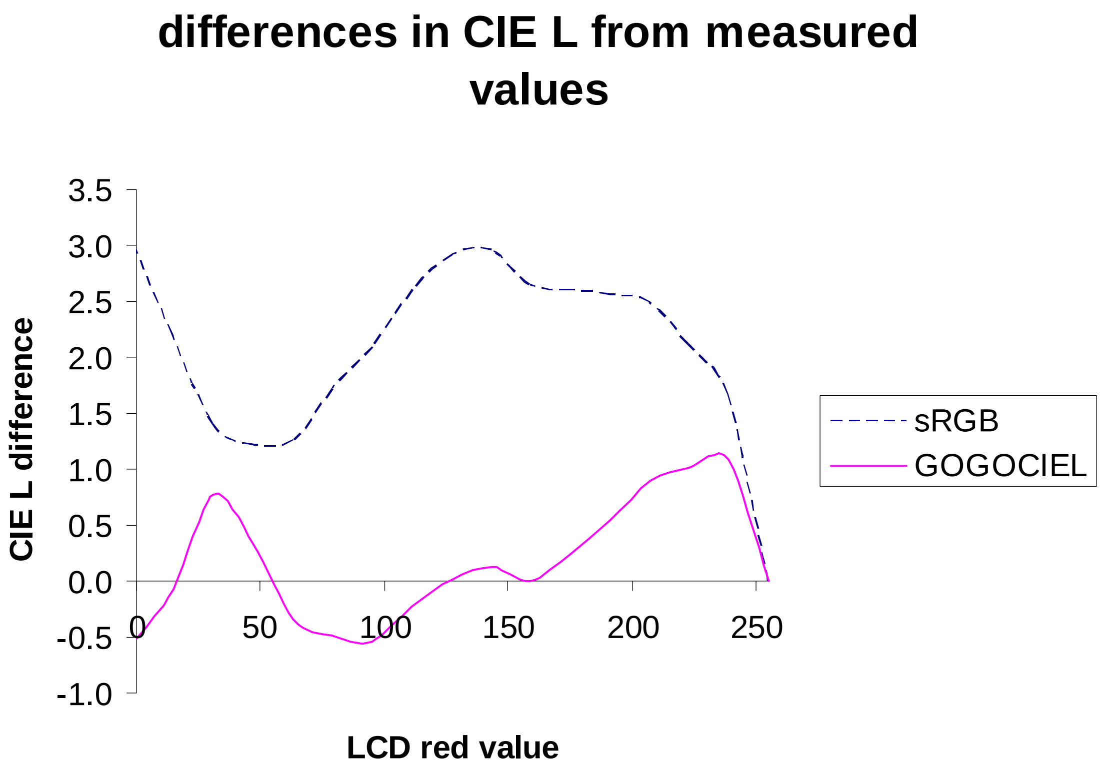


Figure 7.3. CIE Lightness differences from measured LCD Red values with sRGB and GOGO CIE Lightness optimized model fits

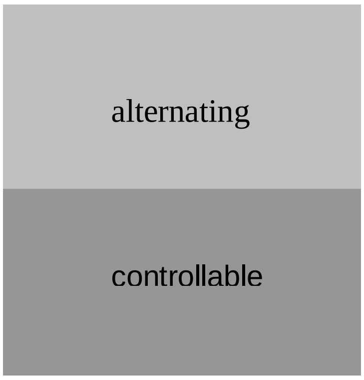


Figure 7.4. Layout of the stimulus in the temporal visual gamma-matching task.

of the screen and of a size one third of the height and one quarter of the width of the screen; the bottom half of the square was the uniform field. For the background, a dark gray was tried out instead of black to reduce the effect of after-images. This did not significantly reduce the afterimages and introduced new artifacts at the field/background separations and the possibility of inverted contrast of stimuli with background.

The use of the temporal gamma brightness-matching task poses several practical problems. The task cannot be used on LCDs. The rise and decay times, *i.e.*, the time it takes to switch a pixel from black to white and back, of LCDs in general, and also for the LCD used in this experiment, are too long. This generally results in a lower than average luminance for the value displayed at half the frame rate.

The task cannot be used in combination with normal (1024 × 768 or 1280 × 1024) and higher display addressability for average graphics cards and CRT displays. For the experimental setup with about 2-year-old off-the-shelf equipment, a 100-Hz frame rate could only be produced with 800 × 600 addressability. Even then, flicker is still visible for most people and a nuisance for some. In the experimental setup, some line flicker was also visible at the border of the alternating and uniform field.

The implementation of the task for Windows-operated PCs is not straightforward. For a stable image, not regarding the 50-Hz flicker, the alternating fields should be written in the vertical retrace time of the frame. Yet there is no overall reliable method to detect the beginning of the vertical retrace period for all the different graphics cards and setups in Windows PCs. Methods that work on one PC can hold operations on another PC indefinitely.

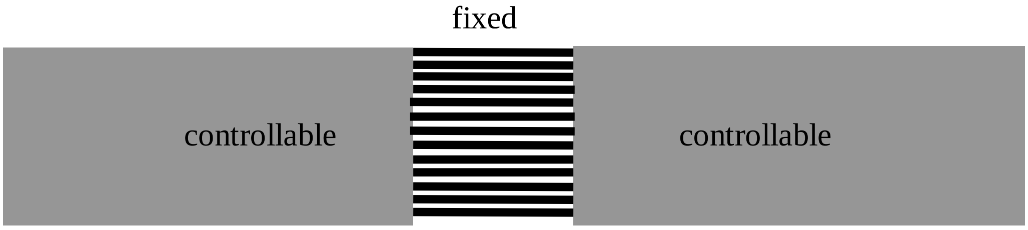


Figure 7.5. Layout of the stimulus in the spatial visual gamma-matching task.

Table 7.5. Spatial and temporal characteristics of the visual gamma-matching tasks on both displays.

condition	viewing distance in cm	approximate angular line distance	frame rate
CRT temporal	50	n.a.	50 Hz
CRT spatial	50	2'47"	100 Hz
LCD spatial direct	50	1'49"	60 Hz
LCD spatial mirror	150	0'36"	60 Hz

In the spatial brightness-matching task, the differing values were produced as alternating horizontal lines in a square of 101 pixels in the middle of the screen. This square was placed in a horizontal band of uniform adaptable brightness of the same height over the whole width of the screen (see Fig. 7.5). This form was chosen to minimize the effects of viewing-angle dependency on the LCD screen. On the LCD, the default addressability of 1280 × 1024 with a 60-Hz frame rate was used, on the CRT the same settings were used as in the temporal matching task to avoid complex display setup operations during the experiment.

On both the CRT and the LCD, the angular distance between lines at normal viewing distance was higher than one arcmin (Table 7.5), and hence the viewers were expected to resolve the individual lines. For that reason an extra spatial task was performed on the LCD with an enlarged viewing distance. Because the experimentation room was rather small the enlarged viewing distance was realized through a mirror on the back wall with the display averted from the user. In the latter condition, the viewing distance amounted to about three times the normal viewing distance of about 50 cm. In combination with the different addressabilities on the CRT and LCD, this produced three levels of resolution for the spatial brightness-matching task (Table 7.5).

For both the temporal and the spatial brightness-matching task, the luminance of the controllable fields on the display could be adjusted by the up and down arrow keys on the keyboard, the best match was to be confirmed with the “Enter”-key. In the experiment, each of the 32 subjects performed eight brightness matches for each of the four colors (red, green, blue, and white) for each of the four conditions in Table 7.5 for the default display sRGB setting and with the brightness and contrast control settings of their own black-level optimization (see Chapter 6). There were no trial sessions as the experiment was meant to study the behavior of untrained users. The participants could complete the task with a total of 256 brightness matches at their own pace.

The order of the conditions (Table 7.6) was balanced over subjects with the restriction that conditions on one display with the same brightness and contrast settings were performed in one block, always followed by a similar block of conditions on the other display. There were only two possible orders of colors per display condition: Either W, R, G, and B, or B, G, R, and W. Because some subjects had severe difficulties in fusing the striped spatial brightness matching stimuli, especially for the most luminous colors W and G for the direct CRT and

Table 7.6. Experimental conditions for visual gamma-matching experiment.

	CRT		LCD	
	default sRGB	user black level	default sRGB	user black level
temporal	B, G, R, W	B, G, R, W		
spatial direct	B, G, R, W	B, G, R, W	B, G, R, W	B, G, R, W
spatial mirror			B, G, R, W	B, G, R, W

Table 7.7. Sequential matches in a visual gamma-matching task per color, display and matching method.

no. of match	low value	high value	matching result	ideal norm. lum.
1	0	255	d1	0.5
2	0	d1	d2	0.25
3	0	d2	d3	0.125
4	0	d3	d4	0.0625
5	0	d4	d5	0.03125
6	0	d5	d6	0.015625
7	d2	d1	d7	0.375
8	d1	255	d8	0.75

LCD viewing conditions, the color order was fixed to B, G, R, and W for the second half of the experiment. It was thought that subjects could get easier accustomed to fusing the dark and bright stripes with the less luminous blue stimuli for which the visual acuity is lower.

The participants were handed written general instructions about the design, course, and purpose of the experiment, and specific task instructions before the first temporal or spatial brightness task started. They were instructed to view the alternating field of the stimulus as having a uniform brightness and adjust the controllable field to the same brightness as accurately as possible.

The order and code-value pairing of the matches for each color is given in Table 7.7. This order was chosen to be able to compare the ideal gamma function single match method for the first match based on Eq. (7.8) and the two-parameter optimization in the code value domain presented in this paper [Eq. (7.9)]. The combinations of code values in the matches were chosen to get a good distribution over the code value and CIE lightness domain for the expected values of gamma.

7.3.5 Hypotheses

Based on the literature overview and the design of the experiment, the following hypotheses were expected to be confirmed by the experiment for the reasons specified.

Preference

1. Subjects should least prefer the two direct spatial matching conditions. With the two direct spatial matching conditions, subjects should have trouble fusing the high and low values in the striped part of the stimulus because the resolution is too low.

Ambient illumination

2. Ambient illumination should not have an effect on the gamma matching results. The luminance resulting from ambient illumination is added to the whole of the screen, so it will not affect the difference in luminance between the alternating and the adjustable field. It might have a small effect on the perceived lightness differences: more average luminance means smaller lightness differences, but this effect should be too small to be of any influence.

Display

3. The TRCs resulting from the gamma brightness matching should adhere better to the gamma power function model for the CRT as for the LCD. The maximum code value of 255 is part of two of the eight brightness matches and the LCD luminance for this code value does not adhere to the gamma model (Fig. 7.3).

4. The gamma resulting from the single brightness match method for the LCD should be too low (see above).

Type of stimulus

5. The results for the LCD with mirror and CRT temporal matching conditions should have less variance than the direct spatial matching conditions (see 1).

Methodology

6. Gamma estimation based on eight brightness matches should be more reliable than those based on the single (0, 255) match. More data should provide a better fit with the possibility to fit offset as well as gamma.

Color and White

7. The gammas for the primary colors of a display should not be equal. The chances that the conversion from voltage to luminance works exactly the same for all three primary colors are very low, especially if the display has been accommodated with different white point settings, which affect the ratio of the primary luminances.

8. The variance for white gamma based on the single (0, 255) match should be larger than for the primary colors. The different gammas of the primary colors should cause color differences between white and gray.

7.4 Results

7.4.1 General

A number of problems became apparent in the course of the experiment. The choice for the use of a mirror to prolong the viewing distance for the second LCD condition in the small experiment room was rather unlucky. The viewing angle for the mirrored LCD condition was minimally vertically 17° off normal, which resulted in considerable differences in brightness and gamma compared to the normal viewing direction.

Subjects had far greater problems in fusing the alternating spatial stimuli, especially for the most luminous white and green colors, as expected. Two subjects caused the experimentation program to crash repeatedly for the direct-view spatial brightness matches on CRT as well as LCD by producing matching code values equal or above the highest code value in the alternating field. These subjects were substituted by other subjects to keep the balanced design intact. However, the results for the 32 subjects who completed the white and green spatial conditions on the CRT were dissimilar to such an extent that they could not be used in statistical comparisons.

One subject’s progress was so slow that completion of the session would have disrupted the experiment schedule: the session was aborted and repeated later with another subject.

7.4.2 Preference

Thirty-one of the subjects were asked which of the four gamma matching methods (see Table 7.5) they preferred: 17 had a preference for the temporal matching method, 11 preferred the spatial matching with mirror on the LCD, and three subjects preferred the spatial matching with direct view on the LCD. The subjects preferences were converted to ranks and analyzed with a Wilcoxon signed-rank test, which revealed significant ($p < 0.05$) differences in preference between the favored methods of temporal matching on the CRT and spatial matching with mirror on the LCD and the lower rated methods of direct spatial matching on

Table 7.8. Overview of results from gamma estimation based on a single brightness match, multiple brightness matches and photometric fits, and regressions to find showing the equivalent single parameter gamma for two and three parameter models. Orange colored background denotes a significant difference with photometric gamma, green denotes no significant difference. Other entries in these columns were not tested because of experiment artifacts (see text).

		single brightness match		multiple brightness match				photometric fits	
	color	mean gamma	std. dev	mean gamma	mean k0	regres. gamma	slope	regres. gamma	slope
CRT temporal	Red	2.286	.0753	2.280	-0.008	2.311	4.118	2.200	3.410
	Green	2.220	.0872	2.182	-0.021	2.300	5.732	2.129	3.735
	Blue	2.176	.1002	2.192	-0.014	2.244	3.817	2.187	3.076
	White	2.294	.1150	2.288	-0.004	2.309	5.894	2.130	3.890
CRT spatial	Red	2.357	.4494	2.269	-0.022	2.367	4.493	2.200	3.410
	Green	2.909	1.9987					2.129	3.735
	Blue	2.262	.2022	2.253	-0.008	2.284	3.814	2.187	3.076
	White	2.637	1.2771					2.130	3.890
LCD spatial direct	Red	1.767	.7244	1.476	-0.118	*1.747	*2.306	2.067	3.121
	Green	1.549	.3534	1.431	-0.119	1.876	3.731	1.995	3.450
	Blue	1.451	.1532	1.319	-0.114	1.677	3.131	1.978	2.559
	White	1.618	.4898	1.386	-0.128	1.738	2.755	2.111	3.669
LCD spatial mirror	Red	1.263	.0831	1.064	-0.113	n.s.**	n.s.**	2.067	3.121
	Green	1.180	.0623					1.995	3.450
	Blue	1.246	.0592					1.978	2.559
	White	1.206	.0762					2.111	3.669

* significance of regression p = .055

** significance of regression p = .273

the CRT and LCD. There were no significant differences found within the favored and the lower rated pairs.

7.4.3 Ambient illumination

The room lighting had no significant effect on the results for single-gamma brightness matches for any combination of display, type of stimulus and color tested in an ANOVA, nor on the results of multiple-gamma brightness-matches methods for any combination of display, type of stimulus, and color that was suited for this method (see below).

7.4.4 Display

The size of the RMSE in code values between the fitted GOG model and the matched values was significantly larger for the LCD spatial condition compared to the CRT spatial condition for the colors red ($t = 3.671$, $p = 0.001$, $df = 27$) and blue ($t = 7.212$, $p < 0.001$, $df = 31$) indicating that the TRC of the CRT adheres better to the gamma model than the TRC of the LCD.

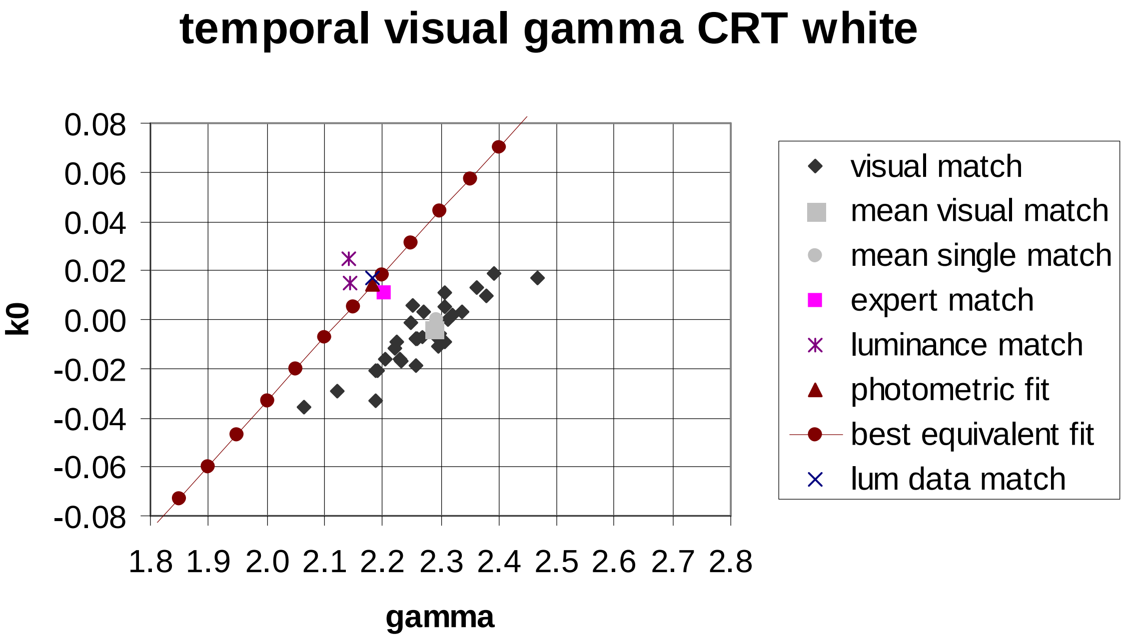


Figure 7.6. Temporal visual gamma CRT white, showing
visual match: Parameter combinations resulting from code value domain fit for temporal gamma multiple brightness matching results for CRT white (combined R, G and B),
mean visual match: the mean of the visual matches,
mean single match: the mean of the gammas resulting from single brightness matches
expert match: result from expert visual match,
luminance match: result from multiple luminance matching with PR650 and L203 photometers,
photometric fit: results from display characterization with L203 photometer with 17 equidistant samples in code value domain,
best equivalent fit: best fit to the characterization data with fixed gamma,

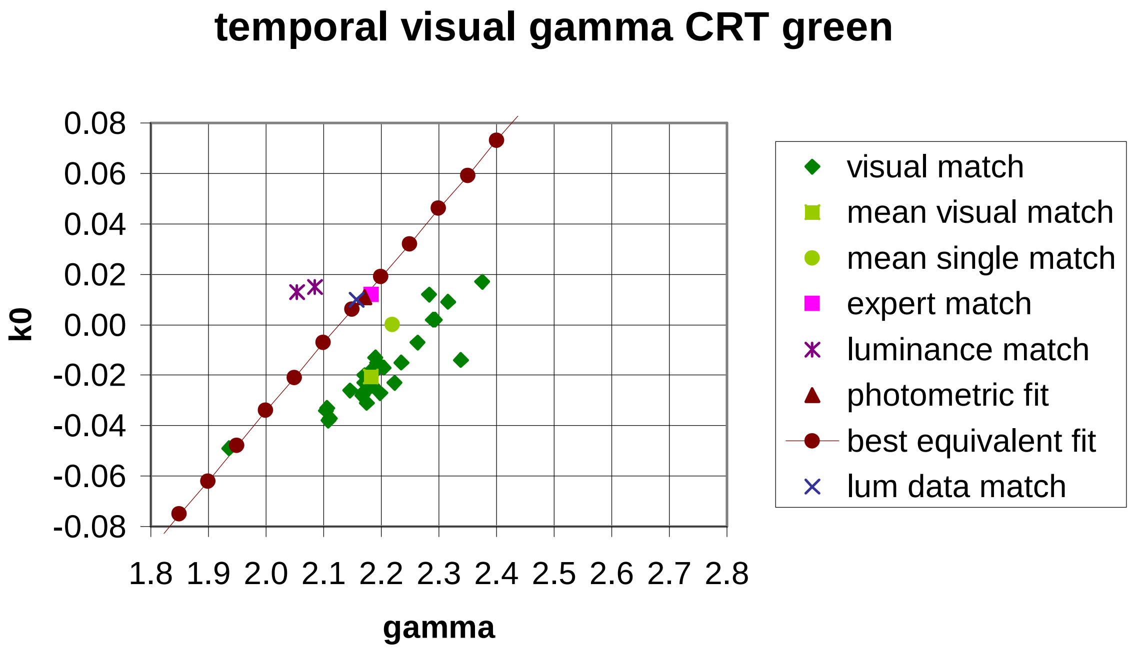


Figure 7.7. Temporal visual gamma CRT green. Same as figure 7.6 for CRT primary green. One visual match outside figure range.

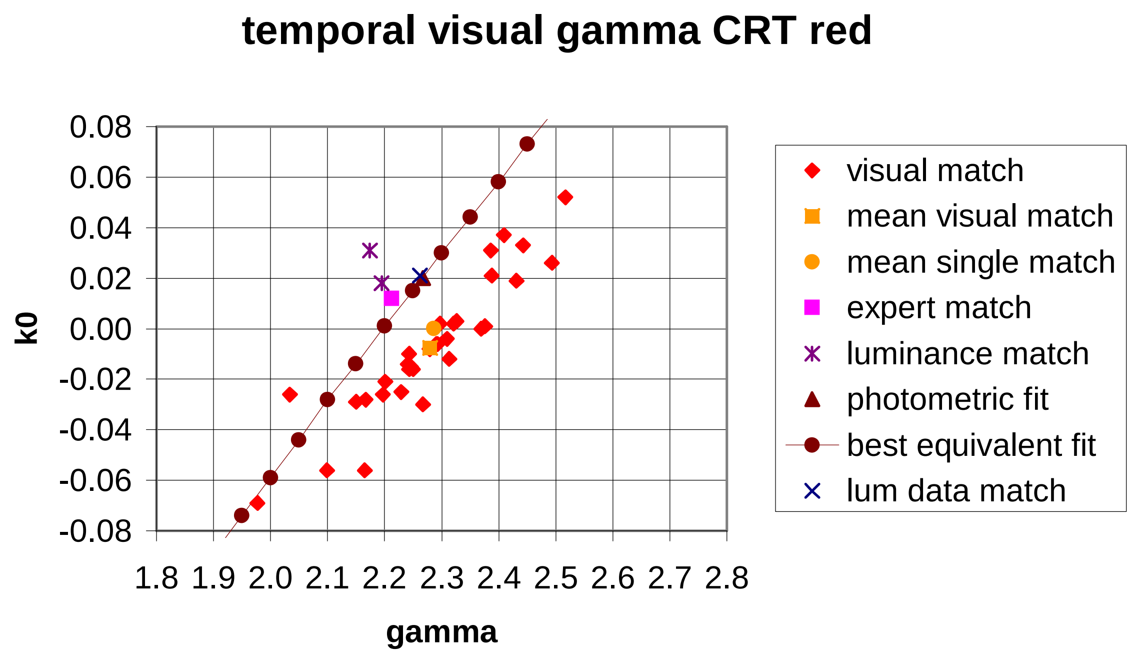


Figure 7.8. Temporal visual gamma CRT red. Same as figure 7.6 for CRT primary red.

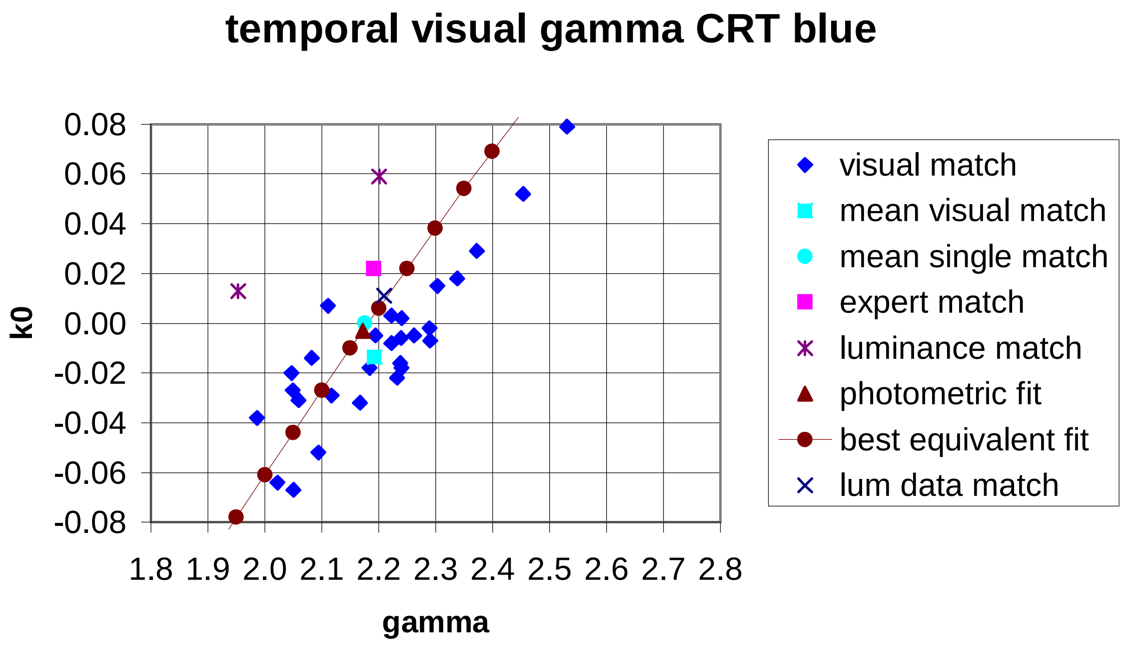


Figure 7.9. Temporal visual gamma CRT blue. Same as figure 7.6 for CRT primary blue. Four visual matches outside figure range.

The gamma estimation for the LCD resulting from a single (0, 255) brightness match was significantly smaller than the equivalent photometric fit with k_0 = 0.0 for all four colors in the experiment, with differences ranging from –0.527 for blue to –0.300 for red. In comparison, the estimated equivalent fit with k_0 = 0.0 based on eight brightness matches performed slightly better with the difference ranging from –0.373 for white to –0.119 for green (*p* = 0.074). Of course, two out of the eight brightness-matching stimuli are composed with the maximum code value 255. In contrast, most of the gamma estimates for the CRT from single brightness matches were larger than the equivalent photometric fit with k_0 = 0.0.

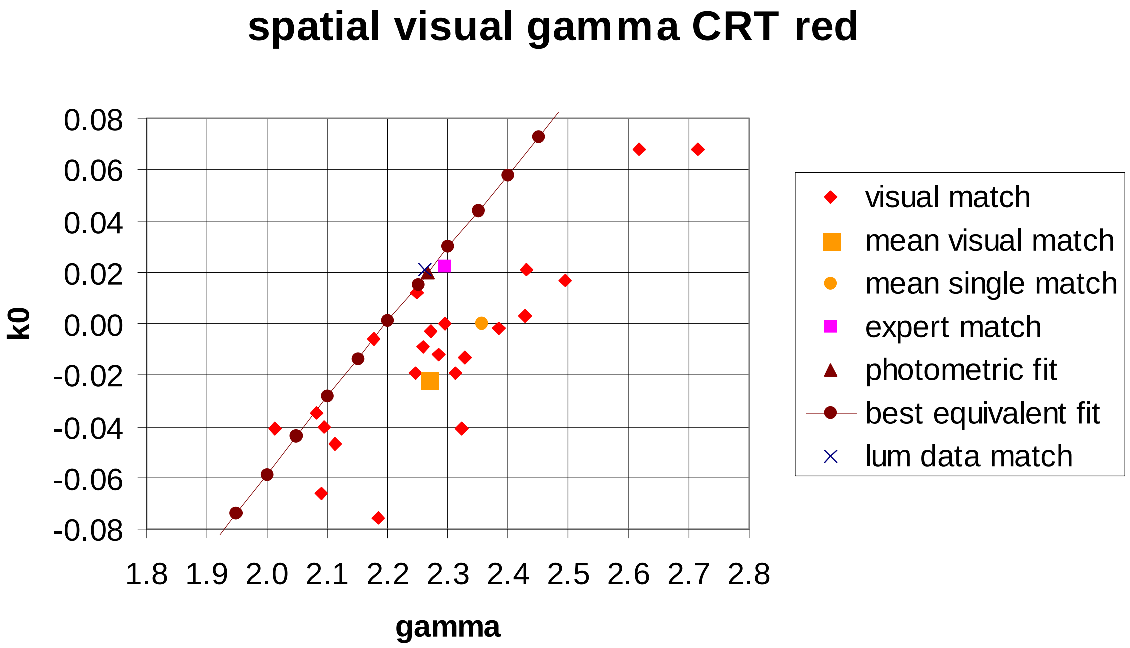


Figure 7.10. Spatial visual gamma CRT red. Same as figure 7.6 for spatial visual matching for CRT primary red. Six visual matches outside figure range, four visual matches would not fit to the gamma model.

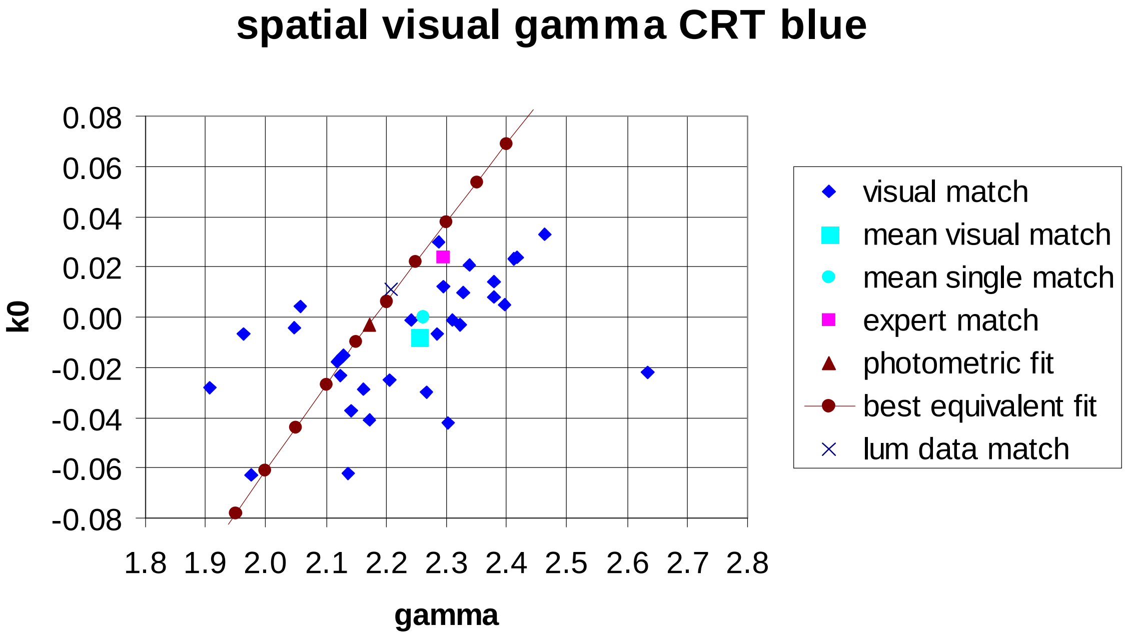


Figure 7.11. Spatial visual gamma CRT blue. Same as figure 7.6 for spatial visual matching for CRT primary blue. Two visual matches outside figure range.

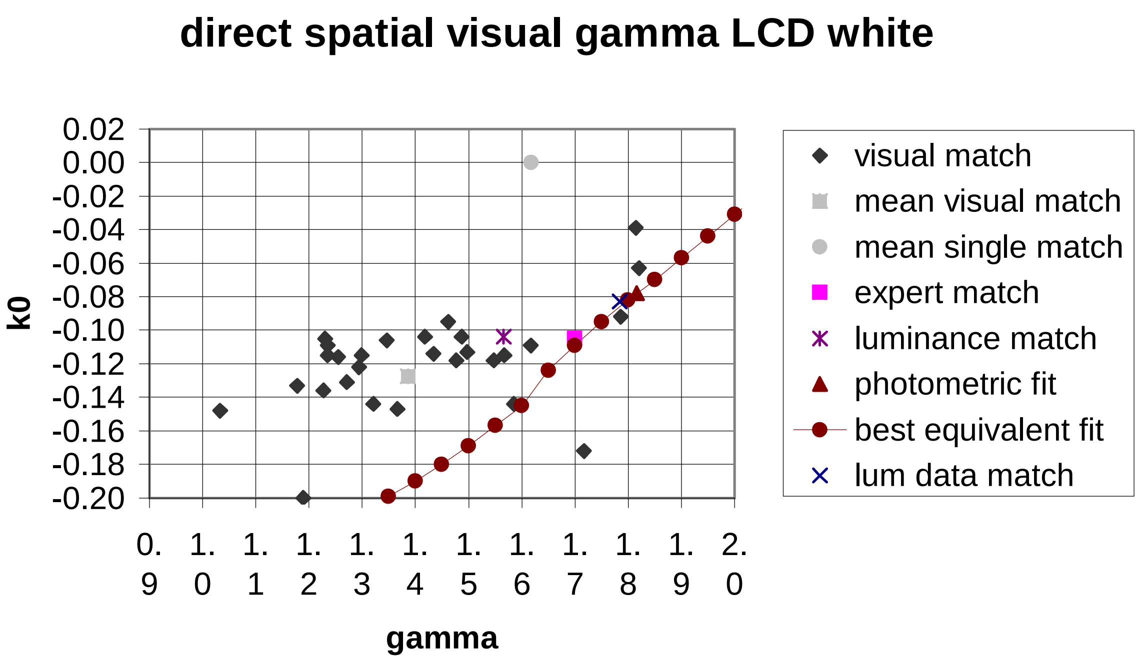


Figure 7.12. Direct spatial visual gamma LCD white. Same as figure 7.6 for direct spatial visual brightness matching on LCD.
No expert matching results, only PR650 used to make luminance match.
Two visual matches outside figure range, two visual matches would not fit to the gamma model.

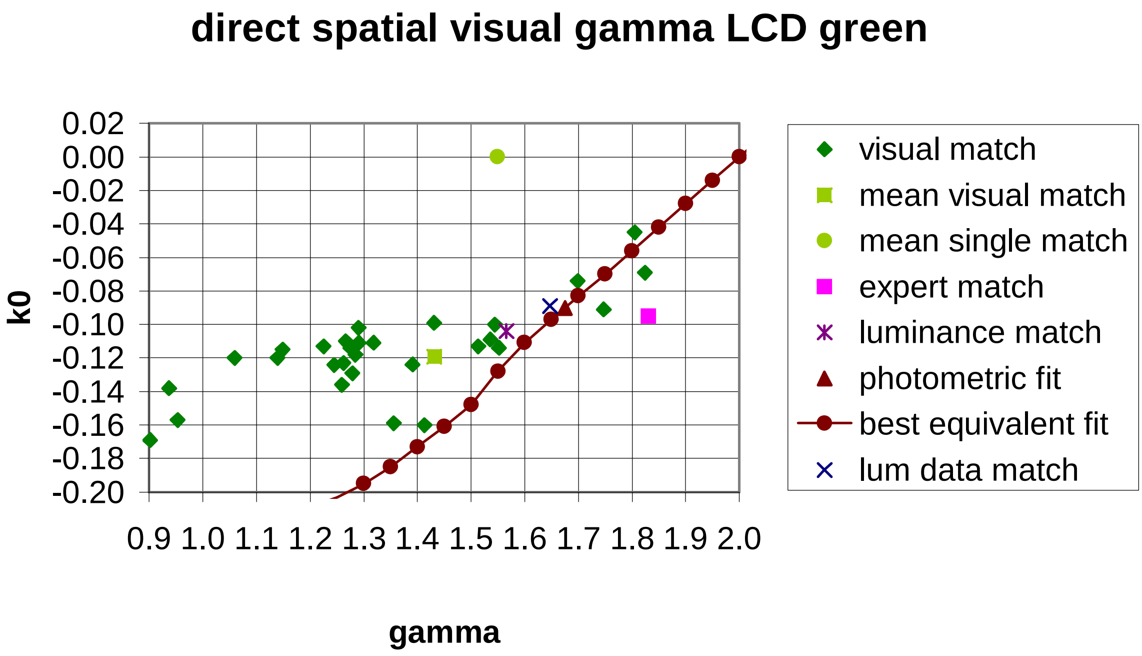


Figure 7.13. Direct spatial visual gamma LCD green. Same as fig. 7.12 for LCD primary green. One visual match outside figure range.

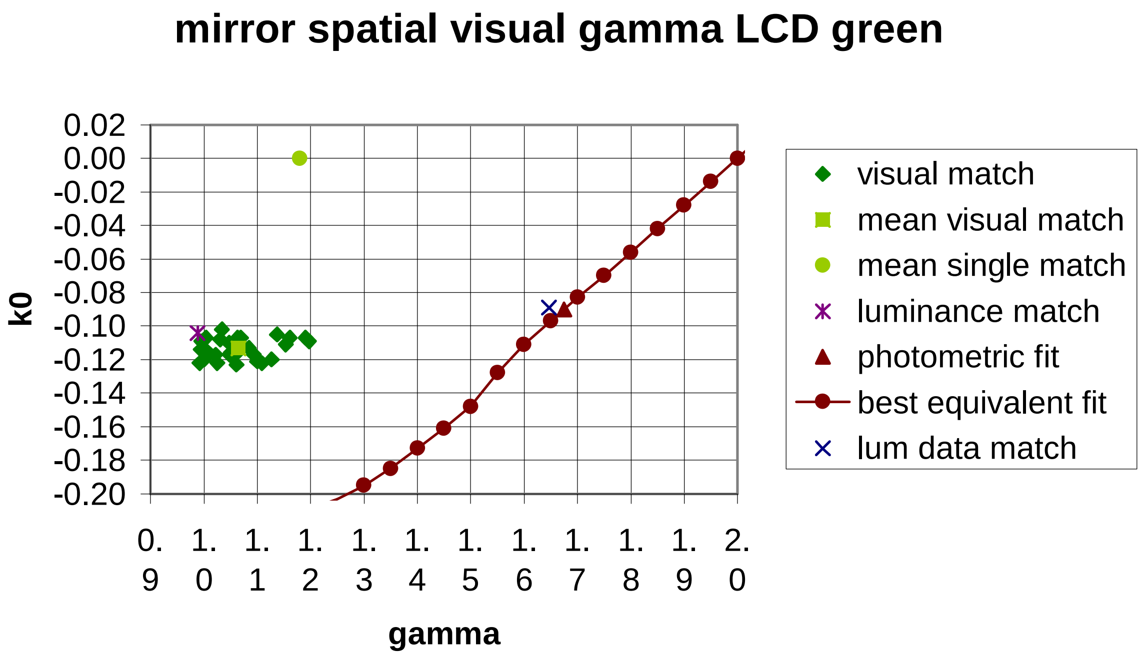


Figure 7.14. Mirror spatial visual gamma LCD green. Same as fig. 7.12 but for visual matches made via the mirror.

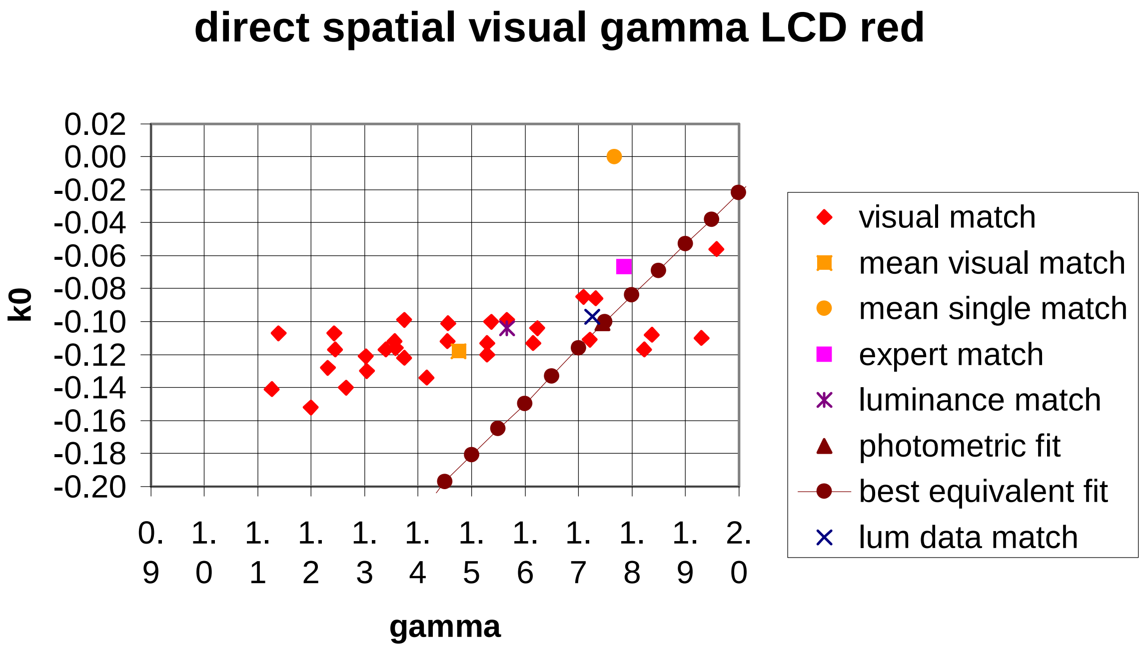


Figure 7.15. Direct spatial visual gamma LCD red. Same as fig. 7.12 for LCD primary red. One visual match outside figure range, one visual match would not fit to the gamma model.

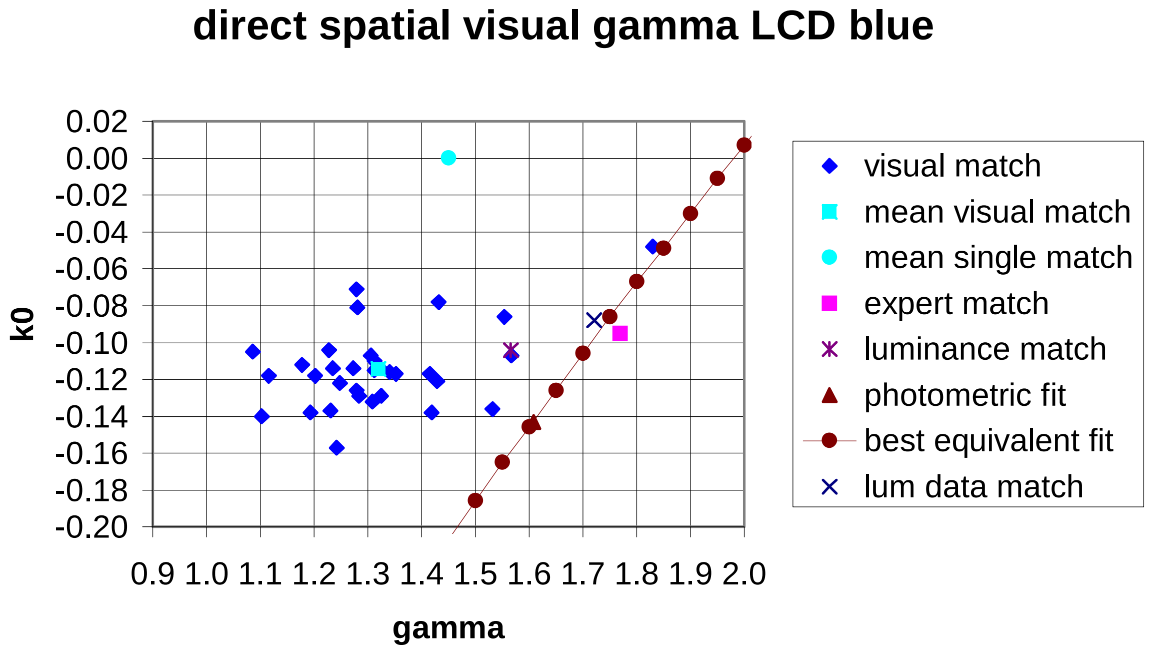


Figure 7.16. Direct spatial visual gamma LCD blue. Same as fig. 7.12 for LCD primary blue.

A comparison of all relevant results can be made from Table 7.8 and Figs. 7.6–7.16. The figures show the results of various fits from visual matches and photometric data to the GOG model [Eq. (7.1) without the luminance offset $Y_{P,0}$] in the gamma *vs.* normalized voltage offset k_0 plane. The same range of gamma and *k*0 values was chosen for the CRT figures and LCD figures respectively, therefore some of the results might fall outside the range, but comparisons between the different conditions are far easier.

The caret markers labeled 'visual match' show the parameter combinations resulting from code-value domain fit for the individual gamma multiple brightness-matching results. The large square marker designates the mean of these visual matches. The round marker of the same color shows the mean of the gammas resulting from a single brightness match. The magenta square marker shows the result from an expert's (the author) visual match. The asterisk marker shows the results from multiple luminance matching (“visual” matching with a photometer) made by measurement with the L203 photometer (CRT) and PR650 spectrometer (CRT and LCD).

For comparison, the brown triangle marker shows the locus of the photometric fit, the results from display characterization with the L203 photometer with 17 equidistant samples in code-value domain, and the brown line with round markers shows the best equivalent fits: the best to the characterization data with fixed gamma. The cross sign marker shows results from luminance comparison for visual gamma matching stimuli from full 256 code-value photometric measurement data.

The viewing-angle dependency of the LCD in the experiment made the results from the spatial brightness matching with mirror unusable, only one example is given for the green primary in Fig. 7.14. The fusion problems that many of the subjects had with the low-resolution spatial brightness matching on the CRT disturbed the results for the luminous green primary and white matchings as is shown in the high means and standard deviations for the single brightness match gamma fits (Table 7.8).

In comparing the figures, it can be observed that the line showing the best equivalent fits to the optimal photometric fits for other gammas, is in all cases nearly linear with a similar slope. Due to the presence of a luminance offset on both displays, the slope is somewhat steeper as the lines depicted in Fig. 7.1.

Another observation that can be made from the figures is that points for the luminance data match, *i.e.,* the multiple brightness-matching fitting procedure applied to the luminance data gathered in the display characterization, always lie close to the points for the photometric fits, showing that the code-value fitting procedure can provide the correct results.

In contrast, the results from the luminance matches with the photometer and spectrometer always show a lower gamma than the photometric fit, indicating that the luminance measured for the alternating stimuli was lower than expected. For the LCD, this corresponds with the results of the visual matches, but for the CRT the visual matches produce higher gammas for the spatial as well as the temporal alternating stimuli.

On the CRT the regression line of the visual matches has about the same slope as the line of best equivalent fits. On the LCD the pattern of the visual matches seems to depend more on the viewing angle: if the eyes of the subject were higher in relation to the normal of the display screen, the resulting gamma fitted to the visual matches was lower.

7.4.5 Type of stimulus

This hypothesis could not be fully tested, although Table 7.8 and the Figs. 7.13 and 7.14 show the LCD with mirror has lower variance, these cannot be counted as valid results. For the CRT, though it is clear that the spatial brightness match with low resolution produces more variance than the temporal brightness match. The fact that some subjects could not perform the spatial brightness match for green and white should be seen as extreme results.

7.4.6 Methodology

A series of one sample t-tests comparing the single brightness match gamma fits to the photometric gamma equivalent for $k_0 = 0$, showed significant ($p < 0.05$) differences for R, G, B, and W on the LCD direct spatial condition and CRT spatial B and CRT temporal R, G, and W, with CRT spatial R and CRT temporal B not significantly differing in gamma.

For the full-brightness match results, the differences for direct spatial LCD R, B, and W were significant and significant for CRT temporal on all colors as well as for CRT spatial R and B. For the displays in the experiment, there is no evidence that the gamma estimation from a multiple brightness match is any better than that from a single brightness match.

7.4.7 All colors vs. white

In a series of paired sampled *t*-tests, the results for temporal CRT, spatial CRT R and B, and direct spatial LCD for the different primary colors were compared. The test showed significant differences between the gammas of all primaries except for the comparison of temporal R and G estimates.

The variance in the single match gamma estimation for W is only slightly higher than for R, G, and B for the temporal stimulus on the CRT, and not the highest for the LCD with direct view.

7.5 Discussion and conclusions

There are many pitfalls in implementing a brightness-matching task to estimate the parameters of a gamma-model tone-reproduction curve for a display. On LCDs, problems with the viewing angle are to be expected, although not so severe as with the older LCD used in this experiment, the effects might be large enough to make the gamma estimation flawed. The spatial gamma-matching task might not be accurate due to pixel dependencies and the temporal task cannot be implemented on LCDs and might not work on all computer systems.

There also appears to be a considerable difference between what instruments measure as equal luminance and what (untrained) subjects see as equal brightness for the stimuli in the brightness-matching task. The subjects seem to have a slight bias for matching a higher brightness in judging the CRT with temporal as well as spatial matching stimuli. For the single brightness match, which is also the first match of the multiple brightness matching, this might have been caused by the display artifacts (jitter, pixel dependency) and perceptual difficulties (flicker and fusion problems), which were most visible for the high contrast alternating stimuli. For the other brightness matches, the contrast reducing effect of the afterimages might play a role. The results of the expert’s visual gamma matching are clearly closer to the photometric fit. The expert was probably better able to disregard the artifacts and better trained to overcome the perceptual difficulties, although age-related decline of visual abilities might also play a role.

For the spatial visual gamma matching, a higher resolution of the alternating stimulus clearly reduces the problems. In practice, this can be achieved by a larger viewing distance, but this might change the reflections on the display and cause problems reaching input devices in practical situations. Using a mirror might be problematic with CRTs because changes in orientation to the earth’s magnetic field affect the displayed image and with LCDs because of possible viewing-angle problems.

A reasonable amount of ambient illumination did not have a significant effect on the results, therefore there should be no problem in performing a visual gamma-matching task in a adequately illuminated office environment.

Multiple brightness matches did not enhance the quality of the gamma estimations for the displays in the experiment. Better results with this task should be expected if the effect of afterimages could be cancelled by a short flash of the opponent color between stimuli. If the tested display has a substantial voltage offset, multiple brightness matches should perform better than the single brightness match. A reduction of variation in the results for all displays might be achieved by letting the user choose the best of a number of simultaneously visible brightness matches representing different gammas. A second measurement centered round the first choice gamma with smaller differences between gammas could be used to improve the accuracy.

This study has produced improvements in the mathematical methods used to estimate GOG-model parameters from multiple gamma brightness matchings and a method has been introduced to effectively compare gamma models with different numbers of parameters. In practice, the visual gamma matching task is vulnerable to a number of artifacts and perceptual problems which make it difficult to make a robust and reliable implementation of the task especially on LCDs.

8 Conclusions and recommendations

The speed of developments in display technology far exceeds the pace of international standards renewal. Nine years after the submission of the paper on standardizing visual display quality (chapter 2) the featured ISO 9241-3 standard published in 1992 is still the current standard in 2007. This standard is being revised and some of the problems addressed in the paper appear to be treated in the new 9241 standard. In the draft standard there are now separate parts addressing optical laboratory test methods and field assessment methods in line with our recommendations in chapter 2 to formulate a separate standard for engineers and display users, and to create a higher degree of independence of the standard from the display technology. Also as a result of our involvement through the contributions of co-author Gerd Spenkelink in the Dutch standardization commission (NEN-ISO TC 159 SC4), different working environments (office, medical, control rooms, production, counter, mobile, and airport/railway station) and tasks are distinguished in the draft version of the new standard 9241-300: "Ergonomics of human-system interaction -- Introduction to electronic visual display requirements". Even a modest amount of gradation of quality features was introduced.

From our experience not enough independent experts are able to contribute to the standardization process on a regular basis, mainly because these contributions are consuming too much time and resources in relation to the available funding. National and European Union research funding institutions should offer more financial support to improve the level of expert input into the standardization process.

The quest for a validated model of image quality based on a broad perspective, combining technological, human vision and ergonomics research, is supported by the introduction of the “Image Quality in Context Cycle” model and the classification framework for image quality factors. These concepts offer a structured approach to the definition and standardization of image quality for different contexts with a well-documented design rationale. The model and the framework should not be seen as definitive solutions carved in stone, but as a next well stocked base station in the expedition discovering image quality space.

The framework can also provide some guidance in display and image quality research. A great deal of the documented experiments is concerned with comparisons between samples of display technologies or variations of technological parameters with effects on multiple quality factors. Understanding the effect that single quality factors and their interactions have on the overall quality can provide fundamental knowledge for all display and image system designers in all technologies.

The usability of the “Image Quality in Context Cycle” model can be demonstrated with the results of the research on the tone reproduction curve presented in chapter 5, 6, and 7:

- The original research question was concerned with the improvement of the user interface to allow the user to adequately adjust the technology parameters of a display.
- The physical model of the display was checked: Improvements were suggested for the physical model and the measurement methods.
- The environmental variable ambient illuminance was an independent factor in the experiments of chapter 6 and 7.
- The presentation of visual stimuli without artifacts proved to be a problem.

- There was a notable difference between the perception of the stimuli between inexperienced users and an experienced user.
- User instructions, stimulus screens and hardware controls did not provide much support in adjusting the display to its optimal settings.
- Standards did not provide criteria for rejection or acceptation.

The two methods suggested for improving the models fitted to the CRT's tone reproduction curve (TRC), i.e. the optimization of the model fit in the CIE Lightness domain and the variable gamma model, clearly proved their value. Besides improving the fit on the dark side of the TRC, the optimization in the lightness domain enhances the perceptual validity of the TRC model. This optimisation could be further refined by making allowances for the viewing intent, that is, the intended view-surround contrast conditions (*e.g.*, office, living room television, or cinema lighting conditions).

The voltage dependency of CRT gamma can have a measurable effect on the TRCs within the luminance operating range of office CRTs. The gliding gamma model provides more accurate fits for these TRCs in the lightness and luminance domain.

The gliding gamma model does not provide improved fits for all CRTs: the CRT used in the black level setting and visual gamma measurement experiment fitted rather poorly to all gamma models. Added circuitry for compensation of CRT artifacts and additional features (in this case the suitability for the Philips Lightframe ™ software) might be to blame.

As shown in chapter 5 the sRGB standard does not do a great job of describing the TRC of at least one sample of the CRT technology it is meant to represent. Rather than trying to fit the TRCs of the new technologies to the technical oddities of the CRT a new standard should be developed. The “physical properties” of the display technologies should not be the base for color management standards. They should be based on human lightness perception, with CIE lightness the most likely candidate. New technologies already have lookup tables to correct their technical TRCs and these could easily be filled with lightness correction curves. With the acceptance of the digital display interface there should be no problem to integrate lightness correction LUTs in the CRT.

From the results of the black-level adjustment experiment, it can be concluded that most of the average untrained users are not able to find the optimal black-level setting with the provided instructions by adjusting the contrast and brightness controls. We suggest e several improvements that can be made to the black-level optimizing procedure:

- The common user is served better by simple instructions presented one at a time.
- Controls that have no real purpose, like the contrast control on the LCD in the experiment, should better be left out.
- Black-level adjustments should be made under low ambient illumination.
- Displays should be provided with a picture line-up generating equipment (PLUGE) signal, which has fields with a +2% and –2% voltage value or preferably an even smaller range, and the black level can then be positioned directly in between.
- The on-screen settings dialogs should be adapted to provide minimum interference with the black-level adjustment task:
  - Less obtrusive dialog luminance and colors
  - More time allowance to adjust the settings.

The study on visual gamma matching methods has produced improvements in the mathematical methods used to estimate GOG-model parameters from multiple gamma brightness matchings and a method has been introduced to effectively compare gamma models with different numbers of parameters.

Visual gamma matching tasks seem to have little practical value especially for LCD displays. The experience gathered by implementing and using the tasks in this study has shown that:

Temporal brightness matching tasks

- are difficult to implement on Windows operated PCs;
- are not suitable for slow switching LCDs;
- suffer from artifacts like jitter and flicker.

Spatial brightness matching tasks

- may suffer from fusion problems for lower resolutions and higher contrasts;
- might cause practical problems if the viewing distance needs to be increased, like lack of room, and short cables for direct view conditions, and change of earth magnetic field, and viewing angle problems with the use of a mirror;
- might not be accurate due to pixel dependencies;
- on LCDs can suffer from problems with the viewing angle; although not so severe as with the older LCD used in the experiment, the effects might be large enough to make the gamma estimation flawed.

General problems with brightness matching tasks are

- focusing problems and afterimages (for multiple brightness matches);
- bias for higher brightness (gamma).

Results from the study imply that there should be no problem to perform a visual gamma matching task in an adequately illuminated office environment. Multiple brightness matches should perform better than the single brightness match, especially if the tested display has a substantial voltage offset.

Based on the experiences in the experimentation room the following suggestions for future research on improving the visual gamma matching task can be made:

- The cancellation of the effect of afterimages by a short flash of the opponent color between stimuli in multiple brightness matching tasks.
- A reduction of variation in the results for all displays might be achieved by letting the user choose the best of a number of simultaneously visible brightness matches representing different gammas. A second measurement centered round the first choice gamma with smaller differences between gammas could be used to improve the accuracy.

In this study the visual gamma matching task has shown itself to be vulnerable to a number of artifacts and perceptual problems which make it difficult to make a robust and reliable implementation of the task especially on LCDs. As the cost of accurate light measuring components is decreasing steadily it seems plausible in our opinion that in the near future most displays will be equipped with some feedback mechanism to optimize the luminance and color output of the display with regard to the applicable standard settings.

Appendix A Image quality factor cells

A list of image quality factors represented in the classification framework for visual display or image quality in chapter 4.

A.1 Level

Luminance level (L1)

Associated terms: black level, white level
Description: Mostly the lowest luminance level or black level or the highest level or white level of the display or imaging system. If one level is set the other level and the average luminance level follows from the combination with the luminance range.
Relevance: Determines in what environment and for what applications the image can be used, e.g. bright sunlight, office work, night shifts on the bridge of a ship. It also determines which kind of content can be adequately displayed (dark scenes, bright scenes). Important for the tuning of multi-panel displays.
References: chapter 6, [BFT03; Rob91]

Spatial level (S1)

Associated terms: size, real image size
Description: The real size of the displayed image in height and width, from which diagonal size and aspect ratio can be derived. Projection displays might have a range of suitable sizes.
Relevance: The real size of the image is important in relation to the application: e.g. a large size is desirable for stadium, meeting room and (home) cinema, and a small size for mobile phone, digital camera viewer and head mounted display. The displayed image should preferably have the same aspect ratio as the original image.
Figure 1 also shows the limit of resolution of the human visual system. Since the vast majority of electronic displays lie below the limit of human ocular resolution, people can actually see and perceive the individual pixels. This is not the case with printed characters or photographs and is one of the reasons that electronic displays are not as comfortable for reading and as pleasing for images as newspapers or photographs. One of the key challenges of display technology is to reduce pixel size to below the point at which humans can perceive individual pixels; accomplishing this is a goal for the future [WR00].[♣] Since 2000 advances in LCD technology have produced displays better than the resolution limit, but it may take some time before the use of these displays is a viable option for the general public.
References: [WR00]

[♣] The remark made about the CRT with the highest number of pixels: "Addressable, not resolvable", could also be made about the 300 and 600-dpi prints on paper. For most black and white printers the neighboring dots overlap for more than 50%, for intermediate gray levels dot patterns are used. This means 300-dpi prints of text and 600-dpi prints of images are usually below the limiting resolution of the human visual system.

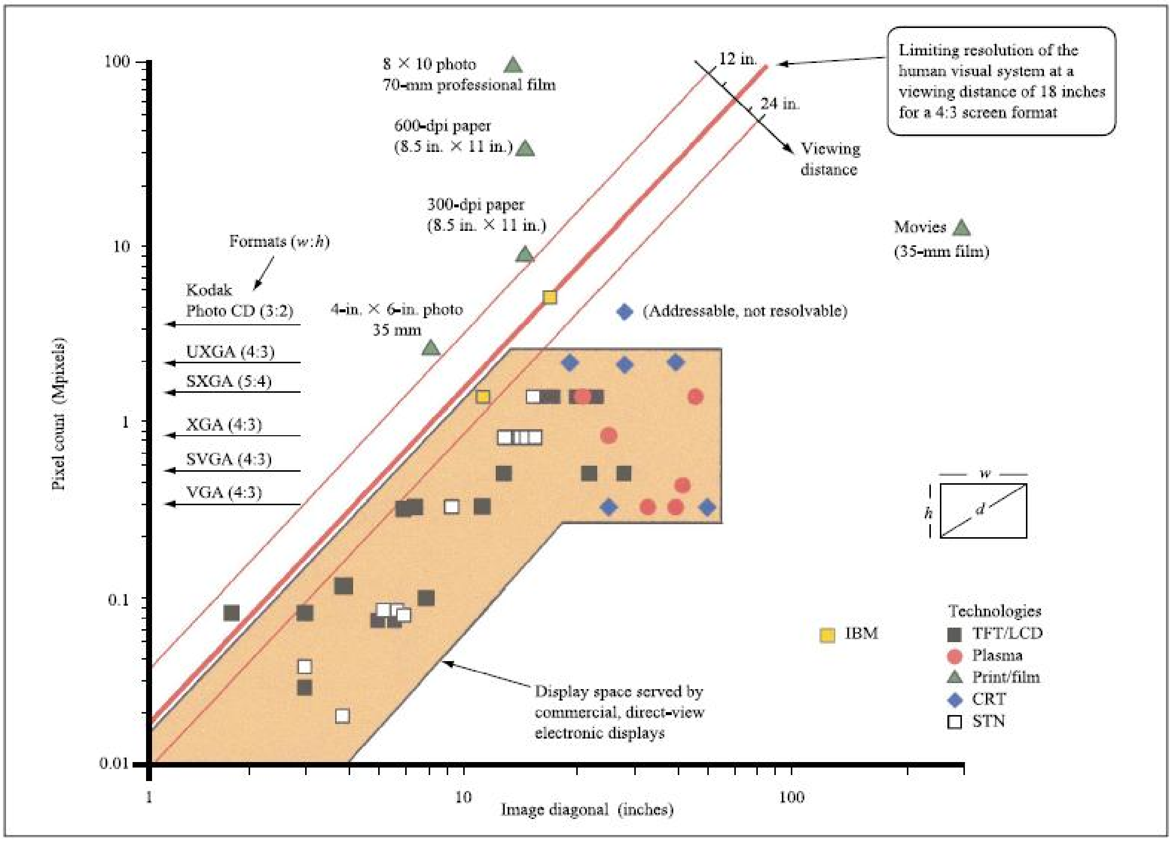


Figure A.1 The relation between real image size, addressability, the viewing distance (only computer display viewing distances depicted), the resolution limit of the human visual system and the relative position of certain computer display and hardcopy applications from Wisnieff [WR00].

Chromaticity level (C1)

Associated terms: white point

Description: Determines the color of white (and preferably all other designated neutral colors) in the image.

Relevance: Important if images have to produce (about) the same color as a neighboring display or hardcopy or as the original that is recorded, like in multi-panel displays. For optimal viewing conditions the white point of the environment should be the same as that of the image. Due to differences in the tone reproduction curves of the primary colors neutral colors may vary in color with luminance.

References: [FHH97; JLHC02; VHSL03]

Temporal level (T1)

There is no level or range of levels of time in relation to human perception. The human visual system is not able to switch to a mode in which it can view as a high-speed camera or take one frame every minute to capture the movement of an opening flower. Any association of temporal level or range with an image quality aspect would be construed.

Depth level (D1)

Associated terms: convergence distance
Description: The distance at which there is no horizontal parallax at the optical axis of the display.
Relevance: In real world viewing vergence and accommodation are normally closely linked visual actions. With stereoscopic displays the eyes must remain focused at the surface of the screen at all times regardless of where the eyes are verged in the stereo monitor. Excessive screen parallax, i.e. the horizontal distance between corresponding points in the left and right eye image, can lead to stereoscopic images appearing out of focus and/or the viewer being unable to fuse the images. A large difference between viewing distance and convergence distance will exacerbate screen parallax.
References: [WDK93]

A.2 Range

Luminance range (L2)

Associated terms: dynamic range, contrast range, contrast
Description: Difference or ratio between the highest and the lowest possible luminance level.
Relevance: See L1. The range also determines how many distinguishable luminance levels can be displayed (L3). In general the luminance range is higher than the contrast between neighboring pixels, which is limited by spatial blur.
References: [WC00]

Spatial range (S2)

Associated terms: angular size
Description: The size of the image in height and width as projected on the retina of a normal accommodated eye.
Relevance: Determines the size of the image on the retina. Is determined by the size of the image and viewing distance. In comparison with the real size a large angular size is desirable for meeting room, (home) cinema and head mounted display, and a small angular size for stadium, digital camera viewer and mobile phone.
References: [WR00]

Chromaticity range (C2)

Associated terms: color gamut, primary color positions
Description: The range of chromaticities that can be displayed.
Relevance: The chromaticity range is determined by the primary colors used and the type of color mixing, additive for displays and subtractive for print. The gamut also depends on the luminance level: for displays the chromaticity range for bright colors is much smaller than for dim colors. The range of visible colors is established in the CIE Standard Observer .
References: [Ber00; EHS04]

Temporal range (T2)

There is no quality factor that can be associated with temporal range (see temporal level, T1). Any associations with frame rate and speed are better represented under temporal addressability T3.

Depth range (D2)

Associated terms: horizontal parallax
Description: the range of depths that can be displayed.
Relevance: The depth range should be limited to avoid eyestrain and problems with focusing and fusion.
References: [WDK93]

A.3 Addressability

Luminance addressability (L3)

Associated terms: gray scale resolution
Description: The number of different intended luminances that can be displayed or printed. (discounting differences between e.g. pixels with the same gray level in the middle and on the side of a CRT that have a different luminance).
Relevance: Determines the visibility of detail esp. in bright and dim areas of the image.
References: [MW91]

Spatial addressability (S3)

Associated terms: resolution, pixel size, dots per inch (dpi), lines per inch (lpi), pixels per inch (ppi).
Description: the number of different intended positions in the basic 2D spatial dimensions that can be separately addressed.
Relevance: The spatial resolution for printing specified in lpi is typically much smaller than the spatial addressability specified in dpi, e.g. a standard 300 dpi printer can usually only produce 65 lpi. A high dpi can help to diminish the visibility of ‘jaggies’ or ‘staircases’ in printed text. A much lower addressability on displays denoted in ppi is needed, because unlike printer dots pixels can display gray values and colors without changing size.
References: [MV85]

Chromaticity addressability (C3)

Associated terms: chromaticity sampling distance
Description: The number of different intended colors that can be produced by a system.
Relevance: This number is often much larger than the number of different colors that can really be produced by the system. If e.g. clipping occurs because the black level of a display is set incorrectly millions of addressable colors might be mapped to the same displayed color.
References: chapters 5 and 6.

Temporal addressability (T3)

Associated terms: update resolution, refresh rate, frame rate.
Description: The frequency at which the display updates the displayed image
Relevance: The refresh rate of the system might be to low. In combination with the luminance level and the amount of temporal discontinuity this might cause the perception of flicker. If the rate, at which the content of moving images is refreshed, is too low, motion discontinuity can be perceived.
The refresh rate of the display system may differ from that of the original recording (movie **ó** NTSC, NTSC **ó** PAL, silent movie **ó** movie). Simple transition methods might lead to

speeded up or slowed down movement (and sound). Temporal up- and down-conversion interpolation techniques are now applied to provide realistic display of moving images.
References: [Rog84]

Depth addressability (D3)

Associated terms:
Description: The number of depth levels that can be displayed.
Relevance: In principle a higher number of depth levels gives a better picture, although human vision is not as sensitive in depth as in the spatial dimensions. The number of depth levels for stereoscopic cameras equals the number of horizontal pixels in the image. Reducing the depth addressability can sometimes improve image quality in cases of temporal discontinuous-depth mismatches (TDL7).
References: [MIJS04]

A.4 Equidistancy

Luminance equidistancy (L4)

Associated terms: tone reproduction curve, gamma
Description: The accuracy with which the luminance differences in the image are mapped onto the JND scale of human lightness perception.
Relevance: A wrong mapping can lead to washed out images or loss of contrast in bright or dim parts of the image.
References: Chapters 5, 6 and 7.

Spatial equidistancy (S4)

Associated terms: geometric distortion
Description: the accuracy with which differences in space are mapped on the image.
Relevance: Pin-cushion and barrel distortion are common artifacts on CRTs, often controls are supplied to try to correct these distortions. Other distortions can occur when images with a smaller rectangular addressability are enlarged to fit the full area of a matrix FPD screen with a higher rectangular addressability, e.g. a 800x600 image is displayed on 1024x768.
References: [VHFC02]

Chromaticity equidistancy (C4)

Associated terms: non-uniform chromaticity sampling
Description: The accuracy with which the chromaticity differences in the image are mapped onto the JND scale of human chromaticity perception.
Relevance: The amount of distinguishable colors that can be displayed is often a fraction of the number of addressable colors.
References: Chapter 5, [HRV97; KS90].

Temporal equidistancy (T4)

Associated terms: non-linear update, judder
Description: the accuracy with which differences in time between image frames at acquisition are mapped onto the timescale at the display of the image.

Relevance: Non-linear update can occur in forms of transfer from a film or video format with one frame rate to a format with a different frame rate. Regularly occurring conversions are from film format (24 frames per second) to PAL video format (25 frames per second), where every 12th frame is repeated for the duration ½ interlaced frame. With NTSC the effect is more pronounced with 5 frames of video for every 4 frames of film. Similar more irregular effects can occur when transmission of signals (internet, satellite, digital broadcasting) is temporarily delayed or disturbed.
References: [LFGC03]

Depth equidistancy (D4)

Associated terms: depth non-linearity
Description: the accuracy with which the differences in depth are displayed in the image.
Relevance: For stereoscopic displays the depth is often stretched between viewer and display and compressed between display and infinity. This can lead to wrongly perceived depth and if the camera system is in motion to false estimations of velocity.
References: [WDK93]

Depth Spatial equidistancy (DS4)

Associated Terms: keystone distortion, depth plane curvature, shear distortion
Description: Depth is not constant over the spatial area of the displayed image. In this respect depth should not be considered as a separate domain but rather as the third spatial dimension.
Relevance: With keystone distortion pixels on the right side of the left-eye view are displayed smaller than those from the left side and vice versa for the right-eye view. This causes an unwanted vertical parallax on the sides of the image. With depth plane curvature objects on the side of the image look further away than objects in the middle of the image, which should be at the same distance. Keystone distortion and depth plane curvature are a result of image acquisition with a toed-in camera configuration, they do not occur with a parallel camera configuration.
On a binocular stereoscopic display when the observer moves sideways the stereoscopic image seems to follow the observer. The movement results in a sideways shear of the image changing the relative distances within the scene.
References: [WDK93]

A.5 Smearing

Spatial smearing (SL5)

Associated terms: blur, spatial blur, spot size, spot shape
Description: The way in which the luminance value (or color value if no chromatic effects occur, i.e. smearing is equal for all primary colors) that belongs to one spatial coordinate in the original is distributed over that coordinate and the neighboring coordinates in the image.
Relevance: Primary source of unsharpness.
References: [Kee02]

Temporal smearing (TL5)

Associated terms: temporal blur, motion blur

Description: The way in which the value that belongs to one frame in the original is distributed over that frame and the following frames in the image.
Relevance: The slow rise and decay times of some LCDs will make moving objects unsharp because past images are not yet faded and the current image is not completed. Sample-and-hold display timing will make moving objects unsharp, because the eye tracking the object will convert the hold period to a spatial smear on the retina.
References: [KV04; Mis04; Slu05]

Chromatic smearing (SC5)

Associated terms: color blur, misconvergence
Description: The way in which the color value that belongs to one spatial coordinate in the original is distributed in different primary colors over that coordinate and the neighboring coordinates in the image.
Relevance: If the primary colors of a CRT are not converged on the same spot color fringes may appear on edges in the image. This can also occur on LCDs if sub-pixel rendering is used.
References: [BNS99; MKL89]

Depth smearing (DS5)

Associated terms: crosstalk
Description: Values belonging to the left-eye view leak to the right-eye view and vice versa.
Relevance: An incorrect head position of an observer (e.g. tilted head) causes annoying image ghosting for linear polarization techniques.
References: [MIJS04]

Temporal Chromatic smearing (TC5)

Associated terms: temporal color blur, color break-up
Description: The way in which the value that belongs to a frame in the original is distributed over that frame and the neighboring frames in the image sequence. Leading and trailing bands of primary colors or primary color pairs can be seen for moving objects.
Relevance: In sequential color displays, where the primary colors are displayed (usually projected) in different sub-frames (time multiplexing), leading and trailing bands of primary colors or primary color pairs can be seen if a mixed-colored moving object is shown.
References: [PNMC98]

Depth Temporal smearing (DTL5)

Associated terms: crosstalk
Description: On temporal multiplexed stereoscopic displays information from one frame may leak to the next.
Relevance: In stereoscopic displays with shutter glasses, where the left-eye view and right-eye view are displayed sequentially in time, values belonging to one view make leak through to the other.
References: [MIJS04]

A.6 Discontinuity

Spatial discontinuity (SL6)

Associated terms: raster separation, pixel separation
Description: the visibility of inactive parts of the image surface between picture elements.
Relevance: In color CRT the shadow mask could be visible for large screens at shorter (than intended) viewing distances. If CRTs are operated with a low horizontal frequency, horizontal raster separations could become visible. For FPDs the active area of the pixels can be too small, making the inactive area visible.
References: [SB92]

Temporal discontinuity (TL6)

Associated terms: flicker, frame rate, duty cycle
Description: the visibility of the fluctuation of luminance in time of a field of constant luminance
Relevance: The duty cycle characterizes the course of the level of luminance in the period of one frame. If the duty cycle resembles a short peak like in most CRTs flicker may be experienced. In LCDs the duty cycle has a constant level and frame flicker does not occur. Lengthy exposure to flicker can cause visual fatigue. The nuisance of flicker depends on the frame rate, the duty cycle of the signal in time, the brightness of the image, the viewing distance and the age of the viewer.
References: [Bac99; BJB01; Burr05; MRB97; RMGB01]

Spatial Chromatic discontinuity (SC6)

Associated terms: color separation, moiré
Description: the visibility of different chromaticities in a field of constant color
Relevance: Repeating small black & white patterns might cause (moving) colored stripes on shadow mask CRTs.
References: [Ami91]

Temporal Chromatic discontinuity (TC6)

Associated terms: color flicker, color flash effect
Description: the visibility of the fluctuation of chromaticity in time of a field of constant color
Relevance: In field sequential color displays, where the primary colors are displayed (usually projected) in different sub-frames (time multiplexing), color flashes can be seen during saccadic eye movements.
References: [BC04; Jär05; PNMC98]

Depth Spatial discontinuity (DSL6)

Associate terms: picket fence effect
Description: the visibility of inactive parts of the image scene between the picture elements.
Relevance: If observers move their head laterally in front of a multiview autostereoscopic display vertical bands may appear due to the black mask between columns of pixels on the display.
References: [MIJS04]

A.7 Instability

Temporal Spatial instability (TSL7)

Associated terms: jitter, ripple
Description: The position of a constant (part of the) image is slightly varying in time.
Relevance: Causes unsharpness of the image and can cause line flicker for horizontal lines on interlaced CRTs. External electromagnetic fields can induce jitter.
References: [KLSH00]

Spatial Color instability (SCL7)

Associated terms: color non-uniformity, mura
Description: The chromaticity and luminance is not constant in a field that should have a constant color.
Relevance: Fields of homogeneous color may look locally discolored. This can appear on (older) CRTs, due to dependency of viewing angle or the exertion of pressure on LCDs.
References: [VHSL03]

Temporal Color instability (TCL7)

Associated terms: color drift
Description: The color of a constant (part of the) image is varying over time.
Relevance: Colors can change as result of uneven loss of luminosity or absorption of the primary color materials (phosphors, dyes). On a shorter time scale the colors might change during the warming-up period of a display.
References:

Temporal Depth instability (TDL7)

Associated terms: temporal discontinuous-depth mismatches
Description: The position in depth of a constant (part of the) image is slightly varying in time.
Relevance: Temporal discontinuous-depth mismatches can occur if objects or parts of an object are assigned to different depth layers in time, which results in a flickering depth percept.
References: [MIJS04]

Appendix B: DICOM GSDF

The DICOM Grayscale Standard Display Function (GSDF) [DIC03] is defined by a mathematical interpolation of 1023 luminance levels derived from Barten's model [Bar92]. The GSDF calculates the luminance, L, in candelas per square meter, as a function of the just-noticeable difference (JND) Index, j:

$$\log_{10} L(j) = \frac{a + c.\ln(j) + e.(\ln(j))^2 + g.(\ln(j))^3 + m.(\ln(j))^4}{1 + b.\ln(j) + d.(\ln(j))^2 + f.(\ln(j))^3 + h.(\ln(j))^4 + k.(\ln(j))^5} \quad \text{(A1)}$$

with ln referring to the natural logarithm, j the index (1 to 1023) of the luminance levels L(j) of the JNDs, and *a* = -1.3011877, *b* = -2.5840191 E-2, *c* = 8.0242636 E-2,
d = -1.0320229 E-1, *e* = 1.3646699 E-1, *f* = 2.8745620 E-2, *g* = -2.5468404 E-2,
h = -3.1978977 E-3, *k* = 1.2992634 E-4, *m* = 1.3635334 E-3.
To apply the above formula to a device with a specific range of L values, it is convenient to also have the inverse of this relationship, which is given by the following equation:

$$j(L) = A + B.\log_{10}(L) + C.(\log_{10}(L))^2 + D.(\log_{10}(L))^3 + E.(\log_{10}(L))^4 + F.(\log_{10}(L))^5 + G.(\log_{10}(L))^6 + H.(\log_{10}(L))^7 + I.(\log_{10}(L))^8 \quad \text{(A2)}$$

with A = 71.498068, B = 94.593053, C = 41.912053, D = 9.8247004, E = 0.28175407,
F = -1.1878455, G = -0.18014349, H = 0.14710899, I = -0.017046845.

Bibliography

[Abo07] (2007, March 1). "US consumers confused over HDTV technology." Retrieved October 25, 2007, from http://www.about-electronics.eu/2007/03/01/us-consumers-confused-over-hdtv-technology/.

[AIM02] (2002, March). Image quality drives online sales. Level of product detail influences customer purchasing behavior (World Imaging Report). *Advanced Imaging Magazine*.

[Ami91] Amidror, I. (1991). The Moiré phenomenon in color separation. *Raster Imaging and Digital Typography II*. R. A. Morris and J. André. New York**:** 98-119.

[Atw07a] Atwood, S. (2007a). "Editorial: HDTV education for consumers should be as clear as the pictures they offer." *Information Display* **23**(2): 2,38.

[Atw07b] Atwood, S. P. (2007b). "Editorial: Welcome to the world of high definition." *Information Display* **23**(11): 2,51.

[Bac99] Baccino, T. (1999). "Exploring the flicker effect: the influence of in-flight pulsations on saccadic control." *Ophthal. Physiol. Opt.* **19**(3): 266-273.

[Bar92] Barten, P. G. J. (1992). "Physical model for the contrast sensitivity of the human eye." SPIE, 57-72.

[BC04] Baron, P. C. and W. Chase (2004). "Variability in susceptibility to the color flash effect in field sequential color displays." *SID Digest*, SID, 338-341.

[Ber96] Berns, R. S. (1996). "Methods for characterizing CRT displays." *Displays* **16**(4): 173-182.

[Ber00] Berns, R. S. (2000). The science of digitizing two-dimensional works of art for color accurate image archives - concepts through practice. Rochester, Munsell Color Science Laboratory**:** 34 pp.

[Bes05] Besuijen, J. (2005). "The shifting gamma perception." *Color Research and Application* **30**(5): 332-340.

[Bes06] Besuijen, J. (2006). "Black level offset: Characterization and correction." *Journal of the SID* **14**(10): 895-903.

[Bes07] Besuijen, J. (2007). "Visual gamma measurement and methods to compare gamma models." *Journal of the SID* **15**(8): 611-623.

[BFT03] Berns, R. S., S. R. Fernandez and L. Taplin (2003). "Estimating black-level emissions of computer-controlled displays." *Color Research and Application* **28**(5): 379-383.

[BJB01] Baccino, T., W. Jaschinski-Kruza and J. Bussolon (2001). "The influence of bright background flicker during different saccade periods on saccadic performance." *Vision Research* **41**: 3909-3916.

[BK97] Berns, R. S. and N. Katoh (1997). "The digital to radiometric transfer function for computer controlled CRT displays." *CIE Expert Symposium '97 Colour Standards for Imaging Technology*, Vienna, Austria, CIE, 34-37.

[BMG93] Berns, R. S., R. J. Motta and M. E. Gorzynski (1993). "CRT Colorimetry. Part I: Theory and practice." *Color Research and Application* **18**(5): 299-314.

[BNS99] Bukal, B., K. J. Nemeth and R. Safaee-Rad (1999). "Novel electron beam profile and convergence measurement method." *SID Digest*, Society for Information Display, 886-889.

[BPI02] Brainard, D. H., D. G. Pelli and A. Ingling. (2002, 12 August '02). "Psychophysics toolbox." Retrieved May 8, 2003, from http://psychtoolbox.org/.

[BPW89] Beaton, R. J., N. A. Perrin and N. Weiman (1989). "Multidimensional scaling of flat-panel image quality." *SID Digest*, Baltimore, Society for Information Display, 120-123.

[BR97] Boschman, M. C. and J. A. J. Roufs (1997). "Text quality metrics for visual display units: II. An experimental survey." *Displays* **18**: 45-64.

[Bri70] Brindsley, G. S. (1970). *Physiology of the retina and visual pathway*. London, Edward Arnold Ltd.

[BS97a] Besuijen, J. and G. P. J. Spenkelink (1997a). "User tests for ISO 9241-3." *Work with display units*, Tokyo, 103-104.

[BS97b] Besuijen, J. and G. P. J. Spenkelink (1997b). Meetrapport onderzoek flat panel displays. Enschede, Department of Ergonomics, University of Twente, 29 pp.

[BS98] Besuijen, J. and G. P. J. Spenkelink (1998). "Standardizing visual display quality." *Displays* **19**(2): 67-76.

[BS02] Besuijen, J. and G. P. J. Spenkelink (2002). "A classification framework for visual display and image quality." *Eurodisplay'02*, Nice, Le club Visu, 795-798.

[BSB$^+$02] Bodrogi, P., B. Sinka, Á. Borbély, N. Geiger and J. Schanda (2002). "On the use of the sRGB colour space: the 'Gamma' problem." *Displays* **23**(4): 165-170.

[Burn06] Burns, M. (2006, Sep 11th 2006). "5 reasons HDTV might NOT boom." Retrieved October 25, 2007, from http://www.engadgethd.com/2006/09/11/5-reasons-hdtv-might-not-succeed/.

[Burr05] Burr, D. C. (2005). "Vision: In the blink of an eye." *BiolCurr* **15**(14): 554-556.

[BZ81] Boesten, M. H. W. A. and E. v. d. Zee (1981). Psychophysical versus psychometric methods in image quality measurements. *IPO annual progress report*. Eindhoven, IPO**:** 67-71.

[CD01] Colombo, E. M. and A. Derrington (2001). "Visual calibration of CRT monitors." *Displays* **22**(3): 87-95.

[CIE06] CIE (2006). CIE Standard S 014-1/E:2006:Colorimetry - Part 1: CIE standard colorimetric observers, Commision International d'Eclairage, 34 pp.

[DF96] Drury, C. G. and D. R. Forsman (1996). "Measurement of the speed accuracy operating characteristic for visual search." *Ergonomics* **39**(1): 41-45.

[DIC03] DICOM (2003). Digital imaging and communications in medicine (DICOM) part 14: Grayscale standard display function. Rosslyn, VA, National Electrical Manufacturers Association**:** 55 pp.

[Dil92] Dillon, A. (1992). "Reading from paper versus screens: a critical review of the empirical literature." *Ergonomics* **35**(10): 1297-1326.

[DK97] Deguchi, T. and N. Katoh (1997). "The effect of user controls on CRT monitor characteristics." *SPIE*, 219-230.

[DK98] Deguchi, T. and N. Katoh (1998). "The effect of user controls on CRT monitor characteristics." *Color Imaging: Device-Independent Color, Color Hardcopy, and Graphic Arts III*, SPIE, 219-230.

[DL98] Daly, S. J. and H. C. Lee (1998). Visual characterization using display model. *US Patent No. 5,754,222*. United States, Eastman Kodak Company, Rochester, N.Y.**:** 18 pp.

[Dow06] Downen, P. (2006). "A closer look at Flat-Panel-Display measurement standards and trends." *Information Display* **22**(1): 16-21.

[Dru95] Drury, C. G. (1995). Designing ergonomics studies and experiments. *Evaluation of human work*. J. R. Wilson and E. N. Corlett. London, Taylor & Francis, 113-140.

[EBU02] EBU (2002). EBU Technical Recommendation R23-2002: Procedure for the operational alignment of grade-1 colour picture monitors., European Broadcasting Union, pp 5.

[EDUT96] Ergonomics Department of University of Twente (1996). Beeldschermonderzoek faculteit WMW, internal report. Enschede, University of Twente, Faculty of Philosophy and Social Sciences.

[EHS04] Engelsen, D. d., I. Heynderickx and A. A. S. Sluyterman (2004). "Color-gamut expansion in CRTs." *Journal of the SID* **12**(3): 241-250.

[Eng95] Engeldrum, P. G. (1995). "A framework for image quality models." *Journal of Imaging Science and Technology* **39**(4): 312-318.

[Eng04a] Engeldrum, P. G. (2004a). "A short image quality model taxonomy." *Journal of Imaging Science and Technology* **48**(2): 160-165.

[Eng04b] Engeldrum, P. G. (2004b). "A theory of image quality: The image quality circle." *Journal of Imaging Science and Technology* **48**(5): 447-457.

[FHH97] Finlayson, G. D., P. M. Hubel and S. Hordley (1997). "Color by correlation." *The Fifth Imaging Conference: Color Science, Systems, and Applications*, SID, 6-11.

[GF00] Gibson, J. E. and M. D. Fairchild (2000). Colorimetric characterization of three computer displays (LCD and CRT). Rochester, NY, Munsell Color Science Laboratory**:** 40.

[GK83] Groot, J. P. d. and A. Kamphuis (1983). "Eyestrain in VDU users: physical correlates and long-term effects." *Human Factors* **25**(4): 409-413.

[Gla97] Glasser, J. (1997). Principles of display measurement and calibration. *Display systems. Design and applications*. L. W. MacDonald and A. C. Lowe. Chicester, UK, John Wiley & Sons**:** 261-287.

[GAB+87] Gould, J. D., L. Alfaro, V. Barnes, R. Finn, N. Grischkowsky and A. Minuto (1987). "Reading is slower from CRT displays than from paper: attempts to isolate a single-variable expression." *Human Factors* **29**(3): 269-299.

[Gui54] Guilford, J. P. (1954). *Psychometric methods*. New York, McGraw-Hill.

[Har87] Hart, S. G. (1987). Development of NASA-TLX: result of empirical and theoretical research. Moffet Field, CA, NASA-AMES Research Centre.

[HB84] Hedman, L. R. and V. Briem (1984). "Short-term changes in eyestrain of VDU users as a function of age." *Human Factors* **26**(3): 357-370.

[Hey07] Heynderickx, I. (2007). Visuele beleving. Delft, Faculty of Applied Sciences, Delft University of Technology**:** 19 pp.

[HRV97] Hill, B., T. Roger and F. W. Vorhagen (1997). "Comparative analysis of the quantization of color spaces on the basis of the CIELAB color-difference formula." *ACM Transactions on Graphics* **16**(2): 109-154.

[Hun05] Hunt, R. W. G. (2005). "Imaging performance of displays: Past present, and future." *Journal of the SID* **13**(12): 983-991.

[HWH+02] Haak, R., M. J. Wicht, M. Hellmich, G. Nowak and M. J. Noack (2002). "Influence of room lighting on grey-scale perception with a CRT-and a TFT monitor display." *Dentomaxillofacial Radiology* **31**: 193-197.

[IEC98] IEC (1998). IEC 61966-3 3CD Colour measurement and management in multimedia systems and equipment, part 3: Equipment using cathode ray tubes, IEC**:** 36 pp.

[ISO92] ISO (1992). International Organisation for Standardization, ISO 9241-3, Visual display terminals used for office tasks - Ergonomic requirements, Part 3: Visual display requirements, ISO.

[ISO97a] ISO (1997a). Comments for ISO/CD Amd. 1 to ISO 9241-3: User performance test.

[ISO97b] ISO (1997b). International Organisation for Standardization, ISO/FDIS 9241-1, Ergonomic requirements for office work with visual display terminals, Part 1: General introduction.

[ISO97c] ISO (1997c). International Organisation for Standardization, ISO 9241-8, Visual display terminals used for office tasks - Ergonomic requirements, Part 8: Requiremetns for displayed colours, ISO.

[ISO98] ISO (1998). International Organisation for Standardization, ISO 9241-7, Visual display terminals used for office tasks - Ergonomic requirements, Part 7: Requirements for displays with reflections, ISO.

[ISO00] ISO (2000). International Organisation for Standardization, ISO 9241-3, Amendment 1, Visual performance and comfort test, ISO.

[ISO01] ISO (2001). International Organisation for Standardization, ISO 13406-2, Ergonomic requirements for work with visual display based on flat panels, Part 2: Flat panel ergonomic requirements, in press, ISO.

[ISO07] ISO. (2007). "Catalogue TC 159/SC 4: Ergonomics of human-system interaction." Retrieved October 24, 2007, from http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_tc_browse.htm?commid=53372&development=true.

[Jär05] Järvenpää, T. (2005). "Measuring color breakup of stationary images in field-sequential-color displays." *Journal of the SID* **13**(2): 139-144.

[JB98] Janssen, T. J. W. M. and F. J. J. Blommaert (1998). "The semantics of image quality." *The fifth color imaging conference: color science, systems, and applications*: 126-130.

[JLHC02] Jun, X., L. Liu, I. Heynderickx and Y. H. Chun (2002). "Influence of white-point setting and contrast on perceived image quality." *SID Digest*, SID, 816-819.

[Kat02] Katoh, N. (2002). Corresponding color reproduction from softcopy images to hardcopy images. Chiba, Chiba University**:** 150 pp.

[KDB01a] Katoh, N., T. Deguchi and R. S. Berns (2001a). "An accurate characterization of CRT monitor (II) Proposal for an extension to CIE method and its verification." *Optical Review* **8**(5): 397-408.

[KDB01b] Katoh, N., T. Deguchi and R. S. Berns (2001b). "An accurate characterization of CRT Monitor (I) Verifications of past studies and clarifications of gamma." *Optical Review* **8**(5): 305-314.

[Kee02] Keelan, B. W. (2002). *Handbook of image quality. Characterization and prediction.* New York, Marcel Dekker.

[Kim07] Kim, C. M. (2007). "The business of displays: Display technology needs to be better explained." *Information Display* **23**(5&6): 6,81.

[Kle93] Klein, S. A. (1993). Image quality and image compression: A psychophysicist's viewpoint. *Digital images and human vision.* A. B. Watson. Cambridge, MA, MIT Press**:** 73-88.

[KLSH00] Klucharev, V., E. Lyskov, M. Sandström and K. Hansson Mild (2000). "Monitor jitter and human-computer interaction." *Displays* **21**(5): 173-177.

[KPK03] Kim, H.-S., S.-O. Park and S.-H. Kim (2003). "Fabrication of brightness optimizer for CRT color monitors." *Color Research and Application* **28**(6): 468-472.

[KS90] Krantz, J. H. and L. D. Silverstein (1990). "Color matrix display image quality: the effects of luminance and spatial sampling." *SID Digest*, Society for Information Display, 29-32.

[KV04] Klompenhouwer, M. A. and L. J. Velthoven (2004). "LCD motion blur reduction with motion compensated inverse filtering." *SID Digest*, SID, 1340-1343.

[LFGC03] Larimer, J. O., C. Feng, J. L. Gille and V. Cheung (2003). "Judder-induced edge flicker at zero spatial contrast." *SID Digest*, SID, 1042-1043.

[Lip86] Lippert, T. M. (1986). "Color-difference prediction of legibility performance for CRT raster imagery." *SID Digest*, The Society for Information Display, 86-89.

[LTT89] Lourens, J. G., T. C. du Toit and J. B. du Toit (1989). "Assessing the focus quality of television pictures." *Human Vision, Visual Processing, and Digital Display*, SPIE, 35-41.

[LXHY04] Liu, L., J. Xia, I. Heynderickx and H. Yin (2004). "Visibility threshold in sharpness for people with different regional backgrounds." *Journal of the SID* **12**(4): 509-515.

[McC98] McCamy, C. S. (1998). "On the number of discernible colors." *Color Research and Application* **23**(5): 337.

[MIJS04] Meesters, L. M. J., W. A. IJsselsteijn and P. J. H. Seuntiëns (2004). "A survey of perceptual evaluations and requirements of three-dimensional TV." *IEEE Trans. on Circuits and Systems for Video Technology* **14**(3): 381-391.

[Mis04] Miseli, J. (2004). "Motion artifacts." *SID Digest*, SID, 86-89.

[MKL89] Milner, N. P., J. L. Knowles and S. P. Lovett (1989). "Ergonomic investigation of misconvergence on colour visual display terminals." *Displays* **10**(2): 81-92.

[MRB97] Morrone, M. C., J. Ross and D. C. Burr (1997). "Apparent position of visual targets during real and simulated saccadic eye movements." *The Journal of Neuroscience* **17**(20): 7941-7953.

[MS92] Muckler, F. A. and S. A. Seven (1992). "Selecting performance measures: Objective versus subjective measurement." *Human Factors* **34**(4): 441-455.

[MV85] Murch, G. M. and L. Virgin (1985). "Resolution and addressability: how much is enough?" *SID Digest*, The Society for Information Display, 101-103.

[MW91] Murch, G. M. and N. Weiman (1991). "Gray-scale sensitivity of the human eye for different spatial frequencies and display luminance levels." *Proceedings of the SID* **32**(1): 13-17.

[MY03] Miller, M. E. and J. Yang (2003). "Visual display characterization using temporally modulated patterns." *Journal of the SID* **11**(1): 183-190.

[Nis90] Nishiyama, K. (1990). "Ergonomic aspects of the health and safety of VDT work in Japan: a review." *Ergonomics* **33**(6): 659-685.

[Ols95] Olson, T. (1995). "Behind gamma's disguise." *SMPTE Journal* **104**(7): 452-475.

[PA98] Pointer, M. R. and G. G. Attridge (1998). "The number of discernible colours." *Color Research and Application* **23**(1): 52-54.

[Pas90] Pastoor, S. (1990). "Legibility and subjective preference for color combinations in text." *Human Factors* **32**(2): 157-171.

[Pic95] Piccoli, B. (1995). The definition and assessment of quality of images on VDUs. Part I: Relationships with asthenopia. *Work With Display Units 94*. A. Grieco, G. Molteni, B. Piccoli and E. Occhipinti. Amsterdam, North Holland**:** 179-182.

[PKB00] Park, S.-O., H.-S. Kim and J.-K. Baek (2000). "Optimum brightness level and simplified characterization of CRT color monitors." *Color Research and Application* **25**: 408-415.

[PNMC98] Post, D. L., A. L. Nagy, P. Monnier and C. S. Calhoun (1998). "Predicting color breakup on field-sequential displays: Part 2." *SID Digest*, SID, 1037-1040.

[Poy98] Poynton, C. (1998). "The rehabilitation of gamma." *Human Vision and Electronic Imaging III*, San Jose, Cal., SPIE, 232-249.

[Poy02] Poynton, C. (2002). ""Brightness" and "Contrast" controls." Retrieved 2006-05-24, 2006, http://www.poynton.com/PDFs/Brightness_and_Contrast.pdf, 5 pp.

[PSB83] Pastoor, S., E. Schwarz and I. P. Beldie (1983). "The relative suitability of four dot-matrix sizes for text presentation on color television screens." *Human Factors* **25**(3): 265-272.

[RB80] Roufs, J. A. J. and H. Bouma (1980). "Towards linking perception research and image quality." *Proceedings of the SID* **21**(3).

[RB97] Roufs, J. A. J. and M. C. Boschman (1997). "Text quality metrics for visual display units: I. methodological aspects." *Displays* **18**: 37-43.

[RMGB01] Ross, J., M. Concetta Morrone, M. E. Goldberg and D. C. Burr (2001). "Changes in visual perception at the time of saccades." *Trends in Neurosciences* **24**(2): 113-121.

[RKHM00] Rash, C. E., V. Klymenko, T. H. Harding and J. S. Martin (2000). "Do FPDs need new image-quality metrics?" *Information Display*(6): 26-29.

[Rob91] Roberts, A. (1991). Methods of measuring and calculating display transfer characteristics (gamma), BBC R&D**:** 13 pp.

[Rob93] Roberts, A. (1993). "Measurement of display transfer characteristic (gamma, γ)." *EBU Technical Review* **Autumn 1993**: 32-40.

[Rob02] Roberts, A. (2002). The film look: It's not just jerky motion... BBC R&D**:** 16 pp.

[Rog84] Rogowitz, B. E. (1984). Measuring perceived flicker on visual displays. *Ergonomic Health in Modern Offices*. London, Taylor and Francis**:** 285-293.

[SB92] Spenkelink, G. P. J. and J. Besuijen (1992). "Display elements and gaps: a comparison of flat panel display characteristics." *Interacting with Computers* **4**(2): 231-245.

[SB94] Spenkelink, G. P. J. and J. Besuijen (1994). "The perception and measurement of contrast: the influence of gaps between display elements." *Behaviour & Information Technology* **13**(5): 320-327.

[SB95] Spenkelink, G. P. J. and J. Besuijen (1995). Brightness: highest luminance or background luminance? *Work with display units*. A. Grieco, G. Molteni, E. Occhipinti and B. Piccoli. Amsterdam, Elsevier Science Publishers B.V.**:** 311-316.

[SB96] Spenkelink, G. P. J. and J. Besuijen (1996). "Chromaticity contrast, luminance contrast and legibility of text." *Journal of the SID* **4**(3): 135-144.

[SB97] Spenkelink, G. P. J. and J. Besuijen (1997). Visual display quality. *Encyclopedia of Microcomputers*, Marcel Dekker. **19:** 383-405.

[SB99] Spenkelink, G. P. J. and J. Besuijen (1999). Visual display quality. *Encyclopedia of Library and Information Systems*. A. Kent and C. M. Hall. New York, Marcel Dekker Inc. **64:** 330-357.

[SBB93] Spenkelink, G. P. J., J. Besuijen and J. Brok (1993). "An instrument for the measurement of the visual quality of displays." *Behaviour & Information Technology* **12**(4): 249-260.

[SBR00] Spenkelink, G. P. J., J. Besuijen and P. Ransijn (2000). "On site measurement of visual display quality." *SID Digest*, Society for Information Display, 176-179.

[ShB95] Sheedy, J. E. and I. L. Bailey (1995). Symptoms and reading performance with peripheral glare sources. *Work with display units 94*. A. Grieco, G. Molteni, B. Piccoli and E. Occhipinti. Amsterdam, North Holland**:** 77-81.

[SHIJ07] Seuntiëns, P. J. H., I. Heynderickx and W. A. IJsselsteijn (2007). "Capturing the added value of 3DTV: Viewing experience and naturalness of stereoscopic images " *Journal of Imaging Science and Technology* **submitted**.

[Shu80] Shurtleff, D. A. (1980). *How to make displays legible*. La Mirada, CA, Human Interface Design.

[Slu05] Sluyterman, A. A. S. (2005). "Dynamic-scanning backlighting makes LCD TV come alive." *Information Display*(10): 28-31.

[SMIJ05] Seuntiëns, P. J. H., L. M. J. Meesters and W. A. IJsselsteijn (2005). "Perceptual attributes of crosstalk in 3D images." *Displays* **26**: 177-283.

[Sny85] Snyder, H. L. (1985). Image quality: measures and visual performance. *Flat-Panel Displays and CRTs*. L. E. Tannas Jr. New York, VNR**:** 70-90.

[Sny88] Snyder, H. L. (1988). Image quality. *Handbook of Human-Computer Interaction*. M. G. Helander. Amsterdam, Elsevier Science Publishers B.V.**:** 437-474.

[Spe89] Spenkelink, G. P. J. (1989). An experiment on actual displays and simulator: a validation of methods and simulator. Enschede, University of Twente**:** 23 pp.

[Spe90] Spenkelink, G. P. J. (1990). A validation study of the high end simulator, Report of the Esprit Project 1593 on behalf of the EC. Enschede, University of Twente.

[Spe94] Spenkelink, G. P. J. (1994). Visual display units: the perception of image quality. Enschede, University of Twente**:** 181 pp.

[SSW90] Spenkelink, G. P. J., H. v. Spijker and T. N. White (1990). "Evaluation of flat panel display properties on a high fidelity display simulator." *Third Int. Conference on Human Computer Interaction IFIP*, Cambridge UK, 721-724.

[STMV00] Stelmach, L., W. J. Tam, D. Meegan and A. Vincent (2000). "Stereo image quality: Effects of mixed spatio-temporal resolution." *IEEE Trans. on Circuits and Systems for Video Technology* **10**(2): 188-193.

[Swa06] Swann, P. (2006, September 8). "HDTV: 5 Barriers to bigger sales." Retrieved October 25, 2007, from http://www.tvpredictions.com/hdobstacles090806.htm.

[Tas79] Task, H. L. (1979). An evaluation and comparison of several measures of image quality for television displays, Air force aerospace medical research laboratory.

[Tin63] Tinker, M. A. (1963). *Legibility of print.* Ames, Iowa, U.S.A., Iowa State University Press.

[TSM92] Travis, D. S., T. F. M. Stewart and C. MacKay (1992). "Evaluating image quality." *Displays* **13**(3): 139-146.

[TZLH06] Teunissen, K., J. Zhang, X. Li and I. Heynderickx (2006). "Method for predicting motion artifacts in matrix displays." *Journal of the SID* **14**(10): 957-964.

[VES98] VESA (1998). Flat panel display measurements standard version 1.0. Milpitas, CA, Video electronics standards association**:** 281 pp.

[VHFC02] Vicario, E., I. Heynderickx, G. Ferretti and P. Carrai (2002). "Design of a tool to benchmark scaling algorithms on LCD monitors." *SID Digest*, SID, 704-707.

[VHSL03] Vogels, I., I. Heynderickx, S. Swinkels and M. De Lameilleure (2003). "The visibility of a local deviation in luminance and white-point of a display." *SID Digest*, SID, 820-823.

[WA05] Watson, A. B. and A. J. Ahumada Jr. (2005). "Spatial standard observer for visual technology." *2005 IEEE International Conference on Systems, Man and Cybernetics*, IEEE, 1623-1627.

[Wat93] Watten, R. G. (1993). "Reinvention of visual fatigue: accumulation of scientific knowledge or neglect of scientific history?" *Ophtal. Physiol. Opt.* **14**: 428-432.

[WC00] Wang, A.-H. and M.-T. Chen (2000). "Effects of polarity and luminance contrast on visual performance and VDT display quality." *International Journal of Industrial Ergonomics* **25**: 415-421.

[WDK93] Woods, A., T. Docherty and R. Koch (1993). "Image distortions in stereoscopic video systems." *Stereoscopic displays and applications IV*, San Jose, Cal., SPIE, 36-48.

[WR00] Wisnieff, R. L. and J. J. Ritsko (2000). "Electronic displays for information technology." *IBM Journal of Research and Development* **44**(3): 409-422.

[WW04] Wen, S. and R. Wu (2004). "Unique gray method for visually measuring the white point of CRT displays." *Displays* **24**: 135-140.

[YK03] Yoshitake, R. and S. Kubota (2003). "The relationship between pixel density and readability on computer displays - Effectiveness of an anti-aliased font on a high density LCD -." *SID Digest*, SID, 296-299.

Curriculum Vitae

Jacobus Besuijen
Geboren 23 januari 1957 te Ens gemeente Noordoostelijke Polder

Opleiding

1969-1972	VWO Streeklyceum, Ede
1972-1975	VWO Thomas More College, Oudenbosch
1975-1984	Studie Elektrotechniek, Technische Hogeschool Delft (nu Technische Universiteit Delft)

1982-1984	Student-assistent, Thoraxcentrum, Erasmus MC, Rotterdam
1984-heden	Werkzaam bij Universiteit Twente, Enschede
1984-1988	Promovendus, afdeling Meettechniek en Instrumentatie, faculteit Elektrotechniek
1988-1994	Medewerker Onderzoek, afdeling Ergonomie, faculteit Wijsbegeerte en Maatschappijwetenschappen
1994-1998	Medewerker Onderzoek en Onderwijs, afdeling Ergonomie, faculteit Wijsbegeerte en Maatschappijwetenschappen
1998-2003	Medewerker Onderzoek en Onderwijs, afdeling Instrumentatietechnologie, faculteit Toegepaste Onderwijskunde
2003-heden	Universitair Docent, afdeling Cognitieve Psychologie en Ergonomie, faculteit Gedragswetenschappen.

Dit onderzoek is uitgevoerd bij de faculteit Gedragswetenschappen van de Universiteit Twente in Enschede.