

Colour

Sense & Measurement

Richard Kirk

Copyright © Richard Kirk 2022

First published in 2022 by
FilmLight
Artists House
14-15 Manette Street
London W1D 4AP
www.filmlight.ltd.uk

To access an online version of this book, and other related resources,
go to:
www.filmlight.ltd.uk/colourbook

All rights reserved. No parts of this work may be reproduced in any form or by any means – graphic, electronic, or mechanical, including photocopying, recording, taping, or information storage and retrieval systems – without the written permission of the copyright owner and the publisher of this book.

Products that are referred to in this document may be either trademarks and/or registered trademarks of the respective owners. The publisher and the author make no claim to these trademarks.

While every precaution has been taken in the preparation of this document, the publisher and the author assume no responsibility for errors or omissions, or for damages resulting from the use of information contained in this document or from the use of programs and source code that may accompany it. In no event shall the publisher and the author be liable for any loss of profit or any other commercial damage caused or alleged to have been caused directly or indirectly by this document.

ISBN (PDF) 978-1-7396839-1-7

Printed and bound in the United Kingdom

Contents

Chapter 1

Introduction: What is seeing?	1
The Numskulls	1
Transcranial Magnetic Stimulation (TMS)	2
The seat of consciousness?	2
The pandemonium architecture	3
Learning to see	4
A simple scene	4
The effect of the surround	6
The effect of the shape	7
The 2-degree standard observer	8
The Checkershadow Illusion	10
Other senses	13
Colour constancy	15
Seeing through a window	17
Summary	18

Chapter 2

Measuring light	21
The colours in the spectrum	21
Early wavelength measurements	23
How much light?	27
The spectroscope	31
The spectrometer	33
Broad spectrum peaks	33
Reflection and transmission spectra	34

Chapter 3

Measuring what we see	37
Mixing light	37
The three-colour approximation	39
Deviations from the three-colour approximation	40
The Bezold-Brücke effect	44

The Abney effect	45
CIE 2006 LMS 2-degree standard observer	48
The CIE 1931 XYZ 2-degree standard observer	51
10-degree standard observers	53
Light units	56
Modern colour measurement instruments	58

Chapter 4

Uniform colour spaces	61
Brightness, hue, and saturation	61
The Munsell Colour Atlas	63
Munsell Value, Hue, and Chroma	64
Ambiguous Colour Terms	65
Yxy colour space	66
Yu'v' colour space	67
White light	68
The Munsell benchmark	71
Display gamma	74
Scene and display	75
Gamma test image	76
CIE L*a*b* (CIELAB)	77
tRGB from CIE 2006	79
tLab derived from tRGB	86
Just Noticeable Differences (JND)	88
Delta E (ΔE)	90
Dominant wavelength	92
Uniform colour spaces	92

Chapter 5

Matching different media	95
Matching contrast	95
Ansel Adams zones	96
Surround, and the crispening effect	97
Correcting for shadow textures	98
Colour gamut correction	99
Film and scene highlights	101
HDR highlights	101
Video zebra stripes	102
The Sun at the Zenith	103
Painting Highlights: Constable & Monet	104
The Empire of Light	105
Highlight compression in TCAM	107

Highlight bleach in TCAM	108
Making HDR from video	109
Compressing the colour gamut	110
Under Colour Removal (UCR)	111
Metamerism	114
Colour Rendering Index	115

Chapter 6

General colour transform tools	117
Quantization	117
1D LUT	119
3D LUT	120
Triangles	123
Capturing LUTs	124
List transforms	124
Zview viewer	126
Spectra	126
Smooth spectra	126
Changing the illuminant	129
Transforming between XYZ and LMS	131
The von Kries model	133
Advanced colour appearance models	134
Camera and scanner calibration	135
Scanning watercolours and acrylics	136

Chapter 7

High gamut and HDR displays	137
Display gamuts and Pointer's gamut	137
Seeing reflection colours	140
New primary wavelengths	142
Colourfulness and the Hunt effect	144
White primary	144
Dithering shadow tones	145
Display calibration	147
Using tRGB as a grading space	149
HDR and high gamut LUTs	149
Correcting for IMAX	150
Limits of contrast	151
Spatial resolution	153
Hyperacuity, Vernier acuity	156

Chapter 8

More on vision	157
The human eye	157
Illusions	158
The blind spot	158
Saccades and the importance of motion	159
Central and peripheral vision	160
The vision pathways	162
Hering opponent colours	165
Adaption illusions	166
Illusions without adaption	167
The McCollough effect	168
Hering opponent based colour spaces	169
Evolution of primate vision	172
Seeing polarization: Haidinger's brushes	172
Seeing magnetic fields: cryptochromes	173
Colour correction and 3D	175
Stereo: the camera adds ten pounds	177
Highlights and flare	178

Chapter 9

Photography and cinema	181
Light sensitivity of early emulsions	181
Colour balance for black and white film	183
Film exposure units	184
Shutter angle and exposure time	184
Black and white film tone curve	188
Reversal film processes	189
Densitometers	190
Cineon negative density space	193
Film lab exposure units	195
The Callier effect	196
Microdensitometers and film scanners	198
Anti-halation backing	199
Chemical edge effects	199
Projector lamp calibration	199
Screen gain	200
Lighting and mireds	202
The camera of the future?	203

Chapter 10

Modelling light and materials	207
Fluorescence	207
Polarization	209
Iridescent colours	210
The colour of the sky	211
Monte-Carlo simulation	214
Beer-Lambert law & additivity failure	215
Phong shading	217
Kubelka-Munk model	218
Neugebauer model	218
Yule-Neilsen model	219
BSSRDF	220
Skin – a hard problem	221

Chapter 11

Light waves and particles	227
Thomas Young and the Wave Nature of Light	227
Maxwell's equations	230
Knife-edge diffraction with waves and particles	231
Bell's inequality	235
What actually happened to 'classical' physics?	238
The photoelectric effect	243
The Stele of Lady Taperet	243

Chapter 12

What is seeing?	245
The Stroop effect	245
The Sapir-Whorf hypothesis	246
Learning to see	250
'That Dress' image and other ambiguities	252
Problems with colour wheels	254
The Natural Colour System (NCS)	258
The colour sphere	259
What is 'understanding'?	260
Can we understand intelligence?	266

What is seeing?	268
Appendix A	
Matrix maths	271
Suffix notation	271
Transpose matrix	272
Identity matrix	273
Inverse matrix	273
Appendix B	
Other tRGB space tools	275
CRI and tCRI	275
Colour interference from a few wavelengths	278

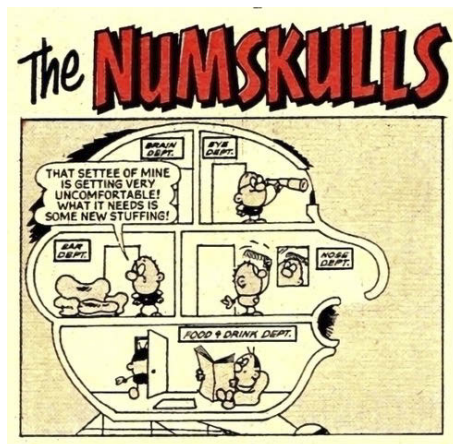
Chapter 1

Introduction: What is seeing?

This is not going to be easy. An easy explanation might liken the eye to a camera and the brain to a computer, talk about “electromagnetic radiation”, then “the visible spectrum” and stuff like that. But this would be wrong: seeing is not just detecting light, it is the whole process that leads from stimulus to understanding. There is a lot to understand, and a lot that is not understood well. Before we climb this mountain, let us have a quick look from a distance, and see whether we can spot a safe route to the top...

The Numskulls

When I was young, ‘The Beezer’ comic (Every Thursday, 4d) had a strip called ‘The Numskulls’. These were a group of little people that lived inside a man’s head...



The Beezer No 409, 16 Nov 1963

The Numskulls each had their own place, and their own job, but they could also move around in the head, and at a pinch one could take over the job of another. Most of the cartoon took place inside the head; but often the final panel showed the man from outside, with a bubble showing his thoughts. The thoughts of the Man, and the actions of the

individual Numskulls inside him were presented as being equally real. The stories were pretty trivial, but the concept was wonderful.

Transcranial Magnetic Stimulation (TMS)

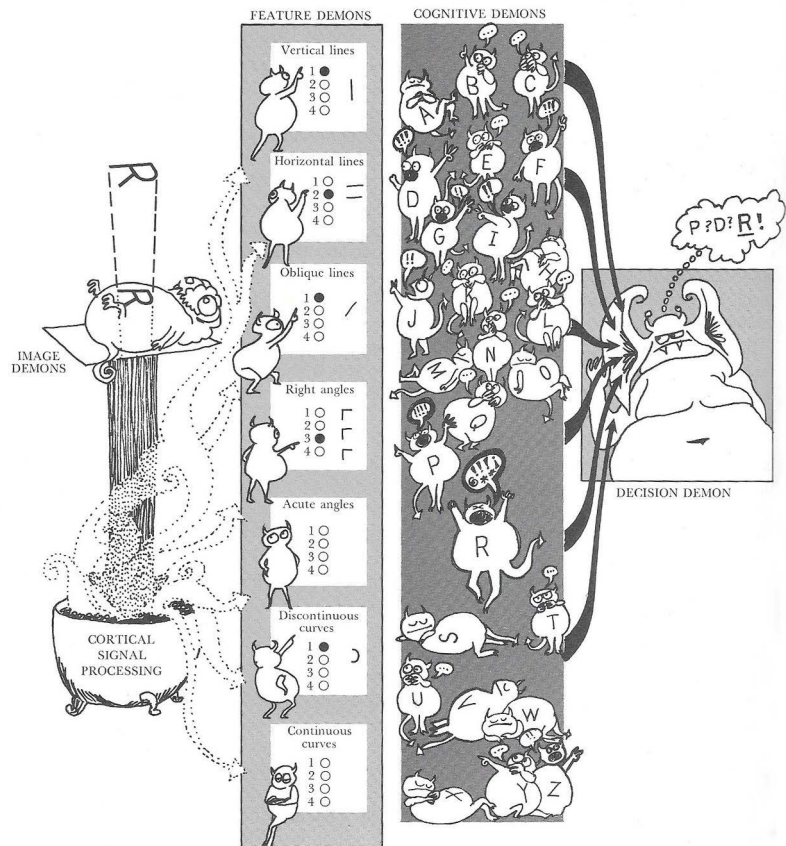
We first learned which bits of the brain do what from people who lost particular skills following very localised head wounds. The modern, kinder equivalent is Transcranial Magnetic Stimulation or TMS. TMS can temporarily stimulate or disable small regions of the brain with sudden changes in magnetic field. Point it at some parts of your head, and you lose the ability to see colour, recognise expressions, or understand letters. Other places cause us to do things, such as put an arm up in the air. And when that happens, you do not feel someone yanks your arm up: it goes up “because I wanted to” (as described to me by someone who had exactly that happen to him). With TMS we can map out what the specific parts of the brain do.

The seat of consciousness?

It is tempting to assume this mechanical information processing is all each bit does. If so, where does the actual thinking happen? Is there some central processing unit in the brain they all report to? If the recognition of colour, shape, motion, and other properties of a single object all happen in different parts of the head, how do the signals come together to give a single understanding?

There is no obvious seat of consciousness within the brain. Descartes thought consciousness was the immaterial soul, communicating with the material brain via the pineal gland. We can assume, as Descartes did, that the conscious lies outside the visible body; or that it may be somewhere physical but just not where we are looking. Neither assumption explains anything: it just puts off having to think about consciousness itself. Nevertheless, the idea that some sharp boundary exists between signal processing and signal understanding keeps coming back to both laymen and experts. Daniel Dennett (‘Consciousness explained’, 1991) likened this concept to imagining a little person (homunculus) in our head, or a little brain within our brain doing the thinking. This is like the conundrum “how do you know you are not a brain in a jar, and all your world is a simulation?”, only turned inside-out. It is a fun train of thought, but it does not explain anything. A better hypothesis, and one that can be proved or disproved, is that consciousness is distributed among the parts that we can see. An ant colony can be smart, but an individual ant is stupid. The consciousness of The Man comes from the actions of the Numskulls within him, yet his thoughts do not belong to any Numskull in isolation.

The pandemonium architecture



'Human Information Processing', P.H.Lindsay and D.A.Norman, 1972

We are not quite done with little creatures in our head. The 'Pandemonium Architecture' is a lower level model for how we might, for example, interpret the signals from the cells in our eyes. This is presented as lots of demons, each doing a very simple and specialised processing task. Each 'demon' is probably a few nerve cells. They are not self-aware. They are probably not as distinct as the model suggests. Nevertheless, describing thought processes as demons, each with a job to do, can help explain how vision might work...

The first level of demons that take the raw image and turn it into signals (the cortical signal processing) are not shown in detail. The next level of demons identify salient features that may belong to letters, and make a noise if they are finding the things they are looking for. The next level of demons above that listens to the previous level, and if what they hear has the right balance for their particular letter, then they make a noise. Finally, there is a decision demon that listens for the loudest demon in the previous layer, and chooses their letter.

Learning to see

You probably do not read this text that way. Someone learning to read starts by recognising letters as individual symbols, but quickly moves to recognising short sequences by their shape, and then anticipating the next bit from what has gone before.

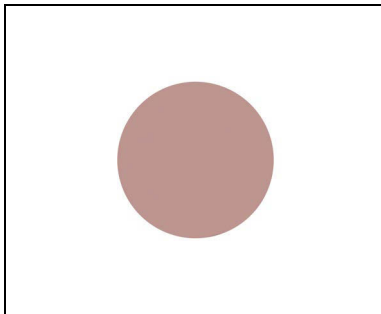
Paris in the the spring-time.

We can tell this by the mistakes we make. Someone learning to read would spot the extra ‘the’ in the text above, but an adult might not. However, the pandemonium model shows why we make particular mistakes, and suggests how a robust character recognition program might work. Artificial intelligence also shows what these simple models can do.

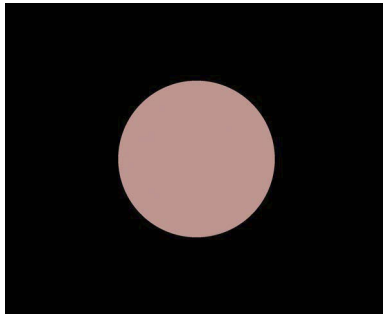
There is another pernicious weed of an idea that keeps coming back. This is that vision is only a bottom-up process, and information only flows one way from low, primitive levels to higher, more sophisticated levels, and eventually to understanding. Information must go the other way too, if only so our ‘R’ demon knows what set of features to look for. If this did not happen, we would have to be born knowing how to read. We even have to learn how to see. People who had cataracts since birth do not suddenly see after an operation: they gradually learn to see. Most of us use similar parts of the brain similarly when we see, so it seems likely the brain’s architecture guides this training process. But there is some real scope for each of us to see in our own way.

A simple scene

Enough theory; let’s look at some colours. Here is a colour. What colour is it?



I see the pinkish-grey colour of sticking plasters that is not quite the colour of anyone's skin. Perhaps we are distracted by the surrounding white. Let us try a black background instead.



This does not look very different. Perhaps the pink looks a bit brighter because of the dark surround. Let us look at the original object against a black background.

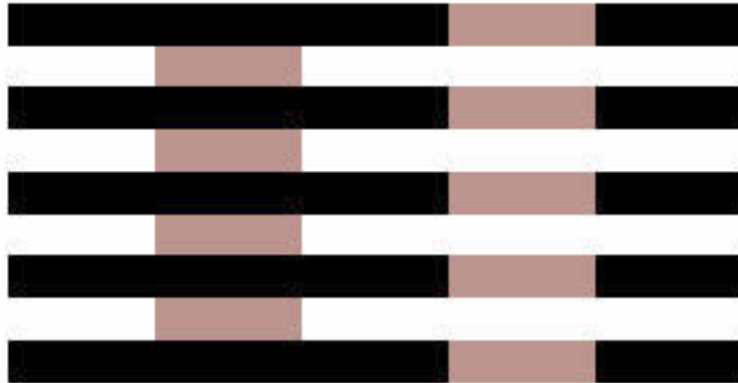


Suddenly the colour looks very different. We know the object is a penny (or some small coin, at least), and it is copper coloured. Fresh copper is bright pink, and goes brown as it tarnishes. It looks shiny here, but this might just be the lighting.



Now we have a real scene, albeit a simple one. The background gives few clues, but the other penny shows the original one is very shiny, even if the date tells us it is over 20 years old.

We started by replacing a white surround with black. This did not make much difference, but there are other arrangements of black and white that do. Here we have the same pinkish-brown colour in an arrangement called White's illusion. The pink-brown patches are all the same colour but a patch on the left looks like the tarnished penny, and a patch on the right looks like the shiny one.



White's illusion

On the left, we see the colour against white, seen through black stripes. On the right, we see the colour against black, seen through white stripes. Changing the background did not make much difference when we were looking at circles, but now it does.

The effect of the surround

We saw the same colour stimulus look like different colours depending on the surround. Suppose we got rid of the surround, and filled the whole field of view with one colour. Would this show us what the colour *actually* is? If you painted a whole wall or ceiling with that colour, it would probably look lighter and pinker than the original colour swatch. I know this because my father did exactly that in the house I grew up in. Twice. The paint in the tin matched the colour swatch, but when we look at a large, featureless patch of colour, we also use our peripheral vision, and this sees colours slightly differently to our central vision. And yet, if we put the small swatch next to the wall, the colours match.

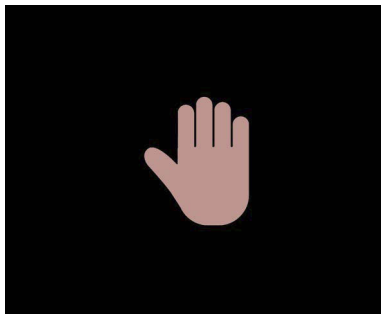
We also saw a similar thing a few years ago when people started using LCD (Liquid Crystal Device) monitors for grading images alongside CRTs (Cathode Ray Tubes). You could set up two computer displays with instruments. You could display a small colour patch on both monitors, and it would match. But if you displayed full-screen white, the LCD looked orange next to the CRT. Not very orange, but enough to worry people

who work with colour. When we looked at the small patch, we used the centre of our vision. When we looked at the whole screen, we used our peripheral vision too. We rarely notice this in everyday life, but it can matter.

The other thing we can take away from this is that our eyes and brain are not at all like a camera and a computer. The full screen matched the small patch on both displays. The two small patches matched each other. So the two full screens must match, right? But they do not. Paradox! If we display an image on the two displays, this hides this difference somewhat: each colour we look at seems the same, but we may still have a nagging feeling that the overall colour balance is somehow out. You get the same thing when trying to match a computer display to film.

The effect of the shape

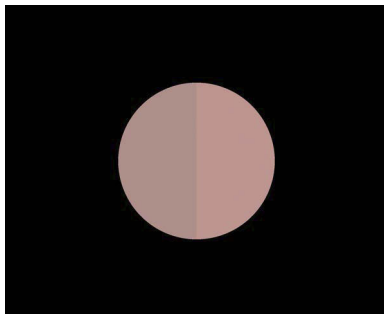
There seems to be no scene that shows us the colour as it *actually* is. However, we can look at a small patch of flat colour against a broad, black surround. If the patch looks about as big as your thumbnail at arm's length, it matches the central region of the eye we use for precision vision.



No, this one won't do. This shape suggests a hand, which can influence our interpretation: we think of flesh colours, and this biases our judgement. Prof. Gegenfurtner et al. showed that patches the shape of fruit and vegetables heavily biased our names for a colour towards or away from the expected colour for the shape. Let us use a circle as our reference shape as a circle is fairly neutral. If we replace a circle with a banana shape, near-yellow hues look more yellow, and greys look more blue (less yellow). This worked for other fruit shapes too. The bunch of grapes shape had two contrasting effects depending on whether people thought of green or purple grapes.

The 2-degree standard observer

The '2-degree Standard Observer' is used for many vision experiments.

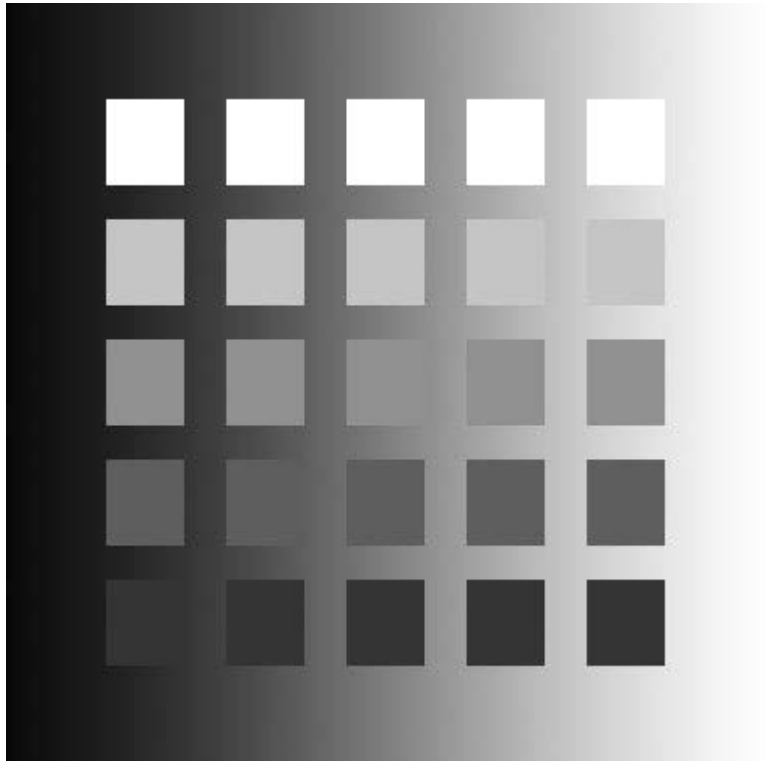


2-degree Standard Observer

The observer sees a circle of a colour with a broad dark surround. The circle may have two separate halves to compare two colours. The observer views the circle, so it subtends an angle of 2 degrees at their eye. This is roughly a thumbnail at arm's length. The black surround does not have to be absolute zero light, provided it is a lot darker than the stimulus. This does not show us what the colour *actually* is. There is probably no arrangement that says what the colour *actually* is, but the 2-degree Standard Observer lets us view any given colour, or see whether two colours look the same, and get a consistent response.

Here are two colours that may look like the pinks from White's illusion earlier. We have the original colour on the right, and a slightly darker version on the left. However, if you took the actual colours from the White's illusion image, they would match because they are the same.

The 2-degree size works well because it is about the size of our central vision region, which we use for most detailed vision. When we look at a 2-degree circle, our eye naturally centres on it. If we use a 10-degree circle to include the peripheral vision as well, the eye wants to wander about within the circle, and it is hard to get a consistent response.



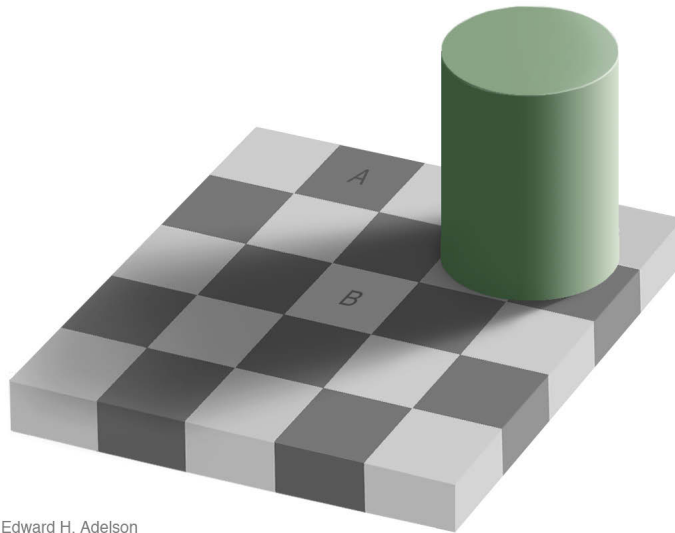
Here we have flat, grey squares against a graded background. The squares in each row are exactly the same colour, but they appear lighter against a dark surround, and darker against a light surround. Surround is still important with our 2-degree observer. We choose black for our 2-degree observer standard as everyone agrees on black, even if black is probably darker than the surround when we see these colours in a scene.

The 2-degree observer view is so simple that there is almost no 'seeing' going on. This is useful because it lets us test whether two colours get the same response without the distractions we normally get from surrounding detail. We use views like this to show our eye uses three sensors for ordinary colour vision; to infer which colours of light these sensors react to; and to see whether these sensors are the same for different people. We shall go into the details of light stimuli and how we measure them in the next chapter.

The Checkersshadow Illusion

Now let's look at an image with recognisable objects...

Here's a famous illusion published by Prof. Adelson of MIT in 1995.



Edward H. Adelson

The Checkersshadow Illusion

The two squares marked 'A' and 'B' are the same colour. If we put the 'A' and 'B' grey into the two halves of the 2-degree observer target, we would not see the dividing line. However, we see them as very different shades of grey. We might write down what we see like this...

- a** There is a two-tone chequered surface. We expect all the light squares to be the same, and all the dark squares to be the same.
- b** There is a green thing. The shape suggests a cylinder. The shading confirms a cylinder, and that the light is coming from the top right. The shading on the sides of the chequered surface backs this up.
- c** The green cylinder is apparently casting a soft shadow on the chequered surface.
- d** Square 'A' is a black square in the light; square 'B' is a white square in shadow.

Seeing is no simple bottom-up process, but many processes all happening at once. We guessed the 3-D shape from the lighting, and the lighting from the 3-D shape. We distinguished the checker pattern from the shadow, perhaps before we were aware there was a shadow. We might know there was a shadow but ignored it since we often want to recognise things rather than appreciate the play of light and shade.

The more sophisticated of us know this is an illusion, and may guess 'A' and 'B' must be the same because that is what illusions do. But they still look different.

At this point, let us consider how we use words when describing patches 'A' and 'B' in the Checkershadow illusion. Some rigour here may help us later.

I might say 'A' **looks** darker than 'B'. Here I am describing how the two colours appear to me in the particular context. Or, I might say 'A' looks **dark grey**, and 'B' looks **light grey**. More rigorously still I might say 'A' looks dark grey, and 'B' looks light grey **to me**, allowing for the possibility that you may see something different...



'That dress'

The famous 'dress' image is a rare example where people can see quite different things. The original dress is blue and black, but I am one of those who consistently see it as being blueish-white and gold.

Suppose I said 'A' and 'B' are the same colour, or 'A' and 'B' matched. If 'A' and 'B' were both printed on paper, I might try to prove this by cutting out the two regions, and sticking them next to each other. I might show them against a plain black background to remove the influence of the surrounding image. I could even cut the 'A' and 'B' to fit in the 2-degree Standard Observer circle. I could then say 'A' and 'B' **matched according to the 2-degree Standard Observer viewing conditions**.

This last statement may seem pedantic: if we see no contrast when 'A' is next to 'B' then they must match, right? However, if 'A' and 'B' were two different materials that needed to match, we might check them under daylight and artificial light. The 2-degree size is also important, as we discussed when painting walls, or comparing white on displays.

I say 'A' and 'B' **are the same colour** if they have the same brightness at all visible wavelengths. If 'A' and 'B' are reflection colours, they reflect the same fraction of light at all visible wavelengths, and will match when they are both viewed under the same illuminant. They are exactly the same colour in all respects, and they will be a match to instruments and people with non-standard vision.

I say 'A' and 'B' **match** if they have no visible contrast, even if they may not have the same brightness at all visible wavelengths. This is an incomplete statement as I have not specified what the matching conditions are, and you should look in the context for the matching conditions.

I say 'A' and 'B' **match to me with the 2-degree viewing conditions** if they do indeed look the same to me under these conditions. Or, perhaps, 'A' and 'B' **match according to me when I cut them out and place them next to each other**.

I say 'A' and 'B' **match according to the CIE 2006 standard 2-degree observer** if their measured spectra give the same values using this particular standard. This standard uses a measuring instrument, typically a spectrometer, instead of my eye, and uses a physical model of human colour perception. Here I use a standard that matches a typical observer with good vision when comparing colours using the 2-degree geometry. The CIE 1931 XYZ 2-degree observer is another, older but still widely-used standard. These standard observer measurements will not exactly fit my vision or your vision, but they are close enough to be useful.

I say 'A' **looks like** 'B', or 'A' **looks darker** than 'B', this is how they appear to me. I have not tried to alter the presentation or to bring the two patches together.

You may not need these exact definitions. Most people describe colours without trying to be rigorous. But if someone says 'A' and 'B' match, ask them what their matching conditions are.

Other senses

Let us take a break from vision, and talk about taste for a bit. Do you like the taste of Brussels sprouts? This is a vegetable that looks like a tiny cabbage, and has a distinctive, strong taste. They can be picked through winter, so they often appear with a traditional Christmas dinner in the UK. If you are not familiar with sprouts, imagine this argument with something else, such as blue cheese, and all will be fine.



Brussels sprouts

Suppose I eat a Brussels sprout, and you eat a Brussels sprout. We both taste Brussels sprout, right? Or rather, we both taste Brussels sprout as mediated by our teeth and our saliva, and the whole of our complex eating and tasting mechanism. Saliva is important; I have known people who have lost saliva for a time, and they have found everything tastes very different, and usually rather nasty. Nevertheless, we could in theory trace the physics and chemistry of a Brussels sprout all the way from popping it into our mouth to the chemical stimulus in the cells of our organs of sense or taste.

We can go a bit further than that. We can probably measure the response of the cells themselves. Most people describe cilantro, or coriander leaves, as having a fragrant, citrus-like flavour. A small but significant fraction of people describe it as 'soapy' or 'foul'. I cannot smell musk, though I can smell most other things. You can try to describe it to me, but the descriptions ('sweaty', 'animal') make no sense to me. There are known genetic variations that cause us to disagree over coriander and musk. We can find which chemicals stimulate a particular response in a chosen detector. We can then map out the territory of all possible tastes and smells. This is easiest if you stick to people with 'normal' senses, but if you can find enough people with a particular abnormality, then you can

predict what they think too. This is not a philosophical fancy: people have done just this, and managed to produce and predict 'taste illusions' and 'scent illusions'.

What happens to the nerve signals then? We can trace them from cell to cell a small way using electrical probes. After tracing the signal through a few nerve cells we rapidly become mazed, and no longer know what each signal means. The signals disappear into our neural tangle, and the sense of taste or smell appears to our consciousness. When the taste of Brussels sprouts appears, it appears with all our personal metadata because we always had Brussels sprouts with Christmas dinner, or we had to eat them at school.

I am not saying that any sensation we agree on necessarily implies a common sensor, and any difference is a consequence of our experiences. Not quite. We do not all agree on the taste of coriander leaves, for example. If we have many sensors, and they are all different in the same way for a few of us, it is likely there is a common cause. If there is a genetic cause, it seems we can often find it. But it is still possible that some of us have our tastes wired up differently for other reasons.

Now back to vision and colour. There are three types of cone cell in the human eye that detect light and turn it into nerve signals. This is the first stage in our colour vision. We can measure the sensitivity of each type of cone cell to each wavelength of light. We can also work out how much of each wavelength is absorbed by the other parts of the eye on the way in, such as the macular dye and the eye lens. We can even go beyond the standard 2-degree observer models, and measure how the cells respond in the peripheral vision.

Exposure makes the cone cells fire. More exposure makes them fire more often. In theory, we can predict the cone cell firing rate from the exposure. In practice this is difficult because the firing rate depends on other factors such as the exposure that has gone before. A cell bleached by high exposure will not fire as much as an unexposed cell. The eyes move, so any particular cell may not point consistently at the same part of the scene.

The signals travel to other nerve cells. We can follow the signals using electrical probes, and we can trace the nerve connections using optical coherence tomography (OCT). At the next level of nerve cells, things get much harder to model, as each signal is modified by the outputs of the nearby cells. Pretty soon we have lost any obvious relation between our nerve signal and the light level exposures at the cone cell. The signals disappear down the rabbit-hole, and the sense of colour appears to our consciousness. If we look at a Brussels sprout, it will look green but with all the overlaying personal modifications because we always had Brussels sprouts with Christmas dinner, or we had to eat them at school.

So far we see no clear distinction between detection and consciousness. The interpretation process may start when two nerve cells meet. It is quite likely that the initial processing levels are fairly mechanical. This is somewhat similar to using neural nets for image recognition. These may have many layers, but instead of training the whole system from scratch, users may start with the initial layers pre-trained for general feature detection. They can then train the last few layers to pick out faces or bicycles or fruit, or whatever they are looking for. Nevertheless, some feature recognition and interpretation begins before the nerve signals leave our retina. If we assume these interpretation processes are wholly mechanical, then we are back with the homunculus fallacy.

We may recognise no sharp barrier between detection and consciousness, but there is a sharp barrier in what we can measure right now. We can measure what stimulates a receptor for taste, vision, touch or whatever, but it is hard to meaningfully trace individual nerve signals. We can generally measure and quantify whatever triggered the original detecting cell, and term it a 'stimulus'. The stimulus goes into a complex process where signals pass from nerve cell to nerve cell. This is where the Numskulls live and have their being, and the consciousness gets an awareness of what is being tasted, seen, or felt.

Colour constancy

This 'complex process' of understanding what we see is not wholly mysterious. We could in theory model any particular cell. We can do experiments using TMS and other tools, and see what is happening at a larger scale — we interpret depth and stereo *here*, and recognise faces and expressions *here*, and so on. We can recognise the mechanical processing, even if we cannot measure the consciousness. Here's an example: we usually recognise objects and see their colours consistently even if the lighting changes. We can do this even if we can't see the illuminant directly. This trick makes the Checkershadow illusion work, and it is called *colour constancy*. If we did this with a colour camera and a computer, the camera would measure absolute RGB, and our computer would scale them to correct for our guess of the lighting white. Our eye and brain probably lost any good measure of absolute exposure even before the signal left the cone cells, yet it still manages to do the same thing, and does it rather well.

Colour constancy also depends on the realism of the display we are looking at. We may have...

- 1 Simple colour shapes with surrounds.
- 2 Simple poster-like images with flat colours.
- 3 Images with shading that look like a view of a scene.
- 4 Images that can be mistaken for a photo of a scene.
- 5 Immersive environments, and real scenes.

If we look at simple shapes, we may be weakly influenced by the surrounding colours. As we go down this list, colour constancy becomes stronger. 'Checkershadow' is 3 on this list: we do not think this is a real scene, but we follow the obvious light and shading cues. I would put 'That Dress' at 4: it is a cropped photo of a real scene, but it could have been constructed as an illusion. We are following non-obvious lighting cues, and making mistakes with close to full colour constancy. It seems we need a realistic image but do not have to believe we are looking at a real scene. Colour constancy is not something that happens in the retina; it requires a deep interpretation of the whole image.

Until we can understand our whole system including our consciousness, I can argue...

- We need to recognise things. We should not be confused if the lighting changes, or if there are several light sources.
- We can model how the lighting affects what a camera would 'see' and correct for it.
- We guess that our eyes and brain need to achieve the same goal. They must do it by completely different means to digital cameras. But if we understand a camera correction process, we may use it to model the eye and brain too.

This is not a great piece of reasoning. People have also argued...

- Old world monkeys had two-channel colour vision.
- New world monkeys evolved separate red-green receptors to have three channels.
- Being able to distinguish reds and green helps us identify fruit amongst leaves.
- Monkeys evolved three-colour vision to gather fruit.

This may not be wrong, but it is flawed logic. Being able to distinguish reds and greens may help when gathering ripe fruit, but this is when fruit is plentiful, and there is enough for all. We need to see whether it helps when food is scarce, or to avoid threats such as poisonous berries or camouflaged predators. It may do all of these, but we have not shown it by experiment.

My argument was over-simple but, unlike the monkey vision argument, we can prove the important steps. Others have shown we can recognise colours very accurately under different lighting conditions. We can argue we have this ability because it is useful, or was evolved for survival, but this argument is no longer as important as before. It seems likely our vision relies on recognising objects and seeing how their colours compare to expectations, while a camera might correct its white point by matching the peaks in its RGB channels. But if both systems seem to do the same job well, we can base a vision model on the simpler camera process that we understand. People have done just this, and shown that we may see the dress as blue and black, or blue-white and gold, but no-one should see any other colours. They have also shown which lighting cues make some of us see the dress as being blue-white and gold, and confirmed this with experiments.

This is not ideal. It would be much better to have an exact model for how we correct for lighting. But that is complex, and I do not expect we will have that any time soon. In the meantime, we model colour constancy as though we were correcting a camera image for uneven illumination. This works and it will do for now.

Seeing through a window

We also appreciate how our eyes and brain work when we see an image that imitates a real scene...



Claude Monet, 'Haystacks'

Monet painted haystacks in the snow, lit by the setting sun and the sky. We know snow is white, yet here it is pink and orange in the sunlight, and blue from the sky in the shadows. The image has the right balance to give a sensible average white. This is sometimes called the ‘grey world’ model. It can be used to give a pleasing colour balance to camera images. However, where do we stop? In this document, the haystack image is surrounded by real white, yet the snow still looks right. We look at the image as though it was a view through a window.

Windows are a significant exception to the global ‘grey world model’. Imagine we are in a room lit with ordinary tungsten-balanced lights, with a window looking out on a scene lit by daylight. A camera with a single ‘grey world’ white balance would show the room as being yellow and the window as blue. But we see the room and the window as separate ‘worlds’ with their own white balance.

We may see a computer monitor as a window. In the 1980s, in the early days of colour management for printing, people tried to match the white point of displays to the ambient lighting. This fitted the grey world model, but it looked odd, with displays looking very orange. If you had no standard lighting booth, it was actually easier and more ‘natural’ to compare a printed sample under ambient office lighting to a computer image balanced to daylight white. Monitors were rarely calibrated back then, but they had knobs you could twiddle to get the right sort of white. Maybe this is what Monet did in his head, so his haystacks looked right on a framed rectangle in a room. I don’t know. But we seem to suspend our overall sense of *what white is* for windows and monitors and other bounding shapes similarly.

Summary

This is a good place to stop. There is much more to say, but we can say it a lot better when we have covered the basics of colour measurement. For the moment, I shall summarize...

- People with normal vision have eyes that react similarly to light. We have standard observer models that predict from spectral measurements which colours should match for the average observer, and these work well.
- We have no measure of our internal sensation of colour. We know the context affects the colours we see in complex ways: the seen colour in a patch depends on whether the patch is banana-shaped, or whether the shape is in shadow. It would be difficult to make a standard observer model that could mimic effects like this.

- The dress image shows we can get different sensations from the same stimulus. This dramatic disagreement is rare, but it can be reproduced and understood. This image is ambiguous because we have to guess the lighting, and different people make different guesses. Real scenes usually have more lighting cues, but there may be more subtle variations.
- We have not yet covered the conundrum “If you and I look at a colour red, do we both see the same red?” We could have very different internal sensations, though I find it hard to see how or why. If we think of green grapes rather than purple when looking at a colour patch like a bunch of grapes, or see the dress as white and gold, we have something to measure. With luck, we may find more cases and learn more. But none of this tells us what red *really* is.
- We make realistic images without reproducing all the colours in the original scene at their original intensities. An image may be dimmer or brighter than the original scene; it may have less colour, and less contrast, and we still feel we have a good match. Monet’s haystacks look good, and he used very few colours. We can get a first-pass match using simple vision models. A skilled artist can take it from there.
- We shall find other problems as we go. We can put numbers to the colours we see with a standard observer model. Illusions show there are many ways to make colours look different, but it is hard to quantify this appearance shift, and we do not know we all see the same thing. There is no Grand Unified Theory that says how everything should look.
- I will try to describe things we can measure: how to make image B look like image A, and things like that. I am not going to talk about colour harmony, or composition, but I may discuss particular images where their composition may illustrate the visual processes that make them work.

Chapter 2

Measuring light

This chapter covers the instruments and techniques we use to measure light. I name the colours I saw, so we can identify the wavelength numbers with visible colour. There are some historical notes on how people have named colours in the past. But that is about it for human vision: this chapter is about measuring, and seeing does not start before Chapter 3.

The colours in the spectrum

I had a tuneable source of single wavelength light. This was an old Perkin-Elmer spectrophotometer I found in a skip. I could not get the measurement bit to work, but the light source was surprisingly accurate. I could turn a dial and get light with a known wavelength. Here are the colours I saw...

Colour	Wavelength	Energy
far red	700 nm	1.771 eV
red	650 nm	1.907 eV
orange	607 nm	2.043 eV
yellow	579 nm	2.141 eV
green	535 nm	2.317 eV
azure	497 nm	2.495 eV
blue	470 nm	2.638 eV
violet	445 nm	2.786 eV
far violet	400 nm	3.100 eV

If physics is not your subject, do not worry about the details. Wavelengths of visible light are often given in nanometres or 'nm'. This typically gives numbers between 400 and 700. We could quote the values in microns, or thousandths of an inch. Think of it as a number for each colour in the spectrum.

Standard wavelengths are measured in vacuum, but we live in air. The refractive index for visible light in air at room temperature is about 1.00027. This would mean that a wavelength of 500 nm in vacuum would be about 499.86 nm in air. The exact wavelength in air depends on the air pressure and humidity. Fortunately, the wavelength difference between air and vacuum is so small that we can quote a wavelength in nm, and not worry whether it is in air or vacuum because the difference will not be visible.

The final column is the photon energy in electron-Volts or eV. This is a less ambiguous measurement than wavelength because it is the same in vacuum, glass, air, or water. It is the energy in a single photon. A typical carbon-carbon single bond has a bond energy of 3.6 to 3.9 eV, so we can see ultraviolet might be harmful to organic compounds. Another good measure would be frequency. Frequency is also the same in vacuum, glass, air, or water; and it gets us straight to one of the fundamental units of measurement. The definition of the SI unit of visible intensity the *candela* (cd) was based on light with a wavelength of 555 nm, and is now based on light with a frequency of 540 THz (555.18 nm). This is the only example I know of light given as a frequency. We shall stick to nanometres from here on.

Here are a few notes on what I saw...

- Far red continued beyond 700 nm, but it got very hard to see. About this point (690 nm according to my notes) the colour appeared slightly more orange than the regular red. This may be real, but you would get a similar effect from a tiny contamination of a shorter wavelength, or even some other process such as the retina responding to the heat in the radiation rather than the photons. Incandescent lamps give off a lot of infrared. Infrared LEDs used for security systems are often quoted somewhere about 900 nm, but have a visible red glow.
- Red looked much the same colour down to about 650 nm. Deep red LED colours look like this.
- There is no obvious 'orange point'. I picked 607 nm as a typical orange.
- There is a fairly clear narrow 'yellow band' with 'golden delicious' green-yellow on one side, and 'butter' orange-yellow on the other side. My chosen 'lemon' yellow was 579 nm.
- There is no obvious 'green' point. There is a range that we would call green. 535 nm is a good 'traffic light' green, somewhere in the middle of this range.
- There was a clear 'azure band' with blue on one side and green on the other. It is not as precise as yellow, but it is still pretty distinctive. I have chosen to call this 'azure' for compatibility with other European languages. 'Cyan' is often used to describe a printing colour, or a display colour, and these two cyans are not quite the same.

- There is no obvious 'blue point'. Half-way between my 'azure' and my 'violet' is 470 nm, which looked to me a perfectly typical blue.
- Violet stopped looking blue at about 445 nm. This and anything shorter is violet.
- Below about 420 nm the violet seemed to lose saturation. We are able to see violet to about 370 nm, but we suffer from the same problems as with far red: we are not sensitive to these wavelengths, and the colour we see may be contaminated by other effects. My experiments were also limited by the light source (incandescent) which has very little violet.

Early wavelength measurements

Thomas Young trained as a doctor and was interested in medicine and anatomy all his life. He was also interested in linguistics, and hieroglyphs and many other things. In 1800, he was preparing a paper on the human voice, and ended up sorting out the fundamentals of sound, then light and vision. In his own words:

“When I began the outline of an essay on the human voice, I found myself at a loss for a perfect conception of what sound was, and during the three years that I passed at Emmanuel College, Cambridge, I collected all the information relating to it that I could procure from books, and I made a variety of original experiments on sounds of all kinds, and on the motions of fluids in general. In the course of these inquiries, I learned to my surprise, how much further our neighbours on the continent were advanced in the investigation of the motions of sounding bodies and of elastic fluids, than any of our countrymen; and in making some experiments on the production of sounds I was so forcibly impressed with the resemblance of the phenomena that I saw, to those of the colours of thin plates, with which I was already acquainted, that I began to suspect the existence of a closer analogy between them than I could before have easily believed.”

This is a good introduction to Thomas Young. He was nicknamed 'Phenomenon Young' when he was at college. He was a well-read polymath, with a clear insight that allowed him to design simple, minimal experiments that clearly demonstrated the basics of light. Unlike many brilliant men before him, he collaborated with other scientists, particularly with Augustine-Jean Fresnel though relations between England and France were less than cordial at the time.

He published the first accurate estimates of the wavelength of light for different colours. His names for the colours, his figures for the

wavelength of light (in wavelength per inch) and my conversions (in nm) are...

Colour	Wavelengths / inch	Wavelength
far red	37640	679 nm
red	39180	659 nm
orange	41610	610 nm
yellow	44000	577 nm
green	47460	535 nm
blue	51110	497 nm
indigo	54070	470 nm
violet	57490	442 nm
far violet	59750	425 nm

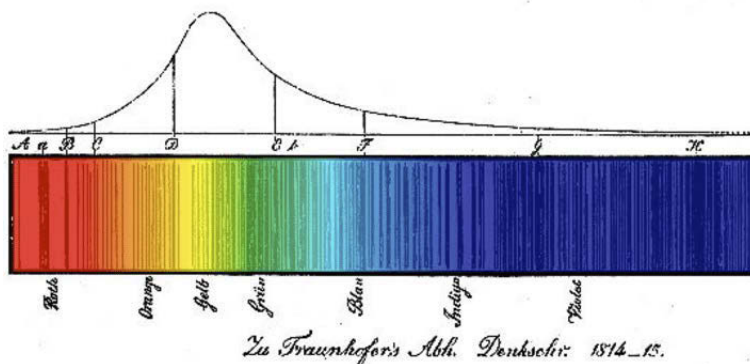
Notice the choice of names for the colours. His 'blue' is my 'azure', and his 'indigo' is my 'blue'.

Isaac Newton in his 'Opticks' of 1730 listed the colours from the prism on page 52 of the Dover reprint as "violet, indigo, blue, green, yellow, orange, red". The inclusion of 'indigo' has been attributed to Newton wanting to have seven colours for numerology reasons. Newton could be mystical, but his 'Opticks' was a plain and straightforward book, written in English rather than Latin. He did have an odd theory that linked colours to a musical scale, so perhaps he wanted colour names to match the notes in a musical scale (this gets a mention in Chapter 12). It may just be that blue, or 'blew' as he spelled it, meant a sky-like 'azure' colour, and he needed a new word for the colour beyond that. In his notebook of 1733 he commented...

"The homogeneous Light and Rays which appear red, or rather make Objects appear so, I call Rubrifick or Red-making; those which make Objects appear yellow, green, blue, and violet, I call Yellow-making, Green-making, Blue-making, Violet-making, and so of the rest."

Newton makes the point that the names belong to the colours we experience. The rays are not coloured, as we might dye cloth. This quote omits 'indigo'. Perhaps this was not a colour word that came naturally to Newton. Maybe 'indigo-making' would have sounded like a dye process. I don't know. Nevertheless, 'Opticks' had a huge impact, and long afterwards when people spoke of colour they often used the same words Newton used.

Here's an illustration engraved by Joseph von Fraunhofer of the solar spectrum he saw in his spectroscope. The colour labels are in German, but 'Blau' has moved up to make room for 'Indigo' (the colours in the diagram are not in his original, but they fit the lines that he labelled).



Fraunhofer's solar spectrum

It is not surprising that Thomas Young used the same colour names that Newton did. But, we no longer commonly name colours that way, so from here on I shall use 'azure' for that colour around 500 nm and 'Blue' for the 470 nm colours.

While we are talking about names, there is another problem colour name. I found a few people who disagreed with the term 'violet'. There is an African Violet with a deep, intense colour, and a European Violet with a pinker colour. There is also Gentian Violet, which is an artificial dye with antifungal properties which is more like my short-wavelength violet. From now on, I am going to use 'violet' for short wavelengths only: we shall meet purple, mauve, magenta, and pink later. And, if you are talking colour with a botanist or a gardener, you may have to check this one.

Back to Thomas Young's measurements: J.D. Mollon in his chapter in "The Science of Colour" (Elsevier 2003) said these figures look very modern. Indeed, if you compare my figure for yellow with my conversion of Thomas Young's figure, we differ by 2 nm, or about 0.3%.

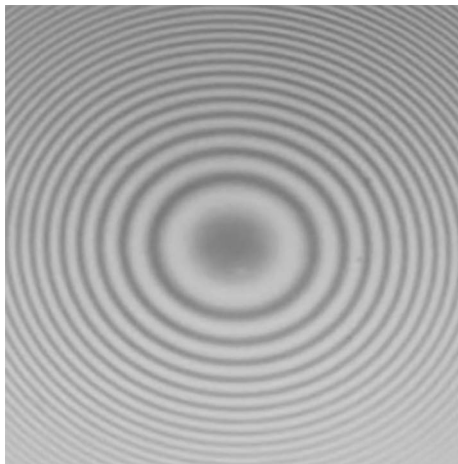
A digression: could Thomas Young actually measure these tiny wavelengths to this accuracy? I find this entirely credible. He was measuring the 'Newton's rings' fringes between a lens and a glass flat.



You can make lenses and mirrors using two thick discs of glass. Mount one disc on a bench, so it can rotate. Put carborundum grit and water between the two discs. Press the top disc and move it back and forth, rotating the top piece from time to time. At intervals, rotate the bottom disc. The carborundum paste grinds the glass, removing more glass where the pressure was greatest. This takes more material from the edge of the lower piece making it convex; and from the middle of the upper piece making it concave. Eventually the ground surfaces approach a convex and a concave sphere: the only shapes that fits no matter how the upper part moved. Once the surface is roughly shaped, repeat with finer grinding powder. The wet surface can be tested for shape and depth of focus. When the surface is the correct average shape, polish it to optical flatness one surface at a time using a polishing lap made of pitch that flows to fit the surface to be polished, and red lead oxide. I have made an eight inch $f/8$ parabolic mirror accurate to about a sixth of a wavelength of light using these traditional techniques on the kitchen table, and it works beautifully.

Making an optical flat is a bit more complicated. You can swap the top and bottom pieces of glass, but it is hard to balance the effect so both pieces are neither concave nor convex. The secret is to use three pieces of glass, and juggle them around.

Young placed a lens on a flat glass surface, lit it from above with monochromatic light, and measured the positions of the fringes he saw. He would have seen something like this...



We have interference between two reflections, one from the bottom of the lens and the other from the top of the flat. The pattern is dark where the two surfaces touch. At the first dark fringe, the lower reflection is shifted by one whole wavelength, so the interference is the same as at the centre. If we count out N fringes, the path difference will be N

wavelengths. The air gap will be $N/2$ wavelengths because the second reflection has to cross this air gap twice.

Young counted out the rings as far as he could, then measured the diameter of the fringe. He had measured the difference in height between the middle and the edges of the spherical surface with a micrometer, so he could predict the width of the tiny air gap from the fringe diameter using simple geometry.

It would not be easy to be that accurate, but it is certainly possible.

How much light?

If you want to skip the physics, the quick story is roughly as follows. You can split your light into different wavelengths. Can you measure the amount of each wavelength of light in some objective way? Shine the light onto something black, and measure the rise in temperature. That tells us how much energy there was in the light.

In 1800, William Herschel delivered a paper titled “Investigation of the Powers of the prismatic Colours to heat and illuminate objects” to the Royal Society in London. William was probably aided by his younger sister Caroline, but you could not tell the Royal Society that women did science back then, so his paper is written in the first person throughout. Let’s credit Caroline Herschel with some of anything he reported as it is hard to know who did what.

William had been preparing for the transit of Mercury, by looking at the sun’s disc through a nine-inch diameter Newtonian reflecting telescope. The heat cracked all the darkened glass slips, so he experimented using a selection of coloured filters that he had to hand. He found two red filters reduced the light to tolerable levels, but he *‘could not bear the irritation, from a sensation of heat, which it appeared these glasses did not stop’*.

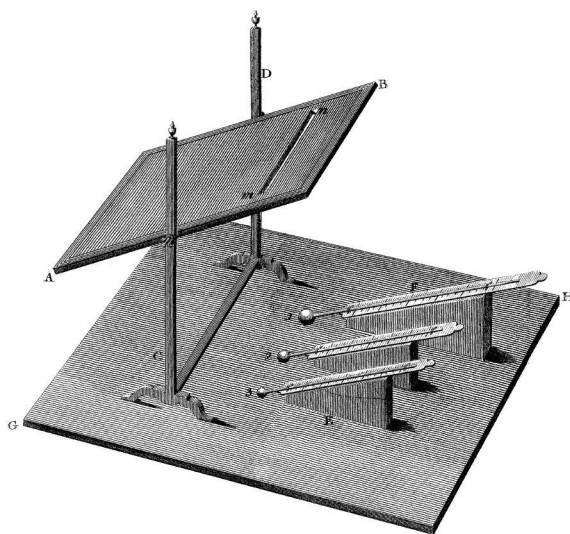
I am not surprised he experienced some irritation. I once made a simple Newtonian telescope with my eight-inch mirror to view the transit of Venus. Instead of viewing the image directly, I was projecting it onto a sheet of paper, using an aperture roughly where Herschel put his eye. My aperture was a piece of wood sprayed black with a hole in it. It caught fire in seconds.

Undaunted, Herschel tried again with green glass. Two green glasses reduced the light level, and also stopped the sensation of heat. He had assumed all visible light contributed equally to the burning-glass effect, yet red light made his eye warm, and green light did not. He built some apparatus to investigate which parts of the spectrum contributed the most heat.

In this separate experiment, he cast a spectrum from a large prism at his window onto a tilting screen with a slit in it. The apparatus could be

arranged so that a band of light from a chosen colour fell upon the bulb of a thermometer. The bulb had been blackened with Indian ink to absorb the light. With this apparatus, he could measure how quickly the temperature rose in degrees per minute.

The illustration below is from his paper (Philosophical Transactions of the Royal Society of London vol 90 (1800) p 262, reproduced with permission of the Royal Society).



Herschel's apparatus

He uses three thermometers. His own thermometer is the large one, and he borrowed two others from a friend. The two small thermometers proved to be a good match, so he used one in the band of light to record the temperature rise, and the other in shadow, so he could correct for other variations in temperature. He also used this large thermometer to verify the other two. He found that the red gave more than twice as much heating as green, and three and a half times as much heating as the violet. Indeed, the peak in heating seemed to lie beyond the visible red, in the infrared.

He then replaced the thermometers with a microscope, and looked at a number of objects such as a coin, some coloured paper, and a nail, noting how well he could resolve detail in the different lights. He found his ability to resolve detail peaked in the yellow-green, and was particularly poor in the violet. This unremarkable yellow-green colour is close to the 555 nm luminance used in the definition of the SI unit the *candela* we briefly mentioned earlier.

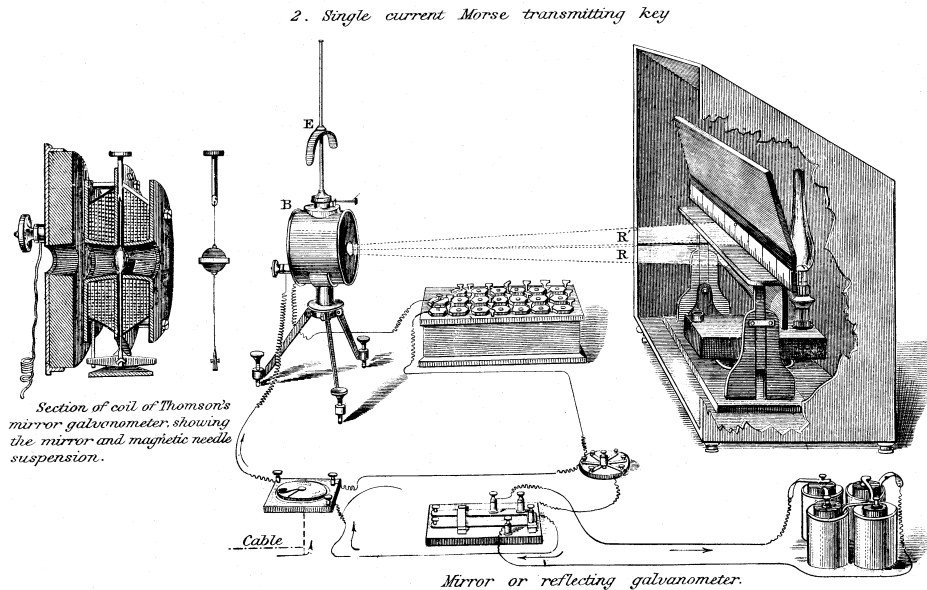
Turning light to heat to measure it is a good idea. Count Rumford had two years earlier noted that mechanical work — in his case the machining of cannon barrels — could be related to the amount of heat produced. In this paper, William Herschel stops just short of equating temperature rise to the amount of radiation absorbed. He carefully acknowledges the possibility that the light he saw, and the temperature rises he measured might be two completely separate forms of radiation that just happened to be spread out by a prism similarly. Nevertheless, the Herschels observed that our ability to resolve detail varies with colour; that the colours we see are part of a larger spectrum; and that we may measure the amount of radiation by converting it to a temperature rise in a blackened body. That is not bad for some crude experiments which might have been done in a single afternoon.

The following year (1801) Johann Wilhelm Ritter, inspired by Herschel's experiments, discovered another form of invisible radiation. Silver chloride was known to go black when exposed to sunlight. This was to become photography later, but at the time there was no way to 'fix' the image. Ritter tried seeing which spectral colours gave the most blackening. This time, red did very little but the darkening increased through the violet, and into the ultraviolet.

Herschel's original experiment was ground-breaking, but his apparatus was not very accurate. He measured only three bands in the spectrum: red, green and blue, and he assumed the sun's brightness did not change significantly. He tried using a smaller slit so he could measure yellow by itself, but he could not get enough light to give a measurable temperature rise. Over the following years, Herschel and many others improved the sensitivity of the detector. They replaced the thermometers with smaller detectors such as thermocouples that needed less light to get warm.

In 1878 S.P. Langley made the first bolometer. This was an instrument that worked in the same way as Herschel's blackened thermometers, but measured the amount of light by the rise in temperature of a blackened piece of platinum from the change in electrical resistance. When the temperature of the platinum increases, the electrical resistance rises. Platinum was a good choice for the detector: it could be made very pure, and very thin, and it did not oxidise. If you have a pure material, the change in resistance is approximately linear with absolute temperature. These days, the most sensitive bolometers are cooled to near absolute zero. Back then refrigeration to those sorts of temperatures was not practical, so they were looking at a 1 degree rise in temperature on top of an absolute temperature of about 300 degrees Kelvin. This may not sound promising, but you can make the platinum really thin, the same absorbed energy gives a bigger rise in temperature. Other advances in electrical measurement for submarine telegraphy — Thompson's galvanometer and the Whetstone's bridge — did the rest. Here is a picture

of a galvanometer from about 1880 for long-distance Morse transmission...



Galvanometer

In those days there were no practical amplifiers. A cable across the Atlantic was one long piece of wire, and a lot of work went into detecting tiny electrical signals at the far end. In the diagram the current passes through a coil, generating a magnetic field. In the middle of the coil is a small slightly concave mirror, with a bar magnet on the back. The magnet hangs by a single silk fibre. When the current flows, the magnetic field generates a torque on the bar magnet, causing the magnet and the mirror to turn until the electromagnetic force is balanced by the torsion from the fibre. As the torsion is very small, you can get a significant deflection for tiny currents. Everything is in a can, so the action is protected from air currents, and variations in temperature and humidity. The light source has a round spot with a vertical line. It is focussed so its light bounces off the mirror, and gives a focussed round spot with a central dark line on the scale. This is like the pointing arm on a conventional ammeter, but lighter and much longer. If you compare voltages with a Whetstone's bridge, you can detect tiny voltage differences from which way the spot swings when you make a connection. You can then short out the coil with a resistor to damp out the swing ready for the next measurement.

Later galvanometers used quartz fibres. The best quartz fibres were made by a process developed by C.V. Boys: a quartz rod was heated until

it glowed yellow. One end of the rod was clamped, and the other end was attached to a bolt fired from a powerful crossbow into an evacuated tube. This gave lengths of extremely fine, strong filament with no tendency to twist when under tension. With these improvements, it was possible in the 1880s to measure steady currents of less than a billionth of an amp. Langley claimed a later instrument should detect the heat from a cow a quarter of a mile away. I love the idea of calibrating your detector using a reference cow, but I do not know if this was ever tested. The detector was used to measure tiny details in the spectra of stars viewed through a large prism of rock salt, so the cow claim is credible.

I have done experiments with similar apparatus, except for the cow. You are unlikely to need to do this today: a modern bolometer from a catalogue may be a compact, hand-held unit with a digital display, that does the hard work for you. However, if you do the measurements with anything like the original apparatus, you appreciate how elegant it all is.

Bolometers are still used today. They are the standard instrument that allows us to measure amounts of radiation of any wavelength in standard units. They are not particularly sensitive, except in the far infrared, when they are as sensitive as anything else. In the visible spectrum, they are often used to calibrate other, more sensitive devices such as silicon detectors. Modern cameras for the far infrared may use arrays of bolometers.

The spectroscope

We saw earlier Fraunhofer's engraving of the solar spectrum. Joseph von Fraunhofer worked as a research manager and director at the Benediktbeuern Optical Institute. He had responsibilities for developing and improving the glass manufacturing processes, and for developing new glass formulas. The dispersive properties of glass — the variation of refractive index with wavelength that makes prisms give rainbows — is a nuisance when trying to make a lens, because it means each wavelength has a different focal distance. The solution to this *chromatic aberration* was to cement a convex and a concave lens together to make an achromatic doublet. The two lenses were made of different glasses, chosen so the chromatic aberration of one half reversed the chromatic aberration of the other half. Hitherto, the experts that made achromatic doublets all worked in London (particularly Dollond of Dollond and Aitchison, established 1750, and only lost Dollond's name when it was taken over by Boots Opticians in 2009). It was Fraunhofer's job to make Bavaria a centre of excellence for achromatic doublet production. This meant he had to produce very precise and uniform prisms and to measure their dispersive properties as accurately as possible.

Fraunhofer needed a set of particular colours for his measurements. It was not enough to divide the light up into seven colours by eye as

Newton did — he needed a precise way of dividing up the spectrum. Sunlight had seven dark lines according to Wollaston in 1802. In 1813 or 1814, Fraunhofer built himself a spectroscope. He used a slit in the shutters of a window, a 60-degree flint glass prism, and a small telescope to view the spectrum, all fitted into an old cigar box blackened with ink. Instead of seeing seven lines, he saw hundreds of dark and faint lines. He gave letters to the ten most distinct lines, and with a better instrument measured precise angles for about 350. These lines have been called the Fraunhofer lines ever since. With these lines, he could then find the same lines with a prism made of a different material, and use the results to design an achromatic doublet.

Fraunhofer did not need the wavelength for his achromatic lens design, but he must have wanted to base his spectra on an absolute measurement rather than rely on some arbitrary formula that fitted the glass he was working with. To measure wavelength, he replaced the prism in his spectrometer with a diffraction grating. Fraunhofer made his first gratings from fine wire wound over fine screw threads. The angle of deviation of each wavelength is determined by the wavelength of the light and the spacing of the grating: if you measure the angle of deviation, you know the wavelength. So, a grating can work a bit like a prism but with an absolutely predictable angular deviation for each wavelength.

Earlier, I suggested Herschel's yellow-green colour might correspond with the 555 nm peak luminance. He saw that the human eye was best at resolving detail in yellow and green, but that might have been true only for the particular prism that he had picked. Another prism might stretch the greens and the yellows, and squash the oranges and the reds: the green-yellows would then look dimmer, and the orange-reds would look brighter. As it turns out, most clear glass spreads out the spectrum fairly evenly, so his yellow-green colour is probably close to the 555 nm peak luminance. There are some types of glass with very uneven spreads. Some can even fold the spectrum back on itself. But these glasses are all deeply coloured, and not useful for lenses or prisms.

We saw how Young's simple apparatus was able to measure wavelength with surprising accuracy, but needed something else to spread out the light into the different wavelengths. How does Fraunhofer's grating, which did both jobs, measure up?

Fraunhofer made ten gratings, and measured the D line for several orders for each grating. He then combined all these figures to produce an average value for the D line of 588.8 nm.

Fraunhofer's absorption D-line is actually the two lines of the sodium doublet, though back then it had not been identified with sodium. The modern figures for the sodium doublet are 588.9950 nm (brighter line) and 589.5924 nm (dimmer line) in vacuum. If we take the wavelength for the brighter line and divide it by 1.00027 — a good average figure for the

refractive index of air — when we get a wavelength of 588.83 nm. Fraunhofer was meticulous, but he must have been a bit lucky too because even the metre was not defined that precisely back then.

By 1823, Fraunhofer had made some finer gratings with up to 300 lines/mm ruled on glass plates. He used these to measure the solar spectra, and the spectra of the stars and planets. The ruling technique for producing gratings continued to be developed throughout the 19th century.

The spectrometer

If we put Fraunhofer's spectroscope and Langley's bolometer together, we get an instrument called a *spectrometer*. This measures the amount of light of each wavelength in fundamental scientific units. A modern spectrometer measures all wavelengths at once with an array of detectors. It will probably not use an array of bolometers for visible light, but it may have been calibrated from a bolometer via other standards.

If I point my spectrometer at a red object and press the button, I get a stream of numbers with low values at the 400 nm end, and high values at the 700 nm end. This tells us what we would expect from something red, but it does not tell us anything about the colours we see.

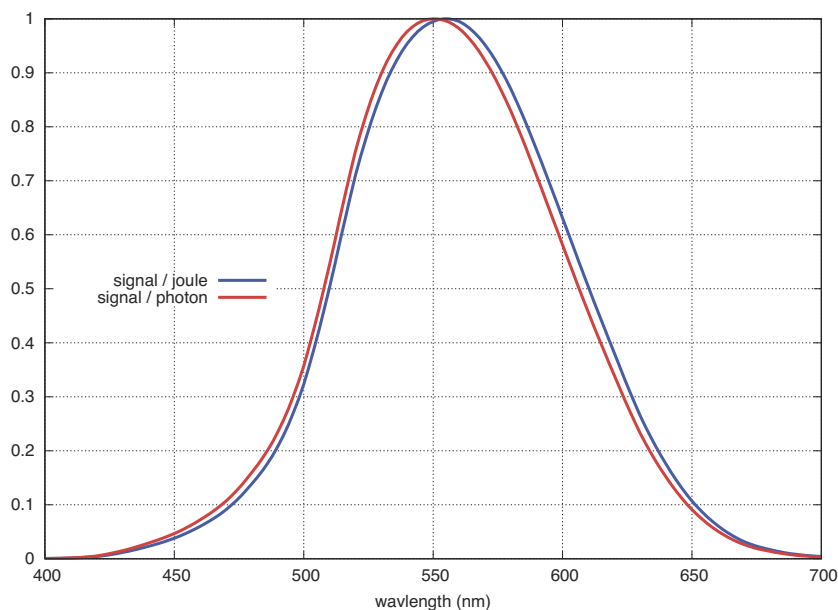
Broad spectrum peaks

A spectrum usually has the spectral bands defined by wavelength in nm, and amount of light in the bands measured in watts (power) or joules (= watts × seconds) for exposures. This is pretty standard, but if you read scientific papers, particularly historical papers, you may find bands of frequency instead of wavelength, and amount of light measured in photons or photons per second.

Does this matter?

The candela unit is based on a wavelength close to 555 nm where the eye has peak sensitivity. The sensitivity is a measure of how much stimulus our eye gets per unit of light. We discuss what 'stimulus' is in the next chapter, so just assume it is how 'bright' the light looks for now.

Sensitivity is often scaled so the peak sensitivity is 1.0, and this is what we shall do here. If a waveband has a sensitivity of 0.5, we need twice as much light to get the same brightness. If a waveband has zero sensitivity, we cannot see it.



Here we plot the sensitivity of the eye for the visible wavebands.

The blue line has the waveband defined by wavelength in nm, and the sensitivity calculated with the exposure measured in joules. The blue line peak is at 555 nm. We meet this curve again in the next chapter when we discuss units of brightness.

The red line has the waveband defined by wavelength in nm, but this time the sensitivity is calculated from an exposure measured in photons. Short wavelength photons have more energy (see the table right at the beginning of this chapter), so a unit of energy has fewer photons per joule. The red line peak is at 550 nm. 5 nm is not a big shift, but you can see the difference between 550 and 555 nm light. If we plotted frequency instead of wavelength on the horizontal axis, the peak would shift again. The peak sensitivity depends on how you plot the graph.

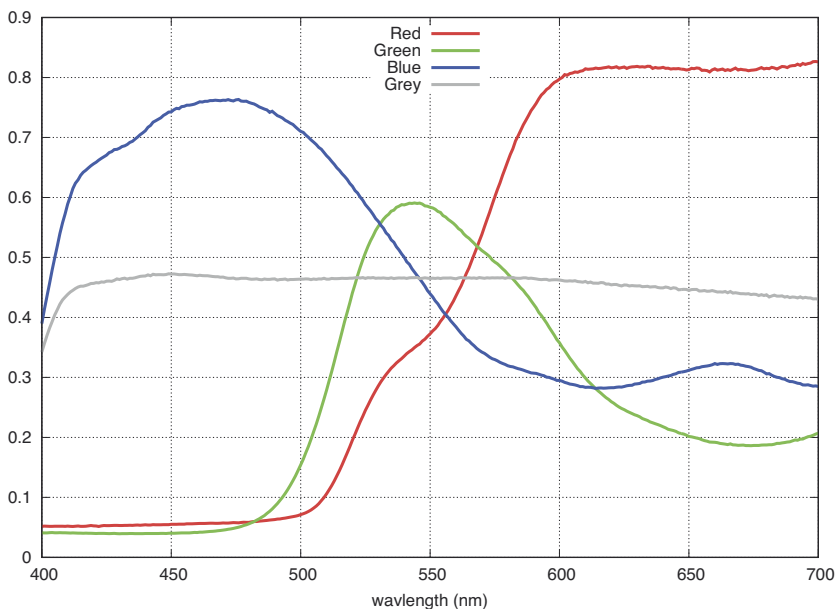
Do not worry. This is like the difference between wavelength in air and in vacuum. I am telling you about it now, so it won't confuse you later, but it is unlikely to be a problem.

Reflection and transmission spectra

The reflection spectrum of a colour is the fraction of the light that is reflected at each wavelength. We measure the spectrum of the colour, and divide that by the spectrum of an ideal diffuse white target under the same illumination. I have not got an ideal diffuse white target, but I do have a barium sulphide tile which reflects between 95% and 98% at all wavelengths, and a reflectance calibration for that tile. This is better than using white paper, which may have fluorescent 'brighteners' to make it

look whiter. We also need a good light source with a smooth spectrum and all the wavelengths.

Here are some reflectance measurements for chips from the Munsell Colour Atlas...



Reflectance of Munsell chips

Reflectance should lie between 0 and 1. The colours all have the same Munsell Value. The green and blue chips reflect less than 50% at their peak, while the red reflects over 70%. This is fairly typical.

There are many mechanisms to make dyes and pigments absorb photons above a certain energy level. Cadmium Yellow (CdS) is a semiconductor with a band gap of 2.4 eV. It absorbs photons with more than 2.4 eV (see the table at the beginning of the chapter), so it is bright yellow. Cinnabar (HgS) is also a semiconductor with a smaller band gap of 2.0 eV, so it is bright red. There are many bright pigments and dyes that absorb short wavelengths, and reflect or transmit the longer ones. There are some pigments and dyes that absorb the long wavelengths or absorb the middle wavelengths; but much fewer materials do this, and the ones that do are generally less reflective, and less bright. Note the green peak is steeper on the short wavelength side; it may be a combination of two passband filters. For much more on this sort of thing see Kurt Nassau's book "The Physics and Chemistry of Colour".

The grey measurement is almost a flat line. There are good whites that reflect most light of all wavelengths, and good blacks that absorb most light of all wavelengths. You can make good, flat greys by mixing these.

A transmission spectrum is a similar set of measurements, but you measure the light transmitted by your test object. You do not need a reference white, as your reference is now the light source without your test object. If your test object is opaque or milky, characterizing it is much more complex. For more on this, see Chapter 10.

Chapter 3

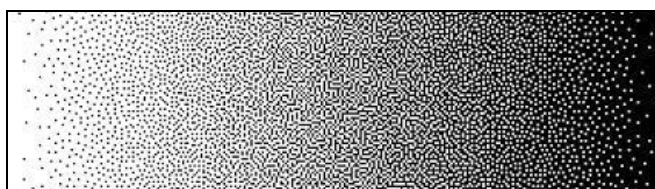
Measuring what we see

Last chapter we talked about how we measure light. We identify monochromatic light by its wavelength in nm. The light we see mostly lies between 400 and 700 nm. We identified how much light there is at each wavelength by turning it into heat, and getting equivalent values in watts. The spectrum of light gives the power for each wavelength. I named some different wavelengths, so we could identify what colour we might be talking about, but we did not discuss human vision much beyond that. This chapter covers how we connect the numbers we measure using the instruments and techniques of the last chapter to the colours we see.

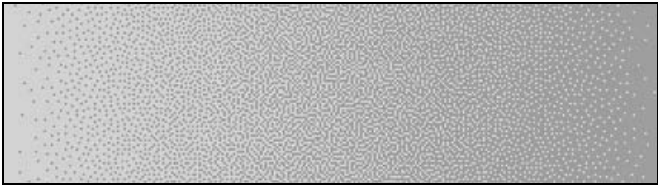
Mixing light

In the last chapter we gave colour names to some single wavelengths. Black is 'zero light'. We have not got to white, grey, pink, brown or light blue yet.

Aristotle considered mixing colours by mixing sands of different colours, or weaving threads of different colours together, or by drawing black lines or dots with charcoal on white marble...



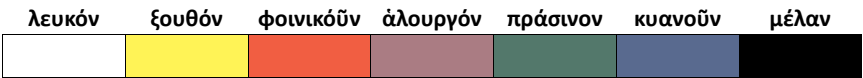
He was familiar with the half-toning tricks used by mosaic artists to get smooth shades from a restricted mineral palette. He noted weavers could work their patterns by lamplight, but always checked their look by daylight, because some colour combinations such as purple and plain changed how they looked. He did not argue that the particular grey you got by mixing black and white was anything absolute, but just the result of a particular mixing process and lighting conditions. But if you picked two tones from this grey scale, you could repeat the mixing process...



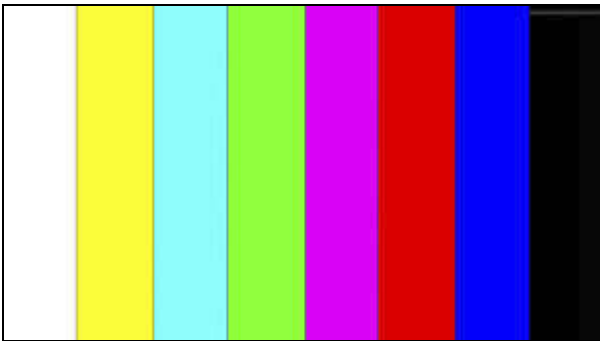
You can always subdivide the gap between any two colours into a set of tones, just as you can divide a line in plane geometry into smaller lines without limit. This was one way people of Aristotle's world visualised infinity. We might not see these infinite gradations as different colours, but they can still be constructed.

Aristotle’s charcoal on marble is actually quite subtle. If you cover half the surface with fine charcoal, you might expect this to reflect half the light. But white looking materials such as paper and marble are translucent. Light bounces around inside it, so the place it went in is not necessarily the place it comes out. This means the charcoal might absorb half the light on the way in, and another half of the remainder on the way out, so it reflects a quarter of the light and not half. We come back to this topic in Chapter 10. Aristotle may not have known these details, but he was wise enough to define his colour mixing process in general terms, so he is right either way.

Aristotle also wondered how many fundamental colours you need to mix all the others. He listed seven colours in order...



This is a different set of seven colours to Newton because he included white and black, and he arranged them in order. Is Aristotle saying that all the colours are mixtures of white and black? He has been interpreted that way. I have difficulty believing this as he already has mixed white and black. We can see a similar sequence in the SMPTE colour bars...



SMPTE colour bars

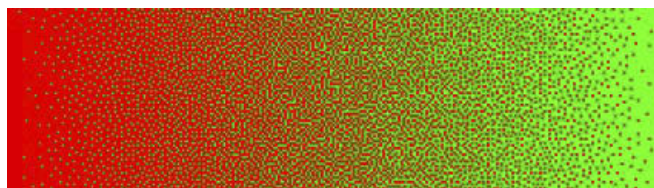
I like to think Aristotle was ordering his colours by how bright they seemed to him. He has got some colours in the middle in a different order, but this may have been because of the pigments he knew (there is no pigment as vivid as video green). He only describes the colours by name. If we have to guess which colour Newton meant by 'blue' in 1730, we may expect problems trying to guess what Aristotle meant back in 350 BC.

Ancient Greek $\kappa\upsilon\alpha\nu\omicron\upsilon\nu$ (kúanon) may once have meant bright, or shining: perhaps bright like the sky, rather than sky blue. Homer used $\kappa\acute{\upsilon}\alpha\nu\omicron\varsigma$ for Hector's hair (elsewhere, he calls it 'yellow'). But Homer is much older Greek than Aristotle. Another interpretation for $\kappa\acute{\upsilon}\alpha\nu\omicron\varsigma$ is for a deep, mineral blue, perhaps from the Hittite 'kuwannan' ("precious stone, copper, blue"). This use came in long after Homer, and is perhaps the one Aristotle understood. If you are going to try to mix all the colours from pigments, you need a dark blue, but can perhaps mix the cyan colour.

Did Aristotle have a hint of the modern understanding of brightness? I don't know. It doesn't matter here, but it is nice to think how close he might have come to adding an eighth colour, and putting these colours on the corners of a cube two thousand years before Thomas Young.

The three-colour approximation

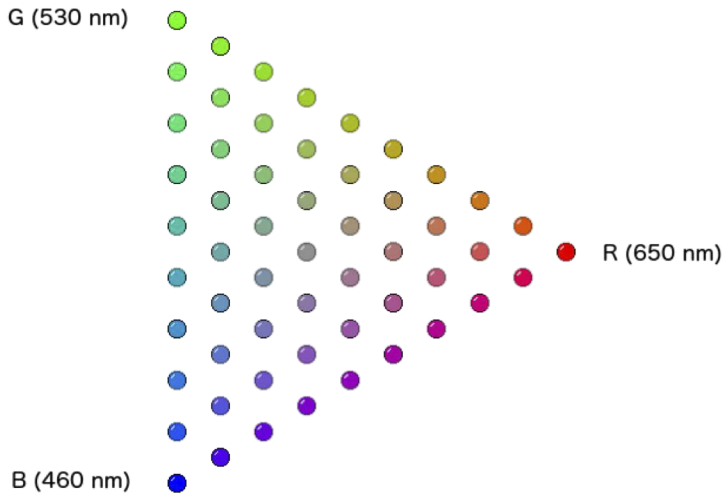
Back to mixing light. We can measure the wavelength of light, and we can measure the amount of light of that wavelength in some standard units. We can then produce a mixture of two wavelengths by varying the amount of light at the two wavelengths. Let us start with a red of 650 nm and a green of 530 nm. The halftone mixture might look like this...



We see bright red and bright green at either end. In the middle we have a brownish colour. The exact colour you see in the middle may depend on how you read this document. It is probably not bright yellow, but it could become yellow if we made it brighter.

Now, let us add a third colour to the mixture. Let us add a 460 nm blue. I could draw this as a continuous, three-way mosaic pattern, but this gets

difficult to interpret, so let us draw small circles for mixtures of the three colours at regular intervals...



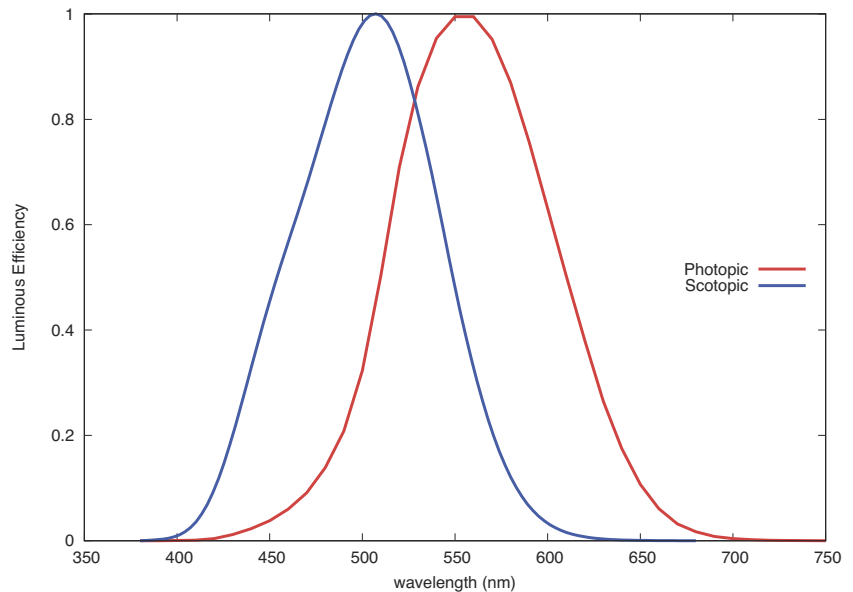
Adding a third colour gives us nearly all the colours. Some colours may look a little drab. We do not have a white, but the point in the middle is grey-ish, which would become white if we increased the brightness. We might not have a colour that is as intense as our pure 500 nm azure, but there are no completely missing colours as there clearly were with just red and green.

Thomas Young was the first person to put this on a scientific basis. He suggested in 1802 that our vision might use three types of colour sensor within the eye. If we had sensors that were sensitive to red, green, and violet, then we might get the intermediate blue-green and yellow colours if a wavelength stimulated two receptors.

I call this *The Three Colour Approximation*. It is an approximation because our eyes are not cameras with arrays of identical RGB sensors. Our eyes are not uniform. We have more than three sensors. You can even see polarization with the unaided eye (Haidinger's brushes). But the three-colour approximation is one of the good and useful tools in the field of vision science where reliable points of reference are rare, and we can live with the errors for typical colour vision.

Deviations from the three-colour approximation

The cone cells give us our three-colour 'photopic' vision during the day. We have a separate set of rod cells that are used at night for monochrome 'scotopic' vision with a spectral sensitivity that peaks around blue-green 500 nm, rather than the yellow-green 555 nm for the cones.

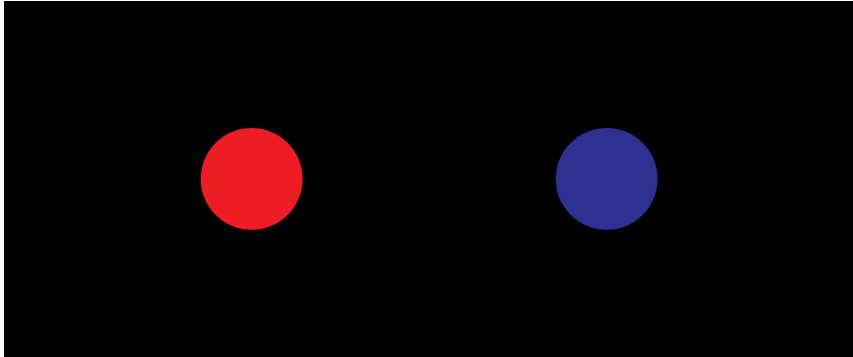


The rod cells contain a pinkish dye called rhodopsin. In 1876, Franz Christian Boll saw how the reddish retina of a dissected frog eye had been bleached by light. He guessed that this bleaching of the dye was how the rod cell measured the amount of light. The cone cells have similar dyes, but they bleach when the retina dies, so we did not see them until we developed computer stabilised microscopes that could image cells in a living eye.

Rod cells can give useable signals at much higher illuminations. In “The Island of the Colorblind” Oliver Sacks tells of the Micronesian island of Pinglap, where many inhabitants were so severely colour-blind that they had no cone cell vision at all because of interbreeding. They had poor vision because they had no working cells in the centre of their vision, and they were troubled by full sunlight, but they could still see in the day using only their rod cells.

There is no evidence we use our rod cells as a fourth colour receptor. At low light levels there is a small ‘mezopic’ vision regime where we use our rods to see, but scrape a bit of useful colour information from the cone cells. The rod cells are more sensitive to short wavelengths than the luminance signal from cone cells, so reds look darker, and blues look lighter. This colour shift with brightness is called the Purkinje shift.

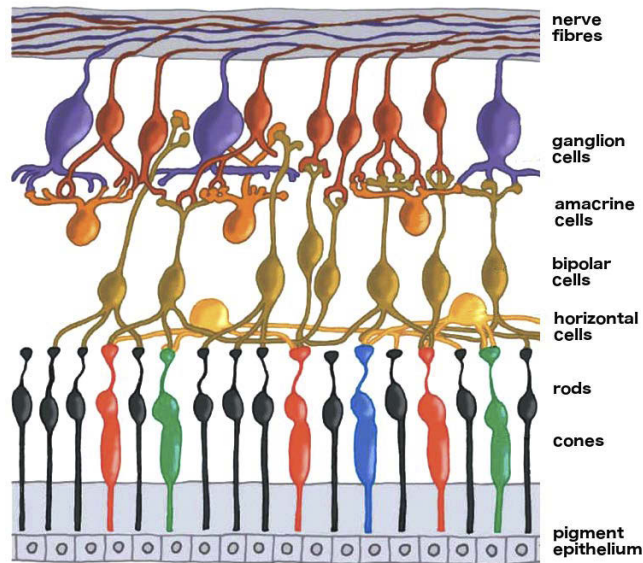
If you print this figure out, the red spot should look brighter than the blue one.



Purkinje shift

If you look at the print by the light of the full moon, you should still see colour, but the blue spot should look brighter. This is not because the light from the moon contains a bigger fraction of blue (it does not) but because our luminance signal now comes from the rods instead of the cones. Mezopic vision is hard to model. In motion pictures we simulate low-light vision by muting the warm tones, and making the mid-tones darker, which gives the right feel.

We have another problem with low light levels. Our cone cells fire in 'tonic mode'. This means that they do not fire just when they get a signal, they will also fire (but less often) when it is dark. This means it is hard to resolve a small signal from the dark signal, particularly after it is combined with other, larger signals from neighbouring cells in the retinal ganglia...



Retinal layers

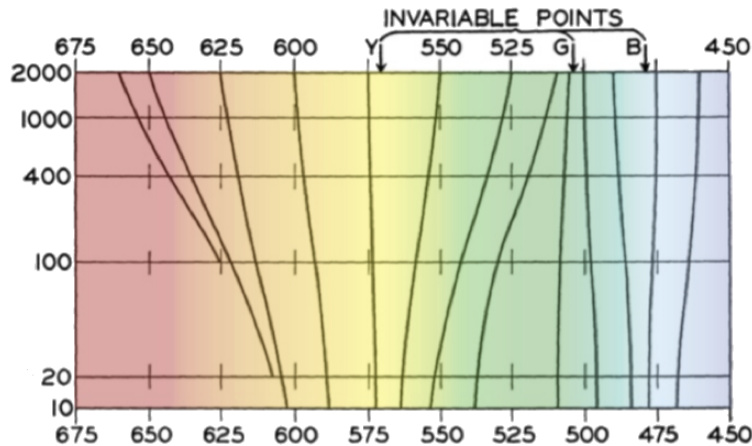
We discuss these connections in more detail in Chapter 8. For now, all we need to note is that the retina is complicated, and there are many layers of cells that combine the signals from the cones with their neighbours to make luminance, colour, edge, and motion signals. The retina is much more than a simple camera RGB mosaic.

It may seem odd that the retina does not preserve the absolute light levels from the RGB detectors. Light comes from the top in the last diagram, and the detector cells are at the bottom. Light has to pass through all the nerve cells to get to the detectors. These cells are fairly transparent, but they scatter light. There are blood vessels on the surface too. The pigment epithelium absorbs some stray light, but the eye is far from black, so there will always be scattered light. So there was probably no good dark level at the cones in the first place.

So far I have talked about red-, green-, and blue-sensitive cone cells. Now is probably a good time to start using the standard names. The three cone signals are called LMS: L for long wavelength red, M for medium wavelength green and S for short wavelength blue and violet.

The Bezold-Brücke effect

We can see red (L) and green (M) in dim lighting, such as moonlight. We should see yellow (L+M) when the L and M signals are roughly the same. We will be less able to distinguish shades of orange or yellow-green when one of the signals is a lot bigger than the other. This gives the Bezold-Brücke effect, where the colour we see depends on the light level.



Bezold-Brücke effect

Bezold and Brücke measured this effect in the late nineteenth century. The diagram dates from about 1940. The horizontal axis is the wavelength in nm, and the vertical axis is the brightness in trolands (td). The colours were added afterwards and show the colour we would see under good lighting conditions. The contours are lines joining the points that look the same colour. Under low light levels, we see more colours as 'red' or 'green', and we lose the shades of orange and green-yellow, just as we might expect.

This shift can be seen when people look at simple test images in a darkened room. When we look at a realistic scene, the colour we see does not depend on the light level. The *colour constancy* process we met in the introduction restores the colours to the ones we know ought to be there. We cannot see a very dim orange colour as orange, but if it shades into a lighter orange that we can see, our vision will assume it is the same.

Our central vision has no S cones, and it is covered by a yellow dye. Our peripheral vision has all three cones and no yellow dye, so it does not match our central vision. Our eye and brain corrects for much of this, just as it deals with the 'blind spot'. We shall discuss this in more detail later.

As well as the rod cells, the eye also has a fifth type of cell that can detect light. Genetically blind mice with no rod or cone cells still have a reflex

that causes the pupil to dilate and contract as the light level varies. They can still detect cycles of night and day. Something is sensing the light levels even when all the detectors we knew about are not there. This is the intrinsically photosensitive retinal ganglion cells. The retinal ganglia are the tangle of nerves next to the rods and cones that usually do simple colour and edge processing. Some of these cells can detect light too. They are not very sensitive, and there are not enough of them to make a fourth channel for a focussed image, but they may have other ways of affecting what we see. They may affect our overall sense of white, or tweak the colour processing in the retinal ganglia in other ways, as well as controlling our pupils. Their peak sensitivity is between 460 and 484 nm.

There are other non-linear effects for really saturated colours, or really bright colours. At high light levels, dazzling highlights give after-images. At low light levels, we have the Purkinje shift, but this only happens when the rod cells have adapted to the low light levels, and that may take minutes to happen. These time-dependent effects are hard to predict. Fortunately, we are usually aware of them, and we seem to know they are not 'real'.

The three-colour assumption holds for normal colour vision. For very dark or very bright colours outside normal colour vision, our vision clings to colour constancy, and seems to fill in the colours that *ought to be there*. Models and simple experiments may predict colour shifts, but they do not turn up when viewing real scenes. We use the three-colour assumption for normal colour vision because it works well. We use the three-colour assumption for abnormal colour vision because we have nothing that works better. And perhaps our visual cortex is doing much the same thing.

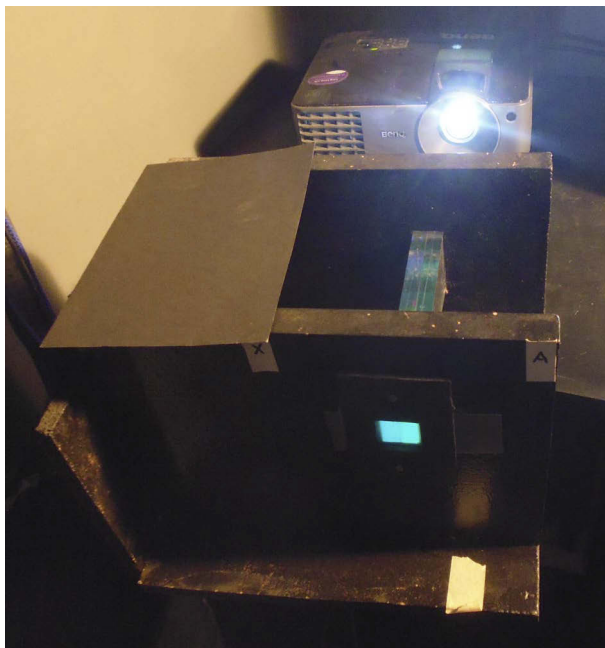
The Abney effect

Here is one of those non-linear effects for really saturated colours. Take a very saturated, nearly monochromatic colour, and mix it with white. Does it have the same hue?

Sir William de Wiveleslie Abney investigated this in 1909. He had a white and a coloured light source. He could let varying amounts of these two light sources into two rectangular glass rods with a diffuser at either end. The light randomised as it bounced around inside the glass rods, and gave a uniform patch of colour at the exit. He saw the hue varied when he added white. This is known as the Abney effect.

Here is a device I built to investigate these saturated colours. It is shown with the cover off. I used a cheap digital projector with the colour wheel removed, so it projected black and white images. This was focussed onto

a graded dichroic filter. I varied the image sent to the projector to create two separate custom spectra at the end of glass rods.



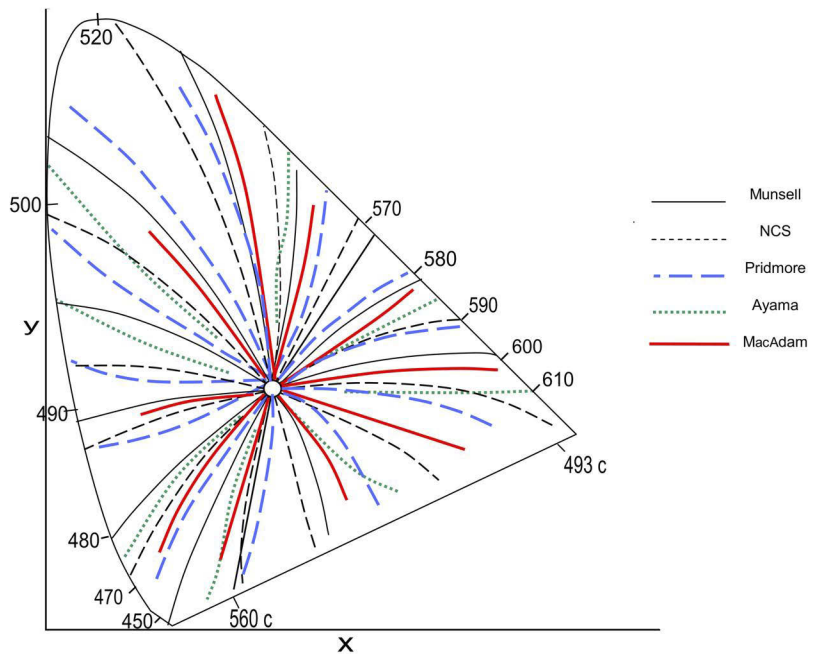
From there on, the optics is like the Abney apparatus. Each spectrum went into a rectangular, glass rod, and produced an even rectangle. You compared the two halves of the square in a darkened room. Here we see blue on the left, and blue-green on the right.

The projector had a mercury lamp, so its spectrum was not uniform for all wavelengths, but I calibrated the program that produced the projector image to correct for that. The brightness depended on what you could get from the mercury spectrum, but it was bright enough to simulate the colours of projected film, which is what I was working with at the time. The wavelength resolution was limited to about 8 nm by the dichroic. Blacks were not fully black. But it worked well enough to give a reasonable simulation of near-monochromatic lights.

What did I see? It was hard to judge hues for the very saturated colours. I felt I could see some hue shifts, but it was difficult to tell in the dark with no reference white whether hues actually matched or not. Imagine being asked to pick a colour swatch to match a sodium lamp. A bit like that.

Here is the Abney effect measured by five authors. The original diagram is in Pridmore RW, Melgosa M (2015) All Effects of Psychophysical Variables on Color Attributes: A Classification System. PLOS ONE 10(4): e0119024. <https://doi.org/10.1371/journal.pone.0119024>. I copy it here

with grateful thanks to the authors, as this must have been a lot of work to make.



The white point is the circle in the middle of the diagram. If there was no Abney effect, the lines of constant hue would be straight lines radiating out from here.

At the white point, the LMS signals are similar, and the hue lines look straight. Close to the spectral locus, non-linear effects become significant and the hue lines bend. You might expect the hue to be constant if one LMS signal was much bigger than the others, or if two were roughly the same. You can probably see straight-ish lines for red, green, and blue primaries, and cyan, magenta, yellow secondaries. Before anyone could measure the LMS primaries, this sort of clue was valuable.

Can we correct an image for the Abney effect? I would hesitate to try. The diagram showed five researchers who all measured different things. Small differences in spectral sensitivity between observers may become significant close to the spectral locus. Any camera has its own primaries and non-linear processes, so it should have an Abney-like effect of its own.

Do we see the Abney effect in real scenes? I doubt we see it clearly. Colour constancy handles reflection colours well, but struggles to cope with these saturated colours. They seem to glow. We feel light is coming from somewhere else. I prefer to stick with the three-colour assumption, and assume hue lines are straight. This is straightforward, and it fits with

the vision models where the vision models give good results. If you want to grade an image so the reds pop, that is up to you. But I would not call it correcting for the Abney effect.

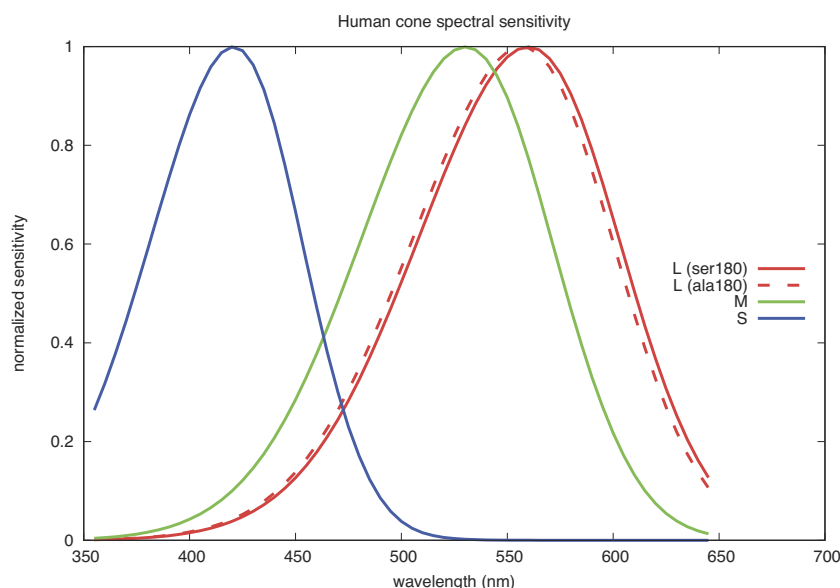
CIE 2006 LMS 2-degree standard observer

The eye has three types of cone sensor, named L, M, and S because their sensitivity peaks in the Long, Medium and Short wavelengths.

A computer-stabilised fundus microscope can image the individual cells in a living eye. We can measure the light reflecting from the eye before and after exposing the retina to bleach the cells. From the colour change, we can identify the cone cells as L, M or S. We can also estimate the peak absorption for each cell.

It is difficult to get good spectral data from a living retina, but all rods and cones detect light using the same bleaching mechanism. We can get good data from rhodopsin in the rod cells because these do not bleach when the retina dies, and we can fit a smooth curve to the noisy data we get from the cone cells. Lamb's chapter 3 from Gegenfurtner and Sharpe's "Colour Vision" gives the details, and the model he used to fit the data.

Here is a plot of the LMS sensitivity as a smooth curve...

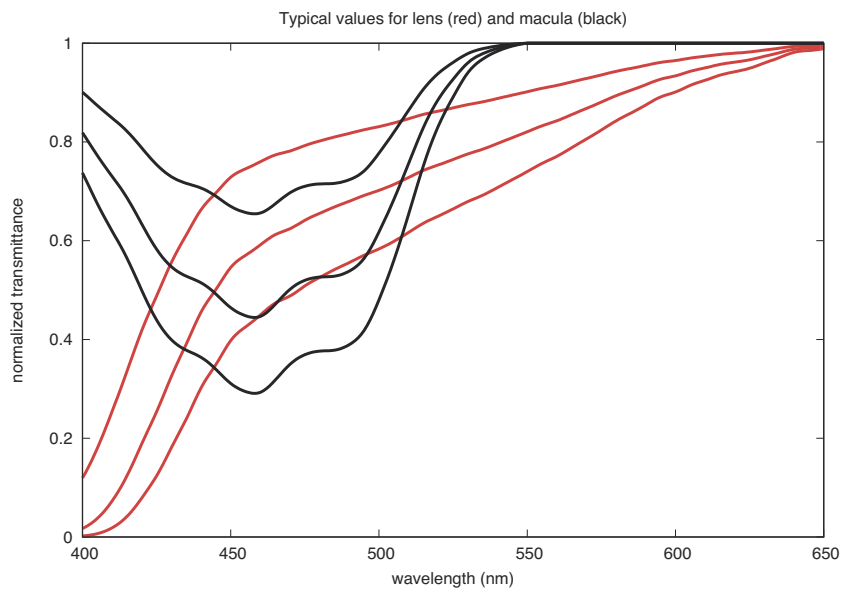


The L-cone pigment is shown with two variants. Approximately 56% of tested subjects have serine and 44% have alanine at codon 180. The difference gives a shift of about 2.7 nm at the peak sensitivity. There are other genetic variants of the M- and L-cones, but these are common, and the ones that have been studied most.

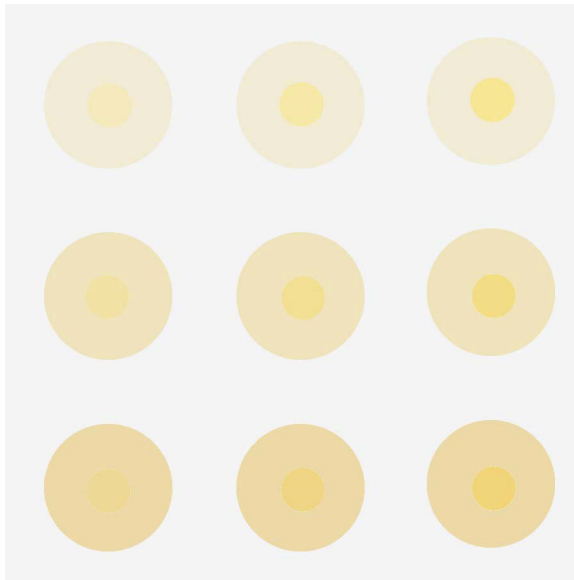
We do not have two sets of standard observers. In practice, the difference between these two red sensitivities is much harder to measure than these cleaned up graphs would suggest. The recent Stockman and Sharpe LMS primaries used for the CIE 2006 standard observer use a weighted average of the two measurements.

This gives the sensitivity of the cell within the eye. We also know the other optical elements of the eye are yellow rather than clear.

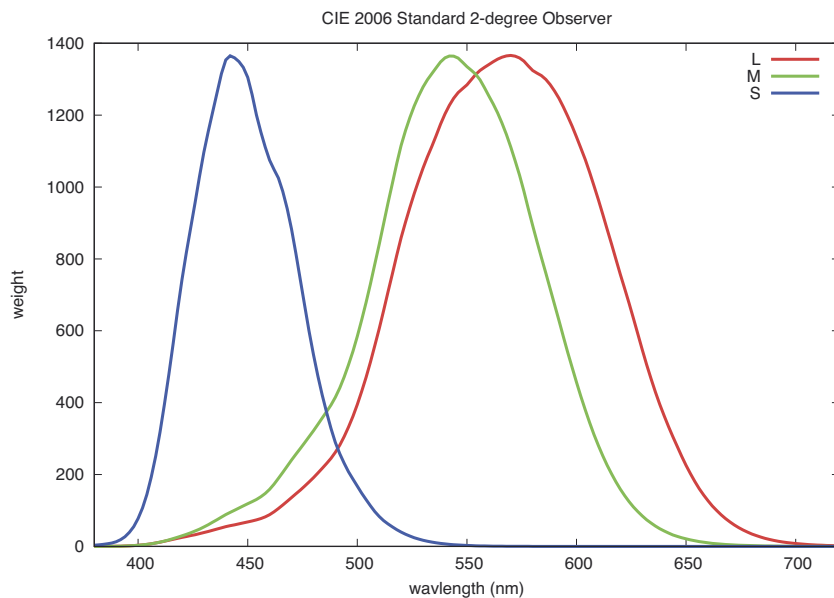
Here is a plot for the low, average, and high densities for the lens and the macular dye for normal human eyes...



Here is what these combinations of lens yellow (down) and macula dye (across) might look like. The background represents a clear lens and macula; it is shown as grey to stop the yellows going outside the video gamut.



If we combine the cell sensitivities with the average transmittance spectra for the lens and the macular dye, and normalise to 1.0, we get the CIE 2006 standard weights...



The lump on the right side of the blue peak comes from the macular dye spectrum.

If you multiply each spectrum value for a colour by the corresponding LMS weights, and sum the values for all the wavelengths, you get three LMS cone cell exposures in the retina for that colour. If two spectra give the same LMS totals, the observer model says the colours will match. And if you have normal vision, they will.

The LMS primaries are not real colours. No spectrum can give you pure M for example; you always get some L or S as well. Think of LMS as the RGB values you might get from a camera matched to the human eye.

The CIE 2006 luminance Y is given by...

$$Y = 0.68990272 \cdot L + 0.34832189 \cdot M$$

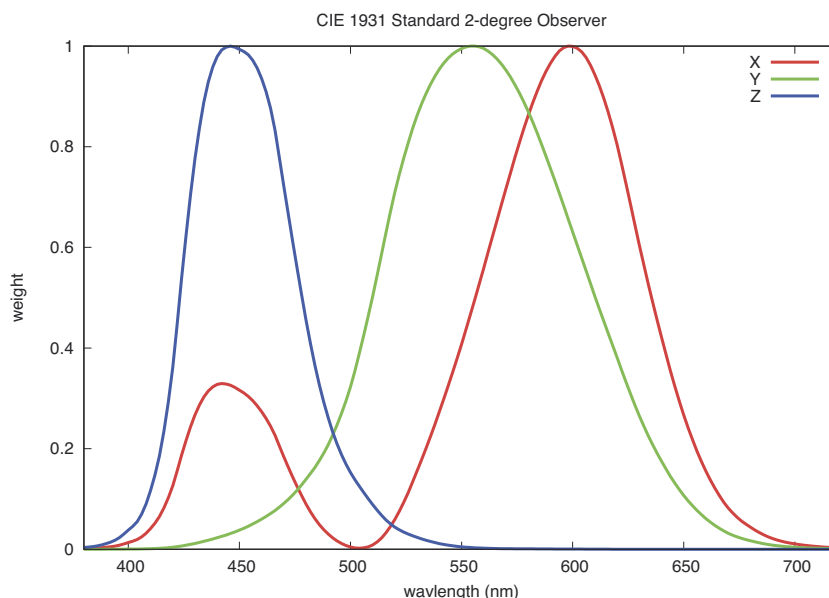
We discuss what luminance is and where it comes from in the next section.

The CIE 1931 XYZ 2-degree standard observer

This is a much older standard. Back in 1931 no-one could see, let alone measure the dyes in the cone cells, so they characterized the eye by what it could see, and derived three XYZ values that could be used to describe colour. We shall start with the middle value Y, as this has the simplest experimental basis.

Y is called *luminance*. We can think of luminance as a measure of lightness or brightness, but it was actually defined using *flicker photometry*. We can see luminance flicker up to about 20 cycles per second. If we switch between two colours with the same luminance, we cannot see flicker much beyond about 5 cycles per second.

Experimenters measured the light levels where they saw no flicker when switching between monochromatic light and a reference light at (say) 10 cycles per second. This is not a good idea if you have epilepsy, and it is not often done, but it did give a sound experimental basis for the luminance weighting for all the visible wavelengths. This is the green Y-curve in the plot below. We can see this curve has a peak at 555 nm, the yellow-green colour that we saw best. So this suggests that this Y signal is not just ‘what we see at 20 Hz’ but also fits well with what we call *brightness*.



When we introduced the three-colour approximation we considered the triangle of colours from mixing 650 nm red, 530 nm green, and 460 nm blue light. This did not give us all possible colours at full saturation and luminance, but it did give all our named colours. The original researchers for the 1931 standard did not have good monochromatic sources, but they did something similar using white light with red, green, and blue filters.

Suppose they tried to match the 500 nm azure wavelength as the target colour. They could match the hue with their RGB primaries, but they could not get the saturation. So, for colours like this, they added a known amount white light to the target colour to desaturate it. They could then get a visual match with their RGB. They then subtracted the equivalent RGB for the white they had added. This gave an RGB mixture with negative values for colours outside the triangle of RGB primaries.

After many colour matching experiments like this, they had a set of XYZ weights. We can calculate the CIE 1931 XYZ values for a spectrum just as we calculated the CIE 2006 LMS cone exposures. If two colours give the same XYZ values then the colours should match.

The 1931 XYZ weights are not the same as the 2006 LMS weights. The people who worked on the standard knew their XYZ values might not be the actual eye primaries. Any linear combination of the XYZ weights would also give three numbers that would match for matching colours. They had no way of telling what the actual eye primaries were, so they chose weights that gave the sharpest functions.

Are XYZ a linear combination of the LMS primaries? They should be close. In fact there are differences because XYZ is based on older and less direct

measurements. They are not large differences, or people would not still use the 1931 standard. Many people put up with the small 1931 errors to remain compatible with the huge body of measurements in the literature since 1931. Many instruments still output CIE 1931 XYZ values directly.

There is a new CIE 2006 based XYZ that is a linear sum of the CIE 2006 LMS. But it is not the same as the old XYZ. Without the original spectra, we cannot accurately convert CIE 1931 XYZ data into CIE 2006 LMS. If you work with legacy data, you may find you have to work in CIE 1931 XYZ. If you are doing new work, I suggest you use the new 2006 standard if you can. Either way, it is sensible to keep any spectral data.

Here is a matrix conversion between XYZ and LMS. It is not a conversion from 1931 XYZ to 2006 LMS. It is part of the CIE 2006 standard: it is usually quoted as the inverse matrix, and it is used to make a set of XYZ primaries that are as close to the 1931 XYZ as possible.

$$\begin{pmatrix} L \\ M \\ S \end{pmatrix} = \begin{pmatrix} +0.21058 & +0.85510 & -0.03970 \\ -0.41708 & +1.17726 & +0.07863 \\ 0.0 & 0.0 & +0.51684 \end{pmatrix} \cdot \begin{pmatrix} X \\ Y \\ Z \end{pmatrix}$$

For those of you not familiar with matrix multiplication: multiply the matrix row values by the XYZ column values and add them to get the output row value. Here the bottom row gives $S = 0.51684 \times Z$. See Appendix A for more on matrices.

In Chapter 6 we fit another matrix...

$$\begin{pmatrix} L \\ M \\ S \end{pmatrix} = \begin{pmatrix} +0.257085 & +0.859943 & -0.031061 \\ -0.394427 & +1.175800 & +0.106423 \\ +0.064856 & -0.076250 & +0.559067 \end{pmatrix} \cdot \begin{pmatrix} X \\ Y \\ Z \end{pmatrix}$$

The bottom line does not have the zeroes we might expect. It is fitted using smooth reflection spectra for a D65 white, so it may give better results for these particular conditions.

10-degree standard observers

By the 1950s, it became clear that our central vision generated much of what we see, but the peripheral vision also plays a part. The outside of the retina has more S-cones, no macular dye, and the light meets the cone cells at a different angle (the Stiles-Crawford effect). These mean our peripheral vision has a slightly different sensitivity to our central vision. The peripheral vision does not directly affect the central vision response, but it may affect our overall sense of the white point, which in turn may influence our interpretation of the colour.

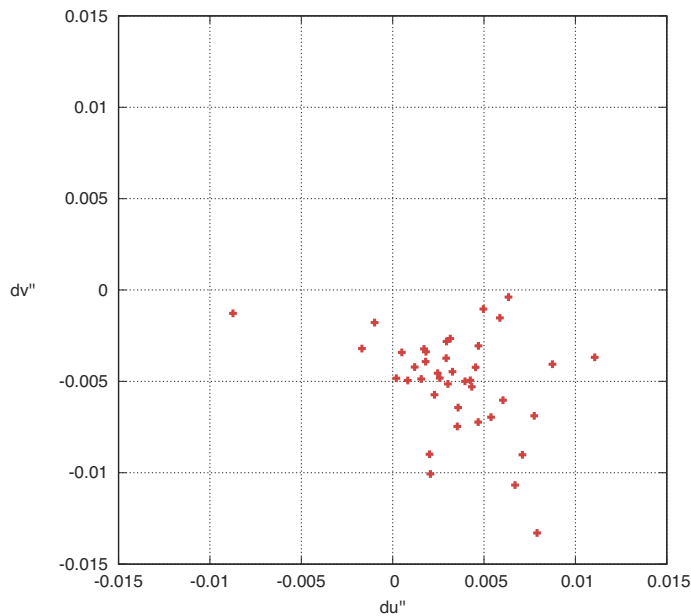
The experimenters repeated the 2-degree visual matching experiments using a 10-degree field, and published the result as the CIE 1964 10-degree standard observer.

I have tried using the 10-degree standard observer. It is not easy. If you stare at the centre of a 10-degree observer target, the central vision has nothing to see, and you have to fight your instincts not to flick your eyes sideways. If you can keep your eyes still enough, you start to lose the surrounding vision (the Troxler effect). Nevertheless, there is a 10-degree observer standard, and it may be better at predicting the colours we see when we paint walls, and other large fields of colour that surround us. It is not often used.

Our peripheral vision is significant when we look at displays. This became clear when we first used liquid crystal displays (LCD) alongside cathode ray tubes (CRTs) for grading motion pictures. We set up both monitors using a Philips probe (a colorimeter) to video D65 white. When we stood back and looked at the monitors displaying full-screen white, the LCD consistently looked slightly orange next to the CRT. We repeated the set-up using a spectrometer. It agreed with the Philips probe, so this was not a problem with the instrument reading. The problem was that we were looking at a full-screen white from the normal viewing distance, so the 10-degree standard observer would have fitted better than the usual 2-degree observer. You would not see the difference if both displays showed a small white patch. You probably would see no difference when both displays were showing a typical image. But there was a difference, and you could see it on some images with a lot of white.

We saw a similar effect with OLEDs. I needed to understand this difference. I put a LaCie CRT and a Sony OLED monitor in a dark room and got a number of people to sit at a comfortable viewing distance and adjust the OLED until it matched the CRT full-screen white. I measured the colours of both monitors using a spectrometer where their eyes had been.

The plot shows the scatter in CIE $u'v'$ space for 31 observers. Even if we ignore a few outliers (and I think all the points are real) the scatter in points is as big as the offset. I saw no significant correlation with the observer's age, so the lens yellow is probably not important. The yellow spot macular dye density varies between observers, so I suspect that.



Observer error matching OLED to CRT

This sort of thing often happens when you try to improve on the three-colour approximation. I hoped to get a correction for the 10 degree observer, but the correction was different for different people. I have a reasonable guess as to what was causing this, but that is no help. In the end, I went back to using the 2-degree observer, with the proviso that it may not give quite the right balance for large images with lots of white.

I did another experiment using just the OLED monitor, with one side covered with a Roscolux coloured filter to change the effective primaries. This was a better experiment because the pixels looked the same, but the spectral shift was smaller, so it was harder to measure. But the results were much the same: different people had different matches.

The most dramatic demonstration of the 10-degree observer I ever saw was when I was trying to match film to a digital projector. We had a Sony 4K projector that used a xenon lamp, and a xenon slide projector to project a matching image from film on the same screen. I could get the white points to match using instruments, but I always saw a colour difference. I could not get a convincing match when adjusting the Sony projector white by hand. In the end, I did away with the image and just looked at a full-screen white. I adjusted the digital projector white to match the slide projector. I could then see a spot about 2-degrees across that moved with my eyes and did not match. It looked a greenish-yellow colour on the digital side, and a lilac colour on the film side. This was the Maxwell spot. The digital projector white was not the same as the slide

projector white: it had a xenon lamp but filtered it, so it used the red, green, and blue bands but not the wavelengths in between.

We cannot correct for the 10-degree observer with just three primaries because we cannot match white for both the central vision and the peripheral vision at the same time. I doubt whether we could match even with a fourth primary. The match will be different for different observers, and depends on where one looks at any instant. I suggest we note this as yet another minor error in the three-colour approximation that is rarely visible with real images; we try to avoid putting displays with different primaries next to each other; and we carry on using the 2-degree observer.

Light units

In the first chapter I briefly mentioned the SI definition of the *candela* as a rare example where a colour was a frequency (540 THz). I have talked about measuring spectra but glossed over the actual units. Now we have covered luminance, which was the Y from CIE XYZ, we can talk about units.

A *radian* (rad) is a measure of angle. It is the length of a circular arc divided by its radius. Radians are like degrees, but there are 2π radians about any point in a 2D plane instead of 360 degrees. Short arcs can be approximated to straight lines: a 1 cm line perpendicular to the line of sight and 1 m away will subtend an angle of about 0.01 radians at the eye. The angle subtended at the eye determines how wide or high something looks.

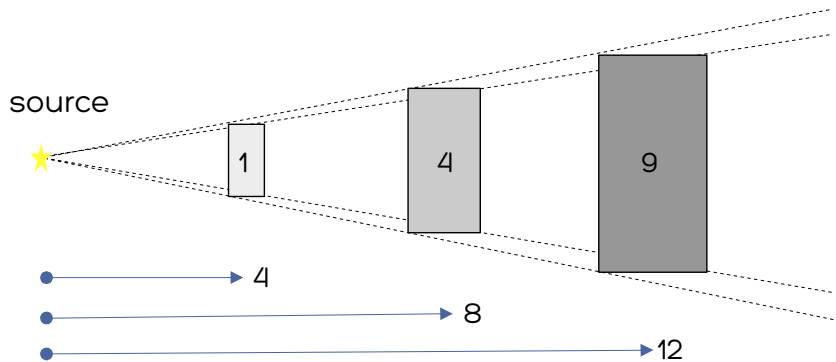
A *steradian* (sr) is a measure of solid angle. It is the area of a segment of a spherical surface of a sphere divided by its radius. Small segments can be approximated to flat areas: a 1 cm square perpendicular to the line of sight and 1 m away will subtend a solid angle of about 0.0001 steradians. There are 4π steradians about any point in a 3D space. The solid angle subtended at the eye determines how big something looks.

The *candela* (cd) is the equivalent brightness of 1/683 watts of 555 nm (540 THz) light per steradian. 555 nm is the peak of the Y curve in the last section. We correct other wavelengths using this Y curve. For example, the weight for 510 nm light is 0.503 because we are about half as sensitive to it. We do not see infrared or ultraviolet light, so they have zero weight.

The 1/683 factor made the candela fit an earlier standard based on a pentane lamp. Not a candle. There is no recorded standard based on candles. Sorry.

In the diagram the single square is one unit area and four units from the light source. Using our small segment approximation, the first square subtends a solid angle of about 1/16 steradian. The second square is four

times the area but twice as far away, so it too subtends about $1/16$ steradian. The same amount of light in watts spreads over a larger area, but the brightness of the lamp in candelas is the same.



The candela measures the total output per steradian of a light source. If the light source is something broad like a flat panel display, we specify the brightness per unit area in *candelas per square metre* (cd/m^2) or *nit*, from the Latin *nitere* – to shine. ‘Nit’ is not an SI term, but it is widely used and easier to type. I shall use nit rather than cd/m^2 from here on. A typical computer display may be 100-400 nit.

A foot-lambert (ft-L) is 3.426 nit. This was still used in the US in 2000, but is rare today. You may meet them in old standards documents. Cinema open gate white was 16 ft-L.

A troland (td) is the brightness in nit divided by the pupil area in mm^2 . This is a figure for the exposure at the retina corrected for pupil size variations. The pupil area may be a measurement of the eye, or the area of a smaller artificial pupil placed in front of the eye. You are unlikely to meet trolands except in research papers. This book has one graph with an axis in trolands. For a typical pupil area of 10 mm^2 , 10 troland = 1 nit.

We specify the total light output in *lumens* (lm), given by multiplying the brightness in candelas by the solid angle in steradians. This is a particularly good unit for devices that emit nearly parallel light, such as lasers, where the solid angle of the beam has little meaning. A small digital projector may output a few thousand lumens (Klm). You can zoom or refocus the projector, and the total output in lumens should remain about the same.

The SI unit of illumination is the *lux* (lx). One lux is an illumination of one lumen per square metre from any angle. 1000 lux is bright indoor lighting e1 lux is low lighting for lecture theatres and telecine suites.

A Lambertian reflector is one that reflects light equally in all directions. If a surface were perfectly white and Lambertian and lit at 1 lux, its reflection would be equivalent to $1/\pi$ nit.

Modern colour measurement instruments

We have mentioned the spectroscope. Here is a list of the different colour measuring instruments we may meet...

A **spectroscope** is not usually a colour measuring instrument because a spectroscope has no detector. You use your eye to see spectral lines and other features. You could attach a camera to some spectroscopes, and measure the exposure. A spectroscope that measures the amount of light at each wavelength is usually called a **spectrometer** (see later).

A **light meter** measures the signal from some light detector, and gives a reading in exposure units.

A **bolometer**, radiometer, or **irradiance meter** measures the amount of radiation falling on the sensor per second. When calibrated, it will read the total power arriving in watts (W) or in watts per square metre (W/m^2) if the area of the detector is taken into account. It may have a filter to restrict the wavelengths it detects, so it can measure IR, UV, or visible light.

A **lux meter** is a light meter with a filter to match the spectral sensitivity to human vision luminance. It may measure diffuse light in lux, or coherent light in nit, or cd/m^2 . An instrument that measures in lux will typically have a white diffuser, so it collects light from all directions over a known area. An instrument that measures in nit will typically have optics or collimation to control the solid angle that it is measuring.

A **colorimeter** is like a radiometer but with three sensors, typically with red, green, and blue filters to match the response of the human eye. These can be good, cheap ways of measuring broadband light, or for measuring some special device such as a video monitor with standard phosphors. They are not good for measuring spectra, particularly if there are sharp spectral bands that coincide with the sides of the red, green, or blue sensitivity peaks. Colorimeters with large detectors can be very sensitive. Most digital colour **cameras** are effectively an array of colorimeters.

A **densitometer** is like a colorimeter with its own light source. Some measure transmitted light; some measure reflected light. They report the measured light in logarithmic units of **density**. Something with a density of 0.0 transmits or reflects all the light; something with a density of 1.0 transmits or reflects 1/10th of the light; something with a density of 2.0 transmits or reflects 1/100th of the light, and so on. Colour densitometers may measure using white light and red, green, and blue filters. These filters may be matched to the dyes in the object being measured such as the status M filters for measuring film negatives; or they may try to match the sensitivities of our eyes such as the Status A filters for measuring positive prints.

A **spectrometer** breaks up the spectrum into many small wavelength bands. A typical solid-state spectrometer will view the light source through a slit, and use a grating to spread the image of a slit onto a line of detectors. This is less sensitive than a colorimeter because the light is collimated to get a sharp spectrum, and then is spread across many detectors. This can limit your ability to measure deep shadow tones. The spectrometer can be calibrated against some standard light source, so each band will have an accurate wavelength and sensitivity.

A spectrometer is not a useful instrument for measuring things like patches of colour in the workplace unless it is comes with optics. A typical colour measurement spectrometer has a lens and an eyepiece, in which case it may be called a **telespectroradiometer**. The sampling region of the spectrometer appears as a dark region on the image in the eyepiece. It can be calibrated to measure the power in watts per square metre steradian nm ($\text{W/m}^2\text{SR.nm}$).

A **multispectral camera** is like an RGB camera but with more channels. These are usually used for machine vision, or for measuring things such as plant health. Multispectral cameras may use a fixed mosaic of filters as most RGB cameras do, or may use a monochrome detector and swap filters to get the different colour channels.

A **hyperspectral camera** returns a spectrum for each pixel. The **hyper-**prefix suggests it goes beyond the visible spectrum. They often measure into the UV and IR, but this is not always the case. There are too many bands to use a filter mosaic, so they scan an image a line at a time, or scan the whole image one waveband at a time. They are not good at capturing motion. They are used for satellite surveying and other scientific measurements. They can also be used to scan historic images such as paintings or early colour films.

So far, none of this tells us what the colour looks like. In the next chapter we explore luminance-chrominance colour spaces, and perceptually uniform colour spaces where equal intervals should correspond to equal visible contrasts.

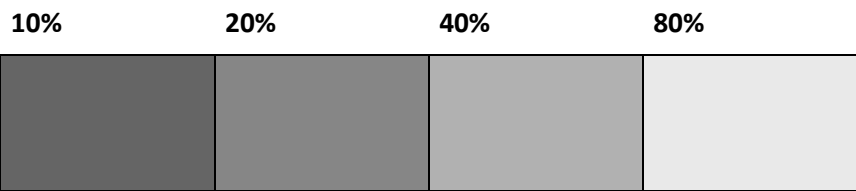
Chapter 4

Uniform colour spaces

In the last chapter we discussed how we measure colours. We have standard observer spaces that should tell us whether two spectral measurements give a visual match. In this chapter we look at colour spaces that are generated from these standard observers, but tell us a bit more about the colours themselves: spaces where similar steps in the values correspond to similar steps in the appearance. Before this, to set the scene, we shall cover some simple luma-chroma terms and colour spaces, and what we understand by the terms *black* and *white*, and then look at the reflection colours in the Munsell Colour Atlas.

Brightness, hue, and saturation

A perfect black absorbs all the light and reflects none. A perfect white reflects all the light and absorbs none. We cannot get perfect blacks and whites, but we can get materials that reflect less than 1%, and more than 99%. We can make grey levels between these limits using halftones...

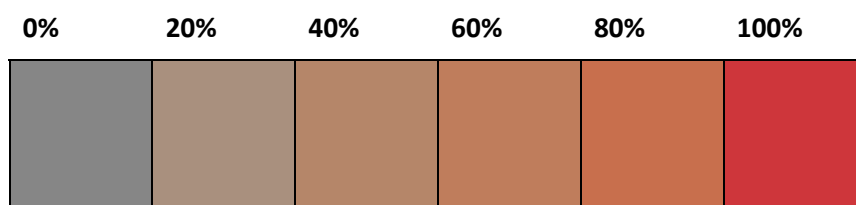


Each patch reflects twice as much as the patch before. I cannot guarantee the actual patches in the diagram you see match this exactly, but they should not be too far off. The visual contrast between any two patches looks about the same. This makes sense when we remember colour constancy from the last chapter. Imagine doubling the light level. You would notice the light getting brighter, but your eyes would adapt, and things would look much as they did before. Here, the contrasts look much the same despite the effect of the white surround. We see equal ratios in light level as equal steps in *brightness*. This is not a perfect rule: we cannot see steps if the light is too dim, and we are dazzled if it is too bright. However, if we ask someone to pick grey colour chips to form equal steps in brightness, this is pretty much what they do.

What about colours? We discussed luminance in the last chapter. If we cannot see flicker when we flip between two colours at 5 cycles per second or more, they have the same luminance. Luminance does not necessarily correspond with brightness as we described it here. However, if we ask someone to pick a grey colour chip with the same brightness as a colour, the results are much the same.

Do luminance ratios look uniform for colours as well as greys? Yes, they do. The argument is much the same as before, but using a coloured light instead of a white one. So, we now seem to have a provisional scale for brightness with uniform looking steps for all colours.

Pick any colour. We can pick a grey with the same brightness. Let us now mix the grey and the colour. Here is an example mixing grey and red, showing the percentage of red...



We have a set of steps from grey to red. We have a brightness parameter going from black to white. This new parameter is often called *saturation*. The first few steps in saturation seem to have similar contrasts. There is no simple argument for generating uniform steps in saturation as we did for brightness, but the contrasts at the patch boundaries look fairly even. If you ask someone to sort coloured chips into equal steps of saturation, this is pretty much what they do.

We could have negative saturations. Any colour has an opposite colour with the same brightness which can mix to give a neutral. The opposite colour to red is cyan. I could make a longer set of patches from mixing cyan and red. The right-hand side would look like the diagram above, but the neutral point would not necessarily lie in the middle.

Chapter 3 has a diagram in the section on the three-colour approximation showing the triangle of mixtures for 530 nm green, and 460 nm blue with the same luminance. The neutral grey is in the centre, and we can see the saturated hues around the edges of the triangle. Our range of colours may not be the same as this diagram, but the general layout of the colours should be similar. The missing third parameter is called *hue*. Hue is often given as a *hue angle* with red at 0 degrees. By convention, we do not have negative saturations; we keep the saturation positive and shift the hue angle by 180 degrees.

We can say what colour lies opposite any given colour, but there is no neat argument that defines a perceptually uniform set of steps in hue. But it seems reasonable to ask someone to pick colour chips with equal

brightness, and arrange them into concentric circles with the same saturation, and colours arranged around the circles arranged in uniform steps of angle.

That last ‘reasonable ask’ may not be possible. You can put the colours into circles of equal saturation, and you might arrange the hues into angles with equal contrast on one circle, but that leaves you with no degrees of freedom for lining up the other circles. It may not be possible to fit a perceptually uniform set of coordinates in a 3D space. However, you can get close: just pick the biggest, most saturated circle with a full set of hues, and make that look right, then line up the other hues.

We have a set of hand-waving arguments to say there should be a perceptually uniform scale of brightness for all colours, and two other axes, called hue and saturation. If enough people took a set of coloured chips, and arranged them into sets of equal perceptual steps, it ought to be possible to agree on a perceptually uniform set of steps in brightness, hue, and value.

Albert Henry Munsell made the first standard set of colour chips, and people have been trying to do just that for over a hundred years. Let’s see how they got on.

The Munsell Colour Atlas

In 1905, Albert Henry Munsell published ‘A Colour Notation’, describing a colour coordinate system designed much along the lines of our last section. In 1913, he published ‘The Atlas of the Colour Solid’ with accurate examples of the colours, and later ‘The Munsell Book of Colour’ with removable colour chips. The Book of Colour is still in production. In the hundred years and more, the set of colours has been expanded, but the underlying system has remained the same.

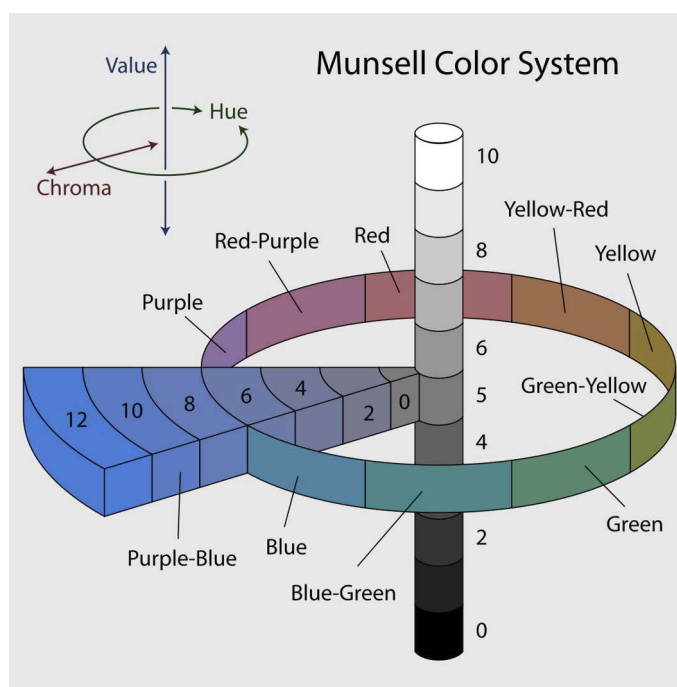
The Munsell colour system has three parameters: Value (lightness), Hue, and Chroma (saturation). The Munsell Colour Atlas has a set of coloured chips at regular intervals in Hue, Value and Chroma, as judged by a human observer under normal daylight conditions. In Prof. Munsell’s original experiments, ‘daylight’ may have been the now obsolete CIE standard Illuminant C, which approximates to 6700K. We probably get a good approximation to Munsell’s original experiments if we viewed the chips in the Munsell Colour Book under 6500K illumination. The Colour Book also looks uniform under a D50, the standard for reflection media viewing booths. These two whites cover most of the viewing conditions for motion pictures, TV, and printing.

The Munsell Colour System might be less perceptually uniform than some other spaces such as OSA-UCS or CIECAM02. It is a set of colour chips that will change colour with time. The illuminant is not specified, though it seems uniform for whites between D50 and D65. The Chroma

(hue) axis has an odd notation. But it has a simple empirical basis: the chips show equal steps in colour because real people experimented until they looked right. Some of the more recent, more saturated chips may have had some help from new vision models, but they still are eyeballed by real people. So, when we meet perceptually uniform colour spaces based on tristimulus XYZ or LMS, we shall use the Munsell Colour Book as a benchmark. If the shapes of the points in these colour spaces looks like the Munsell Colour System, then they are working.

Munsell Value, Hue, and Chroma

At the start of this chapter we described three possible values called brightness, hue, and saturation that might describe colours as we see them. Brightness might have uniform perceptual steps as ratios of light level, but the other two axes might have to be fitted by eye. Here is a diagram for the Munsell colour system.



The vertical coordinate is called Value, and ranges from 0 (black) to 10 (white). Other systems call this *brightness*, *lightness*, or sometimes just *value*. Changing brightness is simple in theory: if we have a colour spectrum, we scale up all the values. The colour stays the same, but you have more of everything.

The circular coordinate is called Hue. Other systems call this *hue* or *hue angle*; hue angle has a value between 0 and 360 degrees. Munsell Hue has ten bands with colour names. Each band has 10 colours 0-9, with 5 at

the centre. Zero is the 10 of the previous band. If there was a 10, it would be the zero of the next band. The Colour Book has 40 hues in Munsell Hue steps of 2.5.

The last coordinate is the radius measured from the neutral axis. The Munsell name for this is *Chroma*. The figure shows a wedge of Chroma values, going from zero (neutral) to 12 (a colourful blue-purple) The Chroma range depends on Value and Hue. Other systems call this *saturation*, or *chromatic intensity*. It is not clear who first chose the term 'saturation' or why, but it is now in wide use. You start from zero saturation (grey, no colour) and go outwards until the colour is *saturated*, and you can go no further.

Ambiguous Colour Terms

Saturated colours are sometimes called 'bright colours'. Please do not use this term yourself, and take care when others do. There is a thing called the Helmholtz-Kohlrausch effect: really bright and saturated colours seem to have a 'glow' about them, so there is a genuine perceptual overlap between saturation and apparent brightness. We meet the Helmholtz-Kohlrausch effect again in Chapter 7. In the meantime, please join me in using 'bright' as the opposite of 'dim', and not a synonym for 'saturated'.

These three terms, *brightness*, *hue*, and *saturation*, are used to describe colours as we see them. They are used in user interfaces, but they are rarely used in digital imaging because they are hard to interpolate. The *hue* axis goes in a circle: 360 degrees is the same hue as 0.0 degrees. If we pass through the neutral axis, the *saturation* does not go negative: it stays positive and the *hue* flips by 180 degrees. It is usually a lot easier to use three Cartesian axes such as XYZ or RGB or LMS when dealing with colour in images.

The meaning of 'brightness' may depend on context. If you ask how bright a display is, you probably will get an answer in nits. Most perceptual measurements of brightness are based on the luminance value, but the nit is not itself a perceptual measure: we might barely see a change of 1 nit on a 1000 nit display, but it would overwhelm shadow detail at 0.1 nit.

There is one other common usage: chroma is generally used for all the colour description that is not luminance. In the Yxy and $Yu'v'$ colour spaces we meet next, we might refer to Y as luminance and xy or $u'v'$ as *chroma*. Munsell uses *Chroma* or *Chromatic Intensity* just for saturation. I shall use chroma with a small 'c' for the general term, and Chroma with a capital 'C' for the Munsell axis. I write chroma axes in **bold** font in this document.

Yxy colour space

Yxy is the simplest luminance-chrominance space. It is derived from CIE 1931 XYZ with...

$$x = \frac{X}{X+Y+Z}$$

$$y = \frac{Y}{X+Y+Z}$$

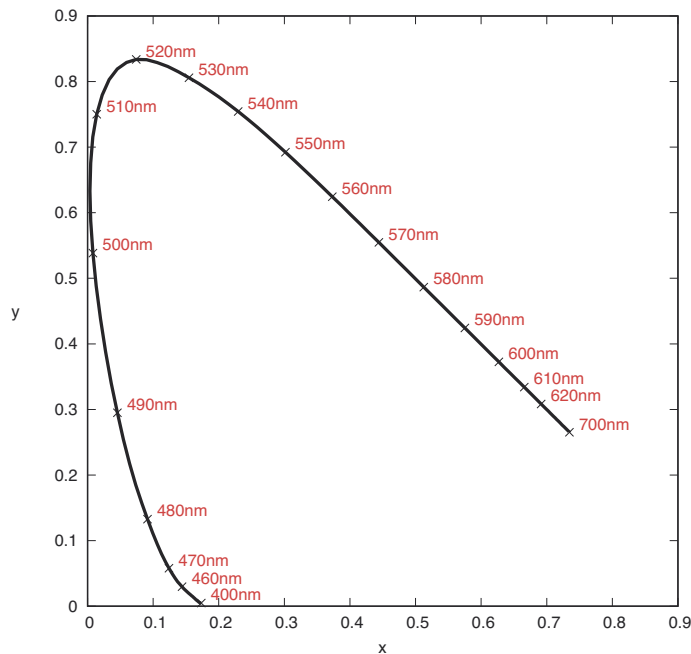
Y is the luminance measured in nit, and the x and y co-ordinates are dimensionless numbers that describe the colour. The inverse transform is...

$$X = \frac{x \cdot Y}{y}$$

$$Z = \frac{(1 - x - y) \cdot Y}{x}$$

The **xy** space has linear mixtures: If you draw a line between two points plotted in **xy**, any point on the line can be made from a linear mixture of the light from the colours at either end. The space is perceptually uneven: a small difference in **x** and **y** in the blue-violet corner is more visible than the same difference in the green corner.

Here are the wavelengths from 400 nm to 700 nm plotted in CIE **xy**...



Yu'v' colour space

Yu'v' is a uniform luminance-chrominance space. It is derived from XYZ with...

$$u' = \frac{4X}{X + 15Y + 3Z}$$

$$v' = \frac{9Y}{X + 15Y + 3Z}$$

Y is the luminance measured in nit, and the **u'** and **v'** co-ordinates are dimensionless numbers that describe the colour. The inverse transform is...

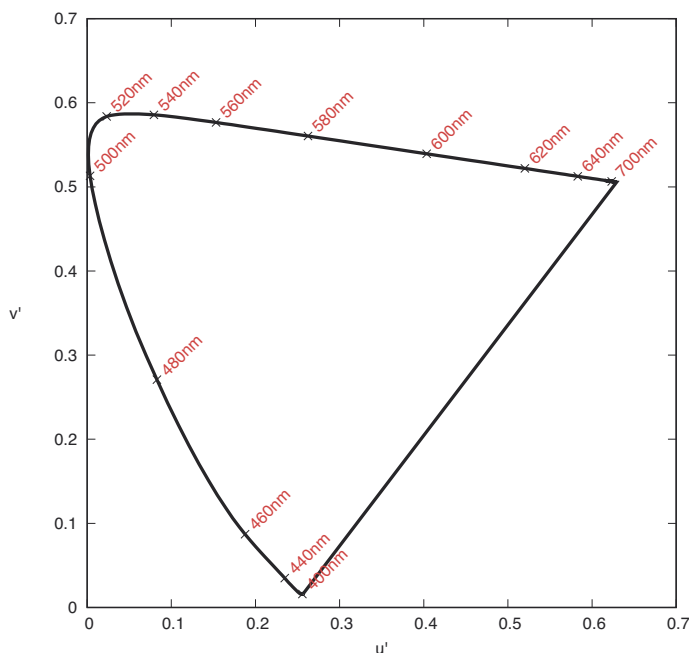
$$X = \frac{9Yu'}{4v'}$$

$$Z = \frac{3Y(4 - u')}{4v'} - 5Y$$

The blue-violet corner has been stretched out, and the green corner has been compressed. It is not perfect, but it is better than **xy**. The **u'v'** space also has linear mixtures.

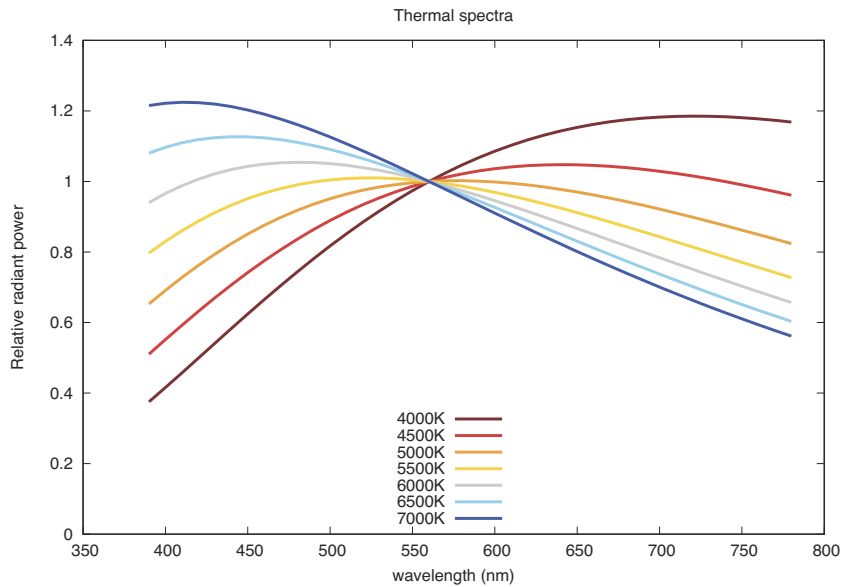
There was a uv space before $u'v'$. It existed for less than a year before it was improved, so any modern reference to Yuv is probably a reference to $Yu'v'$. There is a construction in Yuv for finding the closest thermal white (see next section) to a given near-white colour which is still used. I see no advantage in using this construction in uv , as it is easy to find the nearest point in $u'v'$ with a computer. My calculations were always slightly different to the ones in the literature, and this was why.

Here are the wavelengths from 400 nm to 700 nm plotted in CIE $u'v'$...

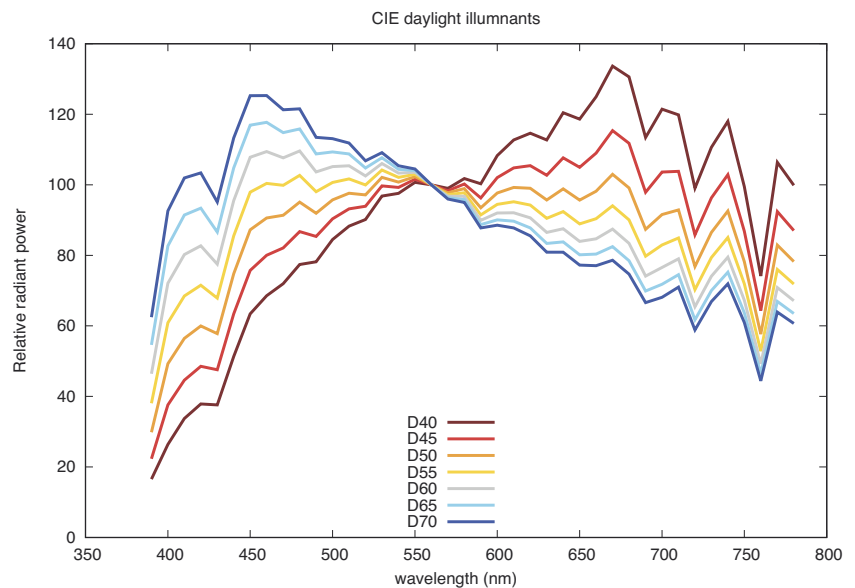


White light

A hot body gives off light. The light spectrum normally depends on the shape of the body, and body material, as well as the body temperature. If we have a large cavity at a uniform temperature, and we sample the light through a small hole so we do not significantly disturb the radiation equilibrium in the cavity, the spectrum then depends on the body temperature alone. These spectra are called blackbody spectra, thermal spectra, incandescent spectra, or Planckian spectra.



The CIE standard daylight illuminants are based on measurements of daylight spectra. The daylight spectrum was measured under a variety of conditions, and reduced to a mean spectrum, and two other spectral functions which can be combined to give a whole family of spectra. These spectra are usually named according to the approximate colour temperature. D65, for example, is a Daylight illuminant with an approximate colour temperature of 6500K.



You cannot buy a light source that fits any of these CIE spectra. The spectral distribution is too complex to reproduce in an artificial lamp. These CIE daylight simulation illuminants only exist as a theoretical ideal

for calculations. When people describe a fluorescent tube as being D65, they probably mean the colour is matched to D65 with a standard observer.

There is no reason why the CIE daylight illuminants should fall on the thermal locus. The points come close, but there are small differences...

Thermal	x	y	CIE	x	y
4000K	0.3820	0.3792	D40	0.3823	0.3836
4500K	0.3620	0.3656	D45	0.3620	0.3707
5000K	0.3460	0.3532	D50	0.3457	0.3585
5500K	0.3330	0.3421	D55	0.3325	0.3474
6000K	0.3224	0.3324	D60	0.3216	0.3376
6500K	0.3137	0.3239	D65	0.3127	0.3290
7000K	0.3063	0.3164	D70	0.3053	0.3215

Here is a list of whites. Which one of these is *white*? The answer is many thermal sources can look white, including all of these...

Light source	Temp
Old monitor white	9300K
Old monitor white	7200K
CIE D65 daylight standard	6504K
Video ITU rec. 709 white	6500K
Kodak Cinema recommendation	6100K
Kodak slide viewer recommendation	5500K
Kodak Cineon view recommendation	5400K
US Cinema standard	5300K
European cinema standard	5200K
ICC print viewing D50	5000K
Art museum viewing standard	3700K
Steenbeck (incandescent)	3000K
CIE illuminant A	2856K
Candle flame	1800K

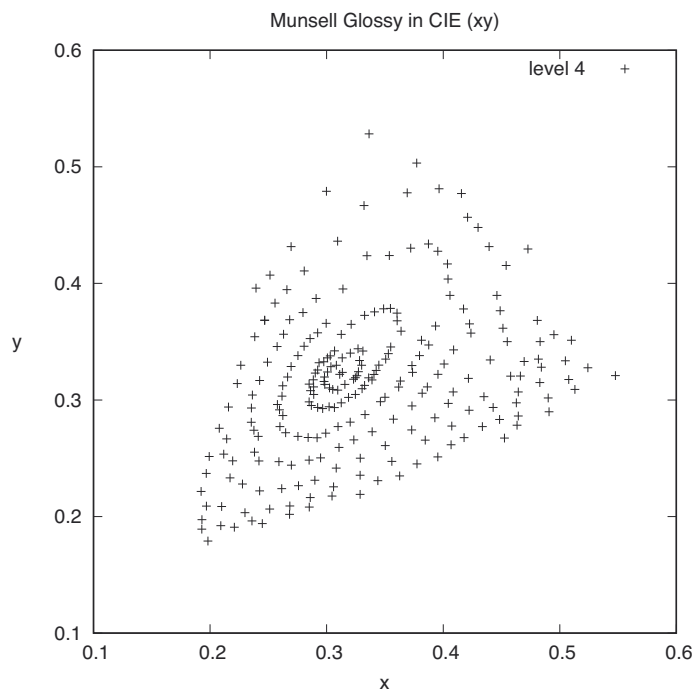
The significant thing here is that both the thermal values and the CIE daylight illuminants white values lie close to a single, common curve. If you go off this curve, the light looks pink or green, and quite unnatural. It seems all natural lighting and most artificial lighting fit the blackbody curve. It is what we are used to and what we like.

I use blackbody curves for most simulations because they are smooth and well-behaved. I get almost the same result with the CIE daylight illuminants. Real light sources such as fluorescent lamps or LEDs may look white, but can have peaks in their spectrum which change the colours we see.

The Munsell benchmark

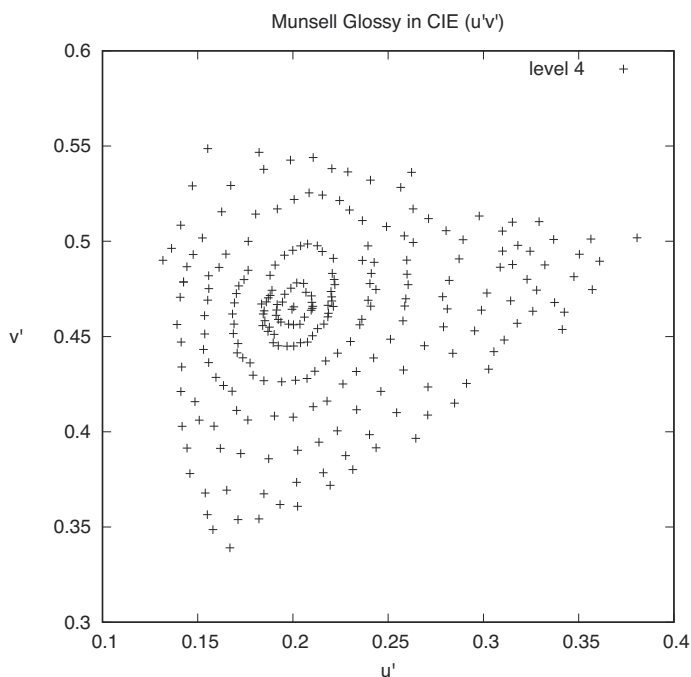
The Munsell Colour Book has a set of coloured chips arranged into uniform perceptual steps of Value (lightness), Hue and Chroma (colour purity or saturation). We cannot measure any of these parameters directly, but they have been widely agreed upon. Suppose we measure the Munsell colours using an illuminant in the 5000K to 6700K range where the Munsell set looks perceptually uniform. If we plot Munsell colours in some other uniform colour space, we should get the same arrangement of Value layers, Hue angles, and Chroma circles.

Here we plot some Yxy values measured from the Value 4 level of the Munsell space. These should be a set of colours with the same Y luminance.



These are experimental values measured with a spectrometer off shiny material. We do not expect perfect circles, but we can see some rather wonky ellipses. With some imagination you may see some radiating lines of points with the same hue.

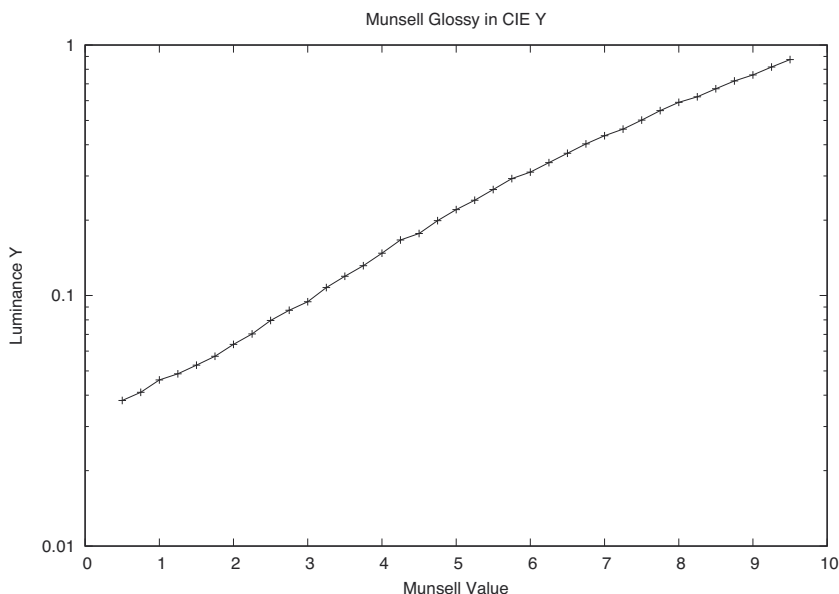
Now let us do the same with the $Y_u'v'$ space...



The $Y_u'v'$ space is still pretty simple, but the colour axes were chosen to give more uniform colour contrasts than Yxy . We can see the ellipses are nearer to circles.

I simulated the illumination using the 6500K blackbody spectrum but the $Y_u'v'$ colour space knows nothing about that. There may be another illuminant that makes things better still. There are certainly illuminants that make things worse.

Finally, let us look at the Y value. This is the same for Y_{xy} and $Y_{u'v'}$ so one plot will do for both...



Equal steps in Munsell Value give similar visual contrasts. We plot reflection Y against Munsell Value for the set of neutral chips. We have a log scale on the Y axis. This is close to a straight line, so equal steps in Munsell Value give similar ratios of Y, at least over the rather limited range of the Munsell chips.

Let us compare Munsell Value against linear light level by eye, just to show how non-linear our lightness perception is. This is risky, but I hope you see roughly what I see, whether you are looking at a printed copy or a digital version on a display. Here is a strip of nine simulated Munsell patches with Value from 1 to 9 against a black background. Below that there is a similar set of patches with uniform steps in light level. The first and last patches in both sets are the same grey.

**Munsell
Value:**

Light level:



The contrast between Munsell patch 1 and 2 on the left is hard to spot on my monitor with ambient light, but the other contrasts look fairly similar. The light level contrast is all squashed towards the left: the second light level patch is almost the same as the middle Munsell patch 5.

Display gamma

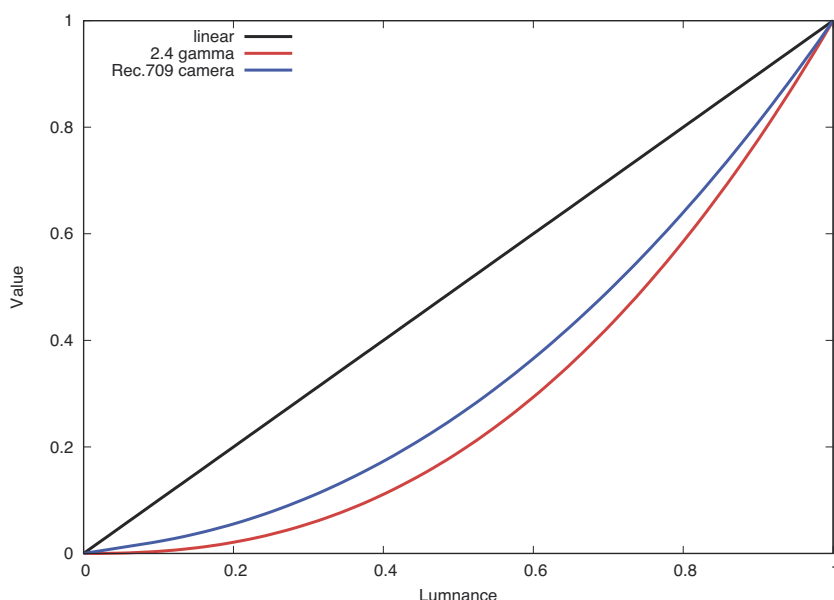
Displays often use a simple power law tone function...

$$\text{luminance} = \text{value}^{\text{gamma}}$$

$$\text{value} = \text{luminance}^{1/\text{gamma}}$$

Formulae may use the Greek letter γ , but it is common to refer to the 'gamma' of a display.

Video uses a gamma value of 2.4. It does not give equal increments for equal contrasts as the log scale did, but it does stretch the contrasts in the shadows similarly. It also goes down to zero nicely, where the log scale never quite gets there.



One problem with power law functions like this is that the curve goes completely flat as you approach zero. This can crush your shadow detail. The inverse function becomes infinitely steep. The curve also stops at zero. Negative light has no real meaning, but you sometimes have camera RGB values with noise, which can give negative RGB light levels. You could clip the negative values because negative light is meaningless, but then the average value in the shadows goes up. A better tone function might go through the origin with a finite slope, and keep going for a bit to let us handle negative values. These 'better tone functions' are used in spaces such as sRGB. But they have their drawbacks too: the choice of slope at the origin depends on the display and the viewing conditions, and the low slope can extrapolate to give large negative RGB values.

We shall talk about viewing conditions and explain the Rec.709 camera curve in the next section.

Scene and display

The Rec.709 camera curve was a specification for TV cameras. The camera tone curve was chosen to give a pleasing image on a general television set. This tone curve clearly does not match the 2.4 gamma of the TV set. Here are the grey levels for the two-tone curves...

2.4 Gamma:



**Rec.709
camera:**

I cannot be sure you see what I see, but I hope this gives a better idea of what the differences look like than the graph. The 2.4 gamma patches are uniform steps from 0.1 to 0.9 representing the TV light levels, and the Rec.709 camera patches are the corresponding light levels that the TV camera would see. The highlights look much the same but darkest shadow tones are darker on the 2.4 gamma TV. If I had gone to lower light levels, the contrast would become more obvious still.

Why does the camera not match the display? This is the effect of the surround. The camera scene is probably surrounded by similar light levels to the scene itself. Even in a lit studio, there will be some bright surround. We usually watch TV with a dim surround. A standard TV grading suite has a neutrally lit grey wall behind the display to match these typical viewing conditions. We see more shadow tones with this dim surround, so we have to make the camera shadows dimmer before what we see matches what the cameraman saw.

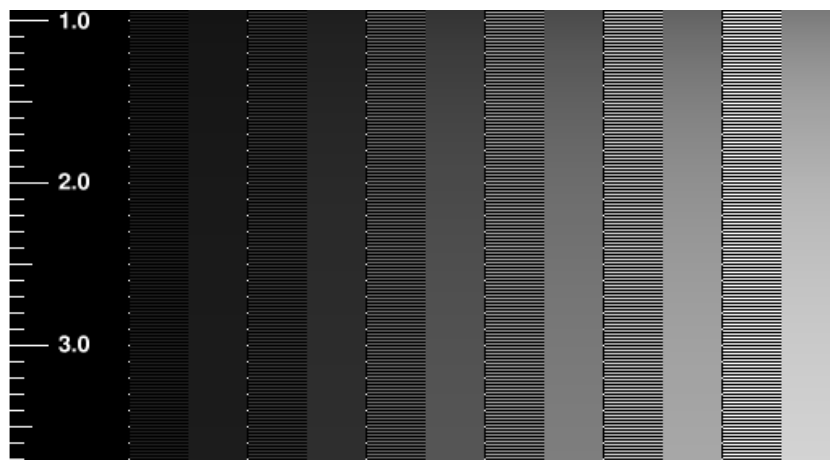
Historically this made sense. A CRT TV had a natural gamma of about 2.4, which is a good match to our greyscale perception. There were many more TV sets than TV cameras, so it made financial sense to put the expensive analogue tone correction into the camera rather than the TV set. Back then people were not going to lug their CRT set into the garden, so the dim surround assumption was good. Today we can watch TV on our mobile phone outdoors, and the OLED or LCD we are watching it on has digital colour correction. If we were to re-invent broadcast TV today we might put much more correction for viewing conditions into the display.

Film has a similar built-in correction for the surround. Motion pictures were shot out of doors or on a brilliantly-lit set, but were always viewed in a dark theatre. Monochrome film labs built on what looked good last time, and converged on a good-looking tone curve. Colour film needed the same tone curve to look right in a dark theatre, but also needed the

right balance between the three colour channels to get neutrals and flesh tones looking right. And they had to do this all with wet chemistry, and no computers: colour film was an extraordinary achievement.

Gamma test image

Here is an image that can determine the gamma on your monitor...



Gamma test image

The image has pixel level detail. If you are looking at this on a computer, you may need to reset any zoom settings. If the left column of each pair looks smooth and even, it is probably displaying correctly.

Note:

To access an online version of this book, and other related resources, go to:
www.filmlight.ltd.uk/colourbook

Look at the column pair on the right. This has black and white stripes on the left, and a grey wedge on the right. Pick the height where the grey wedge is the same average colour as the black and white lines. You may find it easier to determine where the grey is definitely lighter, and definitely darker, and take the middle point. If you are viewing an LCD, you move your head so you are looking straight at this point, as the colours change with viewing angle. Once you have found the point, you can read off the gamma on the scale on the left.

The next column to the left has black and 50% grey stripes. The column after that has black and 25% grey stripes, and so on. I would expect the gamma values to be roughly the same, but they may wander about a bit in the shadows. Eventually, the gamma values may not be right because the chart assumes your 0% black level is black, and your monitor 'black' probably isn't.

If you look at this on a computer, the gamma value ought to be about 2.2. If you look at a printed copy, I hope it would be about the same, but I am much less certain. If you do not get the exact gamma value you expect, do not worry: the document viewer or computer settings may be at fault, and not your display. This still shows you should be able to estimate the gamma on a display using just an image and your eyes.

CIE L*a*b* (CIELAB)

CIE L*a*b* is a space derived in 1976 from CIE 1931 XYZ. Colour is defined with respect to a reference white $X_W Y_W Z_W$ using a function f as follows:

$$f(A) = \begin{cases} A^{1/3} & A \geq 0.008856 \\ 7.787 \cdot A + 0.1379 & A \leq 0.008856 \end{cases}$$

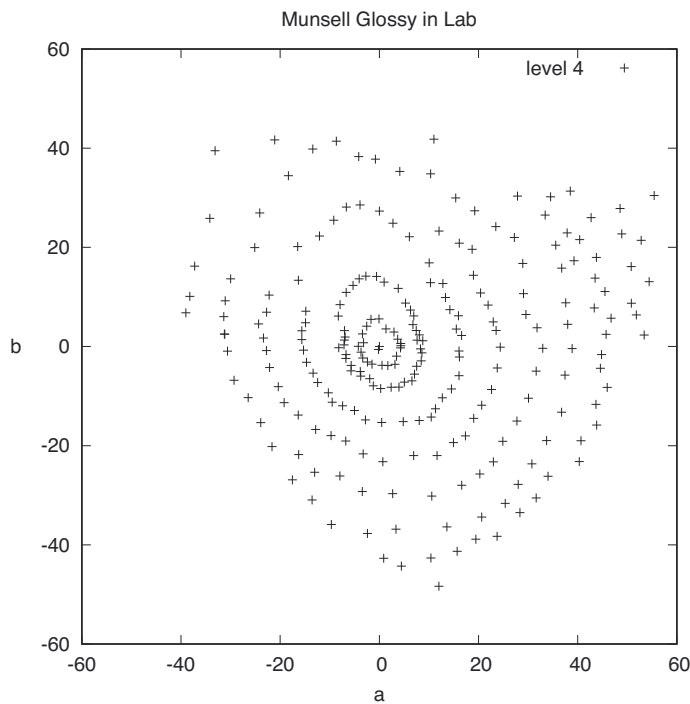
$$L^* = 116 \cdot f\left(\frac{Y}{Y_W}\right) - 16$$

$$a^* = 500 \cdot \left[f\left(\frac{X}{X_W}\right) - f\left(\frac{Y}{Y_W}\right) \right]$$

$$b^* = 200 \cdot \left[f\left(\frac{Y}{Y_W}\right) - f\left(\frac{Z}{Z_W}\right) \right]$$

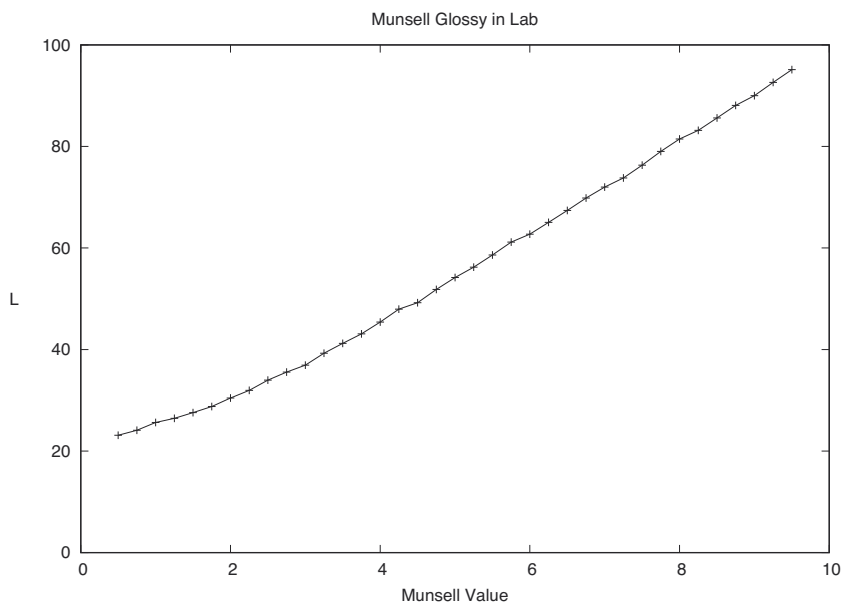
The f function has a cube root region. This roughly fits the log plot we saw for Munsell patches, or the gamma curve for video. It has a linear region for very deep shadows, so it is well-behaved for zero and even for negative values, where the simple log and gamma functions stop.

The f function is applied to X, Y and Z. This is sometimes called the *false von Kries* approximation. In 1904, von Kries published a proposal that the three detectors in the eye might have some common tone curve function, much like colour film. This is sensible, but we should apply it to LMS and not XYZ. It turns out the errors are small for white points near D65. The XYZ to LMS matrix is commonly called the von Kries matrix.



The **a,b** coordinates are not circles, but they are the closest we have come so far. The white point is at the origin.

The *L** value is linear from 10 down to about 3. The values below 2 are pretty dark for reflection material; it is hard to get reflectances below 1% even with glossy stock. The CIE *L*a*b** space is used with displays; a good display should have a contrast ratio of over 1000:1, so it will have to go below 0.1%. It is not surprising we see some differences in the dark end.



The L^* value goes from 0 (black) to 100 (white). L^* values over 100 are handled sensibly, but the range ought to stop at white.

The a^* and b^* values are usually in the -100 to +100 range, but real colours can go well beyond this. These values get smaller if the chroma stays the same but L^* reduces. This stops us getting extreme colour values from noise in the deep shadows.

This $L^*a^*b^*$ space is usually considered to be too simple to be a Colour Appearance Model (CAM), but the same sort of construction is used within more complex spaces such as CIECAM-02.

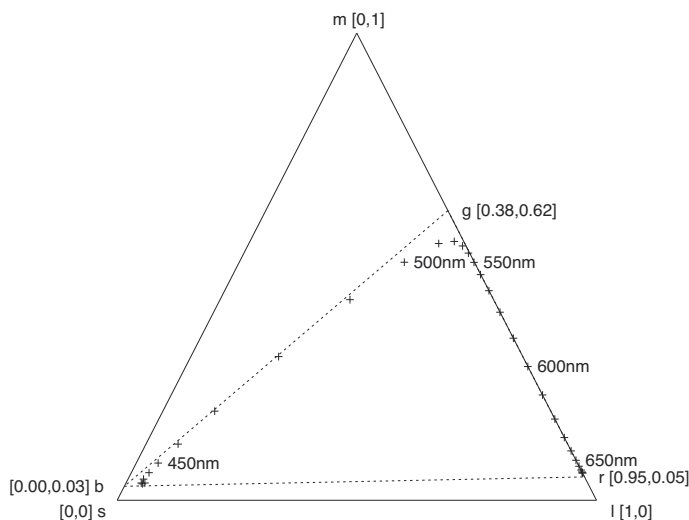
Note:

*CIE $L^*a^*b^*$ is also referred to as CIELAB. This can be handy if the $*$ character gives you text formatting problems. The CIE chose the coordinates L^* , a^* and b^* to avoid confusion with the Hunter Lab colour space, defined in 1948 by Richard S. Hunter. I have never used Hunter Lab, and it is unlikely anyone you meet will have either. If someone refers to 'Lab', I expect they mean CIELAB.*

tRGB from CIE 2006

tRGB, or triangle-RGB is a non-standard space derived from CIE 2006 LMS.

We scale the LMS values to give normalised lms where $l+m+s = 1.0$. We can then show our lms values as a barycentric plot...



In this triangle, 'm' is the height: 0.0 on the bottom line, and 1.0 at the top. The values for 'l' and 's' are measured from the other two edges a similar way. This is an equilateral triangle, so all three values must sum to 1.0.

We plot the points from the LMS values for the single wavelengths from 400 nm to 700 nm within this triangle. This shape is called the

spectral locus. If you join the last point to the first point, you have a closed shape that contains all the real colours. Any point outside this curve is not a real colour: you cannot make a spectrum that will generate the LMS values without using negative numbers.

We draw a triangle just outside this spectral locus. The three corners are labelled 'r', 'g' and 'b'. These are not real colours, but something like 94% of this triangle is filled with real colours.

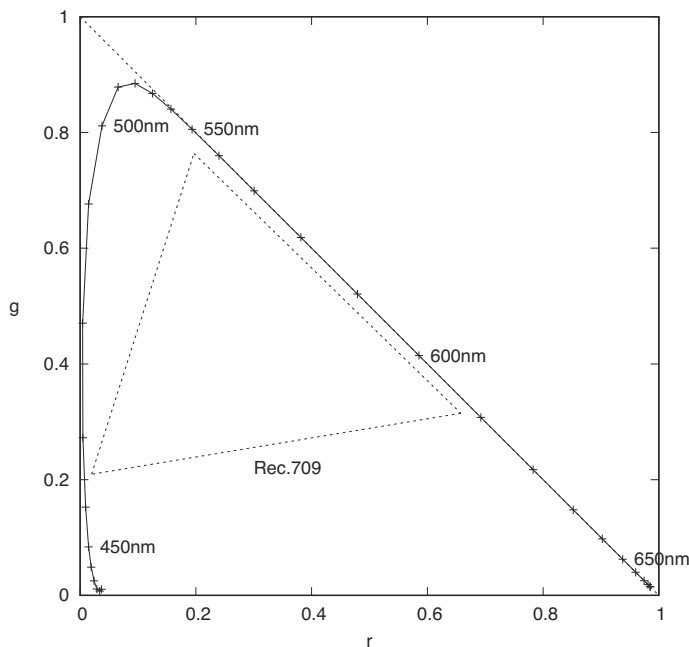
We can convert between *lms* and *rgb* with...

$$\begin{pmatrix} l \\ m \\ s \end{pmatrix} = \begin{pmatrix} 0.95 & 0.38 & 0.0 \\ 0.05 & 0.62 & 0.03 \\ 0.0 & 0.0 & 0.97 \end{pmatrix} \cdot \begin{pmatrix} r \\ g \\ b \end{pmatrix}$$

$$\begin{pmatrix} r \\ g \\ b \end{pmatrix} = \begin{pmatrix} +1.087720 & -0.666667 & +0.020619 \\ -0.087719 & +1.666667 & -0.051546 \\ 0.0 & 0.0 & +1.030930 \end{pmatrix} \cdot \begin{pmatrix} l \\ m \\ s \end{pmatrix}$$

The same matrices convert between unnormalised LMS and tRGB.

The plot below is a stretched version of the previous triangle. I moved the point 'r' to the bottom right, point 'g' to the top left, and point 'b' to the bottom left corner. The top right triangle is all impossible colours. The bottom left triangle is mostly real colours.



Inside this triangle I plotted the video Rec.709 gamut. There is an irregular shape called the Pointer gamut which contains all the colours

from a large set of real scenes, and I plot this out in Chapter 7. The Rec.709 gamut fits inside the Pointer gamut shape. I drew the Rec.709 gamut here instead of the Pointer gamut, as it is a simple shape and makes for a tidier diagram.

The horizontal and vertical axes are the new chroma coordinates r, g . At the bottom of the triangle there is a large region of violet, purple, and deep red colours. The closing edge of the spectral locus is made of mixtures of 400 nm and 700 nm light. We are not very sensitive to either wavelength. If these were reflection colours, they would be very dim. In real scenes you will probably not find a pixel with a g value below 0.1.

We can calculate r, g directly from LMS with...

$$g = \frac{1.7182 * M - 0.0362 * L}{L + M + S}$$

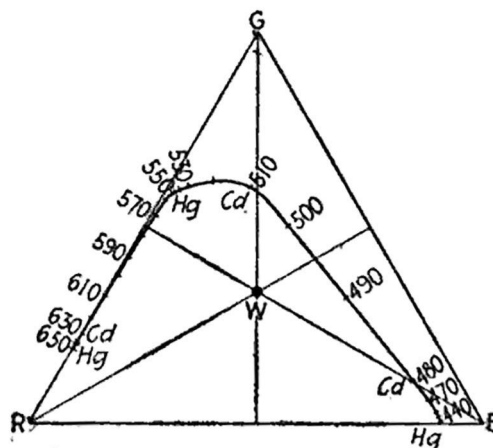
$$r = \frac{1.0671 * L - 0.6873 * M}{L + M + S}$$

This defines the chroma. Chapter 3 had a simple formula for generating luminance Y from LMS...

$$Y = 0.68990272 * L + 0.34832189 * M$$

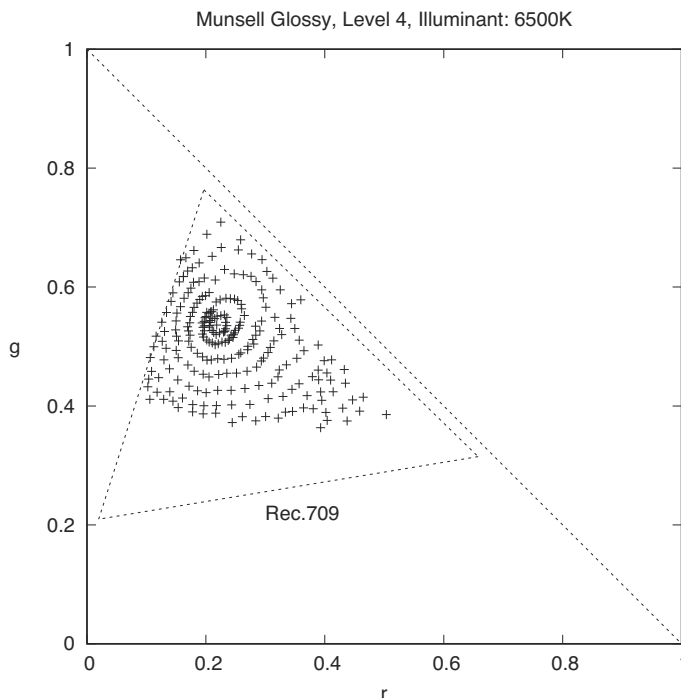
Put them together and we have a Yrg colour space, similar to CIE $Y_u'v'$.

This is not completely new. Here is Maxwell's triangle from a book by Lukiesh from 1921. He is using a barycentric plot, with the B and G corners moved to get his white in the middle: flip it over, and stretch it a bit, and this is our diagram.



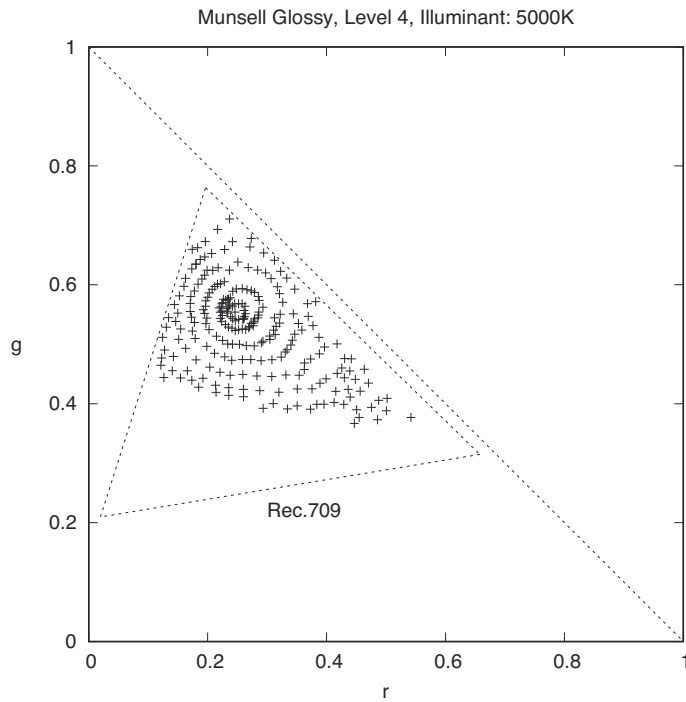
Maxwell's triangle

Here we plot the colours from level 4 of the Munsell Glossy set in the new r, g axes for our 6500K white point...



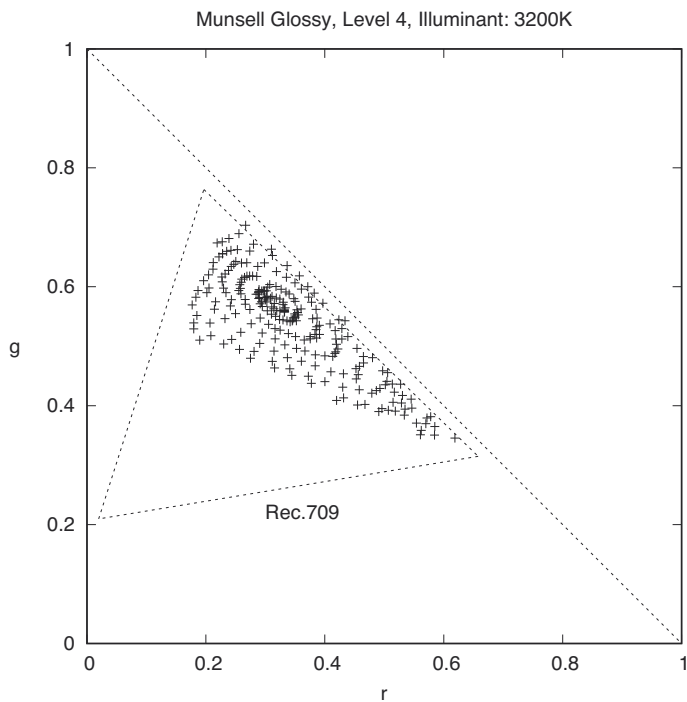
I see no obvious reason why the chroma should come out as circles like a perceptually uniform colour space, but it does. If you put red and blue on the diagonal and have green at the origin, then it doesn't. The r, g axes also mean the long diagonal has red at one end, green at the other, and yellow in the middle, rather than squashing them onto one of the axes.

Our white point is nothing special in this colour space. Let's try a different white...



If anything, the circles look rounder than for D65. I have no evidence that the Munsell space was ever fitted to D50, but D50 has been a standard for reflection viewing booths for a long while, and this may have had some influence.

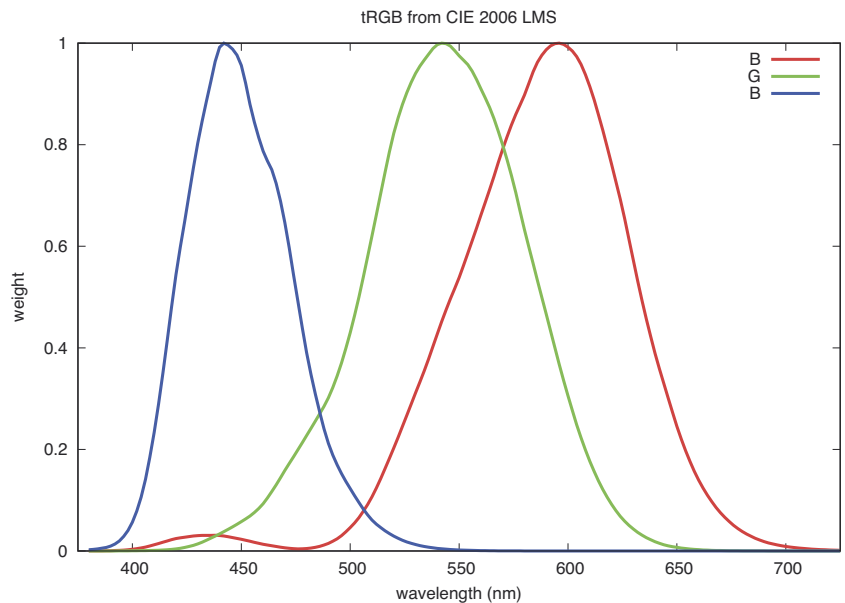
One more, just to see what happens: tungsten light...



We do not expect circles. The Munsell space is uniform for whites between D65 and D50, but it is not suited to tungsten light. We do not get circles, so everything is working as it should.

We do not expect to see much blue contrast. The blue corner is the origin, and we can see the points shying away from it. Some yellow points have crept outside the Rec.709 gamut.

The next plot shows the tRGB weights. Here, tRGB or triangle-RGB primaries have been made from the LMS weights, and the matrix we used to get to the **r,g** chroma.



LMS had little separation between the M and S curves. The XYZ X curve had a large second peak in the blue. Here we see a good separation between R and G, and the R curve has only a small second peak. These curves look like camera primaries. Again, this is a sign that things are working as they should: people have made cameras for years, adjusting the RGB filters to give a good colour separation, but also to 'see' colours as we do.

tLab derived from tRGB

LMS does not have an official luma-chroma space like CIELAB. This is a shame as LMS is a better and more modern set of tristimulus values. So, here goes...

CIELAB calculates L from Y (luminance). Chapter 3 had a simple formula for generating luminance Y from LMS. So, let's use this formula to make Y, then copy the calculation for L from CIELAB...

$$Y = 0.68990272 * L + 0.34832189 * M$$

$$f(A) = \begin{cases} A^{1/3} & A \geq 0.008856 \\ 7.787 \cdot A + 0.1379 & A \leq 0.008856 \end{cases}$$

$$L = 116 \cdot f\left(\frac{Y}{Y_W}\right) - 16$$

The **r,g** chroma axes made nice circles of the Munsell points, so let's borrow that calculation...

$$g = \frac{1.7182 * M - 0.0362 * L}{L + M + S}$$

$$r = \frac{1.0671 * L - 0.6873 * M}{L + M + S}$$

The **r,g** white point is not on the origin like CIECAM **a,b**, so let's subtract the white point values to put it in the right place...

$$\Delta r = r - r_W$$

$$\Delta g = g - g_W$$

The Δr and Δg axes are at 45 degrees to the usual **a,b** axes, so let's rotate them so red lies along the first axis. Let's also scale by L so we get less colourfulness in the dim colours the way CIELAB does. For my first attempt, I scaled up both axes by 4.0 to fit CIELAB **a,b**...

$$a' = 4.0 * L * (\Delta r - \Delta g)$$

$$b' = 4.0 * L * (\Delta r + \Delta g)$$

This gave good Munsell circles for whites from D65 and D50, but we had much less contrast on our b' than on CIELAB b with 3200K tungsten light. My second attempt had a fixed scale for a , and fitted the scale of b to the white point...

$$a = L * (\Delta r - \Delta g) / 0.25$$

$$b = L * (\Delta r + \Delta g) / b_w$$

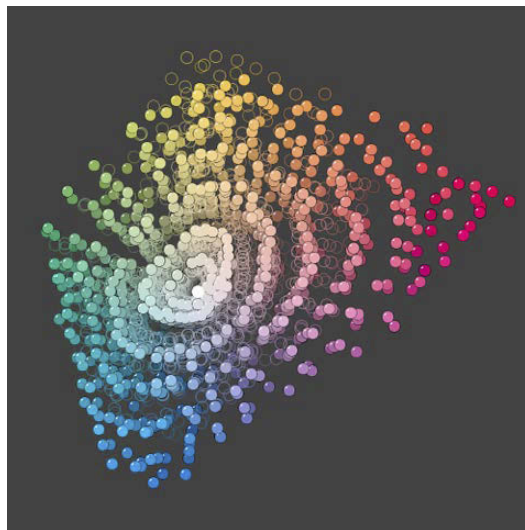
...where...

$$b_w = 1.0 - r_w - g_w$$

The b_w value for D65 is 0.235. The b_w value for 6500K blackbody is 0.243. So this is almost the same as before for white points around D65, and fits much better at 3200K.

Let's see if it works. Let's plot out the Munsell points in tLAB and CIELAB.

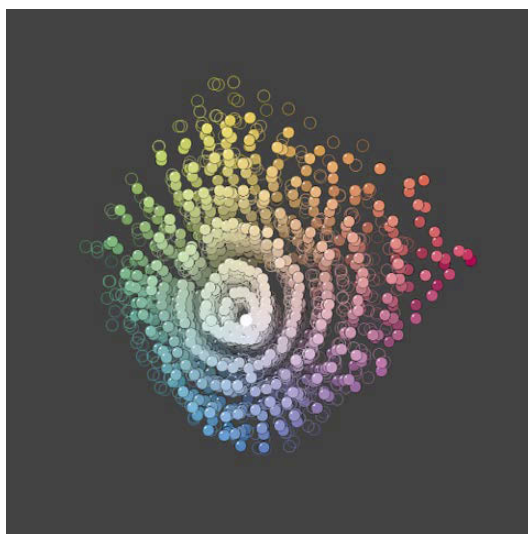
Here are the Munsell colours plotted out for a 3200K white in tLab (filled circles) and CIELAB (hollow circles)...



Munsell colours for 3200K white

The tLab central shape is not a circle, but neither is the CIELAB shape. In fact, they are a similar 'D' shape (not easy to see in this combined plot). The saturated colours seem more evenly spaced in tLab, except for the red corner which perhaps goes too far.

Here is a similar plot with D65 white...



Munsell colours, D65 white

The tLab space is similar to CIELAB. The white point is (100,0,0). The red-green axis is horizontal, and the blue-yellow axis is vertical. Equal steps in L and a, b give similar visual contrasts. We do not see quite the same distribution of points: the cyan edge is squashed in a little more than I would like. I cannot pull it out without doing some non-linear transform. This would be a shame as tLab is linear in a, b chroma, meaning points on a line between a_1, b_1 and a_2, b_2 can be made from mixtures of these two colours. The same is only approximately true in CIELAB.

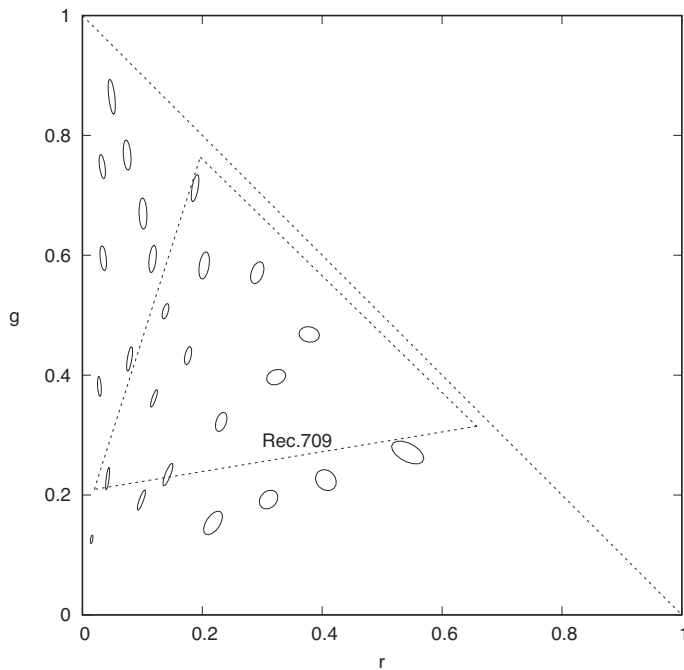
This tLab space is not recognised by any standards body. If you work in LMS rather than XYZ, it is similar to CIELAB, and the formula is simple.

Just Noticeable Differences (JND)

So far we have made colour spaces by dividing the range between white to black, or grey and red, into clearly visible steps with equal colour contrasts.

There is another approach. You can determine the Just Noticeable Difference between a colour and the surrounding colours. This takes a lot of peering at 2-degree Standard Observer circles in darkened rooms. When you plot the locus of colours that cannot be distinguished from the colour at the centre, they usually turn out to be ellipses. These are often called MacAdam ellipses after a 1942 paper by MacAdam.

MacAdam ellipses are usually plotted at some magnification, or they are too small to see clearly. Here are MacAdam's original 1942 ellipses plotted at 5x magnification in our **r,g** colour space.



MacAdam ellipses x5

I do not know the exact spectra that MacAdam used, so there is some possibility for small errors in my conversion to **r,g**. Any errors are probably long-ranged: they may move the ellipses a bit, but they are unlikely to change the shape of the ellipses much.

The ellipses cover more of the colour space than my Munsell Data. I do not have any points as close to the left-hand side or to the bottom. Even with my limited data I see none of the distortions we see in the MacAdam plots. I suspect this is yet another phenomenon that can be measured in a dark room, but can't be seen in a realistic image. I mention JNDs because there are a lot of papers on them, and you might find them in a search, but I do not think they are helpful for what we want to do with images.

Delta E (ΔE)

ΔE is an approximate measurement of colour difference. Each colour space may have its own definition of ΔE.

I am not going to teach calculus, but for the less mathematical, here is a quick summary of how the Greek letter delta is commonly used. Suppose you have a car on a straight road, and the position along the road in metres is **X** at time **T** in seconds...

$\frac{\Delta X}{\Delta T}$	The capital delta is a prefix for a measurable amount. Δ X may be 100 metres, and Δ T may be 10 seconds. The ratio gives us an average speed of 10 metres per second.
$\frac{\delta X}{\delta T}$	The lower case delta is a prefix for a small, hard-to-measure amount. δ X is still an interval of distance, and δ T is still an interval of time, but the ratio between the two is more significant. If δ X and δ T approach zero, their ratio should approach the speed at time T provided X and T are well-behaved.
$\frac{dX}{dT}$	The intervals dX and dT no longer have any separate meaning. The symbol as shown is the speed at time T . It is not a fraction in any meaningful sense. It is read as “d X by d T ”.

ΔE is a contrast we can see, so it should have a capital Delta. If you cannot get this in your text editor, most readers will probably recognise **DeltaE**. The ‘E’ stands for *Empfindung*, German for *sensation*. This can be traced back to Helmholtz and Hering.

The ΔE contrast between CIELAB $L_1^*a_1^*b_1^*$ and $L_2^*a_2^*b_2^*$ is...

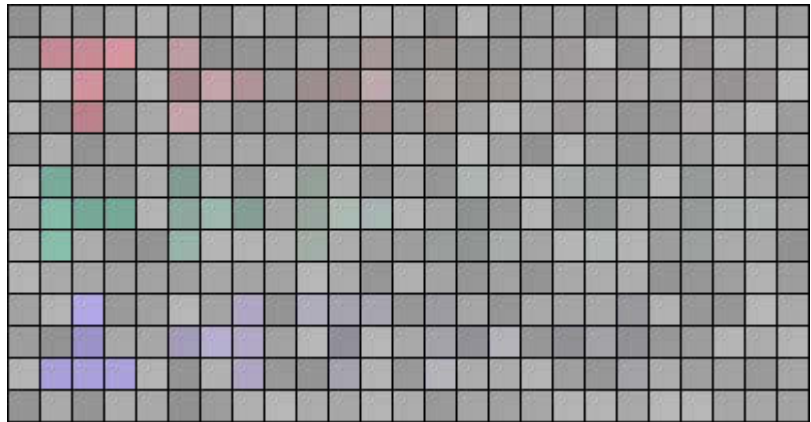
$$\Delta E_{ab}^* = \sqrt{(L_1^* - L_2^*)^2 + (a_1^* - a_2^*)^2 + (b_1^* - b_2^*)^2}$$

A Just Noticeable Difference is a ΔE^{*}_{ab} of about 2.3. In most other spaces, the JND ΔE is a number somewhere about 1.0.

We have a similar formula for tLab...

$$\Delta E_{tLab} = \sqrt{(L_1 - L_2)^2 + (a_1 - a_2)^2 + (b_1 - b_2)^2}$$

The following colour vision test chart has 6 columns of 'T' shapes in random orientations. If the chart is shown in Rec.709 video RGB the contrasts should correspond to ΔE^*_{ab} of 64, 32, 16, 8, 4, 2, and 1. It should be almost the same on a typical computer display.



Colour vision test chart

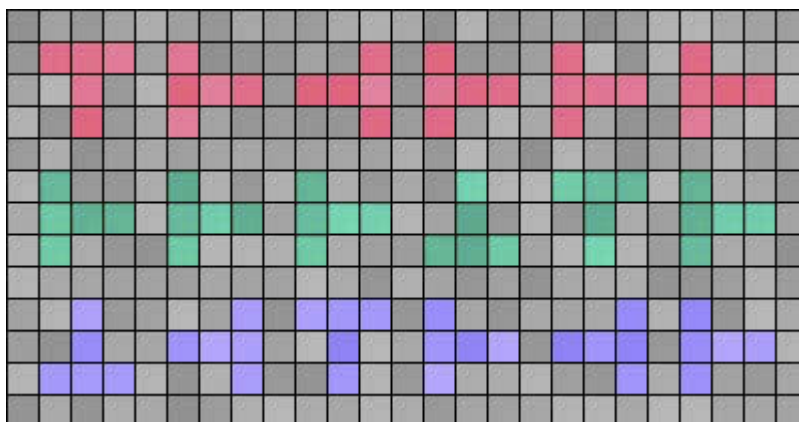
The three rows should give contrasts in the L, M and S channels. You may need to zoom the chart in, particularly for the bottom row as you will need your peripheral vision for the S-cones. The solution is on page 92.

I cannot see the last column clearly. If I did not know the shapes were there, I would not be aware of them. But knowing one of the four orientations are there, it is possible to pick the right one most times.

The 1976 ΔE^*_{ab} was not accurate for saturated colours. There were later modifications in 1994 and 2000 with more complicated formulas. We are good at seeing small deviations from neutral, but less good at seeing differences in saturated colours. We cannot uniformly arrange a set of colours with constant L^* on a flat surface: we want more area in the middle, and less around the edges.

I have not documented the latest ΔE formulas as there will probably be a new one along soon. You can use the simple formula here provided you know it is only accurate near the neutral axis, and you will see less contrast in the saturated colours.

Here's the solution for the colour vision test...



Colour vision test chart solution

Dominant wavelength

Here's one more idea. It often gets into the books, but I have never found it helpful. Suppose you have a colour (light green) and a white point (D65). The light green colour might lie in an x, y plot half-way between the D65 white point and the 520 nm monochromatic green on the spectral locus. We could mix our light green from D65 white and 520 nm green. 520 nm is called the *dominant wavelength* for our light green colour. If we chose a pink colour, then there is no dominant wavelength we can mix with D65 white, but if we extend our line though the white point, it meets the spectral locus at the *complementary wavelength*. This is the wavelength we need to mix with our colour to get D65 white. For a pink colour, our complementary wavelength might be 520 nm green. The dominant and complementary wavelength depend on our choice of white point.

This colour space is not uniform, and the entire red-violet slice is based either on complementary wavelengths or on mixtures of the limiting wavelengths we can hardly see, and probably have not been measured very well.

Uniform colour spaces

This chapter discussed the attempts to fit numbers to our sensations of brightness, hue, and saturation calculated from tristimulus. Tristimulus is based on good science, but the rest has been mostly guesswork: these sensations are in our heads and not something we can measure. Any attempt to make an accurate uniform colour space that matched our sensations is doomed from the start, as colours we see depend on the

surround as well as the colours themselves. Nevertheless, some results are good enough to be useful.

There are two very different ways to estimate colour contrast. The first way is to get people to look at easily visible contrasts and come up with a subjective match. This is easy for brightness, and not too hard for hue and chroma. The results are based on real comparisons, and seem familiar to artists. The second way is to determine the threshold at which very small differences become visible (JNDs). This is how the Macadam ellipses were determined. This is not subjective, but it is a lot harder to do: instead of sorting colour chips in a viewing booth, you are estimating tiny contrasts in a darkened lab under controlled conditions. The results are not familiar to artists. This may be because the detecting threshold is set by signal-to-noise limits very early in the visual chain. This is handy to investigate the retinal ganglia, but less useful for artistic judgements.

The useable spaces were...

Munsell	Based on real colours. No formula, though simulations exist. Depends on illuminant. Useful as benchmark for other spaces.
Yxy	Not uniform but has separate brightness and chroma.
Y $u'v'$	Y not uniform, $u'v'$ approximately so. Simple formula.
CIELAB	Based on old XYZ. Good uniformity for most whites. Widely used.
Yrg	Based on new LMS. Not standard. Y not uniform, r,g approximately so for D50 – D65.
tLab	Based on new LMS. Not standard. Good uniformity for most whites.

Chapter 5

Matching different media

I started working on colour correction on printing images in the 1980s. Photographers back then used reversal slide film, because this allowed them to expose a roll of film, develop it, put it on a light table, and pick the frames they wanted to scan. The print industry used drum scanners back then, so picking which frames you wanted before scanning saved a lot of work.

Slide film has a contrast range of at least 1000:1. You do not get much beyond 100:1 with printing. But the printed image had to look like the slide.

Matching contrast

We saw several tone curves in the last chapter. We saw a log scale give a nice, linear plot for the Munsell patches. This is reasonable. If we are looking at an image at 100% illumination, we would notice if the light level dropped to 90%, but our eyes would soon adapt. Once we had adapted, all the contrasts in the image would look much as before. Provided the absolute light levels do not change so much that things became too bright or too dim to see, the contrasts we see depend on the ratio of light levels, rather than the absolute values.

This makes sense when you consider how the eye works. Our rod cells have dyes that detect the light. Strong light bleaches the dyes, leaving the cells less sensitive. Our pupils expand and contract: this only gives a 5:1 control on light level, but it is faster than the adaption systems in the retina. We want to see objects without being distracted by changes in the light level, and the eye, pupil, and brain all help us do this.

Print and film users often refer to density. Density is given by...

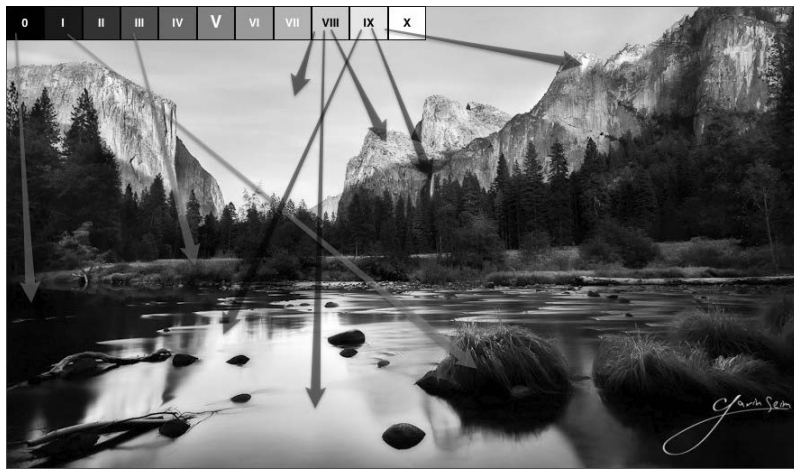
$$\text{density} = -\log_{10}\left(\frac{\text{luminance}}{\text{reference}}\right)$$

Our slide film has a typical density range of 0-3 with respect to the film base. This means the film base had a density of zero. Film that transmitted 10% of the film base had a density of 1.0; film that transmitted 1% had a density of 2.0; and so on. You can get slide film densities beyond 3.0, but it is hard to see detail in film that dark.

Our reflection print image has a typical density range of 0-2 with respect to paper white. The paper white has density zero. The densest black we can get still reflects about 1%. If we transform to density, multiply by 2/3, then transform back to luminance, then our print should look like the slide. It will not have the contrast of the slide, but we have reduced the contrast equally in the highlights and the shadows, which is better than just crushing all the shadow detail.

Ansel Adams zones

There is another way of looking at the same problem. The photographer Ansel Adams divided the tonal range into eleven 'zones'...



Ansel Adams Zones

His aim was to talk about image composition rather than to measure absolute grey levels. He described his zones as follows...

- 0** Complete lack of density in the negative image, other than film base plus fog. Total black in the print.
- I** First step above complete black in print. Slight tonality but no texture.
- II** First suggestion of texture. Deep tonalities, representing the darkest part of the image in which some detail is required.
- III** Average dark materials. Low values showing adequate texture.
- IV** Average dark foliage. Dark stone. Landscape shadow. Recommended shadow value for portraits in sunlight.
- V** Clear north sky (panchromatic rendering). Dark skin. Grey stone. Averaged weathered wood. Middle grey (18% reflectance)
- VI** Average Caucasian skin value. Light stone. Shadows in snow in sunlit snowscapes.

- VII** Very light grey skin. Light grey objects. Average snow with acute side lighting.
- VIII** Whites with textures and delicate values (not blank whites). Snow in full shade. Highlights on Caucasian skin.
- IX** Glaring white surfaces. Snow in flat sunlight. White without texture.
- X** The only subjects higher than zone IX would be light sources; they would be rendered as the maximum white value of the paper surface.

His descriptions are clearly suited to black and white photography and printing on paper. If your original was a black and white glass slide, you may have a contrast range of over 3.0. If you were printing the image in a book, your print range may be less than 2.0. In either case you would still want zone V to be near the middle of the density range, zone 0 to be the medium black and zone X to be the medium white.

But things are not that simple...

Surround, and the crispening effect

Here is an example of what is known as the 'crispening effect'...



Crispening effect

The squares are the same grey level on each row. The left square in each pair is a bit brighter than the right one. This is obvious on the top-right pair, the middle-left pair, and the bottom-middle pair where the grey of the squares is close to the grey of the background.

When we view the original scene, we probably see the mid-tones best. If we see our image in the viewfinder of a camera, the image is surrounded by darkness, so we see more shadow detail. If we project the image or view it on a monitor in a darkened room, we see more shadow detail. If we print the image surrounded by white paper, we lose shadow detail, and more highlight detail. If we want to reproduce the experience of the scene, we must correct the tone curve for the viewing surround.

This is not one of those effects you see with patches viewed in a darkened lab, but vanish when we view a realistic scene, and colour constancy works its magic. If we want to match the appearance of an image under different viewing conditions, we have to correct for this.

Bartleson and Breneman investigated the effects of brightness and surround on black and white image appearance for Kodak in the 1960s. They had a formula that could predict a match between the two viewing conditions if you specified the brightness level and the adaption factor (1.0 for a white surround, 0.5 for grey or more image, and 0.1 for dark). Truelight has similar parameters for the reference and display conditions. Truelight does not use the Bartleson and Breneman formula, but has something similar that fits their data and behaves a bit better with extreme highlights. The Truelight Colour Appearance Model (TCAM) uses the same model.

What about the corners of the image? Presumably the corners of the image have more surround than the middle. Should we have a different correction for the corners?

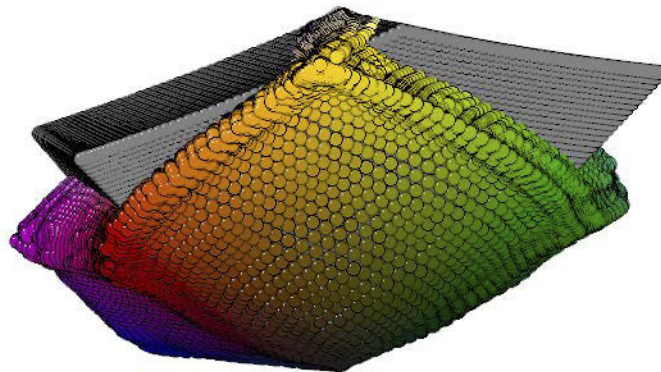
I always put the same correction on the whole image. The good reason is our vision can work with windows — we are happy to apply different rules to different regions in our view when the need arises. It is therefore reasonable to do a global correction for the image right up to the border. The less good reason is that it is hard to do a local correction without the risk of unpleasant side effects. Something that ought to be flat will now have a tone gradient towards the colour. Text and titles may look odd. There are all sorts of ways this might backfire on us.

Correcting for shadow textures

Truelight was the FilmLight product that simulated film processes. There is much less film these days, but it still works. You can take Kodak Cineon film density images scanned from a camera negative, and generate video. To prove it worked we often had to do a side-by-side comparison of film and video, displaying the video on a CRT or a digital projector.

Truelight has a contrast control. This allows us to compress the tone curve just as we did when matching film to paper print earlier in this chapter. It turned out we needed to reduce the contrast on the CRT by 5% to 10% to get a visual match to film, and a bit less to match the digital projector. If Truelight matched the light level on CRT and film, the shadows on the film looked lighter. The film had grain that you could see in the shadows. You could go up to the screen, stick out your hand into a dark region, and cast a shadow from the projector. The CRT gave the same light level but had no grain, and it cast no shadow; so CRT shadows looked black, and film shadows looked dark grey. With the contrast correction, the CRT and film looked the same.

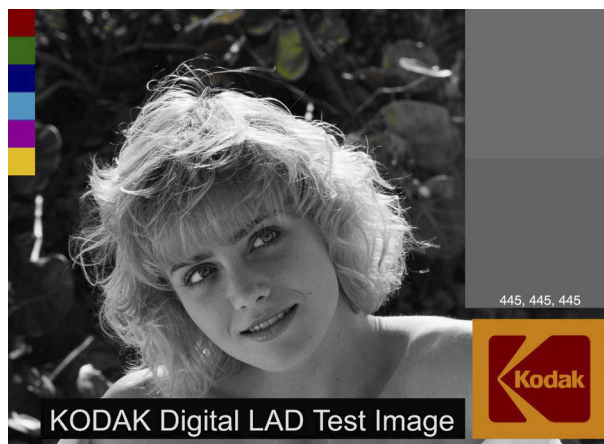
Colour gamut correction



Graphics monitor vs. Kodak Vision film colour gamuts

The diagram shows the colour gamut of a graphics monitor (grey spheres) and of Kodak Vision film (coloured spheres) plotted in CIELAB space. White is at the top, black at the bottom. Some colours, such as bright green, are possible on a monitor (grey corner on the right) but not on film. Other colours, such as deep reds and yellows, are possible on film but are impossible on monitors. Clearly, we cannot provide a match for every colour.

If our medium has not got the colour we want, there is not much we can do. We have a gamut alarm that highlights the colours we cannot achieve on our monitor. Here we can get almost all the colours for the Kodak 'Marcie' image on our video monitor, but the Kodak logo and the test patches cannot be matched.

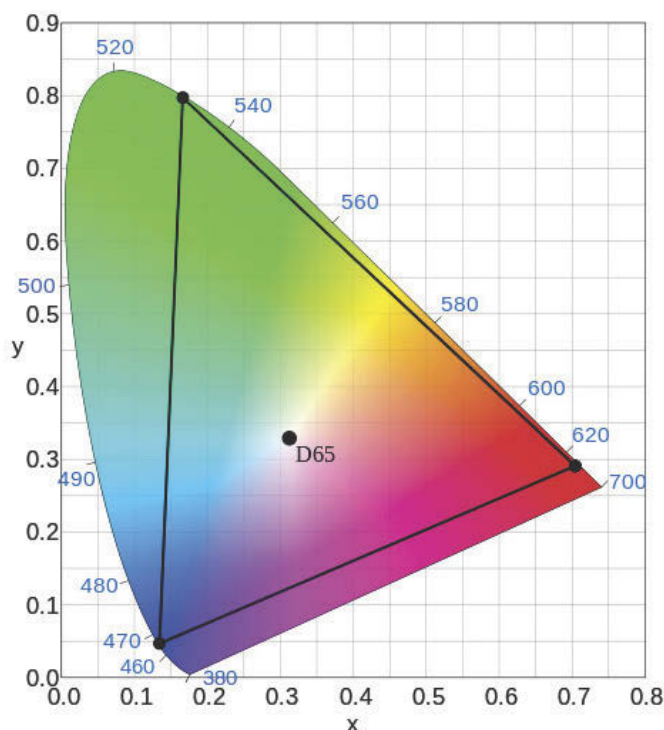


Kodak 'Marcie' image on video monitor

If we had to represent these out of gamut colours, we can reduce the saturation of the whole image. We can reduce the saturation by about 10%, and we may hardly notice. Then, quite suddenly, light flesh tones start looking 'dead'.

We can clip the colour values to the display gamut. The display has a limited range; it probably only accepts RGB values from 0 to 1. That gives a visible band when a value clips. Coloured highlights will show coloured bands as each channel clips in turn.

There are other colours that we may not be able to fix.



Rec.2020 gamut in xy

Here are the colours we can see plotted in the **xy** chroma space. The black triangle shows the Rec.2020 gamut. Rec.2020 is one of the largest three-colour display gamuts. However, if you join the D65 white point to the 500 nm azure colour, the Rec.2020 gamut cuts this line about half-way. There are a lot of azure colours that we cannot get in Rec.2020.

These azure colours are not very common. You find them in underwater scenes, and scenes with ice, and some winter scenes. No-one seems to be making new 4-channel displays just to get a better gamut in the azures.

Film and scene highlights

A traditional cinema film projector had an open-gate white of 50 nit. The brightest white you could get through film would be about 45 nit. Noon in Hollywood can be 90,000 lux. Something white would reflect the equivalent of 30,000 nit. The cinema light level is perhaps 1/600th of that. The light levels would match if the scene were shot 24 times as far from the sun, which would put it somewhere between the orbits of Uranus and Neptune.

HDR highlights

We compared one image medium with another image medium. Our slides may have more dynamic range than our prints, but both have a white point and a dark point.

If our reference is the original scene, our highlights may be a lot brighter...

Computer display:	250 nit
Moon:	2500 nit
White backlit cloud:	10,000 nit
Sun on horizon:	600,000 nit
Incandescent filament:	7,000,000 nit
Sun at noon:	1,600,000,000 nit

You can get daylight readable displays with a 2000 nit white. A few displays go to 20,000 nit. The present display industry has trouble making good, affordable displays over 800 nit with a reasonable lifetime.

Sparkling highlights are an important part of an image: they make fireworks look bright, gems sparkle, and water look wet. Our images can look a lot better for having some extra range for this sparkle. How much brightness do we need?

Standard colour viewing booths run at 2000 lux for the best colour resolution. White in a standard booth will reflect about 640 nit. Beyond this point, more light actually makes colour harder to see. Operating theatres are lit at 4000 lux, but this is to do a fiddly job in a confined space.

It is tiring to work with displays that go over 800 nit. It is particularly bad if you are working on images, and something pops up a white terminal panel at the full display white. High brightness also gives after-images. This might be part of the experience if you were looking at a bright

display, but it would cause problems in motion pictures, particularly if the next scene was at night.

There are no agreed industry figures yet, so here are some guesses. Mass-produced OLED displays can manage a white of 800 nit. They may have burn-in problems with a stationary logo at that brightness, but they should handle sparkle if it keeps moving. Let us assume we can get to 1000 nit because that is a nice, round number. A computer display typically runs at 250 nit with large white non-moving features such as windows. Can we run with 'paper white' at 250 nit and a peak white of 1000 nit, and still get the highlights we want? There is no clear-cut answer to this: we probably can, but we have to make some appropriate grading decisions...

Video zebra stripes

Video cameras have a set-up tool called 'zebra stripes'. These are typically set at 80%. This means you will get the stripes where any of the RGB signals go above 80% of the maximum. This would be about half the peak light level on a video display with a gamma of 2.6 ($0.8^{2.6} = 0.56$).



Zebra stripes on whites

Here is an image of a bronze bust with snow. The sky is above the 80% limit and has been replaced by stripes. The cameraman wanted to keep the snow detail, but not the sky, so they adjusted the camera controls so the stripes masked the sky but not the snow. When it is not snowing, the

levels can be set using a sheet of white paper in the ambient light. This means that white objects and light skin tones do not clip.

We can imagine cameramen doing just the same, but the camera actually recording a lot more in the highlights instead of just going up to twice the exposure and then clipping. The paper white set with the zebra stripes would be shown at 250 nit on our hypothetical display, and the remaining range to 800 nit range would be reserved for the highlights.

The Sun at the Zenith

Here is a picture from the 1850s by Gustave Le Grey, called 'The Sun at the Zenith'. The sky often appears blank in old photographs, because it is much brighter than the subject. This image is in fact two photographs. The image of the sea would probably have had a completely blank sky. The sun is darker than the surrounding clouds from a photographic effect called solarization. The join lies somewhere above the horizon.



Gustave Le Grey, 'The Sun at the Zenith'

Over 170 years later we are still trying to get the clouds in the sky and the sparkle on the sea.

Painting Highlights: Constable & Monet

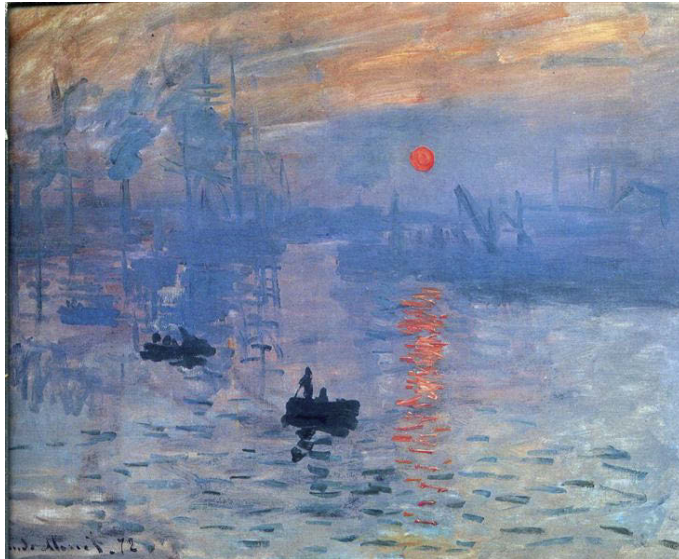
Le Grey was making an HDR image limited by the photographic processes of his day. Painters can do the same thing, but without the limits of collodion. Here are two very different examples...



Constable, 'The Hay Wain'

John Constable's 'The Hay Wain' has bright clouds and sky. The bright sky and the white wall on the mill are almost the same shade, so he is varying the exposure within the image. Constable has reduced the contrast in the white clouds. They are not the weighty blobs we see in some Le Grey photos, but a backdrop to the picture.

Monet's 'The Pool of London at Sunrise' has a different strategy. Monet preserves the colour of the rising sun, which means it is less bright than the clouds above it. It is almost exactly the same luminance as the blue mist that surrounds it. It is this pure colour contrast that makes the edge of the sun so striking.



Monet, 'The Pool of London at Sunrise'

This is how the sun might appear in a motion picture scene. The main actors are in the foreground. The sun is an orange disk in the background. It tells you it is close to sunset or sunrise, but does not intrude on the scene. You should not gaze straight at the sun, but it would do you no harm. If it had a realistic brightness of half a million nits, it would dazzle you, and you would see little else.

The Empire of Light

There is no general algorithm for fixing on a white point. The *grey world assumption* takes the white point from the average light of the whole scene. This can give useful results, but suppose we have a scene with a lot of colour patches: we can shift the appearance of any patch by shuffling patches until it is surrounded by a particular colour, even though the scene average remains the same. We can bias our average towards light colours, which may pick out light sources and white reflectors. We can reduce the weight of very large features such as the sky. We can calculate local sums, which would work for scenes with different light sources. Professor Land's Retinex model averages random paths to any particular point to estimate the local white.

Here is Magritte's painting 'The Empire of Light'. This is a subject he returned to a dozen times or more. What is the white point here?



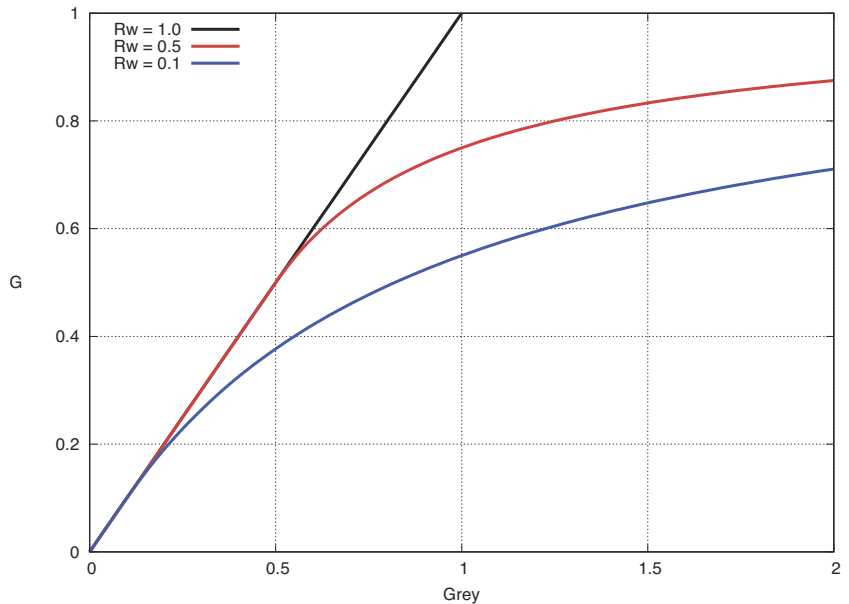
Magritte, 'The Empire of Light'

There is no obvious subject. The clouds in the sky are a likely candidate. They are apparently lit by the sun, though it has set on the foreground. They would probably have a pink tinge as the sun is about to set, but they look white. The street lamp casts a distinctive shadow on the white building with a slight lime-light green. And there are the lit upper windows, with a familiar tungsten yellow. Any of them? All of them? None of them?

A white point is something we use in our colour appearance models. It may exist in our heads the way colour does. Sometimes a single, dominant light sets the white point, but it is not always a real thing you can point to.

Highlight compression in TCAM

Suppose we chose a reference white R_w . The Grey value on the horizontal axis is some perceptually uniform space for the input signal. It may be video RGB or L^* . It is not linear luminance Y . The G value on the vertical axis is the compressed grey value.



Highlight Compression Functions

The black line does what video cameras do — it goes up to 1.0 and then clips.

The blue line has a reference white R_w value of 0.1. All the tones above 0.1 have changed. This is probably most of the image. The blue line goes slowly towards the 1.0 value but no matter how big the *Grey* value is (and it could go well beyond 2.0) the G value never gets to 1.0.

The red line is a compromise. All the tones above 0.5 have changed, but the change comes in gradually. Here is the tone curve applied to a section of an HDR image...



Rw = 1.0



Rw = 0.5



Rw = 0.1



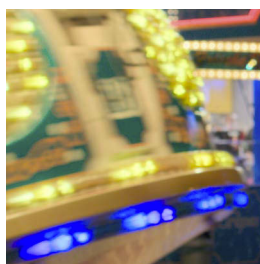
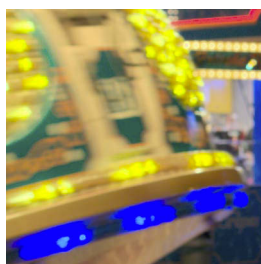
ACES RRT 1.0

Rw=1.0 gives a hard clip: the highlight on the apples has a hard edge. Rw=0.5 looks similar, but the highlights are now soft. Rw=0.1 has gone too far, and the whole image looks dull. We have the ACES RRT for reference. This looks a bit duller than our Rw=0.5 image. This is not a fair comparison as the ACES RRT is fixed for all images, while we can tune Rw. However, it does show that we are doing broadly the same thing as ACES.

This tone curve is part of the Truelight Colour Appearance Model (TCAM).

Highlight bleach in TCAM

TCAM has another control called Highlight Bleach. Here are three renditions of the same cluster of blue LED lights with different settings of Highlight Bleach...



Highlight Bleach

On the left we have kept all saturation there was and reduced the brightness to keep the colours in gamut. This is what Monet did with the rising sun. The change of colour in the centre was probably because the LEDs were too much for the camera. On the right we have reduced the saturation and increased the brightness by reducing the contrast at the

top of the tone curve. This is what Constable did with the clouds. In the middle we have the default TCAM setting halfway between the two.

Supposing we use the $R_w=0.5$ tone curve. If our display white was 1000 nit, our paper white level was 50% and our display gamma was 2.6, our paper white level would be 165 nit. If we said our paper white level was actually 60% and not R_w because the tone curve has hardly bent at all at 60%, our paper white level would be 265 nit.

Where did the rest of the highlight range go? Why do we not miss it? Small, sparkling highlights do not stay in one place on the retina. They bleach any cone cells they hit. This means they probably look whiter than they are. They look bright, but they are so much brighter than their surround that we find it hard to tell exactly how bright they are. We can show them with less brightness and less saturation, and still get the same sense of sparkle, but without the annoying after-images.

Suppose our highlights are not small and sparkling? There are two problems here. They will be dazzling and uncomfortable to look at, and they may burn themselves into the monitor. Some HDR monitors have safety settings to stop bright features harming the monitor by reducing the dynamic range if the brightest features are too large or too stationary. Unfortunately there are no agreed standards on this, and every display handles this differently.

Making HDR from video

This is a dangerous idea unless you know your video has been made from some source with a wider dynamic range. But it can work...



HDR from video (left) and original (right)

Here is a butterfly presentation of two images. On the right we have the original video image reduced to 30% brightness. On the left we have a similar image but stretched using TCAM, so we have the same colours up to the reference white but with 100% highlights.

If the HDR display could do the highlight stretch, we could drive it with an ordinary video signal. This might be a practical way of adopting HDR without a separate broadcast standard.

Compressing the colour gamut

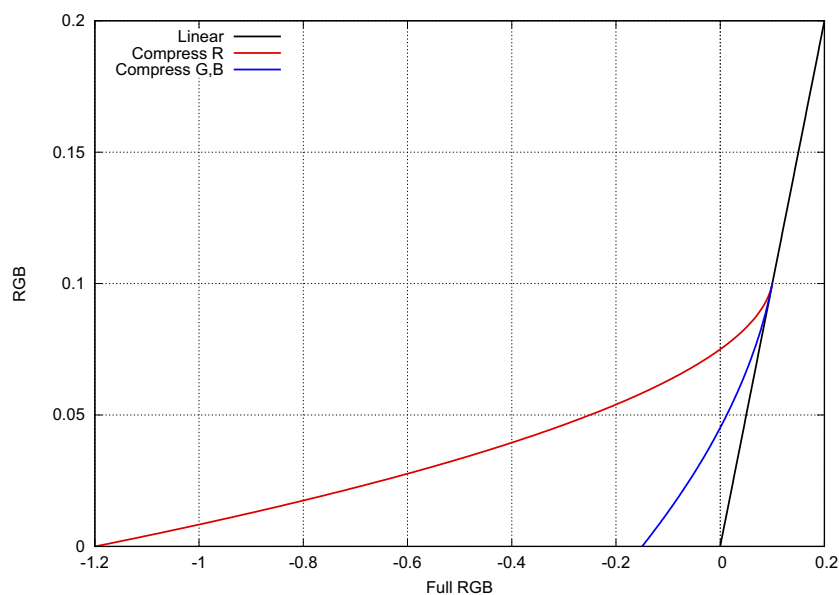
If we can drive ordinary and HDR displays with the same signal, we might be able to do the same for ordinary displays and those with a high colour gamut.

The video colour gamut triangle has corners on the RGB primaries. Outside this triangle, we have negative RGB values. These negative values have no meaning for a video monitor but might correspond to real colours on a high gamut display.

Our RGB colour gamut compression might work like this...

- 1 Scale the RGB values so the largest value is 1.0. If this is a real colour, then at least one of RGB will be positive. If all values are zero, we don't have to do anything.
- 2 Use a smooth function to map the values below 0.1 into the range 0.0 to 0.1. 0.1 has 0.0025 of the linear brightness of 1.0. Values below 0.1 will be hardly visible against the largest value of 1.0, so we can re-map values below 0.1 without changing the look of the image.
- 3 Undo the scaling of step 1.

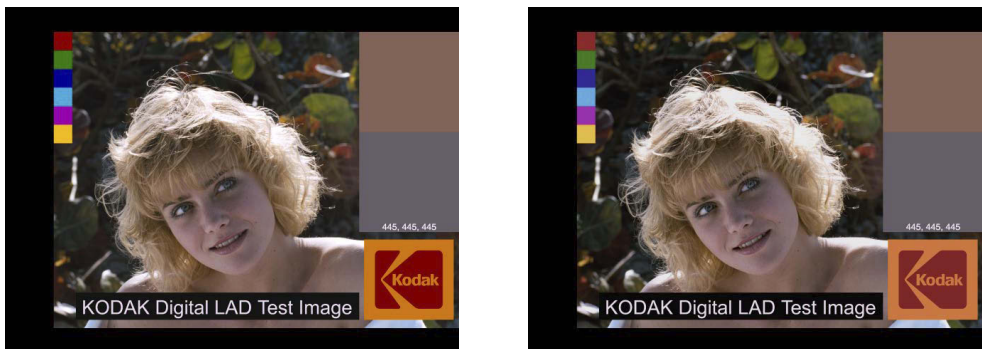
This compression does not affect normal RGB reflection colours. It only changes the very saturated colours. We are not good at seeing shades outside the reflection gamut, so we can afford to compress these colours.



RGB Gamut Compression Functions

The graph shows a set of the compression functions that map all real colours into the 0-1 gamut. The black **Linear** line shows no compression. It is not at 45 degrees because the X and Y axes have different scales. The red **R Compress** curve maps -1.2 onto zero. The blue **G,B Compress** curve maps -0.15 onto zero. At 0.1 the two curves smoothly join the straight line for no compression.

A video display would not decompress the RGB values. A high gamut display could decompress the RGB values, and the decompressed RGB might be real colours when converted to the new primaries.



Compressed and uncompressed RGB

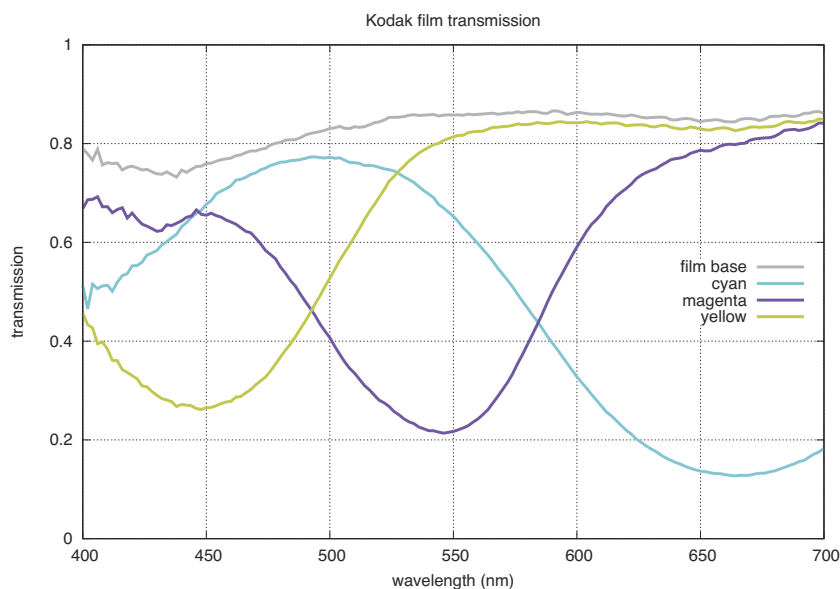
The left-hand image has compressed gamut; and the right-hand image is uncompressed. The colour patches and the Kodak logo are less saturated in the left-hand image, which leaves a bit of room to code all the negative colours.

Under Colour Removal (UCR)

This is a CMYK printing process, but it may be worth discussing because it is not always done the best way. Motion pictures are generally handled as 3-colour RGB spaces. The Cineon standard was CMY negative density, but it behaved a lot like an RGB space. The only exception I can think of (and it is pretty obscure) is Technicolor process 4. This had a grey separation from the monochrome film base that was used to carry the soundtrack, but also carried a grey image in addition to the CMY dye transfers. This made the shadows deeper, but was clear in the light colours. This is (just about) a UCR technique.

What is under colour removal? Suppose you have done an RGB scan of your image. You have density measurements for the original to RGB light. If you want to get the same colour on your print, you want to lay down the corresponding density in CMY inks. Yes, it isn't quite as simple as that and there are all sorts of colour corrections in real practice.

Here are some real dyes measured from cine film...



These dye patches are not particularly dark. The density is about 0.5. They were actual spectral measurements I used for doing Truelight projector lamp calibrations. We can see the spectrum is noisy for the short wavelengths: I was using an incandescent lamp because they are very stable, and they do not have much UV.

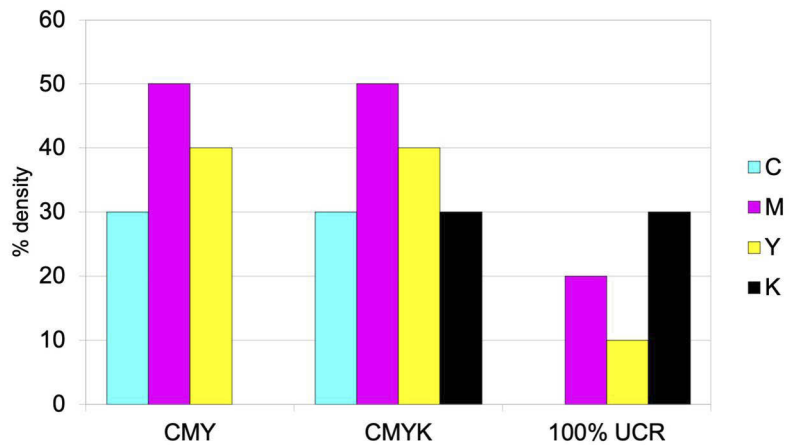
The C and M curves cross at about 585 nm. The Y and M curves cross at about 490 nm. A rule of thumb for CMY inks or dyes is to get these two crossing points at the same density (height) when the CMY overprint is neutral. If we had a much higher density (and the film density could go over 3) the film transmission will have peaks somewhere near these two wavelengths. If the two peaks have a similar transmission, the shadow tones should stay neutral. This deviation from the simple model where we add CMY densities is called *additivity failure*.

Let us consider printing on paper using CMYK inks...

In printing, black has the letter K for 'Key'. A keyline is a black border put around colour images to hide any registration errors for the CMY halftone patterns.

It is generally good to replace overprints of C+M+Y with black. Black ink is cheaper than the CMY overprint, and the printed product will take less heat to dry, which can be the main use of power in a web press. It also makes sure the dark tones are neutral.

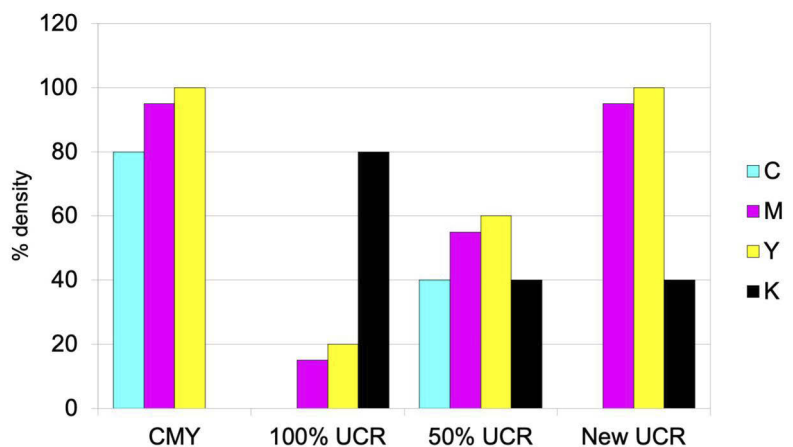
Here we have a reddish-grey colour that we want to print in CMYK...



The vertical axis is “% density”. Let us assume this is scaled so an equal amount of CMY is neutral.

In the first column we have the original CMY that may have come from the scanner. All three inks have a common 30% density. In the second column we have added a 30% level to the black (K) channel. This will make the shadows denser. In the last column we have subtracted the 30% black from the CMY ink levels.

However, this can reduce the colour gamut. Consider printing a dark red...

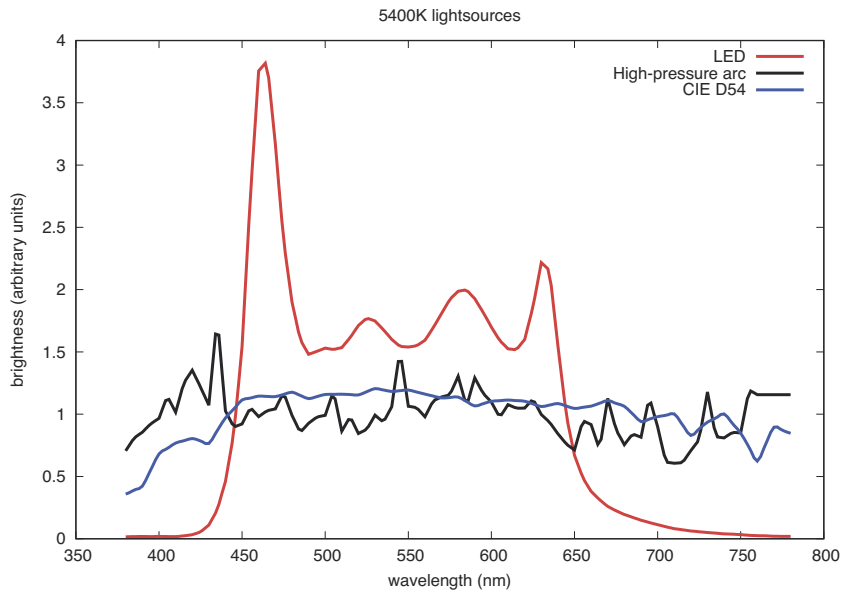


100% UCR greatly reduces the MY ink levels, so the deep red has lost the saturation. 50% UCR is not as bad, but we have still lost saturation. There are smarter solutions where we keep the colour saturated and add black.

Metamerism

We can have two different spectra that look the same colour. This is called *metamerism*. *Metameric failure* is the failure of two colours to match under different conditions. There are several causes of metameric failure: a change in the light source, a change in the reflective or transmissive colour, and a change in the observer (or camera).

Here are three illuminant spectra with almost the same visual white...



All three light sources would look much the same colour to us. A neutral scene lit with any of these would look much the same too. But if our scene had a deep red object that only reflected wavelengths above 650 nm, it would look bright red under the arc lamp and the D54 daylight simulation, but almost black under the LED light source. In Jacques Tati's film 'Playtime' there is a scene in the 'Le Drugstore' restaurant where a tempting red dessert under the counter dramatically changes when placed on the counter due to the different lighting. This is illuminant metameric failure.

I am not saying LED illuminants are bad: I just needed a bad one as an example. Even then, it is not necessarily bad if you wanted to light a room, but some colours may look odd. We are used to seeing colours by daylight or by light sources with a blackbody spectrum, and the CIE standard daylight spectra and the blackbody spectra are quite similar. There is a measure of how close any given light source is to the usual standards. This is called the Colour Rendering Index or CRI, and we will come to it shortly.

You get metamerism from reflection spectra. This is less common; light sources can have sharp peaks in the spectrum, but it is rarer to get sharp peaks in reflection spectra.

Ordinary CMYK printing inks can show metameric failure under different lights with good spectra. Most black inks have a very flat spectrum, so they should look neutral under all lights. I have a printed example, half CMY, half K that matches under office lighting, but the CMY half looks clearly green next to the K half under daylight. Pantone make deliberately metameric stickers where the two halves should match under D65 or D50, so you can check your artwork is viewed under the correct light.

There is also observer metamerism. Human observers can have different colour vision. This is also often used to mean the difference between a camera and a standard observer.

Metamerism can be looked for with more than 3 visual primaries. We can make colours that match for the LMS cones but not match for the rods, to see whether the rods have any effect in daylight (they don't). Or we can make colours that will match for the 2-degree observer but not match for the peripheral vision. This is the experiment with the Sony projector mentioned in Chapter 3 in the section on the 10-degree observer: filtering out the azure and yellow wavelengths shows up the Maxwell spot nicely. People have also tried to stimulate the intrinsically photosensitive retinal ganglion cells while keeping the LMS constant. This is hard to do because the iPRGC cells have a peak between the S-cones and the rods, so there is a fair bit of overlap. But people have tried.

Colour Rendering Index

The CIE's 1995 CRI or *Colour Rendering Index* is an often quoted figure for light sources. You have a test light source and a reference. The reference is typically a black-body spectrum source for sources under 5000K, and daylight matching source such as D65 for 5000K and above. You could have other reference light sources, but your CRI figures would be non-standard.

First, the source XYZ is shifted using a von Kries transform to match the reference (see next chapter — this is effectively just a scaling in LMS). The reference ought to be a similar colour, but the CRI quantifies whether colours look the same under the two lights, regardless of whether they match.

The Test Color Sample (TCS) reflectances are a set of 14 patches taken from an old Munsell Colour Atlas. The digital versions of the reflectances are all you need for the calculation...

Name	Nearest Munsell	Appearance	Swatch
TCS01	7.5R 6/4	Pastel red	
TCS02	5Y 6/4	Pastel orange	
TCS03	5GY 6/8	Pastel yellow-green	
TCS04	2.5G 6/6	Pastel green	
TCS05	10BG 6/5	Pastel azure	
TCS06	5PB 6/8	Pastel blue	
TCS07	2.5P 6/8	Pastel violet	
TCS08	10P 6/8	Pastel magenta	
TCS09	4.5R 4/13	Strong red	
TCS10	5Y 8/10	Strong yellow	
TCS11	4.5G 5/8	Strong green	
TCS12	3PB 3/11	Strong blue	
TCS13	5YR 8/4	Light cream	
TCS14	5GY 4/4	Olive	

The CIE 1931 XYZ standard observer is used for all the calculations. They calculate the ΔE_{UVW} for these colours and subtract the average from 100 to give the CRI.

There are a number of criticisms of this standard. The CIE UVW space is not uniform and not used except for CRI calculations. There is a jump in reference at 5000K. There are proposals to use different colour patches. Using the worst case might be more representative than using the average. However, this is what people currently use.

A good CRI value is 90 or above. The very good light sources are 95 or above. The last few percent may depend on the reference spectrum. For example a 6500K blackbody spectrum is smooth and would make a good light source, but it has a CRI of 98 because it is referred to D65, which is a lumpy daylight simulation.

Appendix B contains some tools based on the CIE 2006 LMS and tRGB colour spaces. This includes a tCRI alternative. This is not standard, but it may be useful.

Chapter 6

General colour transform tools

We can transform from one standard colour space to another using mathematical functions. If we are in DCI XYZ and we want to get to sRGB, we can get from DCI XYZ to linear XYZ with the DCI white point using the inverse DCI tone curve. We can transform from XYZ to Rec.709 RGB using a matrix. We can then get to sRGB using the sRGB tone curve. This is probably the fastest and the most accurate way to transform floating-point image data.

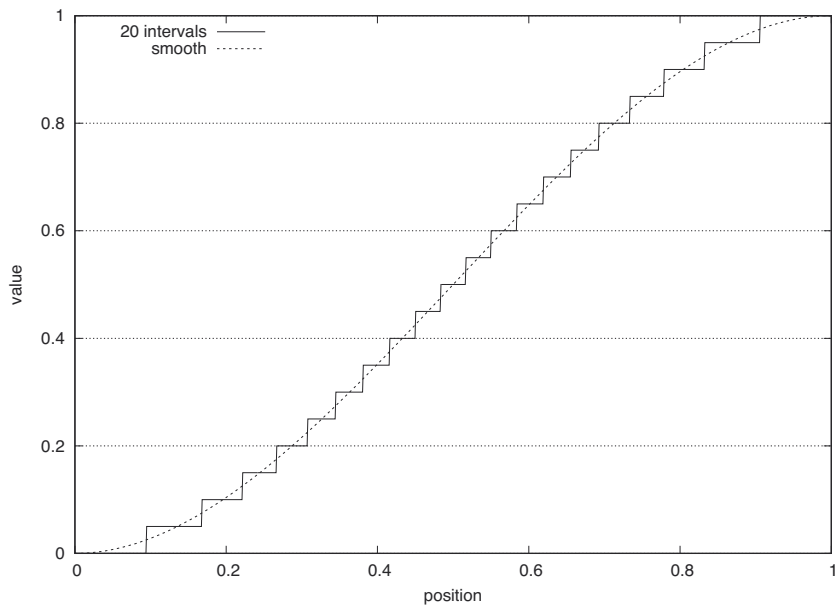
Sometimes we need to do something that cannot easily be put into a formula. We want to make a colour transform that is based on experimental data. Or we want to apply the colour transform as demonstrated by 'before' and 'after'. Or we want to apply a colour transform that is made from many hand edits.

The rest of this chapter describes the tools we use to do this, and a few of the things we can do with these tools.

Quantization

Before we talk about changing image values, we should discuss how digital images store grey levels.

A monochrome digital image is a 2-dimensional grid of pixel values. A colour image may have separate RGB values for each pixel. Newer image formats support floating point values. Older image formats use integers. Here is the difference...



I have plotted out an 'S'-shaped function. The dashed curve is the smooth floating-point function, while the step function is the same function rounded to the nearest level with 20 intervals. Yes, 20 intervals is too few, but I wanted the steps to be visible.

The same function as an image looks like this...

20 intervals:



50 intervals:



100 intervals:



200 intervals:



Smooth:



We can just see some steps with 200 intervals. With real images and the right colour space, 8-bit data (255 intervals) is enough to display a smooth-looking image, but will not be enough if you want to digitally edit the images. Legalized SD video is about 220 levels. Real images also have noise that breaks up and randomises the hard edges we see here. However, if we re-graded the image, or just smoothed or de-noised our image, the quantization can re-appear as coloured bands.

1D LUT

The 1D LUT or Look-Up Table is the simplest transform. We rarely use 1D transforms by themselves, though they used to be an important part of display calibrations. They can be a useful part of a 3D transform, but for now we will describe them as a stand-alone operation.

Suppose your image values go from 0.0 to 1.0. Suppose also we have an 8-bit table of 256 values. The first value `lut[0]` is what 0.0 becomes after transforming. The last value `lut[255]` is what 1.0 value becomes after transforming.

If we want to transform 0.25, the index will be $0.25 \times 255 = 63.75$.

The table only has values at integer values, so we have to interpolate using the integer and fractional part of the index.

The 1D LUT transform of 0.25 is $(1.0 - 0.75) \times \text{lut}[63] + 0.75 \times \text{lut}[64]$.

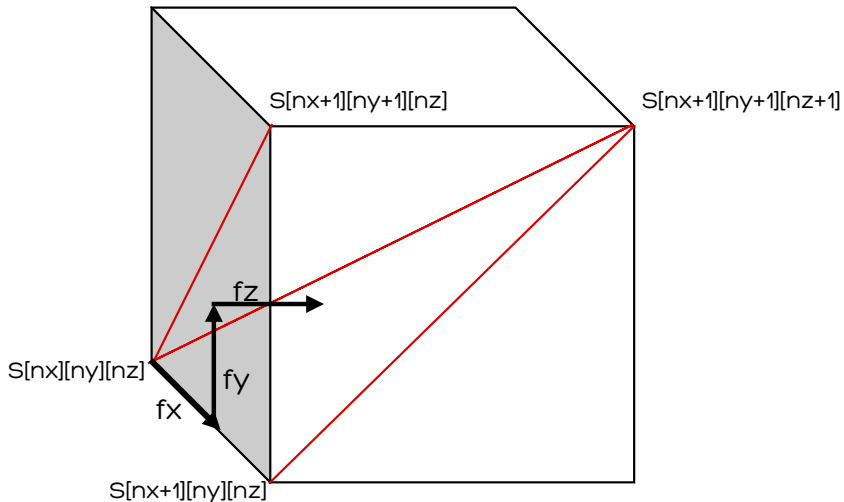
A slightly cheaper calculation would be to return `lut[63]` and to carry the 0.75 over to the next input value. This is called error propagation. It is less common these days because it is not a good fit to the way GPUs work.

An even cheaper calculation would be to return `lut[63]` and forget the remainder. If your LUT is big enough, the errors may be acceptable.

If we wanted to apply the LUT to values ranged from (say) -0.5 to +1.5, we would scale and offset the values so -0.5 and 1.5 became 0.0 and 1.0, then use a 1D LUT as before.

3D LUT

Suppose we have a three-dimensional array of values. Instead of having 256 values and 255 intervals, suppose we have 16 values and 15 intervals per dimension. The point we wish to interpolate might lie in a sub-cube like this...



Here the three indexes have integer values nx, ny, nz , and fractional values fx, fy, fz .

The S values on the corners are the output values we are going to use to interpolate. S is usually a set of three values such as RGB or XYZ, but you can interpolate CMYK or any number of output values if you want.

There are two ways to interpolate...

The best one, particularly for smooth interpolation along the diagonal, is called *tetrahedral interpolation*. For this particular case where $fx > fy > fz$ we have outlined the tetrahedron in red. It has one side parallel to the X-axis, one side parallel to the Y-axis, and one side parallel to the Z-axis. We can interpolate in these three directions like this...

$$\begin{aligned}
 S_{xyz} &= (1-fx) * S[nx][ny][nz] \\
 &+ (fx-fy) * S[nx+1][ny][nz] \\
 &+ (fy-fz) * S[nx+1][ny+1][nz] \\
 &+ (fz) * S[nx+1][ny+1][nz+1]
 \end{aligned}$$

There are six possible rankings for fx, fy, fz , so there are five other cases with tetrahedra that also must be supported.

The other method is called trilinear interpolation. It uses all the corners of the square like this...

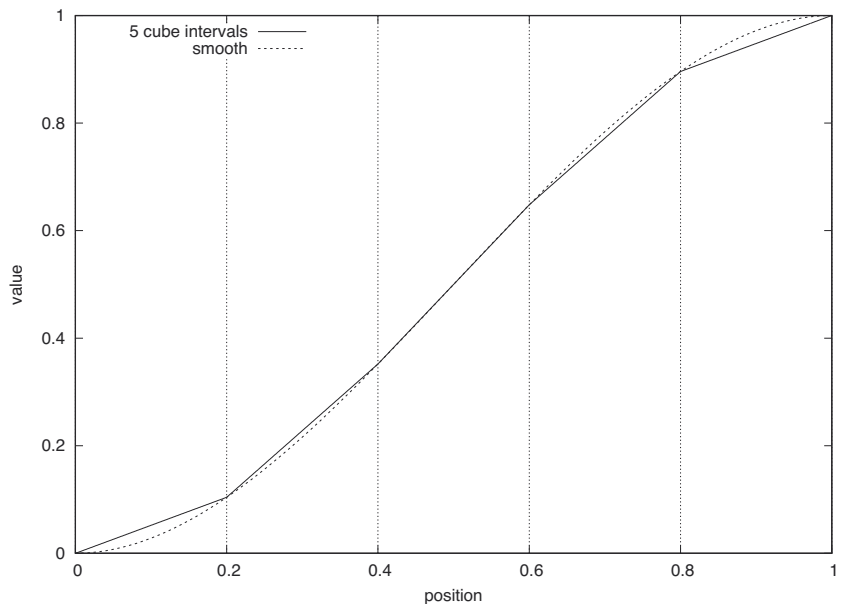
$$\begin{aligned}
S_{xyz} = & (1-f_x) * (1-f_y) * (1-f_z) * S[nx][ny][nz] \\
& + (f_x) * (1-f_y) * (1-f_z) * S[nx+1][ny][nz] \\
& + (1-f_x) * (f_y) * (1-f_z) * S[nx][ny+1][nz] \\
& + (1-f_x) * (1-f_y) * (1-f_z) * S[nx+1][ny+1][nz] \\
& + (1-f_x) * (1-f_y) * (f_z) * S[nx][ny][nz+1] \\
& + (f_x) * (1-f_y) * (f_z) * S[nx+1][ny][nz+1] \\
& + (1-f_x) * (f_y) * (f_z) * S[nx][ny+1][nz+1] \\
& + (1-f_x) * (1-f_y) * (f_z) * S[nx+1][ny+1][nz+1]
\end{aligned}$$

There are no other cases to handle with trilinear interpolation. This is faster to calculate in a GPU, but it is not linear along the diagonal, so you can get coloured bands where you might expect pure grey levels.

3D LUTs work well when you are going from one perceptually uniform space to another. They are less good if the input or output space is linear.

The right number of intervals is important. We could match the image data precision as we might do with 1D LUTs, but the cube files would be very large as the number of entries goes up as the cube of the number of intervals.

Here is the same function we plotted out at the beginning of the chapter, but applied with a cube with 5 intervals. We can see the intervals fit well over the straight part of the curve but do not fit when it bends at either end...



**5 cube
intervals:**



**15 cube
intervals:**



Smooth:



The curve from the cube with 5 intervals actually looks quite smooth as an image, but we can see that the greys at the light end are too dense. If we had colours, we would probably see coloured fringes.

16×16×16 (15 cube intervals) is usually as small as cubes go. You are interpolating over 15 intervals, so you may see bands where the gradient changes. Image noise usually breaks this up, but very smooth images and artificial features may suffer.

You get better results with more intervals, but the amount of data goes up with the cube of the number of values per dimension. 3D LUTs with more than 64 values per dimension are rare.

If the neutral tone curve is not smooth, it can be helpful to add a 1D shaper LUT before the 3D LUT. If you are making a 3D LUT from a list of input and output values, the Truelight cube generator can do this for you. If we want to cover a larger range than 0-1 we can have scaling and offset on the 1D shaper LUT. Not all cube formats support 1D shaper LUTs.

The Truelight cube format supports 1D output shaper LUTs. You might use this if you were making a transform that output linear values. It is not often used, but it can be handy if you want to invert a cube.

The Truelight cube format supports 4D outputs. You can use multiple cubes if you want even more outputs. This may be useful if you have a display with more than four primaries.

The Truelight cube format also supports 4D inputs. This is much less useful, and probably not supported on any GPU code. This works in the same way as the 3D pyramid interpolation, but with 4 dimensions, so there are $4 \times 3 \times 2 = 24$ different possible rankings to code. It is also hard to get enough experimental data to calibrate a whole 4D space. I have done it once for CMYK printing; it was a lot of work and the result was not very useful. If you can reduce your 4D to a 3D RGB-like format, this is probably the thing to do.

Triangles

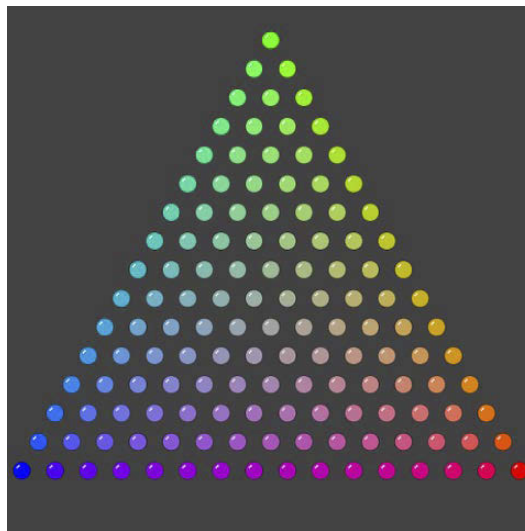
The 1D LUT and the 3D LUT have a finite range and work best on perceptually equal spaces. The Triangle transform has no limit on the luminance range, and works best with linear data. The triangle transform will transform all values with the same chroma equally. This is useful for simulating changes in illuminants, and things like that. It works well with tRGB because the tRGB triangle encloses all the possible colours, and only has a small fraction of unreal colours.

Let us start with three RGB values. They should all be positive or zero. The Triangle transform works like this...

We have a scale $S = (R+G+B)$.

We calculate normalised rgb where $r = R/S$, $g = G/S$, $b = B/S$. We have to handle the special case where S is zero as we cannot divide by zero, but this is easy as we know the output is zero anyway.

We interpolate from a triangular array of input points with the rgb primaries as the corners. Each point has a corresponding set of output values...



The target rgb point should be somewhere in this big triangle, in a small sub-triangle with the nearest three points as corners. We have output values for each of these corner points. We do a weighted average of these three sets of output values to get our target output values. The weights of each point go from 1.0 if rgb lies on the point, to 0.0 if it lies on the opposite side. Let us call the output values x , y , and z .

We then rescale to get the output $X = x.S$, $Y = y.S$, $Z = z.S$.

Capturing LUTs

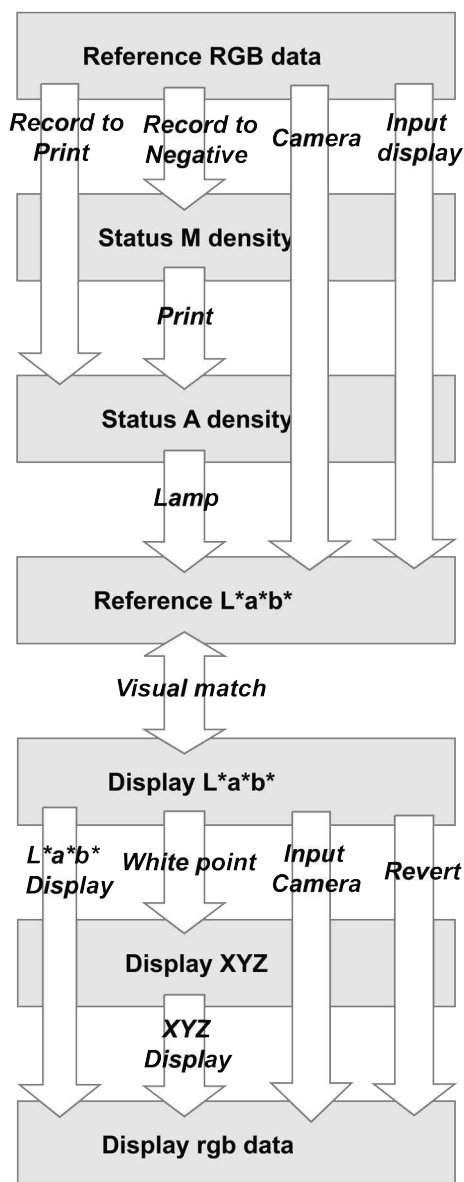
We can generate LUTs from a stack of colour operations. We make a test chart image, based on the 1:1 values in a 1D or 3D LUT, or even a triangle. We apply the colour operations to the image, but skip any spatial operations. We can then read the values back from the patches on the processed chart image.

The chart does not have to be a whole image. A $16 \times 16 \times 16$ cube will have 4096 values, which will be a single line on a larger image.

If the test chart was based on a 1D or 3D LUT, or a triangle, we can just reformat the values to make the transform. In the general case, we get a set of 'before' and 'after' values from the original and the processed chart, and use this to interpolate the transform.

List transforms

The Truelight film simulation had many possible colour spaces and transforms. Each of the arrows in the diagram overleaf was a possible transform, and each of the grey rectangles was a possible space. Some of these colour spaces had additional colour operations within them. We needed a common data structure that would let us combine transforms, invert transforms, and combine transform with formulas.



The Truelight list transform is a list. Each line in the list contains a set of input values and a set of output values. Typically, the transform used 3D inputs and 3D outputs, though it also supports between 1 and 4 dimensions, and also names.

The Truelight library had a routine for interpolating through lists. It does not need the input points to lie on a cubic or triangular grid.

We can make LUTs, cubes and triangles from list transforms, and make list transforms from LUTs cubes and triangles. The list transform is too slow to apply directly to image data in the GPU, but we can make a cube or triangle in the CPU and apply that.

Zview viewer

There is a Truelight tool called 'zview'. This is used to visualise list transforms in 3D. It can display both the input and output values, one by XYZ position and the other by RGB colour. We have used this to generate the diagrams with little coloured spheres in this document. It can also be used to examine lists, delete faulty values, and repair faulty values by interpolating from the remaining good values. It can also display cubes and triangles, as both of these can be made into lists.

'zview' also has tools to let you edit the transform by dragging points on the screen, and preview the look on an image. This was an experiment to see whether people can visualise transforms and edit them in this way. They can, but it is hard to do: I do not expect the industry to go this way.

Spectra

There is a Truelight spectrum format. This is pretty simple: it has two entries per line, a wavelength in nm, and a value. The wavelength values are in order from low to high. The wavelength values usually have a uniform step. The reader handles non-uniform steps. This is handy because some spectrometers output the values from the sites in their detector array with a calibration of their wavelengths rather than cleaning up the data. It will not necessarily make a good job of spectra with missing values or wildly varying intervals.

The spectrum reader in the Truelight library can ignore header lines and lines that do not fit the wavelength-value format. It can accept values in reverse order. If you can dump the readings from your spectrometer in an ASCII format, the Truelight tools may be able to read it without editing.

There is another tool called spectrumCalc that performs a number of the standard operations on spectra, and can output sets of values such as XYZ or LMS or CIELAB generated from transforms as a list transform. These tools are not that exciting. But with the smooth spectra from the next section, they let us simulate changes in illuminant.

Smooth spectra

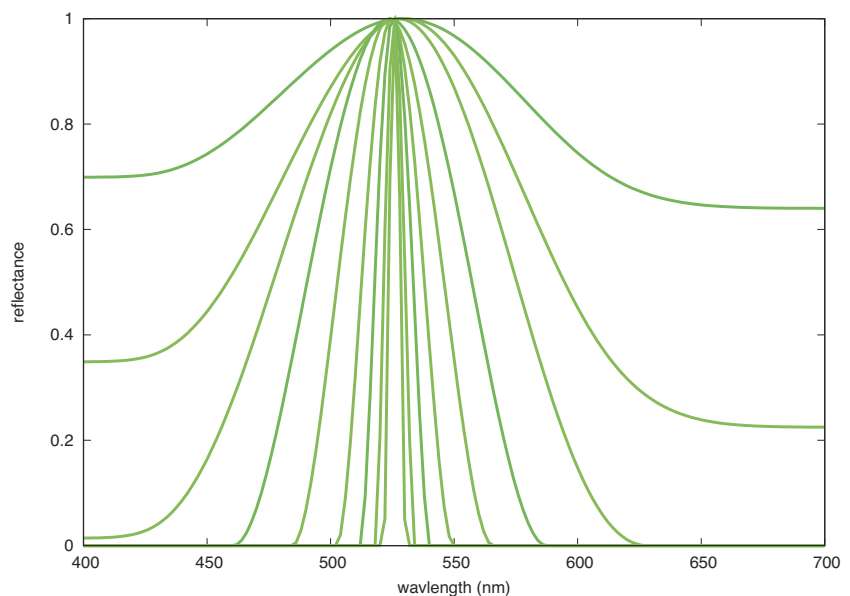
If we have a spectrum, we can calculate the XYZ or LMS values using the standard weights. If we had the XYZ or LMS values and wanted to know the spectrum that fitted these values, there would be many solutions.

There is a unique smoothest spectrum with positive values that matches any set of XYZ or LMS values that belong to a real colour. As we approach the spectral locus, the smoothest spectrum approaches a sharp, monochromatic peak. In the middle of the colour space it will be a smooth curve. If you are happy with that explanation, you can skip to the graph.

We define our smoothest spectrum as the one that gives the minimum sum of the squares of differences of adjacent spectrum values. There is a single solution that can be found using Lagrange multipliers. The problem reduces to inverting a large, square matrix, with rows and columns for each spectrum waveband, plus two to handle the boundary conditions, plus three to cover the three XYZ or LMS sums we are trying to fit.

Light level values cannot be negative. Reflectance values cannot be greater than 1.0. We cannot force this condition on the Lagrange minimization, so we have to apply it by changing the matrix. First, we do the calculation. If there are negative values we find the most negative value and change the corresponding matrix row and column to force that value to zero. If we have reflectance values over 1.0 we find the biggest value and change the corresponding matrix row and column to force that value to 1.0. We repeat until there is either no solution or we have a solution that fits our range requirements.

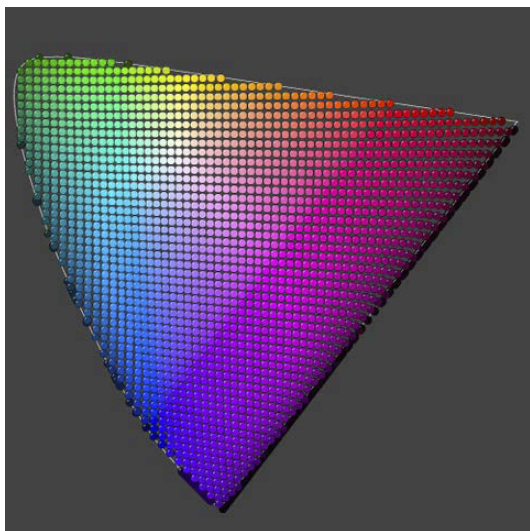
Here is a set of smooth reflection spectra for the same green with increasing saturation...



Smooth Green Spectra

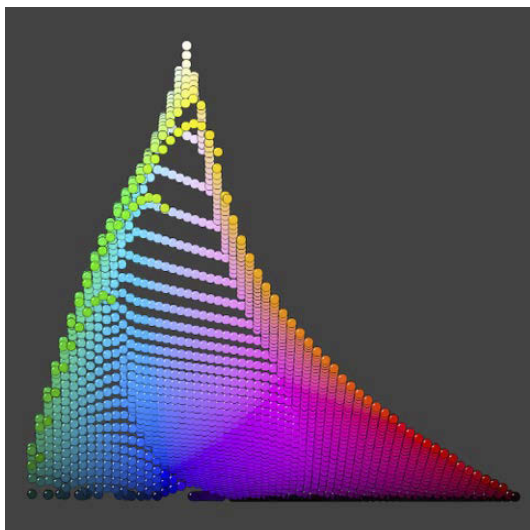
We can see we get a set of peaks around 525 nm. The peaks do not nest perfectly because of the way we avoided negative values, and because our spectra are discrete rather than continuous. But the results are fine for what we need. Here we allowed the spectra to go over 1.0, and then scaled them so the largest value in each was 1.0. This gave us the brightest reflection spectrum for the chroma.

Here is a set of reflection colours as they would appear under a 6500K illuminant plotted in CIE $u'v'$ using zview. The white outside line is the spectral locus. The points close to the spectral locus are dark except for the green-red upper edge. This part of the spectral locus is almost straight.



6500K reflection colours, $u'v'$

Here is the same set of points but viewed from the side, looking towards the top edge...



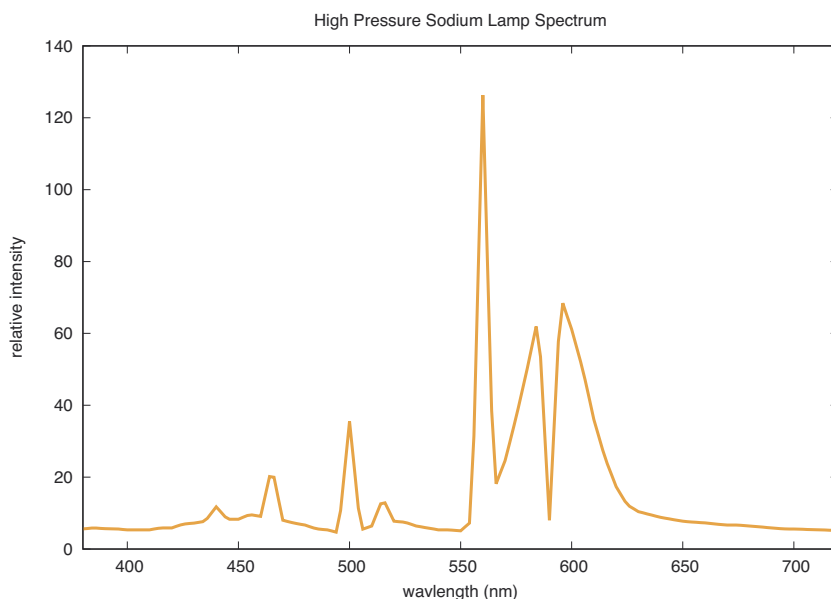
6500K reflection colours, Y vertical

We can see the peak white, and the two distinct edges on the back surface. There is a third edge going from the white point to yellow. We get these edges because we limited the reflectance values to 1.0. The part of the surface with the red corner is limited by the L receptor signal: we want to get the L signal as bright as possible, but we are stopped by the 1.0 reflection limit. On the other side of the crease that runs down to magenta we are limited by the S receptor signal. And on the other side of the creases that lead to azure and yellow, we are limited by the M receptor signal. Those lines should be there, and all is as it should be.

The surface we see bounds the set of real reflection colours. This is a 3D version of the Pointer gamut we mentioned briefly in Chapter 4, and will meet again in Chapter 7. We can get almost to the spectral locus, but where the spectral locus is very bent the colours are very dim, because only a few wavelengths will fit the colour we are after. Even the red-green edge probably has some curvature, but the points are too widely spaced to pick it up. There is no clear gamut line where colours stop like the Pointer gamut, but we could draw contours.

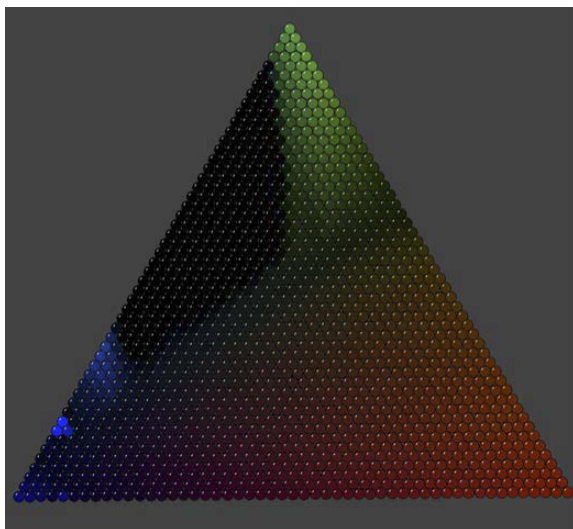
Changing the illuminant

Here is the spectrum of a sodium high pressure street lamp.



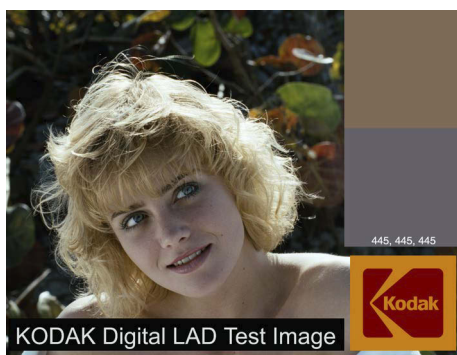
The sodium emissions are below 600 nm, but the high pressure makes them a lot less sharp. The bright peak in the green is from mercury. This is not the solid yellow of the old low pressure street lights: you can see full colour, but it may look a bit odd.

Here is the ‘zview’ plot of the triangle...

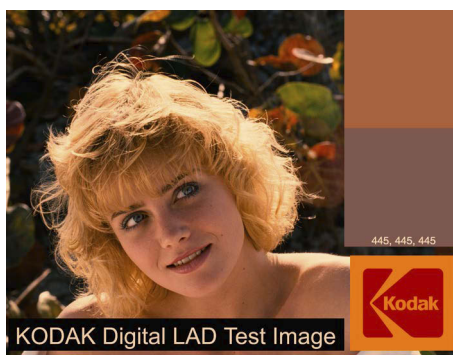


You can see the bright colours from the green peak. Remember the tRGB space green corner does not have real colours, so some of this green is not really there. We can also see the red, the yellow (just!) and the blue and violet peaks from the mercury.

This is what the triangle does to an image...



Before



After

On the left, we have the Kodak ‘Marcie’ image with D65 white. On the right we have the image after converting to tRGB, applying the triangle, and converting back again. We can use this to simulate unusual illuminants. Shooting with D65 and correcting may be more reliable than shooting using sodium lamps and correcting, particularly when the unusual illuminant confuses the camera. It is also useful for correcting studio illumination, so LED floods look like daylight.

Could we have done this with real colours instead of our smooth spectra? It is possible, but it is difficult. We might get a better match on the face if we had actual flesh spectra in our data set as real flesh tones are metameric. Real foliage has two different types of chlorophyll, and we do not know which one we have here. Ideally, we want a set of real spectra where every spectrum was realistic but also fits the nearby colours so the transform is smooth. This is difficult; it may even be impossible.

Do we know what the spectrum of a given colour is? Our image has come from a camera or a scanner. We could measure the spectra with a spectrometer, but we probably do not know the spectral sensitivity of the camera itself. Our ‘flesh pink’ could be skin in daylight or a neutral object taken at sunset. We are not going to get things perfect; the smoothest spectrum set gets us close enough.

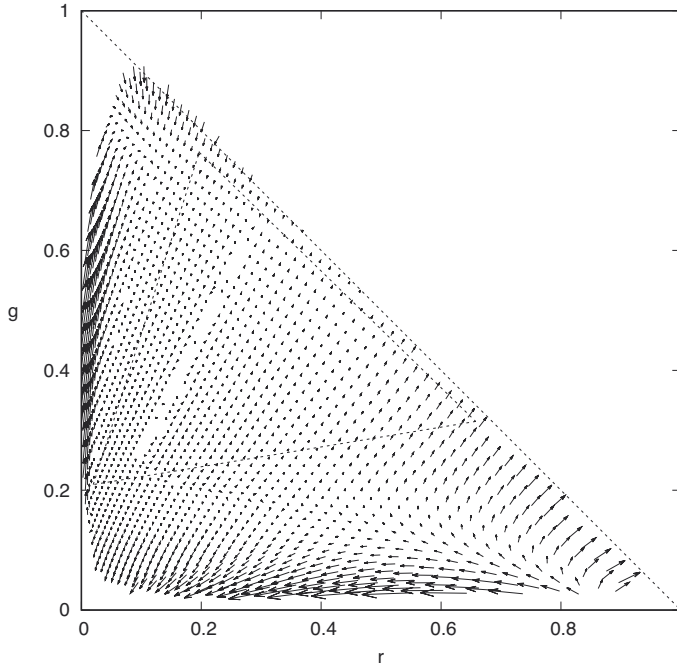
Transforming between XYZ and LMS

In the last section we tried changing the illuminant in an image. This is not something we can do without knowing the full reflection spectrum of everything in the image, but we made a guess using the smooth spectrum set.

In Chapter 4 we had a similar problem. If we had a spectrum, we can make XYZ and LMS values. If we have some historical values in XYZ, we cannot turn them into LMS without some extra information. However, that did not stop us when we changed the illuminant, so let us see if we can get away with the same trick again.

I made an XYZ to LMS list transform from the smooth reflection spectra and an illuminant spectrum. I used a 6500K blackbody spectrum as that is nice and smooth. This list is not enough to span 3D space, as it only has points on the smooth spectrum surface we plotted earlier. We can use this to make a triangle transform as we did in the last section. This will work best when converting from LMS to XYZ, as we can go via **r,g** space, so the triangle is mostly full of real colours.

Instead of a triangle transform, I decided to fit a 3×3 matrix. We have the LMS values for our smooth spectrum set. I can calculate another set of LMS values using the matrix. I could then vary the matrix values until I minimized the errors.



Here is a plot of the errors in r,g space for smooth, reflective colours with a 6500K illuminant. The tail of the arrow is the colour from the CIE 1931 XYZ using the matrix, and the head is the colour from the CIE 2006 LMS. We can see the errors are small for most of the Rec.709 triangle. They get really large only on a few points close to the spectral locus, where we expect the XYZ and LMS values to disagree. We want a matrix that does the best job on the colours we are likely to meet, rather than one that gives a better fit on hypothetical but very dim colours at the bottom or on the left. So I tried fitting only the smooth spectra that were bright enough to have a Y greater than 1% of the white. I varied this threshold: I got much the same matrix from 2% to 20%. 1% seemed to work well, and this is what I have plotted here.

Here is the matrix we fitted....

$$\begin{pmatrix} L \\ M \\ S \end{pmatrix} = \begin{pmatrix} +0.257085 & +0.859943 & -0.031061 \\ -0.394427 & +1.175800 & +0.106423 \\ +0.064856 & -0.076250 & +0.559067 \end{pmatrix} \cdot \begin{pmatrix} X \\ Y \\ Z \end{pmatrix}$$

This is good for converting reflection colours with a 6500K illuminant. It will probably work for D65 and D60 illuminants, but it will not necessarily work for self-luminous objects, or for very different illuminants.

The von Kries model

Von Kries published a notable paper on *chromatic adaption* in 1904. Chromatic adaption is the term for our ability to adapt to different illuminants. If we did this perfectly, it would be called *colour constancy*. He proposed we had three vision channels to the eye, which he called RGB, and they adapted independently. Back in Chapter 3 in the section on XYZ we saw we could not determine the actual eye primaries from colour matching experiments alone. However, von Kries knew that the RGB he sought must be some linear combination of the XYZ primaries...

$$\begin{pmatrix} R \\ G \\ B \end{pmatrix} = \begin{pmatrix} R_X & R_Y & R_Z \\ G_X & G_Y & G_Z \\ B_X & B_Y & B_Z \end{pmatrix} \cdot \begin{pmatrix} X \\ Y \\ Z \end{pmatrix}$$

I shall write this more compactly as...

$$\text{RGB} = M \cdot \text{XYZ}$$

We do not know any of the 9 coefficients in the matrix just yet, but let's keep going. If we had the matrix, we could also calculate an inverse matrix...

$$\text{XYZ} = M^{-1} \cdot \text{RGB}$$

Von Kries guessed the RGB sensors in the eye adapt separately to the changes in overall light level. Suppose we have an original scene lit with D65 daylight, and we want to make a print of the scene that will be viewed in CIE standard illuminant A (tungsten) light. The illuminant white is XYZ_{D65} in the original scene, and XYZ_A when we are viewing the print. We can calculate the corresponding eye exposure values RGB_{D65} and RGB_A using the matrix. If we have a patch in the original scene that is XYZ_{D65} we get the visual match XYZ_A like this...

- 1 Convert XYZ_{D65} to RGB_{D65} using matrix M .
- 2 Convert RGB_{D65} to normalised RGB by dividing by the daylight RGB_{D65} .
- 3 Convert RGB to tungsten RGB_A by multiplying by the tungsten RGB_A .
- 4 Convert RGB_A to XYZ_A using the inverse matrix M^{-1} .

Von Kries did not know what the matrix values were, but he could guess, and he knew that if the guess worked then he may be right. Here is just such a matrix...

$$\begin{pmatrix} R \\ G \\ B \end{pmatrix} = \begin{pmatrix} +0.38971 & +0.68898 & -0.07868 \\ -0.22981 & +1.18340 & +0.04641 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} X \\ Y \\ Z \end{pmatrix}$$

This matrix is often called the von Kries matrix. It is also known as the *Hunt-Pointer-Estevez* or HPE matrix: Estevez used it in his PhD thesis with the von Kries model.

You sometimes see the von Kries matrix quoted as converting XYZ to LMS. The eye primaries are indeed LMS, but the scaling is not right for CIE 2006 LMS. This is obvious if you compare the bottom row of this matrix to the bottom row of the matrix at the end of the last section.

There are several reasons for this...

The simple reason is the von Kries model does not need an absolute scale for its RGB values as it always converts back to XYZ again. By convention, von Kries matrices are usually normalised so each row adds up to 1.0. CIE 2006 LMS has an absolute scale, and this isn't it. So, if you come across a matrix that seems to be converting from XYZ to LMS or vice-versa, and the rows all add up to 1.0, it has probably been normalised.

A more subtle reason is that I was not simply fitting a matrix in the last section to transform the whole XYZ space to LMS. I chose the points that I wanted to fit. So, instead of S depending only on Z, it now has small weights of X and Y. I could force these two weights to be zero, but this would be a different matrix.

Advanced colour appearance models

I do my chromatic adaption in LMS. I do not have a matrix M because I am already working in eye primaries. This is nice and simple. This does not mean all the old XYZ-based von Kries work goes in the bin. The people working with the von Kries model are not just doing a simple fit. They also attempt to model the non-linear aspects of cone adaption, the sampling of the white point from the scene, the effect of inducing colours, the Hunt effect, and many other things.

CIECAM stands for the CIE colour appearance model. There are several versions. The most recent came out in 2016. Before these, there were the Hunt Model, the Nyatani model, and the Bradford University model. There is Land's Retinex model, the Fairchild model, and many others. I have used the Kim vision model as a reference, because I knew this had been tested using some bright images at up to 40,000 nit, which was what I needed at the time.

Do you need to know about these spaces? There is an excellent book by Mark D. Fairchild called "Colour Appearance Models" that describes each of the current appearance models, and compares them. This review is a job that only needs doing once, and I will not attempt to repeat it. The more complex colour appearance models have extra parameters, and it can be hard to choose the best values to use for these. I used CIELAB space within my Truelight film simulation, which is normally considered too simple to be a proper colour appearance model. I also supported the

CIECAM02 Jab space internally to see whether this might help, but I never saw any difference. This was probably because of my particular requirements: I was not changing the white point by much (most cinema whites were about D60, and video was D65), and I had separately included the Bartleson-Breneman calculations for the brightness and the effect of the surround. This probably did not leave much for CIECAM02 to do.

I suggest you use the simplest model you can. If that doesn't work, the way it failed may suggest which of the more complex models should work, and what settings to choose.

Camera and scanner calibration

We calculated the smoothest set of spectra that fitted a particular set of colour values.

Suppose instead we have taken an image of a Macbeth chart, and also measured the reflection spectrum. Can we deduce the sensitivity of the camera from this data?

In theory, we can. First we must get rid of redundant values. If we have good data from the white patch, then the grey patches will not tell us anything more. Some colour patches may be too close to each other, which will make the interpolation noisy.

In practice, the Lagrange minimization we used for the smooth spectra is too precise. Give it measured data from a Macbeth chart and real spectra, and it returns a spectrum that is an exact fit to the data. But the data will often be very far from smooth, and half the values may be negative. It is fine with ideal data, but it is very sensitive to small experimental errors.

I have calibrated cameras and scanners with a different technique. I started off with a guess that the camera has a peak in its sensitivity spectrum (or two peaks for the red channel if necessary). My program then moved the peaks, made them wider or thinner, if necessarily skew them so one side is wider than the other, until we found the best fit. This works, but it needs watching; this is not a process that I would trust to run by itself.

We have another way of calibrating cameras that can run unsupervised. We can capture the camera values for the Macbeth chart patches for a set of exposures, perhaps at 1 stop intervals. We can combine the grey patches to estimate the tone curve from the lightest white to the darkest black. We have pre-measured Macbeth XYZ values for the Macbeth chart patches for some typical white points. We can then correct the tone curve on a frame which has all the colours, and fit a 3×3 matrix. This is not as good as reconstructing the spectral sensitivity, but it is more reliable. It also fits real cases in common practice. Suppose the company hires a

camera for an unusual shot (drone, underwater, high speed). It may have gone back to the hire company before anyone thought of calibrating it, but they often have a test shot with a Macbeth chart. That and a guess of the lighting temperature is all we need.

Scanning watercolours and acrylics

I have not yet managed to calibrate a camera using found colours in the environment, but I have managed to calibrate a Canon CLC500 photocopier to scan Windsor and Newton watercolours. Windsor and Newton kindly gave me a chart with real paint examples. There were about 40 watercolours excluding non-fast colours and Chinese White and Lamp Black (there are now about a hundred). Each patch had a density gradient, so I could get three measurements off each patch using my spectrometer. I could also get a digital scan of the test chart, which gave me the scanner RGB values for each point where I measured the spectrum.

I assumed I could add black to any colour without changing the hue. Lamp Black had a pretty flat reflection spectrum, so it ought to do just this. With the paper white measurement I could calculate the CIELAB values for each patch, and relate it to the scanner RGB. I could also make extra points by simulating adding black.

I also printed out a test chart with a set of RGB colours using the copier as a printer, and scanned that back in. I also measured the chart patches with a spectrometer and calculated the $L^*a^*b^*$ values for each patch. I could then make a cube that corrected the scanner RGB by converting scanned watercolour RGB to CIELAB, and from CIELAB to the print RGB.

I scanned and printed several watercolours, including a painting done by my grandfather. I still have the print of that, and it is hard to tell the print from the original. Two colours (a deep red and a light green) were outside the printer gamut. Vandyke brown is also very saturated, and hard to match.

This process only worked with watercolours. The photocopier in watercolour mode would not give a good match when copying its own output or other printed material. It was able to give a fair match to a Rowney acrylics example chart because many of the pigments were the same.

This suggests that calibrating a camera and colours from the environment may be possible, but something might have to recognise the objects and identify the colours. Possibly AI will be doing that soon.

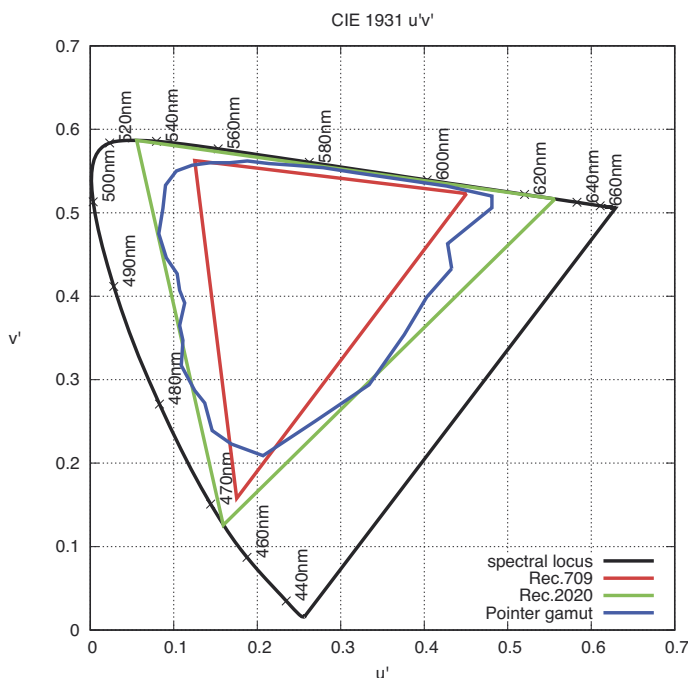
Chapter 7

High gamut and HDR displays

In the last chapter we saw how the smooth spectra reflection colours showed a luminance peak at the illuminant white point, and went down to zero at the spectral locus. We also saw in earlier chapters that we are very good at judging colours and compensating for the illuminant, with rare exceptions.

Display gamuts and Pointer's gamut

Here is a comparison of the old Rec.709 video gamut we are familiar with, and the new Rec.2020 display gamut. Rec.709 was based on commercial RGB phosphors. Rec.2020 is based on three wavelengths: 630 nm for red, 532 nm for green, and 467 nm for blue. You cannot buy a Rec.2020 display, but modern displays can get fairly close. The wiggly line is the Pointer Gamut. This encloses the set of real reflection colours measured by M. Pointer in 1980, and is commonly used to represent the span of colours you may meet in real life.

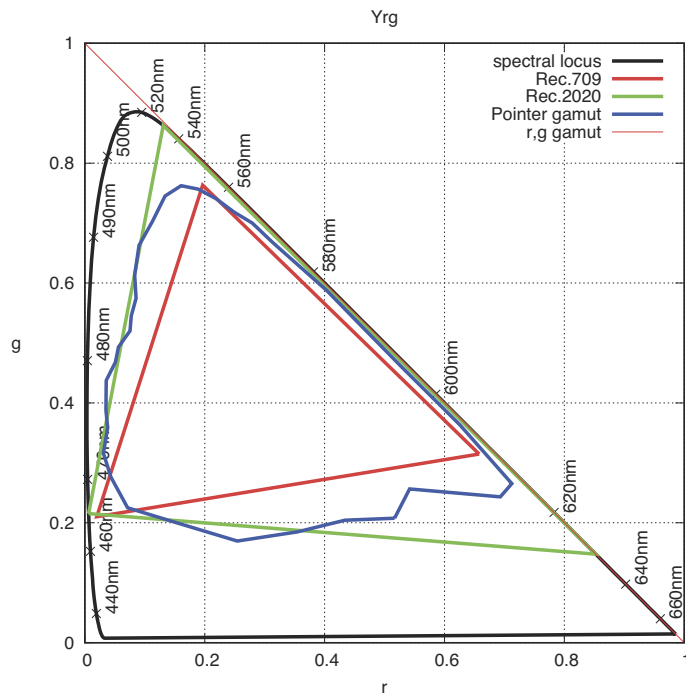


Colour gamuts in u',v'

Pointer's gamut does not necessarily include every reflection colour. We saw in the last chapter that we could get very dim colours almost up to the spectral locus. But it is probably a fair sample of all reflection colours that are bright enough to see. The Rec.709 video gamut is missing some deep reds, pinks, and ice-blue colours. The Rec.2020 gamut is getting just about everything.

Why does Rec.2020 not use 520 nm instead of 532 nm and grab a bit more colour space? There are two possible reasons. In the last chapter we saw that the colours between green and red were bright enough to be seen almost up to the spectral locus; it is more important to get these colours than it is to get the much dimmer colours on the rest of the locus. Also, real displays do not use single wavelengths: they may use a green peak that goes from 520 to 540 nm, with an average that is close to 532 nm. If they used a similar peak going from 510 to 530 nm, the average would not lie on 520 nm, as the 510 nm edge of our peak has left the red-yellow line and started going around the corner.

Here is a similar plot in the CIE 2006 based **r,g** space...



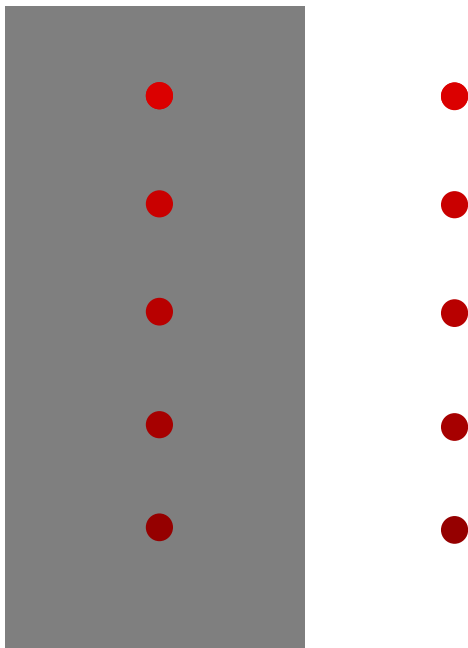
Colour gamuts in **r,g**

I do not have the spectra that Prof. Pointer used back in 1980, so I have had to interpolate using my smooth spectra. The Rec.2020 primaries are based on wavelength, so these are accurate. The Rec.709 primaries are based on phosphors, and I have some measured spectra for these, but these may not represent all video.

In the **r,g** chroma space, Rec.2020 does not contain a few purple colours that lie in the Pointer gamut. This may not be real, as I interpolated the spectra. There are other colours we can get in real scenes that lie outside the Pointer gamut. Underwater scenes, and shots in ice caves can have very intense azure colours, though the light is coming from the sun. We can also have LEDs and other narrow-band lights right on the spectral locus.

Seeing reflection colours

Here we have two sets of red dots. The dots on the right are on a white background. They look like reflection colours...



Luminous colours

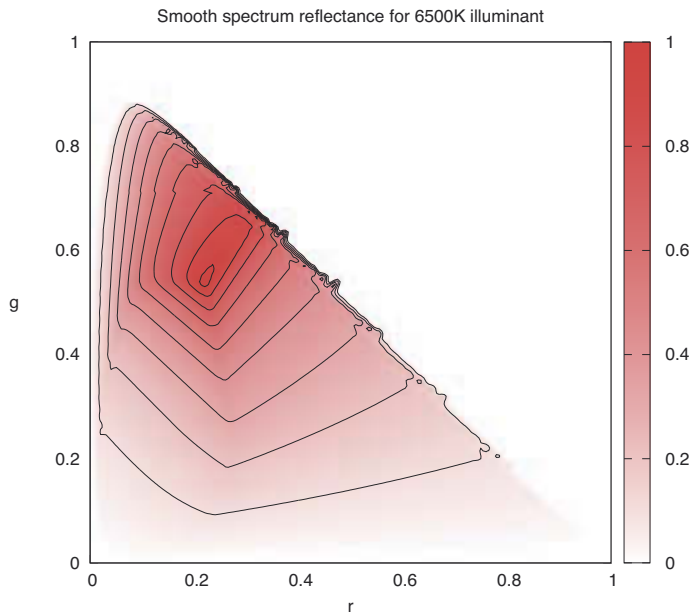
The dots on the left are on a grey background. If the lighting was dim and the background was white then the red dots would be too bright to be a reflection. They would have to be luminous. We would not expect to be fooled by a simple scene like this, but at least the top two dots do have a strange luminous property to them, like Monet's rising sun in the Pool of London. We see luminous colours in real scenes, such as traffic lights, car indicators, LEDs on electronics, and displays; and we have learned to handle them as exceptions to the colour constancy model. If you want to feel how difficult it is to mix luminous and reflectance models, imagine trying to pick a Munsell or Pantone colour to match the hue of a sodium lamp.

If you ask people to rank reflection colours in order of lightness, they often rank very saturated colours as being lighter than their measured luminance. The colours seem to have a 'glow' associated with them. This is the Helmholtz-Kohlrausch effect, briefly mentioned in Chapter 4.

Should we modify our definition of brightness to include the Helmholtz-Kohlrausch effect? This would bring the measurements closer to our language: we could start talking about 'bright colours' again. On balance,

this is a bad idea. Luminance can be measured objectively and works well. The Helmholtz-Kohlrausch effect exists, but it depends on the viewer's understanding of the local 'white point', and is much harder to measure objectively.

Here we plot contours of the reflected luminance as a fraction of white in our LMS derived **r,g** space...



The contours wobble a bit because I have interpolated from a set of smooth spectra and a set of zero points along the spectral locus, which makes interpolating between the two a bit random. But this still shows where the most visible reflection colours lie.

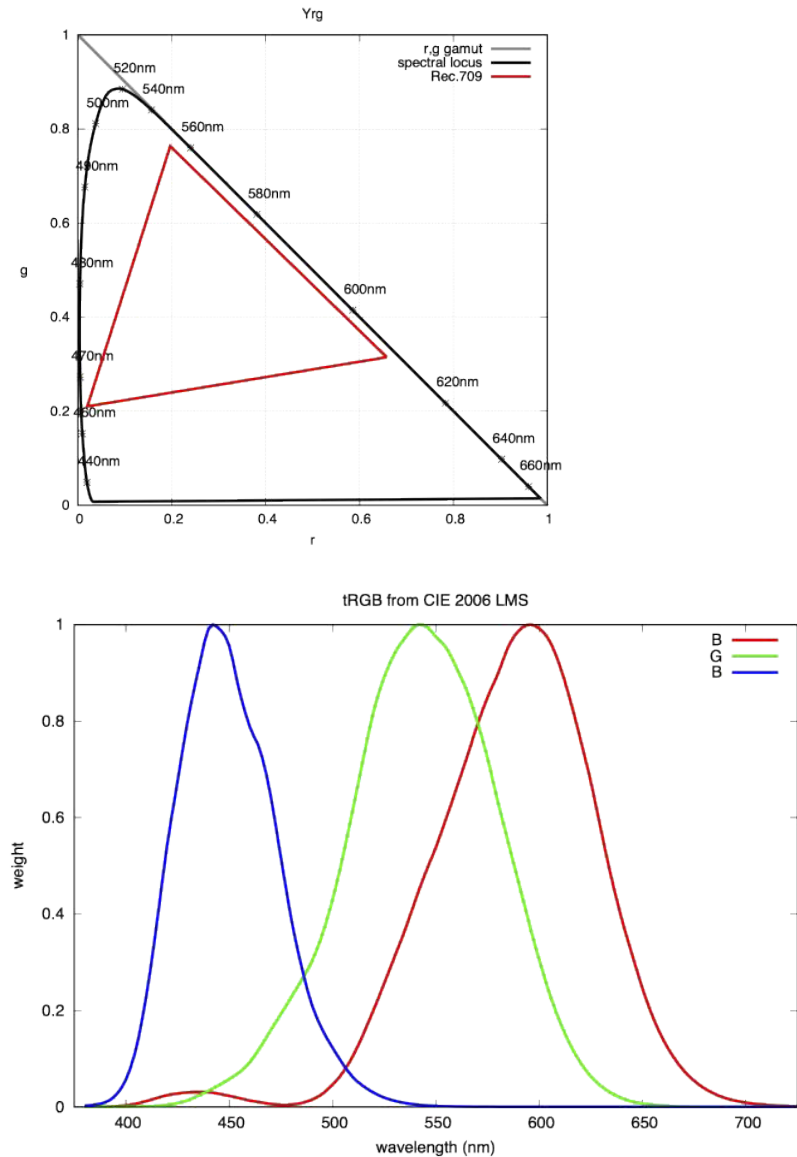
In Chapter 3 we met the Bezold-Brücke effect. This was a hue shift with light levels that can be measured in simple vision experiments, but almost never seen in a realistic scene. We may not be able to tell whether an abstract colour patch is red or orange, but a real scene has cues.

This affects our choice of monitor primary. A large colour gamut would be nice. There are some azure hues in nature that we cannot match on most displays. However, as we approach the left-hand side of the spectral locus, the wavelength band becomes narrow, and the brightest reflection colours are very dim. We saw this in the plot of smooth green colours in the last chapter. The colours on the bottom straight edge of the spectral locus are mixtures of the longest and shortest visible wavelengths, so the reflection colours are very dim here too. The only place where bright colours come close to the spectral locus is the nearly straight section of the spectral locus near the yellows.

If a colour is too bright to be reflective, it looks luminous, and we cannot integrate the colour we see with the rest of the scene. If the colour is too dim, we find other ways of estimating the hue, or we would see the Bezold-Brücke effect. There is no clear point where these two limits cross over. We may want to image luminous colours such as azure Christmas tree lights, but there are diminishing rewards from chasing the last scraps of colour gamut.

New primary wavelengths

Suppose we made a new display. Which wavelengths might we choose? We are not necessarily restricted to having three primaries.



The first plot has a straight line from 535 nm to 700 nm. These wavelengths were ‘green’ and ‘far red’ in the table of Chapter 2. We can mix any colour from ‘green’ to ‘far red’ using these two wavelengths.

Using 700 nm is a bad idea. Our eyes are much more sensitive to 650 nm. 650 nm may give different colours to different people. The R sensitivity curve is steep at 650 nm. Genetic variations may move our R and G peaks by 5 nm or more, so a mixture that appears as yellow to one observer may look orange or yellow-green to another.

610 nm is almost the Rec.709 red. This looks red unless there is some 650 nm red to compare it to. I remember calibrating a Sony BVM CRT in a darkened room, and seeing the calibration patches as a deep red. We then calibrated a Barco DP90P but left the Sony monitor on. The same Sony red now looked almost orange. This may seem obvious as the two wavelengths are far apart on the *r,g* plot, but we are well beyond the range of our Munsell points, so it is nice to have this confirmed by something that emits light. The good news is 610 nm is almost at the peak of the R curve, so genetic variations will have little effect on mixtures with 610 nm red.

535 nm is near the peak of the G curve, so it is efficient, and genetic variations should have less influence. The Rec.709 phosphor is nearer to 550 nm, but it has a broad peak: if we shifted it to 520 nm it would probably no longer sit on the spectral locus. The Rec.2020 green is 532 nm.

500 nm is an azure colour. This is just where the R curve packs up, so we can use it to get blue-green mixtures. This colour is the one that RGB displays cannot get, so we need it for a full gamut.

450 nm is a blue colour. This is close to the peak of the B curve, so it is efficient. There are no known generic variations of our S receptor. The Rec.2020 blue is 467 nm.

The contour plot a page back or so suggests there are no bright colours at the bottom of the *r,g* plot, so we can stop at 450 nm and 650 nm. There are also no bright colours on the left-hand side, so we would lose little outside the 450 nm to 500 nm mixture line.

This gives us five good candidates for narrow-band primaries...

650 nm	Deep red
610 nm	Red
535 nm	Green
500 nm	Azure
450 nm	Blue

Colourfulness and the Hunt effect

Imagine we have a slide projector projecting a colour image on a screen. Now move the projector towards the screen, and refocus. Our image is now smaller and brighter, but the total amount of light is the same. The image will also appear more colourful. The colours have the same saturation as they had before, so we need a new term 'colourfulness' to describe what has changed.

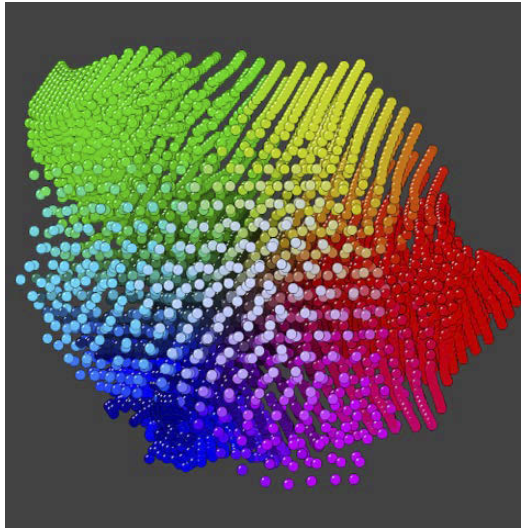
This change in colourfulness with brightness is called the Hunt effect. Red roses viewed on a bright summer's day can give an almost luminous or hallucinogenic sensation of red.

This is not usually an important effect for images on displays. If we turn up the brightness, most images look much the same. But there is an important consequence when dealing with a finite light budget. Suppose you have a white lamp, and you want to project red, green, and blue separations. You get the best colour gamut by filtering out 700 nm red, narrow-band 535 nm green, and deep 400 nm violet; but you will also get very little light. You can use a broadband 630 nm red, 535 nm green, and 450 nm blue and get a lot more light; and get a greater sensation of colour as well thanks to the Hunt effect.

White primary

There is another sensible option for displays: add a white primary. OLEDs can generate broadband white. The OLED colours often include a coloured filter, so it is more efficient to put out pure white than make white out of R+G+B. Replacing R+G+B with broadband white also will reduce the effects of genetic variations. Different people will still see the deep red colours as having different brightness, but this will be confined to the red corner, and not extend all the way to the whites.

It is not clear how commercial OLEDs make their white primary signal. Here we measured the XYZ values for an OLED set up to display Rec.709 using the white primary...



OLED gamut in Rec.709 RGB

A conventional Rec.709 monitor makes all the colours from three primaries. The gamut would have parallel sides and the colours should lie in parallel rows. This display is clearly doing something different.

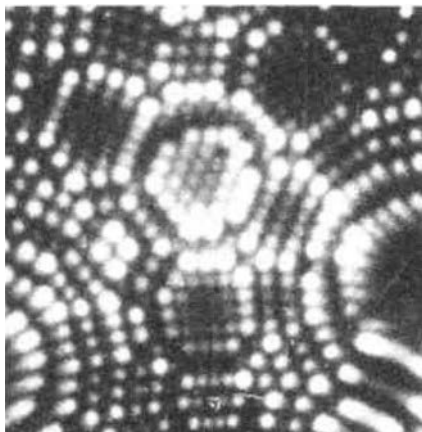
An obvious thing to do would be to make all colours with the same chroma by mixing the same ratio of primaries. These colours would all have the same spectrum, but at different scales, so they would all look the same hue to all of us even if our eyes were not the same. Here it seems the dim colours were made using the RGB primaries, and the white was added for the highlights.

Dithering shadow tones

Many flat panel displays use the same panels. These panels have millions of pixel elements, and these all have to do exactly the same thing. Plasma panels had inert gas filled cells, containing a small tungsten needle that emitted ions, and a phosphor target that glowed when it was hit by the ions. The amount of light emitted depended on the voltage you put across the cell.

The field-ion microscope also used ions emitted from a small tungsten needle. It was the first device to image individual atoms on the tungsten needle. The field was strongest on the atoms at the edge of the

crystalline layers, and this gave the bright spots when the ions hit a phosphor screen...



Atoms through a field-ion microscope

The brightness of the plasma panel probably depended on the position of the last few atoms on its tungsten needle. This meant plasma panel elements never quite matched each other. The later models had many needles instead of just one, but you could still see a fixed pattern in the shadows.

OLEDs are better. The amount of light from an OLED depends on the current, and the OLED current-voltage characteristic is pretty linear over most of the brightness range. However, most OLEDs use the same panels; the difference between the cheap and the expensive ones is largely down to how well they control the tone curve down in the shadows.

We can make grey levels from on-off steps using a Bayer dithering pattern...



Bayer dithering pattern

We can make a time-dependent dither pattern where each dither has the reversed dither on the next frame, so most dithering happens at half the frame rate. There would be a smaller flicker at a quarter of the frame rate and so on. We would spot flicker at a quarter of the frame rate at ordinary light levels, but we are less sensitive at low light levels. This would be best done in the OLED driver, but we have shown it can work by rendering it into frames.

Display calibration

When Truelight first came out, a large post house might have had a hundred CRT monitors. There would be one or two of the expensive Sony BVM monitors which were stable. The rest would run for weeks and suddenly one morning the blacks would all be grey, and no-one knew why. We would have to calibrate several monitors a day as routine, and then to recalibrate any monitors that had gone out of calibration unexpectedly. Monitors also had knobs to adjust 'contrast' and 'brightness', and people often fiddled with those even when we asked them not to.

The Truelight probe was a colorimeter. It had four detectors with dichroic filters matched to the XYZ primaries, with a violet detector for the second X peak. This worked well for several years, because all CRTs and some plasma displays all used the same primaries. We used the same device for calibrating digital projectors. This worked for a while, but digital projectors started using wide gamut primaries. When the red primary shifted from 610 nm to 650 nm, it was half-way down the side of the dichroic peak, so small errors in the dichroic gave XYZ values that were clearly wrong.

The Truelight Projector Probe had a colorimeter and a spectrometer. Spectrometers could measure colour accurately, but they were not sensitive enough to measure all the way into the shadows. The Projector probe had a startup routine. It would do a black calibration to calibrate the dark point of the spectrometer and the colorimeter. It would do a white point balance to calibrate the colorimeter from the spectrometer. If we were using our calibration program, it would also calibrate the RGB primaries and take out the cross-correlation effects.

Back then, displays had a white point of about 50 nit. We needed a 2000:1 contrast range to do a convincing emulation of film, so we needed to measure to within 0.01 nit. This we could do with the colorimeter.

We also had a 'by-eye' calibration option. This was originally a fall-back for when a probe was not available. You could adjust a grey to match alternate lines of white and black; then adjust another grey to match alternate lines of the previous grey and black; and so on.



'By-eye' calibration image

This image needs to match the monitor pixel spacing to look right. If we could do this, then the striped pattern in the middle should match the grey surround in sRGB. This is not the 'by-eye' calibration, but the principle is the same. It could not check the white point, or measure the primary colours, but it could check the tone curve. I could do a by-eye calibration in a couple of minutes.

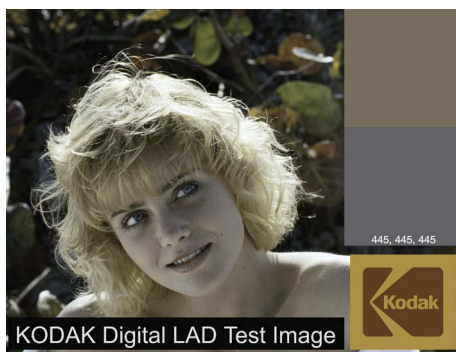
Note: To access an online version of this book, and other related resources, go to: www.filmlight.ltd.uk/colourbook

Displays are now a lot more stable. They may need calibrating once a year, if that. The greatest risk is usually that someone has fiddled with the display controls. Displays have also got brighter (200 nit to 800 nit). The contrast is good, and the dark point is usually right. But we may have more than 3 primaries, so we have to use a spectrometer for all the readings.

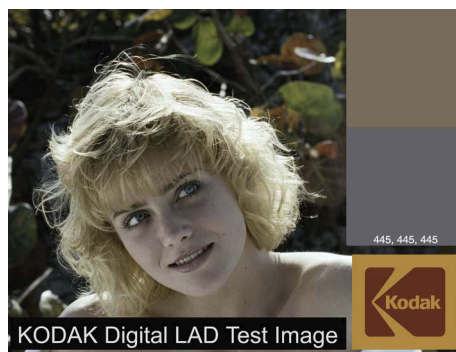
Suppose we have a monitor with a 200 nit white. If we want a 2000:1 contrast ratio we must measure down to 0.1 nit. Suppose we measure down to 100:1. That means we can measure down to 2 nit. This is the sort of level that ordinary spectrometers can do. We can extrapolate the tone curve assuming the gamma value is fairly constant, which it probably is if all pixels go smoothly down to black. We can do the whole calibration with a good spectrometer, and extrapolate for the cheaper ones. If your facility has a digital projector, they probably already have access to a spectrometer.

If you cannot get the shadows looking right, there is still the 'by-eye' calibration.

Using tRGB as a grading space



tRGB



Rec.2020

The image on the left is 'Marcie' converted to tRGB with a gamma of 2.5. The tRGB space contains all possible colours, and only 6% of unreal colours.

The image on the right is 'Marcie' converted to Rec.2020 with the same white point and tone curve. The Kodak logo is a bit more saturated. The 500 nm azure colour would be well out of gamut.

The colours on these images should not look right in this document. Both are probably shown in sRGB. If we made the colours accurate, the saturated colours would have clipped. The point here is that the faces look similar, so tRGB and Rec.2020 are doing the same sort of thing in the middle of the colour space. However, tRGB can get all the colours in the spectral locus, while Rec.2020 can only code a triangle inside it.

HDR and high gamut LUTs

Chapter 6 talked about LUTs and other tools used to capture colour transforms. Can we use the same tools on HDR images?

LUTs are valid for a finite range, typically between 0 and 1. Most display-based spaces are fine. We can drive the display with RGB values between 0 and 1 and measure what the display does. Even Cineon and other film-based colour spaces are effectively display based: we write values to a film recorder, but they eventually become a projection. Even an HDR display can be characterized, though we may need more colour patches, and an instrument that can deal with brighter colours.

A scene is a bit different. We can point the camera at reflection test charts, and measure the test chart response. We can measure the test charts at various exposures, and make a composite set of measurements with a larger exposure range. There are special camera targets with

narrow-band filters that come close to the spectral locus. Even then, most points in the RGB space will lie far from any test patch measurement. This effectively limits the usefulness of 1D and 3D LUTs for high gamut spaces. We can make larger LUTs for larger colour spaces, but if we do not have a simple way to characterize the whole space, it is rarely useful or practical to do so.

The triangle LUT is an exception. We can generate the triangle LUT from spectral measurements, so we can generate data for the whole spectral locus. The triangle transform scales with brightness, so HDR ranges are not a problem. It can only do transforms that scale with brightness, but this is quite a useful subset of all the possible transforms.

Correcting for IMAX

IMAX is a wide screen format. The original film IMAX was like this...

	width	height
35mm	21.9 mm	18.6 mm
IMAX	70.0 mm	48.5 mm

This was projected onto a proportionately larger screen. It is hard to be exact how much bigger IMAX is because that depends on how the theatre was laid out and whether you sat at the front or the back. But in general, the screen should be about 60 degrees high and more than 90 degrees wide to the viewer.

All the important narrative action probably happens within the conventional screen region, but the extended border of the image gives an immersive feel: you can forget you are in a dark room looking at a picture on a screen. This may sound like a gimmick, but it can be quite powerful. I remember seeing a re-master of the 1971 film “North of Superior”. This showed the scenery on the Canadian shore of Lake Superior as shot from a light aircraft. When the plane tilted up and laboured to haul the IMAX camera over the Ontario cliffs, I gripped the side of my seat because I felt the tilt.

In Chapter 5 we discussed the work of Bartleson and Breneman, and their formula for adapting for the different surround. If you have a dark surround, you should use an adaption factor of 0.1. This gives you more visible contrast in the shadows, which fits because you have less peripheral flare and other such effects preventing you from seeing the detail. If you have a light surround, such as a print on paper, this logic suggests you should use an adaption factor of 1.0. I don’t think they do, though.

If the IMAX screen is wide enough, you perhaps ought to use an adaption factor of 0.5. This would give you less contrast in the shadows than for a conventional projection, but it ought to be closer to being there.

Limits of contrast

Back in 2005 or thereabouts, LCDs had a contrast ratio of about 600:1. A few could get up to 800:1, but their blacks were not uniform. They could be used for image processing in an office environment, but in a darkened room they could not give a convincing black like film did. A good CRT could give a contrast range of 4000:1. As CRTs get older, they start having 'tube glow' and the blacks get less dark. However, as a rough guide, we found we could get a good imitation of film with 2000:1 contrast, we could match all but some very dark shots with 1500:1 contrast, but anything less than about 1200:1 was not convincing.

The 2000:1 threshold was a good rule of thumb for matching film. You can still see the film grain noise in projected film black, so we may need more than that to match a general image. However, there are good reasons to believe we may not need much more than 2000:1 to convince the eye. Our eyes are not black, and our retina can reflect and scatter. A cat's eye has a retro-reflective layer called the *tapetum lucidum*. This reflects light back the way it came, so it gets a second pass through the retinal cells, and some of it goes back out through the pupil which is why cats' eyes reflect. Many vertebrate animals have this, but not humans. So the light inside our eyes bounces about and scatters within our eyes until something absorbs it. This scattered light gives a uniform fog on our retina, making it hard to see more than 100:1 contrast. We can see a lot more than 100:1 contrast, but we do this by looking at the highlights and at the shadows, but not both at once, and by darting our eyes about we build up an HDR-like sense of the scene.

Our 2000:1 contrast ratio is a working figure based on the images we know. It is much better than the static human eye. It is a typical figure for a good camera. 2000:1 is a good figure for today, though audiences in the future may demand more. Modern flat panel displays can manage 2000:1. Some technologies such as OLEDs boast much higher figures, but this may be misleading as we need a smooth tone curve all the way to black, and not just a good black. However, it seems likely that dithering the shadows could give us a good tone curve well beyond 2000:1 with current mass-produced displays.

Planetariums have the highest contrast ratio experience. The Zeiss Velvet projector has a contrast ratio of 2,500,000:1, and there are other projectors that go higher. These planetarium projectors do not have to be very bright to simulate the night sky. Their dark tones would probably be completely wiped out by the safety lighting levels in ordinary cinemas.

It seems technology can provide the contrast ratios for our immediate needs. Brightness is a different story. Bright but non-HDR displays of 2020 can go up to 800 nit. The table in Chapter 5 gave the brightness of the solar disc as 1.6 billion nit. So, displays are five million times less bright than something we may see on a good day. Even the Moon at night is 2500 nit. We are not going to mimic absolute light levels of outdoor scenes any time soon. I am typing this on a laptop with a white of about 150 nit. There are two open windows, and it is nearly noon on a sunny day outside. I would not grade scenes like this, but the display is good enough for everyday use. If I was going to watch a motion picture, I might draw the blinds, or watch it at some other time. If my display was brighter, it would consume more power and need more cooling.

Chapter 5 dealt with matching different media. If we can get a 1000 nit peak white and a 250 nit paper white, clever grading might capture most of the appearance of ordinary scenes: water would look wet and gems would sparkle. The 1000 nit display would not look good right next to a 5000 nit display, but on its own it should look fine.

A 5000 nit display has other disadvantages besides needing more power and cooling. It can dazzle us. Our eyes adapt to different light levels, but sudden exposures such as dazzling light on water on a bright day can leave spots before our eyes which last for half a minute or so. The eye pupil reacts to changes in light level in seconds, but it only gives a 4:1 adaption. It can just go to 7:1, but getting to either limit takes more time. The rest of our adaption between starlight and full daylight happens in the retina. Cone cells replace their dyes with a time constant of about 130 seconds. We are not normally aware of this because it happens steadily, and our eyes, irises, and brain try to compensate for it. But this is why daylight lighting in motorway tunnels have much more light near the entrance than the exit. I was reading in the garden, in the sunshine, and came in before writing this. I was having difficulty seeing shadow detail indoors nearly two minutes later, so the time constant of 130 seconds is probably about the actual time it takes to recover most of the way. If we cut from sunlight on water to artificial light, we must limit the brightness of our highlights, or allow the viewer at least 20 seconds to adapt. If we are going from cone to rod vision, the rod recovery time constant is 510 seconds, so all the recovery times go up by a factor of about four.

It is not clear how motion pictures will adapt to HDR displays, as it means changing how we grade. Most home displays may stick to some 1000 nit peak white and compress their highlight brightnesses because that is what the mass-produced display technology can do. Cinemas may want brighter images and higher dynamic range at any cost, if only to outdo the home displays. Either way, it seems likely that we could have a single grade, and compress the highlights differently for different displays.

We have a similar limit on how dim it is sensible to go. We can see with cone vision, but it may take a minute to adapt, and 5 minutes to fully

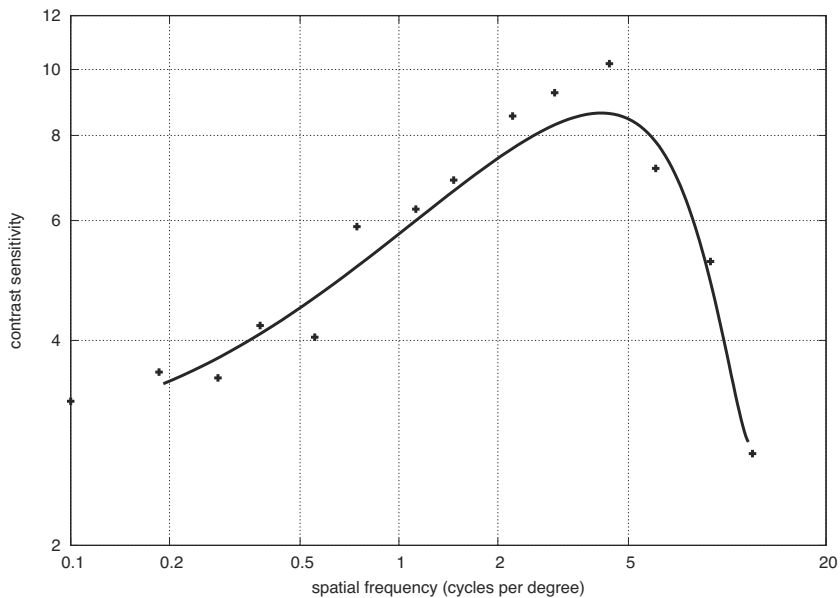
adapt. In fact, the cone cells have several ways of detecting light, and it can take 40 minutes for the most sensitive of these to stabilise. We can turn out the lights or have the searchlight turn on us or drama, but for most normal narrative action we should keep the average light level within the photopic range, and as constant as the narrative allows.

Even within the photopic region, you can have problems from too little light. At the old cinema white of 50 nit, you will see approximately the same colours if the image is slightly brighter or dimmer. However, if you go down to about 12 nit, skin tones look less pink, and less healthy. This is too bright for the Purkinje shift, so clearly there are other non-linear colour effects. I have worked with white points of 12 nit and dimmer: 3D systems often sacrifice a lot of brightness to split the signal between two eyes. I have noticed that at these light levels, people seem to disagree on what colours they see. It may be tempting to use the new dynamic ranges to get closer to the original scene light levels, but this may bring problems.

It is difficult to black out your theatre well enough to simulate night lighting levels. There may be a safety concern if the audience cannot see the way to the exits, though we react much faster to increased light levels than going the other way. Planetariums simulate the night, but the whole programme in a planetarium is simulating the sky at night, and they do not cut to daylight levels. This is probably not the sort of thing that would appeal to your local 7-screen multiplex. If anything, this might be something you would do at home with a head-up display.

Spatial resolution

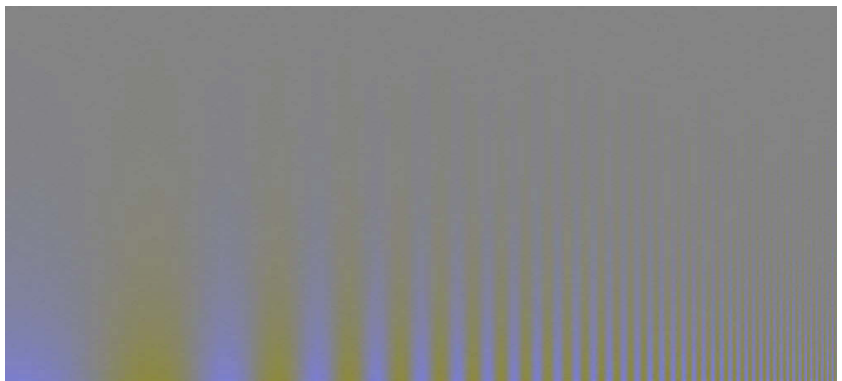
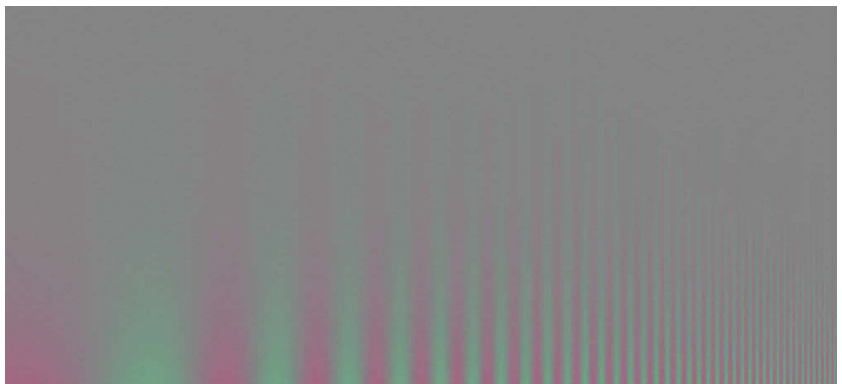
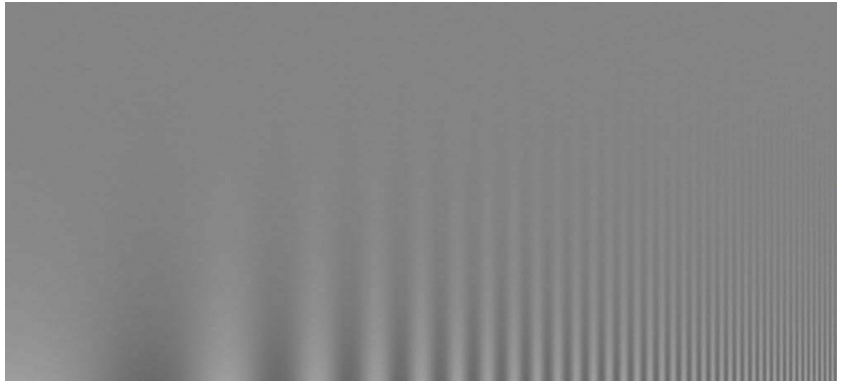
We can define the spatial frequency of a pattern to the eye in cycles per degree. We can present the eye with a simple stripe pattern, and vary the contrast until we find the point where the contrast is just visible. For a monochrome pattern, the eye's response has a peak at about 3 or 4 cycles per degree...



We would expect the resolution of the eye lens and the packing density of the cone cells to set an upper limit. It is less clear what is setting the lower limit, but beneath 0.5 cycles per degree we cannot fit one whole cycle in our central vision.

A popular trick is to make an image with spatial frequency varying in the horizontal direction and contrast decreasing in the vertical direction. If we mentally draw a line around the region we can see stripes, we should get a shape like the curve in the graph. There are three such images on page 155: one for luminance contrast, and two for the two opponent colour contrasts. The spatial frequency you see will depend on the size of your image and how close you are sitting. Try varying your distance from the screen. As you get closer, the shape will stretch out to the right. You

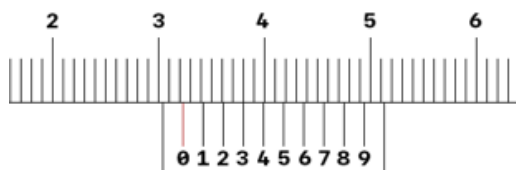
should be able to see the peak in the grey image. The coloured images do not have the same obvious peak.



We see less high frequency contrast in the red-green contrast than in the luminance. We see even less in the blue-yellow, which is reasonable as we have no S-cones in the macula. This is why video coding can save bandwidth on the chroma signals. But both colour contrasts are still visible in the low frequencies where the luminance is only a fraction of the peak.

Hyperacuity, Vernier acuity

The cycles per degree graph gives out at about 20 cycles per degree. The limit of the unaided eye is normally somewhere between 20 and 60 cycles per degree depending on how you measure it. Pierre Vernier published a method of extending the resolution of the unaided eye in 1631. You may have met his scale on the side of calipers.



Vernier Scale

We cannot easily see small features. However, we can see small errors in a large feature. In the diagram we have a Vernier scale. The zero on the red line lies somewhere between 3.2 and 3.3. It would be tricky to divide each interval by eye into tenths. However, on the Vernier scale we can see that the 3 mark on the bottom scale is the one that comes closest to making a continuous line with the mark on the upper scale. This would give us a measurement of 3.23.

This ability to see fine detail is called *hyperacuity* or *Vernier acuity*. We can sometimes resolve detail to 300 cycles per degree. There are several tricks here. On both sides of the Vernier scale we have a line. The eye can get the sideways position of the line by averaging the signals from many cells along the length. Our eyes have edge detecting cells in the retinal ganglia that do exactly this. The lines have a flat surround so our eyes are not distracted by texture or noise. We are looking for a small jog in the line: another thing our eyes do well. Finally, the line is surrounded by other lines with slightly different errors, so we can pick the one that is neither off to the left, nor off to the right. This way we can resolve up to 300 cycles per degree.

We do not need ten times as many pixels on our display. Hyperacuity is not part of everyday vision, but a special case of what the eye can do when the conditions are just right. However, if you had a LED screen that was built from tiles you would have to take extra care in aligning the tiles, as any dislocation in lines or text would be very visible.

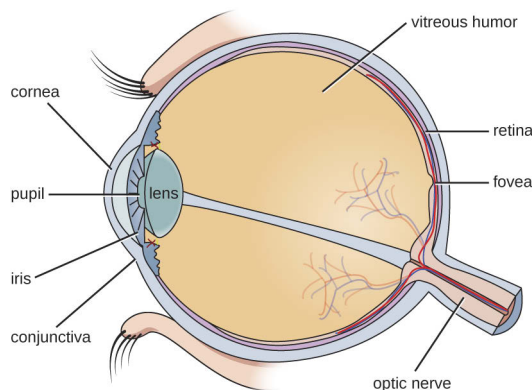
Chapter 8

More on vision

We have already discussed what light is and how we see, and how we make images to be seen. I shied from talking about the physiology of the eye-brain system because it is too easy to point to bits and explain what they do, which ends up equating eye with ‘camera’ and brain with ‘computer’. I am also not the right person to be telling you about this stuff. But, sometimes, a good teacher is someone who only knows slightly more than you, and remembers what it is like to not understand. In which case, maybe I am qualified after all.

Here is the cross-section of the human eye I said I wanted to avoid. In a later chapter we will even discuss ‘electromagnetic radiation’...

The human eye



The eye is a ball filled with gel and surrounded by muscles. There are muscles to enlarge the iris aperture, to pull on the edges of the lens to change focus, and others (not shown here) to rotate the eye in all directions. The dotted line shows the path of light from outside the eye to the fovea where we do all the precision viewing.

The blood vessels and the nerves to the retina come in at the back of the eye, and pass through a hole in the retina. This is the ‘blind spot’ we shall demonstrate shortly. The nerves and blood vessels fan over the top of the retina. Light has to pass through the blood vessels and nerves to get to the rod and cone cells. No, I have no idea why: an octopus eye has very primitive optics but at least their retina is the right way out.

Illusions

‘Illusion’ is a loose term. A mirage or a conjuring trick such as ‘Pepper’s Ghost’ is an illusion. Any stimulus that evokes an intriguing response such as White’s Illusion or Checkershadow is also an illusion. In this book on colour and perception, the illusions are mostly of this second type because these illusions tell us something about how we see. There is no clear boundary between the two: colour constancy experiments, where an apparently simple scene is actually something else, are like conjuring tricks. We don’t really need a rigorous definition of the term ‘illusion’, but it is good to check we use the word in roughly the same way.

The blind spot

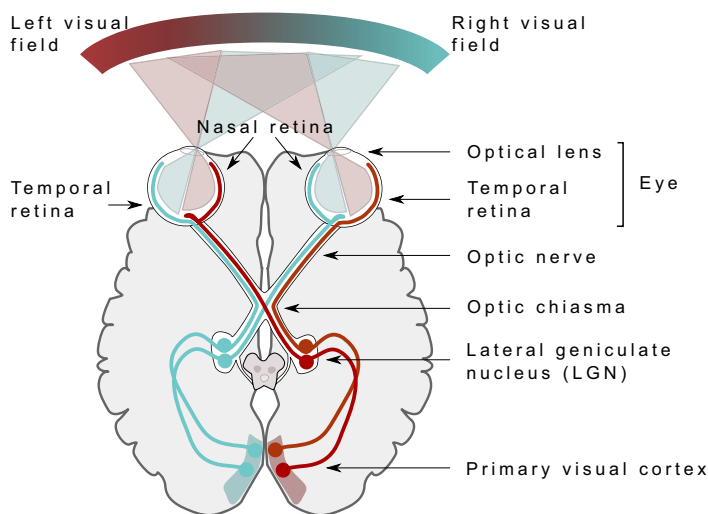
The eye can have no light sensing cells where the blood vessels and the nerves enter the eye. There is a region in each eye where we are quite blind. This is called the ‘blind spot’.

R

L

Put your head about three times as far away as the gap between the two letters. Close your left eye, and look at the letter **R** with your head about three times as far away as the gap between the letters. If you move your head back and forth you should find a position where the **L** vanishes. Close your right eye and look at the **L** with your left, and the **R** should vanish.

There is a gap in our retina, but we do not see it. We probably fill in with data from the other eye.



Human visual pathway

We do not know where this ‘filling in’ happens. The signals from the two eyes mostly cross over at the optic chiasma. There is some sorting of signals in the Lateral Geniculate Nucleus (LGN), and possibly some interpretation of stereo vision, so ‘filling in’ could happen there. The primary visual cortex is a more likely place for a sophisticated spatial operation like this.

Describing this as ‘filling in’ may be a bit misleading: it sounds like something is using the fill tool to flood fill the missing pixels in a picture editor. This is the homunculus fallacy. The ‘filling in’ white instead of the letters probably happens because the parts of our visual cortex that recognise things has never had any input for the blind spot.

Saccades and the importance of motion

Here is another illusion that is so simple that it is barely an illusion. Find something to look at, such as the star below. Keep your eye fixated on this for several minutes without moving. You can blink, but your eye should not move. But first, here’s what to expect. You may notice after a minute or so that your peripheral vision becomes hazy and loses contrast. Then the region slowly shrinks to a small circle in the centre. You know it is still

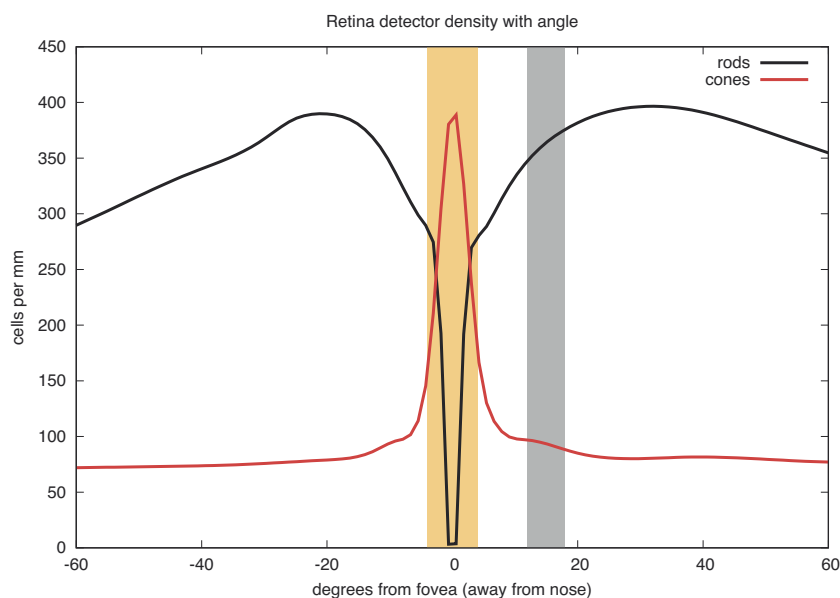
there. The moment you move your eye back it comes. But for the moment we cannot see it.



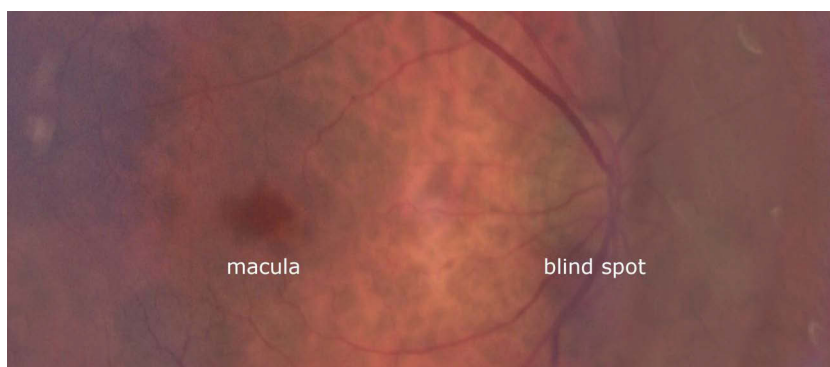
This is the Troxler effect. It is like the 'blind spot' but this time there is no 'filling in' mechanism. We can see with our peripheral vision, but the eye lens has aberrations, so the image is pretty poor. We can detect changes in brightness with our peripheral vision. We are probably better at distinguishing small changes in average brightness with our peripheral vision than our central vision because it has little image detail to distract it. However, if our peripheral vision sees 'something happening to our left' it does not try to interpret the signal: it triggers a strong urge to glance to our left, so our central vision can look at whatever-it-is properly. This is not wholly unconscious: we can control where we look just as we can control our breathing, but it is usually automatic.

Central and peripheral vision

Here is a plot of the cell density on our retina measured out from the centre of gaze...



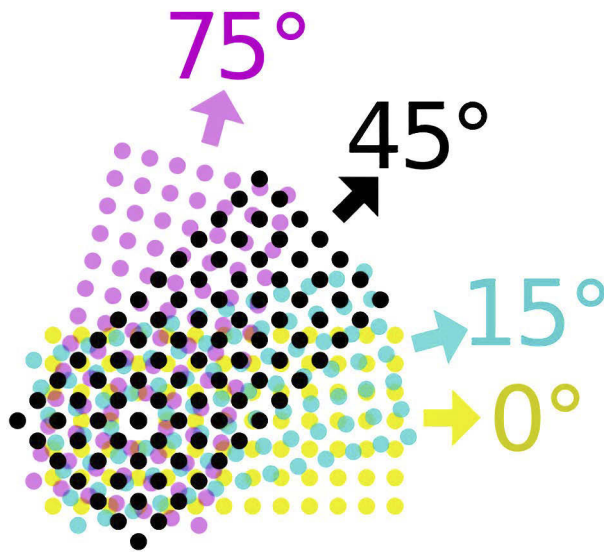
Our colour vision is the red curve; our night vision is the black curve. Most of the colour vision cones lie within one degree of the visual axis. The retina has a pit shape called the fovea, where the cells lie at an angle rather than go straight up-and-down to pack more in. On top of this we have the macula, or 'yellow spot', shown by the orange bar. This is a region of yellow dye that absorbs short wavelengths, and may protect our detailed vision from high exposures. The grey bar shows the approximate region of the 'blind spot' where the optic nerve comes out.



Here is the retina of my right eye as seen at my last examination (Specsavers, St Albans). The macula where most of my detailed vision starts is the rather dull red smudge on the left. The blind spot on the right is where the optic nerve and the blood vessels come in is the lighter circle on the right. The blood vessels lie on top of the retina. Notice how the larger blood vessels surround the macula but do not cross it. The lighting is not even in this image; away from the obvious features, the retina colour is pretty uniform.

The macular dye would stop us seeing blue, but we do not have any blue-sensitive S-cones in the fovea anyway. [This sentence looks blue because it is wide enough to go outside the fovea.](#) The period after 'anyway' just before the blue sentence was also blue, but you probably cannot see that without zooming in.

We cannot see fine detail in yellow on white, or deep blue on black. CMYK screen half-toning takes advantage of this...



CMYK printing screens

Where two halftone screens overlay we get artefacts called moiré or rosette patterns. The cyan, magenta and black screens are at 30-degree intervals, which keeps these patterns from being too visible. The yellow screen is less visible at small scales, so it can be squeezed in at 15 degrees in between C and M. The Y screen is often 8% larger than the others too.

In Chapter 3 we introduced the *three-colour approximation*: the assumption that our vision was a bit like a camera with three channels. We now see we have no S cells in our central vision. The surrounding vision regions have all three LMS detectors, but have no macular dye, so the LM spectral sensitivities here do not match the central vision.

In fact, the three-colour approximation still works. We seem to make a three-channel vision model out of the best LMS data. We can break this model if we make two white colours with the same calculated LMS values but different spectra, but this is pretty hard to arrange, and it hardly ever happens unless we deliberately set out to do that.

The vision pathways

The signals from the eyes get to the visual cortex at the back of the brain via a body called the Lateral Geniculate Nucleus (LGN). We have one on either side of our head. The LGN on our left takes the left half of our visual field from either eye, and the LGN on our right takes the right.

The LGN cross-section has six distinct regions. Two of them have large cells, so they are known as the magnocellular or M-cell pathways. Four of them have small cells, so they are known as the parvocellular or P-cell pathways. There is an additional set of very fine cells, known as the koniocellular (after the Greek word for dust) or K-cells. These were overlooked for a long while, and you will not find them in old books. They fit in the gaps under the six obvious regions.

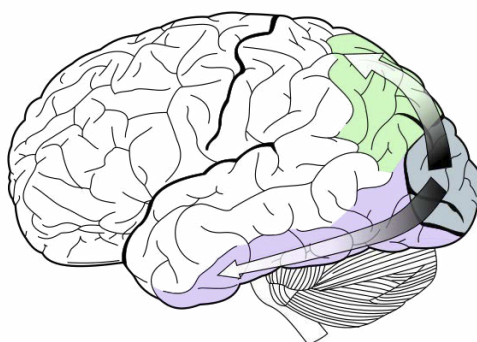
As far as we can tell...

The M-cells carry light/dark contrast information, depth, and motion information. Their signal comes from both eyes. They are sensitive to low spatial frequencies. They can react rapidly to changes with time. They are used when tracking things, resolving edges, and interpreting motion.

The P-cells carry colour and high resolution texture information. Their signal only comes from one eye, so they probably have no depth information. They do not react as fast to changes.

The K-cells are still being researched. They may carry much of the blue channel colour signal. They may be responsible for 'blindsight' phenomena where we react to moving things without being aware of them. The S-cone and K-cell mechanisms are somewhat separate from the rest of the visual system, and some researchers suspect they may be much older.

The M-, P-, and K-cell signals end up in the primary visual cortex at the back of the brain (shown in grey). The information is then processed along two separate pathways...



Ventral and Dorsal pathways

The ventral stream (lilac) is sometimes called the “What” pathway. It is associated with recognising and identifying things, and with long term memory. The dorsal stream (green) is sometimes called the “Where” or “How” pathway and is associated with spatial relationships between objects, and with muscular actions, particularly when coordinating arm

and hand motions, and eye saccades. It has no obvious connection to long term memory.

Each of these streams has a set of vision centres. Here the various M-, P-, and K-cell layers come together. The ventral stream handles old functions such as face recognition and interpreting expressions; and some newer functions such as reading. The vision centres for these functions do not seem to be 'hard-wired' to particular functions, and they cannot be 'hard-wired' for more recent inventions such as reading. However, brain scans show most of us do the same thing in the same place.

The ventral stream gets most of its input from the P-cells. It handles colour and detail, mostly from the central vision. But the ventral stream has a centre for a face and body recognition, and this hardly uses the colour information at all. The part that helps with reading also does not use the colour information. It seems we are quite happy with seeing faces in black and white, just as we are happy reading black and white letters and symbols. Indeed, if we are only using the M-cell signals, we may get a stronger signal from a monochrome signal than we can get from a colour signal.

Many artists prefer to work with black and white. There may be good reasons for not adding colour to 'Casablanca'.

Synaesthesia comes from the ventral pathway. We have not found any use for it so far. It may be some idle cells randomly interpret the P-cells and perhaps K-cells when the rest of the centre ignores them. As well as people associating colours with numbers and letters (I do this a bit), there are also people who associate colours with facial expressions and moods.

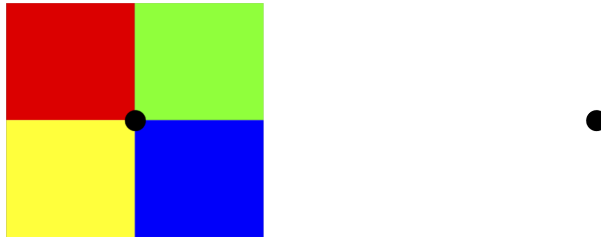
I would not say all this is proven science. People currently agree that the two pathways exist, but I would not be surprised if chunks of this was re-written in the next twenty years. But we are beginning to get a picture of how parts of this might work.

Note:

If you read more on this, you may find the retinal ganglia has 'parachute' or p-cells, and 'midget' or m-cells, named by their shape. The p-cells connect to the M-cell layer in the LGN, and the m-cells connect to the P-cell layer. Confusing, isn't it?

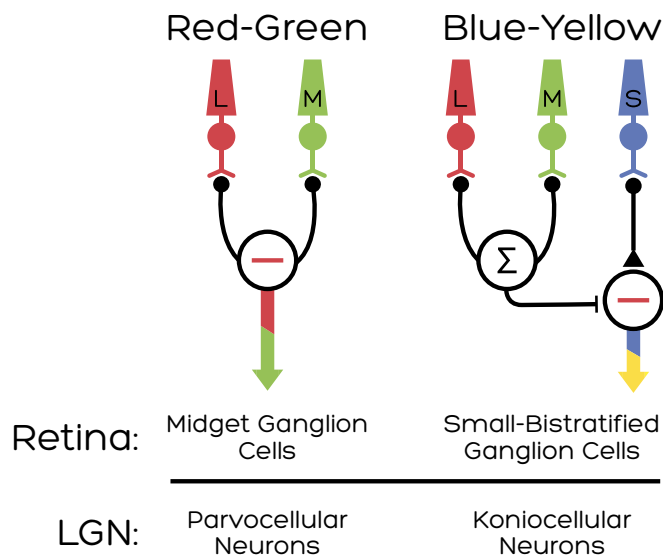
Hering opponent colours

If you stare at the left-hand dot for 30 seconds, and then switch to the right-hand dot, you should see some after-image colours.



If the after-image colours were due to adaption of the individual cones, then we might expect red to give cyan, and green to give magenta. For most people (but not everyone) it seems that red gives green, and green gives red.

The retinal ganglia have cells that generate opponent colour signals. There are red-green opponent cells that are excited by signals from red cones and inhibited by signals from green cones, or vice-versa. There are also yellow-blue opponent cells. There are also luminance cells that are stimulated by red, green and blue, that should give the standard photopic observer response.



What happens if we look at a flashing white light? The LM signals pass through an extra cell, so the LM sum may arrive at the opponent cell after the S signal, which may give us a blue-yellow contrast. The LMS cells have different reaction times, so we could get a red-green contrast too. In fact, a flashing monochrome signal can give the sensation of red, green or blue

colours. This is known as the Fechner colour effect, or Benham's top after a device that gives particularly good colours.

The visual cortex may make some opponent signals. If we flicker between two colours with the same luminance at more than 5 Hz we do not see the separate colours, but fMRI studies show that the visual cortex reacts differently.

Researchers use flicker effects to investigate neural processes. Consider the red-green opponent summation. Suppose we have regularly flickering red and green stimuli at 10 Hz. We would not expect to see anything but the luminance variation. However, suppose we can skew the phase of the stimuli, so the red flashes come before the green flashes in one experiment, and after the green flashes in another. Would we expect to see the same colour? In fact, people can see small colour shifts (see for example Stockman, A., Henning, G. B., West, P., Rider, A. T., & Ripamonti, C. (2017)). This is illuminating, but it is not an important part of our vision. Hearing uses phase signals more than vision: cells stimulated by one ear and inhibited by the other determine which ear is closer to the sound when there are no other cues.

I don't know what we can take away from opponent theory. After-images and colour flicker are not an important part of ordinary colour vision. We see yellow as a distinct colour, while azure is a less obvious band between green and blue. This seems reasonable as yellow is a balance between the L and M cone signals and both cells are in the fovea, while azure is a mixture of M and S and there is no S in the fovea. Also, the S-cone signals are the odd one out because they are carried in the K-cell pathway, while M and S only go via the P-cell pathway. It seems reasonable, but I cannot see how this helps us grade motion pictures.

Adaption illusions

The Hering primaries can be explained by an adaption mechanism that lies close to the retina. The 'Spanish Castle' illusion, where pre-exposure to a reversed colour image adds colour to a black and white image, can be explained the same way. So can the 'lilac chaser', where a moving gap in a ring of green dots has a complementary moving dot provided your eye does not look at it. These are not shown here as they need animation, but a web search will find working versions.

Aristotle noticed that the eye and brain adapt to motion. If you look at an object that is in steady motion, such as a waterfall, then when you look away everything briefly seems to be moving up. This is harder to explain as a simple cell-level adaption, as our eye can track the motion in the waterfall.

In 1999, Webster and McLin (*Psychonomic Bulletin & Review* /999.6 (4).647-653) demonstrated a similar adaption to human faces. If the

viewer looked at a distorted face for 30 seconds, and then at an undistorted face, they briefly saw the second face as having the opposite distortion. The effect does not work if the second face is distorted. It seems to be that we continuously calibrate our sense of what a 'normal face' looks like. Other researchers have seen similar effects for adaption to happy and sad faces, male and female faces, sharp and blurred faces, and other contrasts. Many of these seem to use signals that only exist in the ventral 'what' visual pathway.

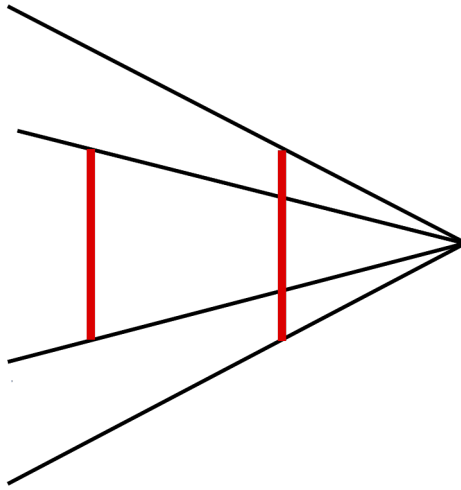
There is another bizarre illusion, called the 'flashed face distortion effect', which may be related. The observer concentrates on a central cross, while images of faces flash up either side. After about 30 seconds, these faces can seem dramatically distorted provided you remain staring at the cross. This may be a side effect of starving the hypothetical 'normal face' calibration of proper data, but this is still poorly understood.

There seems to be a class of illusions that (a) take about 30 seconds of adaption to set up, (b) give an illusion that counters the previous effect, and (c) last only a few seconds when you let your eyes move. That is not to say there is a common process or a common centre in the brain that creates all these illusions. If you put your left hand in cold water, and your right hand in hot water for 30 seconds; then put both hands into tepid water, it will feel hot to your left hand and cold to your right: this seems like another adaption illusion, but it does not involve vision at all.

Illusions without adaption

There is another set of illusions that work without adaption. White's illusion in the introduction needed no staring at a black dot before we could see the colour difference. The crispening effect in Chapter 5 may look like an adaption to reduced contrast, but it is hard to detect any adaption time.

There are illusions that use 3D cues, such as the Ponzo illusion, the Ames Room, and the Müller-Lyer illusion.



Ponzo illusion

None of these have any obvious adaption time: we cannot instantaneously compare the lengths of the two red lines by eye, but as soon as we do the line on the right looks longer, and stays that way.

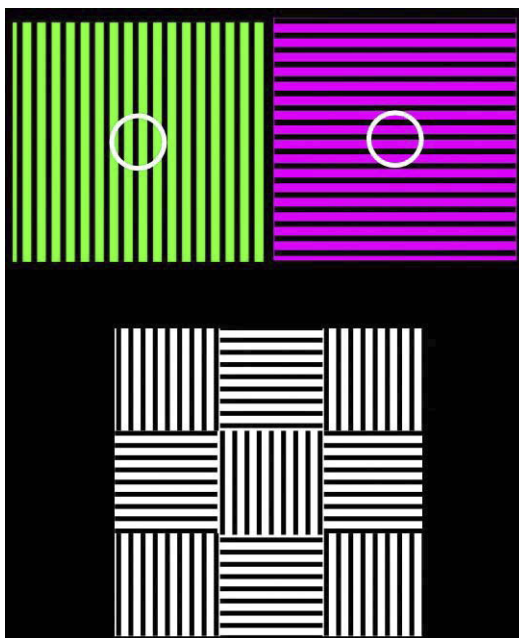
So, what can we take away from all this? There is a subclass of visual illusion effects that seem to have a characteristic 30-second adaption, and a brief (a few seconds) effect at full strength before a rapid decay. If you see an unexpected colour shift, experiment to see if you can find these characteristic times. Does it take some time to appear? Does it fade if you move your eyes? I remember being asked to fix a pink colour shift in a sequence that had a slowly moving shot up a country path at night. I could see the shift, but it only started 15 seconds into the sequence. It was not an error in the colour calibration, but probably some relation of the 'lilac chaser' illusion keyed by the surrounding greenery.

The McCollough effect

The McCollough effect is another effect that has been linked with the opponent cells. This is a spatial effect, but it is of interest to us because of its surprisingly long lifetime. You can try it, but I would just read the description if you plan on grading any images today.

Concentrate on the two coloured patches, tracing out the circles slowly with your eyes. Switch from one patch to the other every 15 seconds or so. After about five minutes, look at the neutral patch below. You should see the vertical stripes tinted pink, and the horizontal stripes tinted

green. This may appear as a subtle fringe next to the bars rather than a uniform tint.



The McCollough effect

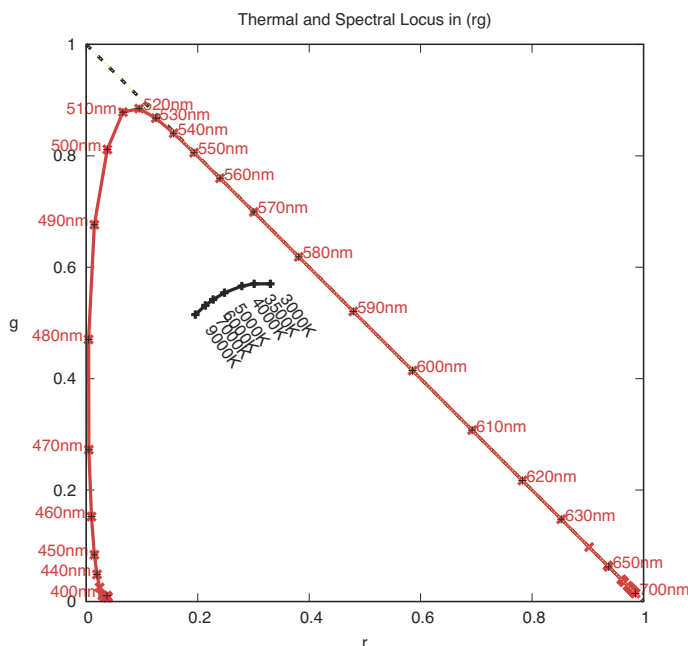
This is not one of your illusions with 15 seconds adaptation, and a few seconds at full strength. The effect is subtle, but it can last for hours, sometimes for several days. People have even reported seeing the colours two weeks later. It seems likely that the process happens in the cells immediately behind the retina because this effect is modified by drugs that affect inhibition in the visual system. There seems to be no purpose to this, and no explanation for the unusual timescales. I have weak synaesthesia that associates colours with letters, and this feels a bit like that. Whatever the cause, there is not much you can do other than to avoid rooms with unusual wallpaper.

There may be other effects like this for all we know. The recent flood of good, new illusions mostly appear as online videos and animations, and effects with longer timescales are much less explored.

Hering opponent based colour spaces

Suppose we want to have a colour space with red-green and blue-yellow axes and neutral in the middle. This sticks the big four named colours on the points of the compass, which seems attractive. But can we square

this with additive mixtures, colour constancy, and the colour spaces in Chapter 3 and Chapter 4?

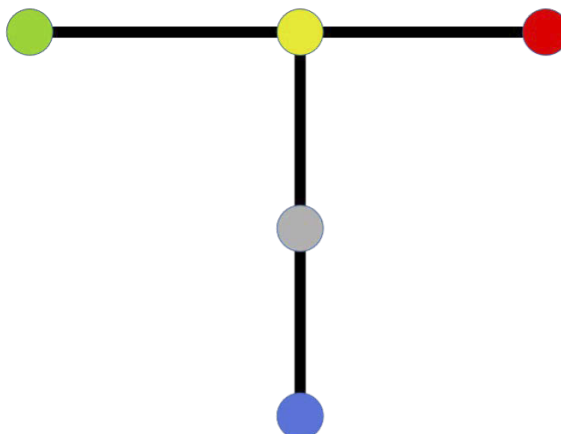


The r,g colour space

There is nothing special about the r,g space for this experiment, other than it does give additive mixtures. Any colour space with this property would do, but this is the one I was using earlier, and the geometry is simple. We have not set a white point, but we have plotted the thermal locus, which is a sensible place for it to be.

In the r,g space, the red-green runs from bottom right to top left. The colour we see as yellow is at 579 nm. The colour we see as blue is 470 nm. A line from 579 nm to 470 nm is almost perpendicular to the diagonal; the thermal locus is a bit off this line to the green side. It is arbitrary which point we take as neutral, but something like D65 daylight should not be far wrong.

If we rotate by 45 degrees as we do for tLab, the red, yellow, green, blue and neutral points form a T-shape like this...



Hering opponents

The Hering opponents are straight lines, but the G-R opponent does not have neutral in the middle. This is not our neutral in the middle, surrounded by the red, yellow, green and blue opponents. If the eye's LMS were three equal channels, and our after-images formed as a photographic negative, then the L-M opponent colours would be red and cyan, with neutral in the middle. But LMS are not three equal channels: the S-channel seems to be the odd one out, as we shall see in the next section.

So, do we want to bend the colour space to get red and green level with neutral? You are welcome to try, but I feel this must be a bad idea. Bending the colour space means we no longer have additive mixtures. Additive mixtures are part of everyday vision. We have additive mixtures with our original LMS exposures within the cone cells. Colour constancy in the visual cortex uses visual mixtures. We use the linear properties of additive mixtures all the time, both in vision and in books like this. Meanwhile, we are almost unaware that our cone signals get transformed into R-G and B-Y opponents within us just as TV signals are separated into luminance and chroma channels, and back again. We can only see this by fixating on a highly coloured test image for thirty seconds, then staring at a blank; and fixating is not really an important part of everyday vision. And when we see an after-image, we usually know it is not real.

Evolution of primate vision

S- and L-cone cells are present in our eyes, and also in jawless vertebrates such as the lamprey. The last time we had a common ancestor was 500 million years ago. The rods may have evolved from the S-cones when mammals were nocturnal. If mammals went nocturnal to avoid dinosaurs, this happened about 200 million years ago. Finally, the L-cones split into L- and M-cones. Old World primates have three-colour vision, and New World primates do not, and these populations split between 30 and 40 million years ago.

The structure of our retina backs up this story. Our S- and L- cones were there in the ancient eye. Perhaps L came first and claimed the visual centre, then S filled in the gaps. Our S-cones look identical to our L- and M-cones, but they can be identified under the microscope by their connections: they are less common than the other cell types, and more highly connected than the L- and M-cones. The rod cells have a separate circuitry that filled in the gaps left by the cone cells. Finally, the L- cones split into L- and M-cones. These and their connections are almost indistinguishable under the fundus microscope unless we can image the dye colour.

In 2009 Katherine Mancuso from the University of Washington gave functional full colour vision to adult squirrel monkeys using gene therapy. Squirrel monkeys are a New World species with no history of three-colour vision. This suggests there is little hard-wired difference between L- and M-cone cells, and no hard-wired circuitry to process the new signal.

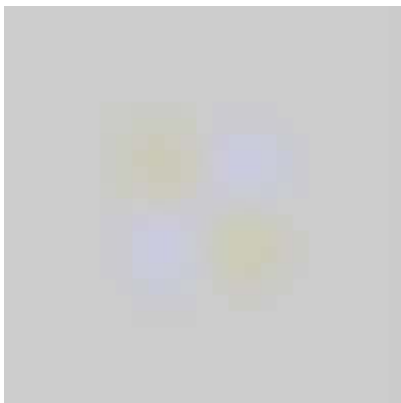
In our LMS vision, S seems to be the odd one out. S-cones are our oldest unaltered cones. They have no genetic variants. Actually, that's not quite true — there may have been a UV version 90 million years ago that could see down to 300 nm, but that's long gone. They had their own K-cell opponent wiring long before L- and M-cones were different things.

What about colour constancy? We could have had a 2-channel form of colour constancy half a billion years ago. In practice, the difference in dye bleaching between the L-S cells might have done the equivalent job well enough. Colour constancy only becomes complicated when we add the third channel, and our visual cortex takes over the miraculous (to me) reconstruction of the linear light levels.

Seeing polarization: Haidinger's brushes

We can detect polarization with the naked eye. If we look at a white region of an LCD, and tilt our head, we may briefly see a figure like the diagram below in the corner of our eye. This figure is known as Haidinger's brushes because the yellow regions resembled the head of an artist's brush. It is caused by polarized light from the LCD interacting with

the anisotropic macular pigment. It is not usually a serious problem when colour matching provided you recognise it and avoid it.



Haidinger's brushes

Haidinger documented this phenomenon in 1844. He discovered it when experimenting with polarizing minerals but found you could also see it in the naturally polarized light of the evening sky at 90 degrees from the sun. The polarization is horizontal, instead of at 45 degrees as for an LCD screen. The yellow 'brushes' point towards the sun.

This is a subtle feature, but it is easier to spot if you can make it move. If you want to learn to spot them, I suggest getting a featureless patch of white on an LCD screen, and then nodding your head while staring straight at the white patch. Look for colours. Other people suggest looking at the sky through a polarizer, and wiggling the polarizer.

This has been suggested as the secret behind Viking navigation when the sun is behind cloud or beneath the horizon. They may have also used the natural birefringence in materials such as tourmaline or calcite ('sun stones'). In the 1950s, SAS transpolar flights navigated using a Pfund skylight compass, which is effectively the same thing.

Seeing magnetic fields: cryptochromes

There is no trustworthy evidence that we can see magnetic fields, but this is one more thing we might do. Cryptochromes are a class of proteins that are sensitive to blue light. They are not obviously related to the opsin proteins we use for vision. They help maintain the circadian rhythms in plants and animals, and they help plants point towards the light.

It also seems possible that cryptochromes may allow animals to sense magnetic fields. Birds, but some mammals, fish, reptiles, even crustaceans and insects seem to be able to sense variations in the earth's magnetic field. The European Robin can migrate large distances. The

British and Irish robins mostly don't bother, but the continental ones do. Experiments have shown robins can point their beaks toward Africa using only the Earth's magnetic field, but only when they could see. This magnetic compass seems to be light-dependent, located in the eyes, and particularly the right eye. Thorsten Ritz et al. (Nature Vol 429 13th May 2004 p177) showed that adding a second field varying at 7 MHz destroyed this inner compass at particular angles, suggesting some radical-pair mechanism. This second field was less than one-third of 1 percent of the Earth's field. If they can sense this at the cellular level, the detection must be extraordinarily precise.

The detection system may use quantum entanglement as used in quantum cryptography. It seems blue light reacts with the cryptochrome molecule to generate an *entangled pair* of free radicals. A free radical is an atom or molecule with an unpaired electron. These two free radicals start off with opposite spins. Free radicals are unstable and highly reactive: they can recombine to pair up their electrons again, or they can react with whatever else they meet. Their spins become randomised with time because the two free radicals are in slightly different positions, and so are in slightly different magnetic fields. The free radical pairs can recombine more easily if they have opposite spin. This pair recombination in cryptochromes does not generate a nerve signal directly, but can influence the processing of other light signals, much like the intrinsic photosensitivity of some retinal ganglion cells as described in Chapter 3. This would explain why it only works when the robin is using its eyes. A robin may see a magnetic field as a modulation in the cells in the right eye.

This may sound strange, but it does make a sort of sense. Cryptochromes were first researched in cryptogams – plants that have no obvious flowers or seeds, such as mosses, ferns, algae, and fungi. This is where the 'crypto' part of the name came from, though they may be hiding more. They are a very old component of life, and common to fungi, plants, and animals. The eye came later, and was honed by selection to a very sensitive organ, working closely with the brain.

The Vikings may have used the weak polarization that comes from the structure of the eye lens to navigate, elevating this from a bug to a feature. It seems reasonable that birds may have exploited a side effect of a chemical originally used for circadian regulation, but with a side effect that affected the visual signal, and done much the same thing. If cryptochromes give a response to magnetic fields within the eye, it seems likely evolution will exploit it for navigation for species that need it.

Colour correction and 3D

People teaching art often say “warm, bright colours advance; cool, muted colours retreat”. Literal people like a younger version of me would put a set-square on the image and say “no, they don’t”. It took me some time before I could get past the words and understand what they are actually saying. I think it is this: warm colours are usually in the foreground, and pale blues and muted tones are usually in the distance; if there are no better cues we may use colour to guess depth. Josef Albers in his ‘Homage to the Square’ played all sorts of tricks to explore this, balancing retreating squares implying distance against bright central colours.

There is also a trick within this trick...



Raphael, 'Woman with a Unicorn'

Here is Raphael's 'Woman with a Unicorn'. The foreground is bright, the mountains in the distance are blue and hazy, and the nearer hills are somewhere in between. Our vision uses many cues to determine 3D shape. The strongest 3D cues come from occlusion: the woman is in front of the wall because she obscures it. If this was a moving image, another strong cue comes from motion parallax, or how things move when our viewpoint moves. Nearby things also look big, and far away

things look small. Parallel lines converge to vanishing points. Stereopsis, the traditional 3D movie effect, is quite weak when compared to these, and it only has effect over a limited range.

There is a weak form of occultation, where the distant hills are partly occluded by the scattering of blue by the atmosphere. This is not a strong cue, but it has the advantage that it is interpreted by our peripheral vision. Our eyes may go straight to the woman's eyes. We may have not noticed she is holding a unicorn yet, but our peripheral vision already sees the surround with no strong features fading into pale blue, so Bang! we get an impression of huge depth. Nice.

We expect to see haze with great distances. Raphael may have just painted what he saw. But it also works close up. Here is Gustave Caillebotte's 'Young Man Playing the Piano'...



Gustave Caillebotte, 'Young Man Playing the Piano'

Compare the contrast at either end of the piano keyboard. Or the near and far piano legs. We would not expect to see this fading with distance. Caillebotte, in projecting this image onto his 2-D canvas, knows he loses the 3D sense we would have had from motion parallax, or depth of focus, or stereo vision. He replaces these cues with another cue: a virtual haze, which restores the sense of depth. We see the chair, the man, the front corner of the piano; and we sense the depth of the far side of the piano and the other chair before we have really seen them.

Motion pictures use smoke on sets to provide a sense of depth as well as 'atmosphere'. The depth of focus is also often kept shallow to separate the action from the foreground and background. 3D cameras provide stereo depth, but they are often heavier and more fragile than ordinary cameras; so they move less, and we lose the motion parallax cues. But if

we show a motion picture produced for 3D in an ordinary theatre, we could replace the stereo with a virtual haze and narrow depth of field generated from the depth information.

Stereo: the camera adds ten pounds

Sir Charles Wheatstone invented the stereoscope, and with the help of his friend Fox Talbot, showed the first stereo photographs. Here is a stereo portrait of him rendered as an anaglyph. I have aligned the separations on the bridge of his nose. We can see a cyan line on the right-hand side of his head. Our right eye is seeing further around the right-hand side of his face than the left. Our left eye is presumably showing a similar effect on the other side, but it is not showing up as well.



Stereo portrait of Sir Charles Wheatstone

If Sir Charles was looking in a mirror, his left eye would see a bit further around the left side of his reflection, and the right eye around the right side. A central camera lens will not see as far around the face as the fusion of the left- and right-eye images. The camera image of the face will therefore seem to bulge toward you. If you want to take a flattering portrait, take the picture from further away. If you are far enough away, your eyes and the camera will see almost the same thing. But there is probably no lens that can match the fusion of the stereo images that happens in our heads when we look at a nearby face. It might be possible to get close by digitally warping a central image using depth data.

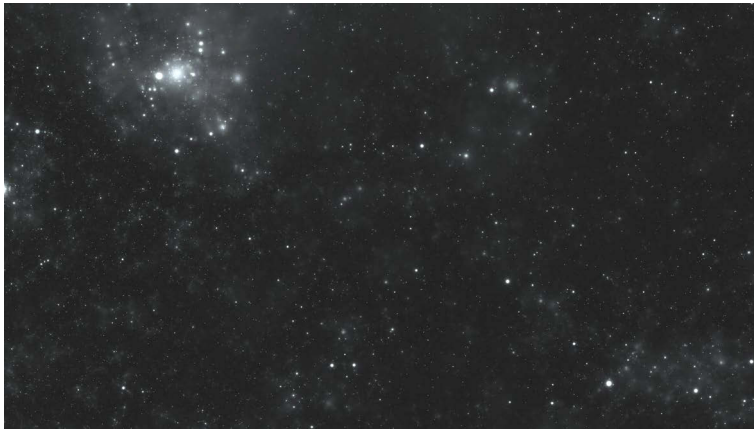
A camera may not see the same subsurface scattering that we do because the geometry where the light grazes the side of the face has changed. This may seem trivial, but we look carefully at faces when we are talking to someone or looking at our own face in a mirror over similar

distances, and we are good at detecting these small differences. See Chapter 10 for more on subsurface scattering.

Highlights and flare

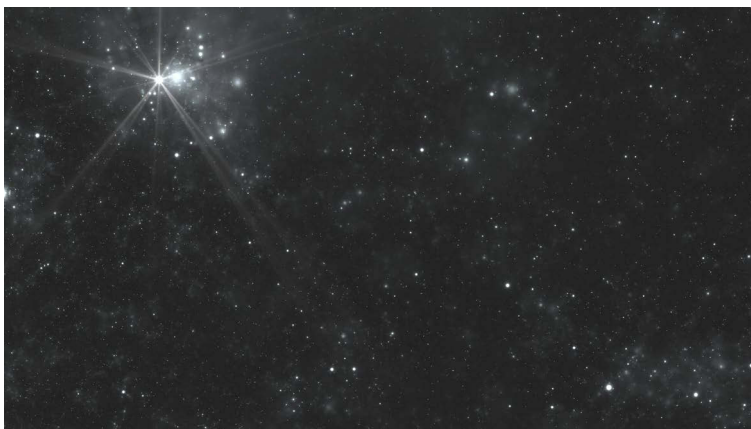
In the last chapter we discussed how we might use the extra brightness on HDR displays if we could get a lot more light. We could have dazzling highlights, but this would leave us with light trails on our retina and after-images. Our eyes might take minutes to recover after a bright scene. Even if we could afford to match the brightness of outdoor scenes, the extra brightness may dazzle us.

Here is an artificial field of stars...



There are some brighter stars up in the top left. Actually, they are not brighter at all — everything has the same white, so the ‘brighter’ stars are actually larger circles. Suppose we wanted to make one of the stars look a lot brighter than the others without using HDR brightnesses. What would we do?

I added artificial flare to one star...

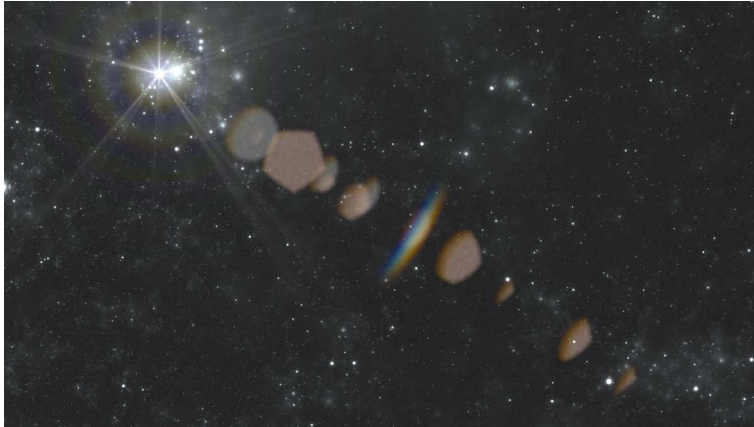


This is perhaps not the star you might have thought was the brightest one in the previous picture: I avoided that one because it seemed to have some flare of its own from the surrounding nebula, and chose the slightly 'dimmer' one to the left of it.

The flare is a set of radiating lines such as you might get from diffraction around the edges of a 5-sided lens aperture. Telescope images often have four flare lines from the supports for the secondary mirror. The flare points to the central star; it draws attention to the right place. But the star continues to look bright, even when we have found it and can see that it is made from the same white as everything else. We get the sense we ought to be screwing up our eyes to look at it.

We have no 5-sided aperture in our eyes. The actual flare we see with the naked eye is much harder to capture in an image. As I write this, there is a car parked on the other side of the road. I get a bright, specular reflection from the bodywork if I glance out of the window, which leave me with lasting spots before the eyes. The flare I see seems to have some radiating lines and some rainbow colours, but the shape is continuously changing as my eye moves. Artists represented star brilliance with radiating lines long before they used lenses.

Here I added an Airy disc about the star, an arc of a halo, and a train of ghosts such as you may get from internal reflections within a complex lens...



I have overdone this to show the various elements. The Airy disc is the coloured circles around the star: this adds to the sense of brilliance. The ghosts do not add much in a still image, but when our lens moves, the train of ghosts pivots about the star, and the association is much more marked. They would also indicate there was something bright in a particular direction even if the star was out of shot. You can use this to show where light is coming from without showing the source itself.

There is something complex happening here. Artists in antiquity drew stars and the sun with rays but without ghosts. Simulating the aberrations of a complex lens train is new. We may be using what our eyes have learned when we use binoculars and telescopes, or we may have learned everything from images. It is not what our eye would see if we were there. And yet we seem to relate to what the camera 'sees'.

The current fashion for obvious lens flare may pass. However, we want to show bright things in an image without actually being very bright, and lens flare seems to do just that.

Chapter 9

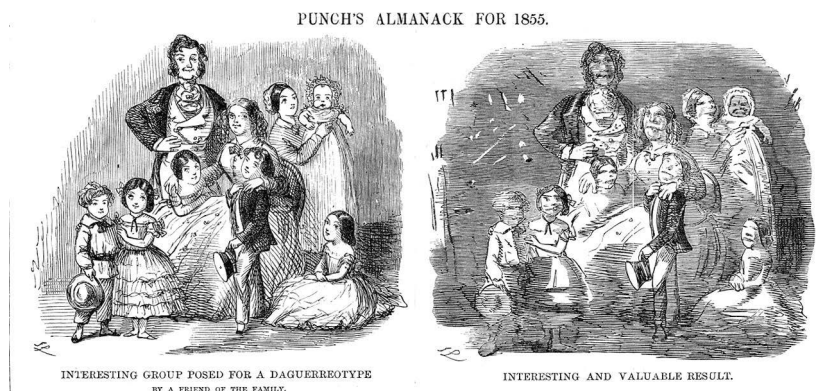
Photography and cinema

The Truelight film emulation shown in the diagram in Chapter 6 may give the idea that you can stack up seven colour transforms and get a match. The devil is always in the detail. In this chapter I will describe some of the more awkward details, and how we coped with some of them. With luck, you may never have to bother with any of this, but it is better knowing then thinking it is easy.

Light sensitivity of early emulsions

The photographic process is not one invention but many. Many people noted that light caused silver halide crystals to go dark. A developer could reduce a silver halide crystal to silver even if it had only been exposed to four or five photons. A fixer could remove the remaining silver halide to stop any further darkening. If your photographic emulsion is in gelatin, a buffer in the fixer keeps the pH at the point where gelatin is neutral, at which point it rejects water and becomes tough. All this could be done with stable materials, so you did not have to prepare the photographic chemistry on site.

‘Punch’ in 1855 lampooned the result...



‘Punch’ cartoon by John Leech, 1855

This is an acute sketch. The plates were insensitive to red, and pretty insensitive to green. Pink complexion came out dark. Lips came out darker still. The exposure took minutes, and any movement blurred the image.

In that year, 1855, James Clark Maxwell suggested that we could reproduce colour images according to Thomas Young's three primary model: expose three plates through red, green, and blue filters, and then project these three plates through red, green, and blue filters. The photographer Thomas Sutton prepared the plates. The red filter would need a long exposure, so he picked a still subject: a coloured ribbon with a fancy bow. After some fiddling, he got what looked like three separate colour signals which could be combined to show something like the original colour image. Maxwell showed the result at the Royal Institute as part of a lecture on vision.

Many wondered afterwards how the red exposure was done. There should have been almost no sensitivity to red. About a hundred years later, the riddle was explained. The red dyes on the ribbon reflected both red and ultraviolet. The red filter on the camera passed both red and ultraviolet. The emulsion could hardly see the red, but it responded to the ultraviolet. The first colour photograph had three channels of blue, blue-green, and ultraviolet. But the principle was sound, and it did give a colour image that looked sort-of right, even if the red plate wasn't quite what you thought it was.

In the early days of motion pictures, sensitivity to red and green was still a problem. You could take a picture with a yellow or red filter to balance the exposures, but this wasted a lot of light.



Stan Laurel

Stan Laurel wanted to act in films, but he had light blue eyes. This was attractive in real life, but he came out on the very early blue-sensitive

emulsions as having no pupils at all. He tried various make-up tricks but to no avail, so he went into directing instead. After a few years the emulsions had improved, he had normal eyes in the movies, and he could go back to acting.



Kodak advertisement

Orthochromatic film was still insensitive to red, so lips came out dark, but this could be fixed with make-up and camera filters. Panchromatic film used dyes on the silver halide crystals to extend the sensitivity into the red.

Intriguingly, this 1920s advert still lists 'indigo' as a colour as Newton did. The make-up all had to change for panchromatic film. A guy called Max Factor was one of the first to get into that.

Colour balance for black and white film

We still want the same colour balance in black and white films as we would get if we were there ourselves. Black and white cameramen used to shoot with no filter, or a 'straw-coloured' filter at dawn and dusk, and progress through yellow, and orange, to orange-red at noon, and back again towards sunset to keep faces looking 'natural'.

Film exposure units

I shall skip this bit. Here is a picture of the light meter that my father and I used...



Light meter

The selenium sensor is face downwards in this image. The meter gives the light level in some internal units. If you added the plastic diffuser, the scale on the meter flipped to another set of numbers. The circular bit is a slide-rule-like calculator. You could enter the film speed (ASA or DIN), the exposure time, and the focal ratio, and it told you what the meter ought to read for a good exposure.

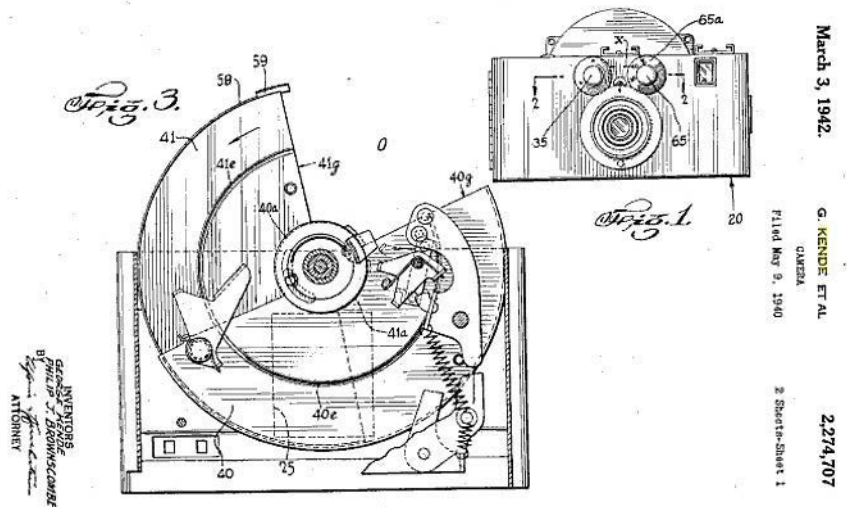
The focal ratio is the ratio of the camera lens aperture to the focal length. You can perhaps read the values on the outside of the disc. The values are $f/8$, $f/5.6$, $f/4$, $f/2.8$, $f/1.4$, $f/2$, $f/2$ and so on in ratios of $1:\sqrt{2}$. If you go to the right in this list, 'going up a stop' doubles the amount of light going through the lens. The exposure times are also marked in approximate powers of two, so it is easy to trade off stops for exposure time.

There were many exposure units if you go back a bit. There were selenium meters like this one, and there were CdS meters which had a different sensitivity (and also needed batteries), and Weston meters. There were different corrections for daylight and incandescent light. We still have stops. We still have exposure times. Unless you are into old school production, the rest is probably history.

Shutter angle and exposure time

A cine film camera cycle exposes a frame, then closes the shutter and advances the film. The term *shutter angle* refers to a rotary shutter. The shutter in the diagram uses two half discs. One of these half-discs is fixed, and is used to cover the aperture when the film advances. The other is moveable to give a shutter angle between 180 and 0 degrees. Here it has been positioned to give a shutter angle of about 85 degrees. These two

half discs spin on a shaft when the camera is running. If this camera was taking 24 frames a second, the exposure would be $(1/24) \times (85/360)$ second.



Adjustable shutter angle

The shutter angle was not always variable as is shown here. A smaller shutter angle gave a sharper image of a moving subject, but it reduced the amount of light for an exposure. A sharp image of a moving subject can look good, but it can also give disturbing aliasing effects, such as wagon wheels in Westerns turning backwards.

It is rare to have spare light when shooting cine film. If you had a variable shutter angle, it was often left at 180 degrees. If you had a fixed shutter plate instead, you could have a shutter angle of more than 180 degrees. This is what home cine cameras did.

Why use a small shutter angle? The answer is it usually looks good.

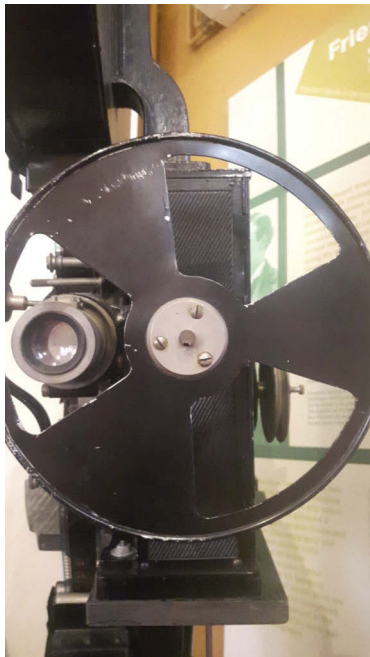
We do not see flicker much above 12 Hz, so it is tempting to think of the eye as a camera that takes 24 frames a second and smears out any motion. The widely-used term *persistence of vision* embodies this false idea. It is false, because our eyes do not stay still, or pan slowly like a cine camera on a tripod, but dart around rapidly.

Suppose we view an image with a fast moving car. We might see detail in the car if our eyes tracked the car, or detail in the background if our eyes stayed still. If the action is fast, we point our eyes largely by instinct, for there is no time for the signal to go to the visual cortex, be interpreted, and to send a signal back to the eye muscles. As things are, our ability to track is hit-and-miss, but our retina is good at picking up flashes of sharp detail. We do not know where our eyes are going in advance, but if our

eyes track a moving object, a small shutter angle image looks crisp wherever we look, and this feels natural.

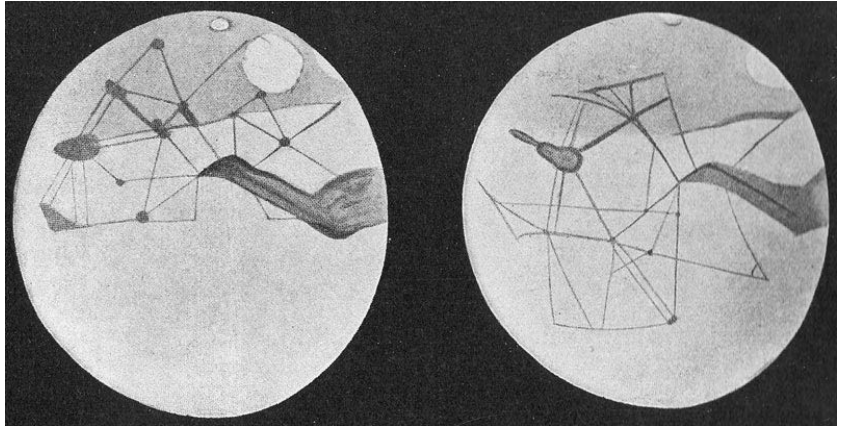
Video cameras capture images with effective shutter angles close to 360 degrees. Early tubes could not wholly erase the previous frame, so their effective shutter angle could go over 360 degrees. This gives video images their characteristic smoothness. This also meant that video could not capture moving detail in the way that film did. This is in part why TV doubled the cinema frame rate. People who want to give their video a 'film look' double every other video frame, to give the sensation of an 180-degree shutter angle. Modern digital cameras have a solid-state shutter that imitates the rotary shutter, and this exposure is often set as a shutter angle.

Film projectors use a similar shutter action, but they usually have two or three gaps per revolution. We can clearly see flicker at 12 Hz. Edison showed film at 16 frames per second. 16 Hz flicker is unpleasant to look at, so he divided each frame exposure into three separate bursts, giving a much more tolerable flicker frequency of 48 Hz. The modern movie standard of 24 frames per second dates from 1929, with a flicker frequency of 48 or 72 Hz depending on the projector action.



Instead of talking about *persistence of vision*, we now describe how movies work in terms of *flicker fusion*. We do not fuse a sequence of sharp images into smooth motion, like processing of old films in a computer. If anything, it is more like the amateur astronomical technique of taking a long video of a planet hoping for one sharp frame where the air is quite still.

This process of 'grabbing all the high resolution data we can when conditions are just right' can also make us see things that are not there. The astronomer Percival Lowell studied and sketched Mars on and off for fifteen years or so with a 24-inch telescope. He published his sketches in books in 1906 and 1908. Here are some of his early sketches of the canals on Mars...



Percival Lowell's sketches of canals on Mars

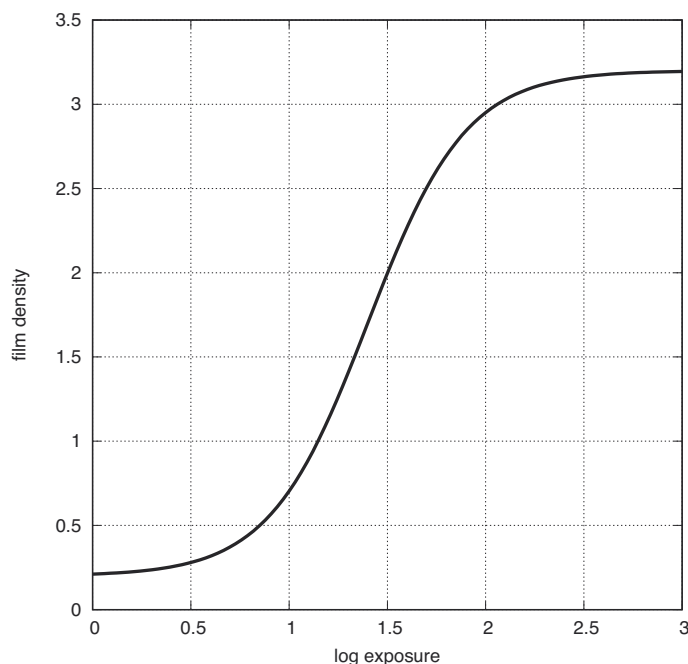
The Italian astronomer Giovanni Schiaparelli had also seen lines when he looked at Mars through a 15" telescope during the opposition of 1877. He had called them 'canali' – an Italian word that could mean natural channels, as well as the artificial features we call 'canals'. However, Percival Lowell also observed the shrinking of the ice-caps, and the changes in colour with the seasons; and interpreted what he saw as a titanic civil engineering project on Mars routing scarce water melting from the poles in their summer to irrigate the equatorial desert regions.

1909, was the year digging started on the Panama Canal. Also that year, the new 60-inch Mount Wilson telescope revealed Lowell's lines to be irregular and natural features, and Lowell came in for much criticism for his fanciful invention. Modern images of Mars make us wonder how he could have seen lines. Nevertheless, Lowell and Schiaparelli were not fools: they were both meticulous and eminent observers. To see what they saw, imagine sitting on a step-ladder for hours, peering through a telescope on a freezing night when vision is best, tracking a wobbling image of Mars. Suddenly, for a fraction of a second, the air holds still and Mars appears crisp. All your retinal cells that detect lines and edges have been starved of any decent signal for the past hour or so, and they all fire at once. You see lines, double lines, sharp features: all kinds of things. Now you try to sketch what you saw in that moment. Nothing you draw looks as crisp as that moment when the air stood still, but put in enough lines, and double some up, and you may feel you are getting close.

Our vision looks for sharp detail even when it is moving so fast that our eyes cannot track it reliably. We may not be sketching Mars, but we like images with a small shutter angle.

Black and white film tone curve

A black and white tone curve looks something like this...



Film tone curve

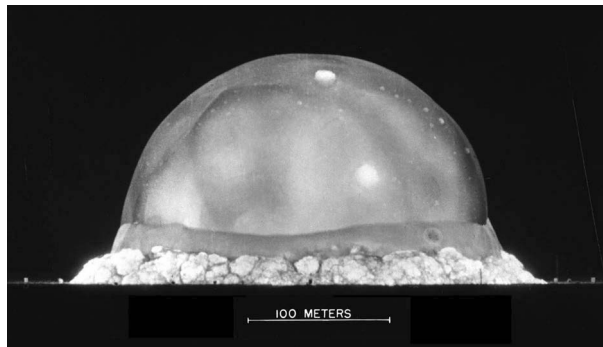
The horizontal axis is 'Log Relative Exposure'. The exposure goes up by a factor of 10 with each step. The zero point on this axis does not mean 'no light' but some arbitrary low light level. The vertical axis is 'Optical Density'. The amount of transmitted light goes down by a factor of 10 with each step. Zero on this axis is the transmission level of the clear film base.

It takes about five photons to make a big silver grain develop. Photons are rare in the shadows, so the chance of a grain having five or more photons goes as the fifth power of exposure. This gives the bottom of the tone curve the characteristic 'toe' to the 'S' shape. If you could add a flat exposure of two photons per grain, the grains only need three more photons, and the toe would become a third power curve. You cannot distribute photons that evenly, but a flat exposure can give a bit more sensitivity for very low exposures. This is called *latensification* (latent image + intensification). Photographers use 'flash' or 'bump' exposures

to alter the film tone curve and reveal hidden detail. It was also used for astronomical images and low dose X-ray work.

At the top end of the graph, almost all the grains have enough exposure to develop; more exposure has little effect and the curve flattens out, giving the characteristic 'shoulder' to the 'S' shape. In between the two there is a nearly straight 'linear' region.

If all grains were the same, this linear region might be rather small. You can extend the linear region with a range of sensitivities. Later black and white films used three emulsion layers with different sensitivities, one on top of the other, to extend the linear region. Extreme forms of black and white film with exposure ranges over a million to one (log relative exposure 6) were used for high speed films of nuclear tests, where the brightness varied rapidly and unpredictably. This is what the Trinity test looked like 25 milliseconds in...



Trinity nuclear test

Reversal film processes

There are two ways of processing black and white film.

With still photography the camera film gave a negative image. The light caused a blackening of the film when you developed it. You removed the rest of the unexposed silver halide when you fixed it. When you printed the film, you put the negative into a projector and exposed some photographic paper. This reversed the image again, so your highlights were white and your shadows were black.

You can also do reverse processing. Here you chemically remove the exposed grains instead of turning them black. You then reduce the remaining silver halide to silver. In the highlights, all the silver halide has gone, and the film is clear. In the shadows, the silver halide becomes silver and the film is dark.

Black and white cine film was usually reverse processed. This was clearly an advantage for home movies when you had the only copy. It was also

used for newsreels. You needed many copies of a newsreel. The prints made from a print had been through the 'S'-shape twice. The highlights and the shadows got flatter, and the linear region gets steeper, so the later prints lost the subtle tones and became more contrasty.

There is also reversal colour film. Kodachrome slide film is an example of that. The film in the camera became the film in the developed slides. Home colour cine film was mostly reversal.

Densitometers

This is the X-Rite TR-310 densitometer...



X-Rite TR-310 densitometer

This was used for the Truelight negative and print calibrations. There are other makes and models, but we found this one had remained much the same for many years and a new one gave almost the same readings as one made 30 years earlier.

It can work in transmission or reflection mode. The measuring head is at the end of the arm. In transmission mode, it used a light source under the surface. It could also work in reflection mode using a light source in the arm. But we were measuring film, so we never used it that way.

A transmission densitometer measures the transmittance of the film in density units:

$$\text{density} = -\log_{10}(\text{transmittance})$$

Typical films scatter light, as well as absorbing it. In an ideal world, we would choose our densitometer optics to match our application. If we wanted to match the appearance of a typical f/2 projection system, we would want an f/2 light source and detector. If we wanted to measure the appearance of a film on a light box, we would want a diffuse light source and a collimated detector. In practice, many densitometers use a 90-degree (f/0.5) input and exit cone. This design makes the densitometer almost blind to pure light scattering artefacts such as fingerprints, but it also means it sees some scattered light that would not be seen in a

projected image. Fortunately, colour film stocks do not scatter much light, so this does not affect the measurement much. Film with silver grains have hard high-contrast edges that scatter much more light, so you may have to take extra precautions when measuring black and white film.

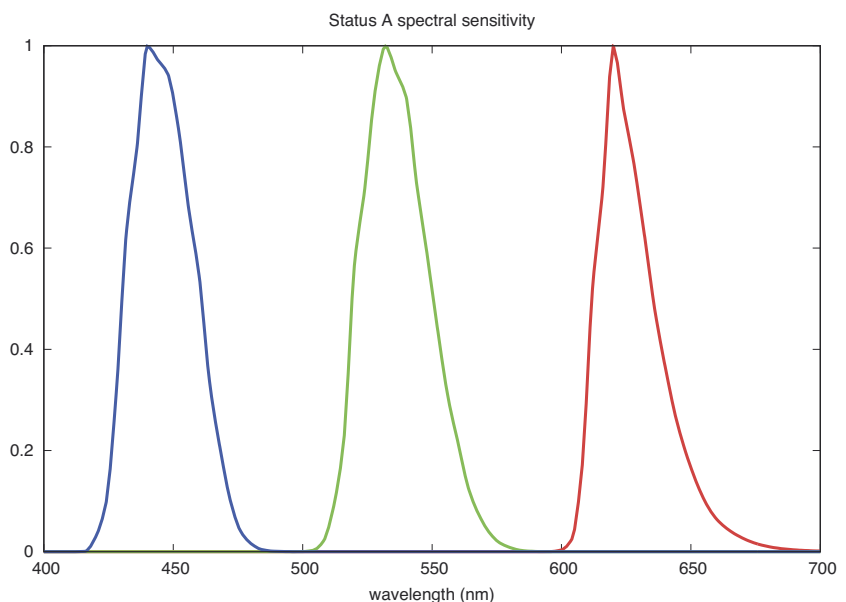
The first film densitometers measured black and white material. Film greys made from silver deposits generally have the same transmittance for all wavelengths, so different light sources should give similar density readings.

Colour film can absorb different amounts of red, green, and blue light. We can measure the density to red, green, and blue light using three monochromatic densitometers with different light sources. Usually, it is more convenient to use a single instrument with one light source and three filtered detectors.

Before making colour measurements, the densitometer should be zeroed. For absolute density measurements, we zero the densitometer without film. For measurements relative to the film base, we zero the densitometer on a bit of clear film base.

There are three ISO standard sets of filters: Status A, Status M, and Status T.

Colour print film has cyan, magenta, and yellow dyes. We use a densitometer with status A (Analytic) filters to measure the amount of dye. These filters have well-separated sharp peaks...



Status A filters are only for measuring combinations of cyan, magenta and yellow dyes. If you measure other materials, you may get misleading results. A material that absorbed 500 nm (blue-green) or 600 nm

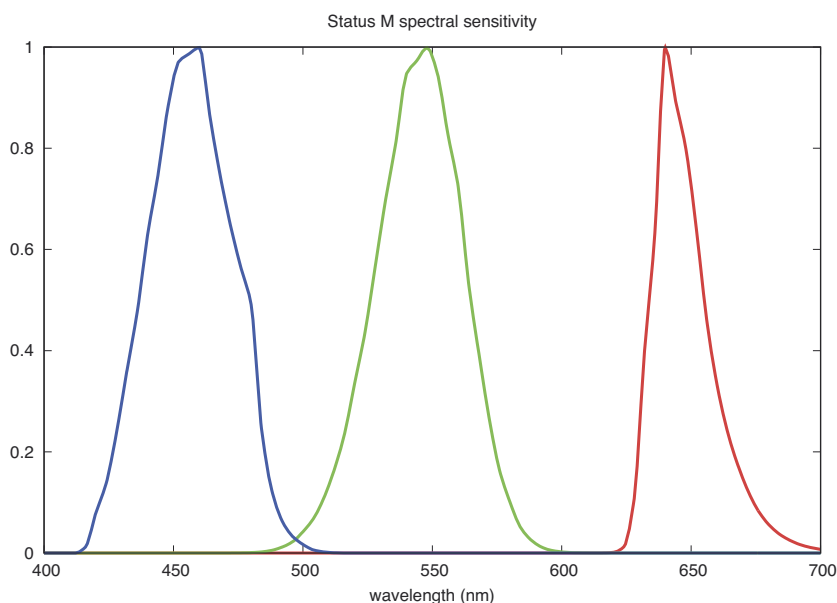
(orange) would give the same status A measurements as clear film, but might look very different. A neutral bit of colour film might have the same status A measurements as a piece of black and white film, but it would look lighter. This is because combinations of cyan, magenta, and yellow dyes have transmission peaks at about 500 and 600 nm.

Status A measurements can be made with an additional heat-absorbing filter. This makes the spectral sensitivity cut off faster as you approach the infrared. Under most circumstances this should make little difference to your measurements.

Status T filters measure positive print densities for CMYK printing colour separations. The peaks are narrower, and more separated.

Status M measurements are useful for predicting how negatives will print. Modern colour film printing uses an additive lamp house. An additive lamp house has three light sources matched to the red, green, and blue sensitivities of the colour print stock. A densitometer that is matched to print conditions should have similar spectral peaks to the additive lamp house.

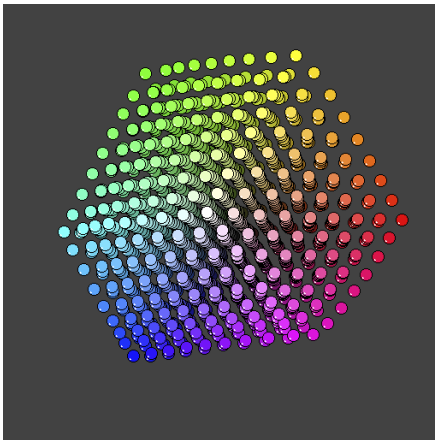
Status M is used for measuring negative and intermediate stock. Both stocks have an orange base. We zero on a clear region of the film base before Status M measurements. The status M spectral sensitivity curves are like this...



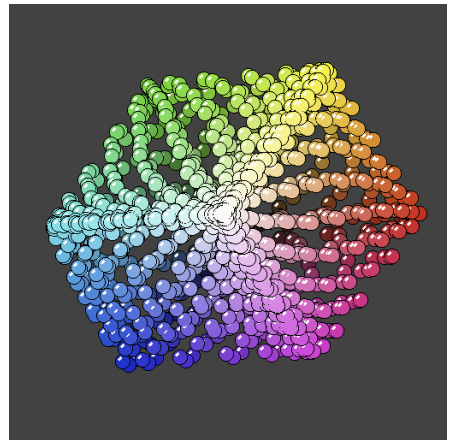
The ISO status M standard is for most negative work, including motion pictures, still pictures, and reversal prints. The SMPTE RP-180 recommended practice paper describes a similar densitometry standard that is aimed specifically at the motion-picture printing industry.

Unfortunately, there is no RP-180 equivalent to the tried and tested Xrite TR-310.

Truelight film calibrations used the TR-310 with a motorized film drive. The client output a standard strip of test images on their film recorder. If they wanted a print calibration, they would get their lab to make a print from that negative. We measured Status M values for each patch on the negative, and Status A values for the print. The negative calibration was a list transform with recorder RGB as input and Status M measurements as the output. The print calibration was a list calibration with Status M measurements as the input and print Status A measurements as the output.



Status M negative values

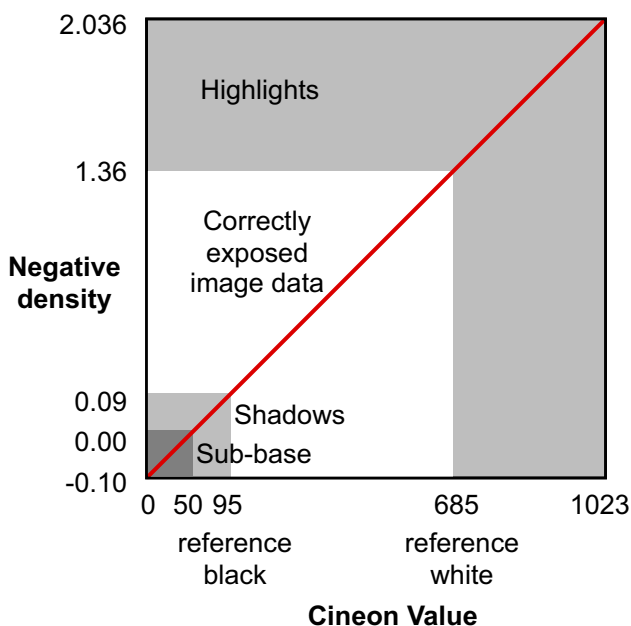


Status A print values

Cineon negative density space

‘Cineon’ is a Kodak trademark for an image processing package, and an image file format. The image file format has a standard way of representing negative densities in 10-bit data. It was widely used in the film industry. Cineon values were 10-bit integers from 0 to 1023.

It is odd to store image data as negative densities, since there are many ways of printing a negative image. Nevertheless, it worked at the time.



The red line shows the relationship between Cineon value and negative density for an aim gamma of 1.0. Each Cineon unit corresponds to a density shift of $0.002 \times \text{aim_gamma}$. Low Cineon values are light on the negative and dark on the print. Cineon RGB images viewed without colour correction on a monitor look like a washed-out version of the print.

Kodak defined absolute RGB 'Dmin' densities for each of their intermediate stocks. The Kodak formula for the absolute density was...

$$D_{\text{absolute}} - D_{\text{min}} = (\text{value} - 95) \times 0.002 \times \text{aim_gamma}$$

The diagram shows the film base density as having a base density of 50. If you stick to the Kodak definition, then the actual Cineon base value will depend on the actual choice of film stock. The actual Cineon base value of 5242 stock is 42.

You may not have Dmin values if you are not using Kodak stock, or you may prefer to measure negative densities relative to base. We used a nearly equivalent alternative formula...

$$D_{\text{absolute}} - D_{\text{base}} = (\text{value} - 50) \times 0.002 \times \text{aim_gamma}$$

Cineon space covers all the useful density range for conventional negatives. A Cineon image usually contains more information than you can see in a single print. You can vary the exposure in a Truelight simulation and see more detail in the highlights or shadows.

There is a subset of the Cineon space, with values between 95 and 685 that is commonly used for 'correctly exposed' image data. Kodak documents call 95 the 'reference black', and 685 the 'reference white'. Fixed conversions from Cineon to display RGB clipped the values outside this range. Old tube based film recorders also used to limit their output range to the 'reference white' or some other similar value to preserve tube life.

You are not restricted to the 'correctly exposed' range. Nevertheless, there are reasons for keeping most of your image within these limits. The negative exposure curve has its 'toe' beneath the reference black. Down at the toe, accurate interpolation is difficult, and the shape of the tone curve can vary with processing chemistry or the age of the negative film stock. You can avoid both these problems by adding 50 to all your Cineon values, which reduces the transmission of your negative by 20%, then increasing the exposure by 20% when you print.

Most print stocks have a tone curve with the 'shoulder' above the reference white. This print shoulder has all the same problems as the negative toe. You can avoid these problems by reducing your white to a light grey on the print. The image will look dimmer when viewed on a light box, but will look almost the same in a cinema.

This is a bit like the soft clip highlights we described in Chapter 5. The 'reference white' was then called 'paper white': the point where you wanted to be sure that texture would be reproduced.

Film lab exposure units

Film labs have controls on their printer lamp house that set the RGB exposure. These controls have units of printer points, with 12 printer points to a stop. Increasing the printer point setting will increase the exposure through the negative, which will make the print darker. Sometimes people quote a figure of 8 printer points to a stop, instead of the usual 12. This is about right when referring to stops on the camera that took the original film, and assuming a typical camera film gamma.

Laboratories work in absolute printer points. The absolute printer point setting to get a given print result on a given stock will be different for each laboratory, and will change with the condition of the chemicals. People

requesting prints from a laboratory will often ask for relative exposure shifts in printer points, and leave the absolute settings to the laboratory.

A shift of -1 printer point is equivalent to....

An exposure filter factor of 0.9439

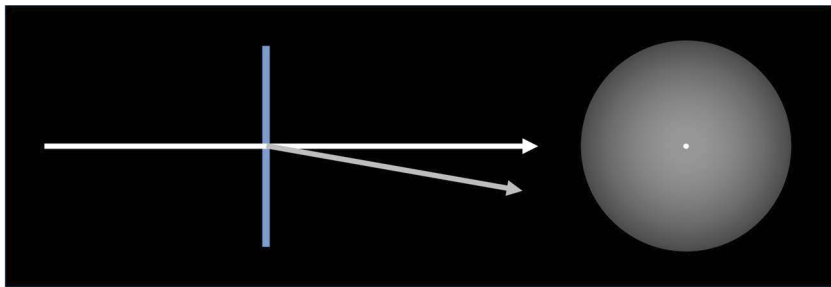
An increase of +0.025 in the negative density

A shift of +12.5 Cineon values

A shift of about -0.08 in typical mid-tone print densities

The Callier effect

Suppose we pass a single beam of light through a piece of film. The film absorbs some light, and scatters some light...



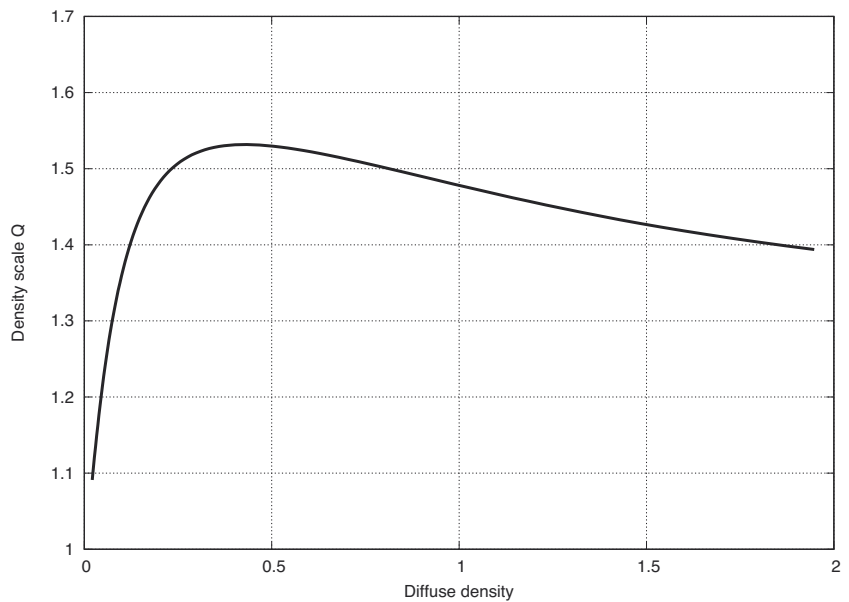
Scattering of light in film

If the film has a uniform density like a gel filter, it absorbs light but scatters very little. If the film is clear but has a fingerprint on it, it scatters light, but absorbs very little. Most users want to measure film dyes rather than fingerprints, so most film densitometers had wide collection angles, so they also measured the scattered light, and they don't see fingerprints.

A film projector may have f/2 optics. If the film scatters light, some scattered light arrives at the screen and can be seen as a halo around the image highlights. The projector optics absorb the rest. A normal densitometer measures almost all its scattered light, so it 'sees' proportionately more light from dark film. We see more contrast on the screen than we measure with a densitometer. This is known as the Callier effect.

We could measure the effective densities on a projected image instead of using a densitometer. This brings its own problems. If you stop the film motion, the frame rapidly heats up. If you keep the film moving you have

to have a long sequence for each patch you want to measure. Your light measuring instrument has to be able to cope with projector flicker. The projector lamp output may not be stable.



Callier effect

Here is an actual plot of the Callier scaling factor Q for monochrome film. If there are no grains then there is no scattering and Q goes to 1.0. The visual effect is very similar to scaling the density throughout by 1.4.

You can measure the Callier effect on black and white film. The effect is large because the sharp-edged silver grains are good at scattering light. You can project a single frame for a few seconds and do a quick monochrome measurement the light level.

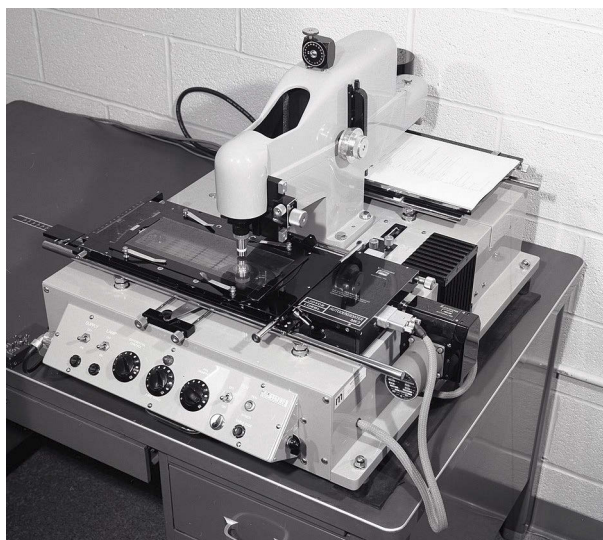
It is much harder to measure Callier factors for colour film. Colour film has dyes in soft-edged clouds, so they scatter less light, and the effect is about a tenth of the size. The colour image loses saturation as the film warms up, though it will recover when the film cools. Kodak recommended limiting showing colour slides to 45 seconds with a powerful projector. 45 seconds is not enough for many spectrometers if the light level is low. I projected a loop of grey patches in a running projector and measured the spectrum with a spectrometer that could correct for flickering light sources. I measured Callier factors between 1.02 and 1.05 for the various film dyes.

You could make better measurements with custom apparatus. I doubt this is worth the effort for calibrating film grading systems. A small density scale in all channels is almost invisible. A difference between

channels can give a subtle but visible colour shift. You can adjust the Truelight Callier coefficients by eye to fit the appearance for a particular print. However, the correction is smaller than the random shifts you would get if you ordered a second print from the same lab on the following day.

Microdensitometers and film scanners

Not all densitometers collect scattered light. Here is a Joyce-Leobl microdensitometer like one I used in the early eighties.



Joyce-Leobl microdensitometer

This can measure the density through a microscope objective using a slit aperture. This could examine minute edge detail with a large depth of focus. It had a lot of collimation on the source and detector, so it was very sensitive to scatter. I doubt if you will meet one of these, but if you measure density through a microscope, you may have to consider the effects of scatter.

Most film scanners are, in effect, microdensitometers. They may have less collimation on the light source, but they will not measure a significant amount of the scattered light. This is probably a good thing if you want to predict how the image will look, but you will lose more light if your film is dirty or the surface is in poor condition. You can reduce the surface scatter by scanning the film surrounded by a liquid with a similar refractive index. Still images were mounted in oil under a clear film on drum scanners. Motion picture film can be scanned under perchloroethylene using a scanner with a wet gate.

Anti-halation backing

A glow about the highlight in early black and white films is often associated with nitrate film stock. This is wrong: apart from exploding or turning to syrup, the nitrate backing has little effect on the image. Early films had haloes, but this was not from scattering but from light passing all the way through the film, and reflecting off the camera pressure plate. The pressure plate in a still camera should be painted black, but black paint would soon rub off the pressure plate in a motion camera. Motion picture camera pressure plates were often chrome plated to resist wear.

The solution for motion pictures was to add a black anti-halation coating to the film that could be removed with an alkaline wash when the film was processed. The same coating also prevented the build up of static, which could be a problem in dry places like Hollywood. The anti-halation coating innovation came in about the same time that nitrate stocks were withdrawn for safety concerns, which is why halation is associated with nitrate stock.

The anti-halation layer is commonly called 'remjet' or 'rem-jet'. I cannot find where this term came from or who coined it. It does not seem to be a trademark. A likely explanation is 'rem' is 'removable', and 'jet' is a standard grade for carbon black pigment.

Chemical edge effects

There are physical effects that can affect the chemical processing. When the lab develops film, the chemicals get more stirring near the film edges from the perforations. The edge of the print film gets slightly more development, and comes out slightly darker. We could measure this.

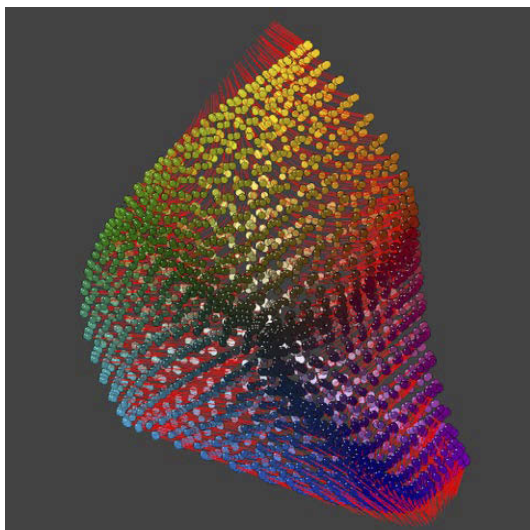
In theory, we could correct for this uneven development. In practice, most film projection was uneven, and ratios as large as 2:1 from the centre to the corners were considered acceptable for cinemas. Light scattering also varied from the centre of the projection to the edge. Compared to this, the darkening towards the film edge was small, and the colour film Callier effect was down in the noise.

Projector lamp calibration

The Truelight lamp calibration transform is typically a list transform going from Status A print densities to CIELAB values of the projected film. We do not measure each projected patch. Instead, we measured the transmission spectrum of the film base and CMY patches with a density of about 1.0, and the spectrum of the open gate white. We then simulated the effects of the light source on film with all the different densities of CMY. In theory the lamp calibration should be a property of the print stock dyes and the lamp spectrum. In practice, Kodak, Agfa, and

Fuji stock gave almost identical results, so the only important variable is the lamp spectrum.

In the early days of Truelight, people used an ARRI tabletop 35 mm projector called the Locpro. This had been developed for angiography, but it had a xenon lamp, so it should have been fine. However, this xenon lamp had an odd spectrum which made flesh tones look grey even though the white point matched. Our Locpro lamp calibration matched what the Locpro did.



Locpro colour shift

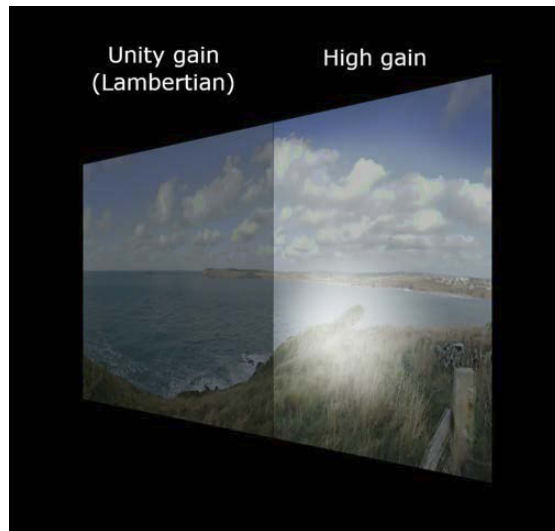
The red vectors show the colour shifts we get if we change from the Locpro xenon lamp to a regular projector lamp. The two lamps had similar colour temperatures. We have adjusted the data to make the whites match. We can see the colour shift is complex.

Screen gain

A Lambertian cinema screen reflects the projector light equally in all directions. This wastes a certain fraction of the light, spilling it onto the ceiling and walls rather than the audience.

A magic screen would reflect light only into the pupils of eyes of the audience. This would make much better use of the available projector light. The rest of the cinema would be darker too, so you would have better contrast. I have often wondered how we could do this. Something holographic, perhaps. However, if we could do this we could send

different images to the left and right eyes for 3D cinema without glasses. And people have tried very hard to do that.



Screen gain

The screen gain is the dimensionless ratio between the peak brightness of the screen and the corresponding brightness of a Lambertian reflector when both are illuminated in a given cinema-like geometry. The screen gain may vary depending on where you sit in the cinema.

A high gain screen is sometimes called a 'silver' or a 'grey' screen' because of its appearance. It reflects more light at the audience, so the projector appears brighter. Unfortunately, the screen gain falls off rapidly with viewing angle, so the viewed image may look dull at the edges, and have a noticeable bright spot — often called a 'hot spot' — centred on the specular reflection of the projector.

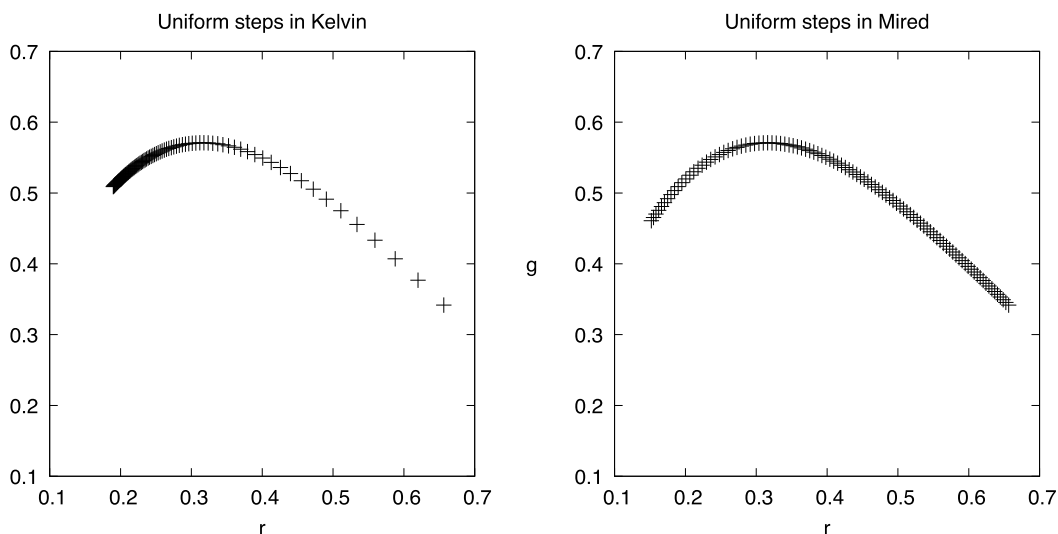
Real cinema screens are a compromise between brightness and evenness. A small cinema usually uses a Lambertian or 'unity gain' screen. The film projector probably has enough light to give the standard 50 nit white with a small unity gain screen, which gives the best looking image with no 'hot spot'.

Cinerama screens were curved. If they were a simple curved sheet then light reflecting off one end of the screen could land on the other end. The screen had vertical strips of material with thin white walls in between. These walls were almost invisible to the audience but trapped the high-angle scatter.

Lighting and mireds

Sometimes colour temperatures are measured in mireds, or micro-reciprocal-degrees. You divide a million degrees by the Kelvin colour temperature to get the mired value. A colour temperature of 2000K is 500 mireds. You can also get colour correcting filters with positive or negative mired values, which will, when added to the mired value of any thermal source, give the mired value of the filtered light. For example, a +100 mired pale blue filter will change a 2000K 500 mired tungsten spotlight to a 2500K 400 mired source, and an orange -100 mired filter can change it back again. This is only approximate, of course. In Chapter 4 we saw the white spectra plotted out for different temperatures, and we can see we could map one temperature to another with a simple left-to-right scaling. This is roughly what these mired filters do.

Colour temperature is not a perceptually uniform measure. Here we have plotted out blackbody spectra for uniform steps in Kelvin, and in Mired in our tRGB **r,g** chroma space...



The Kelvin points go from 1000K to 10,000K in steps of 100K. There are a similar number of Mired points starting at 1000K and going to 100,000K. One more step would get you to zero Mired and infinite temperature, which is a surprisingly ordinary light blue colour predicted by Planck's law.

This also shows that it is easy to interpolate this thermal curve in **r,g** chroma. You can interpolate the Mired value from **r**, and then check the **g** value was close enough to the thermal curve.

The camera of the future?

A camera was a box with film at the back, and a lens at the front. The lens had an aperture which could be varied, so you could let in more light, or stop down the aperture for a larger depth of focus. A shutter would vary the exposure to the film. The film was later replaced with a solid-state array of RGB sensors, which could also be sampled so that it acted as its own shutter.

There have also been...

Stereo cameras	These can record the view from two or more viewpoints. If you are going to send the images direct to the eye then two may be enough. If you want to give the illusion of 3D where you can move your viewpoint you will need more. In the limiting case you would record a hologram, where all possible views are recorded.
Hyperspectral cameras	These can record more than just three RGB channels like the eye. There are cameras with 4×4 filter arrays for each pixel, which can give an approximate spectrum. There are also spectral cameras that can record a full frame for each wavelength. These are usually not capable of recording motion because there is not enough light after you have split it into too many small wavebands.
Plenoptic cameras	The detector is effectively an array of box cameras that record the angular distribution of the light for each pixel. This allows the user to recreate the cones of light, and refocus the image after the data has been gathered. While this sounds great, it wastes a lot of light, and your array of box cameras use a 4K detector to give an SD video resolution image.
High frame rate camera	This has other uses besides capturing fast action. A conventional motion camera takes a set of samples at fixed intervals. This will mean it will alias when viewing any action with a similar frequency, giving you wagon wheels that spin backwards, and flickering monitors. If you record at higher frequencies you can resample afterwards.

The stereo camera is the simplest of these. Early stereo pairs of still subjects were often two plates taken through the same camera, with the photographer putting their weight on the left foot, then the right. For motion, it is often two cameras, side by side. This is easy to develop, and it shares lenses and parts with ordinary cameras. All the others try to

capture extra information through a single lens. If you need a larger lens to capture more light, the cost of a large lens goes up roughly as the cube of its dimensions. If you need custom lenses, the cost will be even higher. And the camera operator has to get used to a new lens.

In Chapter 6 we used our smooth spectra set to make an approximate correction for major changes in illuminant. This cannot resolve metamers, but that probably is not necessary in most cases. If we had a hyperspectral camera, this would give enough information for a grader to make any corrections. We would not use a hyperspectral camera to capture motion picture images, but it might be handy to record what the set looked like.

The plenoptic camera gives an image we can refocus after the event. However, the result is often little better than what we could get from five webcams with fixed focus.

A high frame rate could take out flickering, but it is hardly ever used for that. Usually there are better ways of avoiding troublesome frequency aliasing. Wagon wheels can be blurred. Monitors can be made not to flicker. Al Pacino’s ‘Scarface’ even had the ‘Say Hello To My Little Friend’ machine gun synced to the camera shutter, so the muzzle flashes happened on a frame.

The traditional camera is likely to last in its present form: it is familiar, so people know how to get the best results out of it, and there is a good choice of lenses. However, there is extra information we might want to record, including...

Depth	A lot of depth information can be inferred if the camera moves, but a stationary camera and a stationary subject gives little information. A small, cheap second camera could provide depth from stereo. This could be used to simulate 3D or a shallow depth of focus, or for simulating smoke.
Spectra	A single frame of spectral data will probably record enough for a shot. If you have neon lighting, or metameric flowers in your shot, this should record what the colours ‘actually were’ which would help when grading.
Lighting	It is handy to record the lighting: how much there is and where the light is coming from. This is particularly useful if we want to add rendered details. The highlights can be recorded with a shiny ball and a low exposure. The diffuse lighting can be recorded with a matte white ball and an ordinary exposure.

Wide angle	This is not IMAX: we sometimes need image information that is out of shot. This is necessary if we want to render reflections. It would be handy when blurring the edge of the image, or for patching missing bits of image if we rotate. There may be new formats which support lower-resolution surrounds.
HDR	We could have a second camera that records highlight and shadow detail in low resolution. The lowlight detail would have to use some method of excluding the ordinary light levels or it would be dominated by flare. Perhaps we could use the DLP module from projectors to do this.

One major requirement: this must not get in the way of anything else on set. The whole studio team and actors will rarely hang around patiently while someone fiddles around with a Macbeth chart, a camera, and a chrome ball. The gadget does not have to do everything in this list, but it must be portable and able to sense where the main camera is pointing, or small enough to attach to the main camera without getting in the way. It should sync with the main camera, or run continuously. This may be tricky but not ridiculously so.

A more dramatic solution might be to change the main camera to monochrome. At the moment we sample RGB through the main camera single lens onto a mosaic detector. Our RGB filters typically pass less than half the light of the particular channel. A monochrome camera would have the same hardware but could run on a sixth of the light. Or, if we filtered it to only pass a narrow band, we could avoid chromatic aberration in the lenses. It would not suffer from aliasing of the colour signals at the detector mosaic. We would have to record the colour image through a second camera, and combine that with the high-quality main image. This may be tricky but not ridiculously so.

Chapter 10

Modelling light and materials

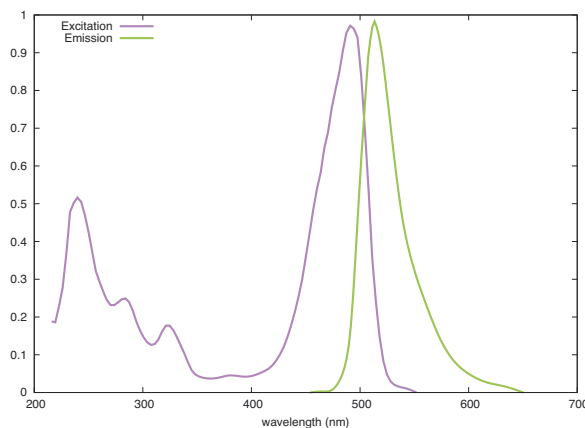
At the beginning of Chapter 3 we touched upon Aristotle, and his experiments with colour. In this chapter we shall briefly describe how we can model the physics of how light interacts with objects.

Modelling is a big subject, and the games and motion picture industries are making it bigger as fast as they can. This chapter will not get you writing your own ray-tracing program. However, it may help our understanding of vision as it is the reverse process: it takes the shape and nature of things and makes a scene that convinces our eyes.

We start with the Monte-Carlo light model. This is not the fastest light model, but it is the easiest to understand. It models how light bounces around in a scene, reflecting and refracting at surfaces, and getting filtered and absorbed by the media it passes through. However, before we start on Monte-Carlo, we shall consider a few things that light can do, but we may not want to have in our model.

Fluorescence

Most models assume that light stays the same colour as it bounces around. There are some colours that do not. Some materials absorb one wavelength and emit another. The compound fluorescein absorbs blue light (and also some deep UV light) and produces green-yellow light...

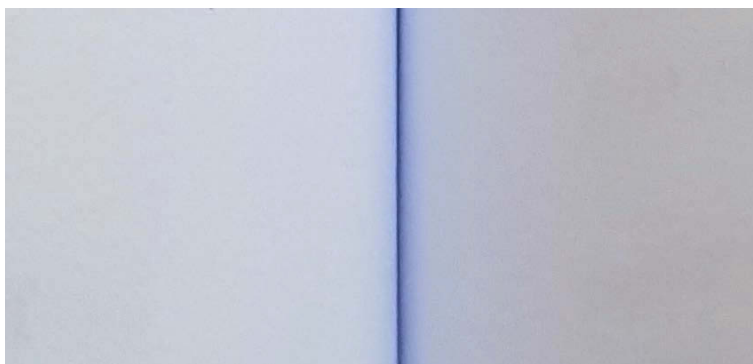


Fluorescence

This is the compound used in ‘day-glo’ or ‘high-vis’ materials, and in spirit levels. It is also used to colour frozen peas in the UK. The compound eosin, which is fluorescein with a couple of bromine atoms, emits in the red, giving us the pink fluorescent colour. I remember making fluorescein with the Friedel-Crafts reaction at school, then trying substituting other halogens – they are all pink-ish, but eosin was the best.

There are other man-made fluorescent materials called ‘optic brighteners’. These are mostly from the stilbene family. They are used in washing powders to offset the yellowing of materials, and in paper-making to make papers appear whiter. They mostly absorb in the UV and emit in the blue. If the paper is on display and exposed to UV the optical brighteners may degrade. Fabrics may replace their optical brighteners when they are next washed. Fluorescent yellow and pink fades on prolonged exposure to UV, but you can counter that by using a lot of it. You can get a good colour gamut with fluorescent inks. I tested this using a fluorescent CMY ink-jet cartridge, which Canon used to make, but the prints fade if you leave them out.

You can sometimes see the optic brighteners as a blueish glow in books where the pages join...



Optic brightener effect

The print industry used to use drum scanners for high quality scans of paper images. The image was illuminated and scanned one pixel at a time through a microscope objective lens. The Crosfield drum scanner had two sizes of drum, and the scan looked brighter when the smaller drum was used. The brightener has a small delay between the absorption of the UV from the scanner’s xenon lamp, and the emission of the blue light. The paper moved faster on the larger drum, and the scanner had moved on to the next pixel before the blue had been emitted. I mention this as a curiosity: this is not a problem you are likely to meet.

Some natural materials fluoresce. Natural paper fibres fluoresce in the blue a bit. Uranium or ‘Vaseline’ glass fluoresces in the yellow-green, and there are other natural minerals that fluoresce in other colours.

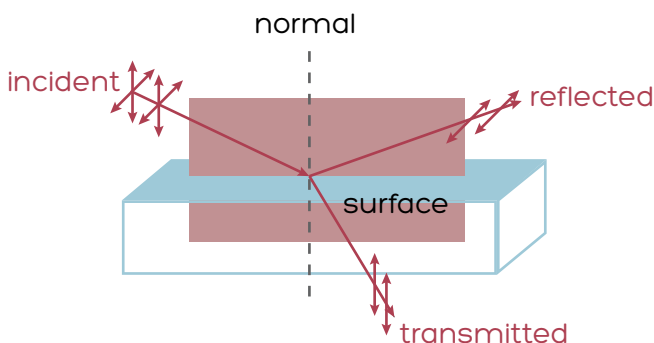
Apparently, ripe bananas fluoresce blue. However, natural fluorescence is rare, rarely efficient, and often swamped by the normal, visible wavelengths, which is why bananas kept their secret for so long.

In theory, you should model the UV light in the scene in case it meets a fluorescent material. In practice this is rarely necessary. Artificial lighting should not have much UV. Sunlight has UV diffusely reflected from the sky, so we see no UV shadows and fluorescent colours outdoors just look like slightly brighter colours: fluorescein is bright yellow, eosin is a bright pink, and optical brighteners are white. Perhaps these colours are a bit brighter than a pure diffuser white would be under the same light, but this is not obvious. Perhaps our eyes assume the white point is a bit brighter if there is a fluorescent colour in the scene. Certainly, if you take a picture of a scene with a fluorescent object, the object still looks fluorescent in the image. This works even if you print the image on paper.

This is good news for light modellers: in general we do not have to simulate invisible wavelengths in case of fluorescence. The two main fluorescent colours, yellow fluorescein and pink eosin, fluoresce under blue light, which we already model as we can see blue.

Polarization

If light grazes off a flat surface, the reflection will mostly be polarized with the electric vector in the plane of the surface. If we wear polarizing glasses that only pass the other polarization, we see 50% of the light in the scene that has this polarization, but a much smaller fraction of the reflection. This is a good thing if you have sunlight glancing off a road or water.



Polarization at a surface

There is a particular angle, called *Brewster's Angle*, where the reflected light contains only one polarization. The light that goes through the surface always has both polarizations. At other angles the reflected light has both polarizations. The diagram exaggerates the separation.

In the diagram we draw the refracted beam as a single line. There are crystalline materials such as calcite where the refractive index is not the same in all directions, and is different for the two polarizations. This is called birefringence. You can also get a weaker form of birefringence by stressing a material or applying a strong electric field. This is the cause of the coloured fringes you may see in a toughened glass windscreen if you wear polarizing glasses.

It is easier to model photons as simple particles. If we model polarization, each photon must have a polarization vector, and each time the photon meets a surface, we have to calculate the components of the vector parallel and perpendicular to the surface, and keep track of what fraction of each polarization was transmitted and reflected. If we have to model birefringence, things get even more complicated. Happily, most scenes do not look very different if you do not explicitly model polarization. If you want to model a view seen through polarizing sunglasses, you can probably get away with only modelling polarization in the last reflection.

Iridescent colours

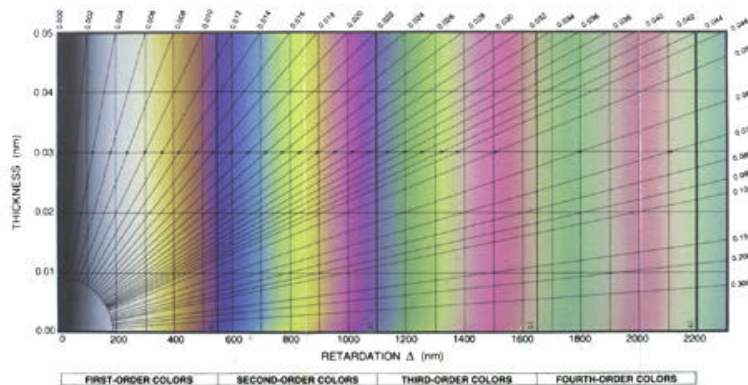
There is another class of colour processes that we do not usually model. Iridescent colours come from the interaction of light with structures with a similar wavelength. Peacock feathers, opals, beetles, and moths are natural examples of iridescence, as well as the interference colours from bubbles and oil films. These colours are generally *goniochromatic* — the colour changes with viewing angle.

Goniochromatism makes iridescence hard to model. We could have a holographic grating that reflects a different colour in each direction. However, most iridescent colours are more subtle: a peacock's tail feathers can appear blue or green or azure and still be recognisable in real life or in an image.

Newton observed these colours in small air-gaps between nearly flat plates, similar to the arrangement Thomas Young used to measure wavelength. He described the coloured rings in Book II, Part I of his *Opticks*, divided into the following orders:

- 1 Black, blue, white, yellow, red
- 2 Violet, blue, green, yellow, red
- 3 Purple, blue, green, yellow, red
- 4 Green, dirty red

We get similar iridescence colours from polarization. You can see these colours on toughened glass car windscreens.



Michael-Levy chart

Put a thin sample of mineral in a polarizing microscope. Where the sample has a single refractive index, it looks dark under crossed polarizers. Where the sample has different refractive indexes for different polarizations, the interference between the polarizations gives colours like Newton's air-gaps did. Rotate the sample to get the greatest retardation as given by the colours on the Michael-Levy chart. If you know the sample thickness, you can read off the birefringence. This is the difference between the two refractive indices. Quartz has a birefringence of 0.009. Most minerals contain some quartz, so this is a useful check on the sample thickness.

Like the colour of peacock feathers, polarization colours can vary with observing conditions. In this case the 'sensitive violet', the colour at the end of the first-order colours, can vary from 536 nm to 557 nm depending on the illuminant even though these were all sensible whites. I remember using the Michael-Levy chart in the Department of Mineralogy and Petrology forty years ago, and I can confirm this. In general, it is a lot of work to calculate the effects of polarization to predict a small colour shift that may not be the same for all illuminants, or for all observers.

If we want to simulate iridescent colours in our model, we can support a more sophisticated version of the Phong reflection model described later.

The colour of the sky

Leonardo da Vinci is perhaps the first person who tried to explain the colour of the sky in terms of scattering. He describes in his notebooks how the sky and objects in the distance looked as he ascended Monte Rosa, between Italy and Switzerland....

"I say that the blueness we see in the atmosphere is not intrinsic colour, but is caused by warm vapour evaporated in minute and insensible atoms on which the solar rays fall, rendering them luminous against the infinite darkness of the fiery sphere which lies beyond and includes it."

https://en.wikisource.org/wiki/The_Notebooks_of_Leonardo_Da_Vinci/VI

He reasoned that the colour of the sky is not because air is transparent blue. If this were so, the colour of the sky would become more intense towards the horizon because you are looking through more air, a bit like stacking blue filters. Instead, he likens it to the blue colour you can see when viewing the smoke from old, dry wood against a dark background...

"If you produce a small quantity of smoke from dry wood and the rays of the sun fall on this smoke, and if you then place behind the smoke a piece of black velvet on which the sun does not shine, you will see that all the smoke which is between the eye and the black stuff will appear of a beautiful blue colour. And if instead of the velvet you place a white cloth smoke, that is too thick smoke, hinders, and too thin smoke does not produce, the perfection of this blue colour. Hence a moderate amount of smoke produces the finest blue. Water violently ejected in a fine spray and in a dark chamber where the sun beams are admitted produces these blue rays and the more vividly if it is distilled water, and thin smoke looks blue. This I mention in order to show that the blueness of the atmosphere is caused by the darkness beyond it, and these instances are given for those who cannot confirm my experience on Monboso (Monte Rosa)."

Any fine suspension of particles scatters light. If the scattering particles are much smaller than the wavelengths of visible light, the shorter wavelength blue will be scattered more than the longer wavelength red. This scattering is usually from nitrogen molecules, though it can also be caused by sulphur compounds after volcanic eruptions. The scattered short wavelengths make the sky look blue, and the absence of these short wavelengths makes the sun look yellow, or orange as it sets. Scattering also gives the blue colour in blue eyes and the exhaust from two-stroke motors.

This scattering is called *Rayleigh scattering* because Lord Rayleigh came up with a precise prediction from electromagnetic theory for particles much smaller than light. It is also known as Tyndall effect after a Victorian scientist who reproduced it long after da Vinci. Rayleigh scattering is polarized, particularly at 90 degrees to the sun.

Mie scattering commonly describes the scattering of particles of micron size and above. This includes pollen, dust, smoke, water droplets, and other particles in the lower portion of the atmosphere. Gustav Mie in 1908 provided the full solution for scattering and absorption by small particles of all sizes in 1908. Kurt Nassau in his book 'The Physics and Chemistry of Colour' (ISBN 0-471-39106-9) describes the red of ruby

glass (Purple of Cassius) as *Mie scattering* though the particles are much smaller than the wavelength of light, and the light is absorbed rather than scattered. So, if you search for Mie Scattering you may get a few surprises. In the high atmosphere, Mie scattering off ice crystals gives us halos, and sun dogs...



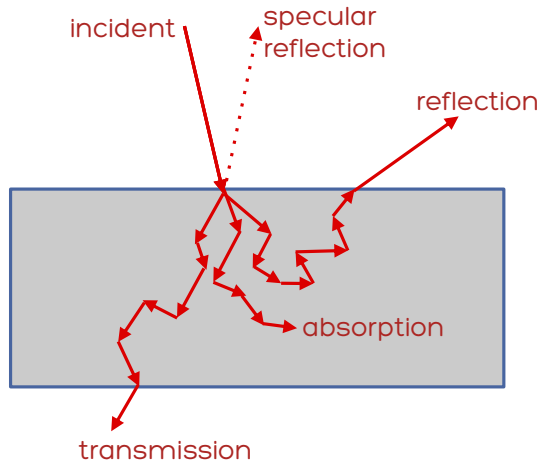
<https://earthsky.org/todays-image/sun-halo-sundogs-sky-optics-canada>

However, atmospheric Mie scattering usually means the scattering by round or irregular particles such as pollen and dust in the lower atmosphere. This is almost always neutral: it gives us the 'glow' about the sun in the picture above, and the appearance of clouds and smoke.

A small fraction of the sky colour is caused by a mechanism similar to fluorescence. UV light is absorbed by the upper atmosphere, and re-emitted at visible wavelengths. This is called airglow. It is not visible during the day but can be seen at twilight. It is usually a greenish colour, similar to auroras, and is particularly visible from the International Space Station, where one can see Earth's atmosphere from the side. Nishita et al. (SIGGRAPH 1993) was probably the first good physical model for rendering skies in scenes. This includes Rayleigh and Mie scattering for random particles. Their 1996 paper includes multiple scattering. It does not include airglow, halos, or sun dogs.

Monte-Carlo simulation

Most of the rest of this chapter deals with simulations where the photons bounce around within some scattering and absorbing medium, but their wavelength does not change, and polarization is ignored. Given a nice, fast computer, we can model what each particle of light does...



Monte-Carlo sum

Here light meets a layer of something such as paint. Some light may reflect at the surface and will be handled by the Phong model described shortly. The rest goes into the material. Within the layer the light may bounce around at random. It may be absorbed by dye or pigment. If it does not get absorbed, it may eventually come back out the top surface, but at a different point to where it went in. Or it may go through the bottom surface, and enter the layer below. We have not drawn any layer below, but there is still a chance the light may come back up again.

This is one ray of light with a particular wavelength, coming in at a particular angle. We have to sum over all wavelengths, all angles, and all possible paths to see what the reflection looks like. Or we can just fire off photons at random, and sum what they do. This is less efficient, but it is much simpler to program. This will only predict the appearance if we have good values for all the light scattering and absorption processes. It might do a good job on Aristotle's original example of charcoal on Pentelic marble. If we simulate polished wood, or skin, our model will be a lot simpler than the real thing.

In the summer of 1913, the roulette ball at the Casino de Monte-Carlo fell on black 26 times in succession. Unlikely-looking sequences can happen, but we know we usually come out of a Casino poorer than when we went in. If we make lots of bets, our losses probably approach some statistical model that is much easier to calculate. The same applies to the

predictions of the Monte-Carlo method. If you get noisy results, average more calculations. If more averages make no visible difference, you are probably done.

The Monte-Carlo method is not only its own benchmark: it is also a useful benchmark for other, simpler calculations that came long before computers made the Monte-Carlo method a practical tool.

Beer-Lambert law & additivity failure

The Beer-Lambert law describes how light is attenuated by transparent media. If you have two layers of translucent material, you add the densities. If you have twice as much material, you double the density. If you have the same thickness but have diluted the colour 50:50 with something translucent, you halve the density.

As history would have it, this was initially discovered by Pierre Bouguer when on a holiday in Portugal while looking at his glass of local red wine, not beer. He wrote this up in 1729, and Lambert cited this paper. August Beer had a similar law that involved dilution. The Beer-Lambert Law encompasses both of these.

The modern version has all sorts of extra requirements. The material must not scatter light. It must not change the wavelength of the light. The light must be monochromatic, or at least sufficiently narrow band that the wavelength are all attenuated by the same amount. Here I will use the Beer-Lambert Law in something like its original sense.

Here I simulate the appearance of different densities of a yellow dye. This was measured from a projection of a patch of Kodak Vision film stock, and the white reference was clear film base. The yellow patch had a measured density of about 1.0. This chart simulates the colours we get with up to six times the original density in sRGB...



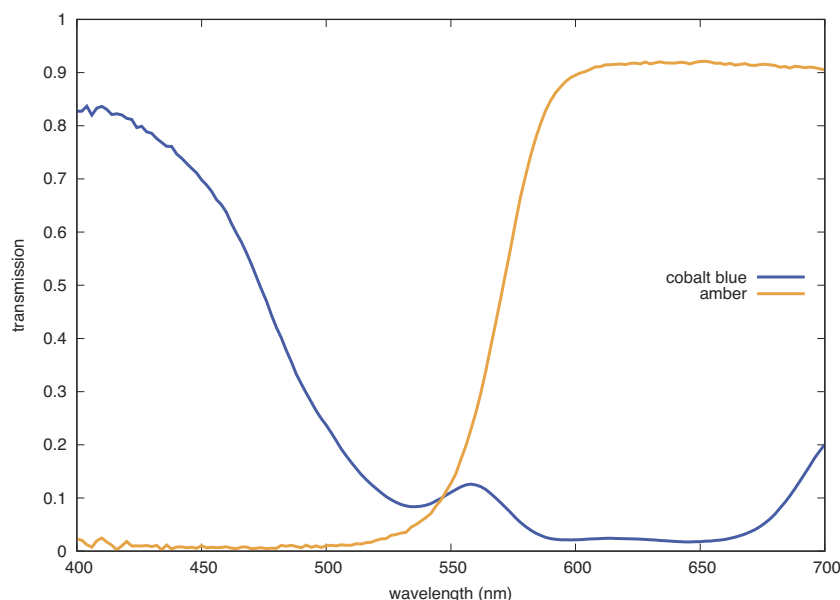
Additivity failure

The colour changes from bright yellow (up to second row left) to orange (bottom right). The yellow dye removed all the blue wavelengths earlier; adding more yellow now takes out the longer wavelength greens. Eventually we end with red.

It may be difficult to see the contrast because our vision sees this as a wedge of one colour, and tries to correct for it. Those of us that work with printing and test charts a lot may be better at seeing the actual colours. But the hue shift is there. The old printing term for this shift was 'additivity failure'.

This is a more dramatic effect with the cyan dyes, but sadly I cannot show it as the cyans cannot be rendered on most displays. It might show in a printed copy, but that would depend on the local printing inks. We do not often need full gamut displays, but this is one of the times it would have been handy.

We can also get unexpected colour changes...



Transmission

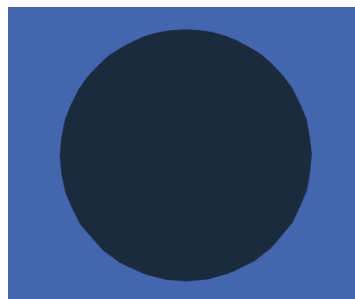
If you stack the Amber glass and Cobalt Blue, and look at an incandescent light, it looks red. Cobalt Blue transmits a small amount of the red above 650 nm, and this red is all that is left when Amber cuts off all the shorter wavelengths. These effects are hard to show in a chart table as I did with the yellows because the added colours are very dim. But it is quite striking when you see it.

Phong shading

Phong shading is a simple model for reflections off surfaces. Bui Tong Phong wrote this model for his Ph.D. dissertation in 1973. Here is a demonstration of Phong shading for a red sphere under a single light source with a blue background...

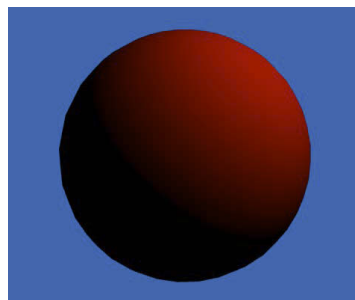
Ambient reflection.

This is a uniform light. This simulates the ambient light from the surrounding scene, or the glow of the object itself. Here, the ambient reflection is a dull blue.



Diffuse reflection.

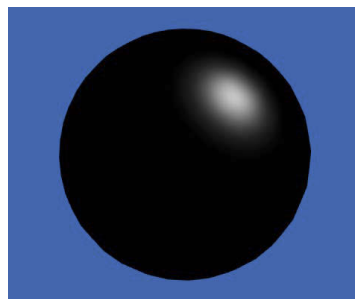
The lighting is white and the ball is red. The shading varies as the cosine of the illumination angle, as though the ball was matte, and reflecting light equally in all directions.



Specular reflection.

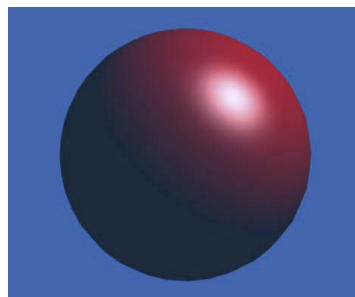
The lighting is white and the reflection is white. The shading is similar to the diffuse reflection, but the cosine distribution has been raised to the *shininess* parameter.

This image uses a shininess of 10. This gives a narrow highlight, as though the ball was slightly rough. A higher value makes the ball look smoother.



Full model.

All these components have been added together.



Phong shading is quick to calculate and gives a realistic render for opaque objects. If necessary, we can add extra parameters for directional textures (turned metal) or iridescence (mother-of-pearl).

Maxwell's equations predict how a reflection splits the two polarizations. Phong shading and its variants do not usually model the polarization because that would make things a lot more complicated.

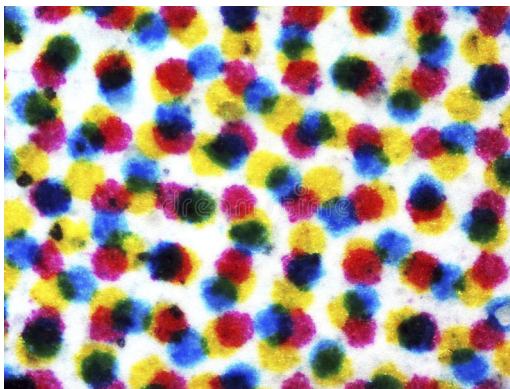
Kubelka-Munk model

The Kubelka-Munk model predicts the reflection of pigments and glazes, and other uniform layers. This is a one-dimensional up-and-down model. There is an absorption constant K , and a scattering constant S . If light travels a small distance δx , a fraction $K \cdot \delta x$ is absorbed, and a fraction $S \cdot \delta x$ is scattered to the opposite direction. This gives us two coupled differential equations for the inward light flux i , and the outward light flux j .

This may seem like an absurd simplification of 3-D scattering, but it actually fits quite well once you determine the constants S and K for each wavelength. The equations have analytic solutions. There are simple solutions for a single layer with a reflecting backing, for the case where the thickness is large, and for when S is small compared to K . It is still used.

Neugebauer model

The Neugebauer model was a simple model for colour printing. Suppose you have a halftone pattern printed in CMYK...



CMYK halftones

Note:

Printers give black the letter K for 'Key'. A keyline is a black border that printers put around colour images to hide any registration errors for the halftone patterns.

Suppose we print the yellow separation first (this is common as yellow inks can scatter light). Our yellow dots cover an area fraction Y of the paper, leaving a fraction $(1-Y)$ of the paper white. We then print the magenta separation. Unless the halftone patterns align, the area fractions should combine at random, giving us...

Area Fraction	Colour
$(1-Y).(1-M)$	Paper white
$Y.(1-M)$	Yellow
$(1-Y).M$	Magenta
$Y.M$	Red (Yellow + Magenta)

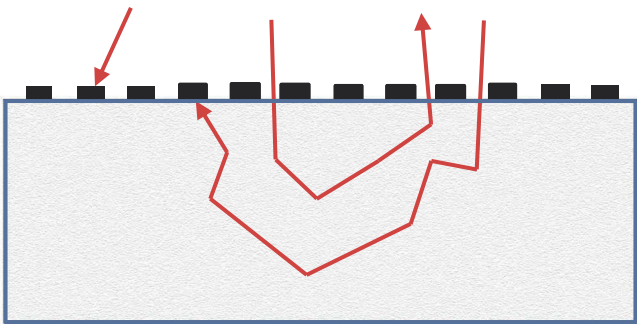
With CMYK printing we have 16 possible overprints. If we make a patch with each possible overprint, and measure the RGB density, we can then calculate the RGB reflection of the halftone. This is not particularly accurate for fine halftones, but it works for big dot patterns.

Yule-Neilsen model

So far we have not modelled the sideways scatter. The Yule-Neilsen model assumes any light getting to the paper scatters about in the paper at random, and comes out at some point near where it went in. Suppose we have a black halftone pattern with an area fraction of K .

Suppose the scattering distance is small compared to the ink pattern. We then have the Neugebauer model. The ink will absorb a fraction K , and the paper reflects the $(1-K)$ remainder.

Now suppose the scattering distance is large compared to the ink pattern....



Yule-Neilsen model

The ink absorbs a fraction K of the light on the way in. The $(1-K)$ remainder bounces around within the paper and comes out at a different part of the ink pattern. The ink pattern absorbs a fraction K of the $(1-K)$ on the way out. The fraction of reflected light is not $(1-K)$ but $(1-K)^2$. This is probably a good model for Aristotle's charcoal lines on marble.

If we have 4-colour printing, we have the Neugebauer fractions on the way in and on the way out.

The Yule-Neilsen model and the later Clapper-Yule model predict reflections where the light scatters within the paper. I have worked on conventional and stochastic screening techniques. These reflection models are handy to explore the nature of the colour effects, but we rarely have precise data for the ink distribution in the dots or the scattering in the paper. I might start with a model, but I would end up fitting to many measurements of colour patches.

A fine pattern can give a different colour to a coarse one. With coloured inks, we can get a change in colour from additivity failure. If we have a cyan ink with a large dot pattern, our light will pass through the cyan ink on the way in and on the way out. If we have cyan ink with a small dot pattern, our light may pass through the ink only once. We get less additivity failure from going through the ink once, so the cyan appears more saturated: more cyan, less blue. This change in colour can be quite clear, particularly if you print a stochastic screen patch next to a conventional one.

BSSRDF

A bidirectional reflectance distribution function or BRDF models surfaces which absorb, reflect, or scatter light at that point in the surface. These are like some simple light transport models we described earlier, but adapted for rendering 3D scenes.

We seem to be very sensitive to subsurface detail. In 2001 Jensen, Marschner and Hanrahan rendered glasses of skimmed and whole milk, and other white liquids. Skimmed milk has proteins that scatter light. Whole milk also has fat globules that scatter. The difference is subtle, but we see it clearly. A white liquid with no subsurface scattering looked like a glass of correction fluid: we would not drink that. It is important to get this subsurface scattering right. And, to get this right, we have to be looking at the subject from the right angle.

A more general model uses a bidirectional surface scattering reflectance distribution function or BSSRDF. This is similar to the BRDF except that the light may leave the surface from a different point to where it went in. The Kubelka-Munk model had explicit solutions for subsurface light transport in one dimension. There is no nice solution for subsurface transport in three dimensions, but it is possible to fit a model to a

uniformly lit homogeneous slab. A ray of light striking the surface at a point produces reflections and scatters from a fuzzy region about that point. Objects rendered with BSSRDF look ‘softer’. Edges need their own model: light striking the upper surface of (say) a glass of milk gives a bright band below the surface when we view it from the side. This may sound like a subtle detail, but we are good at spotting things like this.

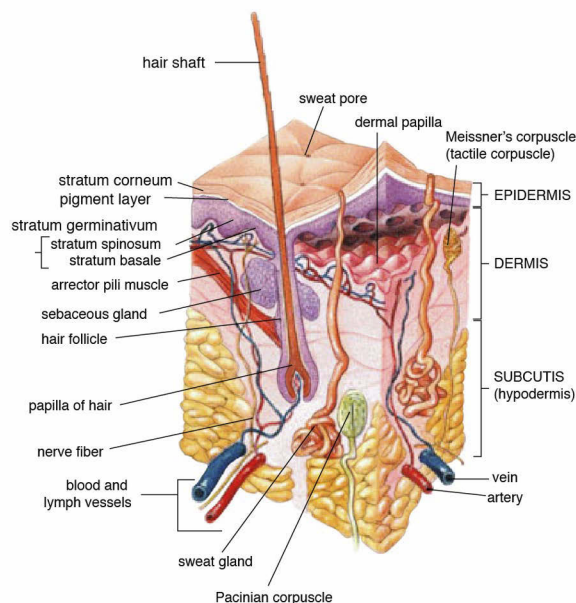
BSSRDF is effectively a surface calculation: there is no explicit calculation of light in the bulk, so it can work fast enough for interactive graphics.

There is a lot of work on new render techniques. Any summary I write will swiftly be out of date. For more on BSSRDF see

<https://graphics.stanford.edu/papers/bssrdf/bssrdf.pdf>.

Skin – a hard problem

This is about as bad as it gets. Skin is metamereric. It is transparent with several coloured layers. It has 3D structure. And it is important we get it right because we are good at spotting errors in faces. At least it does not fluoresce, as far as we know...



Skin structure

I wanted to see if I could simulate the average colour of skin, using some simpler models. This would be handy for retouching skin tones, or for checking that skin tones in images were realistic. I started with the skin model from...

http://graphics.ucsd.edu/~henrik/papers/skin_bssrdf/skin_bssrdf.pdf

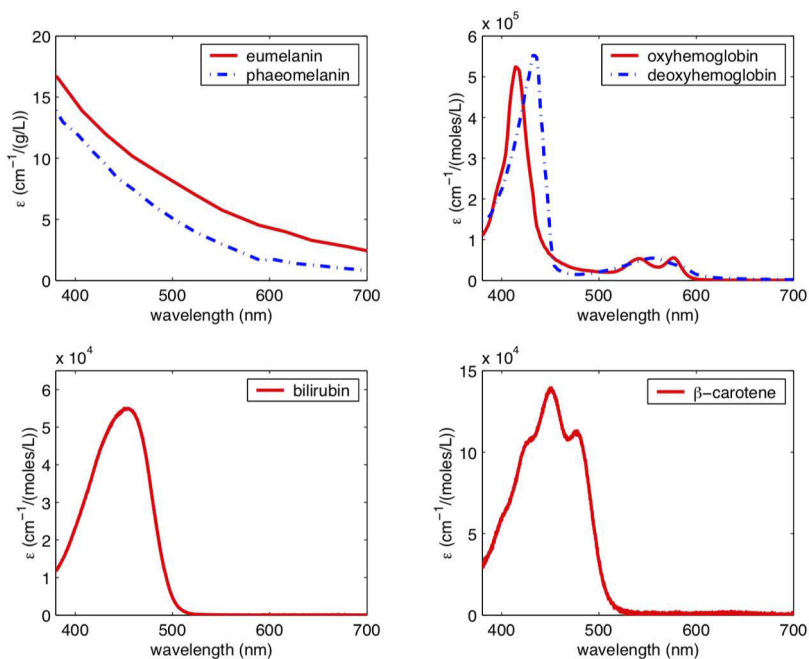
Starting from the lowest level: there is definitely structure we can see below the dermis: I can see the blood vessels in my hands as I write this.

Donner & Jensen assume a semi-infinite dermis thickness, and factor the blood colour (oxy- or deoxyhemoglobin depending on the circulation level) into an average dermis colour. The tiny dermal blood vessels give most of our average light skin colour. We can see the deeper blood vessels, but unless we model them in detail, their average contribution to the skin colour is similar to having a semi-infinite dermis with a slightly different colour. It won't be Monte-Carlo accurate, but it may do. The Donner & Jensen dermis has an average scattering model to cover all the internal features. I used a Kubelka-Munk type model for the dermis because I was not trying to match the sideways light transport.

Donner & Jensen model the epidermis as uniform layer 0.25 mm thick, with a light mean free path modelled on scattering theory. The scattering mean free path for red light is about 0.25 mm according to their model, so there will be significant scattering. The epidermis also contains two types of melanin: eumelanin and phaeomelanin, with different colours. The melanin is in grains rather than dispersed as a dye, so it ought to add to the scattering, but this is not modelled.

The surface properties depend on the oil level. Donner & Jensen modelled this. This seems to be an important part of matching the light into the skin surface. If I left this factor out, my predictions of skin colour were too light and too neutral, which suggests I overestimated how much light got scattered back by the skin surface.

These plots show the absorption of the skin pigments...

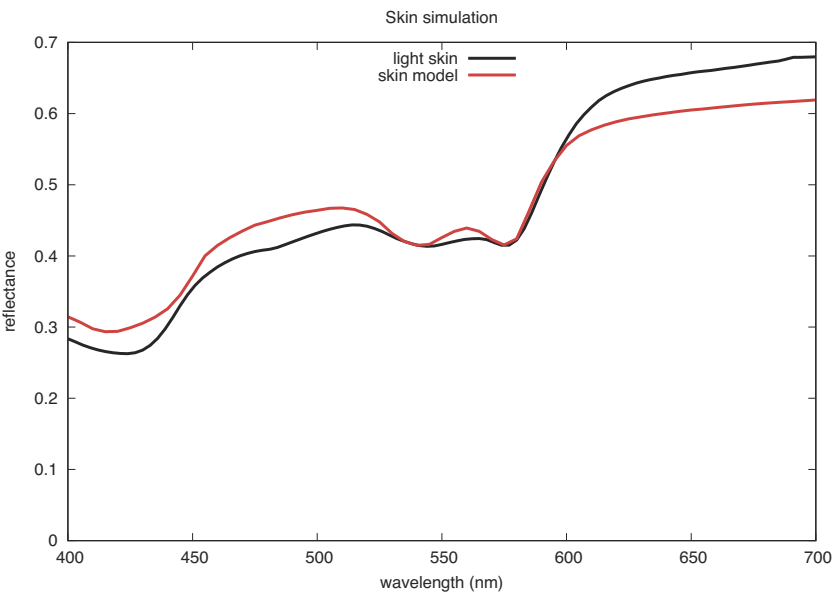


Skin pigments

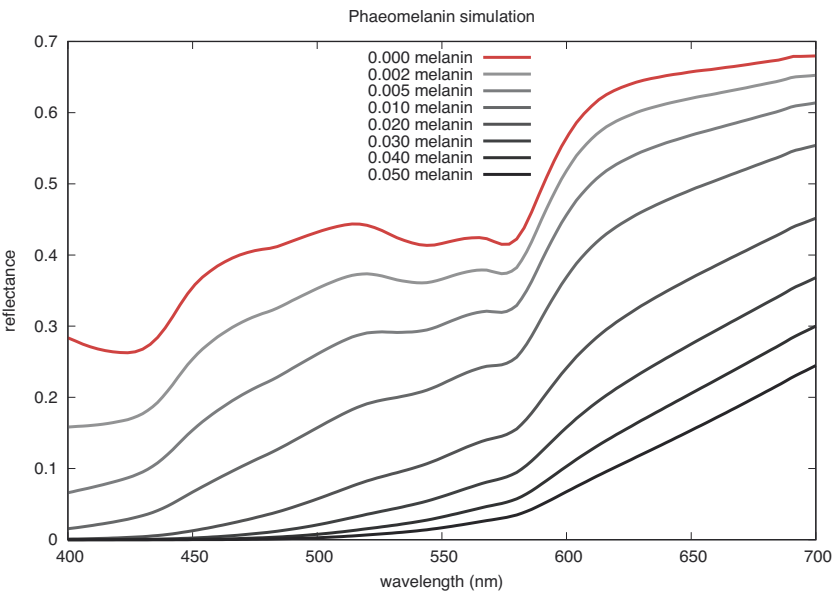
A low value means the wavelength is not absorbed. Eumelanin, for example, absorbs more blue than other wavelengths, so it looks yellowish. Bilirubin and beta-carotene should not be present in healthy skin, but can make it red if you have a rash. These were not modelled by Donner & Jensen, and they are not usually significant in a healthy person.

I had some measured skin reflection spectra. I took the lightest skin tone I could find, so I was not modelling melanin. It was not pink so bilirubin and beta-carotene should not be a factor.

Here is the best match I could get using Donner & Jensen’s data...



This may look a rough fit, but this is no triumph as all the main features were all there in the original spectra. What if I had modelled darker skin? Here is the same model with varying amounts of added phaeomelanin...



The melanin in the model is in the epidermis, so it masks the colours in the underlying dermis. As we add melanin, the characteristic double-peak of oxyhemoglobin disappears, followed by most of the other spectral features. Eumelanin does much the same. If we have separate

parameters for the two types of melanin, we could probably find a convincing match even if the science was no good. If you need a reflection spectrum for skin, go to a database of skin measurements and interpolate from the closest matches. And the best of luck to you.

Chapter 11

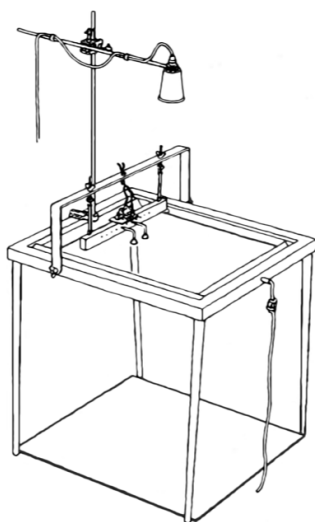
Light waves and particles

I have been talking about light as a wave and light as a particle without explaining how it can be both. This takes you into quantum physics. This is counter-intuitive, but a willing suspension of disbelief can be worthwhile. If you are here just for the colours, you can skip all this without missing anything you actually need. Or you can skim it for the general story. There isn't going to be a test. Really.

Thomas Young and the Wave Nature of Light

We met Thomas Young in Chapter 2. He is perhaps one of the first people to see light and sound and water waves as a whole family of similar physical phenomena.

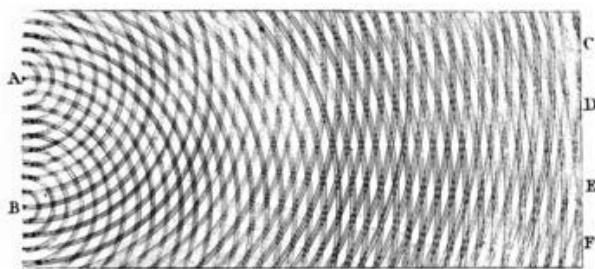
To help him investigate general wave-like phenomena, Young invented a device called a ripple tank. This is similar to the demonstration still used in schools today. A shallow glass-bottomed tank stands on legs. Light from a small source above the tank passes through the water, and onto a white screen below the tank. The waves, if there are any, act as lenses, generating patterns of light and dark on the screen below.



Ripple tank

A bar hangs from springs above this tank. In the illustration, two dippers hang down from the bar into the water. If we tap the bar, it bounces on the springs, generating a regular set of circular waves from the two dippers. The modern school ripple tanks also have a small electric motor on the bar to keep the vibration going, but Young did not have this.

Here are Young's own drawings from his paper to the Royal Society in 1803. With two small dippers A and B, the pattern you get is something like this...



Thomas Young, 'A Course of Lectures on Natural Philosophy'

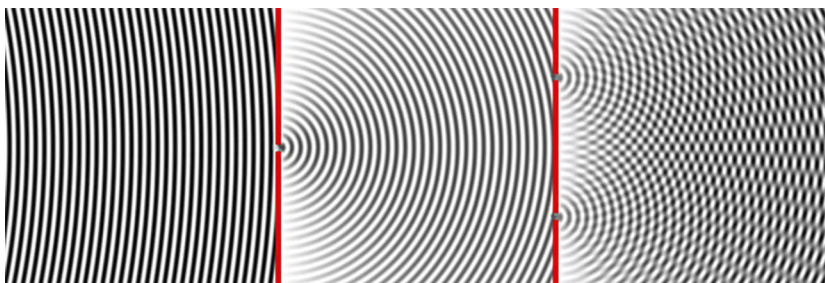
The two dippers are on the left-hand edge of the image. Each dipper is generating a circular set of waves, and the two sets interfere with each other. At the points C, D, E and F, the crest of the wave from A matches the trough of the wave from B, and the waves cancel each other out.

According to Cambridge legend, the experiment was inspired by the interference pattern in the wake from two ducks on Emmanuel College pond. Some sources cite swans instead of ducks. I reckon Emmanuel pond does not look big enough for swans, but it does have ducks. Swans or ducks, the pattern would be different as the source of the waves is moving, but it would look similar.

Thomas Young came in for a lot of criticism in the Royal Society for daring to contradict Newton and his corpuscular theory of light. In retrospect, this seems strange as Newton understood well that light had some wave-like properties, and said so. Newton's rings, right?

Newton's and Young's models do have slightly different predictions. Newton's corpuscular model implies you get no light within the geometric shadow. Young's ripple tank experiments suggest that a small amount of light is able to creep around the corners. At the time, there was no good way of measuring the light levels within the shadows accurately enough to make a convincing case either way. Young needed

to come up with something convincing. In 1807, he published his famous two-slit experiment...



Young's two-slit experiment

Waves go from left to right. Monochromatic waves from a distant source arrive at a first screen with a single slit. A wave spreads out from the other side of this slit, and arrives at a second screen with two slits. The pattern after the second screen is similar to Young's engraving.

It is easy to have two synchronized waves in the ripple-tank. We cannot synchronize two light sources in the same way, but we can make two light sources out of one like this. Young could have used a mirror or a prism to give an image of a second light source, but that would mean one source was the original, and the other was a copy: this way both the sources are equivalent.

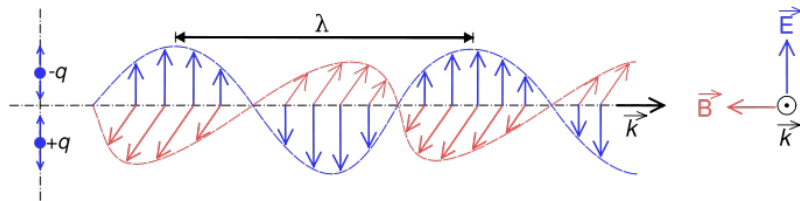
I have done this experiment. You take a microscope slide, blacken it with a sooty flame, and then score a fine line in the soot with a pin. This is the first screen. You do the same with another slide, but rule two lines, perhaps a millimetre apart. You put a sheet of tracing paper on one side. I had a sodium lamp on the other side. You see a dim series of lines on the tracing paper screen. Old low pressure sodium street lights were monochromatic enough to show this, but it does not work with the new ones.

Now comes the critical bit. Cover up one of the two slits, and you get one broad fringe, which is the diffraction pattern from a single slit. Cover up the other slit, and you again get one broad fringe. You only get the interference pattern when you have both slits showing. This means that something must have to pass through both slits in order to make the interference pattern. Whatever light is made of, it has to pass through both slits in order to make the interference pattern.

Maxwell's equations

James Clark Maxwell, the same one who showed the first colour photograph, published a paper in 1865 called 'A Dynamical Theory of the Electromagnetic Field'. By this time, the speed of light had been measured in the laboratory, both in vacuum and in media, and Huygens' theory that light travelled slower in refracting media had been confirmed. Young's experiments had been verified, repeated, and improved upon.

In this paper, Maxwell started with four previously known electromagnetic laws: Gauss' law for electrostatics, Gauss' law for magnetism, Faraday's law of induction, and Maxwell's own modified version of Ampère's law for circuits. He then pulled an enormous live rabbit out of an utterly empty hat by finding useful solutions of these equations for empty space. The solutions were called electromagnetic waves.



Maxwell's equations

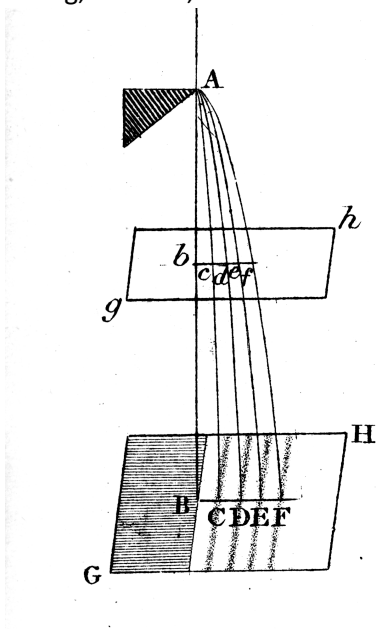
The waves consist of two components at right angles: an electrostatic E , and a magnetic B . The speed of the wave comes from two constants: one for E and one for B , for the material (or vacuum) that the wave is travelling through. These constants can be determined from electrostatic and magnetic experiments, and had no obvious connection with the speed of light. And it also explained polarization, and how reflection separated the polarizations.

Newton was justly famous for using his laws of motion and gravitation to explain why the Moon did not fall or go off in a straight line, but moved around the Earth in a stable orbit. This was even better than that. And the journalists of the day wrote it up every bit as badly as they would write up relativity and quantum physics a few decades later.

If light was a particle, it would have to go through both slits in Young's experiment to give the interference pattern. That is just silly. At this point, waves seem to have won. But...

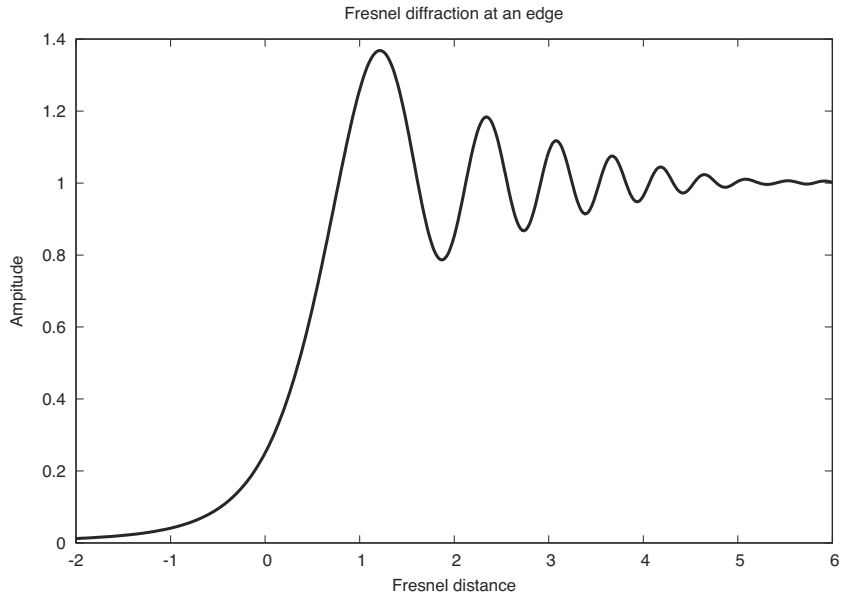
Knife-edge diffraction with waves and particles

Francesco Grimaldi (published posthumously in 1665) recorded seeing faint fringes close to a knife edge. He was able to identify four fringes CDEF on the outside of the geometric shadow B of edge A. Grimaldi moved his screen GH up and down and determined the three-dimensional shape of these fringes. Grimaldi's results were familiar to Young, Newton, and Fresnel.



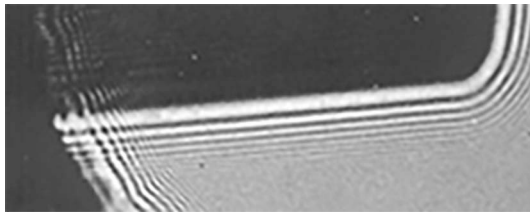
Grimaldi's knife-edge

Fresnel calculated the diffraction of a knife-edge by a wave and found the ripples that Grimaldi had seen. Young demonstrated knife-edge diffraction with the ripple tank.



The graph vertical axis gives the light level: the horizontal axis is the offset in units scaled to the diffraction plane. The geometric shadow edge (***b*** and **B** in Grimaldi's diagram) is at zero on the horizontal axis. We see the ripples on the right, but the shadow of a pure edge on the left has no visible features.

Here are some knife-edge ripples seen in an electron microscope...



Electron knife-edge interference

Yes, electrons are behaving like waves. The microscope has been focussed on a plane below the target, just as Grimaldi did. You can also get knife-edge diffraction effects with ions, neutral atoms, and molecules. The De Broglie wavelength is given by...

$$\lambda_B = \frac{h}{p}$$

... where **h** is Planck's constant, and **p** is the momentum of the particle.

On a traditional transmission electron microscope (TEM) the image is made by electrons hitting a phosphor screen. This is what you look at

when you are focussing the instrument and scanning the target. The electrons are still arriving as particles and dumping their energies at one point on the phosphor. But the wave model maps the probability that the electron will turn up at any particular point. The interference pattern is a real thing you can look at.

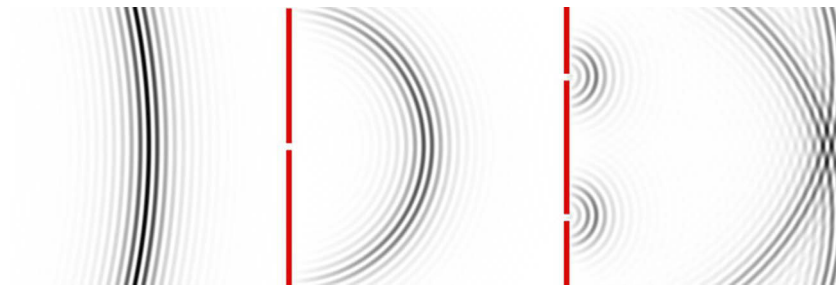
Suppose we turn down the beam current in the microscope. The electron microscope column is probably about a metre high. The electrons in the microscope had been accelerated to 80 KeV, so they are doing more than half the speed of light. If we have a beam current of 1 microamp, the microscope will have 20,000 electrons in the column at any time. If we turn it down to a pico amp then the column is empty 98% of the time. The image on the phosphor is now too dim to see, but if we take a long exposure, the image we see is just the same as we would get with light.

There are people who have done a full two-slit experiment in an electron microscope, and you get light-like interference fringes, which are not affected by the beam current. The electron as a particle may not go through both slits, but each electron wave function certainly does. It's not as mad as it sounds.

We do not see real objects as waves. Our 80 KeV electron has a De Broglie wavelength of only 4 pico metres, or 1/100,000th of the wavelength of violet light. Real objects like molecules do not have a noticeable wavelength unless they are moving very slowly and their temperature is near absolute zero. It's a bit like relativity, but at the other end of the scale: the very little things are a bit odd, and the ordinary sized and familiar remains just the same.

On the other hand Maxwell's laws, beautiful as they were, had all sorts of problems even at the beginning of the twentieth century. Rutherford backscatter had shown how tiny the nucleus was. If electrons were orbiting a nucleus like planets orbiting the sun, they ought to be emitting radiation and spiralling in. Elements had sharp lines in their spectra when electrons changed energy levels, rather than slowly rising or falling. A good model for blackbody radiation fitted beautifully in the red but went completely wrong in the UV. There is no going back.

The Young's slits experiment gets a bit less paradoxical when we start thinking about real particles. The De Broglie wavelength will only apply to an infinite wavelength. Let us replace our infinite waves with small wave-packets like this...



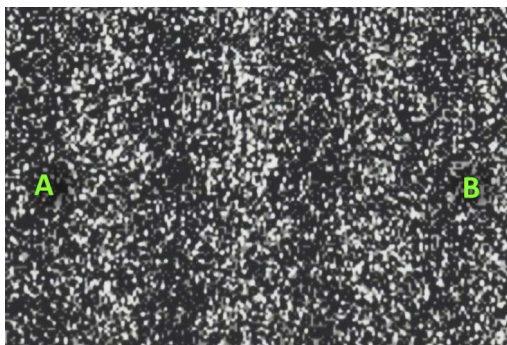
The two-slit experiment with wave-packets

As before, the waves move from left to right. A wave packet approaches the first screen. A second wave packet spreads out from the slit in the first screen. Two wave packets spread out from the slits in the second screen. Finally, two wave packets meet the screen on the right.

In the middle of the screen the two wave packets overlap, and we get interference fringes. At the top and the bottom the two waves hit the screen at different times, the two wave packets do not overlap, and we do not get interference fringes. The wave-packets do not have a single wavelength like our original, infinite waves; they have a spectrum with a peak. This spectrum peak is narrow for a broad wave packet, and broad for a narrow wave packet. We can know the wavelength of a broad wave packet accurately, but not know accurately where it is. We can say accurately where a narrow wave packet is, but not know its wavelength accurately. If it was a particle, we could not know its position and its speed. If you have heard of Heisenberg's Uncertainty Principle, this is where it comes from.

Remember: we cannot get to see or measure the wave-packet itself. All we ever get to see is the point where the particle arrives. If we repeat the experiment with many particles we can plot out the intensity of the wave-function, but we never get to know the phase of the wave.

Here's a real two-slit experiment with electrons...



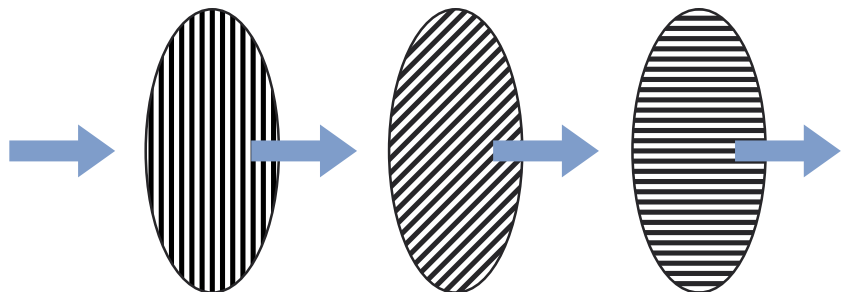
Dr. Tonomura's two-slit experiment using electrons

This is an image from Dr. Tonomura's experiment which he kindly gave to the public domain. **A** may be closer to the left-hand slit than **B**. If we get interference, the wave packet from one slit must be long enough to overlap with the wave packet from the other slit. This is good: we need no faster than light processes to 'see' the other slit, or to check there is no third slit even further off. If we slowly increase the path difference or reduce the length of the wave packet, the wave packets overlap less, the fringes lose contrast, and we approach the single slit distribution.

If we detect the electron at **A**, we know at once it cannot turn up at **B** or anywhere else. Is this a faster than light process? If the wave function only determines the probability that the electron will turn up at any point, there must be something else that makes sure that only one electron turns up in total. Or can we get two electrons, or none? How would we know?

Bell's inequality

In 1964 John Stewart Bell came up with a theory — *Bell's inequality* — along with the proposal for an experiment for testing it. To show how this works, let's look at light passing through several polarizers...



Bell's polarizers

Light is coming from the left. The first polarizer filters out the vertically polarized light. A very fine set of wires running up and down could polarize light, as the conducting wires would absorb or reflect all the light with the electric vector parallel to the wires. This is not how light polarizers usually work because the wires would have to be very fine, but this does work for microwaves.

The second polarizer is at 45° . If this was aligned with the first polarizer, it should pass all the light. If this was at right angles to the first polarizer, it should pass no light. It is half-way between the two. If the second polarizer acts as we would expect for waves, it would transmit only the component of the electric vector parallel to the wires. This transmitted vector would be $\cos(45^\circ) = 1/\sqrt{2}$ of the incident vector. The wave power goes as the square of the electric vector, so the second polarizer

transmits half the light. If we measure light in photons, the second polarizer should transmit half the photons on average.

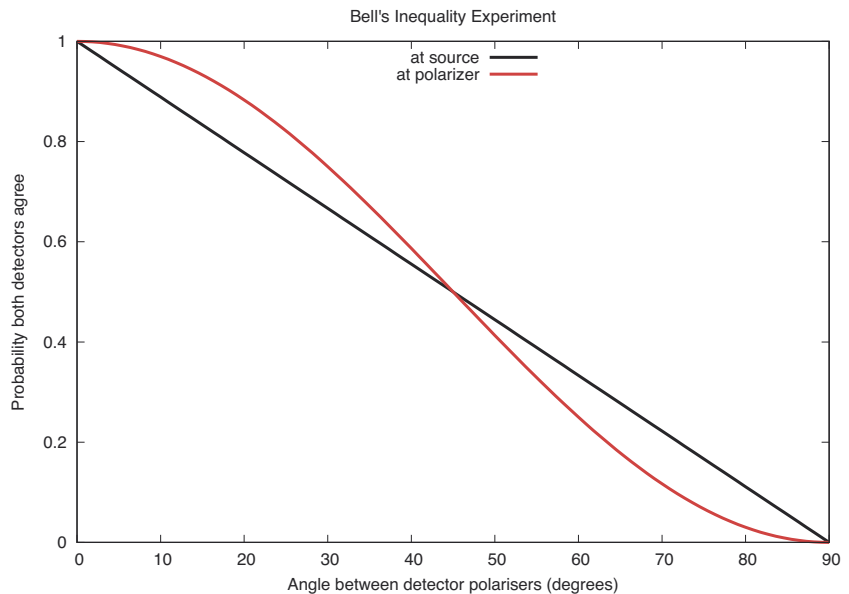
The third polarizer is at 45° to the second polarizer, so it should pass half the light again. If we took out the middle polarizer, the first and last polarizers are at 90° , and will transmit no light. But all three polarizers should transmit a quarter of the light.

Now we dim the light, so we have separate, single photons. If each photon has its own polarization state, any photon that gets through the first polarizer should not get through the last one. But photons do get through when we have the middle polarizer. Maybe a polarizer 'twists' the polarization of the photons that get through? Bell avoided having to consider that possibility because he started with a pair of photons with the same polarization.

Non-linear optics uses a special crystal that can turn one photon into two photons. This is a very inefficient process, so it plays no real part in colours we see, but it is just what we need here. If we put in a 400 nm violet photon, we should get out two 800 nm near infrared photons to conserve energy. The photons must come out at a particular angle if the electric vectors are conserved. And the two photons have the same polarization as the original photon.

Bell's experiment is neat as Young's two slits. Two photons come from the same source, so they have the same polarization, but we do not know what it is yet. We look for the two photons using two detectors with polarizers at different angles. We do not have to worry that the polarizer somehow 'twists' the polarization of the photon because each photon only sees one polarizer.

Bell saw there were two likely results...



The black **at source** line shows what would happen if creating the photons fixes their polarization. The red **at polarizer** line shows the outer limit of what happens if the polarization of the two photons is fixed when the first one meets a polarizer. Any significant deviation from the black line in the direction of the red line shows creating the photons did not fix their polarization.

We get the same result when the angle between the polarizers is 0° , 45° or 90° . If we use another angle such as 22.5° and average many events, the probability is significantly different, and a real experiment approaches the **at polarizer** value. We can make the left-hand photon meet its polarizer before the right-hand one, or vice versa: it makes no difference.

That last bit is important. Some people argue that quantum mechanics is not compatible with relativity. Quantum mechanics is all point actions and events. Relativity is about the warping of time and space, and how we cannot say A necessarily happened before B. But there is no real conflict here. Whether photon A reached the polarizer before photon B depends on our viewpoint, but that doesn't matter for Bell's experiment as the results are the same either way. You only have a paradox if your physical model requires some information to zip between the photons to keep the polarization balanced. Maybe space and time just got lucky here. Maybe someone will find a real incompatibility. If they do it, will be big news. But this isn't it.

This strange linking of two photons, where what happens to one affects the other, is called *entanglement*. Entanglement meant the robin might

'see' tiny differences in magnetic field in Chapter 8. Robins have two free radicals with different spin rather than two photons with different polarizations, but the fates of the two particles from a single event are entangled in the same way.

You may still feel uneasy about a lot of this. Is a wave-function 'real'? What actually stops a particle turning up in two places? You are in good company. The literature is often little help. The *Copenhagen interpretation* said wave-functions *collapse* into particles when something is *observed*. The definition of '*observed*' is deliberately vague: it sounds as if we change things by looking at them, but maybe the possibility that they can be observed is enough. The *Schrödinger's cat* thought experiment was an argument against the Copenhagen interpretation, and not a description of the wonder and mystery of it if you could really make the box it needs. Einstein famously said "*God does not play dice*" but he was against the Copenhagen interpretation, and his side lost. Maybe all outcomes are equally 'real', but the others exists in *parallel universes*? This is fine science fiction, but if these universes do not interact with ours other than just by existing so our one is the *other one*, can they meaningfully be said to exist? Some argue that the wave-function is *holographic*, or contains *hidden variables* so it can be deterministic and yet give us the quantum result. Wondering about things such as this is a proper part of science, but the consensus seems to be *this seems to be the way it is*, and *we see no paradoxes*. Woo abounds: take any literature with *collapsing wave-functions*, *hidden variables* and *parallel universes* with a large pinch of salt.

Very small things aren't quite like bigger things. Your common sense trained on bigger things, and it may not be a good guide. Get used to it.

What actually happened to 'classical' physics?

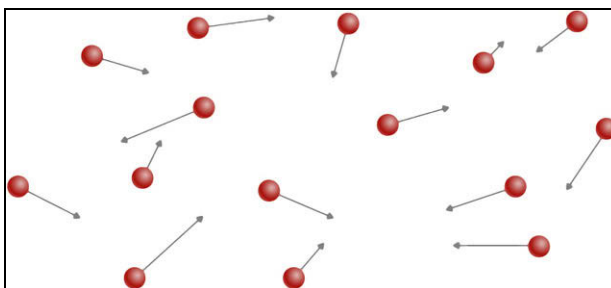
The mathematician W. W. Rouse Ball in his 'Short Account of the History of Mathematics' (4th edition, 1908) said light seemed to be a wave and Newton's corpuscular theory with some wave-like properties 'is now known to be untenable'. Just when everything looked nice and tidy, things started to go very peculiar indeed.

Hot atoms emitted particular wavelengths. Cold atoms could absorb the same particular wavelengths. A sodium lamp emits yellow wavelengths, and sodium atoms in the sun's photosphere absorb the same wavelengths, giving us the Fraunhofer D lines in the solar spectrum. In broad terms this seemed to make sense. A tuned oscillator such as a piano string or a bell which emits certain frequencies when struck will also vibrate when excited by the same frequencies. It seemed sensible that atoms could absorb or emit characteristic frequencies like a piano string or a bell.

When you looked at the details, things were less clear-cut. Some held that atoms did not exist. Others held that hard electrons orbited a central nucleus like planets orbiting a sun. As an electron went from one orbit to another, it emits light with a very particular wavelength. If an electron orbited a hydrogen nucleus (proton) like a planet orbiting the sun, the two would provide an oscillating electric and magnetic field. Classical mechanics suggests they would radiate energy, and spiral towards each other, and the radiation frequency would rise. However, the atomic structure is stable, and electrons do not spiral into the nucleus, so people accepted that there was some other unknown process that operated at the atomic scale, and left it at that.

In 1894 a physicist called Max Planck started investigating the spectrum of incandescent lamps. The electricity companies wanted to know what was the best operating temperature of a light bulb, which would be helped if you could predict the spectrum you would get.

At the time, people thought atoms and molecules were convenient theoretical constructs, but maybe little more than that. Chemists made ball-and-stick models of chemical compounds, but many physicists of the day doubted that atoms were actual, physical things rather than a mental construct: if you tried to model them as hard things, then you got silly results. Maxwell thought they were real, and so did an Austrian called Ludwig Boltzmann. He came up with the kinetic gas model, which leads to the equipartition of energy, which in turn gets us to blackbody radiation, so we will make a small digression into kinetic gas theory...



Kinetic model of an ideal gas

Suppose you have a box of an ideal gas made of small, round atoms bouncing at random against each other, and the walls of the box. Each atom has a random amount of energy, but the average energy determines the temperature. If we know how much thermal energy is in the gas, we know on average how fast the atoms are going, and hence how much pressure they will exert when the atoms bounce off the walls of the box. Boltzmann also calculated the distribution of energies you ought to get if the energy distributed itself randomly. The

Maxwell-Boltzmann distribution drops off exponentially with energy for large energy levels.

Helium comes pretty close to this ideal gas at room temperature, and from measurements of the helium, Boltzmann was able to determine that the gas atoms moved up and down with an average energy given by...

$$e_z = k \cdot \frac{T}{2}$$

...where k is the Boltzmann constant ($1.380\,6504(24) \times 10^{-23}$ J/K) and T is the absolute temperature of the gas in Kelvin. This is true not just for Helium but for any simple, ideal gas. The gas atoms have on average the same amount of energy in all three axes, so the total energy of each atom would be on average...

$$e_{xyz} = 3k \cdot \frac{T}{2}$$

This gives you the energy needed to raise the temperature of a known number of atoms of an ideal, monatomic gas by one degree at constant volume. All the noble gases (He, Ne, Ar, Kr, Xe) are a good fit to the kinetic gas model, and have almost the same specific heat for constant volume.

Now, suppose your gas is diatomic, like nitrogen or oxygen. You still have three translational degrees of freedom with $kT/2$ energy, but the gas molecule can also rotate about two axes. Boltzmann's equipartition theory says that each degree of freedom should have on average the same energy, so each molecule of oxygen or nitrogen would have five units of energy on average...

$$e_{xyzab} = 5k \cdot \frac{T}{2}$$

Some diatomic gases come close to this figure: hydrogen, and carbon monoxide are within 10%, and nitrogen is only just outside, but some others are not so close. In its early days, many people distrusted the kinetic theory and the principle of equipartition. Why were there only two rotational degrees of freedom? The diatomic molecule — even a single atom — has some width, so why not include that too?

There was another famous problem. Lord Rayleigh and Sir James Jeans applied Boltzmann's equipartition theory to an incandescent blackbody. A solid, hot body will not have the kinetic energy degrees of freedom that the ideal gas has, but it will have a number of vibrational modes, and each of these modes ought to have an energy of $kT/2$. The black body is three-dimensional, so the number of vibrational modes should rise as the

square of the frequency. Classical electromagnetic theory suggests that each vibrational mode should emit the same power, so the power in each frequency band should go up as the square of the frequency too.

An incandescent lamp spectrum fits the Rayleigh-Jeans theory in the infrared, but it reaches a peak and then decreases in the ultraviolet. The Rayleigh-Jeans theory predicted the emission should continue to rise through the ultraviolet and beyond. We saw earlier that classical physics said electrons ought to spiral into the nucleus, so it is no great surprise that the Rayleigh-Jeans theory should go wrong as the wavelengths approached the wavelengths we see with atomic spectra. This discrepancy was melodramatically named “the ultraviolet catastrophe” (SFX: sinister organ chords, roll of thunder, manic laughter).

Planck realized that something was going wrong with the very short wavelengths, and this was probably connected to the fact that electrons only gave off certain characteristic wavelengths when they changed their energy states. Suppose the wave packet had a fixed amount of energy for a fixed wavelength? If this energy came from a vibrational mode in a black body, and the vibration mode had a Maxwell-Boltzmann energy distribution, the odds it had enough energy drops rapidly as the energy goes over $kT/2$. If you need four times the average thermal energy to make one ultraviolet light packet, then you are not going to get many of them.

This turns out to be the solution to Boltzmann’s problem with the kinetic gas model too. You can put energy into rotating a gas molecule, but only in discrete steps. With a diatomic molecule, you have two ways of rotating the molecule lengthwise, so the discrete energy steps are small, and you almost get Boltzmann’s prediction of $kT/2$. The molecule can also rotate about its axis, but the discrete energy steps are much larger. The Boltzmann probability of having enough energy to get to even the first energy step is small, so the average energy in this mode is much less than $kT/2$.

Planck assumed the energy E was proportional to the frequency ν (Greek letter ‘nu’) as follows...

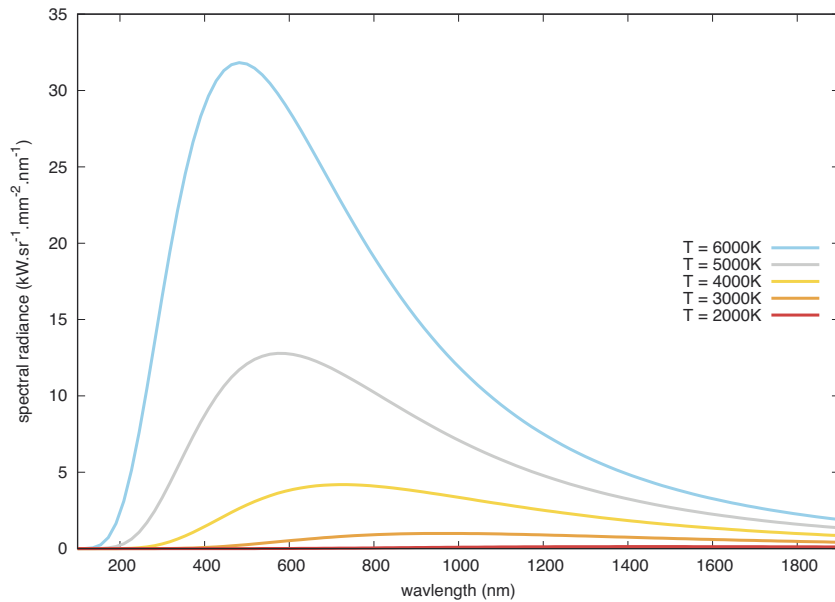
$$E = h \nu = \frac{hc}{\lambda}$$

...where h is Planck’s constant (modern value 6.626×10^{-34} J/s).

By 1900 Planck had a solution for the intensity I at wavelength λ that fitted the experimental data nicely...

$$I(\nu, T) = \frac{2hc^2}{\lambda^5} \cdot \frac{1}{e^{\frac{hc}{\lambda kT}} - 1}$$

...and he could work out the best design for a light bulb. He also provided the first successful assumption that went beyond classical electrodynamics and is seen as the birth of quantum theory. He was given a Nobel Prize for Physics for this work in 1918.



Planck's blackbody spectra

Planck, however, had a deep respect for classical electrodynamics. He had assumed that light came in little packets or 'photons', without any belief that this might be real, but just to fix Boltzmann's equipartition principle which most people of the time did not believe anyhow. He spent the rest of his life trying to get the monster he had created to fit in with classical physics.

Boltzmann suffered from depression later in life. He became worried that he had somehow 'broken' classical physics, or sent it in the wrong direction. He studied philosophy to see how this might help make his quantization ideas acceptable to himself and the world in general. He became depressed again, and in 1905 while on holiday in Italy with his wife and child, he hanged himself.

If you feel we should go back to the 'classical' science of 1900 and somehow fix things, remember Boltzmann and Planck: they were the smart guys who got it right at the time. It is all rather sad. But science probably took the right path in the end.

The photoelectric effect

My introduction to quantum effects did not follow the actual history very closely: I skipped straight to the 'all particles are like waves' bit as that was the bit we needed. The real history had many of the scientists of the day trying to patch up particle theory.

The first serious evidence for light as a particle actually came from work on light detecting valves, and the discharge of gold-leaf electrosopes. If you shone light at a metal surface, it emitted electrons for some wavelengths of light and not for others. The light seemed to arrive in little packets with an energy that depended on the wavelength. If your packet had more than some threshold amount of energy, it could lift an electron from the surface. If it had energy to spare then the electron came out with some kinetic energy. If it did not have enough energy then no amount of light would produce an electron. Einstein got his Nobel for this work in 1921, and not for relativity.

The Stele of Lady Taperet

The original is in the Louvre. It dates from some time in the Third Intermediate Period of Egypt between 1096 and 664 BCE. A woman worships Ra-Horakhty, who blesses her with light.



The Stele of Lady Taperet

I particularly like the way they drew light as discrete, coloured quanta.

Chapter 12

What is seeing?

This title brings us back to the beginning. Those of us fortunate to have good colour vision have always had it. Vision is a process that probably formed as our own consciousness did. We use the eye and brain to examine other things. Can it examine itself?

The Stroop effect

It is worse than that. We cannot untangle the parts of the brain that process colour from the parts of the brain that are reading this...

control	compatible	incompatible
fish	red	red
leg	blue	blue
car	green	green
seat	yellow	yellow
plate	blue	blue
book	red	red
hat	blue	blue
cake	green	green
cat	yellow	yellow
paint	red	red

Try to name the colours of the words in each column. The first two columns should give no difficulty, but you may stumble or slow down on the last column, where the colour words contradict the actual colours. There are several theories for why this happens, but they mostly reduce to a common model. Our brain wants a word for the colour. We look at the colour of the word, and we read the word of the colour at the same time. In this case, reading seems to be faster than looking. In the

pandemonium model, the reading demon shouts out the answer before the vision demon is ready.

Why is reading faster? Reading uses the faster, parvocellular processing. Perhaps reading directly generates a link to a word, while colour gives a stimulus that we then have to name. There are lots of theories and lots of research on the Stroop effect.

Is seeing affected by the other senses? Yes it is. We process sound much faster, in part because the nerve pathways are shorter. This means what we hear ought to be about one tenth of a second ahead of what we see. Yet, when we see people speak, the sound and the vision appear in sync. Ventriloquists have used this for ages, putting their speech out of sync with the motions of their puppet. The puppet has large lips, so our brain cannot resist locking the speech to it, and we miss the timing of the speakers p's and b's.

The Sapir-Whorf hypothesis

The Sapir-Whorf hypothesis, otherwise known as *linguistic relativity*, holds that the language we use, and its structure affects what we understand.

The *strong* hypothesis says we think in words, and language determines what we understand. If we do not have words for something, then we do not understand it; perhaps we are not even aware of it. We can list, for example, how many colour terms any language has, and use this to rate how 'advanced' the culture is. Yes, this is rapidly going somewhere nasty: languages that only have words for 'light', dark', and 'red' are the most primitive, while English has lots of words for colour, hurrah. The strong hypothesis, or *Whorfianism* was in fashion around about 1900 but has fallen from favour.

The weak hypothesis is that we do a lot of thinking in words, and the tools in our language may help us understand. Some quotes of Sapir suggest he favoured this weak hypothesis.

First, let's get rid of the spectre of Whorfianism...

In 1969 Berlin and Kay published 'Basic Colour Terms, Their Universality and Evolution'. They asked speakers for the colour terms in their language, and to identify which colours from a chart, with hue horizontal, and lightness vertical. English looked like this...

	SR	10R	5YR	10YR	5Y	10Y	5GY	10GY	5G	10G	5BG	10BG	5B	10B	5PB	10PB	5P	10P	5RP	10RP
9			Pink		Yellow															
8	•			•																
7			•																	Pink
6			Orange																	
5														Blue		Purple				
4	•			Brown				Green						•						
3		Red															•			Red
2																				

Berlin & Kay: English colour terms

The colour divisions are mostly vertical, but 'Pink' is light 'Red', and 'Brown' is dark 'Yellow'. 'White' would be at the top, and 'Black' at the bottom. There is no 'grey' as the chart has no saturation axis.

Berlin and Kay ranked languages in seven stages according to the colour terms they contained...

- 1 Dark-cool and light-warm (much looser than 'Black' and 'White')
- 2 Red
- 3 Either green or yellow
- 4 Both green and yellow
- 5 Blue
- 6 Brown
- 7 Purple, pink, orange, or grey

Berlin and Kay's methodology has been criticised, and they have since loosened their definitions of the stages. However, other people have repeated their work and the general conclusions seem to hold. You do not find a language which has a word for 'blue' unless it already has words for red, green, and yellow.

	5R	10R	5YR	10YR	5Y	10Y	5GY	10GY	5G	10G	5BG	10BG	5B	10B	5PB	10PB	5P	10P	5RP	10RP
9	3	2	5	Wap	2				1	1			Wap	1	5	12	6	3		2
8				9	6	2	3										1			
7			2	5	4	4	1	1		2										
6		Meh		2	1			2	3		1									
5	6	2			Wor	1		6	7	4	Nol	2	2					Meh	3	
4	19								5			3								11
3	2			Kel							1		1		1					1
2		1	1	3	4	6	12				2	1	4	3	4	4	Kel	2		

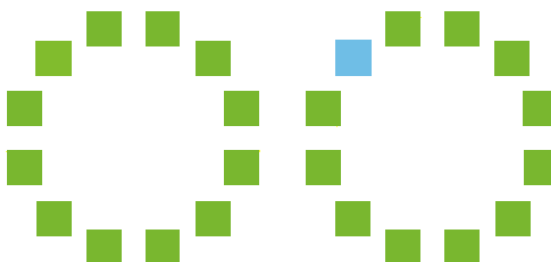
Kay and Regier: Color naming universals: the case of Berinmo
<https://www1.icsi.berkeley.edu/~kay/berinmo.pdf>

Here is one famous (or notorious) exception: the Berinmo language of Papua New Guinea. This may have colours ‘Wor’ that covers Green and Yellow, and ‘Nol’ that covers Blue and Green. Berlin and Kay did not study Berinmo. Indeed, one of the criticisms of Berlin and Kay’s original study was that most of their subjects were bilingual speakers from technical cultures, who lived in places you could fly to. Such people may filter their ‘colour words’ because of the other languages they know. However, people who wanted to repeat the study have had trouble finding the original Berinmo again. For more objections see Kay and Regier’s paper.

What constitutes a bona-fide colour word? There is no colour between ‘Green’ and ‘Blue’ though most English-speakers would understand ‘Turquoise’. A linguist recalling her experiences in ‘New Scientist’ (I have tried searching for her name, please supply it if you can) remembered discovering that a rainforest tribe she had been talking to for months had a word for ‘violet’. The word existed, but it was rarely used, and they had not thought of it when she had first asked them about names for colours.

So, what might this actually mean? I guess each language probably has some sort of equilibrium, where needed words catch on, and redundant words are forgotten. The sky is blue for everyone, and plants are all green, but ‘Red’ is named first, and perhaps ‘Yellow’ second depending on their salience. If you only use ‘Blue’ for the colour of the sky because you have no blue materials, then the word will not get used much. People may not use the rare terms consistently. For example, in English, ‘livid’ is a colour. It came from the Latin, *lividus*, a blue-grey colour, like bruised skin. It could also mean purplish. But bruises change colour. In medieval times ‘livid’ was used for pale orange, yellow, or ivory. Nowadays *livid* often means scarlet or red. But it is also used as a term for a ‘wrong’ or ‘unhealthy’ colour: J.R.R. Tolkien and P.D. James both use the term ‘livid green’, and they are both good enough writers not to confuse *livid* with *vivid*. So, if you ask someone to list all the colours in their language, what you get depends on how you ask and how patiently you listen.

The Himba people live in Northern Namibia. They were the subject of a 2011 BBC documentary “Do you see what I see?”, which claimed they had less experience of blues, but were very adept in spotting the ‘odd colour out’ between different greens.



‘Do you see what I see?’ colour discrimination test

They claimed the Himba could see the yellower green out in the left-hand diagram (the one in the same position as the obvious mismatch on the right) but not see the bluer one in the right-hand set. The ‘experiment’ in the documentary was a dramatic reconstruction. I do not have the images the Himba saw, but interpolating from video RGB values, the DeltaE on the left is about 4, and on the right is over 100. The Himba have several words for different greens: *dambu* for light yellowish-green, *zuzu* for dark green, and *buru* for all colours between bluish-green and violet. Maybe this is important: a nomadic people may need to recognise small differences in plants, but do not need an everyday word for blue. Maybe their *buru* has been contaminated by our *blue*. Maybe this is all Whorfism.

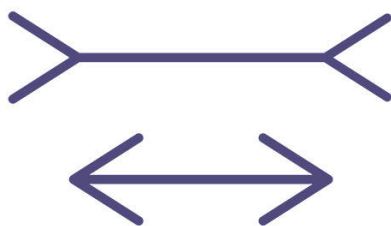
I could spot the odd green out as fast as the Himba on the program. I guess most people who grade motion pictures could do the same; but your average person may have trouble. This is why I wanted to get the word ‘azure’ into common English in Chapter 2; it won’t make our eyes better, but having a word for this colour might mean we interpret what we see more effectively.

Why might Sapir and Whorf think colour depends on language and experience when we all have similar eyes? We all have similar ears, but I cannot distinguish subtle differences in exact vowel sounds while others can. I recognise accents in spoken English, but I have difficulty identifying exactly what the differences are. I probably learned to filter out important words and meanings, and discard the exact vowel sounds when I first learned to speak. I could probably learn to ‘do’ accents and learn to hear these differences, but I have not done yet. It is reasonable to suspect the colours we see may depend on our training and experience. If we work with colour a lot, we learn to appreciate subtle differences in hues and tones, but as far as anyone can tell we do not see any new colours.

Learning to see

Suppose we both look at the same subject. If neither of us has defective colour vision, we both get a similar stimulus within the eye. There are genetic variations in our colour receptors, but these are probably smaller than the differences due to the yellowing of our eye lenses as we age, or the differences in our 'yellow spot' dye. These are all things we can measure and predict, and so don't make for good philosophy. The real problem is how we get from the stimulus to the perception of colour. How can we know whether our perceptions are the same or not?

In time, we may develop new tools for measuring perceptions, and some factors in this problem may cross the boundary from metaphysics to physics or, 'things we can measure and predict'. But much will probably remain mysterious for some time at least. I cannot show that our perceptions are the same, but I might ask what grounds we have for saying they are different. The human visual system has some built-in features, and some freedom to learn. We seem to see in similar ways with similar performances. We exhibit the Stroop effect and other effects to similar levels. This suggests we use similar parts of the brain in similar ways. There are even cases where people have suffered damage to their vision centres, and somehow their brains have re-purposed a neighbouring piece of brain to see as they saw before. There seems to be a particular stability to the way our vision works.

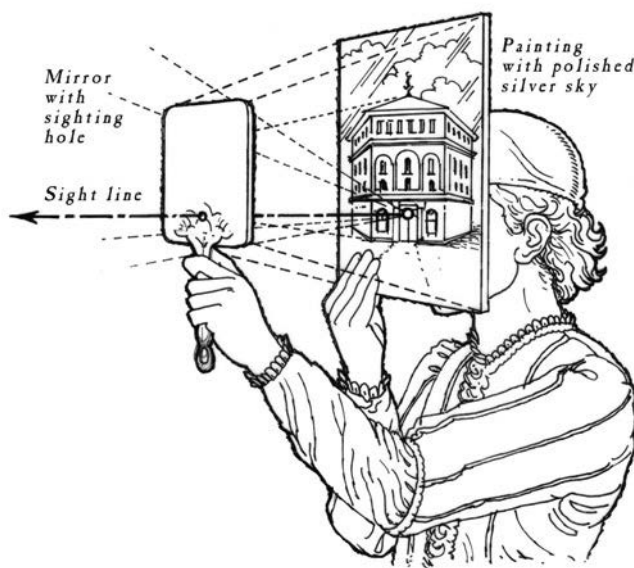


The Müller-Lyer illusion

This does not mean we are identical. The upper line in the Müller-Lyer illusion above will probably look shorter to us if we have straight lines in our environment. This is probably because we see the upper configuration in convex, closer edges, and the lower one in concave, more distant edges. Some people who live in round houses do not see edges in this way, and are not fooled. Experience is important.

We also use our experience when we look at images. People want to draw parallel lines as parallel lines, and expected to see them as parallel in a painting. Brunelleschi in the 15th century had an ingenious

demonstration. He made a drawing of the baptistry at Florence (a very geometric building) with linear perspective. You looked at the Baptistry through a hole in the picture. If you held up a mirror, you could see the reflection of the picture. If you moved the mirror, you could see the edges of the drawing in the mirror fitted the edges of the baptistry itself.



Reconstruction of Brunelleschi's demonstration

Brunelleschi was a practical man. He would show stonemasons how he wanted stones to fit together for his buildings by carving models out of turnip. I imagine his apparatus had the picture and the mirror on a stand, and the mirror slid on a rail. No part of his original apparatus survives.

This demonstration was not to investigate linear perspective; Brunelleschi already understood that well when he did the baptistry drawing. He was showing people, probably rich, artistic patrons, that it was okay to draw buildings on a flat surface like this, and the buildings were not wonky or falling over.

To a lesser extent, we expect colours to be represented in a particular way. We know the sky is blue, and people expected it to appear as blue in their holiday photos. Colour photographs would often render skies as white or grey. Part of this was down to the tone curve of the photographic process, but we also remember the sky as being bluer than it was, and photo print shops had to make skies bluer to match our expectations. This may be because skies are larger than photos, and our peripheral vision can make them seem bluer. But the colour people like to see in photos does not match the actual colour the sky was.

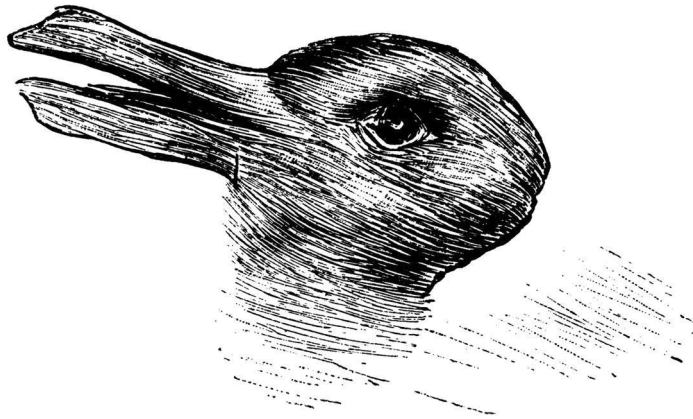
We have probably got used to a particular 'day for night' look from motion pictures. The effect known as 'American Night' used tungsten-balanced film. Indoor lighting used tungsten lamps. Outdoor lighting used dim, diffuse daylight, which looked blue on tungsten balanced film. The Purkinje shift mentioned in Chapter 3 also predicts that night scenes should look blue, but only slightly blue. At even lower lights we lose all sense of colour, but objects we know are blue still appear lighter. But we have got used to 'night equals blue' and so we use the same colours for digital colour correction of night scenes.

We could show night scenes at night light levels. But this is not practical for motion pictures where we may need to tell a story without time for our vision to adapt to the changing light levels. This would also mean we see no detail with our central vision where there are no rod cells. So, we need some convention to represent night with practical light levels. Darkening the mid-tones and adding a subtle steel blue cast would probably be closer, but the vivid nighttime blue is what we have learned to expect.

'That Dress' image and other ambiguities

So, what about this 'dress' image? People with normal vision see the dress as either blue and black, or gold and white. They have the same eyes, so the difference must lie in the brain. But this does not mean there are two sorts of vision. The actual dress always looks blue and black. The 'dress' image was heavily cropped, and there are confusing lighting cues. The dress is either blue and black lit with a bright yellowish light, or it is yellow and gold, lit with a blueish light. We expect blue light, yellow light, or anything along the thermal locus, so both interpretations are reasonable. We do not expect green light or pink light, and versions of this image with the channels swapped do not seem ambiguous. Maybe someone raised in a world where pink and green lighting is the norm might see it, but we do not.

Here is another famous ambiguous image: 'Rabbit and Duck' from the 23 October 1892 issue of *Fliegende Blätter*...

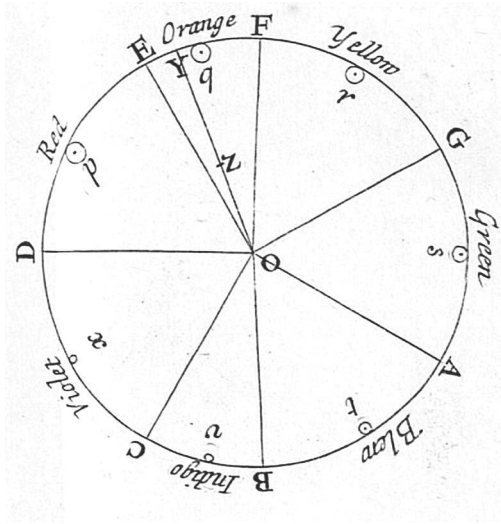


Rabbit and Duck

The duck faces left. The rabbit faces right. We can see it as a rabbit or a duck but not both at once. We interpret it as a rabbit or a duck because we know what rabbits and ducks look like. In this case there is no real answer. If you search for it, there is an unretouched digital image of a black cat that looks like a crow. In this case, the animal really is a cat, and it looks to most of us like a crow. This is the real equivalent of the 'dress' image. We do not say we have two types of vision that see different animals: we have similar vision and common experience, and this causes us to view the image as one thing or the other but not both. It is amusing, even fascinating, but nothing mysterious.

Problems with colour wheels

The colour wheel has been invented and re-invented many times. Here is Newton's Colour Wheel from his *Opticks* about 1706...

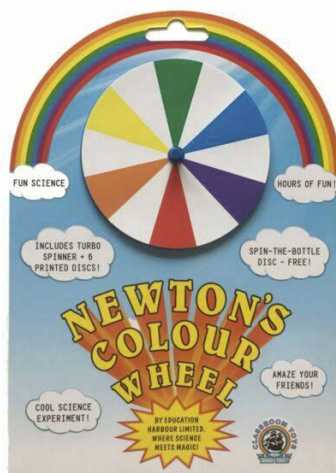


Newton's Colour Wheel

What does this diagram show us? Newton describes it as showing the amounts of the seven spectral colours go from red through blue to violet. There are purples, pinks and other colours that do not occur in the spectrum. They can be made from mixing the limiting colours in the spectrum. But Newton only talks about spectral colours, so we should not include them here. So, why did he join violet to red? The A-G letters around the edge are musical notes. Starting at D we have a musical scale in the Dorian mode. Orange and Indigo are narrow because they are semitone intervals. Violet does look like blue with a bit of red, so as you return to D. Madness! The worst thing is it is almost right: deep red is about 760 nm, and deep violet is 380 nm, so deep violet may actually be an octave up from deep red as a frequency; but violet does not resemble red because it is a harmonic.

In Newton's description, the sectors showed the proportions of the seven colours needed to make white at the centre. There was no way to objectively measure amounts of light back then, so the proportions may be lengths in a projected spectrum. Huygens argued (correctly) that it should be possible to mix white from two opposite colours if light added like vectors. Newton said you could not make a pure white from just two colours. To give him credit, he did try and could not convince himself that

the white was 'pure'. But why did he make a plot with white in the middle if you weren't supposed to mix colours like this?

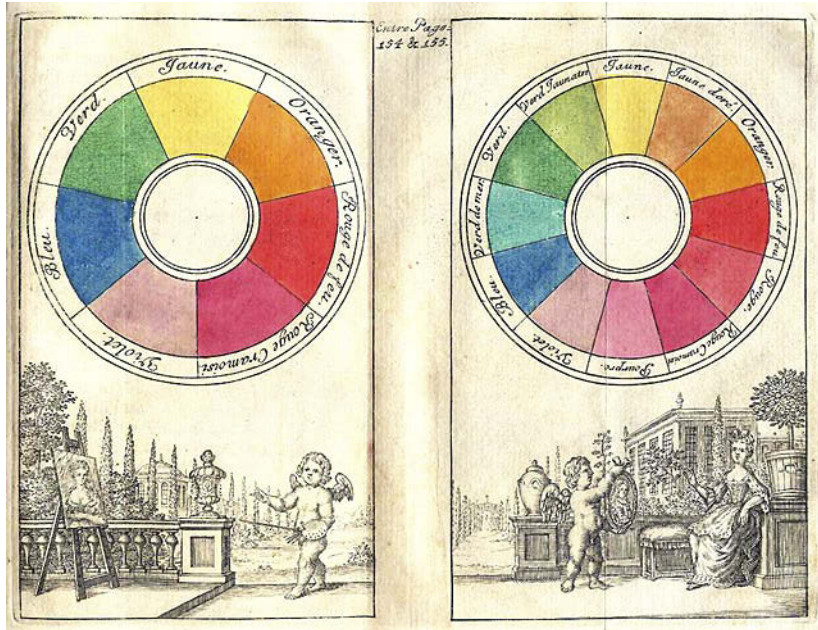


Newton's Colour Wheel hand spinner

There is no evidence that Newton intended you to average the colours by spinning the disc, though that is a possible reason for the sectors. He did describe a different apparatus with prisms, a lens and a large moving comb with teeth that flickered between spectral colours. But not a spinning top. The real pioneer of using rotating discs was Maxwell, who used a spinning wheel to show you could make a grey from spinning red, green, and blue sectors that matched a grey mixed from white and black sectors. It is a beautiful demonstration. But there seems to be a mistaken belief that this is how Newton's wheel worked. I am not even sure the colours opposite each other in Newton's wheel were complements. This could have been proven with Maxwell's colour wheel, but as far as I know, no-one ever did.

Newton's colour wheel appears in his *Opticks* of 1704. This book was in English and much more accessible than his 'Principia'. It contains many fine examples of the experimental method, particularly his 'experimentum crucis' with the prisms. It was widely read and

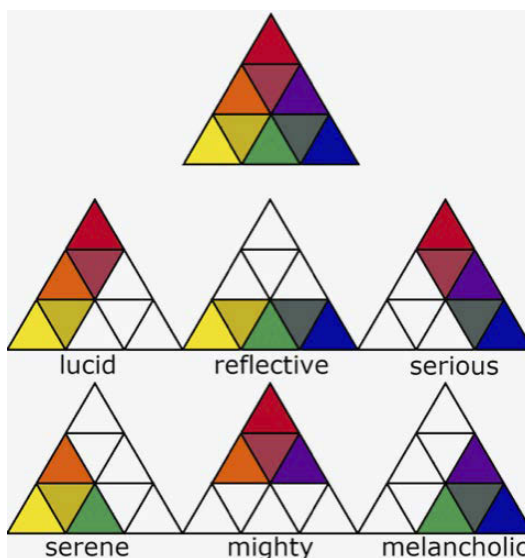
understood. And within four years, colour wheels started appearing in other books...



*C.B. (probably Claude Boutet or its editor, Christophe Ballard) - *Traité de la peinture en mignature* (The Hague, 1708), reproduced in *The Creation of Color in Eighteenth-Century Europe**

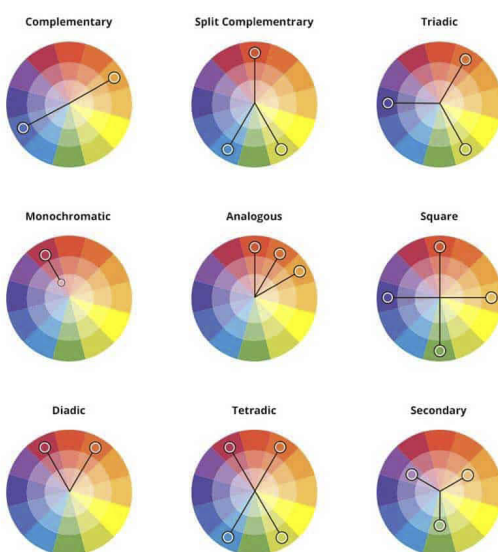
The left-hand colour wheel has seven sectors like Newton. But they are all equal sizes. 'Indigo' has gone, and there is a new 'crimson' that fills the gap between red and violet. This is a natural way of closing the colour circle for anyone who mixes paints or dyes. Intriguingly, Jacob Le Blon was also experimenting with CMY and CMYK printing in the Netherlands at just this time, and he called his primaries 'Red', 'Blue' and 'Yellow' in his 'Coloritto' of 1725. So, instead of having seven primaries, it seems a lot neater to have six or twelve. The three primaries are at 120 degrees to each other, and each colour has a complementary colour opposite it.

This hollow ring of colours is helpful. It gives you a sense of the colours there are. But it does suggest all sorts of geometrical relations between the colours...

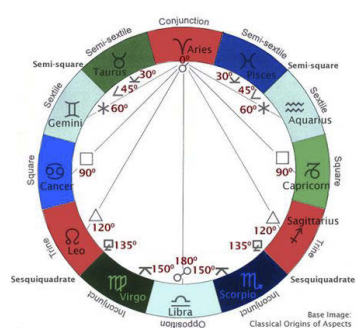


Goethe's colour harmonies

Goethe started from the 6-fold colour wheel. His primaries were red, yellow, and blue. His secondary colours were green, orange, and violet. And there were tertiary colours. He ascribed universal moods and natures to each of these colours. There were also colour harmonies between colours with particular spatial relationships. But dividing the circle into 12 gives all sorts of links into astrology, numerology and worse...



Determining Compatibility



Main Aspects

- Conjunction (same sign) - 0°
- ◐ Opposition (sign directly opposite) - 180°
- ◑ Square (same quality - not opposition) - 90°
- △ Trine (same element) - 120°
- * Sextile (similar element - not opposition) - 60°

Other Aspects

- ≍ Semi-sextile (adjacent signs) - 30°
- ⋈ Quincunx/Inconjunct (opposition adjacent signs) - 150°
- ⌒ Semi-square (halfway between semi-sxt & sxt) - 45°
- ◌ Sesquiquadrate (halfway between trine & inconjunct) - 135°

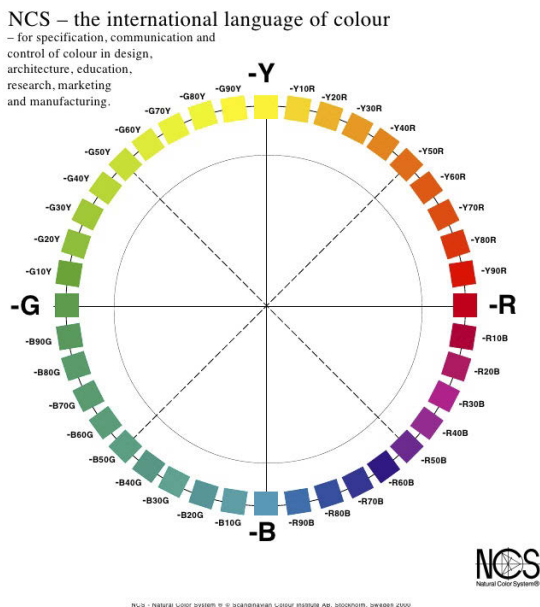
Examples of 'Colour Theory'

Some of this sort of thing is taught as ‘Colour Theory’. Use this for inspiration if you are stuck for colour schemes, but it is not science as I know it.

The Natural Colour System (NCS)

We met the Hering colour opponents in Chapter 8. Red gives an after-image of Green and vice-versa. Blue gives an after-image of Yellow and vice-versa. If we were looking for colour complements we might expect Green to complement Magenta, and Red to complement Cyan. But no, most people see Red pairs with Green; if you do not believe it, try it for yourself with the test in Chapter 8.

The Natural Colour System is an attempt to make a creative colour system based on the Hering primaries. I am not sure how helpful this is. After-images are not an important part of everyday vision; they are something our vision tries to correct for. Red, Green, Blue, and Yellow are also significant named colours. Here, the Green-Red axis is horizontal, and the Blue-Yellow axis is vertical.



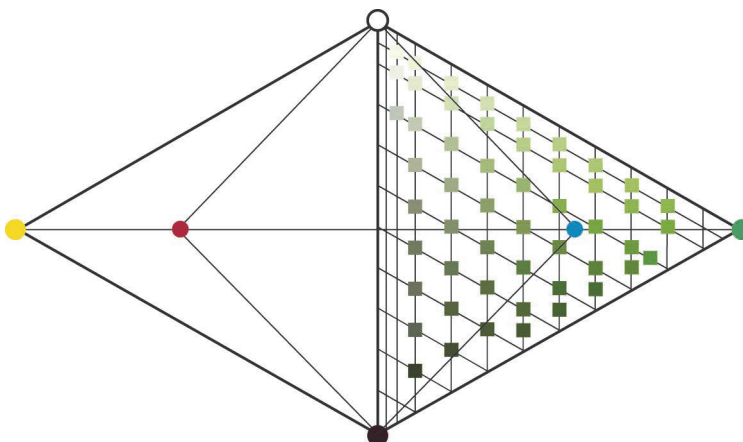
Natural Colour System (NCS)

We know we can get yellow from mixing red and green light. Red, yellow and green should be in a straight line. They are in a straight line on the CIE x,y plots and in our r,g plots. So a blue-yellow line and a red-green line should make a ‘T’-shape, not a cross (Chapter 8 again). A horizontal line through the middle of the blue-yellow would run from green to magenta,

or from red to cyan, or somewhere between the two with additive mixtures. Munsell space has red opposite blue-green, and yellow opposite blue-violet. Munsell space also had incomplete circles where the colour patches could not be made.

The NCS is based on Hering opponents, not on additive mixtures. It can have a circle of colours with red and green, blue and yellow at the four points of the compass if it wants. However, additive mixtures look smooth, and it is hard to keep that smoothness and warp the red-green line to a semicircle.

Here we see something like cyan where blue ought to be, a blue-green where green ought to be, and a slightly pink red instead of red. We do not have a problem with an empty circle. That would be just a ring of colours. But when we try to go into the centre we have problems. And the NCS system is a three-dimensional colour space with neutral down the main axis, like this...



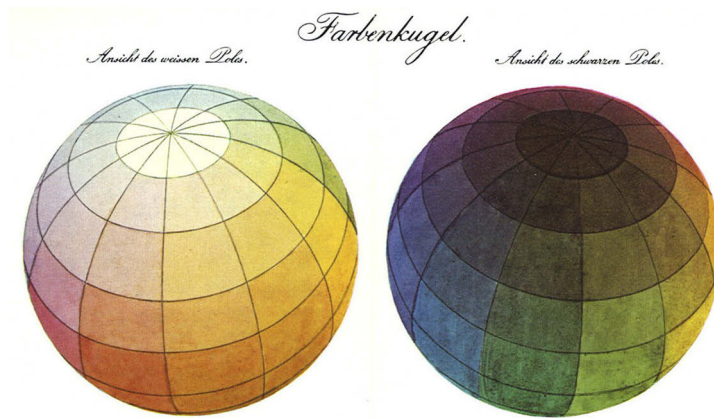
Natural Colour System colour space

I am not saying that the NCS system is wrong or is not useful, or that colour wheels do not serve their purposes. Many people use them to good effect. But it seems whenever people draw colour circles, facts get bent to fit the circle, rather than the other way around. The human vision colour space is nearly triangular. That takes a lot of bending.

The colour sphere

The Munsell colour solid has one point (white) at the brightest end, and one point (black) at the darkest end; and more colours in the middle. Philipp Otto Runge in 1810 described all the mixtures of colours as a

sphere. White is at the top of the sphere and black at the bottom. In the middle we have a brighter range of colours.



Runge's colour spheres

Munsell was aware of Runge's work. Munsell's book of colour in 1929 still had a diagram that described his space as spherical. However, Munsell tried to put Value, Hue, and Chroma on an experimental basis. He used a colour wheel similar to Maxwell's one, and matched Value by looking for non-flickering mixtures. He saw that the shape formed by the colours he could mix was not spherical. In particular, the high-Value yellows, and the low-Value purple-blues went well beyond his sphere. So he abandoned the sphere, and showed his Munsell colour solid for the odd shape it was.

What is 'understanding'?

The Stroop effect shows how entwined the sensation of colour within our heads is with the expression of colour in words. Presumably the process we call 'understanding' is part of the tangle too. I am currently using my 'understanding' and 'words' to describe 'the sensation of colour'. I think I can, so they are not hopelessly tangled. But I have been working with colour and colour measurement for almost forty years, and the way I see colour may have been subtly changed by this.

Perhaps the influence works the other way. I was at a Colour Group meeting where the presenter described migraine aura. He asked how many of us had experienced this. About a third of perhaps forty, including me, put our hands up. The normal rate is about 5% for women, and 2% for men. Is becoming a colour scientist comorbid with migraine aura? We don't know.

The previous chapter, which you may have skipped on my advice, described the crisis in science after 1900, when we had to abandon, or at least modify the search for an absolute understanding of physical phenomena. Quantum mechanics has given us an ability to predict,

sometimes to extraordinary accuracies, but only by summing the possibilities of many, unobserved events. If you want an idea of how well this can work, search for the latest fine-structure constant calculations. The fine-structure constant is a single dimensionless number. Experiment and theory now agree to better than one part in a billion, but the theory takes some epic calculations to get there.

Mathematics and philosophy had a similar crisis in the twentieth century. Alfred North Whitehead and Bertrand Russell produced the *Principia Mathematica*. This first came out in 1910-1913, and a second edition came out in 1925-1927. I had a copy when I was working on printing in the eighties that I read in my lunch hour. I got through the first volume. I certainly remember the notorious footnote from about page 370: "From this proposition it will follow, when arithmetical addition has been defined, that $1 + 1 = 2$." It was a sincere attempt to derive all mathematics from the fewest and most primitive axioms. What the reader does not see is the frantic, endless re-drafting whenever Russell realised his axioms were not primitive enough. For this story, I can recommend 'Logicomix', a graphic novel by Apostolos Doxiadis, Christos Papadimitriou, et al.

We could call this 'Classical Mathematics'. This age may have been brought to a close by Kurt Gödel, and his 'incompleteness theorem', or by the Church-Turing thesis, and the beginnings of the age of computers. But there is another, perhaps more tangible factor here. Whitehead and Russell expressed the axioms of mathematics using some very specific symbols. They defined lexical or typographical rules which generated new 'true' statements from the existing pool of 'true' statements. And from this came all integer mathematics. Or did it? As a reader, I was left with the nagging feeling that we had not explained anything: instead of counting on our fingers, we were using typographical rules to count symbols, but nothing had really changed.

"Nineteen hundred and fifty seven. I don't think I've got enough beans," said the wolf gloomily.

"I don't see how beans come into it," Polly said. "It's years you're counting in, not beans."

"It's beans while I'm actually counting" the wolf said firmly.

Catherine Storr, 'Clever Polly and the Stupid Wolf'

This section is wandering into philosophy, so I am going to stick in some quotes. Sometimes a short quote like "information wants to be free" can be wrong in every particular, and yet convey an idea that would take many, many more words, and still be wrong.

Centuries earlier mathematicians had re-expressed the old, geometrical rules in coordinate geometry and algebra. Whitehead and Russell were

re-expressing familiar mathematics using unfamiliar axioms. Both forms are just as ‘true’: neither are ‘absolutely true’. About that time of the *Principia Mathematica*, some philosophers abandoned trying to understand eternal verities from pure reason, and started looking more closely at the language and intellectual tools we use to converge on an understanding,

And yet the old idea that only an intelligence could say what was ‘true’ or ‘right’ lived on. The four colour theorem proof came out in 1976. Humans had previously shown that you could prove or disprove the four colour postulate (it was not a theorem, until proven) if you tried it on a very large but finite set of maps. So, people wrote a computer program to grind through this set of maps, and check whether four colours was enough. This is a poor proof: an elegant proof would be shorter and might also tell us something about why four colours was enough. However, there were two teams of computer programmers working on the problem, the calculation took months, and the least elegant program had the head-start. But it is still a proof, right?

I would have thought so. The same calculations could be done by enough humans with enough time. But humans get tired and make mistakes, and computers, while not perfect, work to much higher standards. But there were many people at the time who felt that a proof that was not written, not readable, and not wholly understood by an intelligence, was not a proof in their familiar sense.

I do not know if this attitude still exists. I would be happy with the idea that all theorems are in some interim state. If they are true, they should be true for all intelligences, and in all notations. For now, we have human intelligence, aided by computers. One day we may have true computer intelligences, or meet alien intelligences; in which case we can ask them what they think.

“Young man, in mathematics you don't understand things. You just get used to them.”

John von Neumann

“When we talk mathematics, we may be discussing a secondary language built on the primary language of the nervous system.”

John von Neumann

If Bertrand Russell was a late example of ‘classical mathematics’, John von Neumann is a fine example of ‘post-classical mathematics’ if there is such a thing. He was an eminent mathematician and early computer scientist. In ‘The Computer and the Brain’ he clearly saw the coming overlap between computing and intelligence. I find his *“You don't understand things”* applies to much more than mathematics. The ideas of quantum

physics were wholly alien and repellent to the physicists of the early twentieth century. They are still weird, but we can get used to them. Mathematics and physics do not seek only eternal absolutes; they accept some fuzzy boundaries set by our intelligences, and the tools we use.

In chapter one I named particular colours and gave their wavelengths. I remember as a child meeting cyan and magenta as colours for the first time as ball-point pens, not as paints. Having a physical example is good. Having a common name helps. This is how most science starts: you sort things into groups of similar things, and you give them names.

Astronomers started off by naming anything that didn't look like a point or a planetary disc as a 'nebula'. They had elliptical nebulas, spiral nebulas, barred spiral nebulas, planetary nebulas, and irregular nebulas. Much later they worked out that the elliptical, spiral, and barred spiral nebulas were much bigger than the other sort, and re-named them 'galaxies'. Later we worked out what they are and how they work.

Mathematicians use symbols to represent sets of numbers...

$\emptyset$: the empty set

$\mathbb{N}$: the set of 'natural' numbers 1,2,3...

$\mathbb{Z}$: the set of positive and negative integers

$\mathbb{Q}$: the set of rational numbers such as $4/7$

John Conway, the man who gave us 'the game of life', gave us a particularly neat construction for the set of 'Surreal Numbers', starting from the empty set, and defining numbers in terms of pairs of previously defined numbers. This generates the sets of natural numbers, and their mathematical operations, the signed integers, the rational numbers, and even irrational numbers like π . For a summary of Surreal Numbers in Socratic dialogue, I recommend Donald Knuth's 1974 book "Surreal Numbers: How Two Ex-Students Turned on to Pure Mathematics and Found Total Happiness". This does not mean Conway re-defined integers: they are (probably) just the same as they ever were. But mathematicians can use symbols to point to a shared idea without necessarily citing a particular derivation.

"Rules are for the obedience of fools and the guidance of the wise."

Douglas Bader

If I had been asked before I wrote this whether I used 'the proper scientific method', I would probably have answered yes, I did. I would then have wondered what 'the proper scientific method' was, and then whether there was actually any such thing.

You may remember some method from school science lessons: hypothesis, control, experiment, results, and things like that. This is not

always how science works. Astronomers look at astronomical phenomena, and try to explain them in physical terms. They guess at the existence of ‘black holes’, ‘gravity waves’, or ‘dark matter’, and then look for evidence for or against. But they cannot do astronomical-scale experiments. Perhaps they are using the results of ‘natural experiments’. But they are ‘doing proper science’, whatever that may be.

A key feature of science is that experiments should be repeatable. This is not always possible given finite resources. People using the Large Hadron Collider may do billions of experiments, and push their confidence levels out to six sigma, but all their results may be wrong if their single apparatus is inaccurate. The researchers are aware of this. Any complex piece of apparatus is many experiments in one, and capable of checking itself to some extent. However, there is a cautionary tale in Robert Weber’s ‘A Random Walk in Science’. Michaelson spent a lot of his life measuring the speed of light. If you measured the speed of light in the 1930s, you compared your figure to Michaelson’s latest one. If your result was low, you wondered what factors might make it higher. If your result was high, you wondered what factors might make it lower. A lot of supposedly independent researchers converged on the same slightly wrong value for the speed of light. The error was discovered in the 1940s when long baseline radar triangulation tested the speed of light value in a completely different way.

Philosophers and psychologists sometimes make a distinction between *qualia* or things we sense or feel (singular *quale*), and *quanta* or things we can measure. It is sometimes implied, though rarely stated, that qualia and quanta are different things: philosophers and psychologists study qualia and scientists study quanta. This is not necessarily so. For this document, it is probably enough to say qualia are things we feel, and quanta are things we measure; and sometimes we can both feel and measure.

A lot of medical research is all about whether someone feels *better* or *worse* when given a trial drug or a placebo, and *better* and *worse* are qualia. Sometimes there is something we can measure to gauge the level of cure, but in a lot of cases it goes on subjective opinion. Medical trials are some of the most structured examples of ‘the proper scientific method’: not because pharmacists are fools but because of the difficulty in making an objective measurement, and the huge consequences of a wrong decision. For a good introduction to this, I suggest Ben Goldacre’s book ‘Bad Science’.

Colour science is also research on qualia. We are trying to predict the connection between the quanta of what we are looking at and the qualia of the colour we experience. There were a few bad railway accidents in the early days with colour blind people confusing red and green signals, but these days the worst that happens is two colours do not match.

What is proper scientific method? I don't know. But science can take many forms, and it feels wrong to box it in with unbending rules.

"All models are wrong. Some models are useful."

George Box

(a statistician, talking about scientific models; not an exact quote)

What is the state of colour science? I have described some colour spaces. I have tried to use the latest and best colour models, such as preferring the CIE 2006 LMS over the 1931 XYZ. However, in real terms, the difference is not that big, and many people have not yet bothered to switch. I avoided the more complex models; search on CIECAM if you want to know more on them. They are probably better at predicting what we see than the simpler models. However, I prefer a nice, simple model that can be wrong when you go beyond its initial premise, rather than something general that can fit anything if you tweak the parameters enough. No model yet gives a quantified explanation for White's illusion in the introduction. For these illusions, we are still at a very early stage of science: we are collecting them, naming them, and sorting them into groups.

Here is my list for what we have...

- The CIE vision models and the three-colour approximation seem to work well for typical illuminants and white points for display brightnesses between 20 and 1000 nit. The newer LMS models are better than the older XYZ ones, and we should use them.
- Beneath about 15 nit we start to see errors: some flesh colours seem less pink, more grey. Below about 1 nit we start seeing the influence of the rods (Purkinje shift) and the non-linear effects in the colour processing (Bezold-Brücke effect). The eye takes a minute or so to adapt to low light levels. There are useable colour looks that can convey the impression of low light levels on a conventional display.
- Daylight scenes can be much brighter than a 1000 nit display. It is tiring to look at displays brighter than about 1000 nit. The eye takes a minute or so to adapt back to the ordinary light levels. There are good tone mappings that can convey the impression of high light levels on conventional displays. There is no way to get the extra colourfulness as we are limited by the display gamut, but this is not usually that important to the viewer.
- We are not good at seeing or remembering colours outside the gamut of reflective colours, or when the illuminant is far from white. This is not a problem if we are trying to reproduce a scene on a display: we are less sensitive to errors.

- There are many illusions where the colour we see is not what we would expect from the three-colour approximation. We have no way to predict and correct for these effects. It is not clear whether it would be helpful to do so.

Here's one more example. I had been puzzled by a quote from the painter Renoir's notebooks that you could see a bright band in the sky just above hills, and a dark band in the hills just below the sky. Was he aware of the spatial filters within his own eyes and brain? Or was this just him remembering what should be there? He carried on drawing into his nineties, though someone had to tie a pencil into his hand because he could no longer grip firmly. One day, when travelling on holiday, I saw these bands myself. It was in a roadside café. The top of the windows were tinted. I was the right height, so the edge of this tint almost came on the horizon. When he got old, Renoir used to wear glasses with a graded tint. Maybe these are part of the effect. I have tried to reproduce this, but I cannot get it to happen to order. So, I describe it to you, and ask "Have you seen something similar?"

This may not sound like science as it is usually reported. But that is what science is like as it happens, rather than when it has all been done and agreed on, and the story has been cleaned up and put into books.

Can we understand intelligence?

Intelligence is complex. I have described retina signals going into a tangle of nerve cells, and said these cells work together to reach a state called 'understanding'. We know the individual nerve cells behave predictably, but the collection of cells is very complex. Is there any way to untangle this complexity?

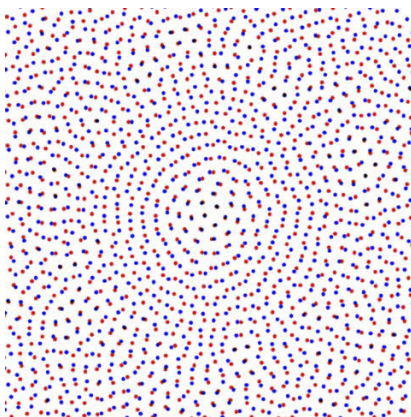
There is some hope. Neural networks, both real and artificial ones, can have common emergent properties. Some animals can see from birth, but humans seem to come with very little programming. I have seen a week-old child raising its hands, and learning that when they do *this* (unidentified muscle movement) then *that* happens (a flesh-coloured shape appears). We all start from a basic state, yet in the end we all see alike. Vision is complex, and may differ for all of us, but the goals are the same. If we guess these goals, we may be able to model what this bit of our neural network is doing, even if we do not know in detail how it is doing it.

Consider the three-colour approximation. We want to recognise objects. We do not want to be distracted by the illumination. Our retina has no accurate measurement of absolute LMS exposures, or even a good zero for black. And yet the raw LMS signals go in, and the network seems to recreate the illuminations, and the reflective colours of objects in the scene. We see luminance ratios as contrasts. We see L:M:S ratios as hues. We may then ignore or downrate data we do not need: we look at

Chequersshadow, and see the chequers, but not the shadow. We can see colours through a window or on a monitor, and we correct for different white points without needing to think. We can see contrasts as luminance ratios. The three-colour approximation is a useful and reliable tool, which suggests some processes I described may be real.

Supposing this were not so. Our vision actually builds its understanding from darting the eyes about the subject and integrating the information it gets. I would have to correct the image for every possible path the eye could take across it. I would have to account for contractions of the iris and adaption of the retina. I would have to model the scattering in the eye, the properties of the retina, and the neural processing. Instead, I forget those complexities, and treat the eye as a digital camera: not because it is, but because it has the same goals. This is not perfect, but it will do.

Vision has many other tricks. For instance, we are good at recognising textures. This is good for identifying complex objects. We do not encode the precise textures, but we can remember the shape, the direction and other parameters of textures. We can spot motion of textures, such as hair or trees waving in the wind. People landing an aircraft unconsciously judge their height by the scale of the tyre marks on the runway. And so on.



What is this texture? I took an image of blue-noise random dots, and superimposed a second copy rotated by 5° . We spot the centre of rotation immediately, and see pairs of dots out to about half-way. Beyond this the rotation is larger than the dot spacing, and the structure is harder to spot.

We might spot the two-dot texture. We might see the texture goes in a circle, and there is a quiet region in the middle. But how, and why do we do it so quickly? This is not something we need in ordinary life. I guess it is somehow connected to how we see flow and motion, but I don't know.

Computed neural nets are very different to the ones in our heads. The nodes are in neat layers. They are controlled by a clock. They are trained, or they predict. We can train some layers and freeze others. These differences are there mostly, so we can do experiments, and understand the results. But they can recognise features in an image as we do. They can also do some surprising new tricks. For instance, we can train a neural net on physical simulations of liquids moving with turbulence. We can then get it to predict how a liquid might flow in another situation not in its training set. The results are probably not physically accurate: even if the network could rediscover the laws of turbulent fluid motion, it is coming up with the answers too fast for any known simulation method. But the simulations look splashy and liquid-like. Even if we do not know quite what it is doing or how it is doing it, knowing that it *can* do it is a step forward.

Can we understand intelligence? Maybe. I expect advances and CAT scanning, and neural net simulation will reveal larger chunks of our mental process. Intelligence is not the mystery it once was.

What is seeing?

We probably have a ruler in our desk drawer, a weighing machine in the bathroom, and a set of scales in the kitchen. Our bathroom cabinet may have a thermometer, and the central heating will probably have a thermostat. Our watch or mobile phone can tell us the time. A car may have gauges for speed, battery charge, oil pressure, water temperature, and engine speed. An electrical meter can measure amps and volts (AC and DC) or resistance. We are surrounded by coloured objects, and we need to match colours when decorating, yet there is no instrument for measuring colour. Why is this?

There is such a device. Many of us have one. A digital camera or mobile phone can do this. It measures the light levels in three colour channels at millions of points at once. A good digital camera can show us an image which looks good when we can still see the original. But it seems it cannot help us pick curtains in a shop that match the wallpaper we have at home.

‘I see’ also means ‘I understand’ in English. This is also true in other languages. The acts of seeing and understanding are entwined, and common language does not separate them. The raw RGB numbers in our camera may not mean much to us, but the camera presents them to us in the form of an image. We see the context for the colour, and we see the frame around the image that limits the local white balance. Our colour constancy kicks in, and we automatically correct the view we see.

Suppose we took a picture of our wallpaper at home. The camera corrects for the lighting, so we may want to include some reference white patch. We then go to the shop and take pictures of the curtains with the

same white patch. If we have matched the whites, and the curtains match the wallpaper, then we buy. And when we get them home we find they don't match. It may be the lighting spectrum is different. It may be there are small differences between our vision and the camera. We may be looking at a larger patch of colour with our peripheral vision. Probably all three. If we had taken in a sample of the wallpaper into the shop, we might still not get a match we liked.

A frustration with colour science is how good our eyes and brain are at correcting the signals from our eyes, and how hard it is to do the same with science. Digital cameras are not poorly made: they must be amazingly consistent to get matching results from the millions of pixels. Our eyes are not nearly that good: different points within the same eye have significantly different spectral sensitivities. But the brain has had a lifetime's training on which differences matter, and which can be ignored. So, you spend all morning calibrating a monitor with instruments, then your eyes say "whoah, that's completely wrong" straight away.

Another frustration is this adaption and correction is instinctive. Brunelleschi painted colour pictures. He could have used his sliding mirror to show that his colours were correct but no-one felt the need for that. Perspective needs to be thought about, but we just know when colours are wrong. Or, maybe, the people looking at Brunelleschi's sliding mirror saw a flat drawing differently to the way we do now. Perhaps the Müller-Lyer illusion would not work for them back then. We don't know. We have no access to our instinctive processing, so we have to use all sorts of ingeniously indirect methods to find out what our heads are doing, while our heads fight back and refuse to co-operate.

Colour science is a strange job. You are the experimenter, the apparatus, and the lab rat. And with those wise words, I shall leave you. Good luck!

Appendix A

Matrix maths

Here is a matrix conversion from XYZ and LMS, as it appears in Chapter 3.

$$\begin{pmatrix} L \\ M \\ S \end{pmatrix} = \begin{pmatrix} +0.21058 & +0.85510 & -0.03970 \\ -0.41708 & +1.17726 & +0.07863 \\ 0.0 & 0.0 & +0.51684 \end{pmatrix} \cdot \begin{pmatrix} X \\ Y \\ Z \end{pmatrix}$$

This is the same as the three separate equations...

$$L = +0.21058 \times X + 0.85510 \times Y - 0.03970 \times Z$$

$$M = -0.41708 \times X + 1.17726 \times Y + 0.07863 \times Z$$

$$S = +0.0 \times X + 0.0 \times Y + 0.51684 \times Z$$

This is how I write matrix equations through this document. This is not the only way matrices can be written. You may find the same matrix written with the rows and columns swapped. To avoid any confusion, here is a brief introduction to matrix maths.

Suffix notation

If we call our matrix M, we might write the equation as...

$$LMS = M \cdot XYZ$$

If XYZ and LMS are vectors and M is a matrix, the multiplication and summing operations are implied. However, we can make the summing a bit more explicit with suffix notation...

$$LMS_i = M_{ij} \cdot XYZ_j$$

LMS and XYZ have only one suffix, so they are vectors. M has two suffixes so it is a matrix. Any suffix that only appears on one side of the equation is summed over.

Suppose our XYZ has been derived from RGB with another matrix N...

$$XYZ_j = N_{jk} \cdot RGB_k$$

We can write our transform from RGB to LMS as...

$$LMS_i = M_{ij} \cdot (N_{jk} \cdot RGB_k)$$

We can ignore the brackets as it does not matter what order we do the multiplications in. Or we can group the two matrices together into a single matrix...

$$LMS_i = (M_{ij} \cdot N_{jk}) \cdot RGB_k$$

$$LMS_i = MN_{ik} \cdot RGB_k$$

The 'k' suffix only appears on the right-hand side of the equation, so we can replace it with any other letter that is not already in use.

Transpose matrix

Suppose we had swapped the matrix suffixes. Suppose we had written...

$$T_{ji} = M_{ij}$$

$$LMS_i = T_{ji} \cdot XYZ_j$$

T is the transpose: it is matrix M with the rows and columns swapped. We do not sum over either suffix in the first equation because they both appear on both sides. We sum over 'j' in the second equation.

Now suppose we want to write the matrix out in the original style. Let us re-arrange the first equation so the repeated suffix 'j' is the one being summed over...

$$LMS_i = XYZ_j \cdot T_{ji}$$

In the original notation, the row and column value from A·B was made by multiplying the corresponding row of A and the column of B, and summing. Our equation now should look like this...

$$(LMS) = (XYZ) \cdot \begin{pmatrix} +0.21058 & -0.41708 & 0.0 \\ +0.85510 & +1.17726 & 0.0 \\ -0.03970 & +0.07863 & +0.51684 \end{pmatrix}$$

LMS and XYZ are the same three numbers, but written as rows not columns. This is equally correct in this notation. The original matrix-vector arrangement looks neater, so this is how colour transforms are often written. This is how I write them. But you may find ones written the other way.

Identity matrix

The identity is the equivalent of the number one: multiplying by the identity matrix has no effect. It is usually called 'I'. For 3 dimensions, we have...

$$I = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Inverse matrix

This is the matrix equivalent of division. The inverse of matrix M is often written as M^{-1} .

$$I = M \cdot M^{-1}$$

We can express this equation as a set of simultaneous equations. If we have matrix M, we can solve for the elements of M^{-1} . Some matrices cannot be inverted. This happens if the simultaneous equations are not all independent. For example, we can convert colour RGB to monochrome using a matrix, but there is no inverse transform.

Appendix B

Other tRGB space tools

The tRGB space in Chapter 4 is not recognised by any standards body. Any tools mentioned here may have been used by myself alone. However, until there are recognised standard tools based on CIE 2006 LMS instead of 1931 XYZ, these ones may have to do.

CRI and tCRI

The CIE colour rendering index (CRI) is a figure that is used to give a high value for an illuminant that match a reference illuminant. The reference is a blackbody spectrum for colour temperatures up to 5000K, and for sunlight as represented by the CIE standard illuminants beyond that. The CIE 1995 standard CRI is calculated as follows...

- 1** Find the reference illuminant, which is the nearest blackbody or daylight simulation to the illuminant we mean to measure. This is done using the old Yuv space.
- 2** Find the von Kries (LMS) scaling to map the test illuminant to the reference illuminant.
- 3** Calculate the reflectance spectra for a set of 14 reflection colours for both illuminants. The reflection spectra were originally measured from 14 chips from a Munsell colour atlas. They include eight pastel colours, strong RYGB colours, a pink-yellow, and an olive green.
- 4** Calculate the XYZ colour values for both illuminants. Shift the test illuminant reflectance white XYZ values, so the white point matches the reference white XYZ using the von Kries transform.
- 5** Calculate the ΔE values between the test white XYZ and the reference white XYZ. This is done in the now rarely-used CIE UVW space.
- 6** The CRI is 100 minus the average ΔE .

There are 14 standard reflectance spectra. I have probably not described the CIE calculation in enough detail to perform the calculation. There have been several modifications since 1995. There are a number of criticisms of the CRI, the most serious perhaps being that it does not actually correlate with the light sources that give good colour matches. The reflection spectra are flat between 600 and 700 nm, so the standard

is not sensitive to different red wavelengths. Three-band LED lights that have no deep reds can have a high CRI. There is also a discontinuous jump between the blackbody fit below 5000K and the daylight simulation above it. I have not found out what you are supposed to do at 5000K, though it may be written down somewhere.

I have tried to come up with some equivalent CRI calculation based on CIE 2006 LMS. Because it is based on the visual triangle tRGB spaces described in Chapter 4, I have called it tCRI.

We calculate tCRI as follows...

- 1** Calculate the Yrg 'r' value, and use it to find the closest blackbody value. See the graphs in the Lighting and Mireds section at the end of Chapter 9 for how to do this. This does not necessarily return the closest point on the blackbody, but the blackbody curves are all smooth so the result after the von Kries correction will be very similar. There is no sensible way of finding the 'closest' white when the error is in the green-pink direction anyhow.
- 2** Calculate the reference blackbody spectrum, scaled to have the same Y as the original spectrum. Record the temperature, as anyone comparing the light source with others will want to know this. I have chosen to only use the blackbody spectra because these are smooth and have a simple formula, and some real light sources are close to them. We could do a similar calculation for daylight simulations if we needed it.
- 3** Calculate the worst reflection metamer. This has a reflectance of 1.0 for any wavelengths where the test illuminant is brighter than the reference, and 0.0 otherwise.
- 4** Blur the reflection metamer. I used a 25 nm half-height, half-width bell-shaped blur. This makes it roughly as sharp as the CIE reflection spectra, and similar to the real spectra we may have to match.
- 5** Calculate the LMS values for both illuminants, and both illuminants with the worst metamer.
- 6** Scale the test illuminant values so the white LMS matches the reference LMS. This is a von Kries correction as we are already in LMS.
- 7** Calculate the tLab ΔE_m for the metamer values and ΔE_{m-1} for the complementary metamer values. We can get the complementary LMS by taking the metamer LMS from the white LMS.
- 8** $CRI = 100 - \sqrt{(\Delta E_m^2 + \Delta E_{m-1}^2)}$.

The tCRI values have a similar range to the CIE CRI values...

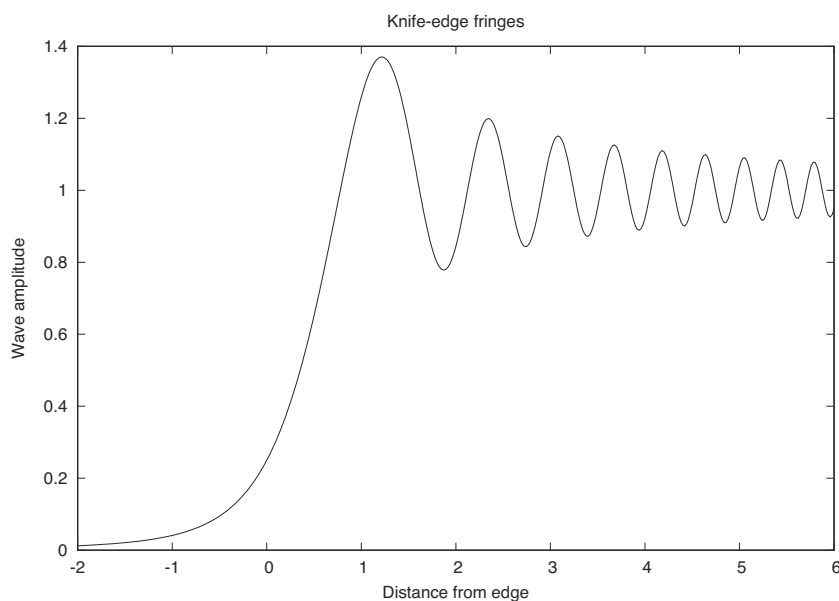
tCRI	CIE CRI	Blackbody	Illuminant
100	100	Any	Ideal blackbody
99.7	99.1	2309 K	Real incandescent
99.1	96.8	4230 K	NKT supercontinuum white laser
98.8	98.7	4983 K	Cermax Xenon 300W
98.6	95.1	6481 K	CIE FL15 daylight
98.4	98.8	2958 K	Halogen truck headlights
97.8	96.7	6253 K	Osram Biolux Fluorescent 965
96.3	80.4	2833 K	Luxeon White Warm LED
96.3	91.8	6538 K	Kino-Flo daylight
93.8	100	6540 K	Daylight, CIE D65
93.2	94.4	3245 K	ARRILITE 300
91.9	96.5	6228 K	Xenon arc
90.0	98.0	5905 K	Hive Wasp 1000 Sulphur plasma
75 – 90			Good fluorescent and LED lighting
60 – 75			Bulk fluorescent and LED
45.9	69.1	2897 K	Sodium high pressure
43.5	44.9	2897 K	Coated Mercury vapour
Negative			Sodium low pressure, monochrome LED

Daylight has a tCRI of 93.4. If we were trying to match daylight, then this would be the best value. In practice, a D65 light is usually only a colour match to D65, but we can get real light sources that come close to blackbody spectra.

The Hive Wasp 1000 is a sulphur plasma lamp. This has a published CRI of 98.0. It has less blue-violet than the blackbody so our tCRI is only 90.0. However, if we fit a CIE daylight spectrum instead of a blackbody, we get a tCRI of 96.2.

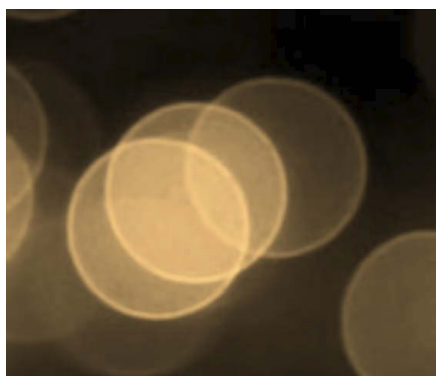
Colour interference from a few wavelengths

In Chapter 2 we described how Thomas Young measured the wavelength of light. In Chapter 11 we described the knife-edge diffraction observed by Grimaldi. These are both examples of light interference.



Here is a plot of the waves that Grimaldi saw at the knife edge. Zero on the horizontal axis is where the knife-edge is. We do not often take out of focus images of knife-edges, but there is an effect called 'bokeh' which is the shapes you get when distant lights are out of focus...

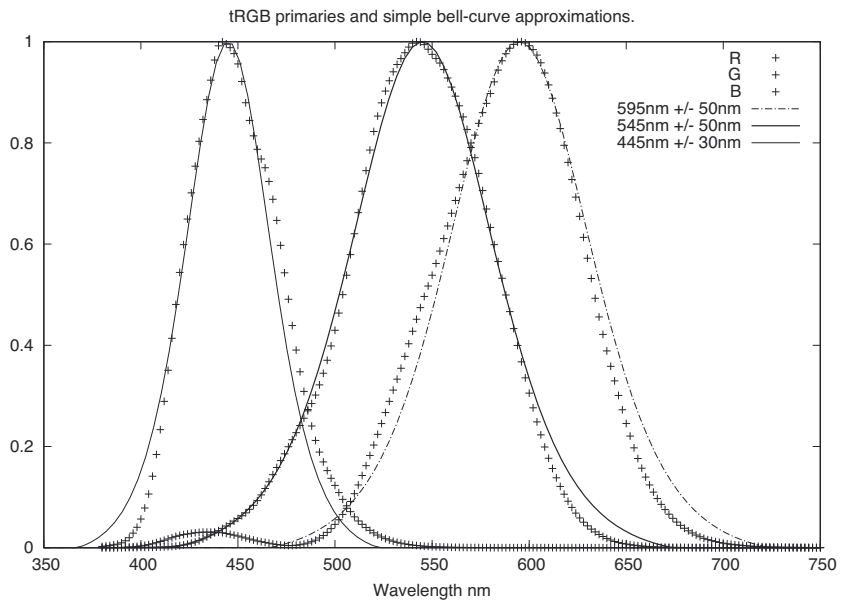
Here is a sample from a photograph. The edge of the bokeh ring has the same pattern of fringes. We can see the bright outer edge of the bokeh, and just about see the dark ring beyond it, but the second ring is hard to see. What has happened to all the rings in the graph plot? The image is not sharp and the fringes are approaching the size of the camera pixels, but we ought to see something.



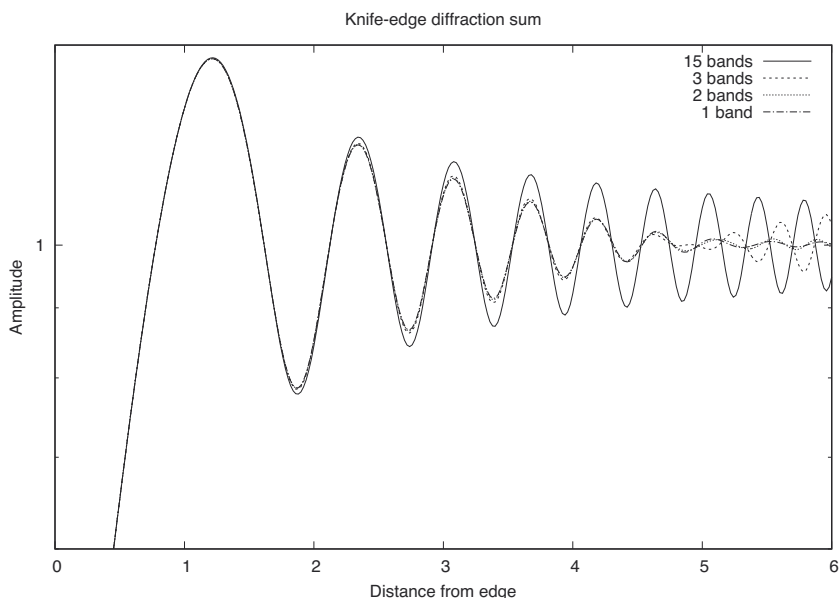
Bokeh

The graph plotted the interference pattern for a single wavelength. The photo is summing many wavelengths. The edge and the first fringe or two are in much the same place for each wavelength, but the higher order fringes will average out.

Supposing we wanted to simulate circular bokeh. We can calculate the diffraction pattern for each bokeh ring for each wavelength. We cannot just resize for each wavelength because the zero lies just outside the bright fringe. In the image, the bokeh circles are all roughly the same size, so we could rubber-stamp the fringe pattern. But in a more complicated case we could not.



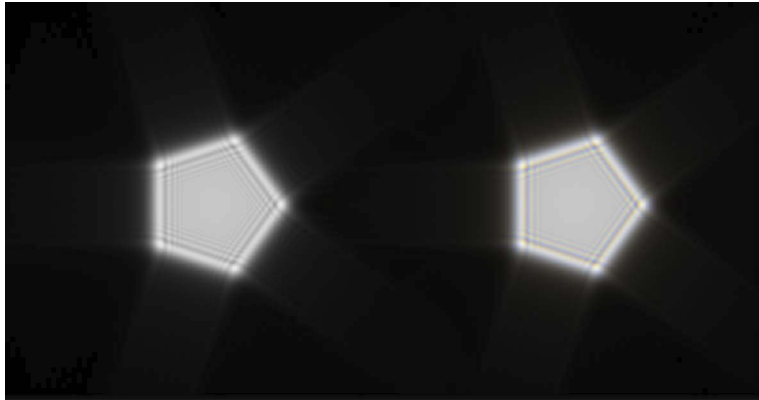
The graph shows the tRGB primaries, and a set of simple Bell-curves I fitted to each of these. This is not an attempt to 'clean up' what the eye does, but I need to simulate diffraction without the distraction that coloured fringes could come from lumps on the spectral curves.



This graph shows just the fringe detail. The '1 band' curve shows the interference pattern for the single peak wavelength (545 nm in the case of the tRGB green). The '15 band' curve is the weighted sum of 15 different wavelengths on the green peak. We can change the number of wavelengths and the results are almost the same: as the distance from the edge increases, the peaks on one curve cancel out the troughs in another, and the ripples disappear. For the case of the tRGB peaks, where the peak half width is about 8% of the peak wavelength, we can get almost the same result with three bands. At 2 bands, the approximation breaks down: we can see the curve goes through a minimum as the two sets of fringes get out of step, but they get back in step again.

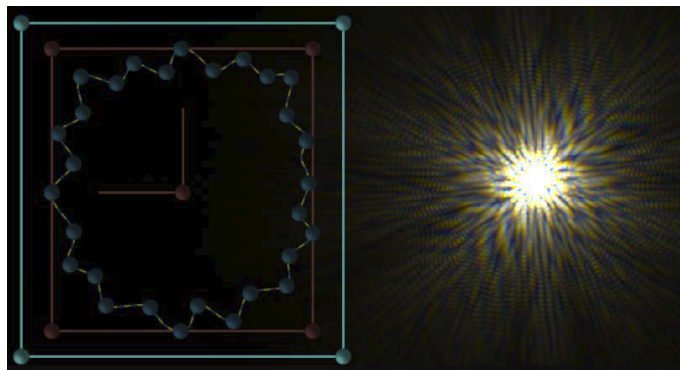
This only works for broadband illuminants. We can calculate the diffraction pattern for 495 nm, 545 nm, and 595 nm, and average them 1:2:1. We get a good approximation to the tRGB green if the spectrum was smooth. If our illuminant was a 545 nm laser then the fringes should all come back.

Here is a simulation of bokeh from a 5-sided lens aperture. On the left we have a 545 nm calculation. I can see five, perhaps six bright fringes. On the right we have a simulation in tRGB using three sets of 3 wavelengths. This gives us realistic looking colours, but it is hard to see anything beyond the third fringe. This is realistic, and fits simulations using many more wavebands.



5-sided aperture bokeh simulation

Here is a simulation of the flare we might get from the iris in our eyes. The iris shape is a hand-drawn shape shown on the left. The flare pattern has rotational symmetry because I have assumed the iris is in the focal plane, which it probably isn't. But the results look similar to the flare I can see on a sunny day looking at the parked cars outside.



Eye iris bokeh simulation

Index

10-degree observer	53	bolometer	29, 58
1D LUT	119	Boltzmann	239
2-degree observer	48	bottom-up	4, 10
2-degree standard observer	8	Boys	30
3D LUT	120	Bradford University model	134
Abney effect	45	BRDF	220
acrylics	136	Brewster's Angle	209
additivity failure	112, 215	bright colours	65
Adelson	10	bright displays	152
after-images	101	brightness	51, 61, 64, 152
airglow	213	Brunelleschi	250
Airy disc	180	Brussels sprout	13
Albers	175	BSSRDF	220
American Night	252	by-eye calibration	147
Ames Room	168	Caillebotte	176
Ansel Adams zones	96	calcite	173
anti-halation	199	calibration	135, 147
Aristotle	37	Callier effect	196
azure	21	camera calibration	135
Barco DP90P	143	canals	187
Bartleson and Breneman	98, 150	candela (cd)	22, 56
Beer-Lambert law	215	candelas per square metre	57
Bell's inequality	235	capturing LUTs	124
Benham's top	166	Casablanca	164
Berinmo	248	Checkershadow Illusion	10
Berlin and Kay	246	chroma	65
beta-carotene	223	chromatic aberration	31
Bezold-Brücke effect	44, 141	chromatic intensity	65
bilirubin	223	Church-Turing	261
birefringence	210	CIE 1931 XYZ	51
black and white film	189	CIE 1964 10-degree observer	54
blackbody spectrum	68	CIE 2006	50
blind spot	44, 158	CIE 2006 LMS	48, 132
blindsight	163	CIE L*a*b*	77
bokeh	278	CIE standard daylight	69
		CIE UVW	116
		CIECAM02	135

CIELAB	77	electron-Volts (eV)	22
cilantro	13	entanglement	174, 237
Cineon	193	eosin	209
Cinerama	201	epidermis	222
Clapper-Yule model	220	epilepsy	51
CMYK	111, 162, 218	evolution	172
colorimeter	58	exposure	185
Colour Appearance Model	79	exposure units	184
colour appearance models	134		
colour balance	183	Fairchild model	134
colour constancy	15, 44, 172	Fechner colour effect	166
colour gamut	110, 141	field-ion microscope	145
colour matching	52	filling in	159
Colour Rendering Index (CRI)	115	film tone curve	188
colour sphere	259	flare	178
colour temperature	202	flicker fusion	186
colour vision	14	flicker photometry	51
colour vision test chart	91	fluorescein	209
colourfulness	144	fluorescence	207
complementary wavelength	92	foot-lambert	57
compression	107, 109	four colour theorem	262
cone cell	14, 48	fovea	161
consciousness	2	Fox Talbot	177
Constable	104	Fraunhofer	25
contrast range	151	frequency	22
Copenhagen interpretation	238	Fresnel	231
CRI	275	fundus microscope	48
crispening effect	97, 167		
cryptochromes	173	gamut	99
		gamut alarm	99
D65	69	gamut compression	110
De Broglie wavelength	232	Gegenfurtner	7
Delta E (ΔE)	90	Gegenfurtner and Sharpe	48
demon	3	gene therapy	172
densitometer	58, 190	Gödel	261
density	95	Goethe	257
dermis	222	goniochromatism	210
Descartes	2	grating	32
display calibration	147	grey world	18, 105
dithering	145	Grimaldi	231
Dollond	31		
dominant wavelength	92	Haidinger's brushes	40, 172
dorsal stream	163	haze	176
dress image	11	HDR	101, 137, 149, 152
		HDR from video	109
electron microscope	232	hearing	166

- Helmholtz-Kohlrausch effect 65, 140
- hemoglobin 222
- Hering opponents 165, 169, 259
- Herschel 27
- high gain screen 201
- high gamut 137, 149
- highlight bleach 108
- highlight compression 107
- highlights 101
- Himba 249
- Homer 39
- homunculus 2
- hot spot 201
- hue 62
- hue angle 64
- Hunt effect 144
- Hunt Model 134
- Hunter Lab 79
- Hunt-Pointer-Estevéz (HPE) 134
- Huygens 254
- hyperacuity 156
- hyperspectral camera 59, 203
- identity matrix 273
- illuminant 129
- illusion 158, 167
- IMAX 150
- incandescent spectrum 68
- infrared 28
- intrinsically photosensitive 45
- inverse matrix 273
- iridescence 210
- irradiance meter 58
- ISO standard filters 191
- Just Noticeable Differences (JND) 88
- K-cell 163
- keyline 112, 218
- Kim vision model 134
- knife-edge diffraction 231, 278
- Kodachrome 190
- koniocellular 163
- Kubelka-Munk model 218
- Lagrange multipliers 127
- Lamb 48
- Lambertian 200
- Lambertian reflector 57
- lamp calibration 199
- Langley 29
- latensification 188
- Lateral Geniculate Nucleus (LGN) 159, 162
- Le Blon 256
- LGN 159, 162
- light meter 58, 184
- lightness 64
- lilac chaser 166
- list transforms 124
- LMS 43, 51
- Locpro 200
- Lowell 187
- lumens (lm) 57
- luminance 51
- luminous colours 140
- lux (lx) 57
- lux meter 58
- MacAdam ellipses 88
- Macbeth chart 135
- macula 161
- macular dye 49
- magnetic fields 173
- magnocellular 163
- Magritte 106
- Mars 187
- matrix conversion 53, 271
- matrix maths 271
- Max Factor 183
- Maxwell 182
- Maxwell spot 55
- Maxwell's equations 230
- Maxwell's triangle 81
- Maxwell-Boltzmann 240
- McCollough effect 168
- M-cell 163
- melanin 222
- metameric 221
- metameric failure 114
- metamerism 114
- mezopic 41
- Michael-Levy chart 211
- Michelson 264

microdensitometer	198	Pfund skylight compass	173
Mie scattering	212	Philips probe	54
migraine aura	260	Phong shading	217
mireds	202	photocopier	136
modelling	207	photoelectric effect	243
moiré	162	photography	181
Monet	18, 104	photons	33
monkeys	16	photopic	40, 153
Monte-Carlo	207	pineal gland	2
mosaic	37	Pinglap	41
Müller-Lyer illusion	168, 250	pitch	26
multispectral camera	59	Planck	239
Munsell	65, 259	Planckian spectrum	68
Munsell benchmark	71	planetarium	151
Munsell Colour Atlas	35	plasma panel	146
Munsell Value	73	platinum	29
musk	13	plenoptic camera	203
		Pointer gamut	80, 129, 137
nanometres (nm)	21	polarization	40, 172, 209
Nassau	35, 212	Ponzo illusion	168
Natural Colour System (NCS)	258	primary visual cortex	159
negative toe	195	primary wavelengths	142
Neugebauer model	218	Principia Mathematica	261
Newton	24, 210, 254	print industry	95
Newton's rings	25	print shoulder	195
nit (nt)	57	printer points	195
nitrate	199	Punch	181
Numskulls	1	pupil	95
Nyatani model	134	Purkinje shift	41, 153, 252
		Purple of Cassius	213
OLED	102, 144		
optic brighteners	208	qualia	264
optical coherence tomography	14	quanta	264
optical flat	26	quantization	117
Opticks	24, 254	quantum cryptography	174
		quantum physics	227
panchromatic	183		
pandemonium	3	radian (rad)	56
parvocellular	163	Raphael	175
P-cell	163	Rayleigh scattering	212
peacock	211	Rayleigh-Jeans	241
penny	5	Rec.2020	100, 138
pentane lamp	56	Rec.709	132, 138
peripheral vision	6, 54, 160	Rec.709 gamut	80
persistence of vision	185	reference white	195
perspective	251	reflection colours	140

- reflection spectrum 34
- remjet 199
- Renoir 266
- retardation 211
- retinal ganglia 42
- Retinex 105, 134
- reverse process 189
- rhodopsin 41
- ripple tank 227
- Ritter 29
- rod cells 95
- rosette 162
- Rowney 136
- Rumford 29
- Runge 259

- saccade 164
- Sacks 41
- Sapir-Whorf 246
- saturation 62, 65
- scanner calibration 135
- Scarface 204
- scattered light 151
- scattering 211
- scent illusions 14
- scotopic 40
- screen gain 201
- selenium 184
- shadow textures 98
- shape effect 7
- shaper LUT 122
- shininess 217
- shutter angle 184
- silver halide crystals 181
- skin 221
- sky colour 211
- slide film 95
- smoke 176
- smoothest spectrum 126
- SMPTE colour bars 38
- SMPTE RP-180 192
- sodium 129
- Sony BVM CRT 143
- Spanish Castle 166
- spatial frequency 153
- spectral locus 80
- spectrometer 33, 59
- spectroscope 25, 31, 58
- spectrum 21, 126
- spectrumCalc 126
- sRGB 74
- Stan Laurel 182
- status A 191
- status M 192
- stele of Lady Taperet 243
- steradian (sr) 56
- stereo 177
- stereo camera 203
- Stiles-Crawford effect 53
- stimulus 15
- Stockman and Sharpe LMS 49
- stop 184
- Stroop effect 245
- subsurface scattering 177, 220
- suffix notation 271
- sun dogs 213
- sun stones 173
- surround 6, 9
- synaesthesia 164

- tapetum lucidum 151
- taste 13
- taste illusions 14
- tCRI 275
- telespectroradiometer 59
- The Empire of Light 106
- The Hay Wain 104
- The Pool of London at Sunrise 104
- The Sun at the Zenith 103
- thermal spectrum 68
- Thompson's galvanometer 29
- three-colour approximation 39, 162
- tLab 86
- TMS 2
- tone compression 109
- tonic mode 42
- Tonomura 235
- tourmaline 173
- Transcranial Magnetic Stimulation 2
- transmission spectra 36
- transmission spectrum 36
- transpose matrix 272
- tRGB 79, 85, 275, 279
- triangle 130, 150

Triangle transform	123	yellow spot	161
troland (td)	57	Young	23, 25, 40, 182, 227
Troxler effect	54, 160	Yrg	81
Truelight	98, 124, 181	Yu'v'	67
Truelight cube format	122	Yule-Neilsen model	219
Truelight probe	147	Yxy	66
Truelight Projector Probe	147		
Truelight spectrum format	126	zebra stripes	102
tungsten light	84	Zview	126
two-slit experiment	229		
Tyndall effect	212		
ultraviolet	29		
ultraviolet catastrophe	241		
Under Colour Removal (UCR)	111		
vacuum	22		
value	64		
ventral stream	163		
ventriloquists	246		
Vernier acuity	156		
viewing surround	97		
violet	25		
vision pathways	162		
von Kries	77, 133		
von Kries matrix	134		
wagon wheels	185		
watercolours	136		
waterfall	166		
wavelength	23		
wet gate	198		
Wheatstone	177		
Whetstone's bridge	29		
white	70		
White's illusion	6		
Whorfianism	246		
window	17		
Windsor and Newton	136		
Woman with a Unicorn	175		
xenon	55		
X-Rite TR-310	190		
XYZ to LMS	131		
yellow dye	44		

About the Author



Richard Kirk developed a colour calibration system for proofing colour pages on film recorders in 1985. Since then he has worked continuously on colour and imaging technology: with Crosfield Electronics for the printing industry; with Canon for colour photocopiers and desktop devices for Canon; with Framestore-CFC for film post-production; and finally for FilmLight, where he developed the Truelight colour calibration system, the standard for film previewing and visualization. He holds over 30 patents on colour and imaging techniques, and an AMPAS Scientific and Technical Award (Oscar) for the Truelight system.

