Human Image Perception

4c8 Media Signal Processing

Ussher Assistant Professor François Pitié

2016/2017

Electronic & Electrical Engineering Dept.,
Trinity College Dublin
pitief@tcd.ie

Overview

We have seen the need for compression.

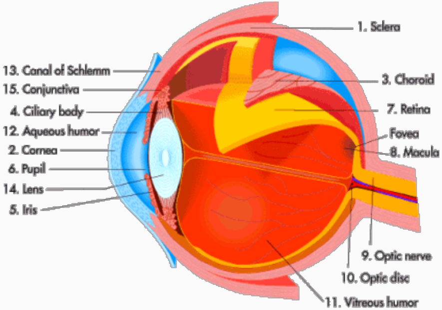
We can use what we have learned in information theory to exploit spatial and temporal redundancy but it is not enough

We must determine ways in which we can exploit redundancy in the way we perceive images.

To do so it is important to understand some relevant aspects of the Human Visual System (HVS).

Colour Spaces

Vision : The Human Visual System (HVS)

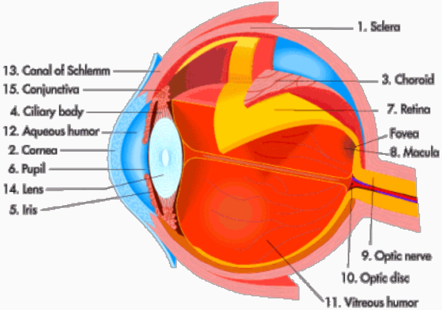


The retina has two types of cell:

The cones: sensitive to colour and luminance, located near the centre of the retina (fovea)

The rods: located near the periphery of the retina, more sensitive to light and motion, luminance only, less resolution

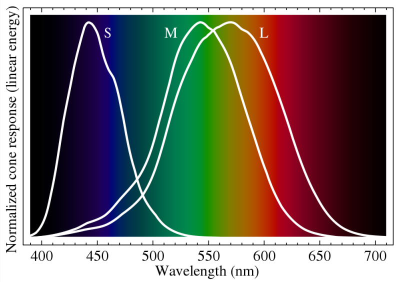
Vision : The Human Visual System (HVS)

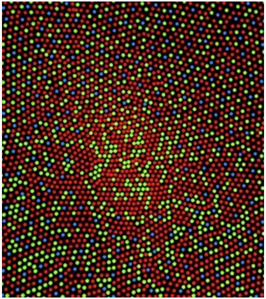


Electrical Impulses from the retina are channelled by the optic nerve to the Visual Cortex.

The Visual Cortex does a whole bunch of smart things, including filtering, object recognition, edge detection and stereo fusion.

Cone Cells



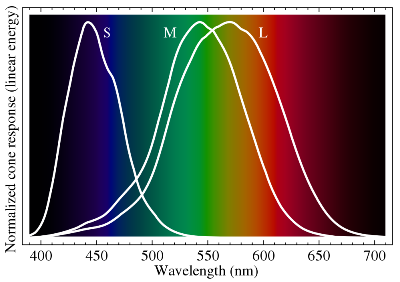


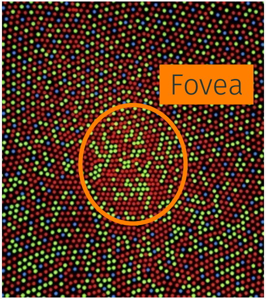
Attribution: Mark Fairchild

The S (short), M (middle) and L (long) cone cells convert wavelengths into 3 signals known as a tri-stimulus.

The cone responses are quite spread out, with $S \approx$ blue, $M \approx$ green and $L \approx$ yellow-red.

Cone Cells





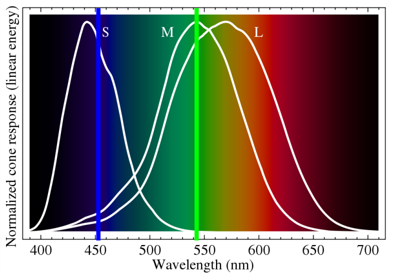
Attribution: Mark Fairchild

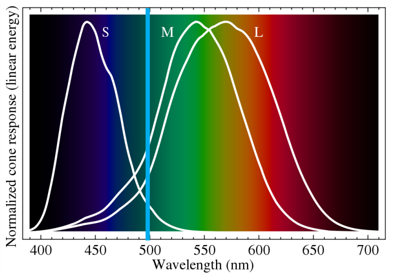
The S (short), M (middle) and L (long) cone cells convert wavelengths into 3 signals known as a tri-stimulus.

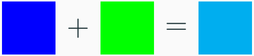
The cone responses are quite spread out, with $S \approx$ blue, $M \approx$ green and $L \approx$ yellow-red.

The Fovea has very little S Cones.

Metamerism



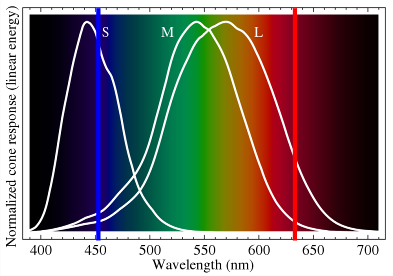




Light is a spectrum and different combinations of wavelengths can result in the same signals (S, M, L) and appear to have the same colour.

This is called metamerism.

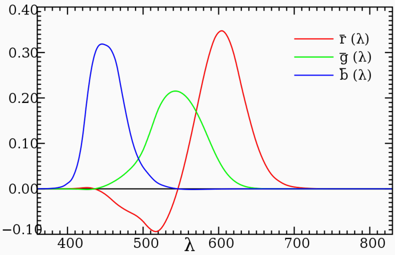
Purple



Purple/pink is an interpretation of our brain when we have S (blue) and L (red) but no M.

How come we see purple in rainbows? The L cones have a long tail response and $L > M$ at short wavelengths, hence the perceived purple colour.

Colour Matching Functions

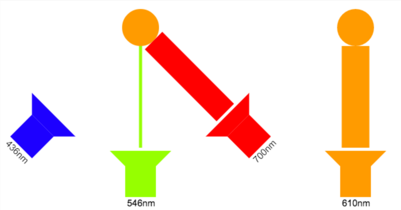


CIE RGB red = 700 nm, green = 546.1 nm, blue = 435.8 nm

How do we perceive a mono-chromatic light source as a function of 3 primary colours?

These functions obtained from perceptual studies in the 1920's are known as colour matching functions and can be used to estimate RGB values for any combination of colours.

Colour Matching Functions



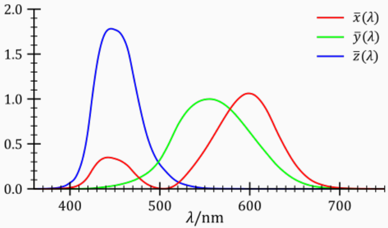
see Chandler Abraham [`https://goo.gl/vn8Wuv`]

The matching functions were obtained by allowing participants to combine the 3 r,g,b stimuli to match the appearance of colour from a single wavelength.

Negative values arose because no combination of r,g,b provided a good match. Participants had to add some amount of red stimulus to the target wavelength.

XYZ Colour Space: The Standard Reference

People working on this were bothered with the negative values, hence they derived a new colour space: XYZ. XZY is simply a linear transformation of RGB. XYZ now serves as a standard reference for building other colour spaces.



Y = roughly similar to response of Cones M (green stimulation)

Z = Cones S (blue stimulation)

X is a mix of cone responses.

XYZ Colour Space: The Standard Reference

For reference, the actual transformation between XYZ and RGB is:

$$\begin{pmatrix} Y \\ Y \\ Z \end{pmatrix} = \begin{pmatrix} 0.49 & 0.31 & 0.20 \\ 0.17697 & 0.81240 & 0.01063 \\ 0.00 & 0.01 & 0.99 \end{pmatrix} = \begin{pmatrix} R \\ G \\ B \end{pmatrix}$$

This is mathematical construction. Note all values of (X, Y, Z) are physically realisable (eg. some may correspond to negative values of R, G, or B).

The Chromaticity Diagram

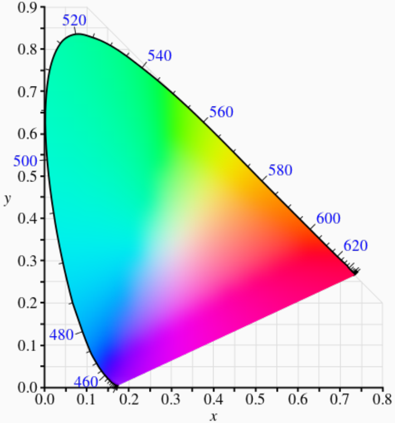
XYZ is used to graph the chromaticity diagram, which is used for comparing the gamut, or range of values, that different colour spaces can represent.

Normalising for luminance:

$$
\begin{aligned}
x &= X/(X+Y+Z) \\
y &= Y/(X+Y+Z)\,,
\end{aligned}
$$

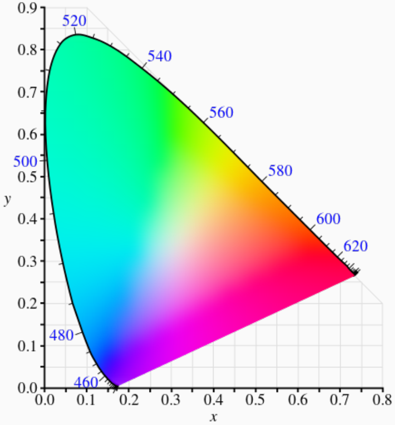
the diagram show all the colours visible by an average human on the x-y plane.

This is the human gamut.



The Chromaticity Diagram

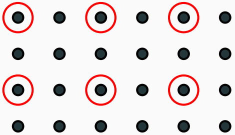
All monochromatic lights (ie. a pure hue of a single wavelength), lie on the spectral locus.



Chroma Subsampling

In the next slide, we will subsample the U and V chrominance channels and leave the Y channel alone.

For instance, a 2:1 subsampling will look as follows:



(*i.e.* we are only keeping the circled pixel sites)


original

Colour Perception

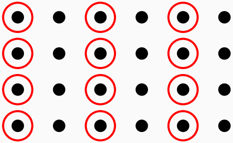
Colour Perception

Colour Perception

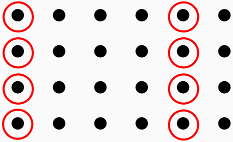
Colour Perception

Chroma Subsampling

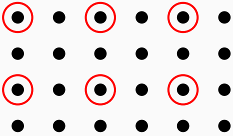
You will often see ratios in the description of codecs:



4:2:2



4:1:1



4:2:0

Activity Masking

Noise harder to see in Textured areas due to reduction in contrast sensitivity at higher spatial frequencies.

Activity Masking

A 100×100 block of noise has been added to each image at two locations. Because of activity masking it is much less visible in right image. Hence perceived quality of the right image should be higher.

Activity Masking

A 100×100 block of noise has been added to each image at two locations. Because of activity masking it is much less visible in right image. Hence perceived quality of the right image should be higher.

Putting it Together

There is a lot of statistical redundancy in images. For instance, in local image regions, say 8x8 blocks, the data tends to be flat or typically homogenous much of the time. This redundancy can be removed without affecting the image substantially.

The HVS response to image stimuli implies that one can introduce artefacts into images without them being seen. The colour subsampling illustrated this idea. Thus techniques that remove statistical redundancy can apply that concept heavily in regions where the resulting defects will not be noticed.

Efficient coding techniques can be used to represent any data as a more compact stream of digits. This technology can be used both for compression and error-resilience.

Quality Metrics

How to Assess Picture Quality?

Compression: how bad are the artefacts introduced? ?
Restoration: is the picture really better?

Subjective assessment (see ITU-R BT.500-11 recommendations), the subjects use a 5 point scale:

1. very annoying
2. annoying
3. slightly annoying
4. perceptible, but not annoying
5. imperceptible

Lots of subjects, tedious, complex calibration process.

Objective Metrics

Here are a few popular *objective* metrics. $I(\mathrm{x})$ is the image pixel, $G(\mathrm{x})$ is the ground truth/reference pixel and N is the number of pixels.

Mean Square Error

$$MSE = \frac{1}{N}\sum_{\mathrm{x}}(I(\mathrm{x}) - G(\mathrm{x}))^2$$

Mean Absolute Error

$$MAE = \frac{1}{N}\sum_{\mathrm{x}}|I(\mathrm{x}) - G(\mathrm{x})|$$

Objective Metrics

Peak Signal-to-Noise Ratio (PNSR) is used widely in image compression. This is the log of the ratio between the peak signal (image) power and the noise power:

$$PSNR = 10 \log_{10} \frac{255^2}{MSE}$$

Typically PSNR ranges:

30dB: annoying/very annoying.

35dB: perceptible.

40dB: barely perceptible

SSIM

The structural similarity (SSIM) index was introduced in 2004 to predict the perceived quality of an image.

It is based on some of aspects of the HVS discussed in this lecture, including Weber's law and activity masking, and is shown to perform better than standard metrics such as MSE or PSNR (see previous slide).

It is now widely used in Broadcast, but is still a matter of debate within the compression community.

In-depth study of SSIM is done in 5C1.