Human Image Perception

4c8 Media Signal Processing

Ussher Assistant Professor François Pitié

2021/2022

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

Introduction

We have seen the need for compression.

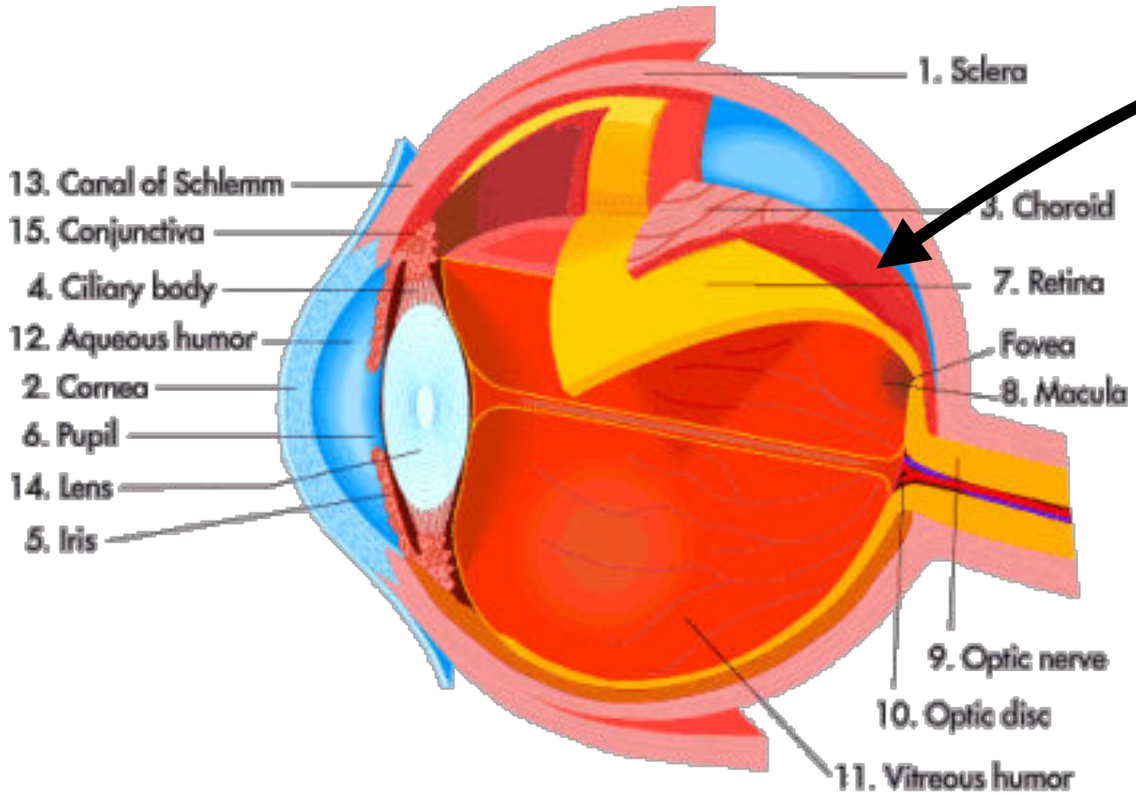
We can use what we have learned in information theory to exploit spatial and temporal redundancy but it is not enough

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**.

Colour Spaces

The Human Eye

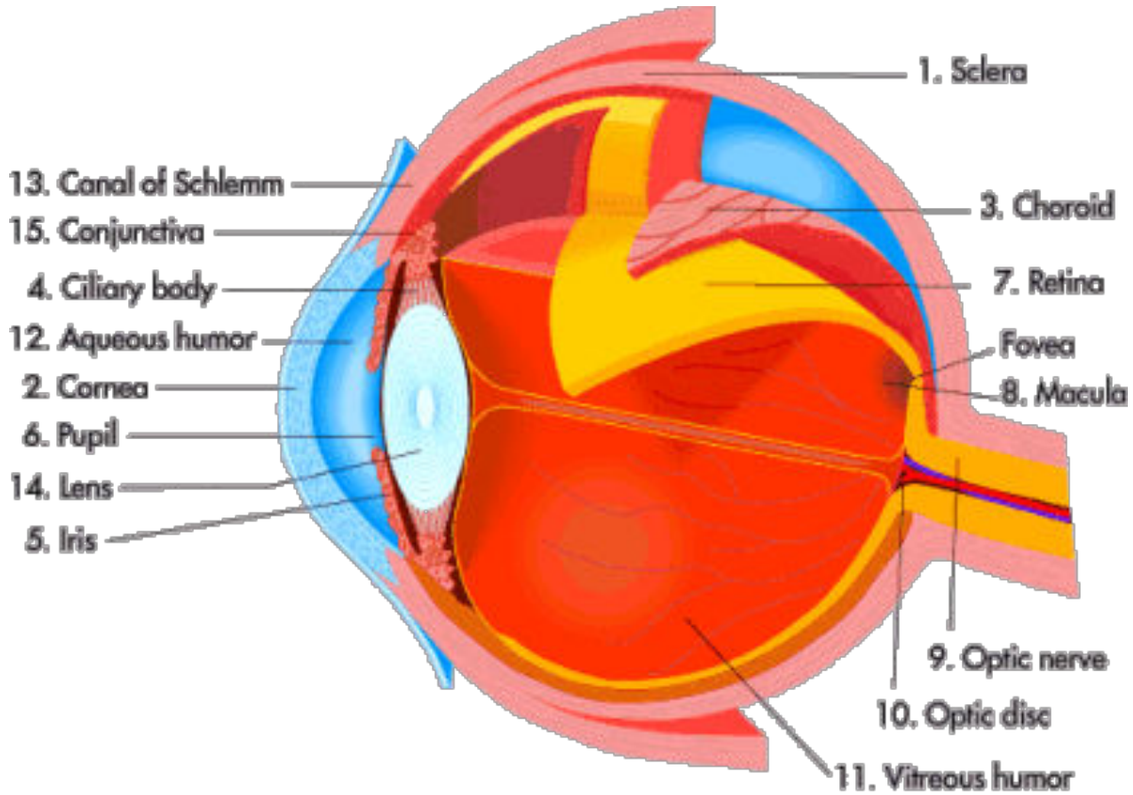


Light is focused onto the retina

CONES – sensitive to colour and luminance, located near the centre of the retina (fovea)

RODS – located near the periphery of the retina, much more sensitive to light, luminance only, more sensitive to motion, less resolution

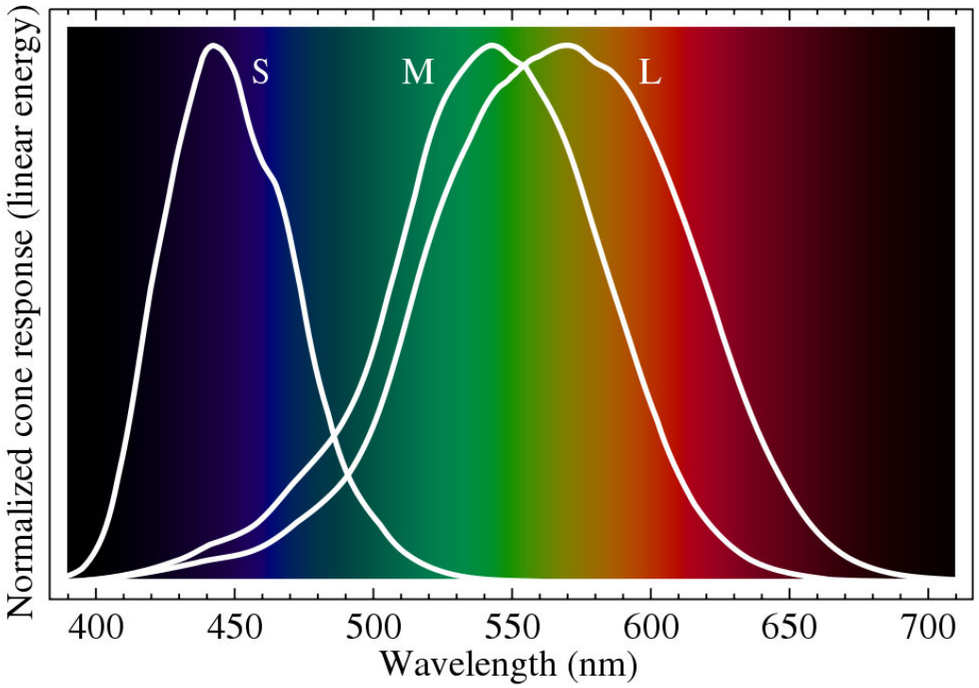
The Human Eye

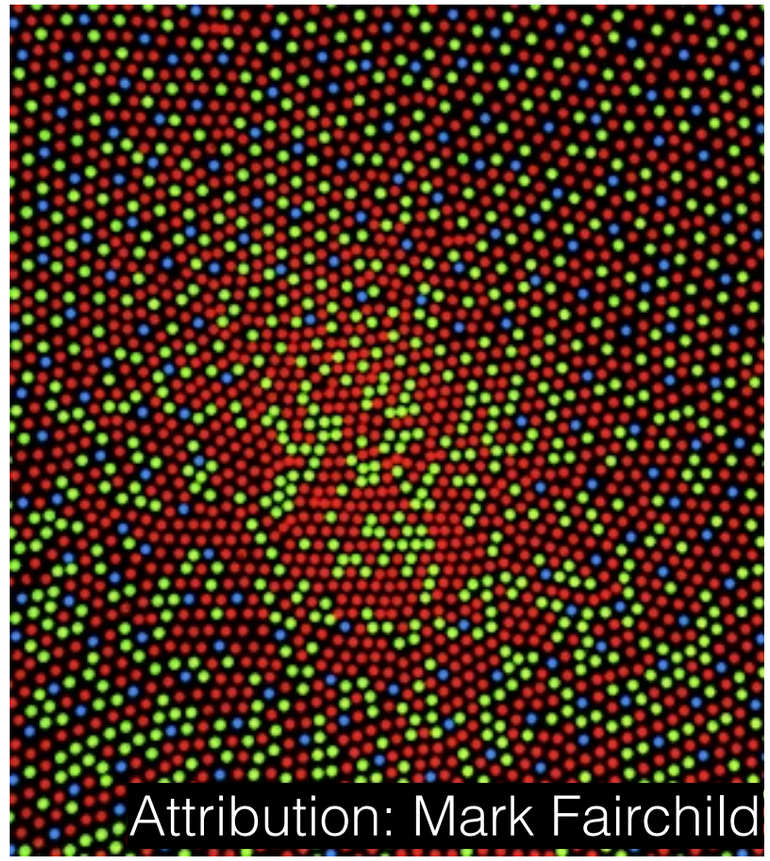


Electrical Impulses from the retina are channelled by the optic nerve to the Visual Cortex

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

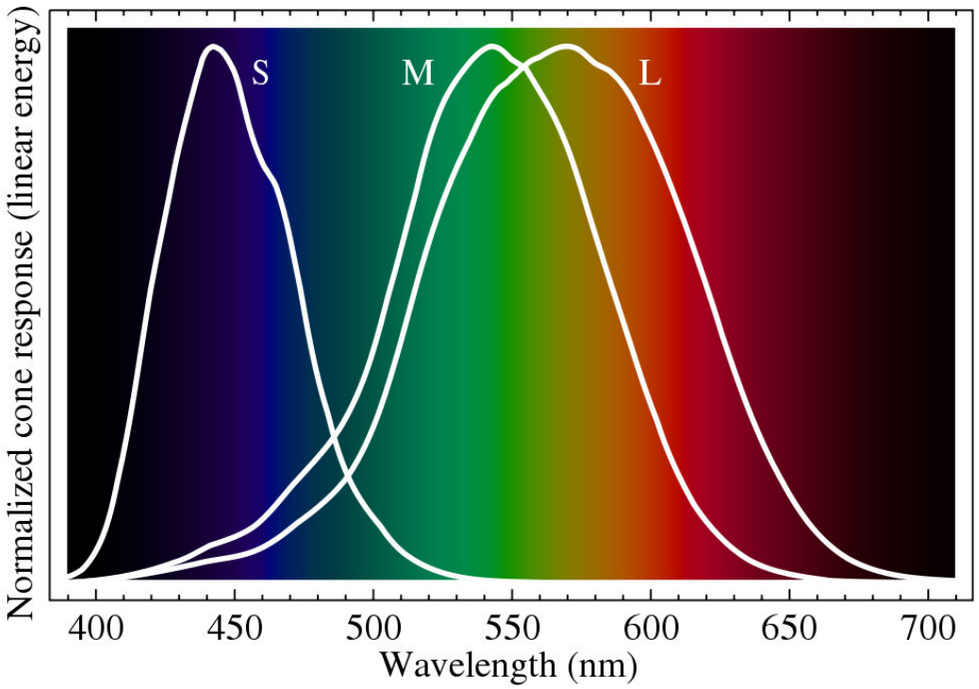
Cone Cells

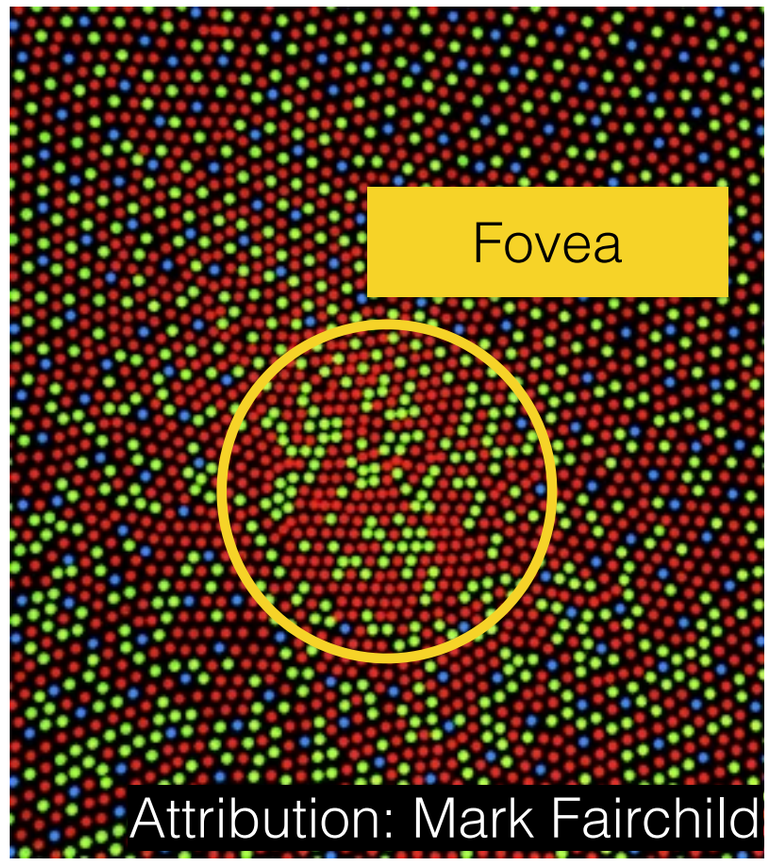




Cone Cells in the eyes convert wavelengths of life into 3 values known as a tri-stimulus. S (short), M (middle), L (long) cone cells.

Cone Cells

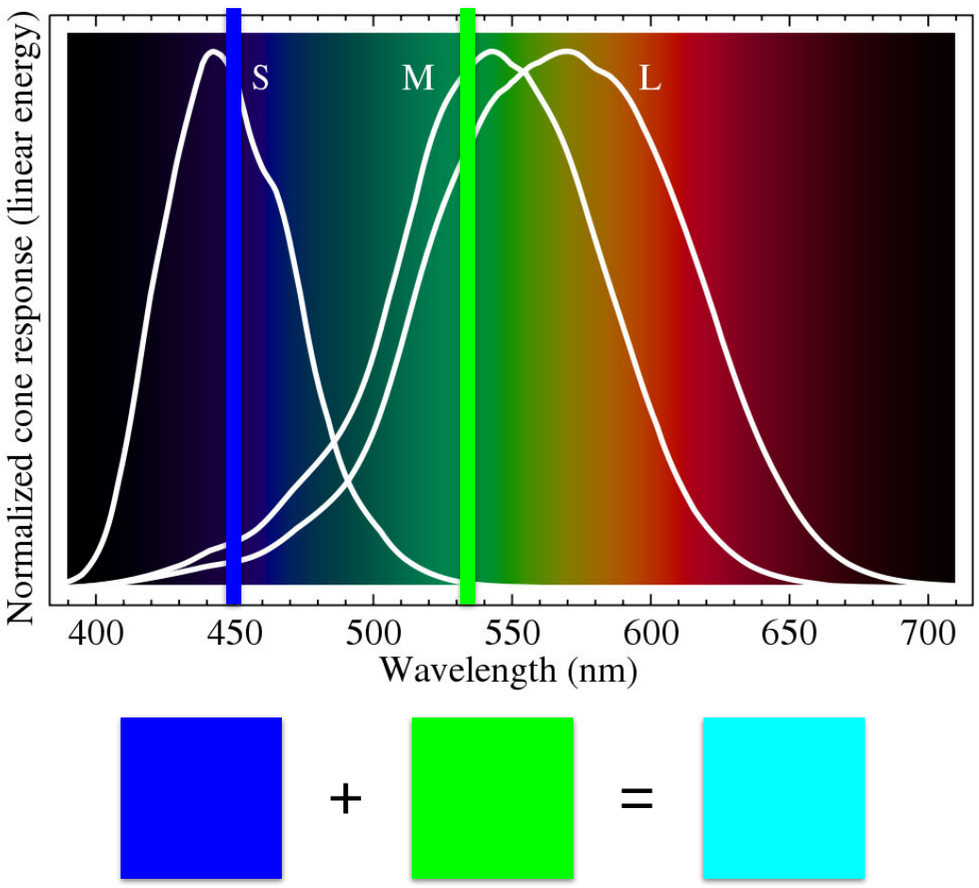


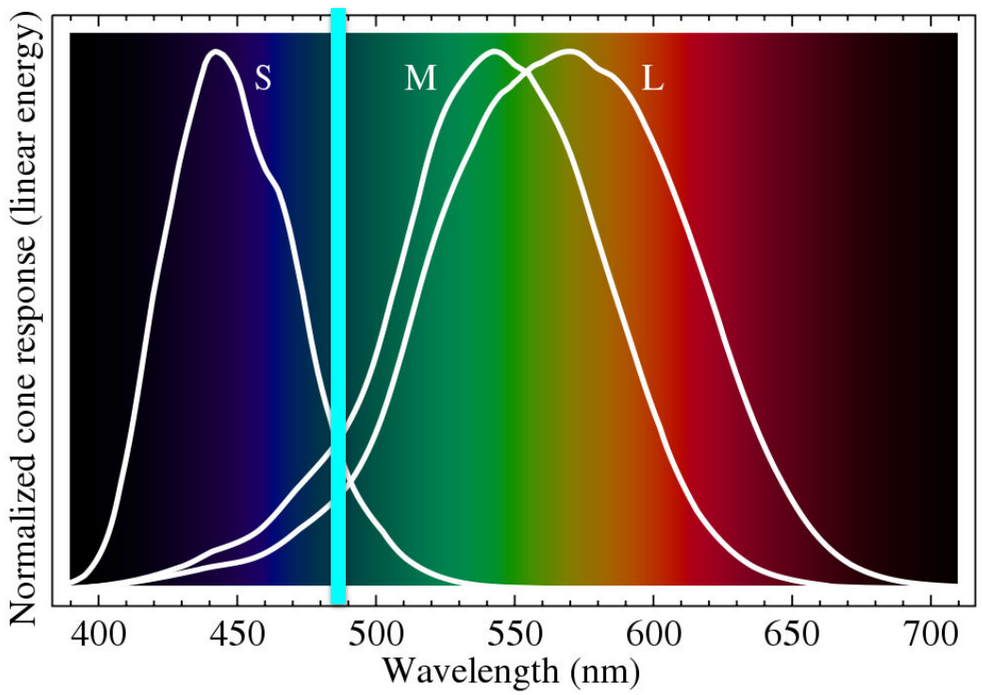


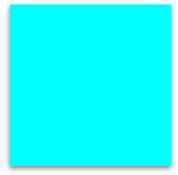
Cone Cells in the eyes convert wavelengths of life into 3 values known as a tri-stimulus. S (short), M (middle), L (long) cone cells.

The Fovea has very little S Cones.

Metamerism



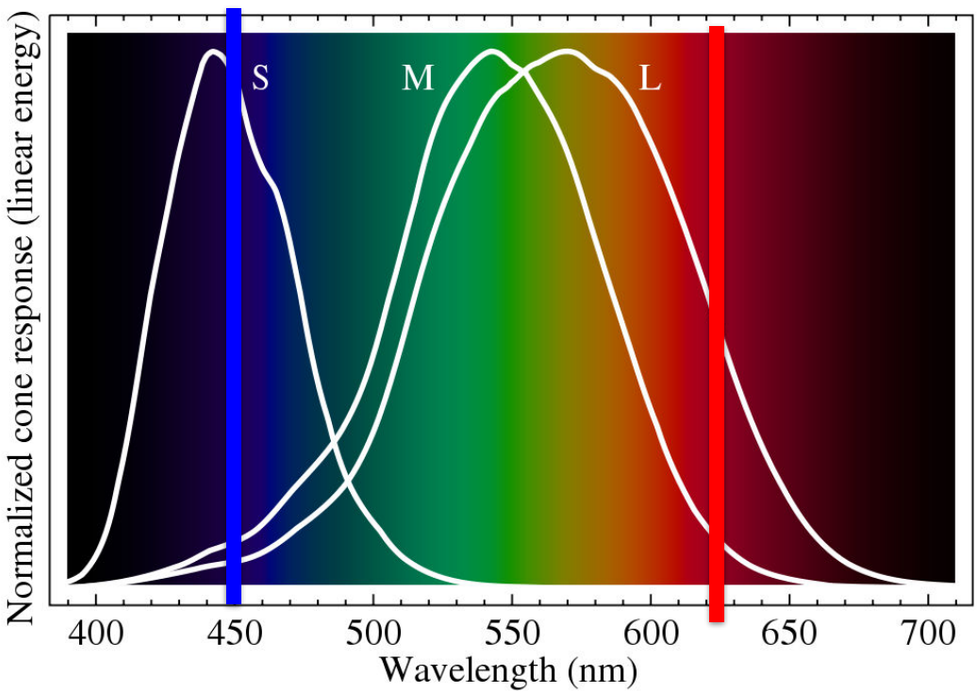


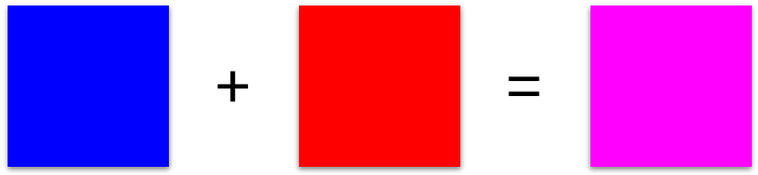


Light is a combination of wavelengths.
Different combinations may appear to be the same colour.

This is called metamerism.

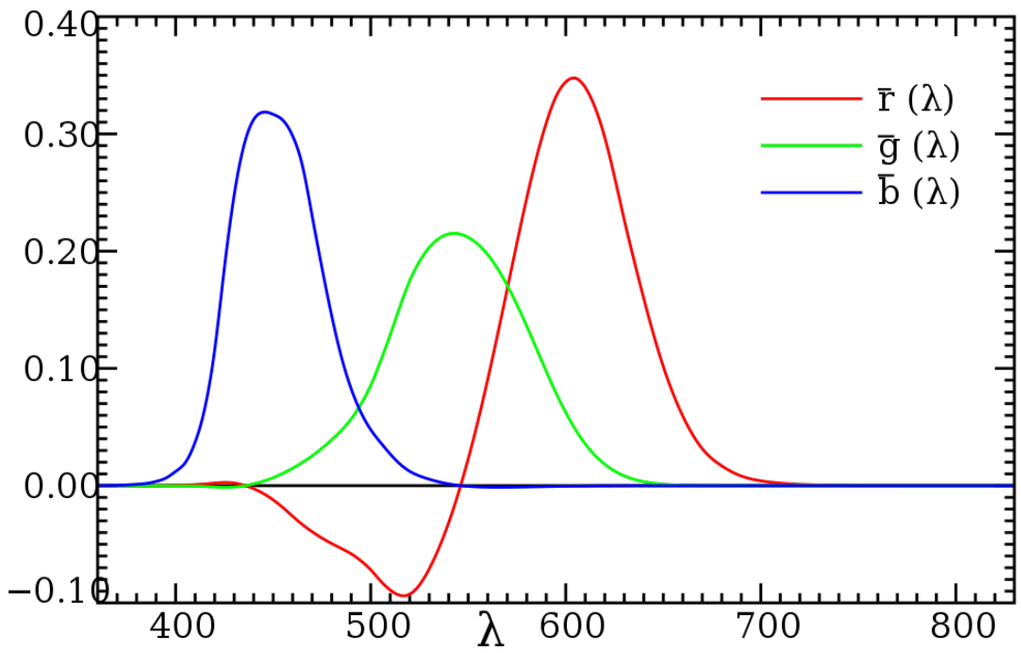
Purple





Purple is a combination of blue and red wavelengths.
There is no purple wavelength.

CIE-RGB Colour Space

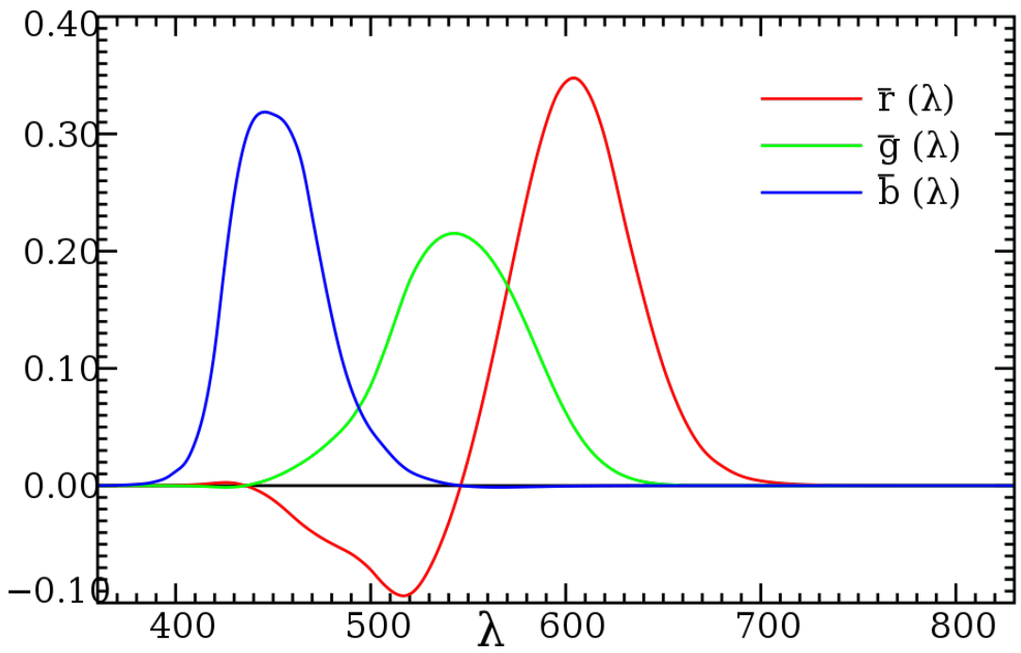


CIE RGB red = 700 nm, green = 546.1 nm, blue = 435.8 nm

How do we perceive a mono-chromatic light source as a function of 3 primary colours? (perceptual studies in the 1920's)

These functions are known as colour matching functions and can be used to estimate RGB values for any combination of colours.

CIE-RGB Colour Space



spectral power distribution

wavelength

matching function

$$R = \int_0^\infty S(\lambda)\bar{r}(\lambda)d\lambda$$

$$G = \int_0^\infty S(\lambda)\bar{g}(\lambda)d\lambda$$

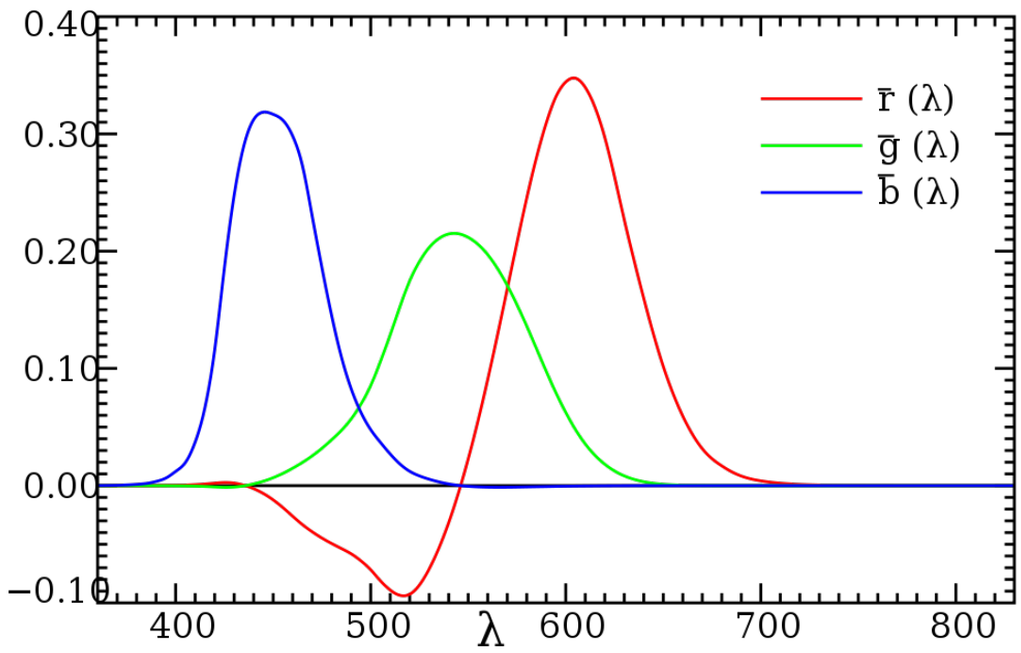
$$B = \int_0^\infty S(\lambda)\bar{b}(\lambda)d\lambda$$

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? (perceptual studies in the 1920's)

These functions are known as colour matching functions and can be used to estimate RGB values for any combination of colours.

CIE-RGB Colour Space



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

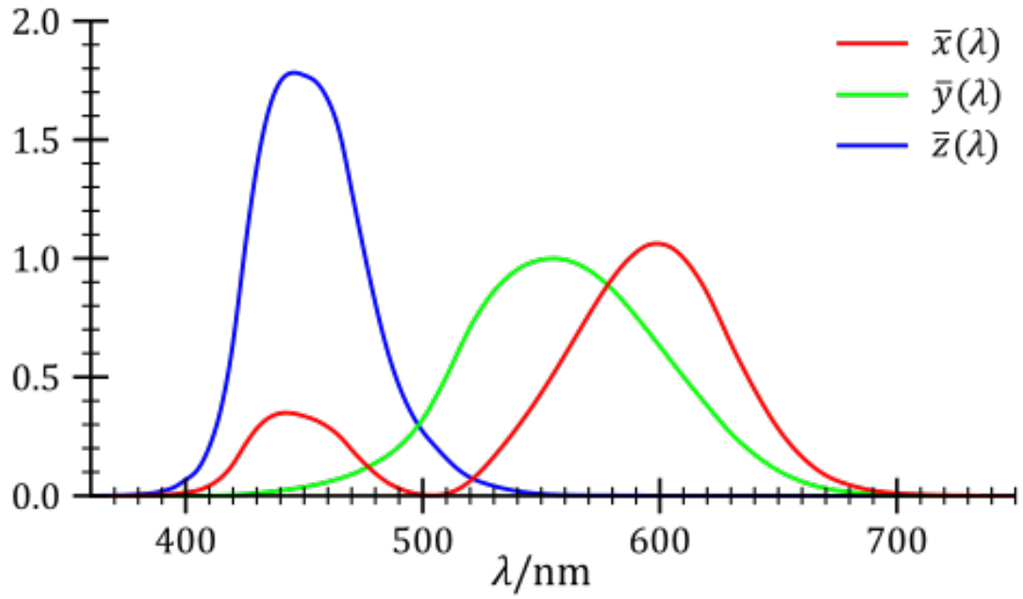
Negative values arose because no combination of r,g,b provided a good match. Participants had to add some amount of red or blue stimuli to the target wavelength.

CIE-RGB Colour Space

Links:
- Chandler Abraham [https://goo.gl/vn8Wuv]

CIE 1931 XYZ Colour Space

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.

CIE 1931 XYZ Colour Space Chromaticity Diagram

XYZ is used to graph the chromaticity diagram.

let's normalise for luminance:

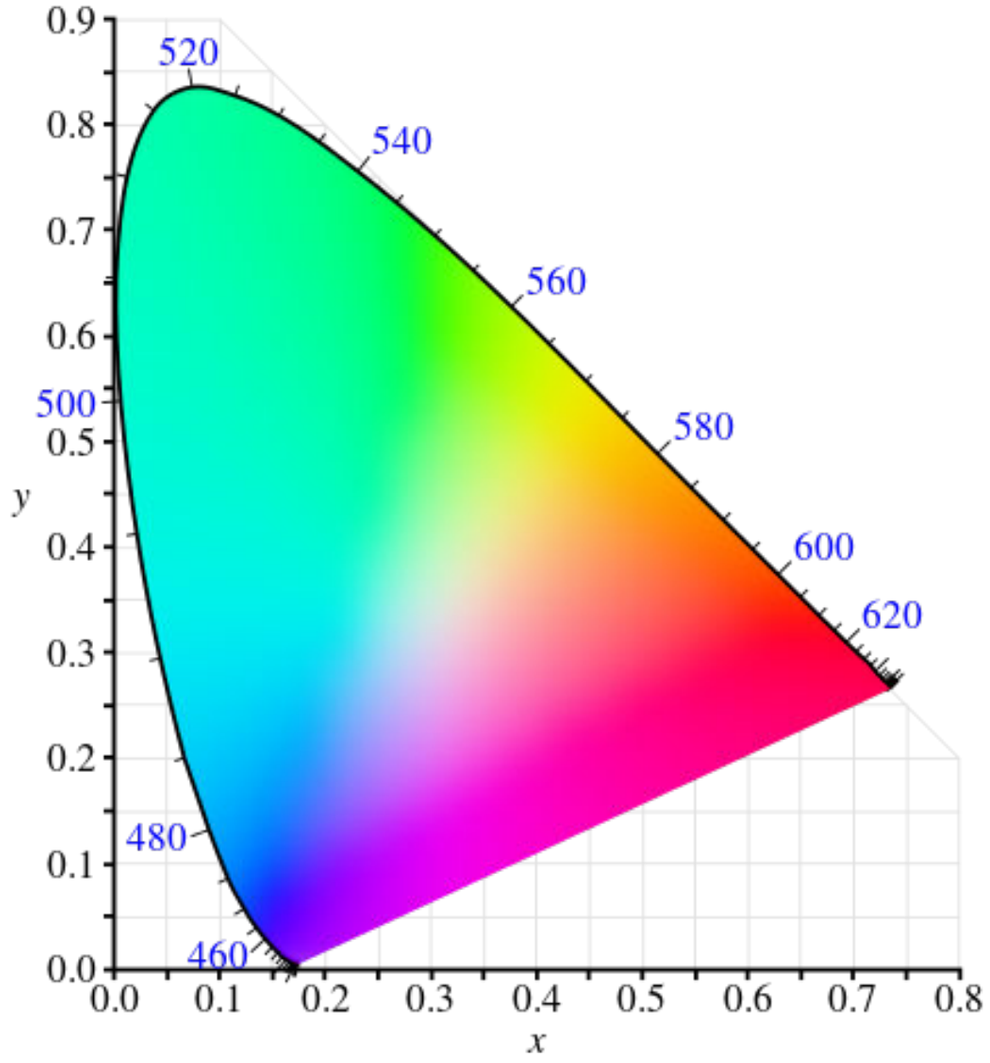
$$x = X / (X+Y+Z)$$

$$y = Y / (X+Y+Z)$$

Here are shown all colours visible by an average human on the x-y plane.

This is the human gamut.

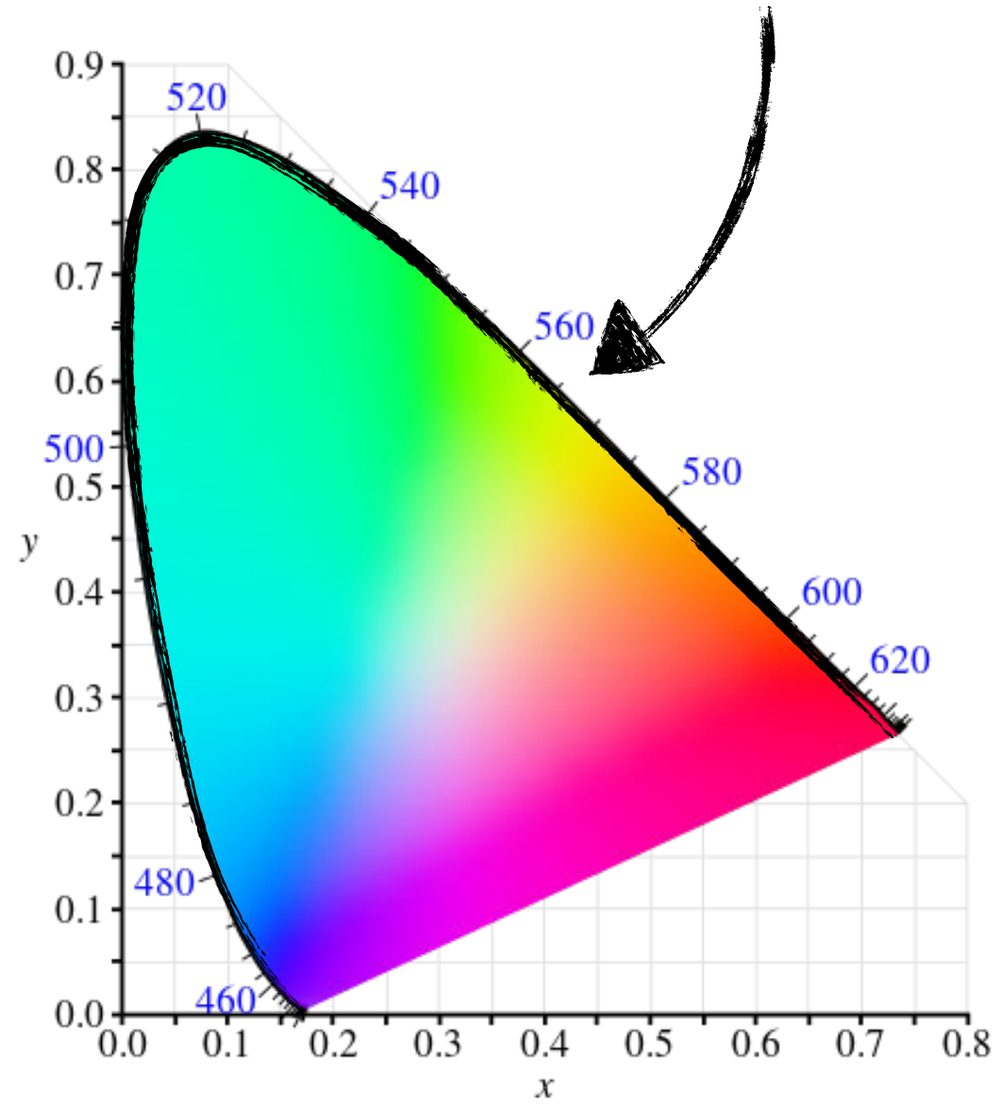
This chromaticity diagram is used for comparing the gamuts of different colour spaces.



The Human Gamut (XYZ Colour Space)

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

spectral locus

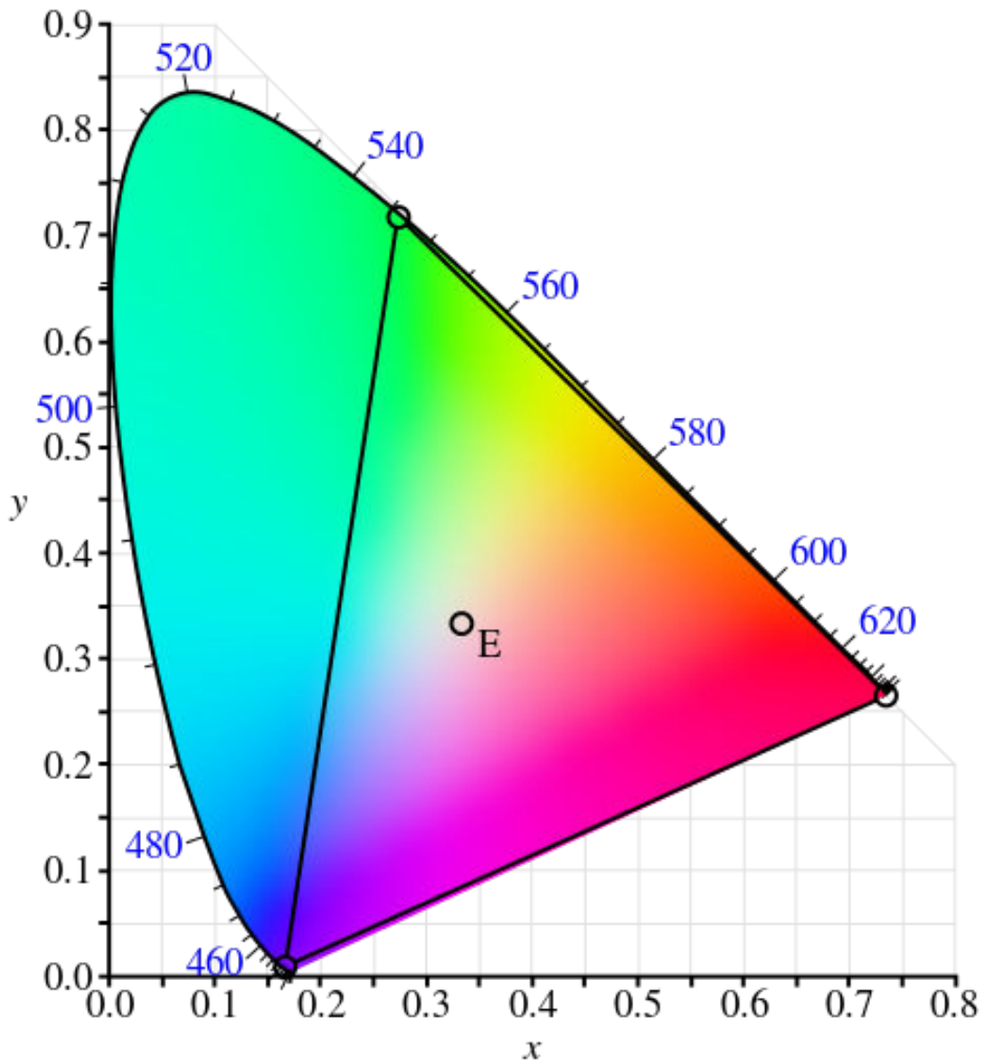


The RGB Gamut

The RGB gamut is a triangle that fits inside the human gamut.

Each combination of RGB values,
lie inside that triangle.

The vertices of the triangle
correspond to the 3 primary colours
Red, Blue, Green.

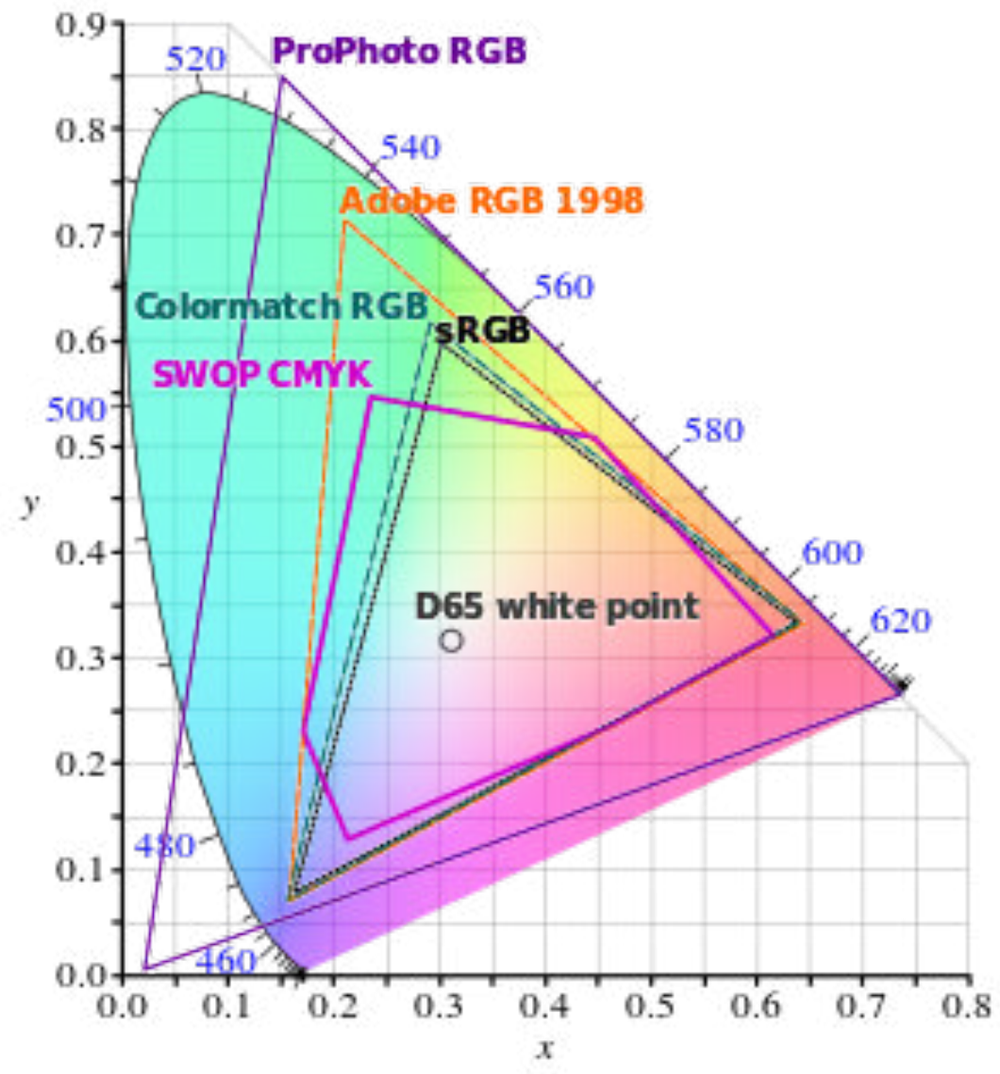


The RGB Gamut

There are many different definitions of RGB.

Each RGB colour space defines its own primaries.

That is, what is the bluest blue, redest red and greenest green.



YUV Colour Space (Broadcast and Compression)

By convention, colour spaces for broadcast use a tristimulus of 1 luminance (Y, same as in XYZ) and 2 chrominance values (U, V) to represent colour. YUV was used so that TV colour signals could be backward compatible with black and white TV sets.

The luminance is set as:

$$Y = 0.3\ R + 0.6\ G + 0.1\ B$$

Remark: exact values can vary

The higher weight for green reflects our sensibility to the green wavelength.

YUV Colour Space

The chrominance values U and V are defined as follows:

$$U = 0.5(B - Y)$$
$$V = 0.625(R - Y)$$

putting everything together, we have a linear relationship between RGB and YUV:

$$\begin{bmatrix} Y \\ U \\ V \end{bmatrix} = \begin{bmatrix} 0.3 & 0.6 & 0.1 \\ -0.15 & -0.3 & 0.45 \\ 0.4375 & -0.375 & -0.0625 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix}$$

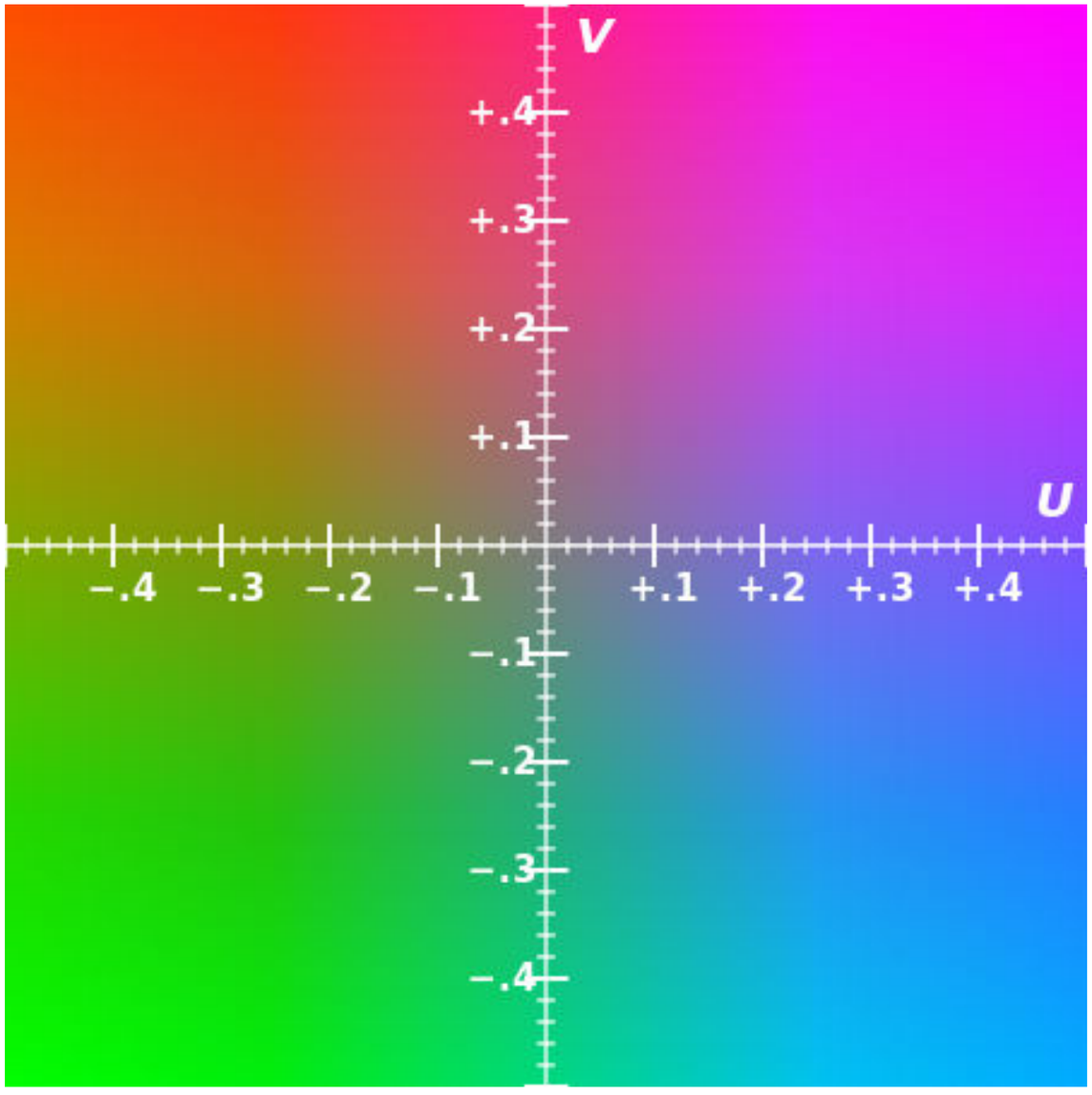
Remark: exact values can vary

YUV Colour Space

Using Floating point values

(ie. 0-1 range)

The UV plane, for Y=0.5



YUV Colour Space

A few examples (0-255 range)

Black	rgb = [0 0 0]	yuv = [0 0 0]
White	rgb = [255 255 255]	yuv = [255 0 0]
Gray	rgb = [x x x]	yuv = [x 0 0]
Red	rgb = [255 0 0]	yuv = [76.5 -38.3 111.6]
Green	rgb = [0 255 0]	yuv = [153 -76.5 -95.6]

Note: It is common to scale the U and V components so that it fits inside the range 0 to 255 (add 128 to both values)

YUV Colour Space

There are many variations on the YUV colour space

YUV – used in PAL colour TV

YIQ – used in NTSC colour TV

YDbDr – used in SECAM colour TV

YCbCr/YPbPr – used for digital TV and still image / video compression

Conversion from RGB to each of these colour spaces is linear but the conversion coefficients can vary slightly.

YUV Colour Space

Y

U

V

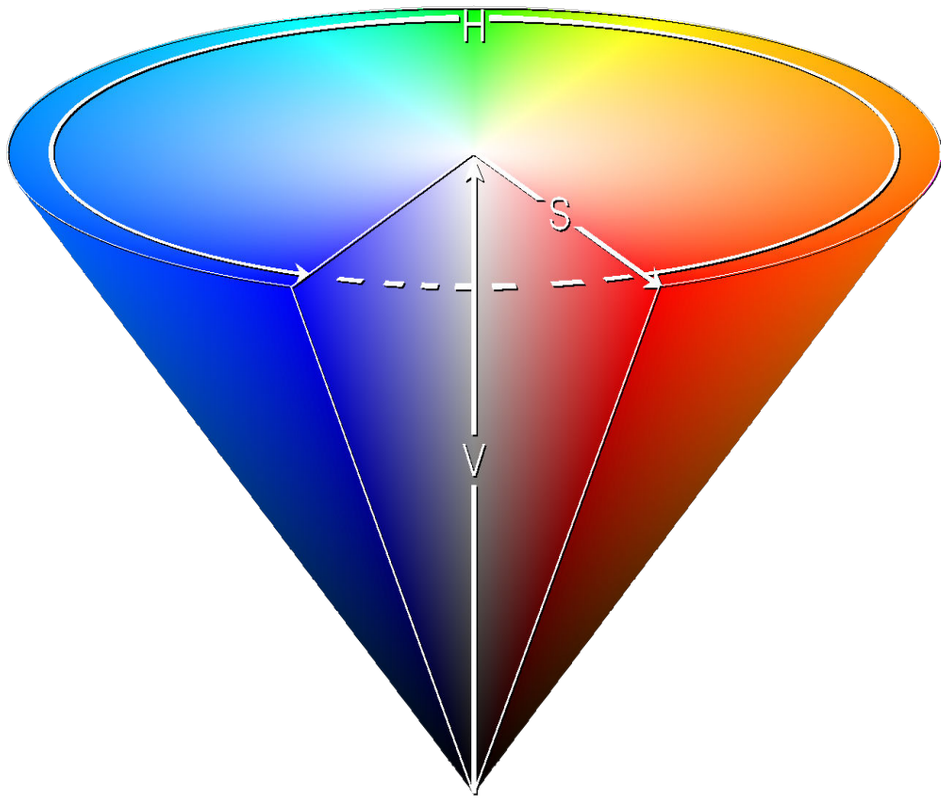
HSV Colour Space

Sometimes used in Image analysis

H – hue = the shade of a colour
(red, green, purple etc.)

S – saturation = colour depth
(from “washed out”/grey to vivid)

V – Value = brightness of the colour



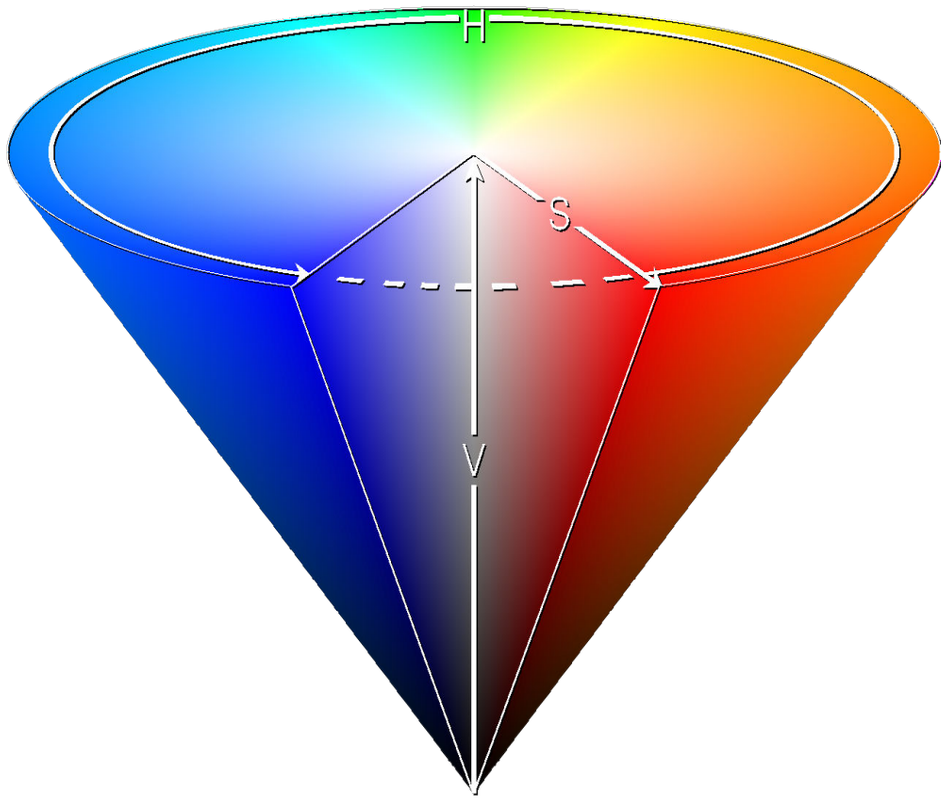
HSV Colour Space

Conversion

$$V = \max(R, G, B)$$

$$S = V - \min(R, G, B)$$

$$H = \begin{cases} \frac{G-B}{6S} & V = r \\ \frac{2S+B-R}{6S} & V = g \\ \frac{4S+R-G}{6S} & V = b \end{cases}$$



HSV Colour Space

Hue

Saturation

Luminance

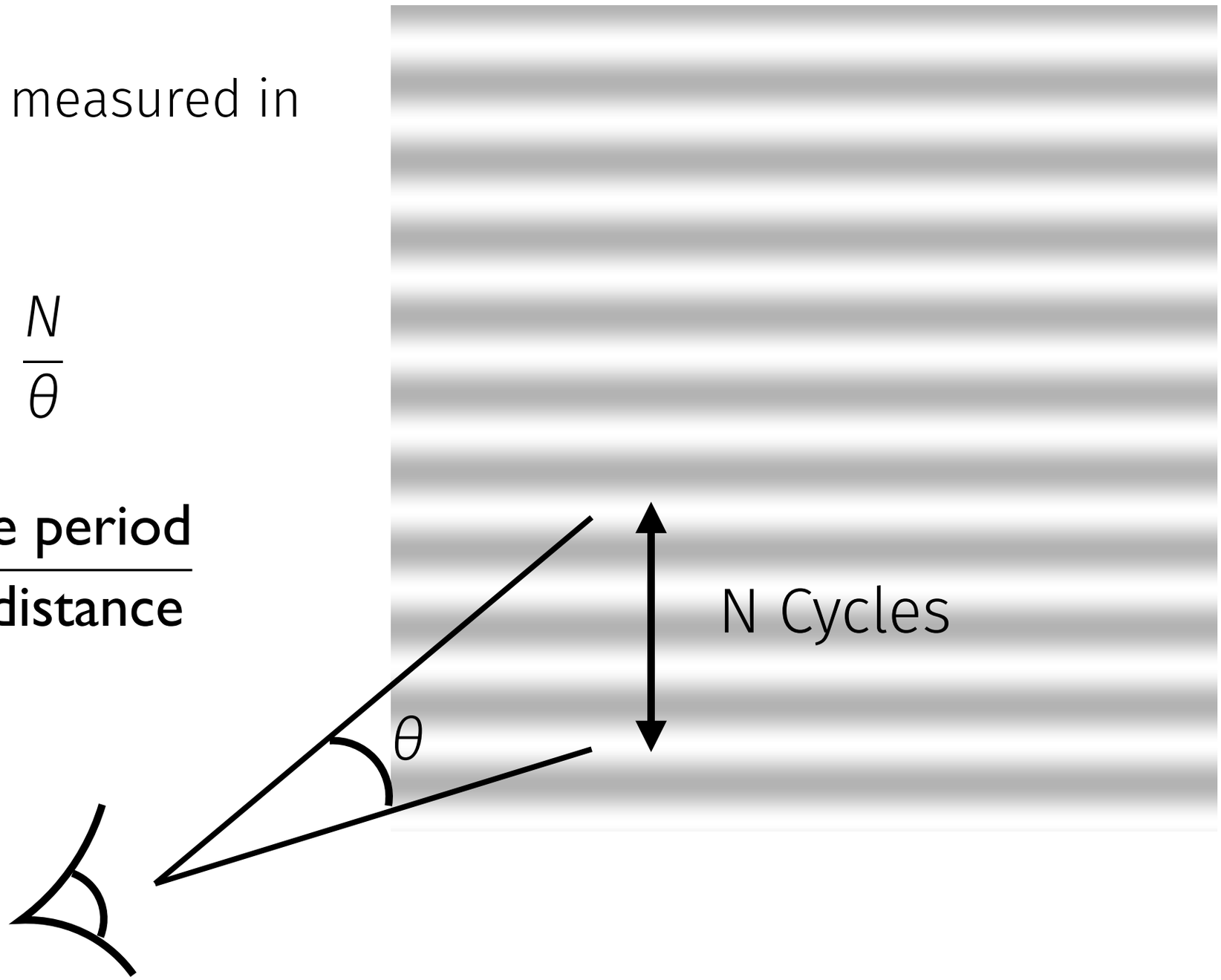
Contrast Sensitivity

Spatial Frequency Sensibility

Spatial frequency is measured in cycles per degree.

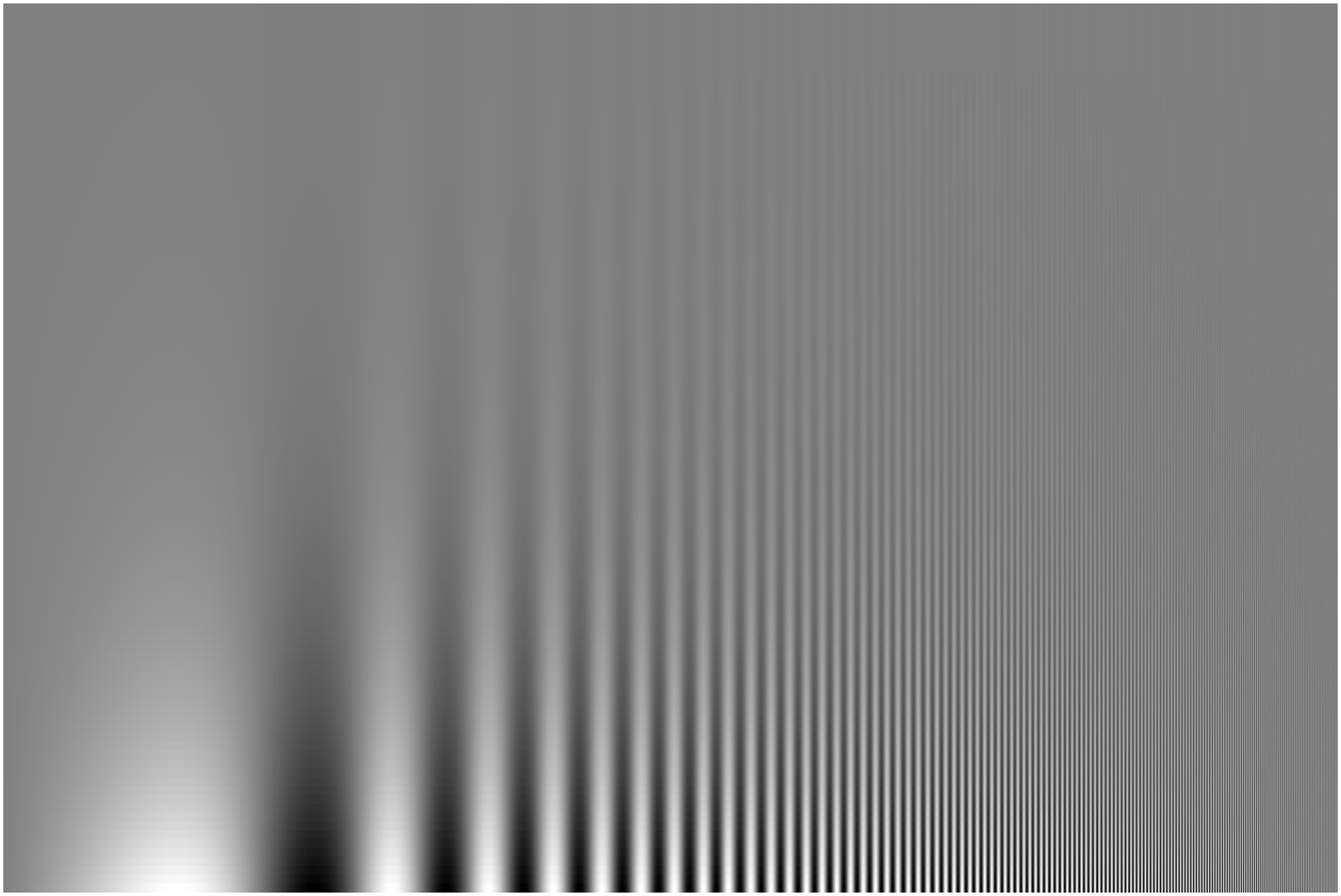
$$\textbf{spatial frequency} = \frac{N}{\theta}$$

$$\tan(\theta) = \frac{N \times \textbf{cycle period}}{\textbf{viewing distance}}$$



Spatial Frequency Sensibility

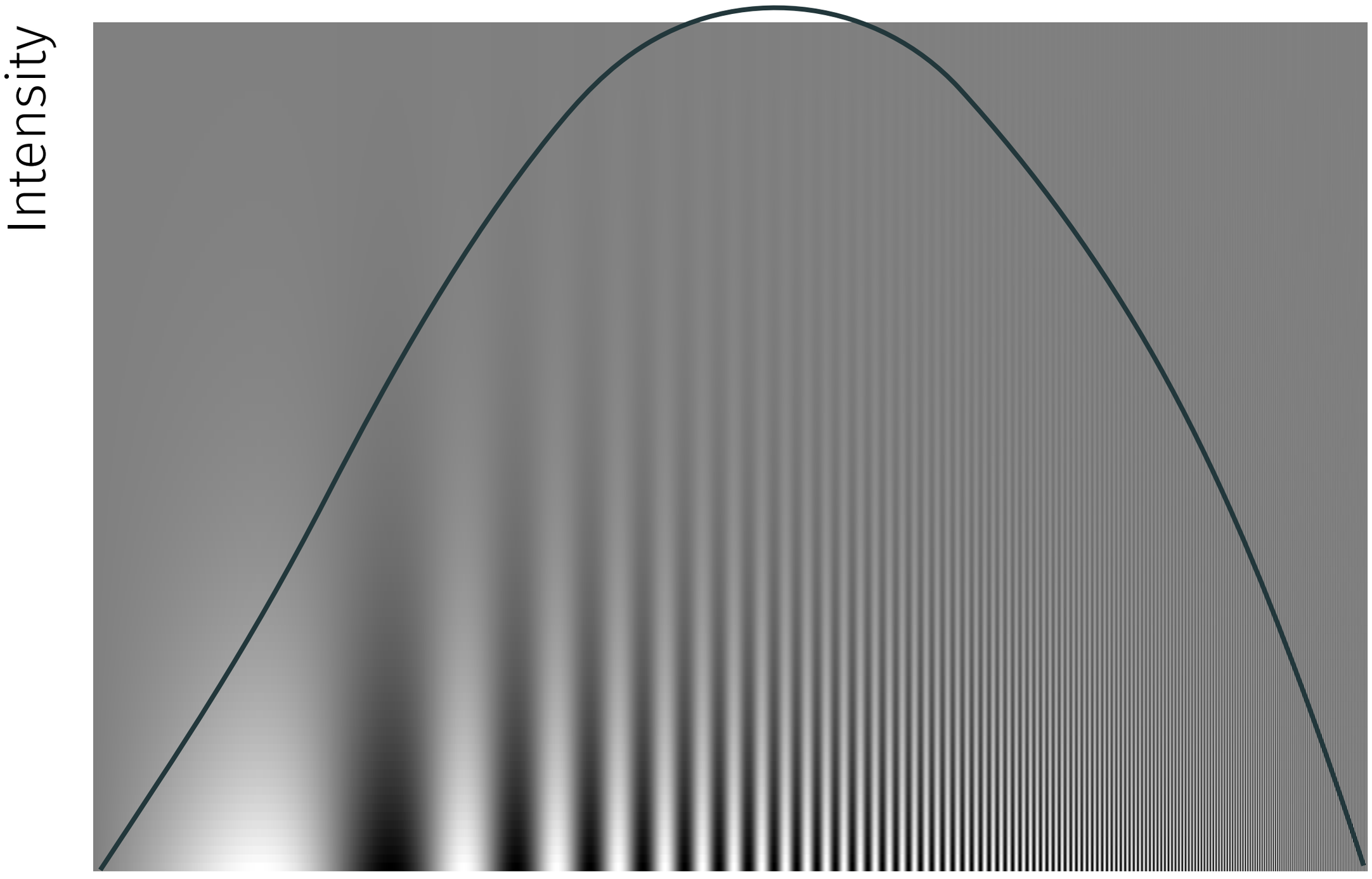
Intensity



Campbell-Robson Diagram

Spatial Frequency

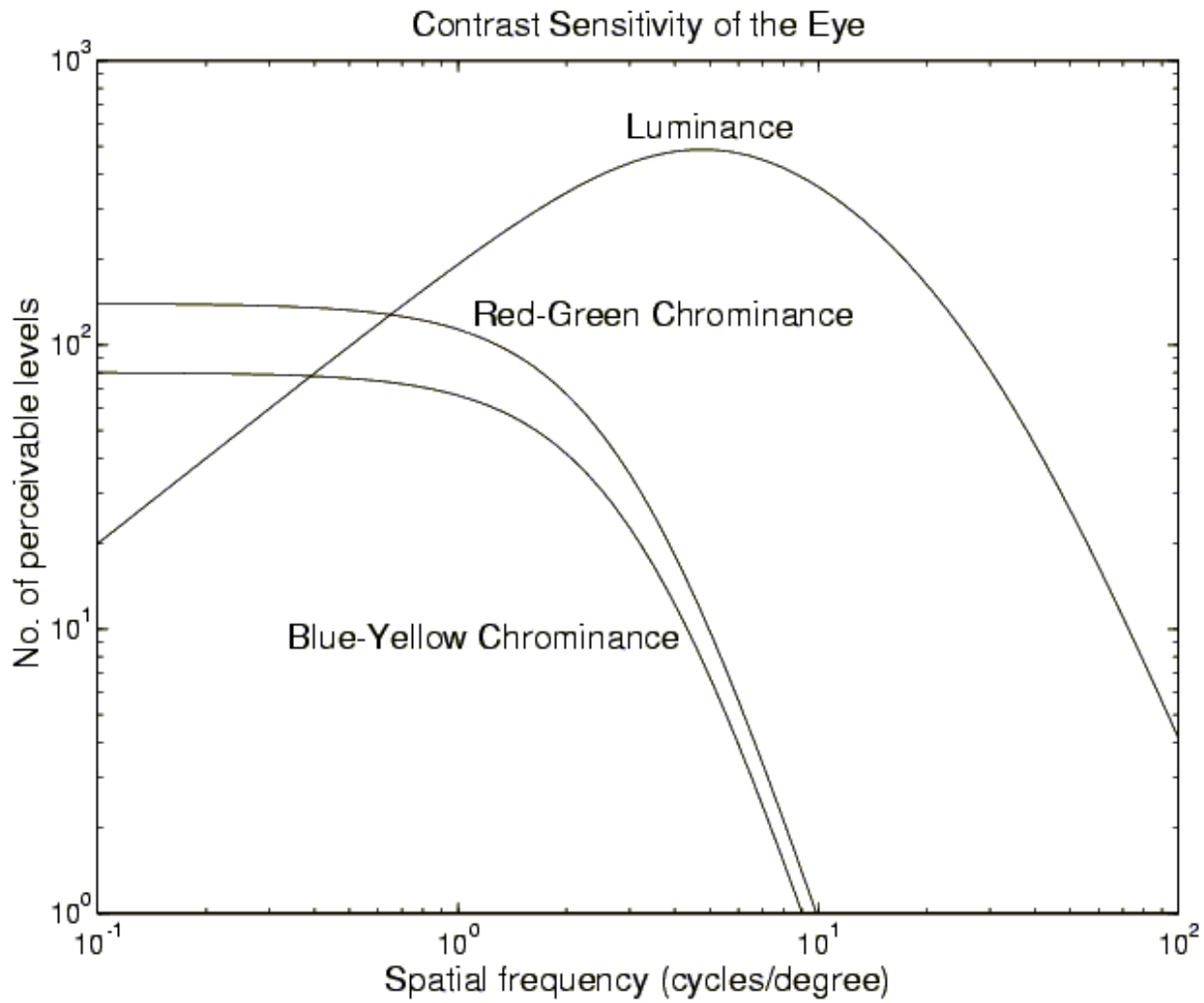
Spatial Frequency Sensibility



Campbell-Robson Diagram

Spatial Frequency Sensibility

using log scale, we get these contrast sensitivity graphs for Y,U,V



HVS is less sensitive to higher frequencies.

HVS is less sensitive to chrominance than luminance

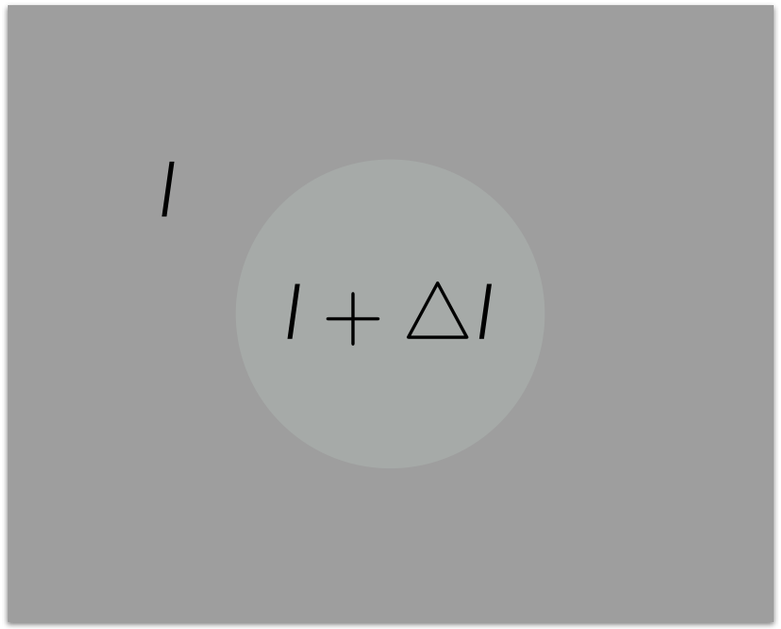
Threshold Visibility: Weber's Law

Weber's law relates the perceived brightness of an object to the brightness of its background. The law can be derived by measuring the 'Just Noticeable Difference' (ΔI) between two visual stimuli.

just noticeable difference

$$\frac{\triangle I}{I} = k$$

constant



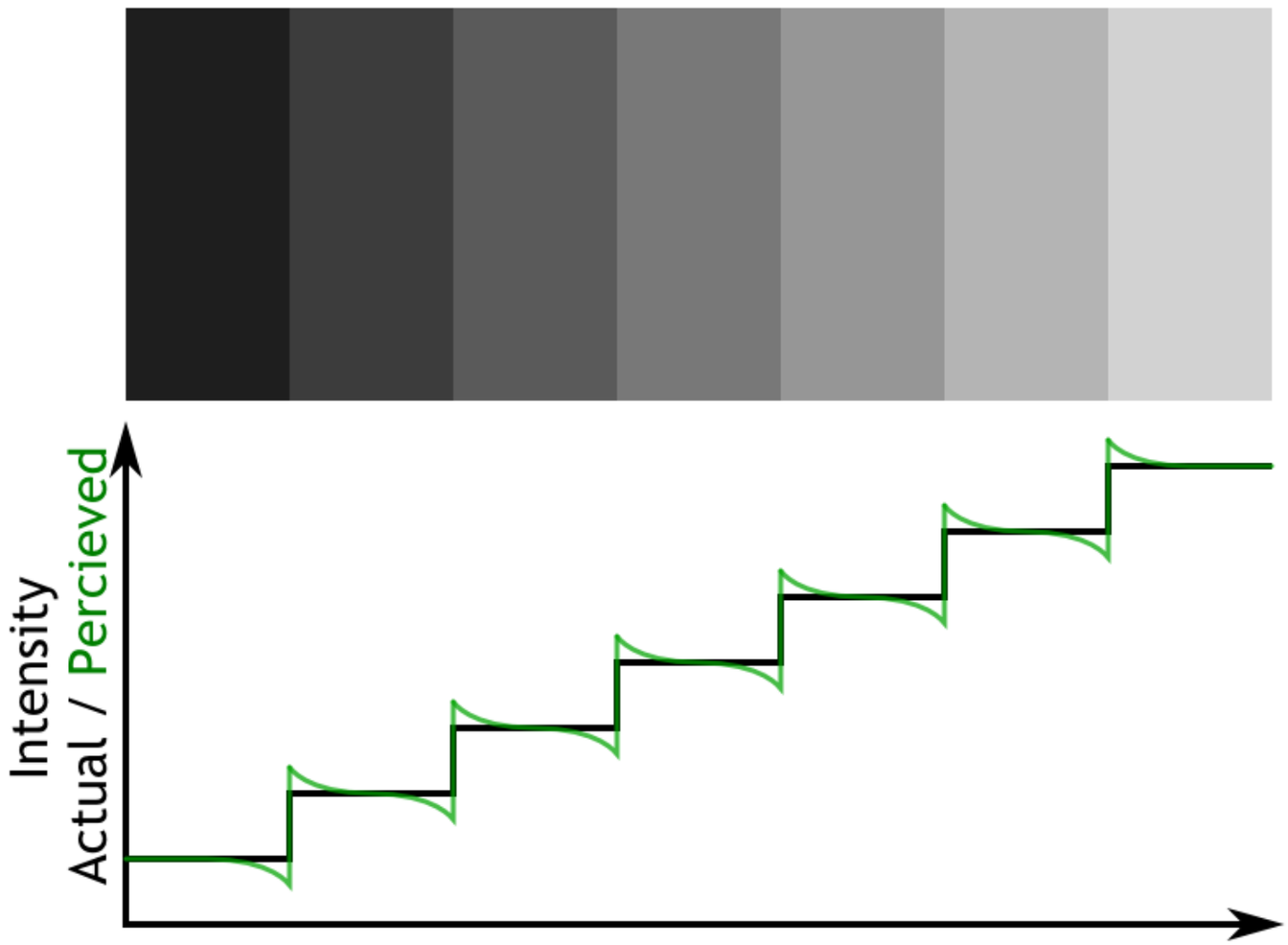
Weber's law implies that you need more Brightness difference to resolve an object against a bright background than against a dark background.

Weber's law is a measure of threshold visibility.

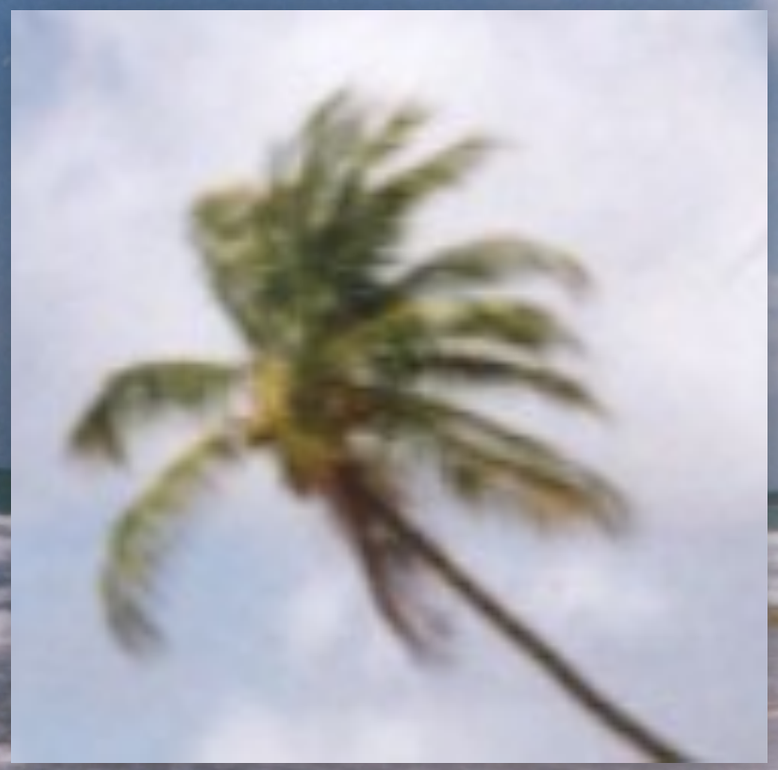
Contrast Sensitivity: Mach Banding

The HVS perceives that each vertical stripe looks brighter on the left and darker on the right.

This is an effect known as ‘Mach banding’. It is as a direct result of spatial filtering in the visual cortex.



Consequences of Spatial Frequency Sensitivity: Chroma Sub-Sampling



Original Image

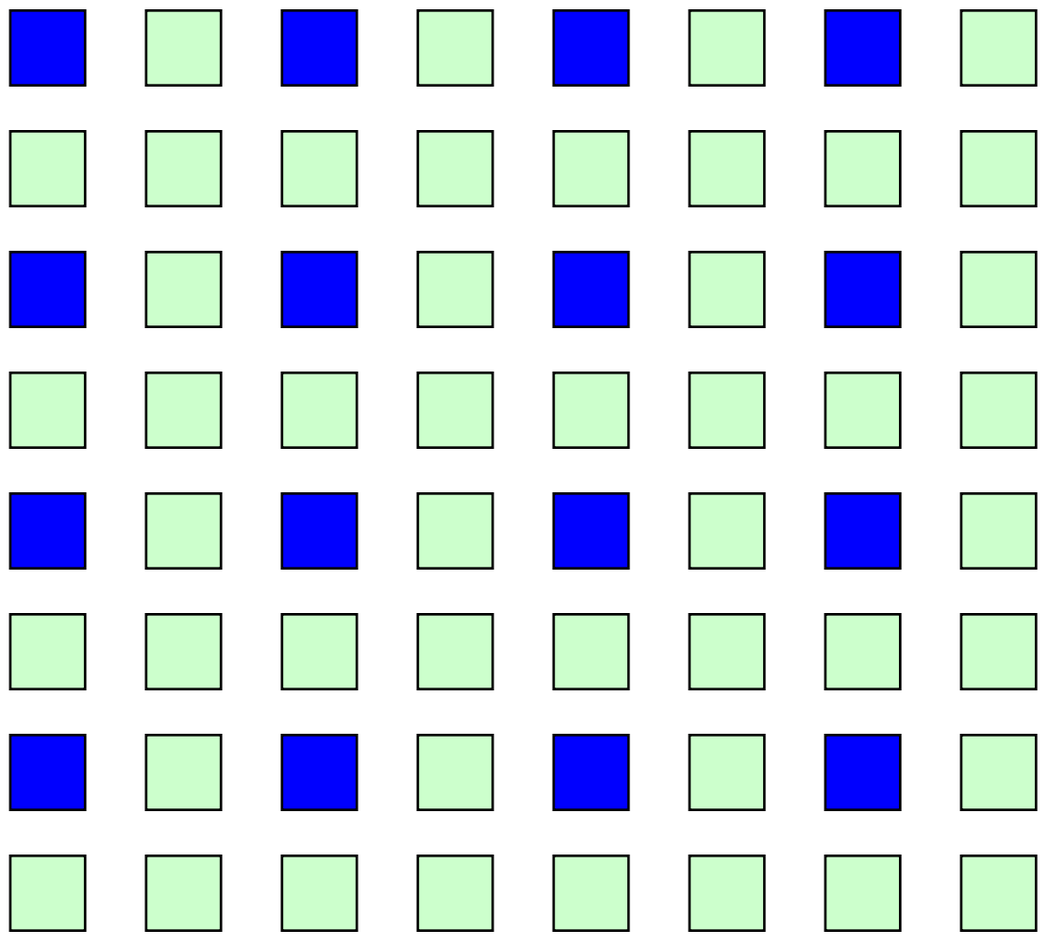
Chroma sub-sampling

We subsample the U and V chrominance channels and leave the Y channel alone

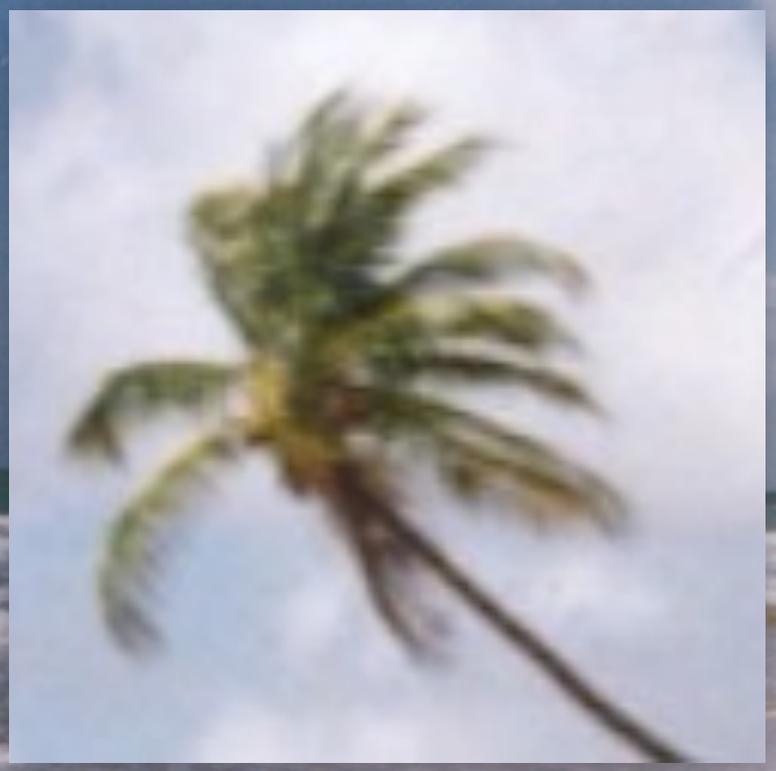
2:1 in both directions

Keep

Discard

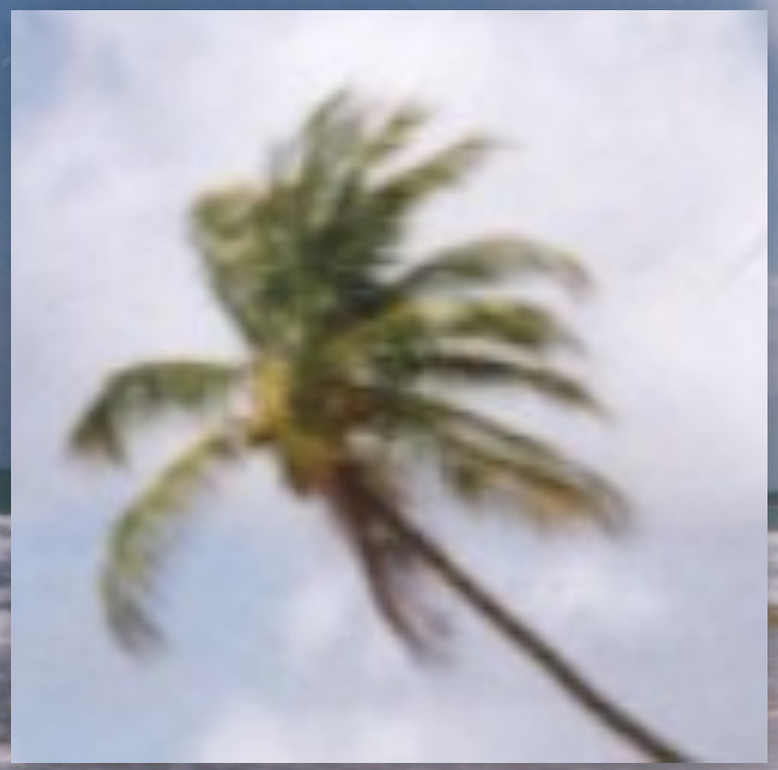


Chroma sub-sampling



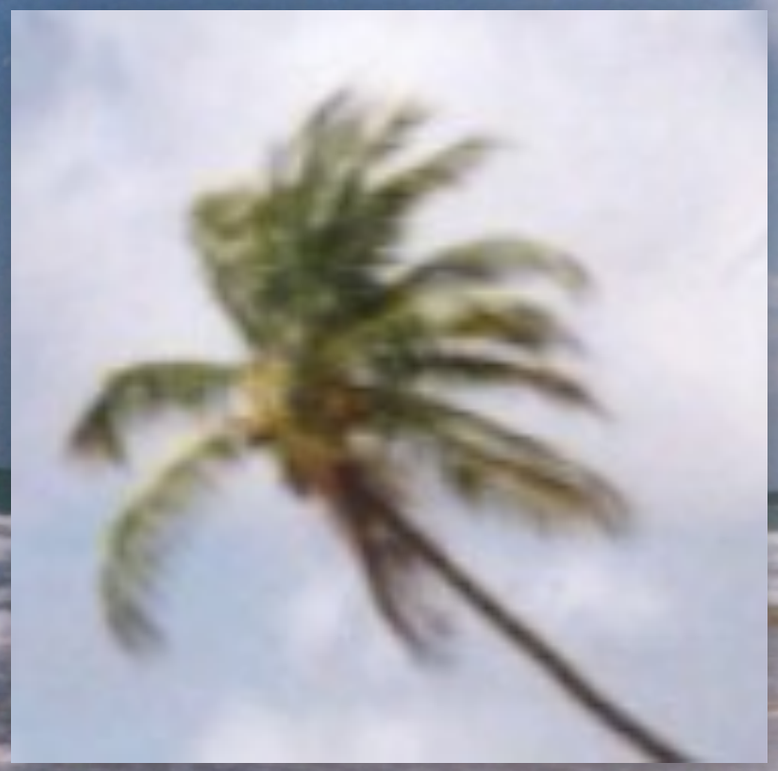
4:1 chrominance (U,V)
bandwidth: 50%

Chroma sub-sampling



Original Image

Chroma sub-sampling



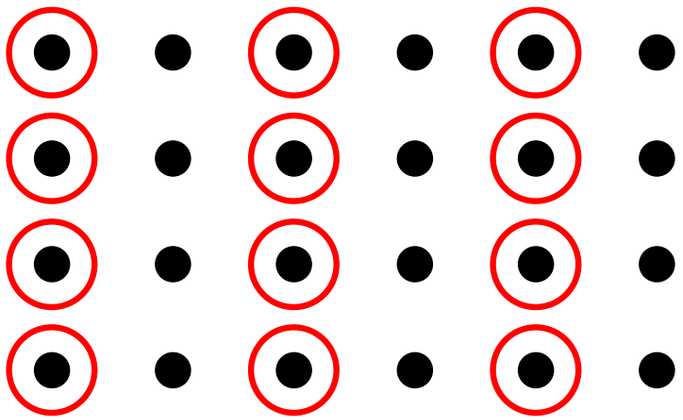
16:1 chrominance (U,V)
bandwidth: 37.5%

Chroma sub-sampling

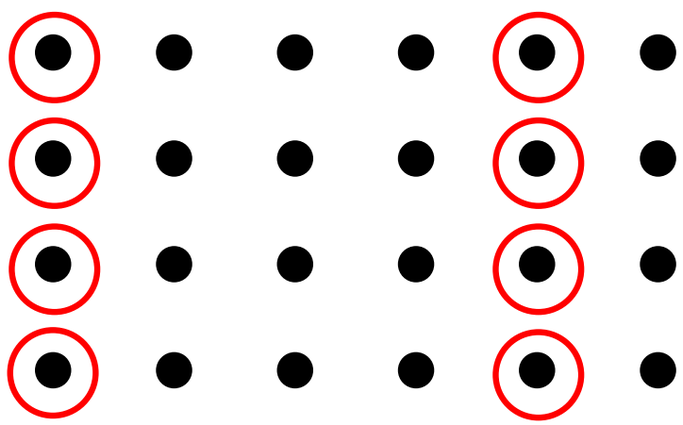
16:1 - only Y
bandwidth: 68.75%

Chroma sub-sampling

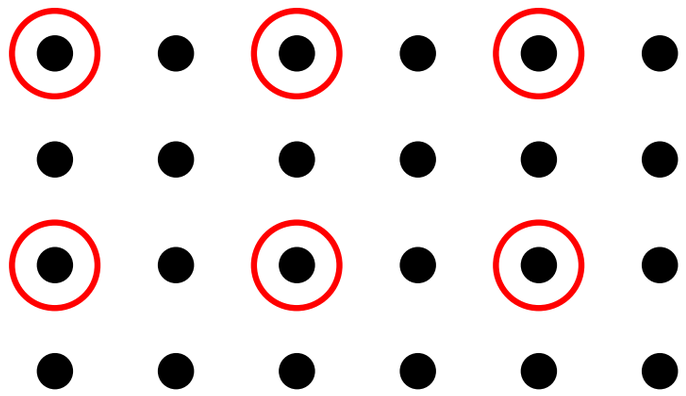
You will often see ratios in the description of codecs:



4:2:2



4:1:1



4:2:0

Consequences of Spatial Frequency: Activity Masking

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

Consequences of Spatial Frequency: Activity Masking

A 100 x 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.

Consequences of Spatial Frequency: Activity Masking

A 100 x 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.

Consequences of Spatial Frequency: Activity Masking

A 100 x 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

Putting it together: what makes compression possible?

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 actually 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...

Mean Square Error

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

$I(x)$ is the image pixel, $G(x)$ is the ground truth/reference pixel and N is the number of pixels.

Mean Absolute Error

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

Objective Metrics

The **Signal-to-Noise Ratio** is another popular measure and it has units in decibels (dB).

$$SNR = 10log_{10}\frac{\frac{1}{N}\sum_{\mathbf{x}} I(\mathbf{x})^2}{MSE}$$

This is a ratio between the signal power, measured as the sum squared intensities in the original image I, and the noise power measured as the MSE of the error.

Peak SNR (PNSR) is used widely in image compression. This is the log of the ratio between the peak signal (image) power and the noise power.

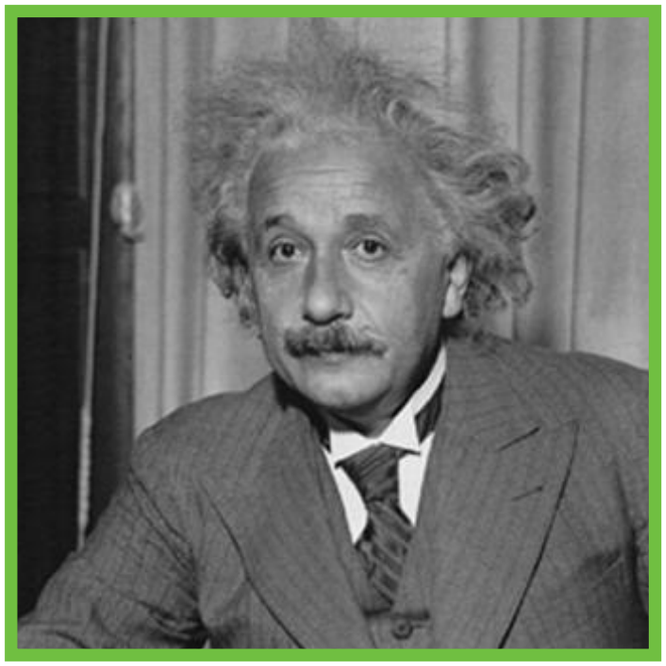
$$PSNR = 10log_{10}\frac{255^2}{MSE}$$

The issue with Objective Metrics

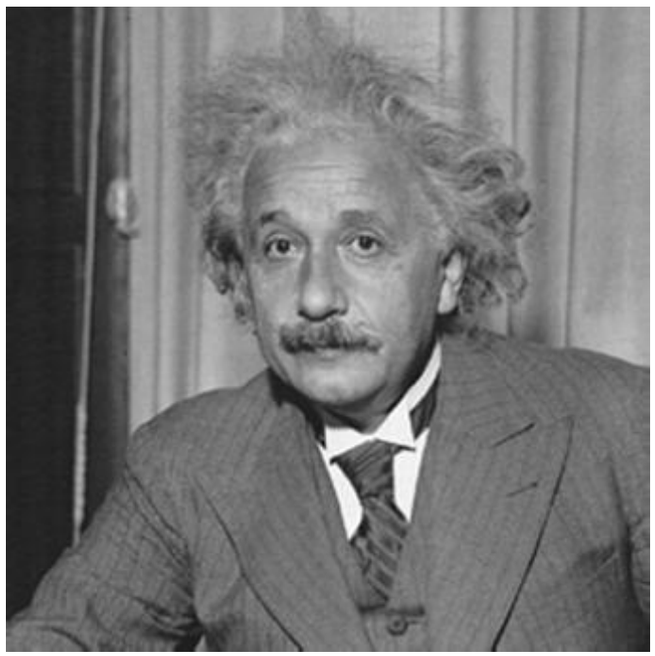
However, these metrics do not take into account the HVS.

The issue with Objective Metrics

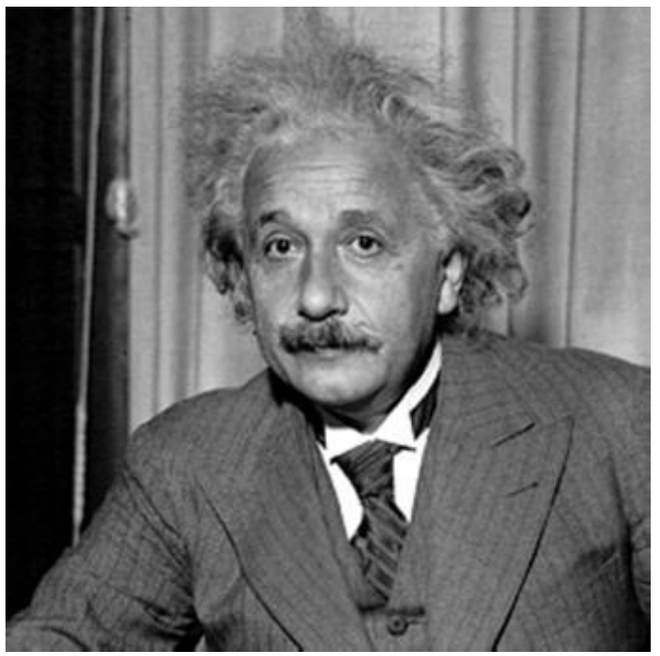
All these 5 degraded images have similar MSE score



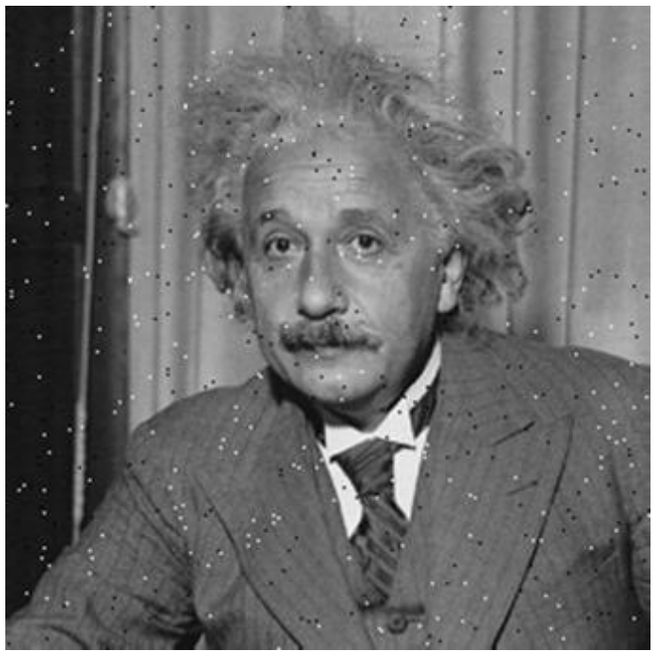
Original, MSE = 0; SSIM = 1



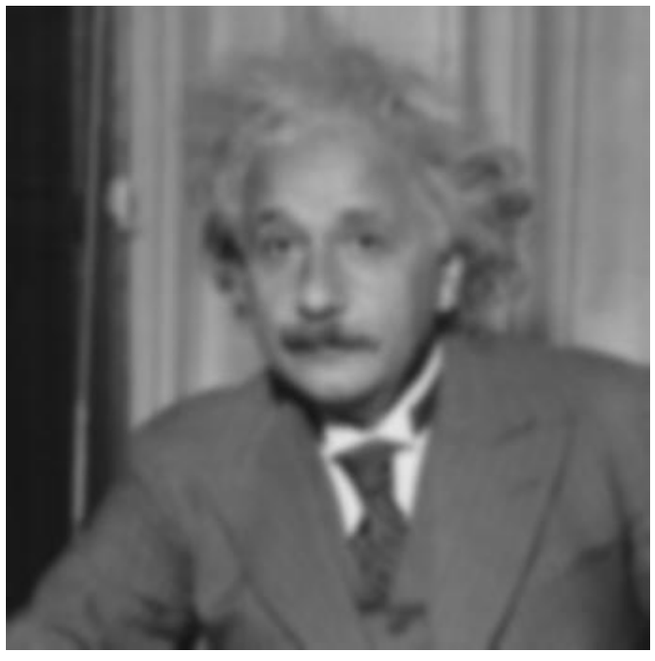
MSE = 144, SSIM = 0.988



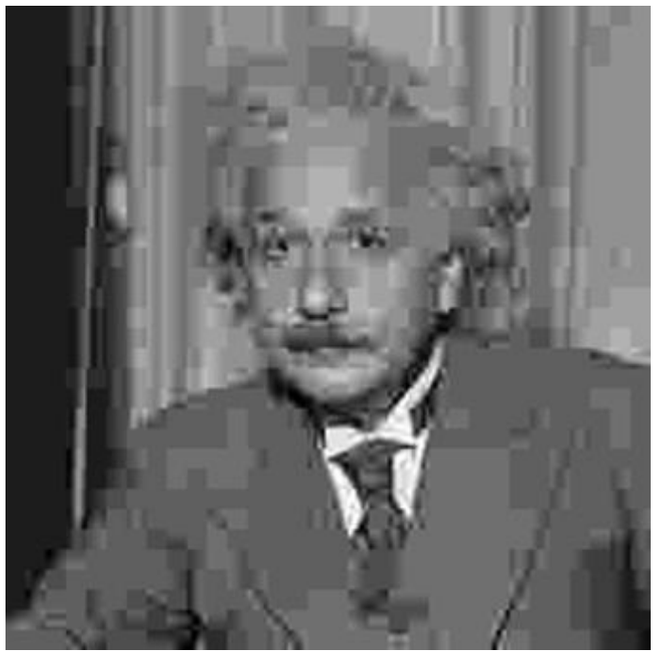
MSE = 144, SSIM = 0.913



MSE = 144, SSIM = 0.840



MSE = 144, SSIM = 0.694



MSE = 142, SSIM = 0.662

The issue with Objective Metrics

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

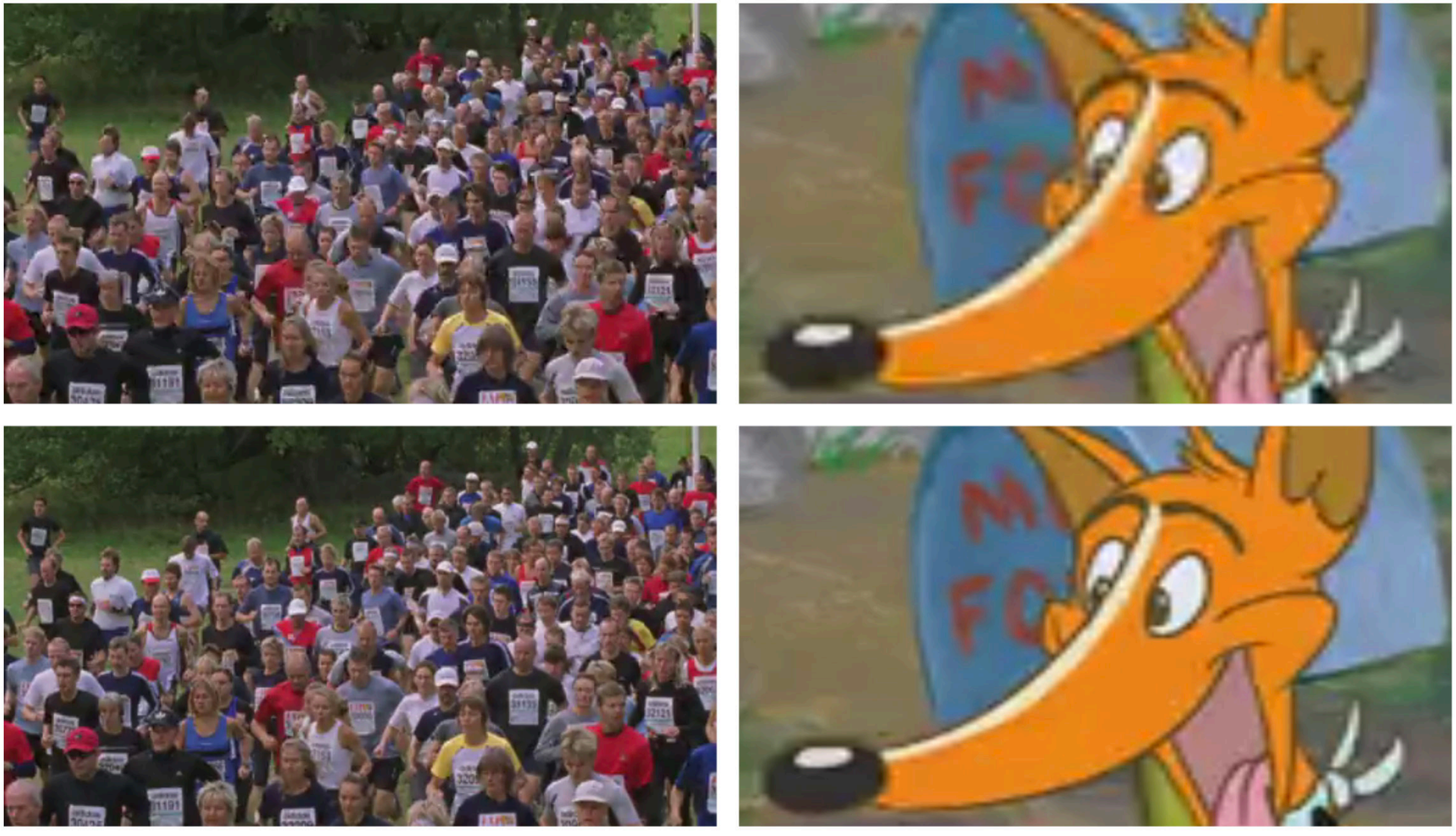
It is based on some of aspects of the HVS discussed in this lecture, including Weber's law and activity masking, and is shown to perform better than standard metrics such as MSE or PSNR (see previous slide).

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.

The issue with Objective Metrics

In videos, the problem is similar:



Frames from 4 videos; the two videos on top have a PSNR of about 31 dB, the bottom two have a PSNR of about 34 dB.

The issue with Objective Metrics

For videos, you can look at VMAF, a new perceptual video quality assessment algorithm, which was recently introduced by Netflix [https://github.com/Netflix/vmaf].

It combines multiple metrics using Machine Learning to match human ratings.

See Netflix Blog entry [https://goo.gl/dtdLTZ]

Summary

Summary

We introduced the RGB and YUV colour spaces

We discussed HVS factors that influence compression:

- Contrast sensitivity drops as spatial frequency increases.
- Contrast sensitivity is less for chrominance than for luminance.
- Activity masking

We discussed ways of measuring image quality necessary to quantify levels of degradation in compressed images.