History of Computer Vision and Human Vision System

簡韶逸 Shao-Yi Chien

Department of Electrical Engineering

National Taiwan University

Spring 2023

History of Computer Vision

1960s--1970s

Marvin Minsky, MIT
Turing award, 1969

"In 1966, Minsky hired a first-year undergraduate student and assigned him a problem to solve over the summer:

connect a camera to a computer and get the machine to describe what it sees."

Crevier 1993, pg. 88

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

PROJECT MAC

Artificial Intelligence Group July 7, 1966
Vision Memo. No. 100.

THE SUMMER VISION PROJECT

Seymour Papert

The summer vision project is an attempt to use our summer workers effectively in the construction of a significant part of a visual system. The particular task was chosen partly because it can be segmented into sub-problems which will allow individuals to work independently and yet participate in the construction of a system complex enough to be a real landmark in the development of "pattern recognition".

1960s--1970s

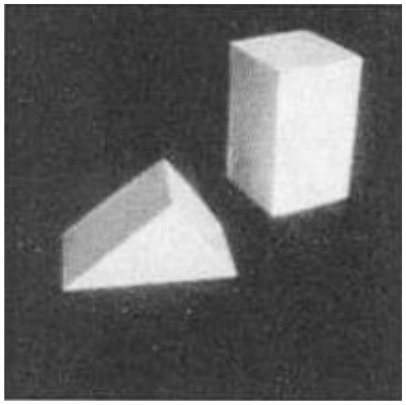
Marvin Minsky, MIT
Turing award, 1969

Gerald Sussman, MIT

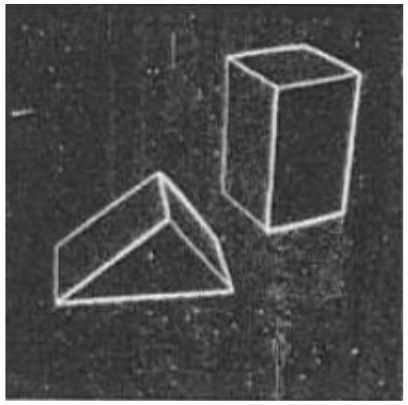
"You'll notice that Sussman never worked in vision again!" – Berthold Horn

1960s--1970s

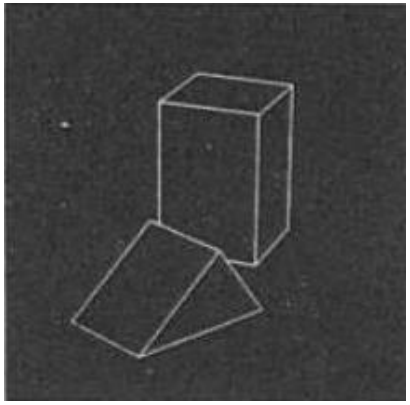
Larry Roberts
"Father of Computer Vision"



Input image



2x2 gradient operator

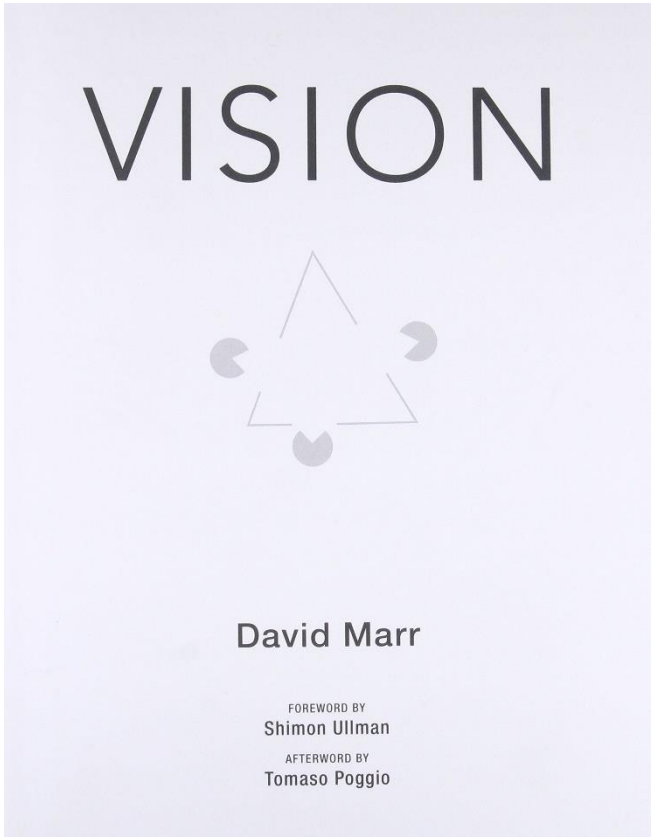


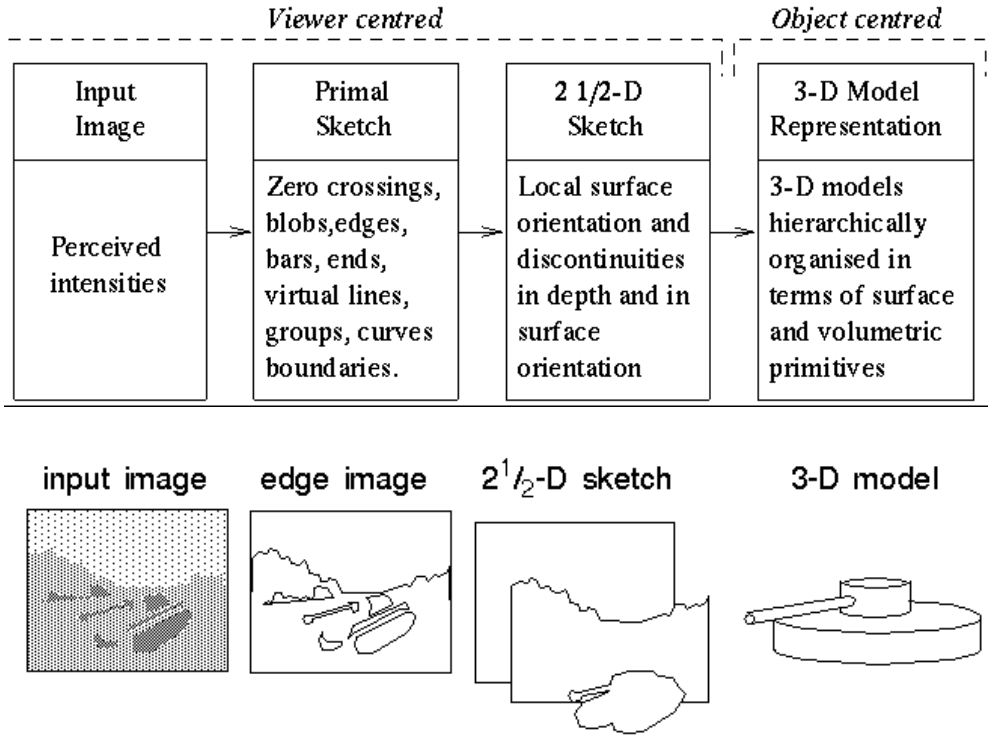
Computed 3D model rendered from new viewpoint

Larry Roberts PhD Thesis, MIT, 1963,
Machine Perception of Three-Dimensional Solids

Slide credit: Steve Seitz

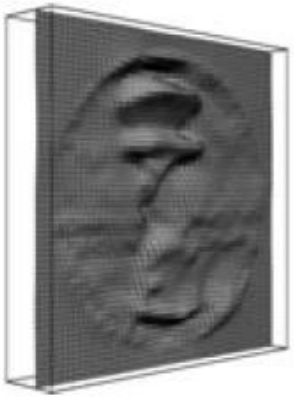
1960s--1970s


VISION
David Marr
FOREWORD BY
Shimon Ullman
AFTERWORD BY
Tomaso Poggio

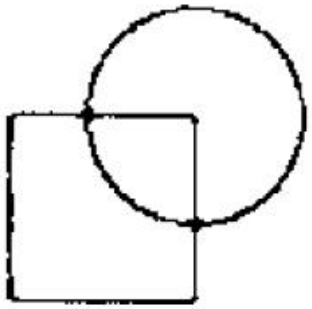

Viewer centred
Object centred
Input Image
Primal Sketch
2 1/2-D Sketch
3-D Model Representation
Perceived intensities
Zero crossings, blobs,edges, bars, ends, virtual lines, groups, curves boundaries.
Local surface orientation and discontinuities in depth and in surface orientation
3-D models hierarchically organised in terms of surface and volumetric primitives
input image
edge image
2 1/2-D sketch
3-D model

1980s: sophisticated mathematical techniques for performing quantitative image and scene recognition

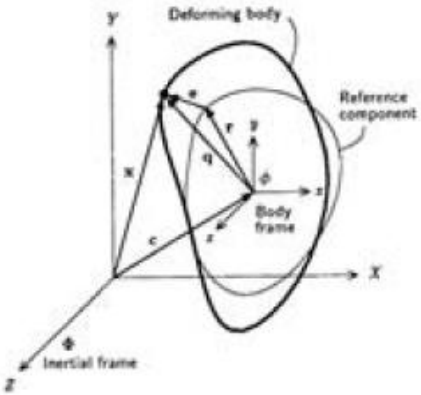
(a)



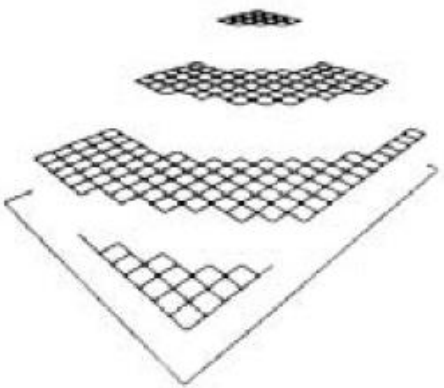
(b)



(c)



(d)



(e)

(f)

Image credit: Rick Szeliski

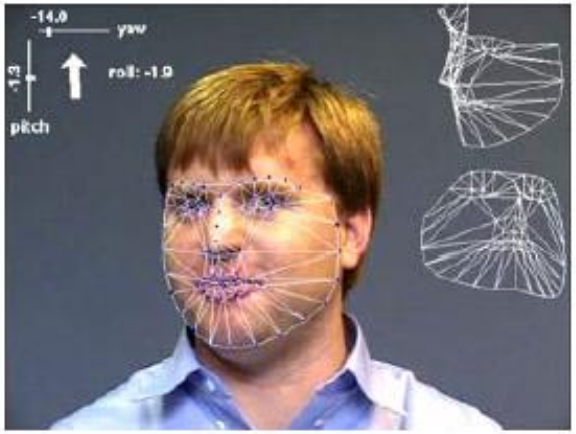
Image credit: Rick Szeliski

1990s

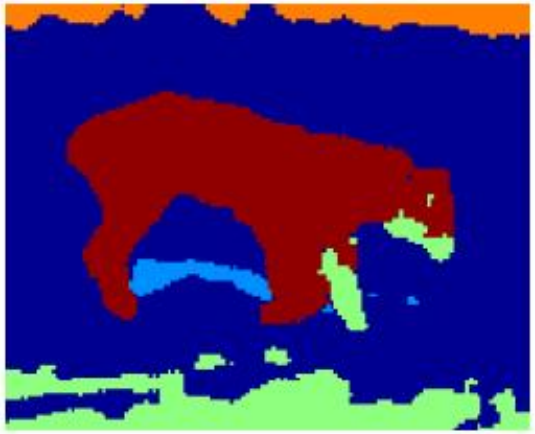
(a)

(b)

(c)



(d)



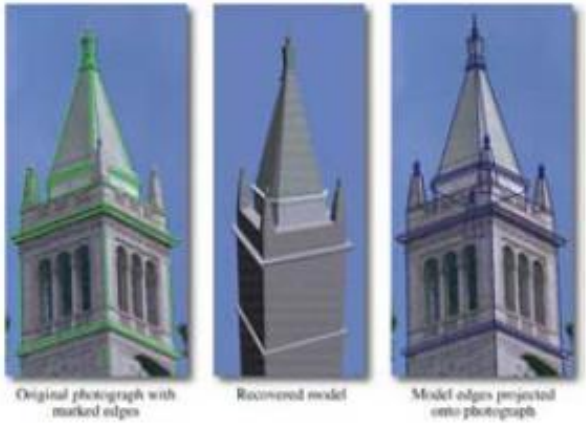
(e)

(f)

Image credit: Rick Szeliski

2000s: Vision+graphics, feature based technique, global optimization

(a)



(b)

(c)

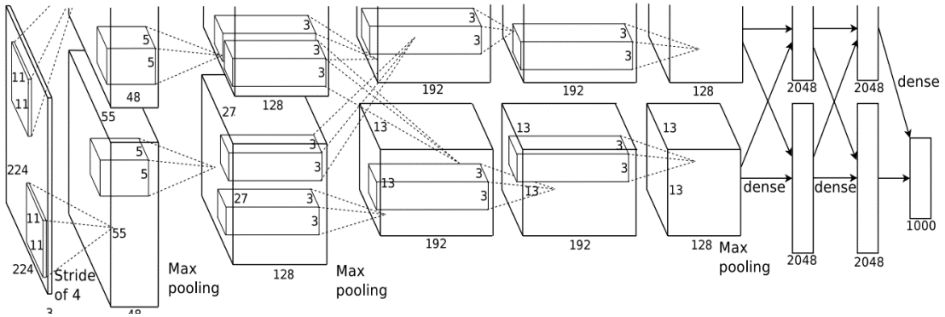
(d)

(e)

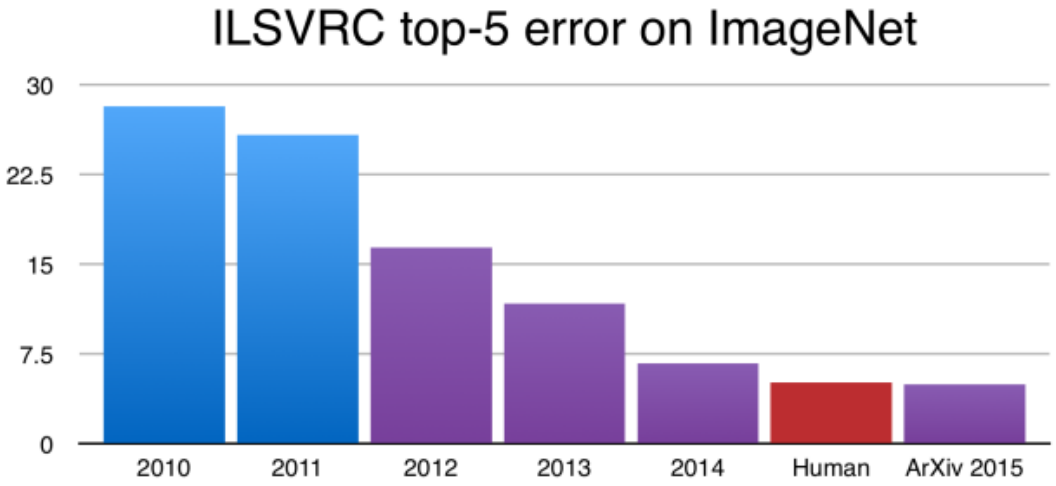
(f)

Feature Based Recognition

2010s: Deep learning



[AlexNet NIPS 2012]



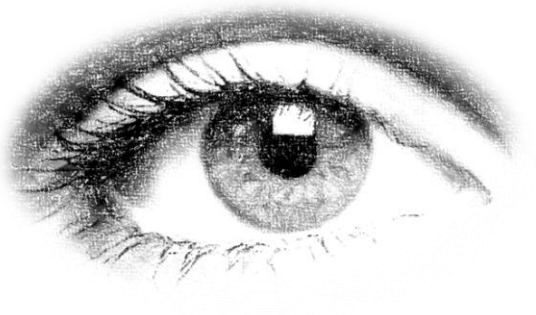
Not deep-learning

Deep-learning based

Source:
https://devblogs.nvidia.com/parallelforall/mocha-jl-deep-learning-julia/
https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/

Dark Side of Deep Learning Based Method?

- Training data
- Robustness to adversarial settings
- Computations (and bandwidth)
- …



Human Vision System

H. R. Wu and K. R. Rao, "Chapter 2: Fundamentals of Human Vision and Vision Modeling," *Digital Video Image Quality and Perceptual Coding*, CRC Press, 2006.

Some references from: http://www.hf.faa.gov/Webtraining/VisualDisplays/HumanVisSys2a.htm

Outline

- Human vision system
- Color vision
- Luminance and the perception of light intensity
- Spatial vision and contrast sensitivity
- Temporal vision

Outline

- Human vision system
- Color vision
- Luminance and the perception of light intensity
- Spatial vision and contrast sensitivity
- Temporal vision

Human Visual Spectrum

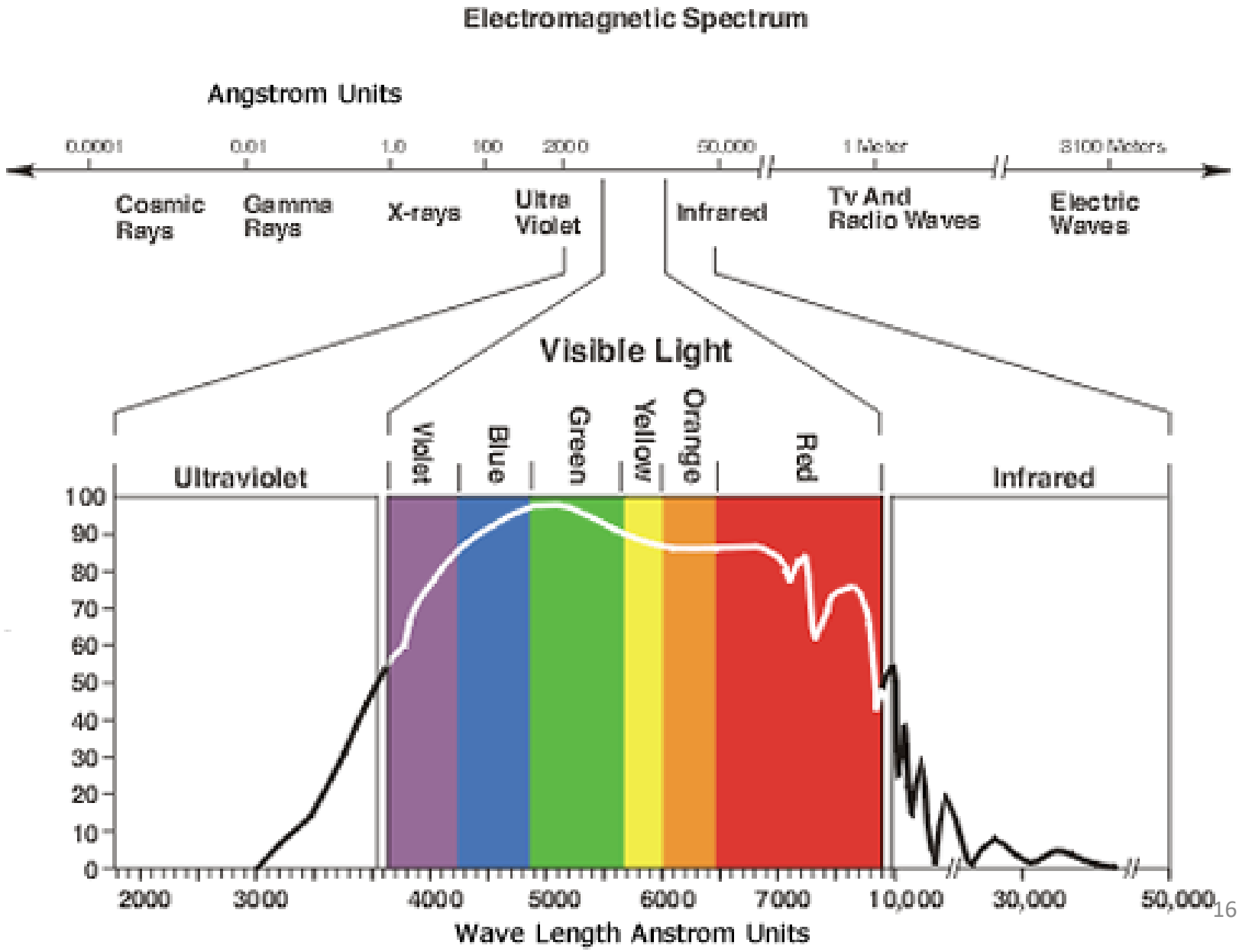
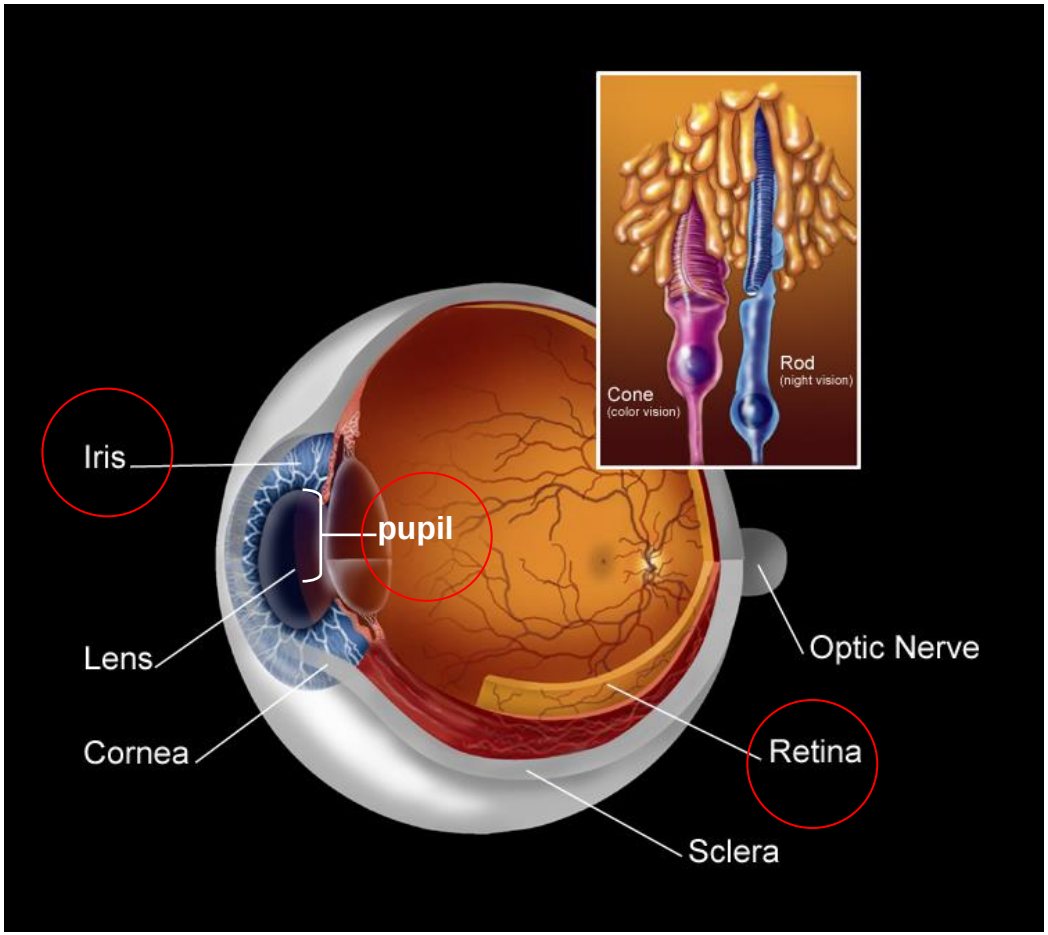
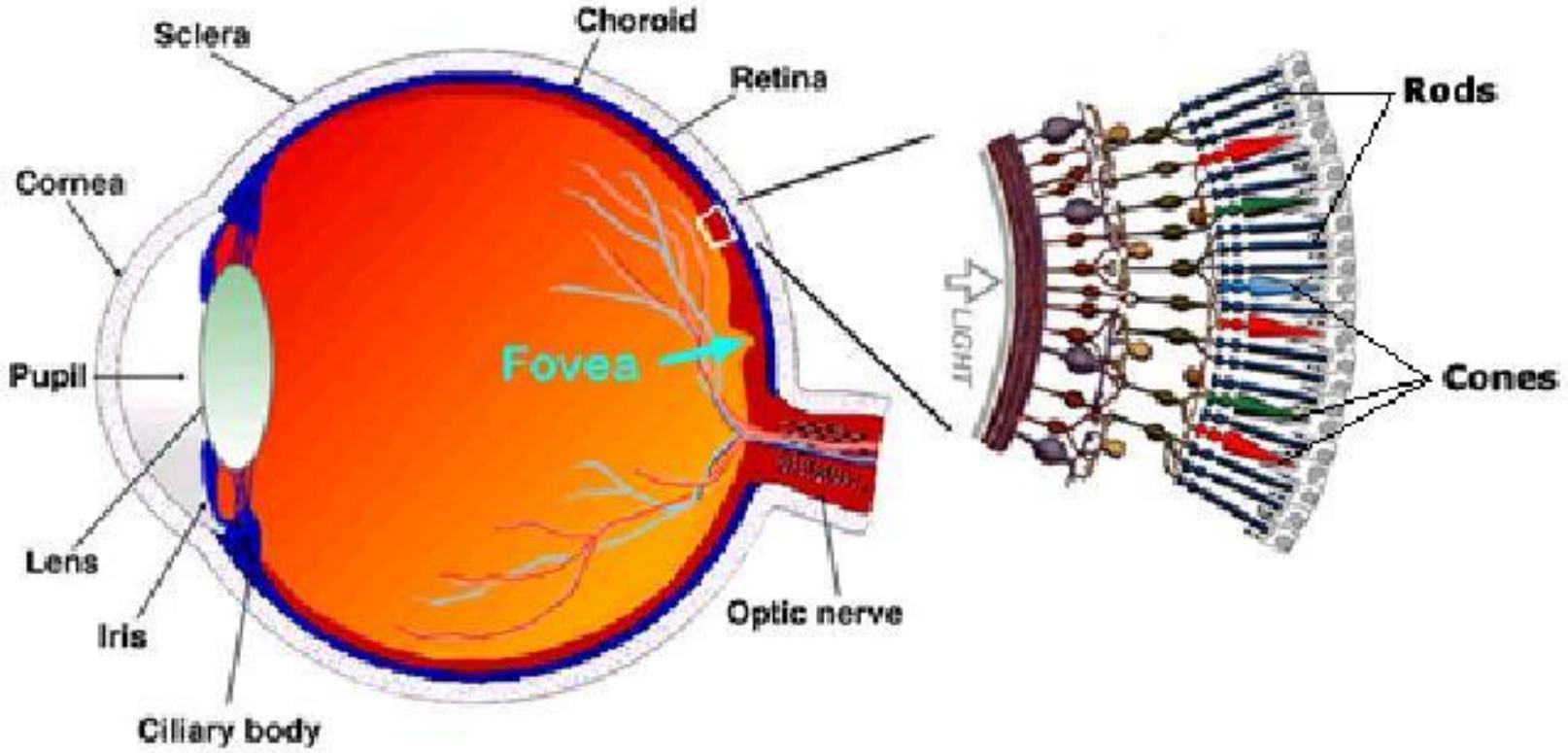


Image Formation at Human Eye

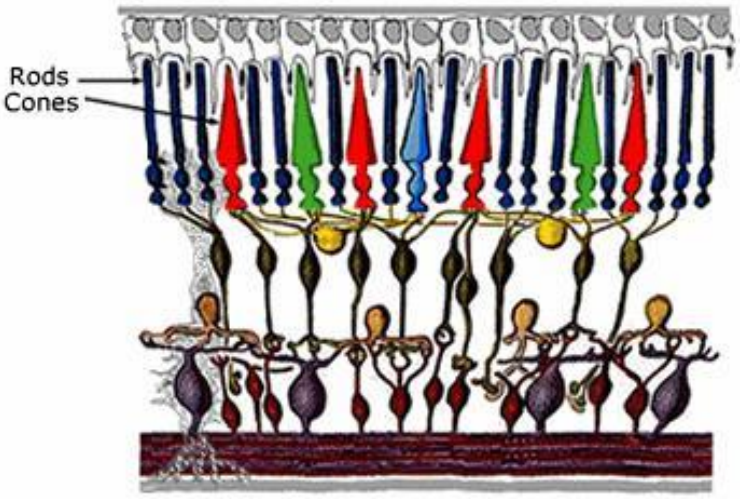




- The retina of each eye has approximately **126 million receptor cells: 120 million rods and 6 million cones**. Rods are responsible for our night vision and cones for our day and color vision.
- Rods are extraordinarily sensitive to light and can respond to a single photon, the smallest quantity of light.
- There are 0.8 million nerve fibers to transfer nerve impulse.

Rods and Cones

- Cones are responsible for our color vision and respond in moderate to bright light to what we perceive as red, green, and blue.
- These two systems contribute to the extreme range of light intensity levels that humans can perceive. This range, from one photon to glare tolerance limit, is in the order of **$1:10^{16}$**.
- There are slightly more red receptors in the eye than green and **very few blue receptors** compared to red and green. The ratio is 1 blue to 14 red and green in the peripheral retina, 1 to 20 in the fovea, and none in the foveal pit.



Rods

- Rods - provide "scotopic" or low intensity vision.
  - Provide our night vision ability for very low illumination
  - Are a thousand times more sensitive to light than cones
  - Are much slower to respond to light than cones
  - Are distributed primarily in the periphery of the visual field

Cones

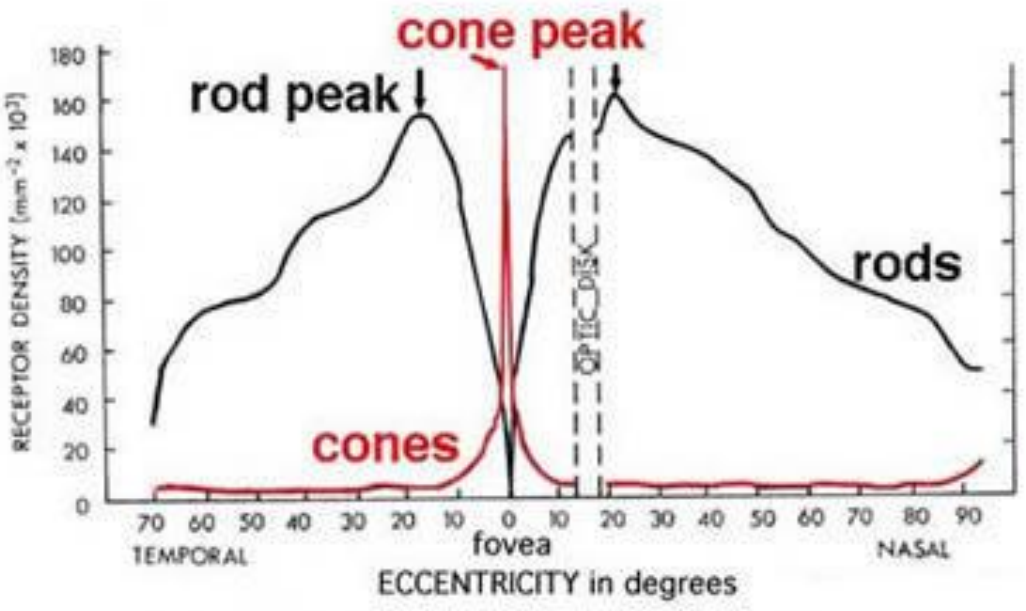
- Cones - provide "photopic" or high acuity vision.
  - Provide our day vision,
  - Produce high resolution images,
  - Determine overall brightness or darkness of images,
  - Provide our color vision, by means of three types of cones:
    - "L" or red, long wavelength sensitive,
    - "M" or green, medium wavelength sensitive,
    - "S" or blue, short wavelength sensitive.
  - Since cones do not function in very low light, we have no color vision at night or in other very low illumination environments.

Rod and Cone Densities

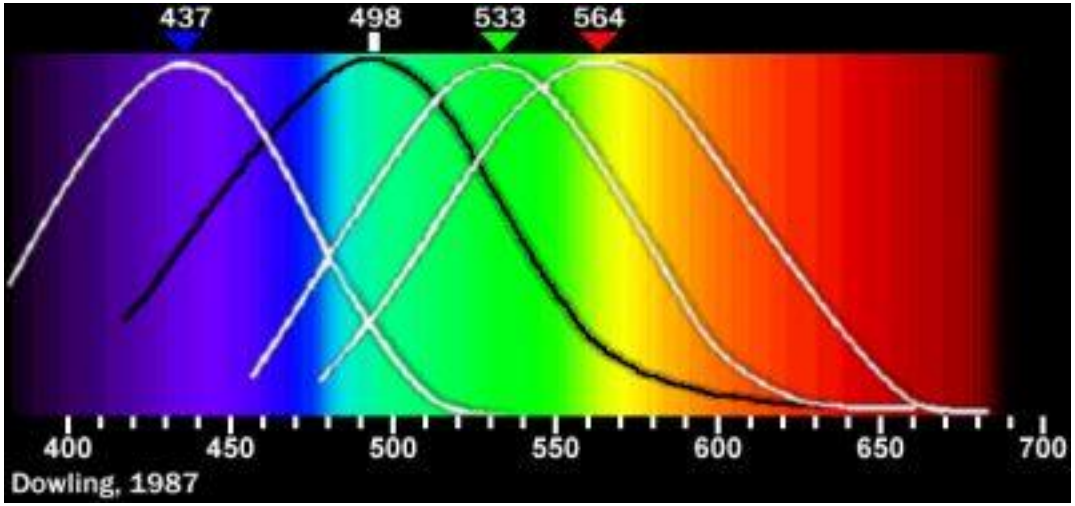
- **Only cones are present in the fovea**. Cone density decreases rapidly outside the fovea and then falls to a fairly even density in the peripheral retina.
- **The highest density of "M" and "L" cones, but the lowest density of "S" cones are found in the fovea**. S-cones form only 3-5% of the cones in the fovea. There are no "S" cones in the center of the fovea, the fovea centralis, where visual acuity is highest. The maximum density of "S" cones,15%, is found 1 degree from the fovea. The remainder are dispersed unevenly throughout the retina where they make up 8% of the cones.
- The small number of cones sensitive to blue, the "S" cones, in the fovea and in the rest of the retina is why pure blue should not be used for small text, lines, or symbols. **Blue also has little perceived brightness which is why it should not be used against black or a darker shade of blue as background, like this.**

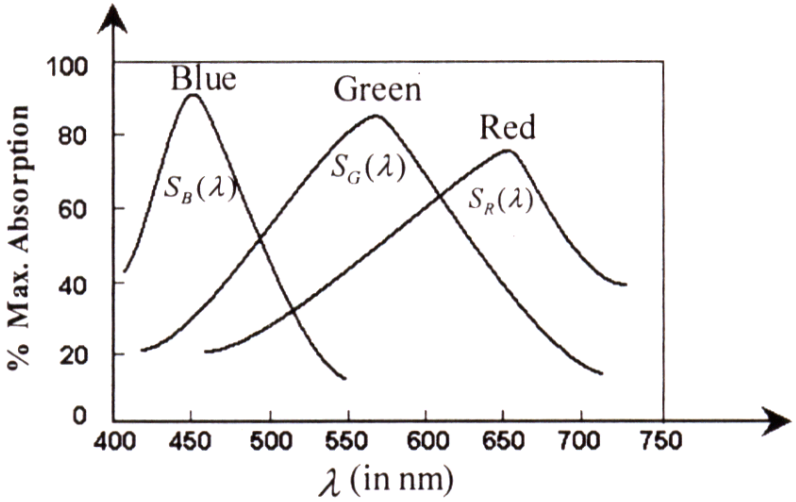
Rod and Cone Densities

- This figure shows the relative density of rods and cones in the retina. There are no rods in the fovea where vision is most acute. Rods are most dense in the periphery of the retina; cones are most dense in the fovea.



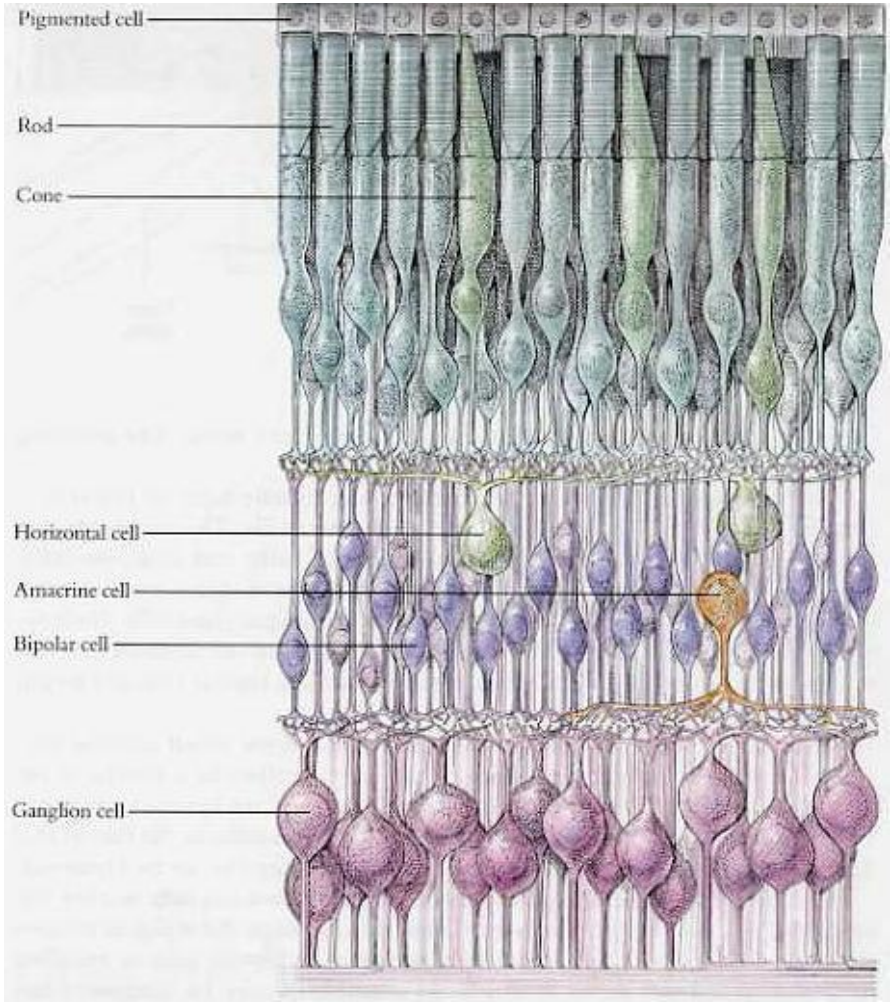
Wavelength Sensitivities of Cones and Rods





- This diagram shows the wavelength sensitivities of the different cones and the rods
  - Note the overlap in sensitivity between the green and red cone.
- Rod Sensitivity- Peak at 498 nm.
- Cone Sensitivity
  - Red or "L" cones peak at 564 nm.
  - Green or "M" cones peak at 533 nm.
  - Blue or "S" cones peak at 437 nm.

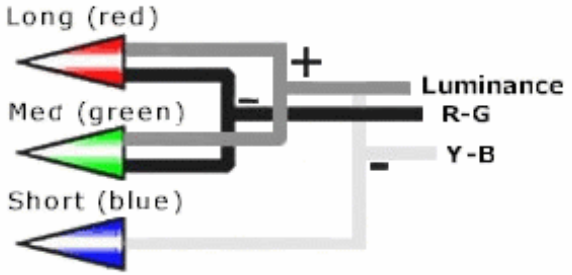
Retina



- Bipolar cell
- Ganglion cell
- Horizontal cell
- Amacrine cell
- Many Rod → one Bipolar, many Bipolar → one Amacrine, then to Ganglion
- 120 million rods and 7 million cones
- 1 million ganglion cell
- 127:1 compression!

Cone Cell Output

- human visual system detects color by comparison between the output of a minimum of two similar receptors.
- These pathways are:
  - L + M = Luminance, achromatic
  - L - M = Red/Green, chromatic
  - S – (L+M) = Blue/Yellow, chromatic

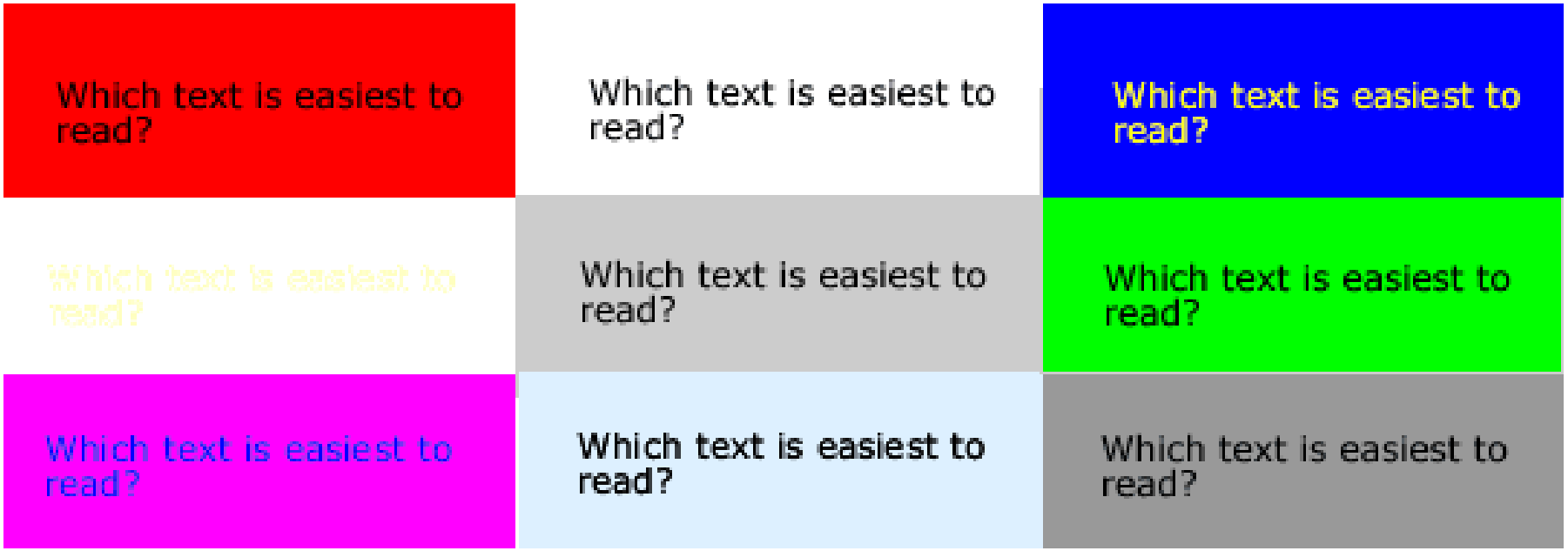


- The signals from the L - cones (red) is summated with the signals from the M-cones (green) to form the luminance channel.
- The red/green color channel results from the subtraction of the M-cone signals from the L-cone signals.
- The blue/yellow channel is derived from the subtraction of the summation of the L-cone and M-cone signals from the S-cone signals (blue)

Color Difference

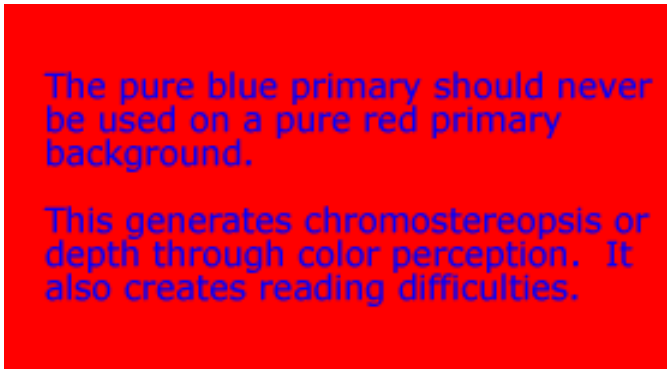
- It is generally thought that S - cones do not contribute to the luminance channel
- Short-wavelength light (blue) adds chromatic information, but does not affect brightness
- This is why yellow characters on a white background on an electronic display are difficult to distinguish. The yellow and white differ only in that the white contains blue and the yellow does not. The blue in the white has low luminance, therefore the yellow text has little luminance contrast when compared with the white background.

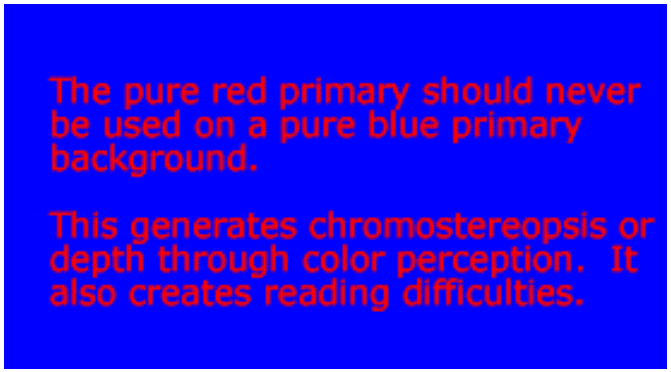
Visual System - Luminance & Chrominance



Color Difference

- Thirty percent (30%) of viewers will perceive blue as closer, and ten percent (10%) will perceive both red and blue as being in the same plane.
- This example illustrates why red and blue primaries should never be used one on the other in displays.



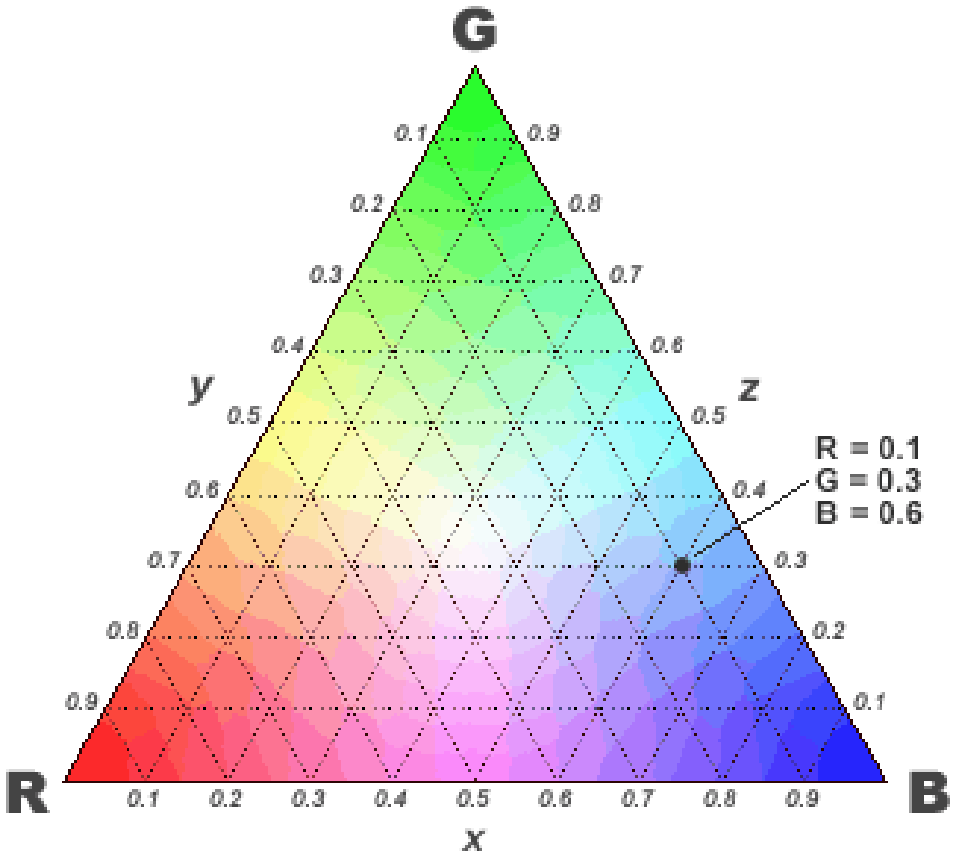


Outline

- Human vision system
- Color vision
- Luminance and the perception of light intensity
- Spatial vision and contrast sensitivity
- Temporal vision

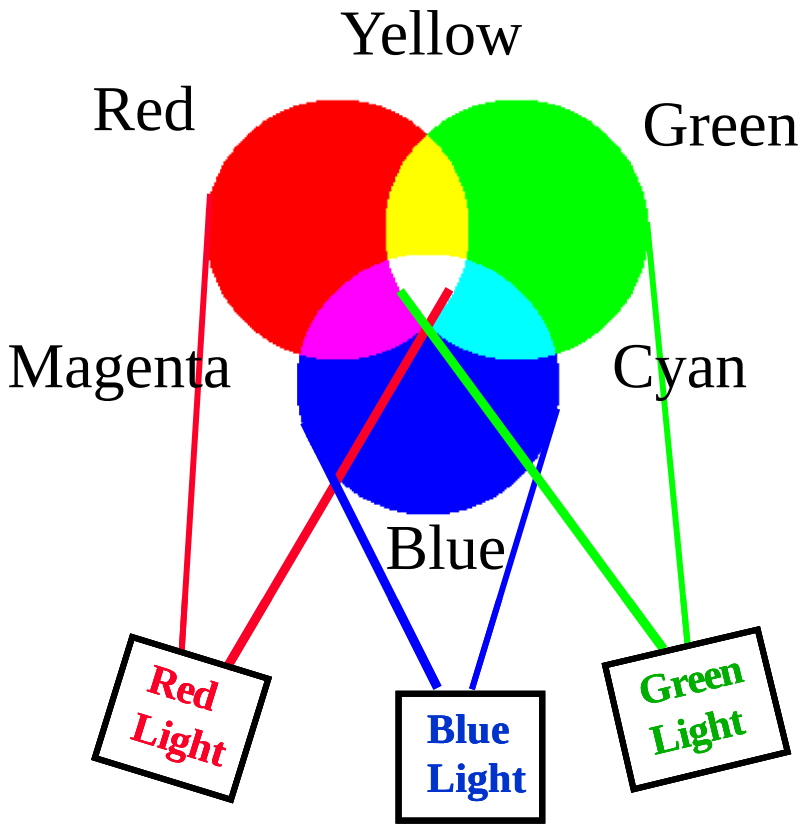
Maxwell Diagram

- Color matching



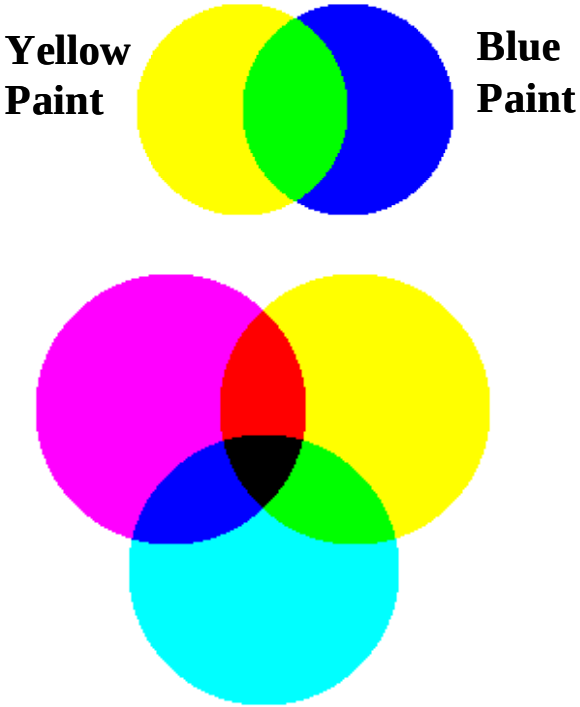
Target Color = 0.1R+0.3G+0.6B

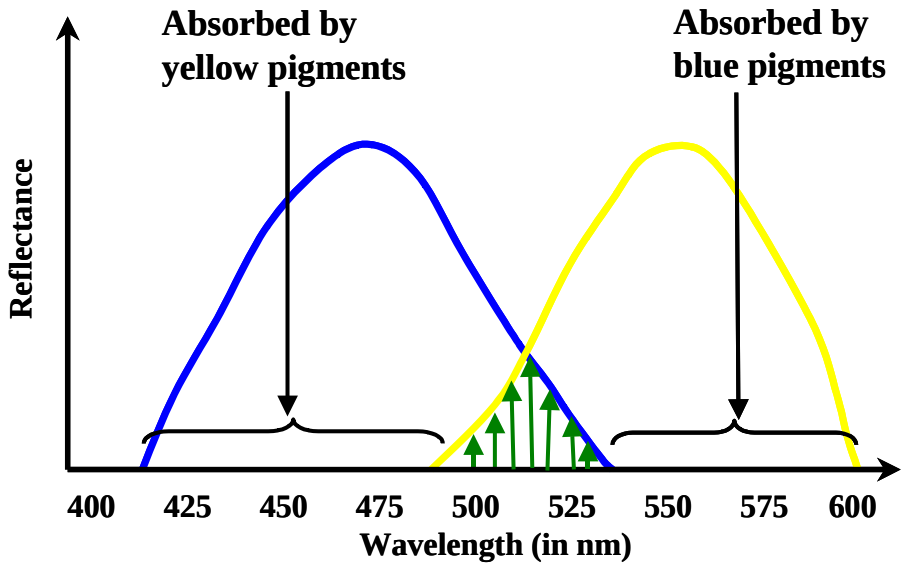
Additive Color Matching



- Primary colors can be added to obtain different composite colors

Subtractive Color Matching

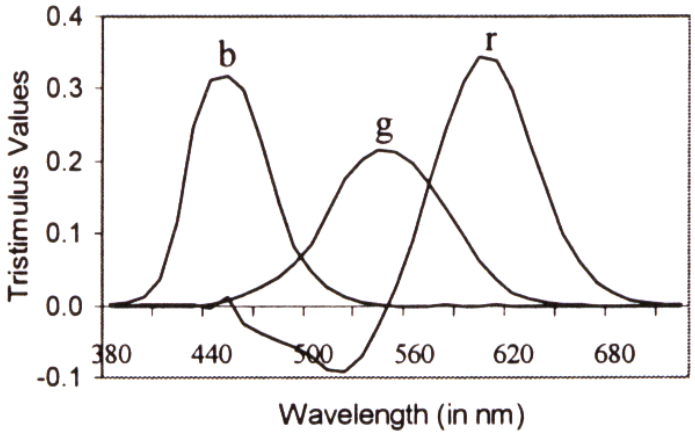




- Subtractive color mixing. a) mixture of yellow and blue paint produces green color, b) composite color is the difference between two added colors, c) mixture of cyan, magenta, and yellow colors.

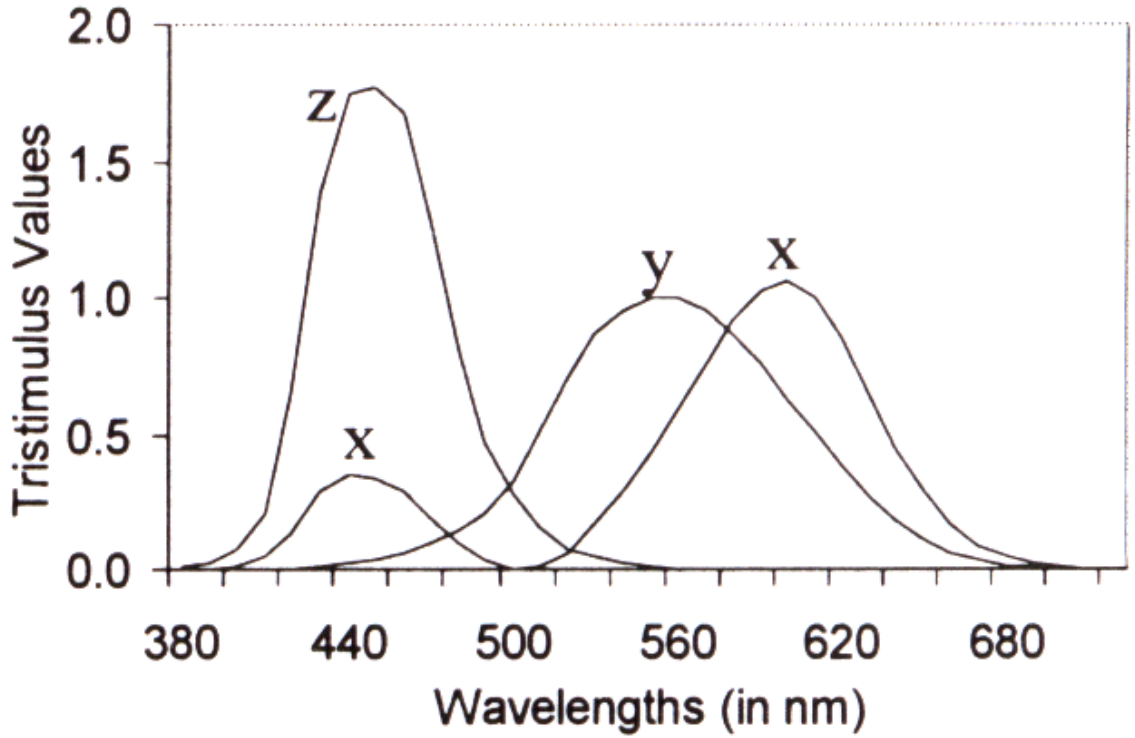
Tristimulus Values

- The primary sources recommended by CIE are three monochromatic colors at wavelength 700nm (red), 546.1nm (green), and 435.8nm (blue).
- Let the amount of the k-th primary needed to produces a color C, and reference white color be denoted by α and β, respectively.
- α/β is called the tristimulus value of color C.



This figure shows the necessary amounts of three primaries to match all the wavelengths of the visible spectrum.

CIE {X, Y, Z} system



$$X = \int_{\lambda} S_{\lambda} R_{\lambda} \bar{x}(\lambda) d\lambda$$

$$Y = \int_{\lambda} S_{\lambda} R_{\lambda} \bar{y}(\lambda) d\lambda$$

$$Z = \int_{\lambda} S_{\lambda} R_{\lambda} \bar{z}(\lambda) d\lambda$$

Keep all tristimulus values all positive.
X, Y, and Z roughly correspond to supersaturated red, green, and blue, respectively.

Chromaticity Diagram

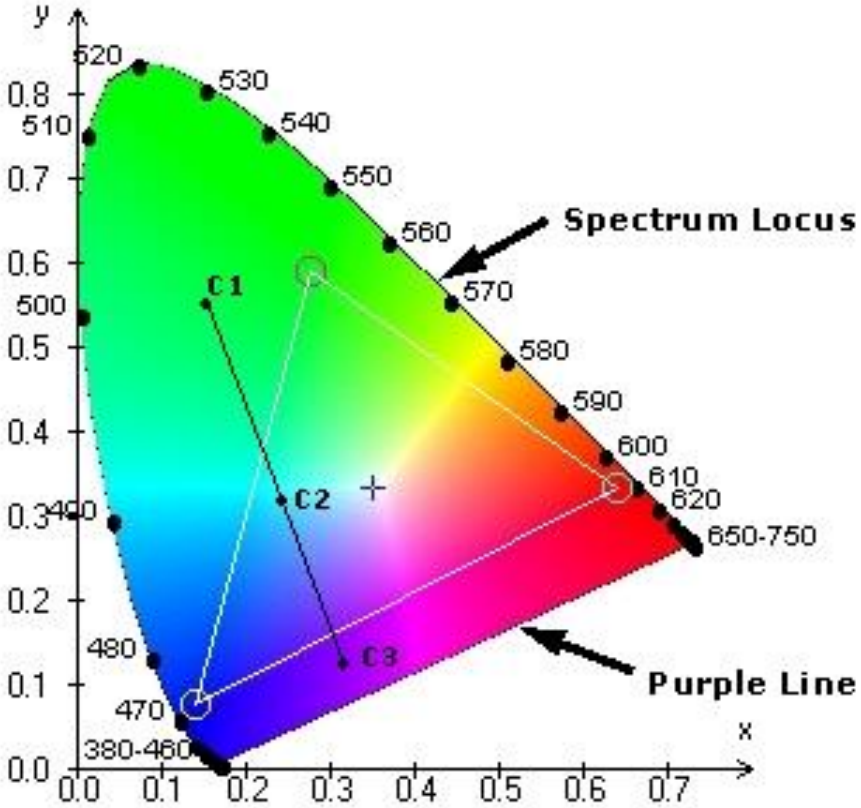
- Any color can be defined by its Tristimulus values (X, Y, Z) or chromaticity coordinates (x, y, z).
- White lies at or near the middle of the enclosed figure.
- Mathematical conversions are used to convert between x,y,z values and XYZ.

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

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

$$z = \frac{Z}{X+Y+Z} = 1 - x - y$$

x,y,z represents the proportions of the X primary, Y primary, and Z primary respectively in a given color mixture.

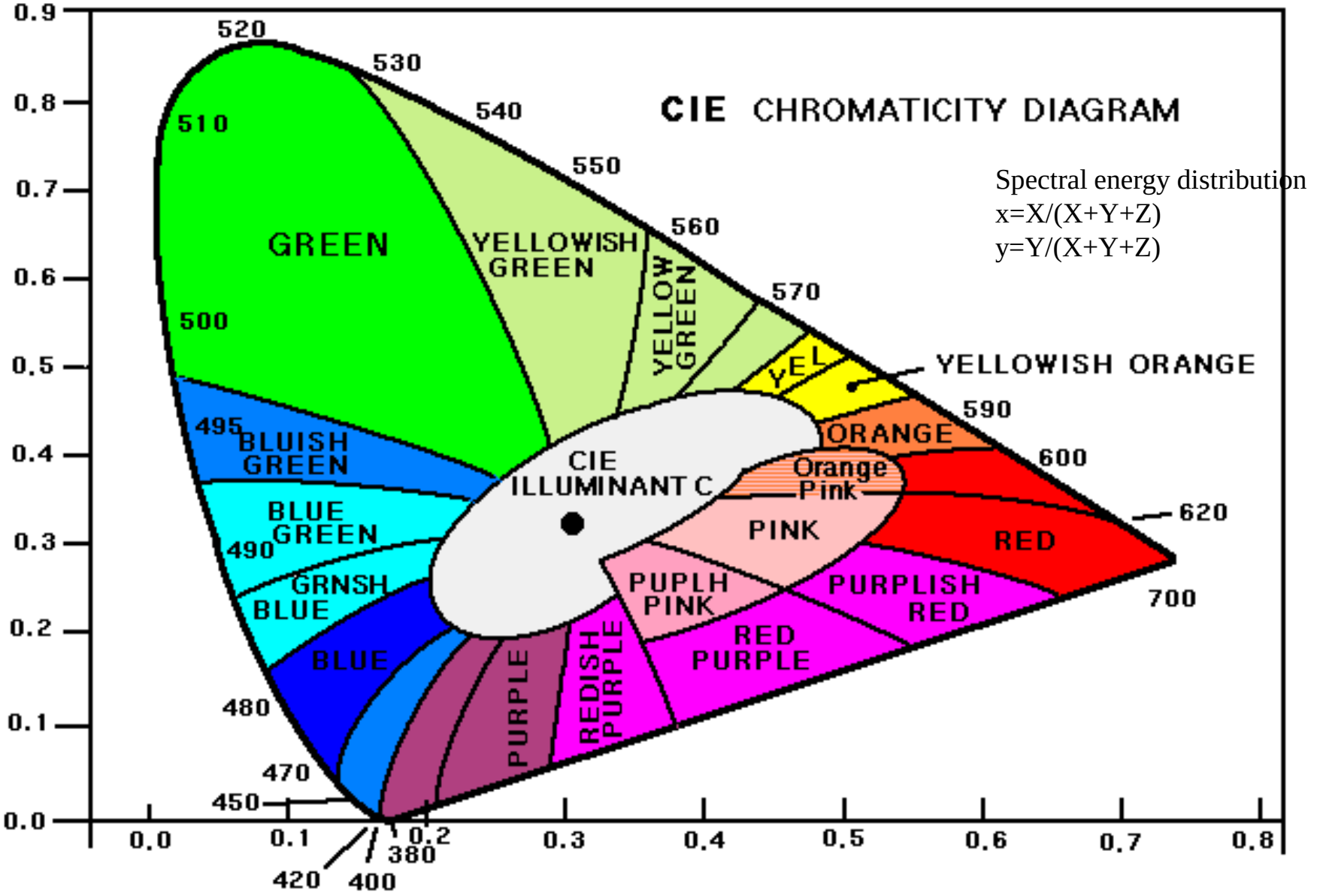


- In 1931, the Commission Internationale de l'Eclairage (CIE, International Commission on Illumination) developed a light measurement standard. The CIE conducted extensive color matching experiments with colored lights to develop a system based on human color perception (red, green, blue).
- The 1931 standard is known as CIE XYZ. Colors in the XYZ color space are specified by projection onto a two dimensional plane.
- All colors that can be humanly perceived can be plotted within this space.

CIE Chromaticity Diagrams

- CIE chromaticity diagrams graphically present useful electronic display information with regards to color and luminance.
- The horseshoe shaped line is termed the "spectrum locus."
- All pure spectral, visible wavelengths lie on this line. The line which closes the horseshoe shape is termed the "purple line."
- All pure purples lie on this line.
- All colors that can be humanly perceived lie within this shape.
- A color plotted closer to the center is more desaturated, that is, contains more white. All colors that can be created by mixing two colors lie along a line between those colors.

Commission Internationale de L'Eclairage (CIE) Diagram (1931)



International Commission on Illumination

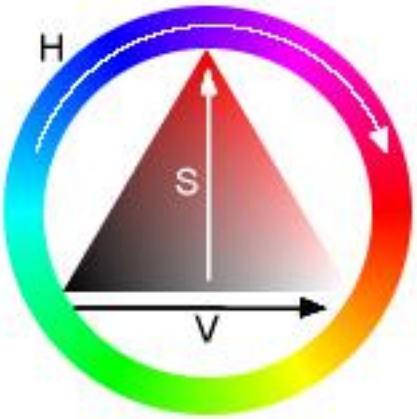
Color Appearance

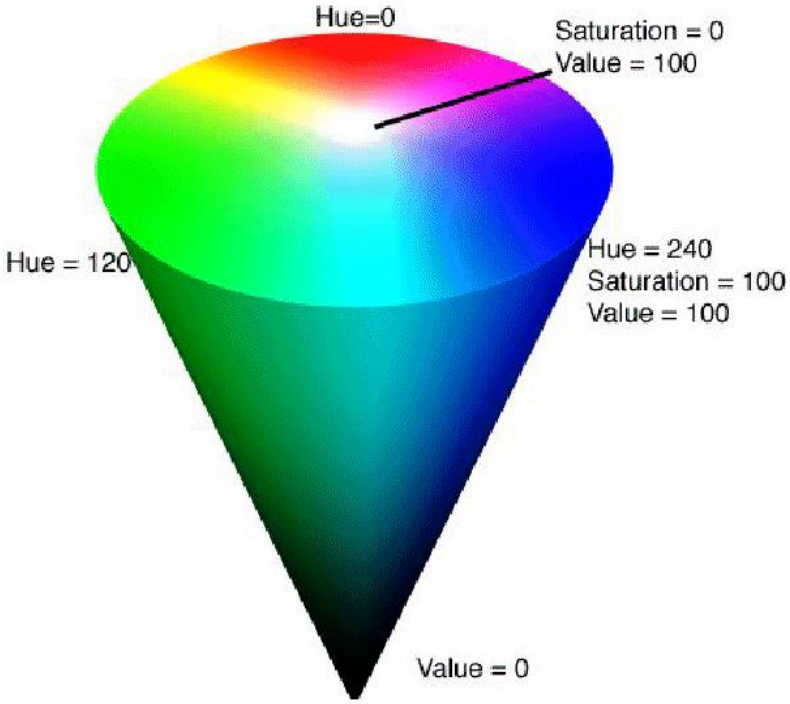
- Five perceptual attributes
- **Brightness:** the attribute according to which an area appears to more or less intense
- **Lightness:** the brightness of an area relative to a similarly illuminated area that appears to be white
- **Colorfulness (Chromaticness):** the attribute according to which an area appears to be more or less chromatic
- **Chroma:** the colorfulness of an area relative to a similarly illuminated area that appears to be white
- **Hue:** the attribute of a color denoted by its name such as blue, green, yellow, orange, etc.

- Increasing the illumination increases the brightness and colorfulness of a stimulus while the lightness and chroma remain approximately constant

Color Representation

- There are three main perceptual attributes of colors: *brightness (V), hue(H), and saturation(S)*.





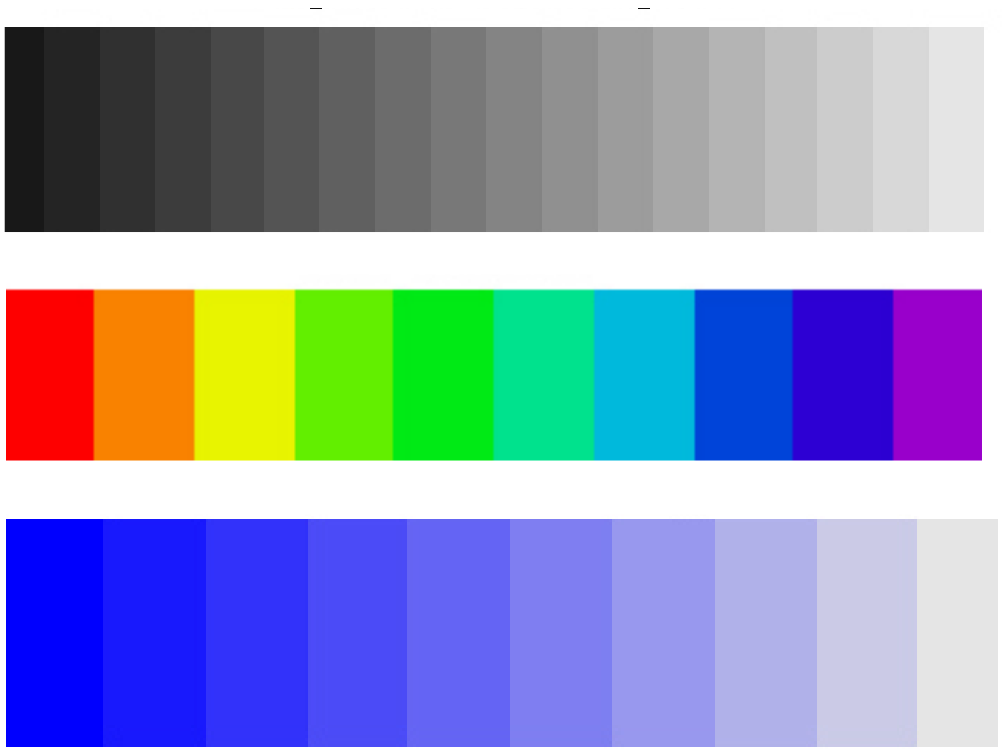
The HSV Cone

Brightness is the perceived luminance.

Hue is an attribute we commonly describe as blue, red, green, etc.

Saturation is our impression how different the color is from achromatic (white or gray) color.

Color Attributes

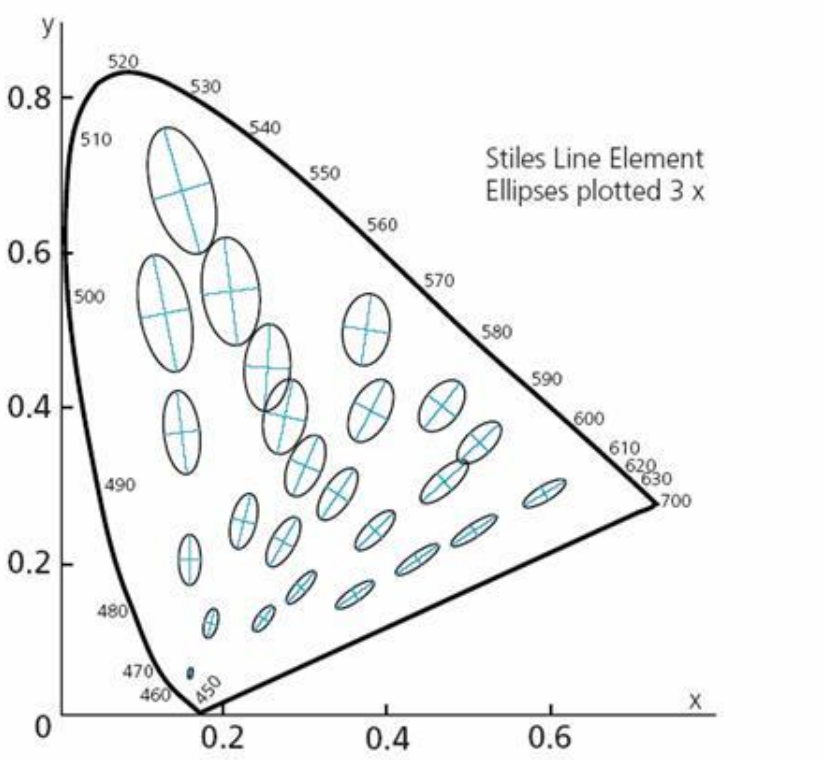


- Example of perceptual attributes of color. a) different brightness levels (dark to bright), b) different hues (red to violet), and c) saturation (the dark blue at the left side is highly saturated whereas the faded blue at the right side has low saturation).

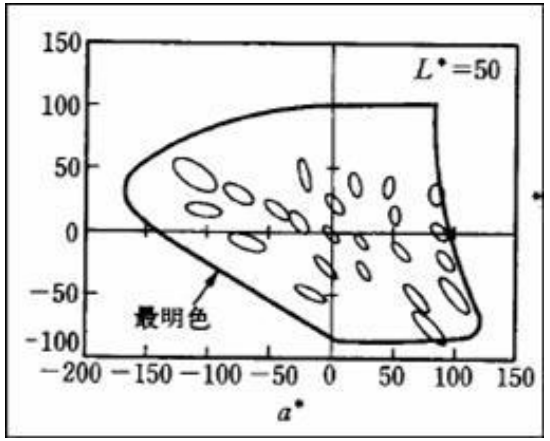
CIE XYZ and L*U*V* Color Spaces

- The problem was that equal geometric steps in CIE XYZ did not correspond to equal perception steps. Therefore, it was not possible to use the CIE 1931 diagram to determine what colors were the most different from each other perceptually. Ideally, colors selected for a display should be maximally different from each other to avoid color confusion.
- To address this problem, in 1978, the CIE issued a modification, the CIE L*U*V*Color Space. This color space more closely approximates uniform perceptual differences as the geometric distance between two colors.

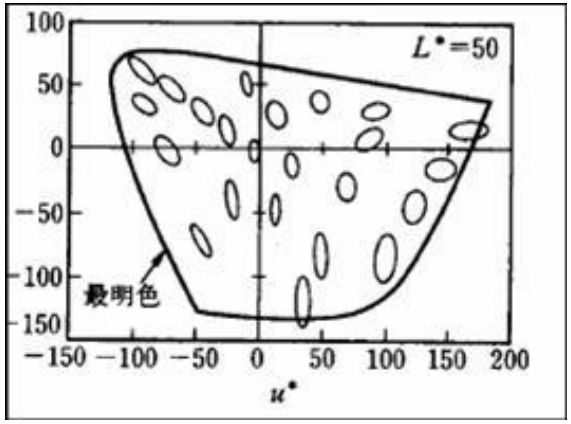
Uniform Color System



CIE XYZ



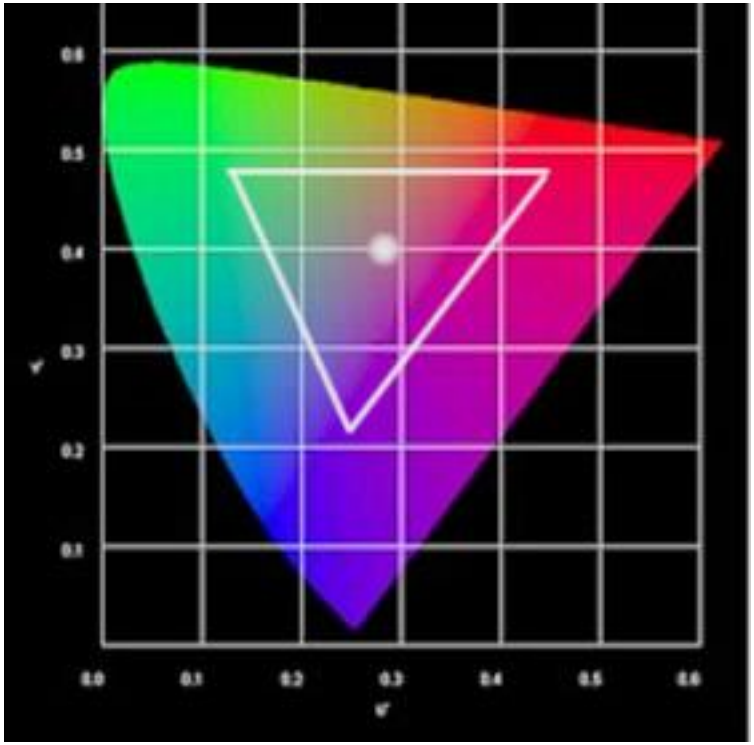
Lab



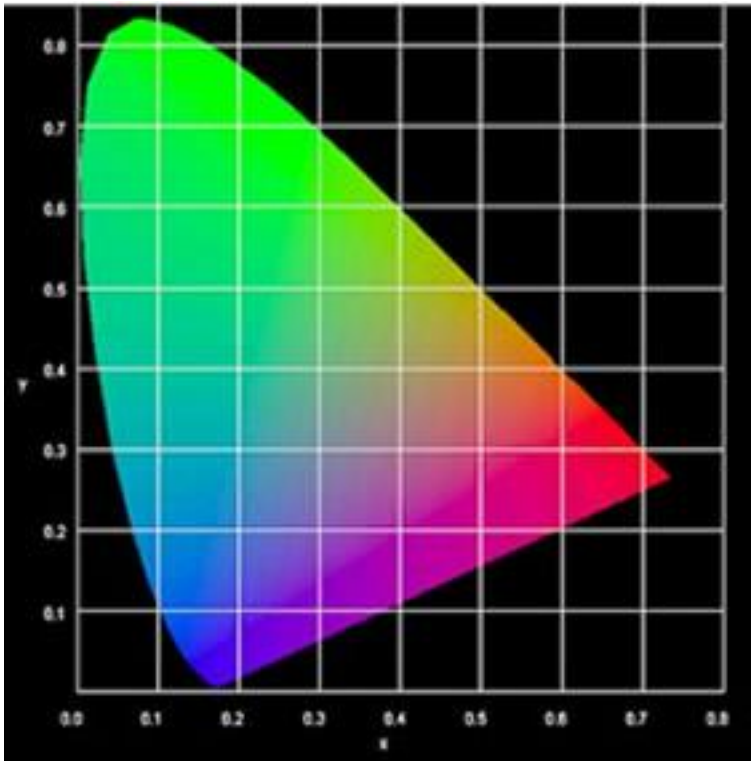
Luv

CIE L*U*V* Color Space

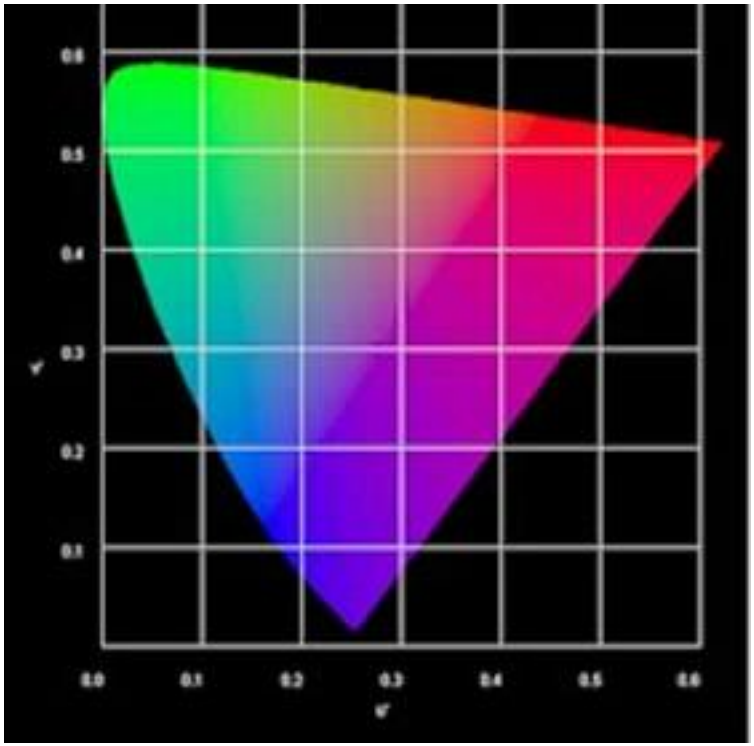
- CIE diagrams are used to determine the "gamut" or range of colors that can be displayed by a particular monitor by plotting the red, green, and blue primaries and joining them with straight lines to form a triangle. All colors that can be produced by the display lie within the triangle.



CIE XYZ and L*U*V* Color Spaces



1931 CIE XYZ Chromaticity Diagram



1978 CIE L*U*V* Chromaticity Diagram

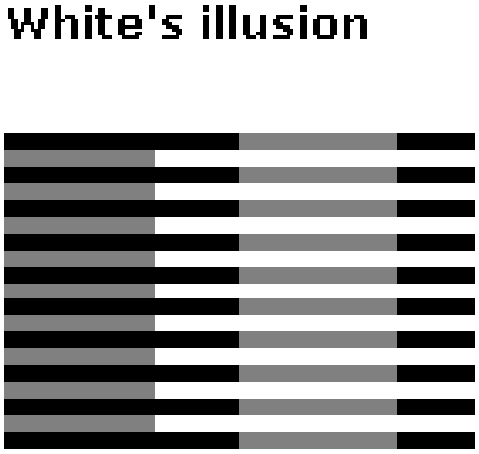
$I = V\cos 33^o - U\sin 33^o \quad Q = V\sin 33^o + U\cos 33^o$

Outline

- Human vision system
- Color vision
- Luminance and the perception of light intensity
- Spatial vision and contrast sensitivity
- Temporal vision

Luminance and the Perception of Light Intensity

- Weber's law
  - The ratio of the increment to the background or adaptation level is a constant that $\Delta I/I=k$
  - ΔI: just noticeable difference, JND
- Fechner's law
  - $S= K \log(I)$
  - S: sensation magnitude; K: constant
- Steven's power law
  - $S=kI^a$

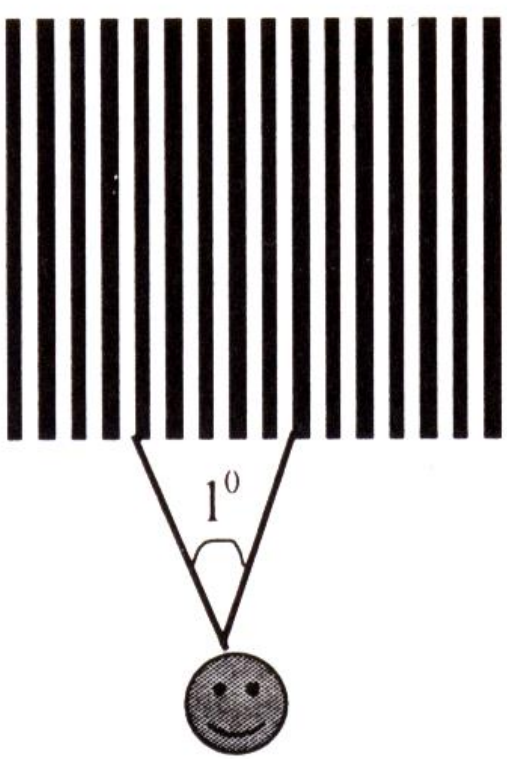


Outline

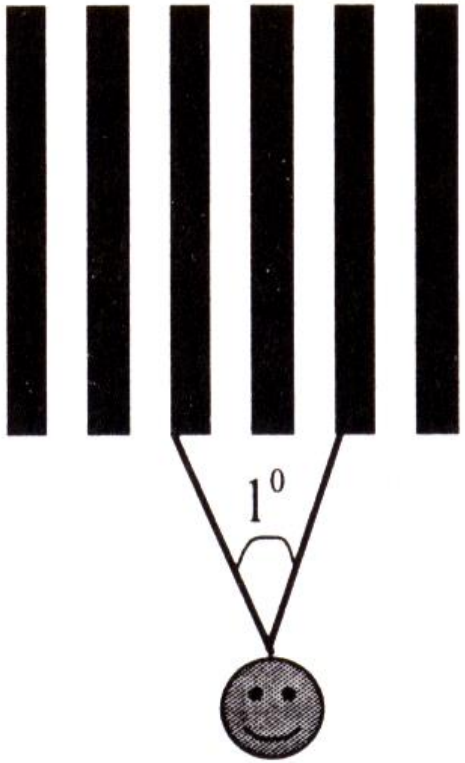
- Human vision system
- Color vision
- Luminance and the perception of light intensity
- Spatial vision and contrast sensitivity
- Temporal vision

Spatial Frequency

- Spatial frequency is generally expressed in



5 cycles/degree



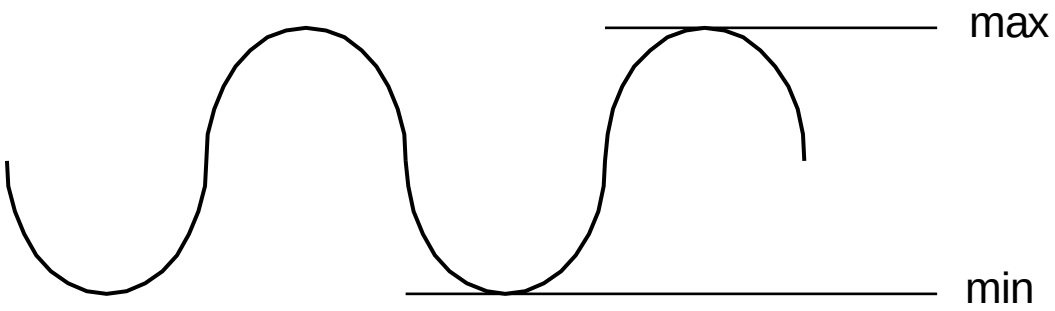
2 cycles/degree

Human eyes' resolution: ~60cpd

Sampling frequency ~120cpd

Modulation Transfer Function (MTF)

- MTF is the spatial frequency response of an imaging system or a component
  - Contrast at a given spatial frequency relative to low frequencies
- Spatial frequency
  - Measured in cycles or line pairs per millimeter (lp/mm)
- Contrast



$$Contrast = \frac{\max - \min}{\max + \min}$$

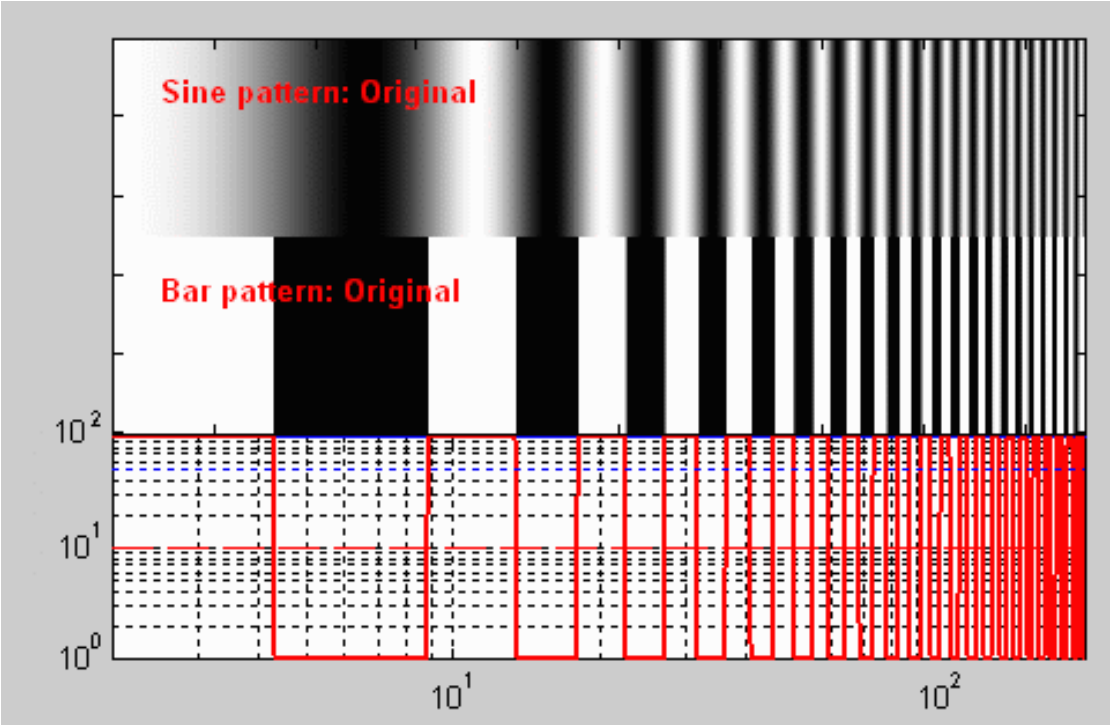
Some references are from:
http://www.normankoren.com/Tutorials/MTF.html

More Precise Definition of MTF

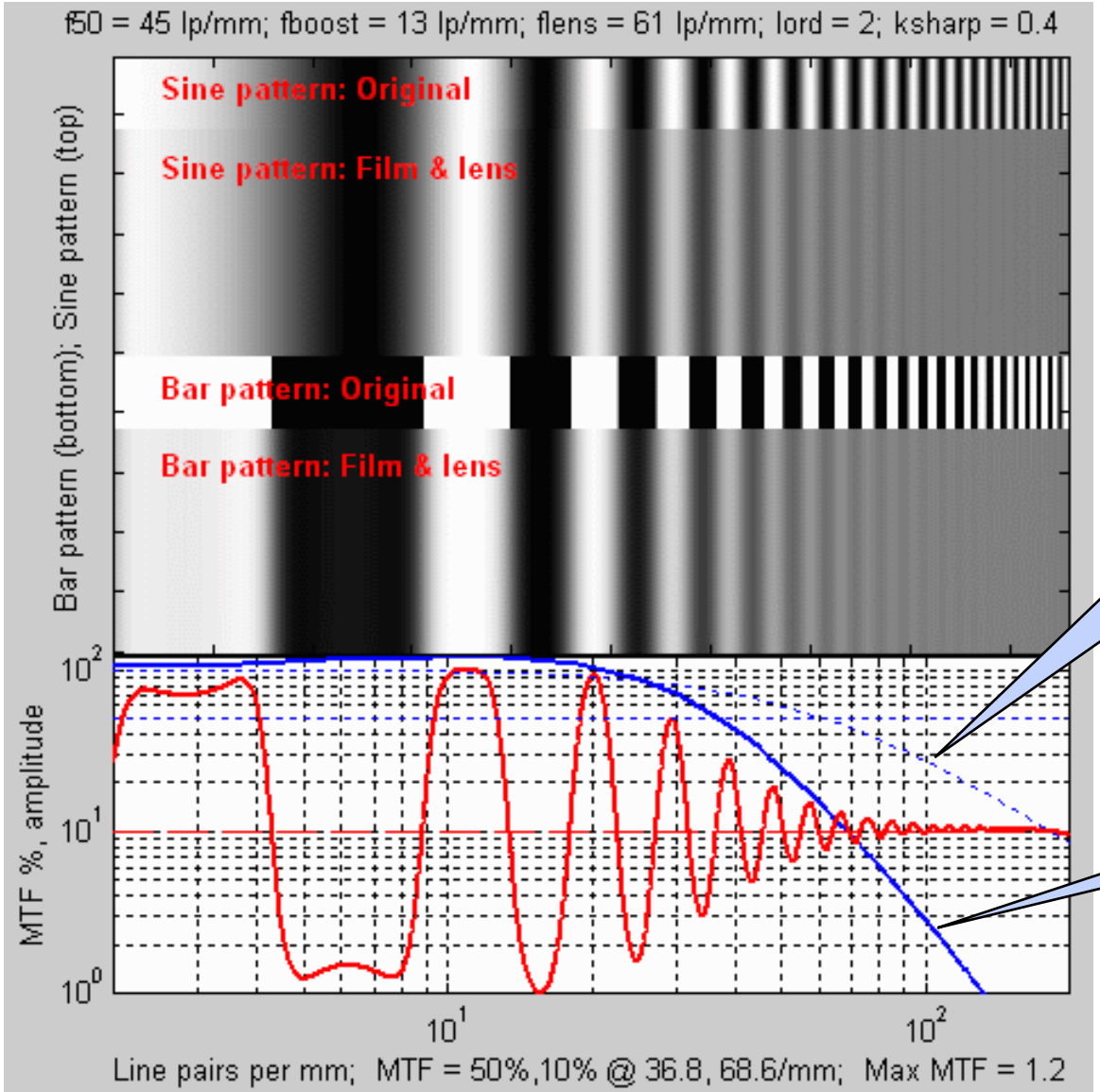
- V_B: the minimum luminance for black areas at low spatial frequencies
- V_W: the maximum luminance for white areas at low spatial frequencies
- V_{min}: the minimum luminance for a pattern near spatial frequency f
- V_{mix}: the maximum luminance for a pattern near spatial frequency f
- $C(0)=(V_W-V_B)/(V_W+V_B)$ is the low frequency contrast
- $C(f)=(V_{max}-V_{min})/(V_{max}+V_{min})$ is the contrast of a given frequency
- $MTF(f)=100\%*C(f)/C(0)$

An Example of MTF (1)

- The target is 0.5mm in length on film
- Lens: Canon 28—70 f/2.8L
- Film: Fuji Velvia
- Input:



An Example of MTF (2)

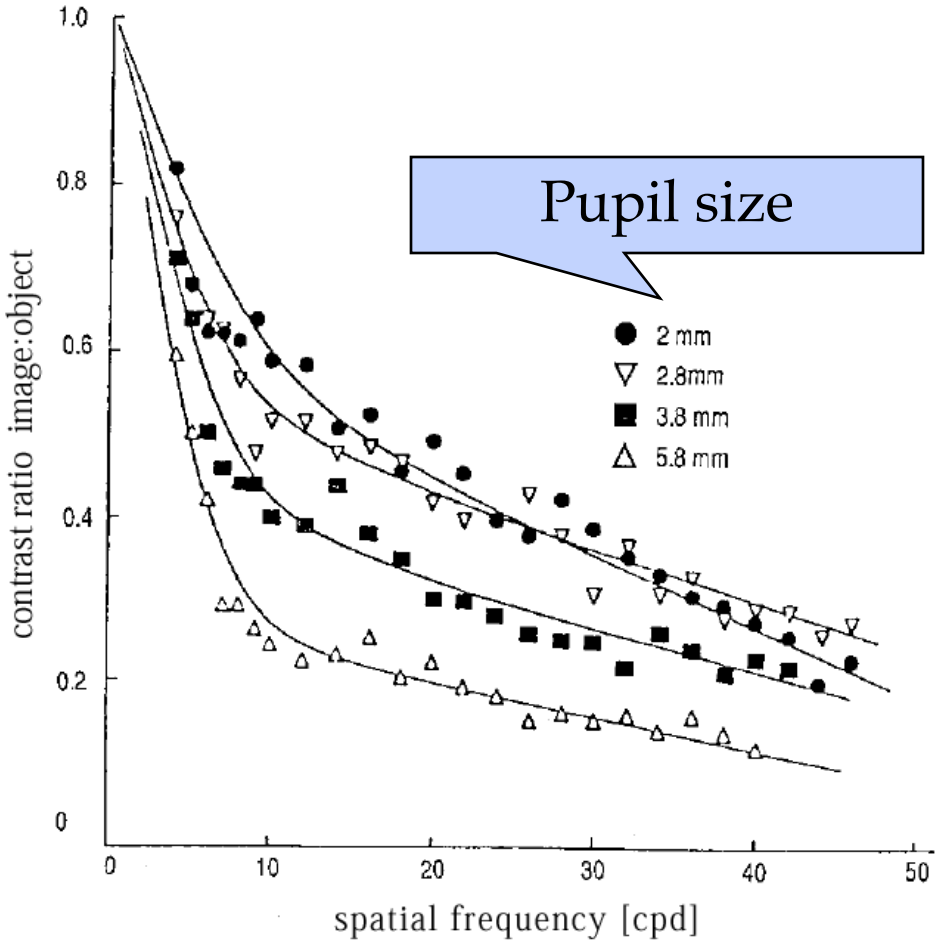


- Output

MTF of the lens

MTF of the lens+film

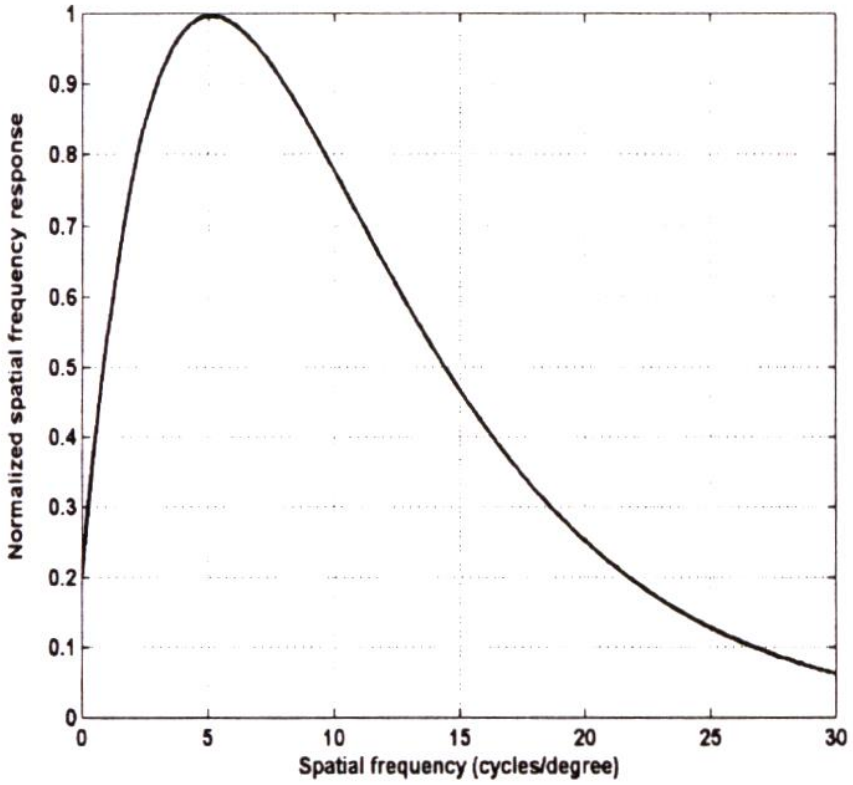
MTF of Human Eyes

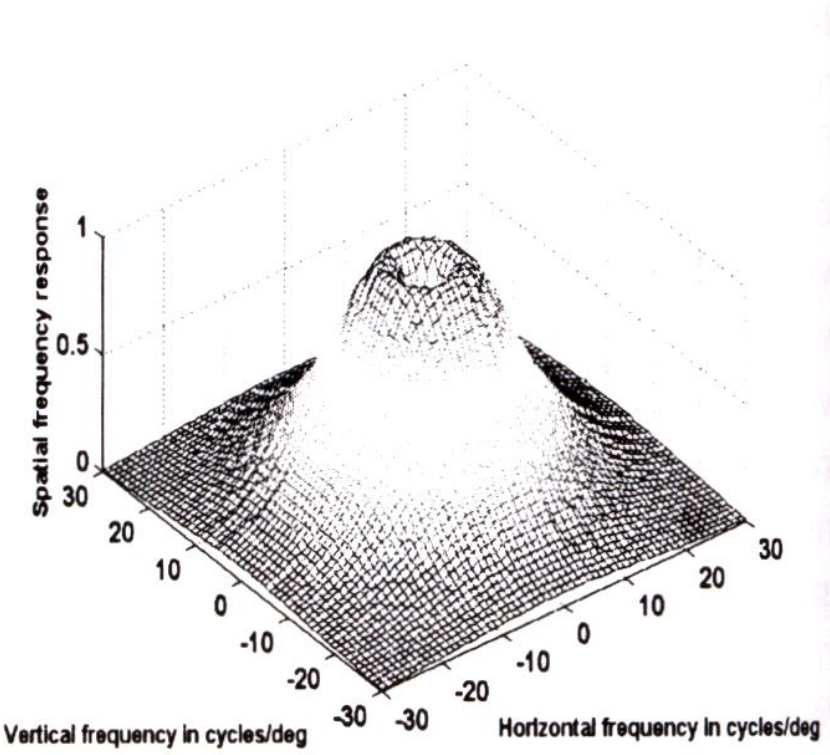


- 20-20 vision
  - Can distinguish patterns with a feature size as small as one minute of an arc
  - 30 cpd
- The MTF of an imaging system is designed following this property
  - For a 35mm imaing system: 55 lp/mm

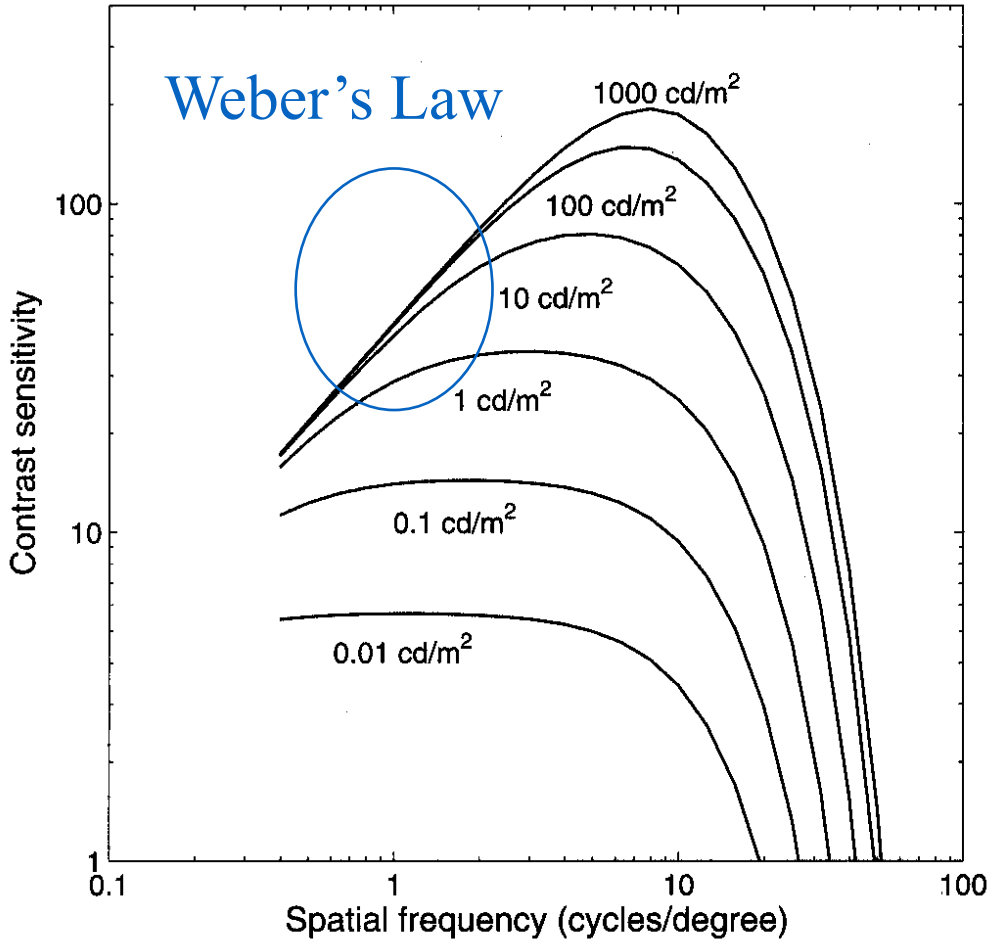
The data is referred to the handout of the course Image and Video Compression by Prof. Bernd Girod

Frequency Response of Eye





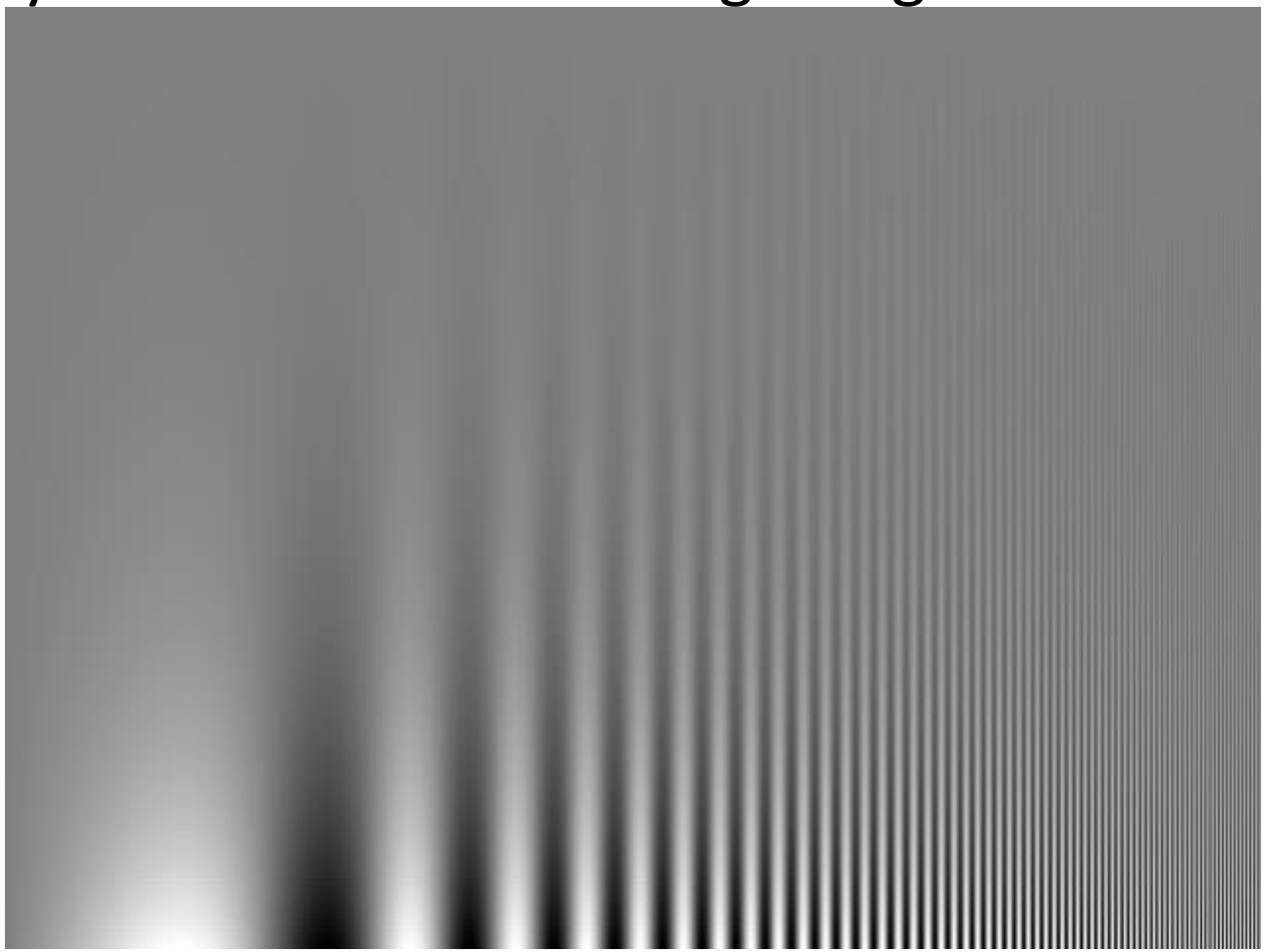
Contrast Sensitivity Function (CSF)



- Contrast sensitivity: the reciprocal of the threshold contrast needed to detect sinusoidal gratings

How to Measure?

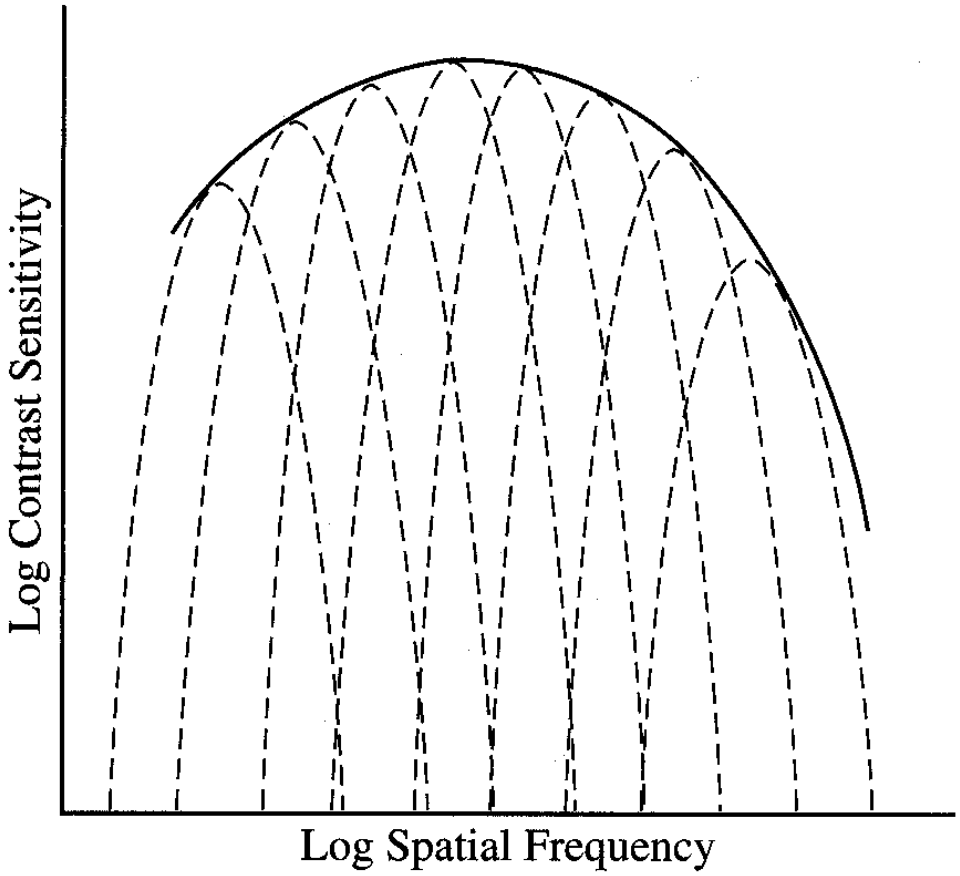
- By modulated sine wave grating



Oblique Effect

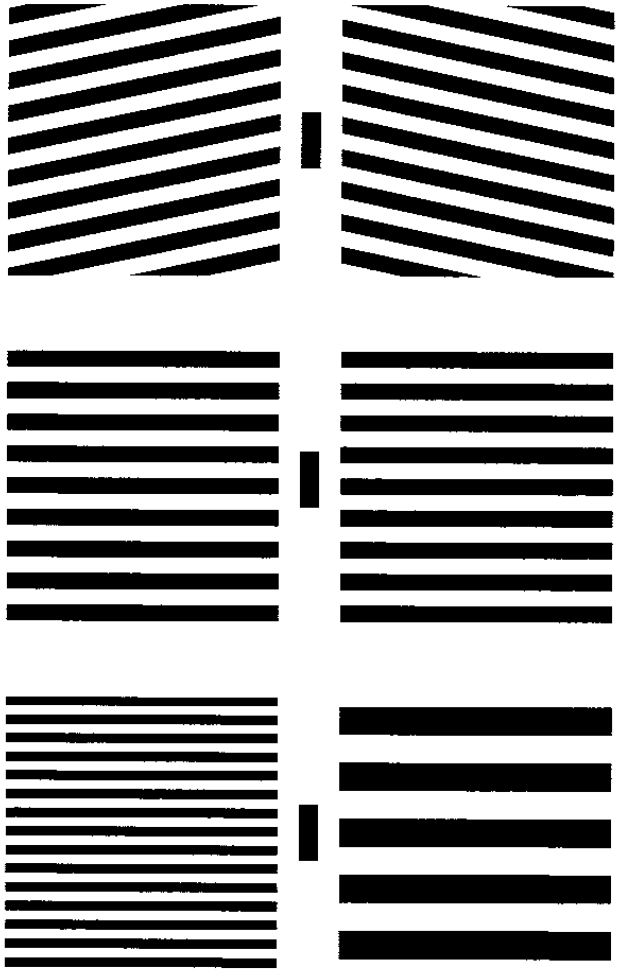
- The oblique effect is the term applied to the reduction in contrast sensitivity to obliquely oriented gratings compared to horizontal and vertical ones. This reduction in sensitivity (a factor of 2 or 3) occurs at high spatial frequencies.

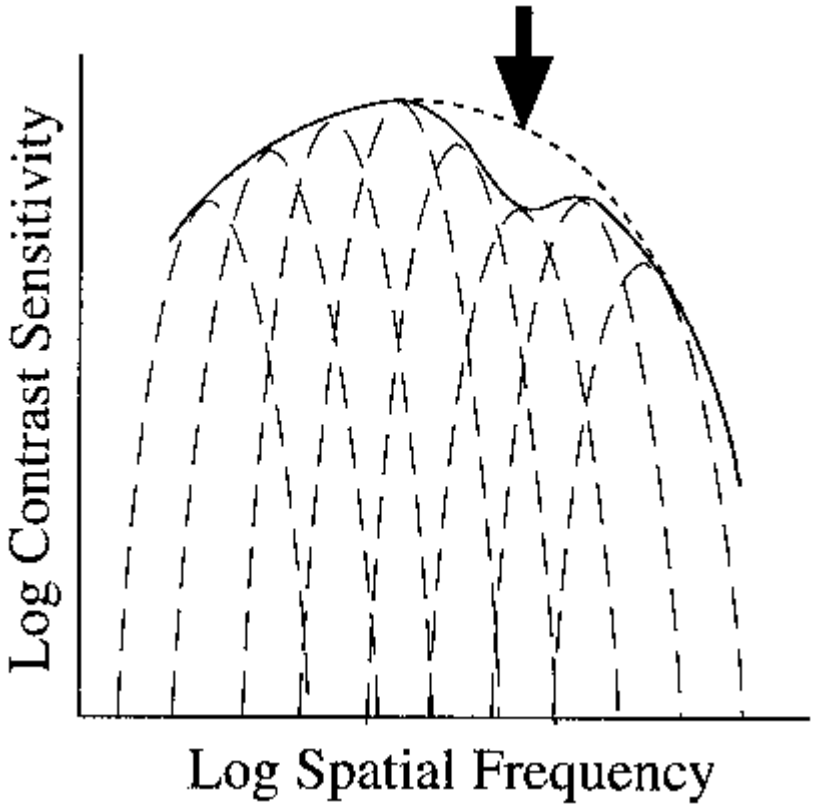
Multiple Spatial Frequency Channels



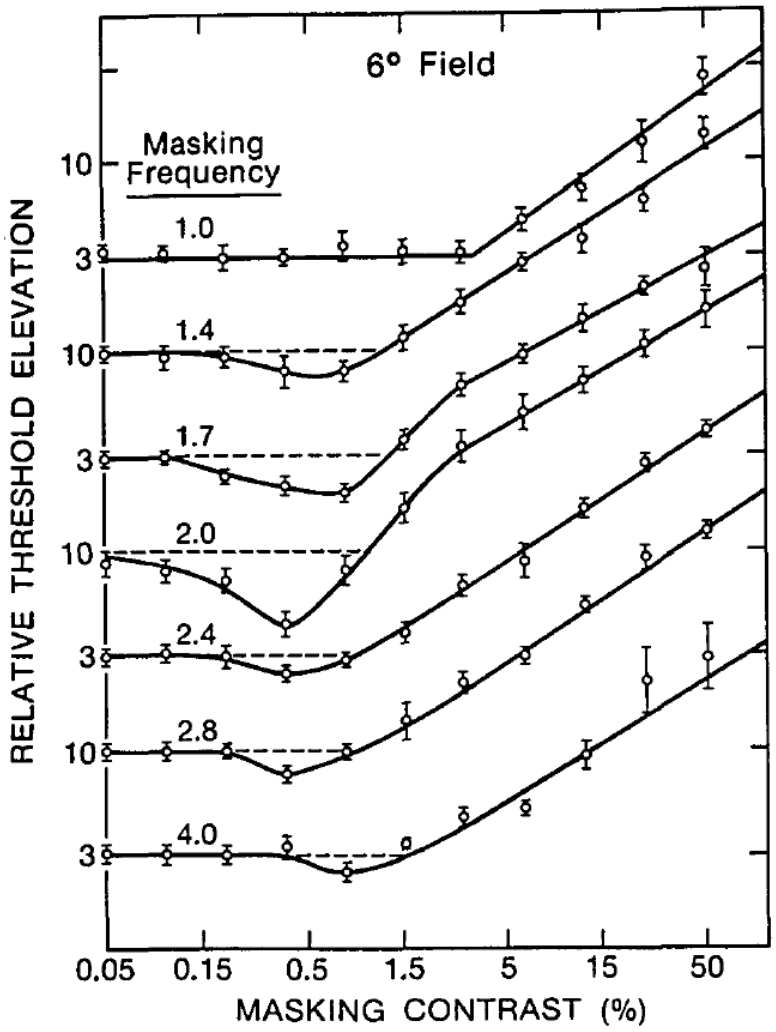
- The CSF represents the envelope of the sensitivity of many more narrowly tuned channels
- This multiple channel property exists for spatial frequency and orientation
- Multiresolution representation

Pattern Adaptation



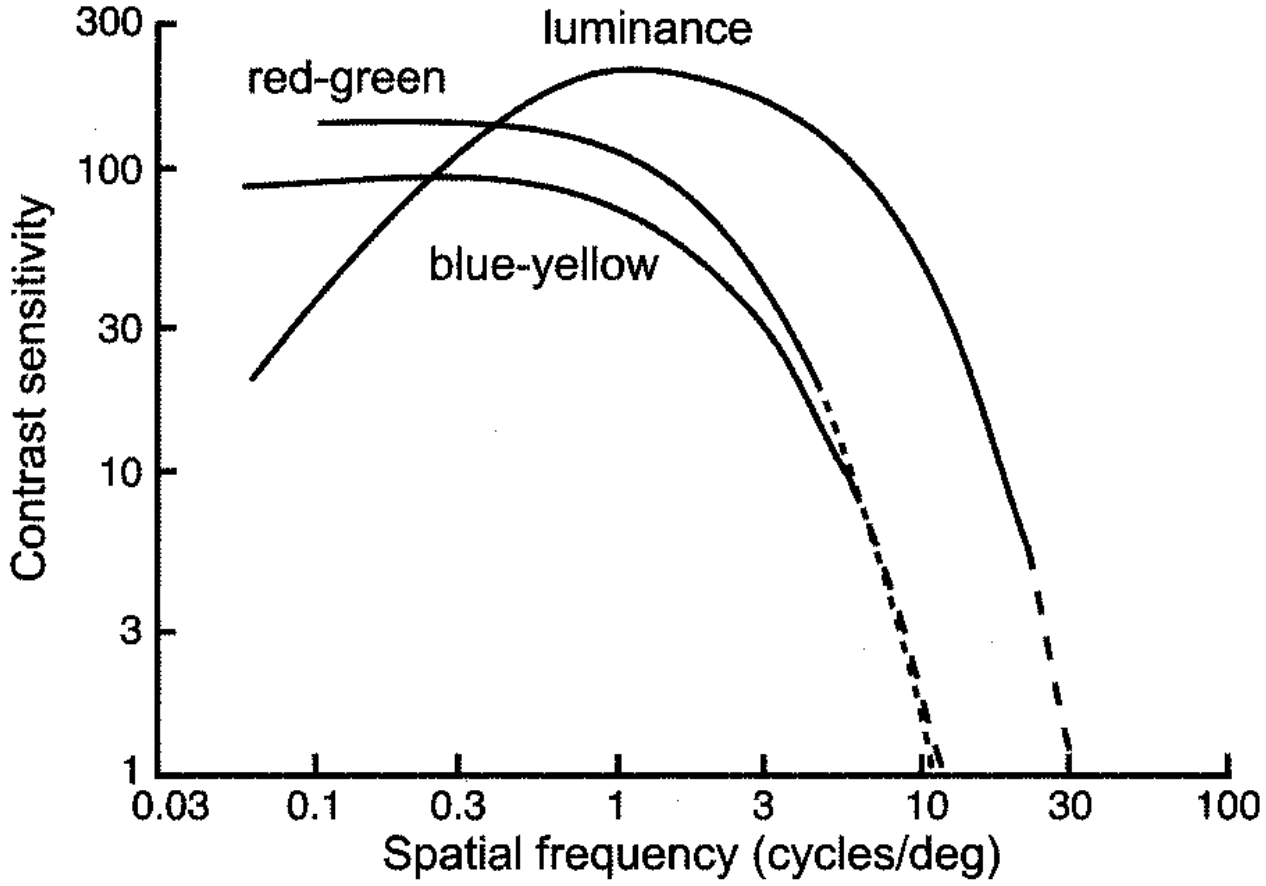


Masking and Facilitation

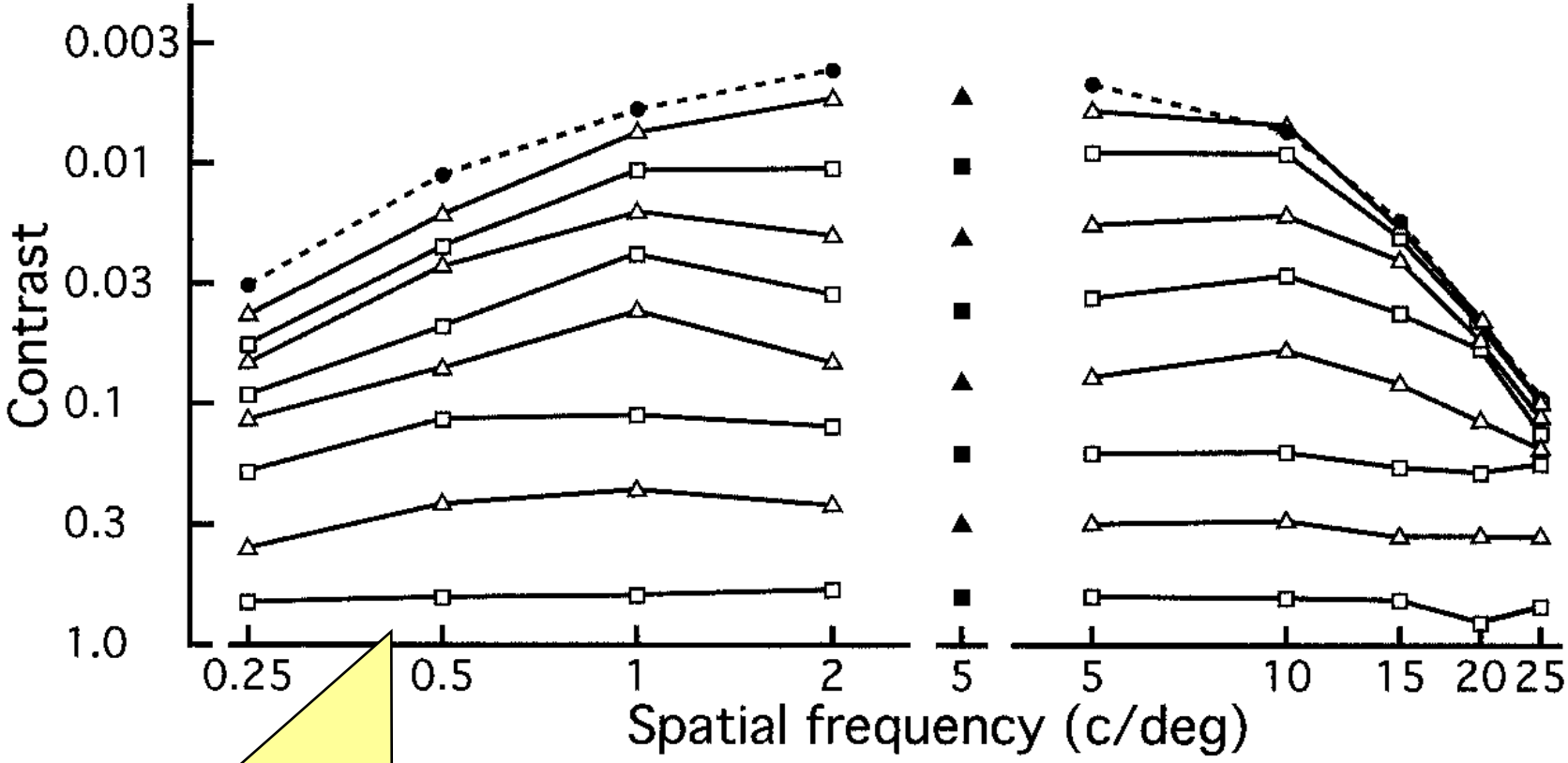


Target test grating: 2cpd

Chromatic Contrast Sensitivity



Suprathreshold Contrast Sensitivity



Outline

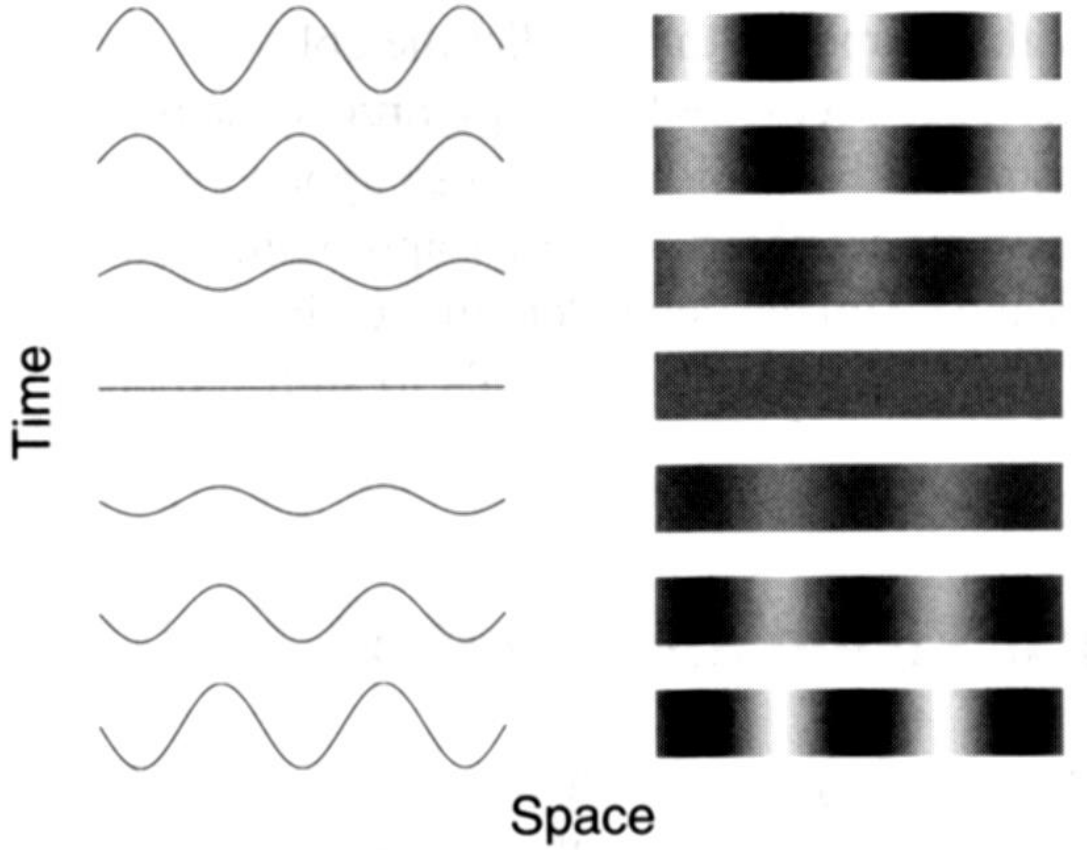
- Human vision system
- Color vision
- Luminance and the perception of light intensity
- Spatial vision and contrast sensitivity
- Temporal vision

Critical Flicker Frequency (CFF)

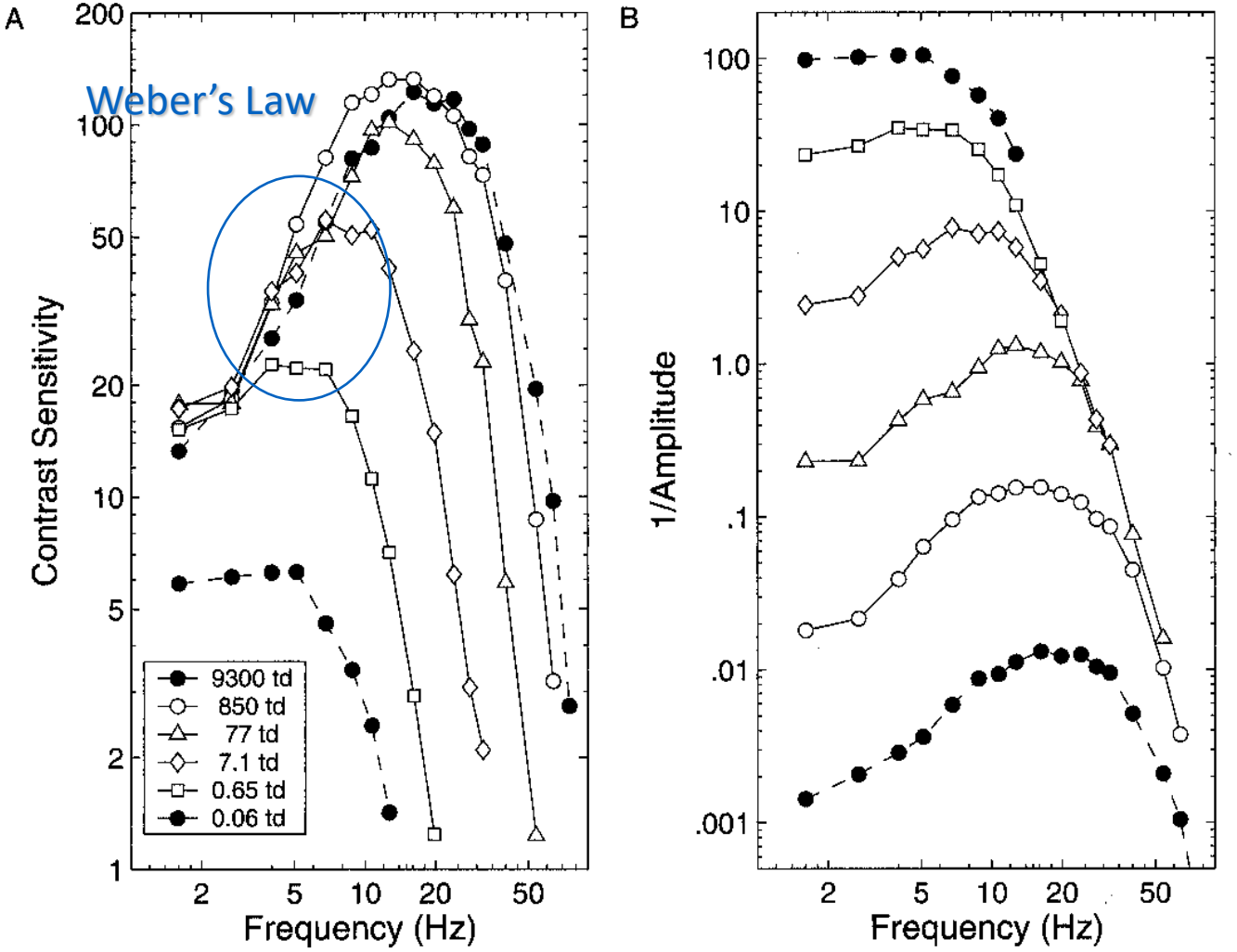
- Ferry-Porter law
  - CFF rises linearly with the logarithm of the time-average background intensity
- Talbot-Plateau law
  - The perceived intensity of a fused periodic stimulus is the same as a steady stimulus of the same time-average intensity

Temporal CSF

- How to measure?



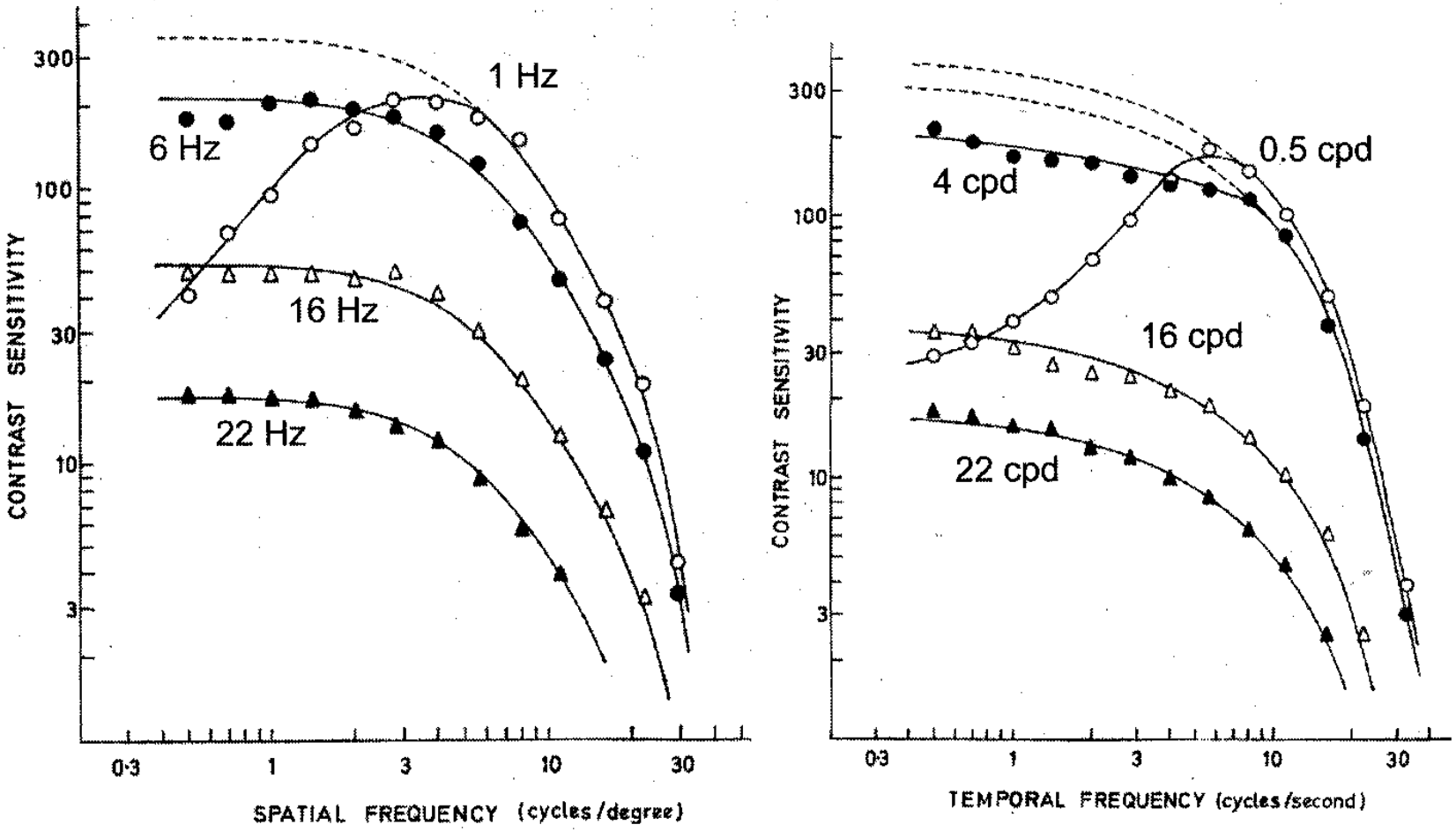
Temporal CSF



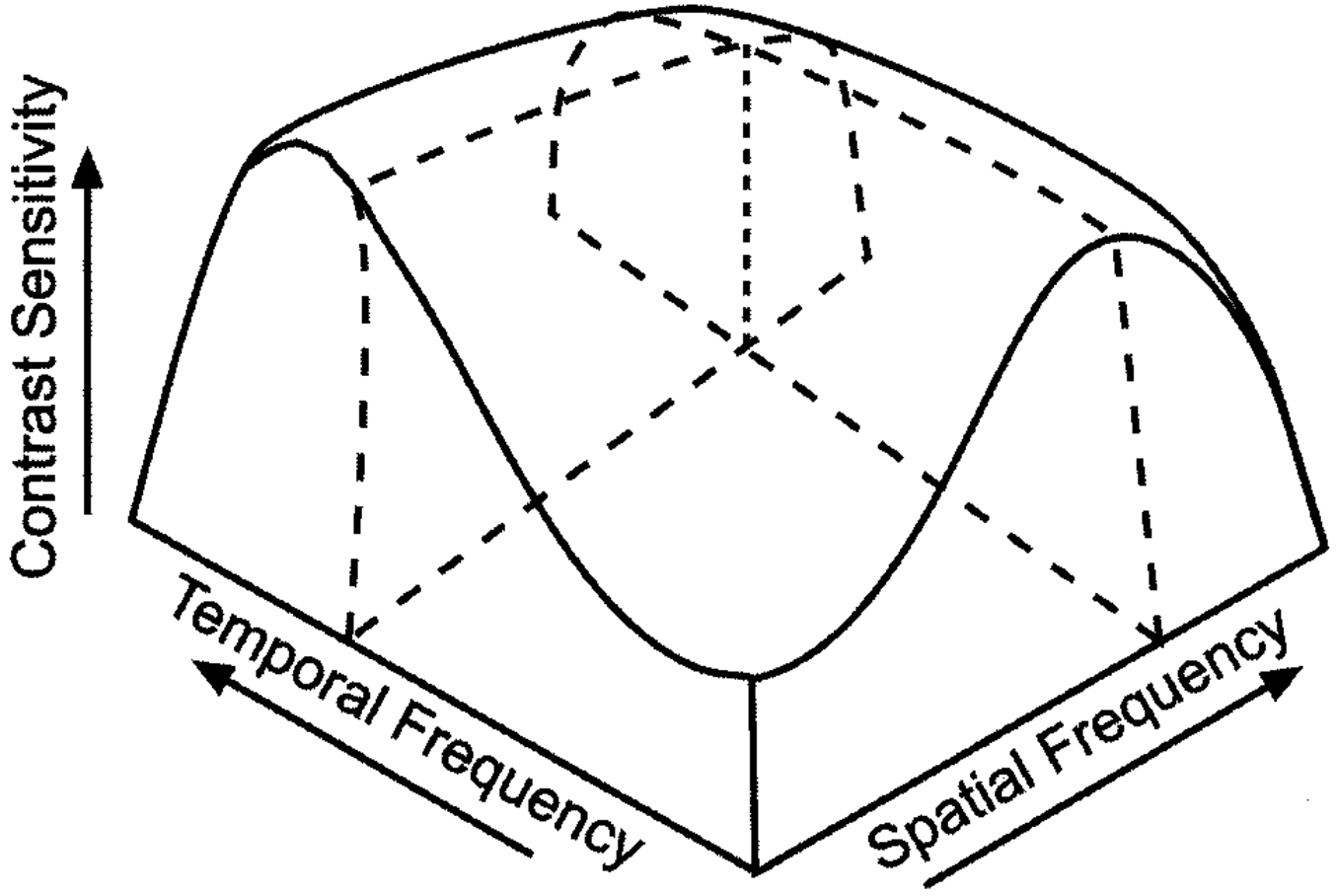
The eye is more sensitive to flicker at high luminance

The flicker sensitivity is negligible above 50/60 Hz

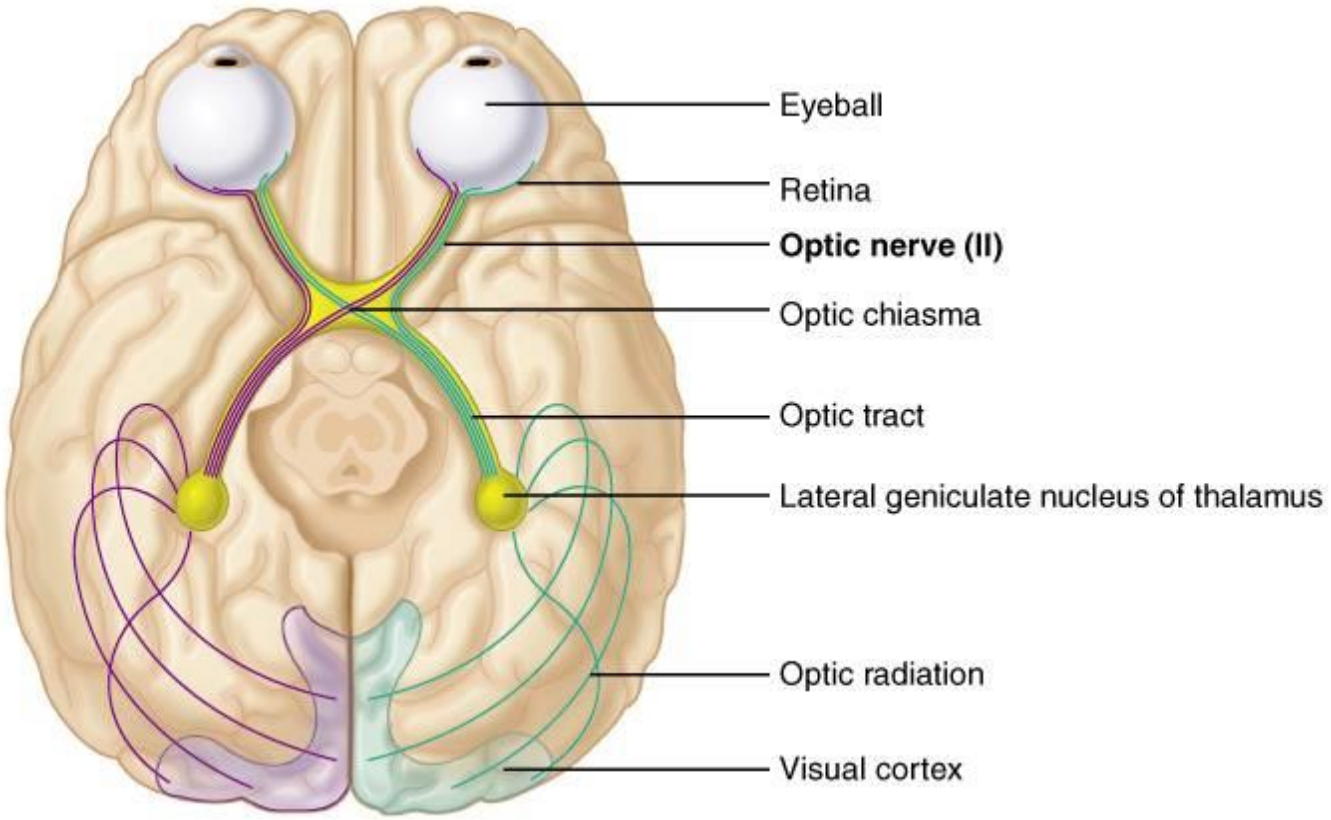
Spatial-Temporal CSF



Spatial-Temporal CSF

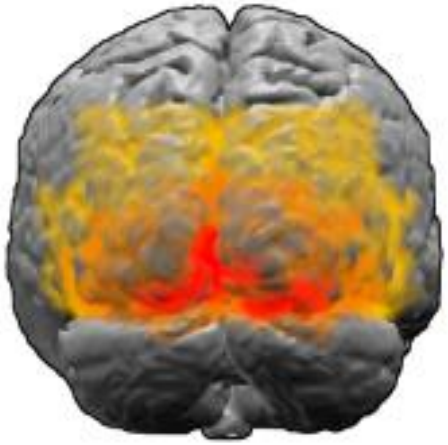


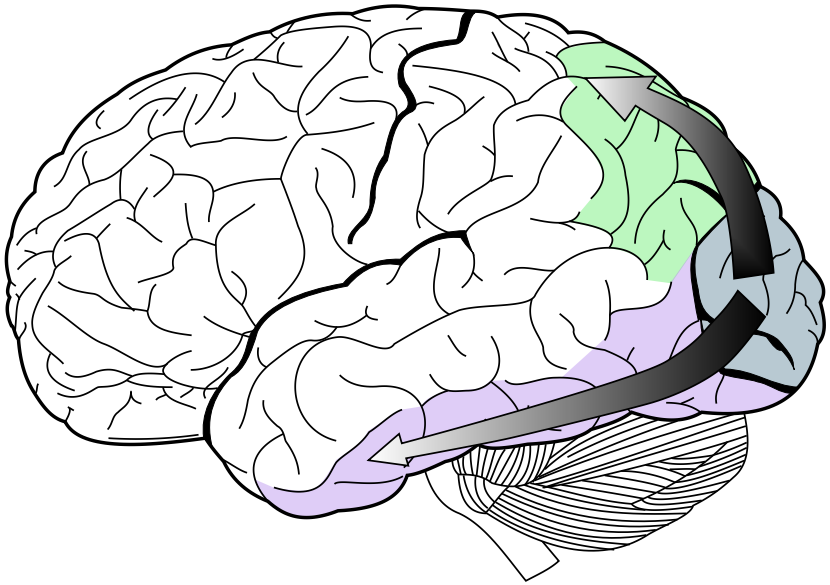
Path after Retina



Visual Cortex

- V1, V2, V3, V4, V5
- Dorsal stream: V1→V2→V5→Parietal Lobe (頂葉)
  - Where pathway
- Ventral stream: V1→V2→V4→ Temporal Lobe (顳葉)
  - What pathway





Visual Cortex

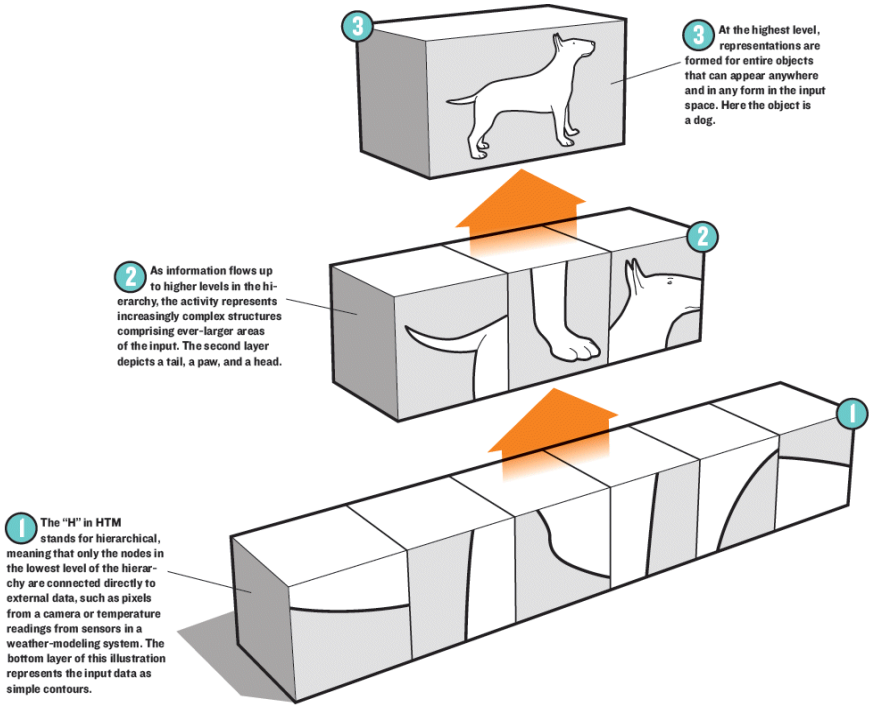
- V1: Receive all the information from retina. Composed of spatiotemporal filters. Contrast based instead of intensity based.
- V2: Get signals from V1 and V3—5. For more complex attribute, such as contour
- V3: Global motion
- V4: Connect to V1 and V2. Recognition, attention
- V5: Connect to V1. Gaze, the temporal information generation

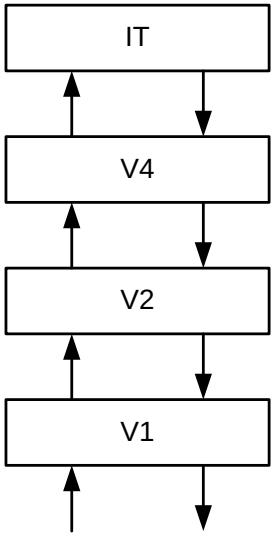
Image shown to subjects	40ms	80ms	107ms	500ms
	"Possibly outdoor scene, maybe a farm. I could not tell for sure."	"There seem to be two people in the center of the scene."	" People playing rugby. Two persons in close contact, wrestling, on grass. Another man more distant. Goal in sight."	"Some kind of game or fight. Two groups of two men. One in the foreground was getting a fist in the face. Outdoors, because I see grass and maybe lines on the grass? That is why I think of a game, rough game though, more like rugby than football because they weren't in pads and helmets..."

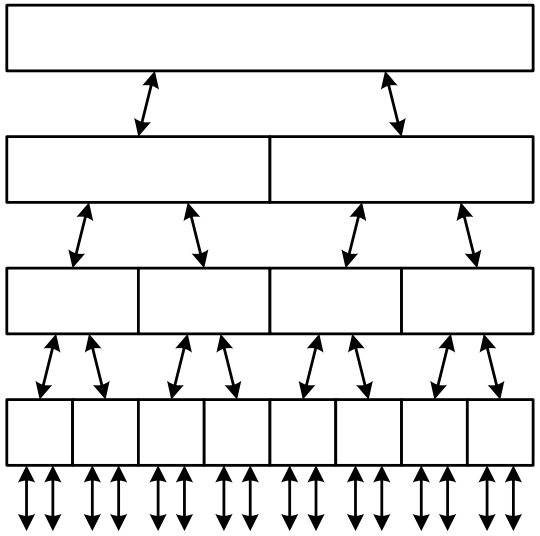
Figure 2. Human subjects reporting on what he/she saw in an image shown for different presentation durations (PD=27, 40, 67, 80, 107, 500ms). From Fei-Fei and Perona, JoV 2007 [26].

Fei-Fei, Iyer, Koch, Perona, *J. Vision,* 2007

Visual Cortex – HTM Model

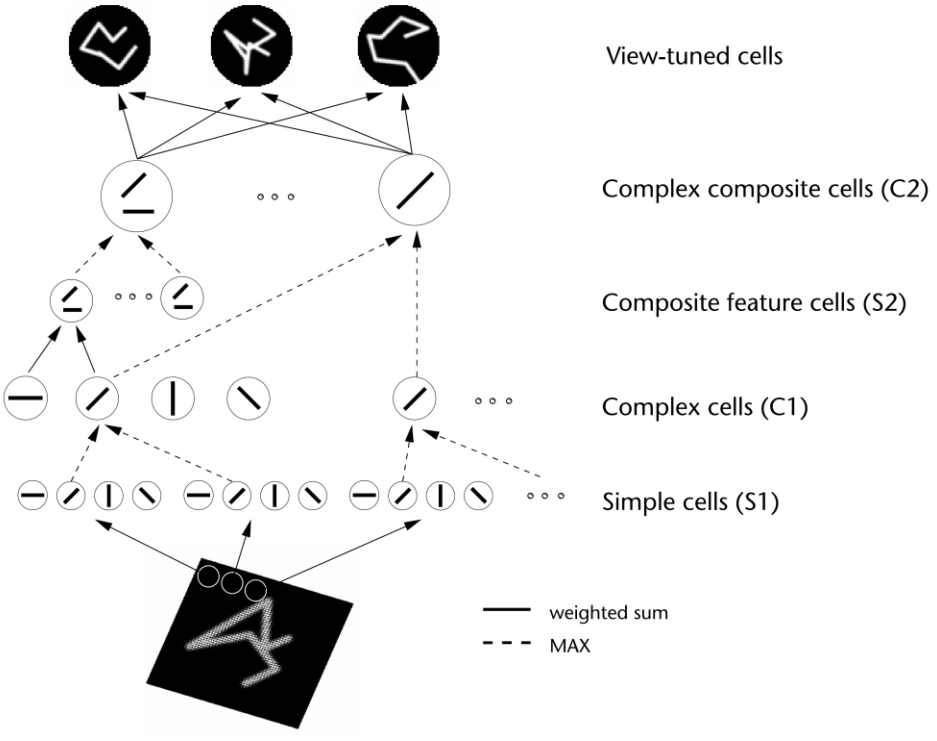


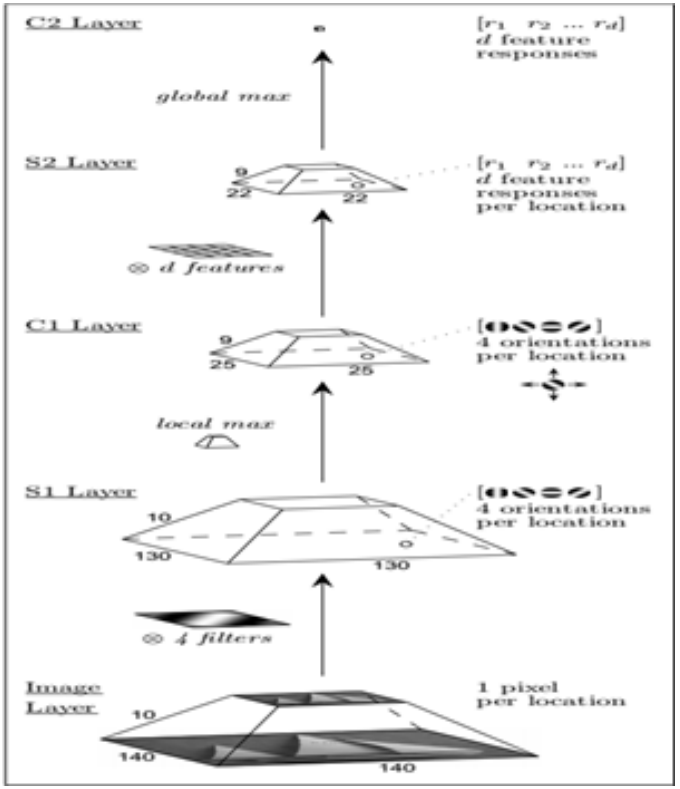




J. Hawkins, "Why Can't a Computer be more Like a Brain?" *IEEE Spectrum*, vol. 44, no. 4, pp. 21-26, 2007.

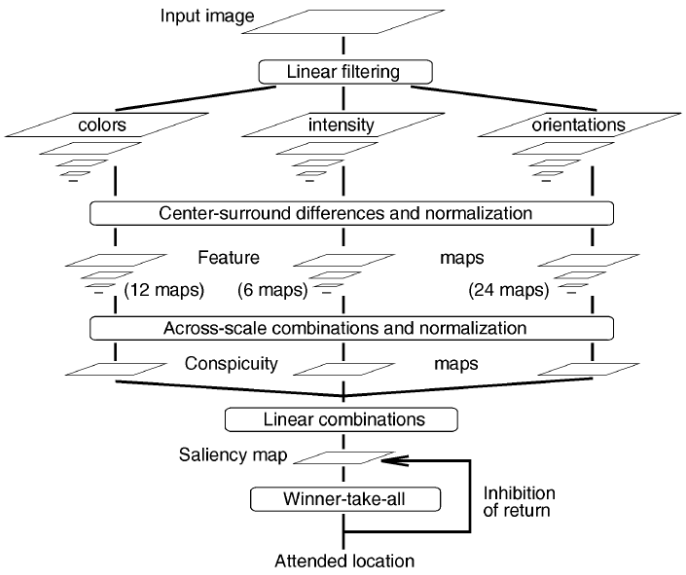
Visual Cortex – HMAX Model



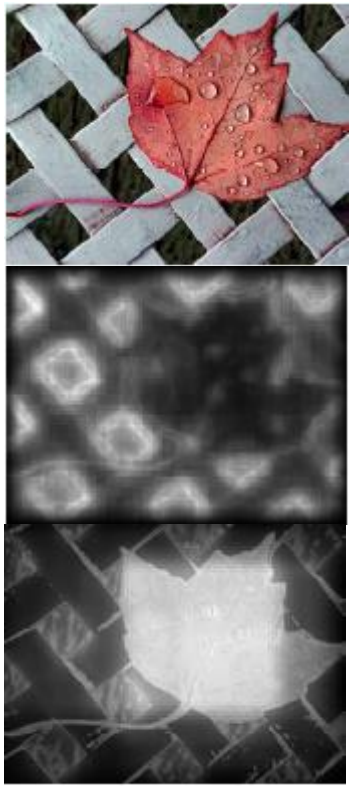


M. Riesenhuber and T. Poggio, "Why Can't a Computer be more Like a Brain?" *Nature Neuroscience*, vol. 2, no. 11, 1999.

Attention Model

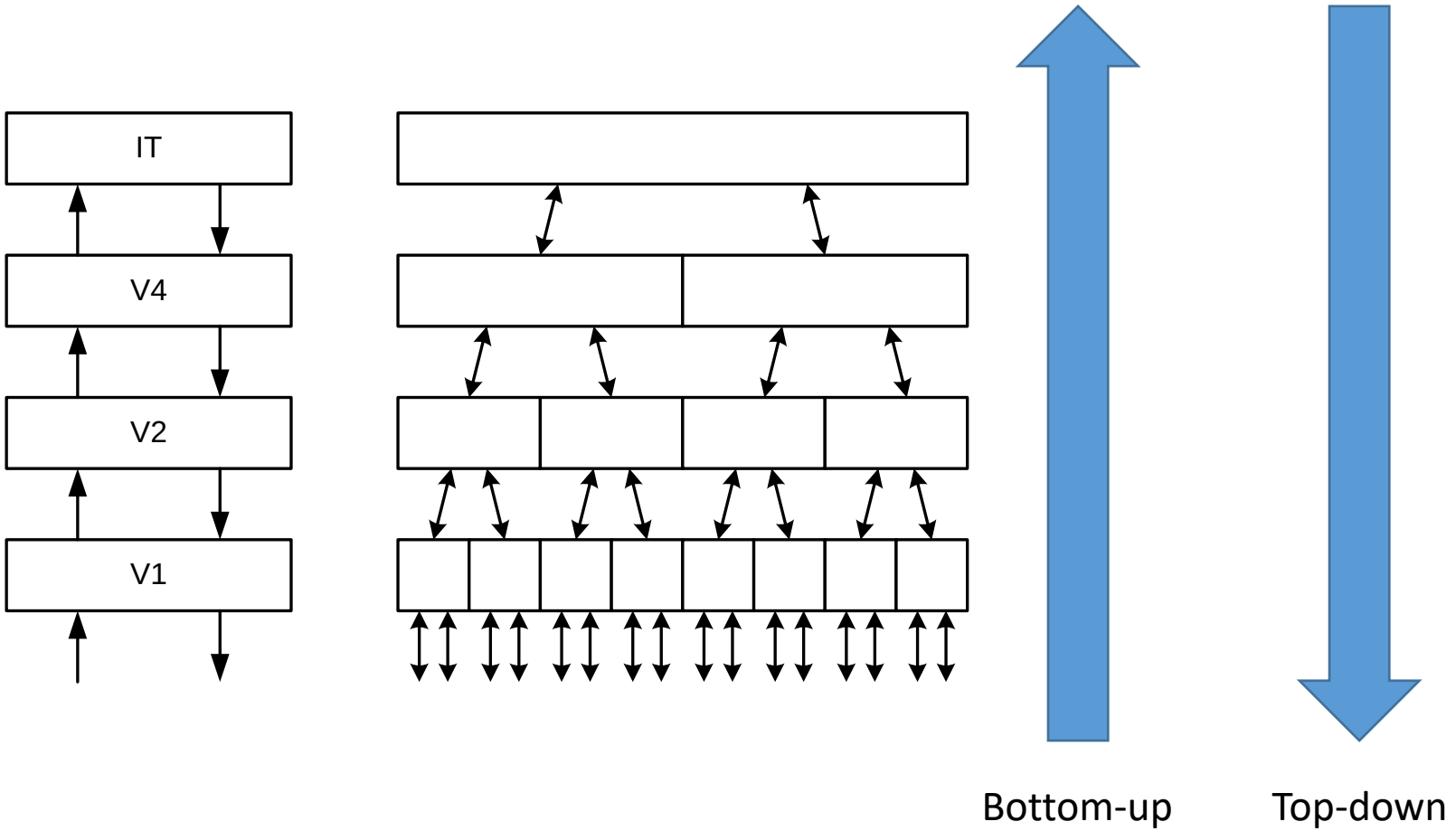


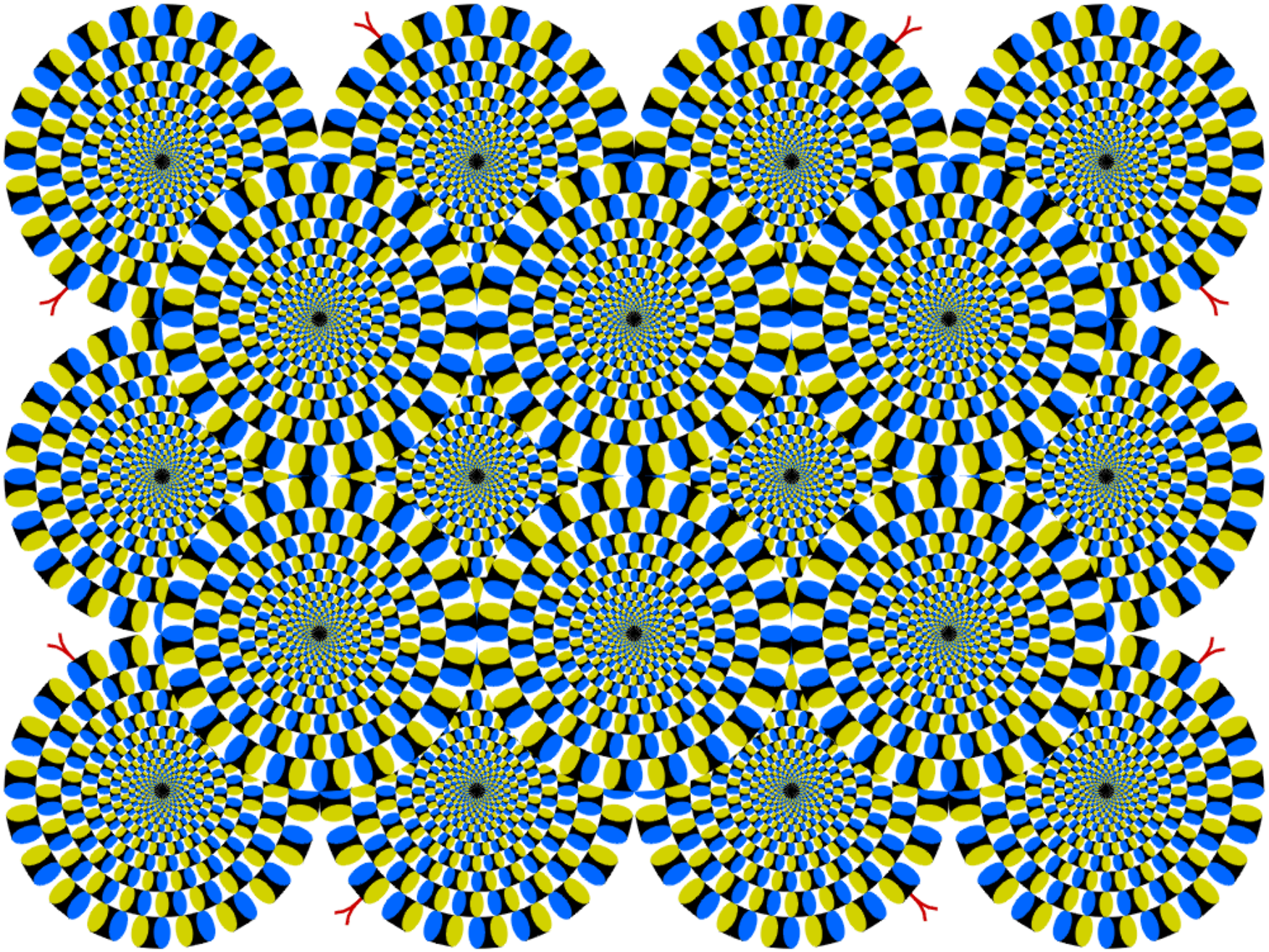
[Itti 1998]



Ref: Wei-Chih Tu , Shengfeng He, Qingxiong Yang, and Shao-Yi Chien, “Real-time salient object detection with a minimum spanning tree,” *CVPR2016*.

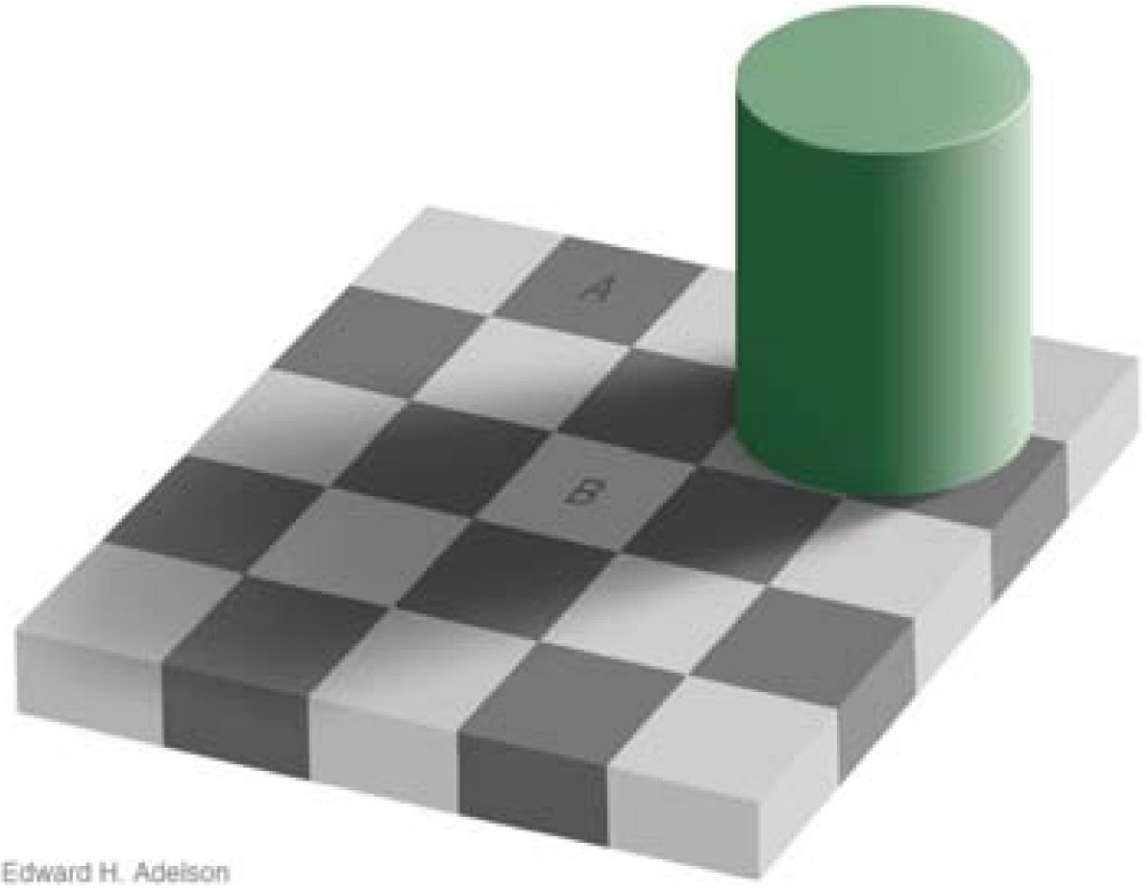
Top-down and Bottom-up





Copyright A.Kitaoka 2003

Which one is darker? A or B?



Which one is embossed?