Computer Vision

James L. Crowley

M2R MoSIG option GVR

Fall Semester
13 October 2011

Lesson 2

Visual Perception in Man and Machine

Lesson Outline:

1 Projective Camera models.

1.1 Projection between two planes (Homographies)

The projective transformation from one plane to another is called a homography.
A homography is bi-jective (reversible).

A planar homography is expressed as a 3x3 matrix :

$$\begin{pmatrix} x_b \\ y_b \\ 1 \end{pmatrix} = \begin{pmatrix} wx_b \\ wy_b \\ w \end{pmatrix} = \begin{pmatrix} h_1^1 & h_2^1 & h_3^1 \\ h_1^2 & h_2^1 & h_3^1 \\ h_1^3 & h_2^3 & 1 \end{pmatrix} \begin{pmatrix} x_a \\ y_a \\ 1 \end{pmatrix}$$

$$\mathrm{x_B} = \frac{wx_B}{w} = \frac{h_{11}x_A + h_{12}y_A + h_{13}}{h_{31}x_A + h_{32}y_A + h_{33}} \qquad \mathrm{y_B} = \frac{wy_B}{w} = \frac{h_{21}x_A + h_{22}y_A + h_{23}}{h_{31}x_A + h_{32}y_A + h_{33}}$$

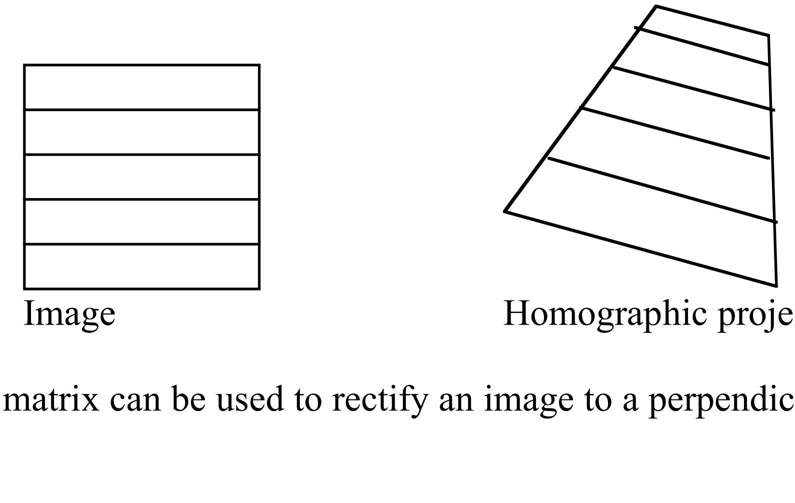
In tensor notation

$$\boxed{Q^B = \mathbf{H}_A{}^B \; P^A}$$

$$\begin{pmatrix} q^1 \\ q^2 \\ q^3 \end{pmatrix} = \mathbf{H}_A^B \begin{pmatrix} p^1 \\ p^2 \\ p^3 \end{pmatrix} = \begin{pmatrix} h_1^1 & h_2^1 & h_3^1 \\ h_1^2 & h_2^2 & h_3^2 \\ h_1^3 & h_2^3 & h_3^3 \end{pmatrix} \begin{pmatrix} p^1 \\ p^2 \\ p^3 \end{pmatrix}$$

$$x_B = \frac{q^1}{q^3} \qquad y_B = \frac{q^2}{q^3}$$

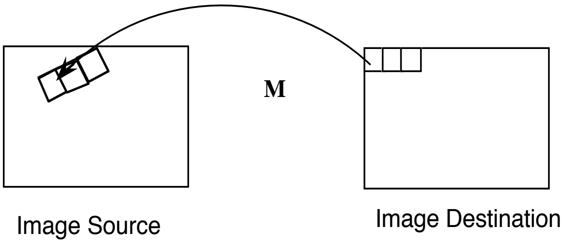
The projection of a plane to another plane is a degenerate case of the the project transform. In this case, the transform is bijective and reduces to a 3 x 3 invertible



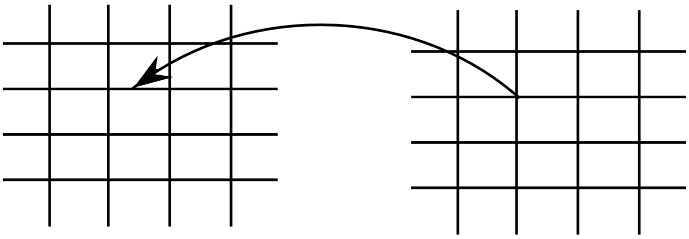
This matrix can be used to rectify an image to a perpendicular view.

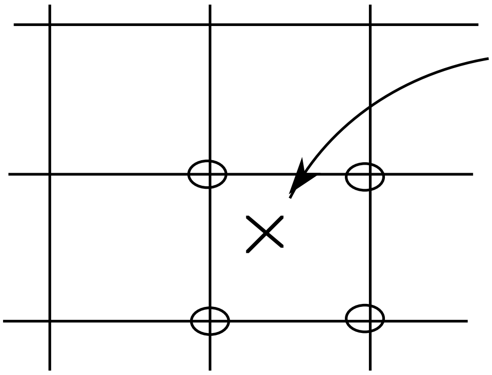
1.2 Image Rectification

For each pixel in the destination image, (x_d, y_d) compute its position in the source image(x_s, y_s)



Determine the appropriate pixel value (intensity or color) for the source image and put this pixel value in the destination.

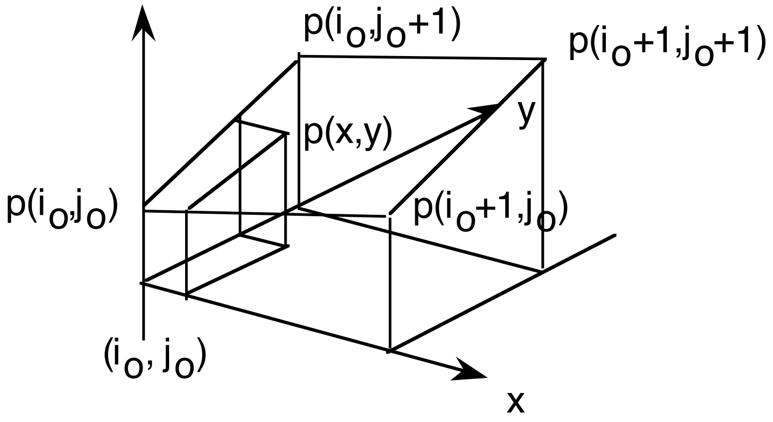




The problem is that the calculated pixel is a real number.
To obtain a destination pixel value we need to interpolate. This can be done by

zeroth order:	Nearest neighbor
First order:	Linear or bilinear interpolation
second order	Cubic spline.

1.3 Bilinear Interpolation



The mathematical form is a hyperbolic parabaloid.

$$p(x, y) = a\,x + b\,y + c\,x\,y + d.$$

For an integer position array p(i,j), the interpolation at real position x, y can be given as an offset from the adjacent pixel p(i,j) where i =trunc(x) and j=trunc(y)
where coefficients are given by the slopes m_x, m_y, m_{xy}

$$p(x, y) = m_x \cdot(x-i) + m_y\cdot (y-j) + m_{xy} \cdot (x-i)\cdot (y-j) + p(i,j).$$

where :

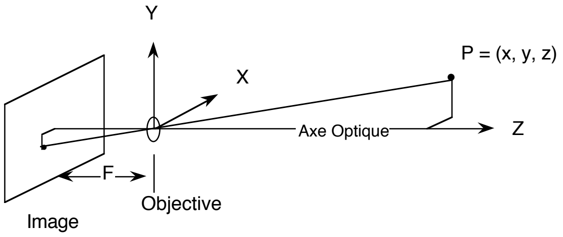
$$a \equiv m_x = \frac{\Delta P}{\Delta x} = p(i+1, j) - p(i, j)$$

$$b \equiv m_y = \frac{\Delta P}{\Delta y} = p(i, j+1) - p(i, j)$$

$$c \equiv m_{xy} = p(i+1, j) + p(i, j+1) - p(i, j) - p(i+1, j+1)$$

$$d = p(i, j)$$

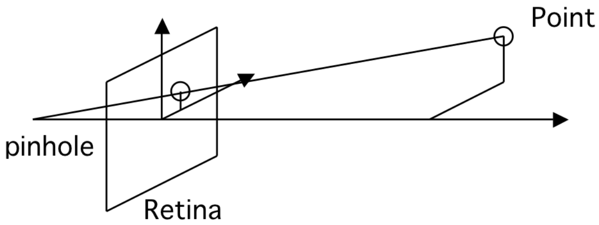
1.4 The Pinhole Camera



Points in the scene are projected to an "up-side down" image on the retina.

This is the "Pin-hole model" for the camera.

The scientific community of computer vision often uses the "Central Projection Model". In the Central Projection Model, the retina is placed in Front of the projective point.



We will model the camera as a projective transformation from scene coordinates, S, to image coordinates, i.

$$\vec{Q}^i = \mathbf{M}_s^i \vec{P}^s$$

This transformation is expressed as a 3x4 matrix:

$$M_s^i = \begin{pmatrix} m_{11} & m_{12} & m_{13} & m_{14} \\ m_{21} & m_{22} & m_{23} & m_{24} \\ m_{31} & m_{32} & m_{33} & m_{34} \end{pmatrix}$$

composed from 3 transformations between 4 reference frames.

1.5 Extrinsic and Intrinsic camera parameters

The camera model can be expressed as a function of 11 parameters.
These are often separated into 6 "extrinsic" parameters and 5 "intrinsic" parameters:
Thus the "Extrinsic" parameters of the camera describe the camera position and orientation in the scene. These are the six parameters:

Extrinsic Parameters = $(x, y, z, \theta, \varphi, \gamma)$

The intrinsic camera parameters express the projection to the retina, and the mapping to the image. These are :

F : The "focal" length
C_x, C_y : the image center (expressed in pixels).
D_x, D_y : The size of pixels expressed in pixels/mm.

1.6 Coordinate Systems

This transformation can be decomposed into 3 basic transformations between 4 reference frames. The reference frames are:

Coordinate Systems:

Scene Coordinates:

Point Scène: $P^s = (x_s, y_s, z_s, 1)^T$

Camera Coordinates:

external world: $P^c = (x_c, y_c, z_c, 1)^T$

Retina: $Q^r = (x_r, y_r, 1)^T$

Image Coordaintes

Image: $Q^i = (i, j, 1)^T$

Using "tensor notation" to keep track of our reference frames, the transformations are represented by projective transformations in homogeneous coordinates:

$$\vec{Q}^i = C_r^i \vec{Q}^r \qquad \vec{Q}^r = P_c^r \vec{P}^c \qquad \vec{P}^c = T_s^c \vec{P}^s$$

These express

1) A translation/rotation from scene to camera coordinates: T_s^c

2) A projection from scene points in camera coordinates to the retina: P_c^r

3) Sampling scan and A/D conversion of the retina to give an image: C_r^i

When expressed in homogeneous coordinates, these transformations are composed as matrix multiplications.

$$\vec{Q}^i = C_r^i P_c^r T_s^c \vec{P}^s = M_s^i \vec{P}^s$$

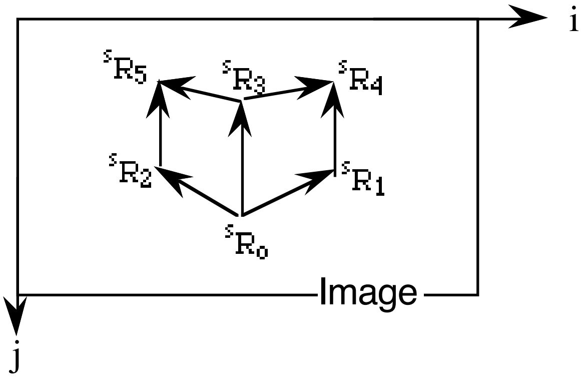
1.7 Calibrating a Camera Model

How can we obtain the model $\mathbf{M}_s^i$? By a process of calibration.

Observe a set of at least 6 non-coplanar points whose position in the world is known.

$R^s{}_k$ for k=0,1,2,3,4,5 (s are the scene coordinate axes s=1,2,3)

For example, we can use the corners of a cube. Define the lower front corner as the origin, and the edges as unit distances.



The matrice $\mathbf{M}_s^i$ is composed of 3x4=12 coefficients. However because, $\mathbf{M}_s^i$ is in homogeneous coordinates, the coordinate m_{34} can be set to 1.

Thus there are 12-1 = 11.
We can determine these coefficients by observing known points in the scene. ($R^s{}_k$).
Each point provides two coefficients. Thus, for 11 coefficients we need at least $5\frac{1}{2}$ points. With 6 points the system is over-constrained.

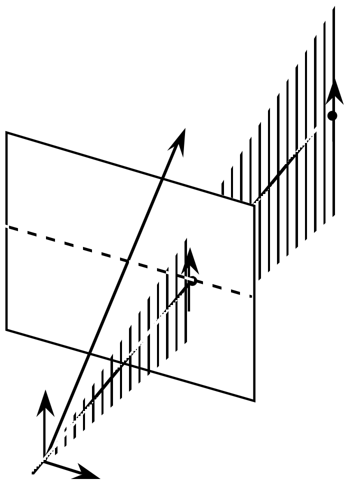
For each known calibration point $R^s{}_k$ given its observed image position $P^s{}_k$, we can write:

$$i_k = \frac{w_k\, i_k}{w_k} = \frac{M_s^1 \cdot R_k^s}{M_s^3 \cdot R_k^s} \qquad j_k = \frac{w_k\, j_k}{w_k} = \frac{M_s^2 \cdot R_k^s}{M_s^3 \cdot R_k^s}$$

This gives 2 equations for each point.

$$(M_s^1 \cdot R^s{}_k) - i_k\,(M_s^3 \cdot R^s{}_k) = 0 \qquad (M_s^2 \cdot R^s{}_k) - j_k\,(M_s^3 \cdot R^s{}_k) = 0$$

Each pair of equations corresponds to the planes that pass though the image row and the image column of the observed image point $P^s{}_k$



The equation $(M_s^1 \cdot R^s{}_k) - i_k\ (M_s^3 \cdot R^s{}_k) = 0$ is the vertical plane that includes the projective center through the pixel $i = i_k$.

The equation $(M_s^2 \cdot R^s{}_k) - j_k\ (M_s^3 \cdot R^s{}_k) = 0$ is the horizontal plane that includes the projective center and the row $j = j_k$.

In tensor notation

given $P^i = \begin{pmatrix} wi \\ wj \\ w \end{pmatrix}$ we write : $P^i = \mathbf{M}_s^i\ R^s$

with k scene points, R_k^S and their image correspondences P_k^i we can write

$$P_k^i = \mathbf{M}_s^i\ R_k^s$$

with $i \cdot w = P_k^1 / P_k^3$ et $j \cdot w = P_k^2 / P_k^3$ for each image point k, there are two independent equations

$$\begin{pmatrix} p^1 \\ p^2 \\ p^3 \end{pmatrix} = \begin{pmatrix} wi \\ wj \\ w \end{pmatrix} \text{ donc } \begin{pmatrix} i \\ j \\ 1 \end{pmatrix} = \begin{pmatrix} p^1/p^3 \\ p^2/p^3 \\ 1 \end{pmatrix}$$

and with $P_k^3 = \mathbf{M}_s^3\, R_k^3$

$$i = p^1/p^3 = \mathbf{M}_s^1\ R_k^s / \mathbf{M}_s^3\ R_k^s \Rightarrow i\, \mathbf{M}_s^3\ R_k^s - \mathbf{M}_s^1\ R_k^s = 0$$
$$j = p^2/p^3 = \mathbf{M}_s^2\ R_k^s / \mathbf{M}_s^3\ R_k^s \Rightarrow j\, \mathbf{M}_s^3\ R_k^s - \mathbf{M}_s^2\ R_k^s = 0$$

We can write this as:

$$\begin{pmatrix} R^1 & R^2 & R^3 & 1 & 0 & 0 & 0 & 0 & -iR^1 & -iR^2 & -iR^3 & -i \\ 0 & 0 & 0 & 0 & R^1 & R^2 & R^3 & 1 & -jR^1 & -jR^2 & -jR^3 & -j \end{pmatrix} \begin{pmatrix} \mathbf{M}_1^1 \\ \mathbf{M}_2^1 \\ \mathbf{M}_3^1 \\ \mathbf{M}_4^1 \\ \mathbf{M}_1^2 \\ \mathbf{M}_2^2 \\ \mathbf{M}_3^2 \\ \mathbf{M}_4^2 \\ \mathbf{M}_1^3 \\ \mathbf{M}_2^3 \\ \mathbf{M}_3^3 \\ \mathbf{M}_4^3 \end{pmatrix} = 0$$

For N non-coplanair points we can write 2N equations.

$$\mathbf{A}\ \mathbf{M}_s^i = 0.$$

We then use least squares to minimize the criteria:

$$\mathbf{C} = \ \| \mathbf{A}\, \mathbf{M}_s^i \ \|$$

For example, give a cube with observed corners

$P^L{}_o = (101, 221)$	$P^L{}_1 = (144, 181)$	$P^L{}_2 = (22, 196)$
$P^L{}_3 = (105, 88)$	$P^L{}_4 = (145, 59)$	$P^L{}_5 = (23, 67)$

Least squares will give:

$$\mathbf{M}_s^i = \begin{pmatrix} 55.886873 & -79.292084 & 1.276703 & 101.917630 \\ -22.289319 & -17.878203 & -134.345576 & 221.300658 \\ 0.100734 & 0.038274 & -0.008458 & 1.000000 \end{pmatrix}$$

2 The Physics of Light

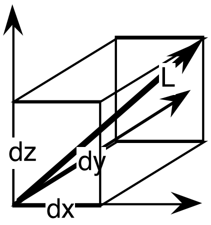
2.1 Photons and the Electo-Magnetic Spectrum

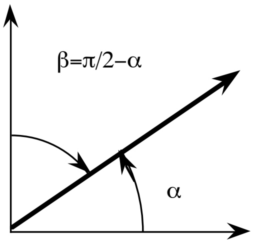
A photon is a resonant electromagnetic oscillation.
The resonance is described by Maxwell's equations.
The magnetic field is strength determined the rate of change of the electric field, and the electric field strength is determined by the rate of change of the magnetic field.

The photon is characterized by
1) a direction of propagation , $\vec{D}$,
2) a polarity (direction of oscillation), and
3) a wavelength, λ, and its dual a frequence, f : $\lambda = \frac{1}{f}$

Direction of propagation and direction of polarity can be represented as a vector of Cosine angles.

$$\vec{D} = \begin{pmatrix} \cos(\alpha) \\ \cos(\beta) \\ \cos(\gamma) \end{pmatrix} = \begin{pmatrix} \Delta x / L \\ \Delta y / L \\ \Delta z / L \end{pmatrix}$$





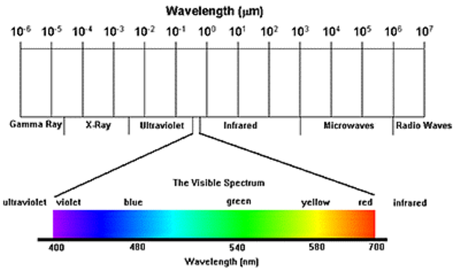
Photon propagation is a probabilistic phenomenon, described by Quantum Chromo-Dynamics. Photons are created and absorbed by abrupt changes in the orbits of electrons. Absorption and creation are probabilistic (non-deterministic) events.

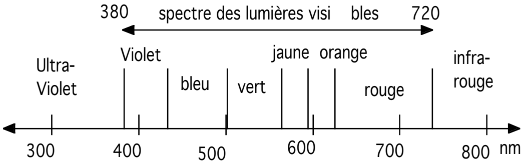
Photons sources generally emit photons over a continuum of directions (a beam) and continuum of wavelengths (spectrum). The beam intensity is measured in Lumens, and is equivalent to Photons/Meter^2.

A lumen a measure of the total "amount" of visible light emitted by a source.
The lumen can be thought of as a measure of the total photons of visible light in some defined beam or angle, or emitted from some source.

The beam spectrum gives the probability of a photon having a particularly wavelength, S(λ).

The human eye is capable of sensing photons with a wavelength between 380 nanometers and 720 nanometers.



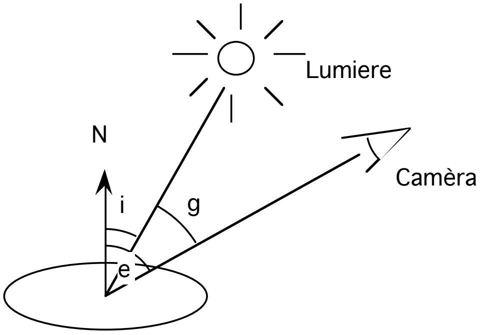


Perception is a probabilistic Phenomena.

2.2 Albedo and Reflectance Functions

The albedo of a surface is the ratio of photons emitted over photons received. Albedo is described by a Reflectance function

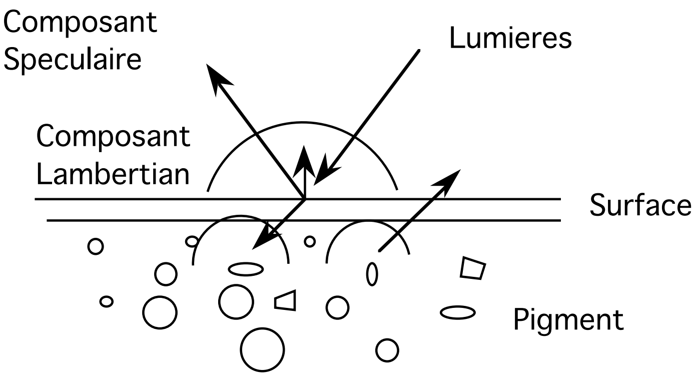
$$R(i, e, g, \lambda) = \frac{\text{Number of photons emitted}}{\text{Number of photons received}}$$



The parameters are

i: The incident angle (between the photon source and the normal of the surface).
e: The emittance angle (between the camera and the normal of the surface)
g: The angle between the Camera and the Source.
λ: The wavelength

For most materials, when photons arrive at a surface, some percentage are rejected by an interface layer (determined by the wavelength). The remainder penetrate and are absorbed by molecules near the surface (pigments).



Most reflectance functions can be modeled as a weighted sum of two components: A Lambertian component and a specular component.

$$R(i, e, g, \lambda) = c\, R_S(i, e, g, \lambda) + (1 - c)\, R_L(i, \lambda)$$

Specular Reflection

$$R_S(i, e, g, \lambda) = \begin{cases} 1 & \text{if } i = e \text{ and } i + e = g \\ 0 & \text{o}\textit{therwise} \end{cases}$$

An example of a specular reflector is a mirror.
All (almost all) of the photons are reflected at the interface level with no change in spectrum.

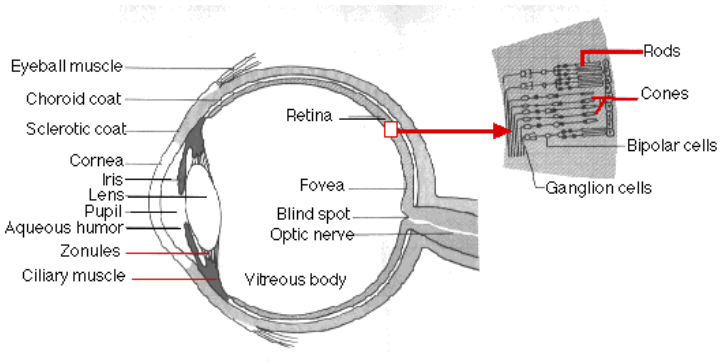
Lambertion Reflection

$$R_L(i, \lambda) = P(\lambda)\cos(i)$$

Paper, and fresh snow are examples of Lambertian reflectors.

3 The Human Visual System

3.1 The Human Eye

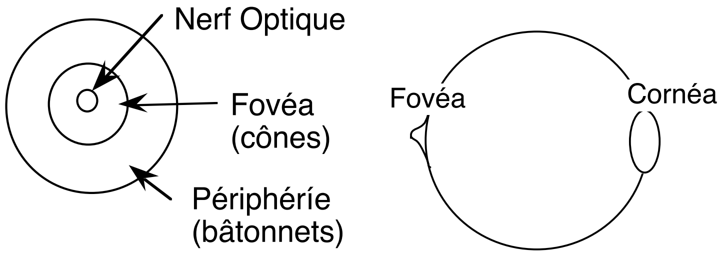


The human eye is a spherical globe filled with transparent liquid.
An opening (iris) allows light to enter and be focused by a lens.
Light arrives at the back of the eye on the Retina.

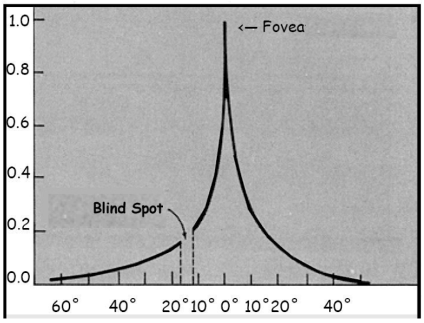
3.2 The Retina

The human retina is a tissue composed of a rods, cones and bi-polar cells.
Cones are responsible for daytime vision.
Rods provide night vision.
Bi-polar cells perform initial image processing in the retina.

Fovea and Peripheral regions



The cones are distributed over a non-uniform region in the back of the eye.
The density of cones decreases exponentially from a central point.
The fovea contains a "hole" where the optic nerve leaves the retina.



The central region of the fovea is concentrates visual acuity and is used for recognition and depth perception. The peripheral regions have a much lower density of cones, and are used for to direct eye movements.

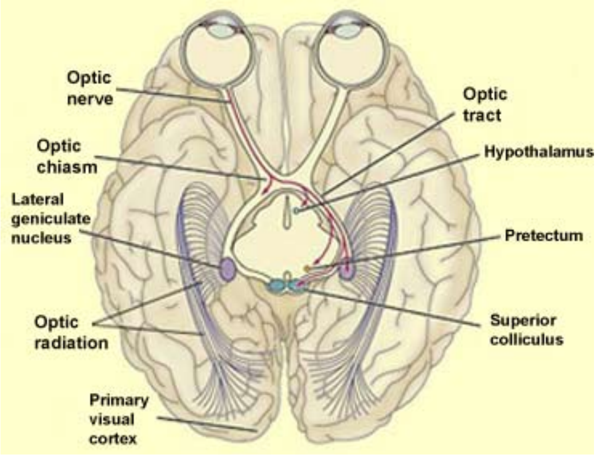
The eye perceives only a small part of the world at any instant. However, the muscles rotate the eyes at

The optical nerves leave the retina and are joined at Optic Chiasm.
Nerves then branch off to the Lateral Geniculate Nucleus (LGN) and the Superior Colliculus.

Nerves branch out from the LGN to provide "retinal maps" to the different visual cortexes as well as the "Superior Colliculus".

Surprisingly, 80% of the excitation of the LGN comes from the visual cortex!
The LGN seems to act as a filter for visual attention.

In fact, the entire visual system can be seen as succession of filters.



3.3 The Superior Colliculus

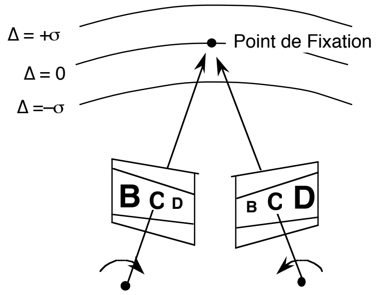
The first visual filter is provided by fixation, controlled by the Superior Colliculus.
The Superior Colliculus is a Feed-Forward (predictive) control system for binocular fixation. The Superior Colliculus is composed of 7 layers receiving stimulus from the frontal cortex, the lateral and dorsal cortexes, the auditory cortex and the retina.

3.4 Vergence and Version

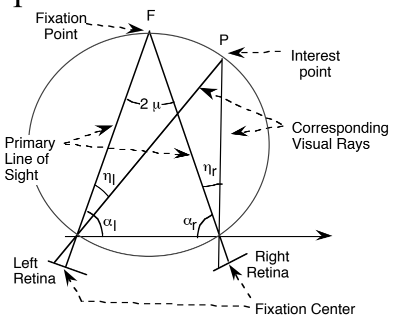
At any instant, the human visual system focuses processing on a small region of 3D space called the Horopter.

The horopter is mathematically defined as the region of space that projects to the same retinal coordinates in both eyes. The horopter is the locus of visual fixation.

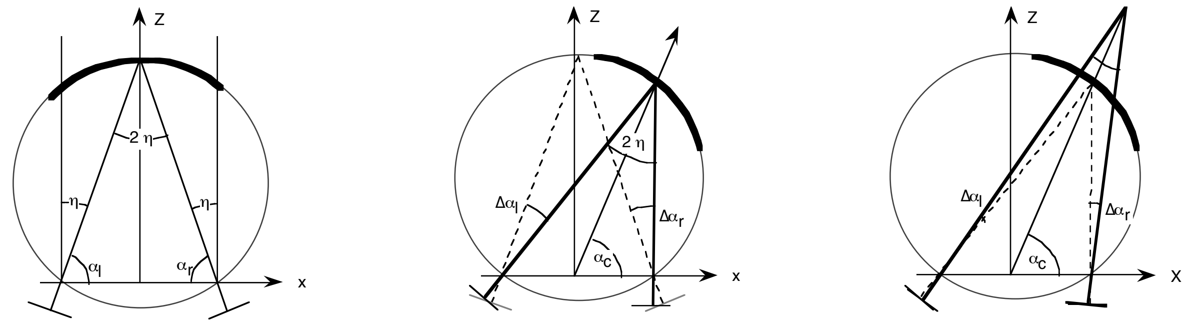
The horopter is controlled by the Superior Colliculus, and can move about the scene in incredibly rapid movements (eye scans). Scanning the horopteur allows the cortex to build up a composite model of the external world.



Eye movements can decomposed into "Version" and "Vergence".
Version perceives relative direction in head centered coordinates.
Vergence perceives relative depth.



Vergence and version are described by the Vief-Muller Circle.
Version (angle) is the sum of the eye angles.
Vergence (depth) is proportional to difference.

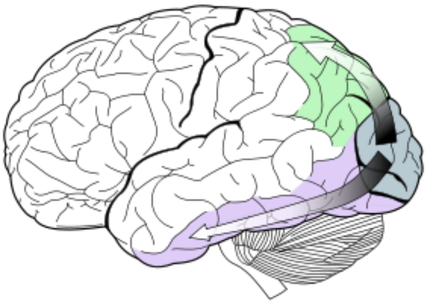


Symmetric Vergence   Vergence   Version

Vergence and version are redundantly controlled by retinal matching and by focusing of the lenses in the eyes (accommodation).

3.5 The Visual Cortex

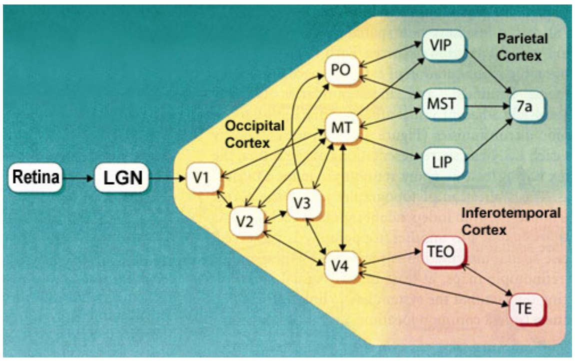
Retinal maps are relayed through the LGN to the primary visual cortex, where they propagate through the Dorsal and Lateral Visual pathways.

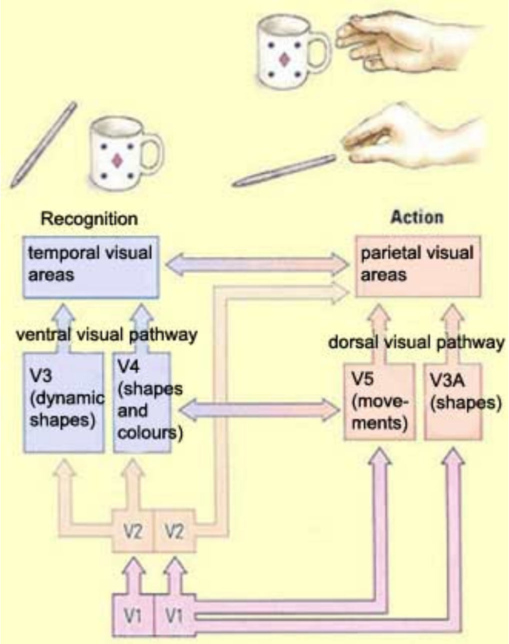


Dorsal visual pathway (green) is the "action pathway".
It controls motor actions. Most of the processing is unconscious.
It makes use of spatial organization (relative 3D position), including depth and direction information from the Superior Colliculus.

The ventral visual pathway (purple) is used in recognizing objects.
It makes use of color and appearance.

These two pathways are divided into a number of interacting subsystems (visual areas).





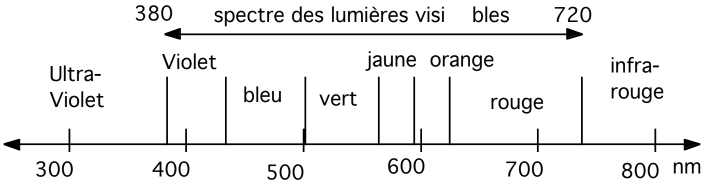
Most human actions require input from both pathways. For example, consider the task of grasping a cup. The brain must recognize and locate the cup, and direct the hand to grasp the cup.

4 Color Spaces and Color Models

4.1 Color Perception

The human retina is a tissue composed of rods, cones and bi-polar cells.
Cones are responsible for daytime vision.
Bi-polar cells perform initial image processing in the retina.

Rods provide night vision. Night vision is achromatique. It does not provide color perception. Night vision is low acuity - Rods are dispersed over the entire retina.

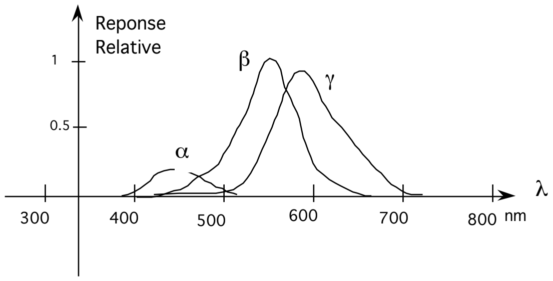


Rods are responsible for perception of very low light levels and provide night vision. Rods employ a very sensitive pigment named "rhodopsin".

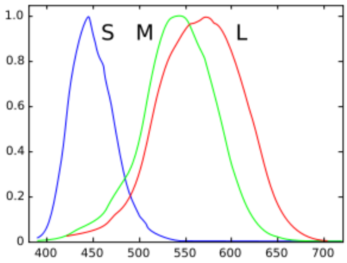
Rodopsin is sensitive to a large part of the visible spectrum of with a maximum sensitivity around 510 nano-meters.

Rhodopsin sensitive to light between 0.1 and 2 lumens, (typical moonlight) but is destroyed by more intense lights.

Rhodopsin can take from 10 to 20 minutes to regenerate.



Relative Sensitivities

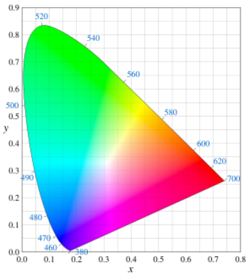


Normalised Sensitivities

Cones provide our chromatique "day vision". Human Cones employ 3 pigments :

cyanolabe α 400–500 nm peak at 420–440 nm
chlorolabe β 450–630 nm peak at 534–545 nm
erythrolabe γ 500–700 nm peak at 564–580 nm

Perception of cyanolabe is low probability, hence poor sensitivity to blue.
Perception of Chlorolabe and erythrolabe are more sensitive.



The three pigments give rise to a color space shown here (CIE model).

Note, these three pigments do NOT map directly to color perception.
Color perception is MUCH more complex, and includes a difficult to model phenomena known as "color constancy".

For example, yellow is always yellow, despite changes to the spectrum of an ambiant source

Many color models have been proposed but each has its strengths and weaknesses.