

Computer Vision CITS4240

School of Computer Science & Software Engineering
The University of Western Australia

Introduction

Image formation in the eye and the camera

Biological vision is the process of using light reflected from the surrounding world as a way of modifying behavior. Generally, with humans, we say that the surrounding environment is *interpreted* by visual input. This usually implies some form of conscious understanding of the 3D world from the 2D projection that it forms on the retina of the eye. However much of our visual computation is carried out unconsciously and often our interpretations can be fallacious.

In this section we will briefly overview the human visual system and try to understand the ways in which this system uses *computation* as a means of interpreting its input. To do this, we will also try to understand how images are formed on cameras, how images are stored in computers, and how computations can be carried out efficiently. Although not strictly correct, this analogy between machine vision and biological vision is currently the best model available. Moreover, the models interact in an ever increasing fashion: we use the human visual system as an existence proof that visual interpretation is even possible in the first place, and its response to optical illusions as a way to guide our development of algorithms that replicate the human system; and we use our understanding of machine vision and our ability to generate ever more complex computer images as a way of modifying, or evolving, our visual system in its efforts to interpret the visual world. For example, image morphing scenes would be difficult to interpret for the naive viewer.

The eye

Any understanding of the function of the human eye serves as an insight into how machine vision might be solved. Indeed, it was some of the early work by Hubel and Wiesel [5] on the receptive fields in the retina that has led to the fundamental operation of spatial filtering that nowadays dominates so much of early image processing. There are many good references to the function of the eye, although Frisby [2] gives an excellent overview with a computational flavour.

The eye is considered by most neuroscientists as actually part of the brain. It consists of a small spherical globe of about 2cm in diameter, which is free to rotate under the control of 6 extrinsic muscles. Light enters the eye through the transparent *cornea*, passes through the *aqueous humor*, the *lens*, and the *vitreous humor*, where it finally forms an image on the *retina* (see Figure 1).

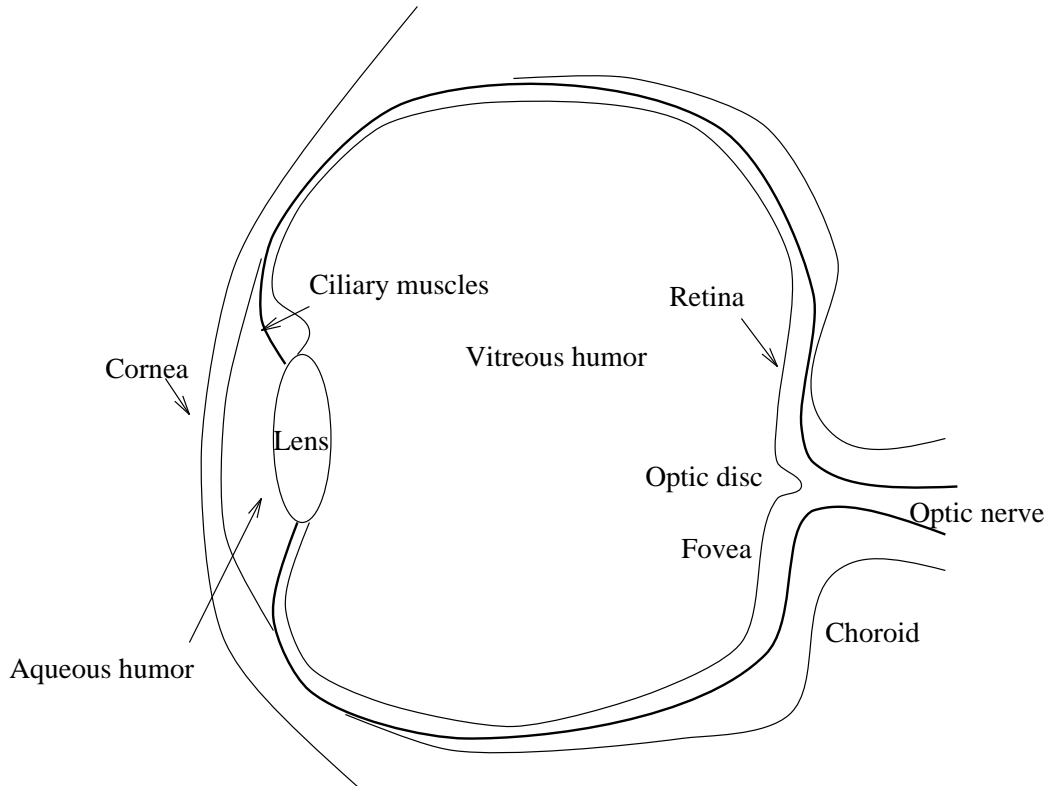


Figure 1: Sketch of a cross-section of the eye (schematic after Nalwa).

It is the muscular adjustment of the lens, known as *accommodation*, that focuses the image directly on the retina. If this adjustment is not correctly accomplished, the viewer suffers from either nearsightedness or farsightedness. Both conditions are easily corrected with optical lenses.

The retina itself is a complex tiling of photoreceptors. These photoreceptors are known as *rods* and *cones*. When these photoreceptors are stimulated by light, they produce electrical signals that are transmitted to the brain via the *optic nerve*. The location of the optic nerve on the retina obviously prohibits the existence of photoreceptors at this point. This point is known as the *blind spot* and any light that falls upon it is not perceived by the viewer. Most people are unaware of their blind spot, although it is easy to demonstrate that it exists. And its existence has been known about for many years as it is reputed that the executioners in France during the revolution used to place their victim so that his or her head fell onto the blind spot, thus illiciting a pre-guillotine perception of the poor character without a head.

The rods and cones do not have a continuous physical link to the optic nerve fibres. Rather, they communicate through three distinct layers of cells, via junctions known as *synapses*. These layers of cells connect the rods and cones to the *ganglion cells*, which respond to the photostimulus according to a certain *receptive field*. We can see from Figure

2 that the rods and cones are at the *back* of the retina. Thus the light passes through the various cell layers to these receptive fields, and is then transmitted via various synaptic junctions back towards the optic nerve fibre.

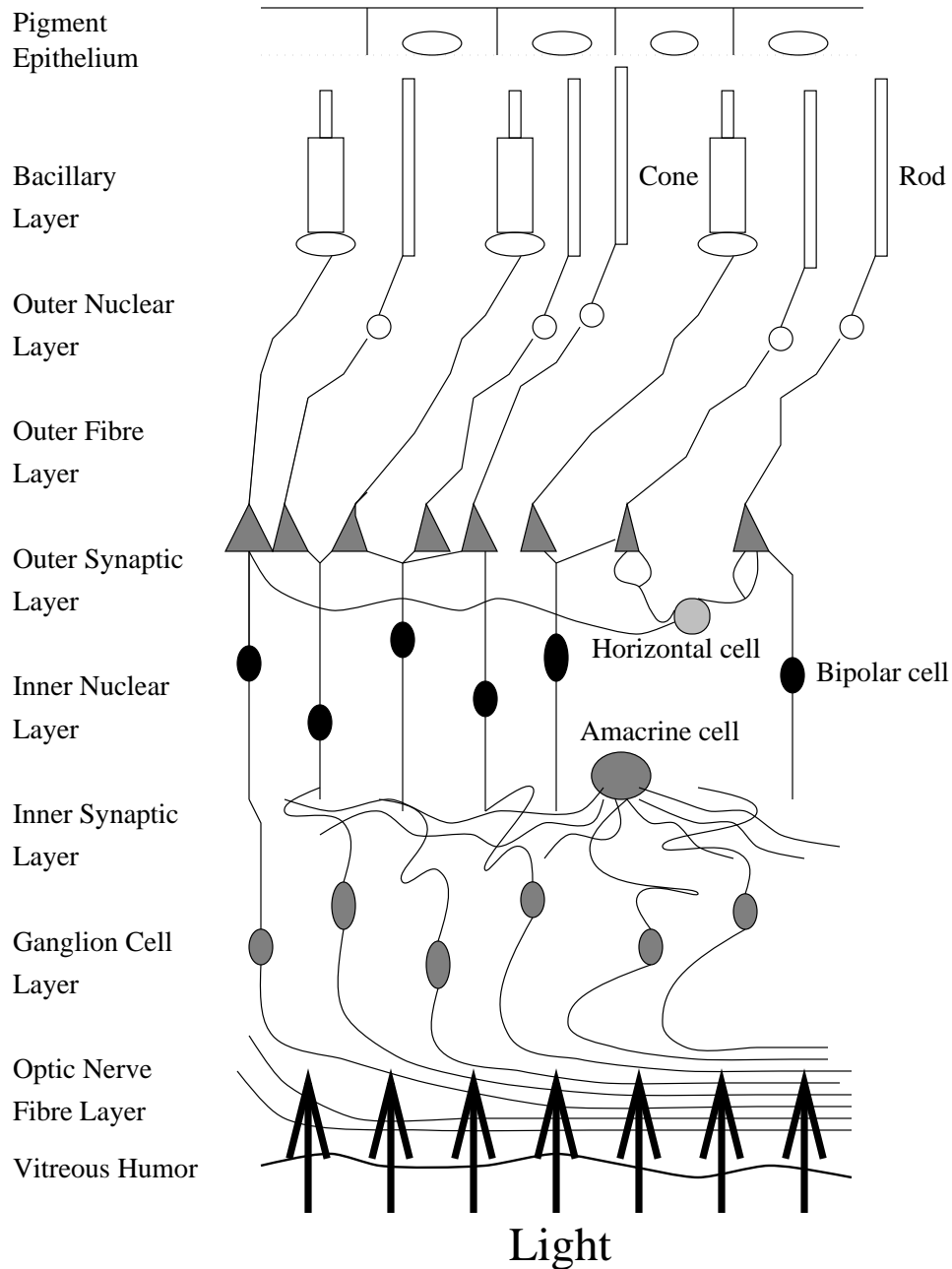


Figure 2: Schematic of a cross section of the retina (after Nalwa).

It has long been known that the spatial organization of some of the receptive fields of the retinal ganglion cells is circularly symmetric, with either an excitatory central region and an inhibitory surround, or with an inhibitory centre and an excitatory surround. Such cells are known to computational vision researchers as “mexican-hat operators”. Their existence inspired Marr [6] to develop a computational approach to physiological vision understanding, since their response to a light signal was analogous to the convolution of the signal with the second derivative of a Gaussian. Marr’s theory of edge detection will be explored in a later lecture.

However, there are also other spatial organisations of receptive fields. There are orientationally selective receptive fields, with an excitatory lobe to one side and an inhibitory lobe on the other, so as to form an antisymmetric field. These cells exist over a wide variety of orientations. Both the even and odd symmetric receptive fields are known as *simple cells*. They respond strongly to luminance edges and line features in images, respectively.

Other types of receptive fields, known as *complex cells*, are also found in the retina. Their behaviour is more complex, combining in a non-linear fashion the responses of the even and odd symmetric filter responses. And *end-stopped* cells appear to act as simple differentiation operators. How the behaviour of these cells might be combined into a single computational model will be outlined in a future lecture.

The human eye is a remarkable organ, whose sensitivity and performance characteristics approach the absolute limits set by quantum physics. The eye is able to detect as little as a single photon as input, and is capable of adjusting to ranges in light that span many orders of magnitude. No camera has been built that even partially matches this performance.

Very little is known about what happens to the optic signal once it begins its voyage down the optic nerve (see Figure 3). The optic nerve has inputs arriving from both the left and right sides of both eyes, and these inputs split and merge at the *optic chiasma*. Moreover, what is seen by one eye is slightly different from what is seen by the other, and this difference is used to deduce *depth* in stereo vision. From the optic chiasma, the nerve fibres proceed in two groups to the *striate cortex*, the seat of visual processing in the brain. A large proportion of the striate cortex is devoted to processing information from the fovea.

Camera models

To understand how vision might be modeled computationally and replicated on a computer, we need to understand the image acquisition process. The role of the camera in machine vision is analogous to that of the eye in biological systems.

Pinhole camera model

The *pinhole camera* is the simplest, and the ideal, model of camera function. It has an infinitesimally small hole through which light enters before forming an inverted image on the camera surface facing the hole. To simplify things, we usually model a pinhole camera by placing the image plane *between* the focal point of the camera and the object, so that the image is not inverted. This mapping of three dimensions onto two, is called a *perspective projection* (see Figure 4), and perspective geometry is fundamental to any understanding of image analysis.

Perspective geometry

Euclidean geometry is a special case of perspective geometry, and the use of perspective geometry in computer vision makes for a simpler and more elegant expression of the

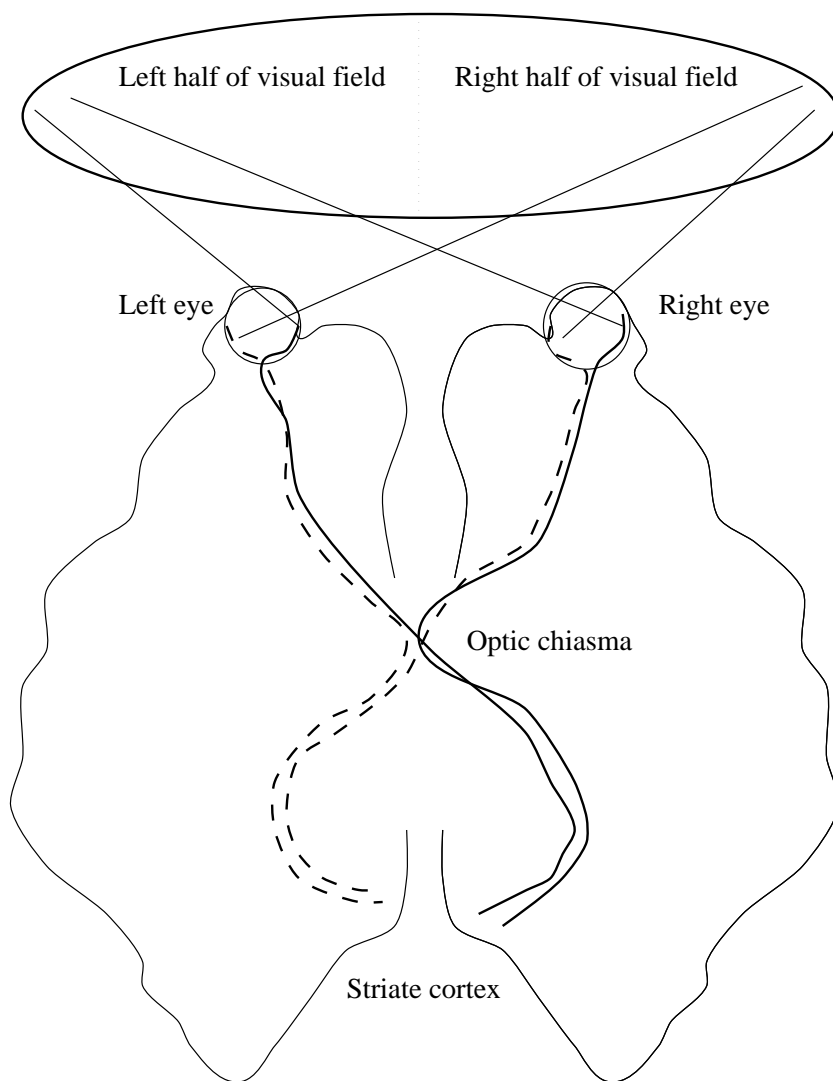


Figure 3: Overview of the visual pathways from eyes to striate cortex (schematic after Nalwa).

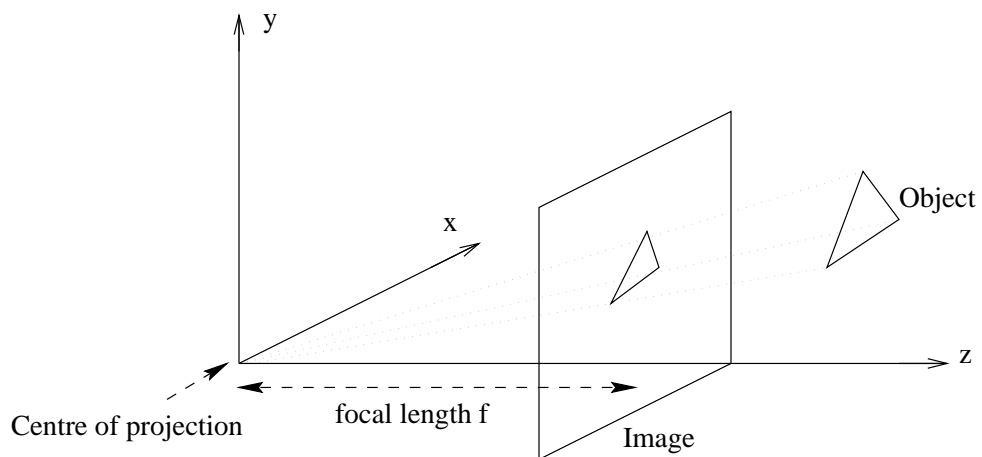


Figure 4: Perspective projection in the pinhole camera model.

computational processes that render vision possible. A superb overview of the geometric viewpoint in computer vision is given by Faugeras [1].

A perspective projection is the projection of a three-dimensional object onto a two-dimensional surface by straight lines that pass through a single point. Simple geometry shows that if we denote the distance of the image plane to the centre of projection by f , then the image coordinates (u, v) are related to the object coordinates (x, y, z) by

$$u = (f/z)x \text{ and } v = (f/z)y.$$

These equations are non-linear. They can be made linear by introducing *homogeneous transformations*, which is effectively just a matter of placing the Euclidean geometry into the perspective framework. The linear version is simply

$$\begin{bmatrix} U \\ V \\ S \end{bmatrix} = \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}.$$

Here $u = U/S$ and $v = V/S$, if $S \neq 0$.

Thus, in the general projective representation, each point in the n -dimensional projective space is represented by an $n + 1$ vector (Sx_1, \dots, Sx_n, S) , where $S \neq 0$.

Simple lens model

In reality, one must use lenses to focus an image onto the camera's focal plane. The limitation with lenses is that they can only bring into focus those objects that lie on one particular plane that is parallel to the image plane. Assuming the lens is relatively thin and that its optical axis is perpendicular to the image plane, it operates according to the following *lens law*:

$$\frac{1}{u} + \frac{1}{v} = \frac{1}{f},$$

where u is the distance of an object point from the plane of the lens, v is the distance of the focussed image from this plane, and f is the focal length of the lens (see Figure 5).

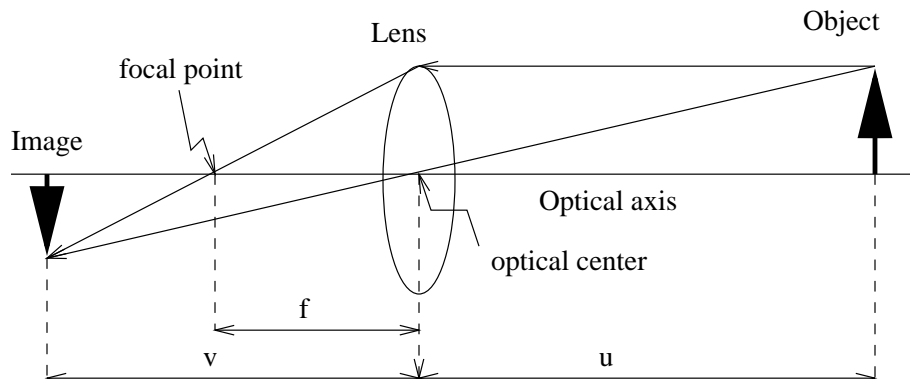


Figure 5: Simple lens model.

Camera calibration

To deduce three-dimensional geometric information from an image, one must determine the parameters that relate the position of a point in a scene to its position in the image. This is known as *camera calibration*. Currently this is a cumbersome process of estimating the *intrinsic* and *extrinsic* parameters of a camera. There are four intrinsic camera parameters: two are for the position of the origin of the image coordinate frame, and two are for the scale factors of the axes of this frame. There are six extrinsic camera parameters: three are for the position of the center of projection, and three are for the orientation of the image plane coordinate frame. However, recent advances in computer vision indicate that we might be able to eliminate this process altogether. These new approaches will be discussed in a later lecture.

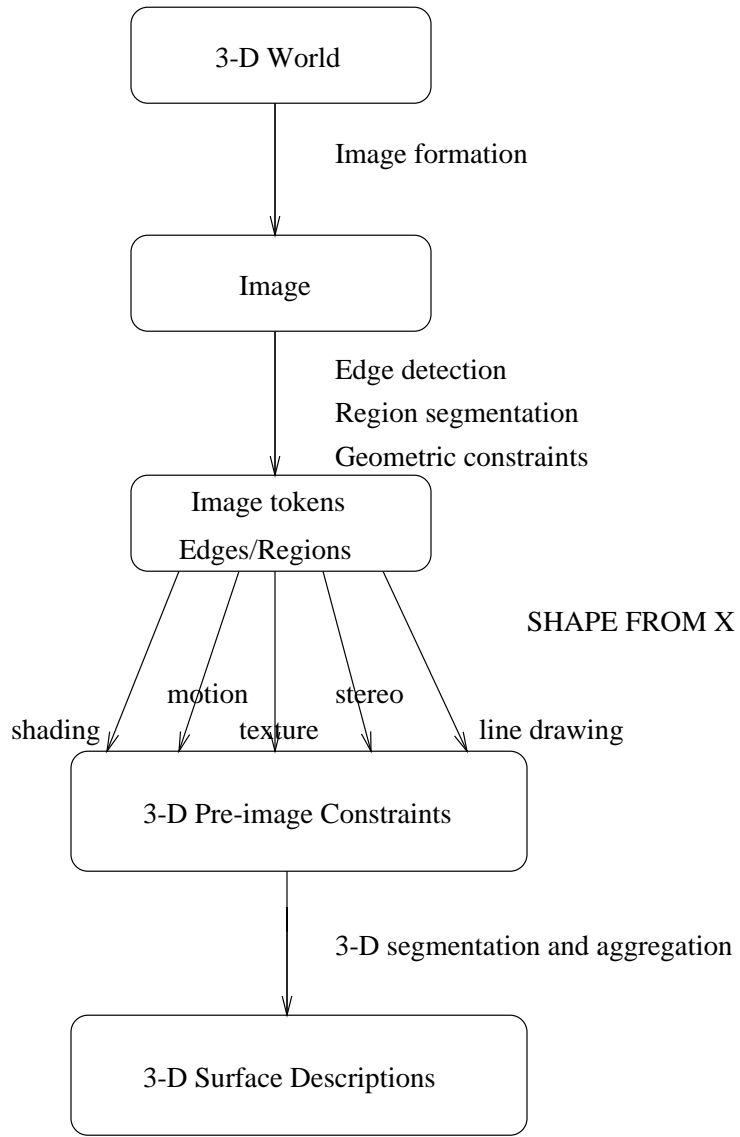
The current model of visual computation

Because very little is known about how the brain processes visual information, we will instead look at the computer-vision paradigm and highlight methods of computation that are biologically inspired. Much that is done in computer vision has an engineering approach - that is, find a solution, regardless of how it relates to any biological system. However, there are some techniques in computer vision that are inspired by what is known about how biological systems function, and it is these analogies that we shall draw upon in the following lectures.

The next five lectures will concentrate on classical image processing techniques, both in the spatial and the frequency domain. We will begin with binary images, where the mathematical modelling is simplest. Most industrial applications of computer vision still attempt to extract as much information as possible from an impoverished binary image, and indeed much can be done with this data. Thus, lecture 2 will cover the geometrical and topological properties of binary images and closely follows chapters 2 and 3 of Horn's "Robot Vision" [4]. Continuing with image topology in lecture 3, we will also study connected components and the theory of morphology, as developed by Serra [8]. Gonzalez and Woods' "Digital Image Processing" [3] covers this work in Chapter 8. Fourier theory is introduced in lecture 4, and then the image enhancement techniques developed in lectures 5 and 6 are also covered by Gonzalez and Woods.

In the computer-vision paradigm (see Figure 6), an image is first acquired and then some form of data-reduction is performed to help in the subsequent analyses. This in itself is biologically inspired: the retina has something like 120 million rods and 5 million cones, but the total number of nerve fibres leaving the eye is only about 1 million. Some form of compression must be applied very early in visual processing; the compressed format must be high in information, and low in redundancy, whereas we know that raw image data is highly redundant.

Most researchers agree that the early stages in vision perform some sort of *edge detection* (lecture 7). In lectures 8–10 we will study one particular model of edge detection, known as the *local energy model*. Aspects of this model that are biologically inspired will be outlined, along with their mathematical foundations.



Recognition and Prediction

Figure 6: Current model for computer vision.

Following the data compression stages, image interpretation needs to be undertaken. How do we infer information about the 3D world from mere 2D projections? Mathematicians would say that the problem is ill-posed, in the sense that there is an infinite number of solutions or interpretations, but the physical constraints imposed by the real world limit these interpretations substantially. Thus, after the compression stage comes a series of modular, but perhaps interlinked, stages of what is known as *Shape-from-X*. Examples include shape from stereo, shape from motion, shape from texture, or shape from shading. In all these cases, additional information renders an ill-posed problem soluble, and allows us to interpret the 3D geometry of the scene under view.

Having obtained the 3D geometry of the scene under view, we then need to interpret it, or match it with our existing knowledge of the world. We understand what it is we see, even when we see that object from a completely novel view. In order to match the geometry of the objects in our scene with that of the objects in our known database, we use *invariants*, that is geometrical descriptions of the object that are invariant to the types of transformations that the visual process performs.

In lectures 11–13 we will consider the problems of calibration, stereo and motion, and comment upon the biologically plausible nature of their solution. We will also briefly outline the recognition process that then follows, given either the 3D data constructed by the stereo module, or using geometric information directly available from the 2D scenes.

References

- [1] Olivier Faugeras. *Three-Dimensional Computer Vision: A Geometric Viewpoint*. M.I.T. Press, 1993.
- [2] J. P. Frisby. *Seeing: Illusion, Brain and Mind*. Oxford University Press. Walton Street, Oxford, 1979.
- [3] Rafael C. Gonzalez and Richard E. Woods. *Digital Image Processing*. Addison-Wesley Publishing Company, 1992.
- [4] Berthold Klaus Paul Horn. *Robot Vision*. MIT Press, 1986.
- [5] D. H. Hubel and T. N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology*, London, Vol. 160, 106-154, 1962
- [6] D. Marr. *Vision*. Freeman, 1982.
- [7] Vishvjit S. Nalwa. *A Guided Tour of Computer Vision*. Addison-Wesley Publishing Company, 1993.
- [8] J. Serra. *Image Analysis and Mathematical Morphology*. Academic Press, 1982.