Computer Vision

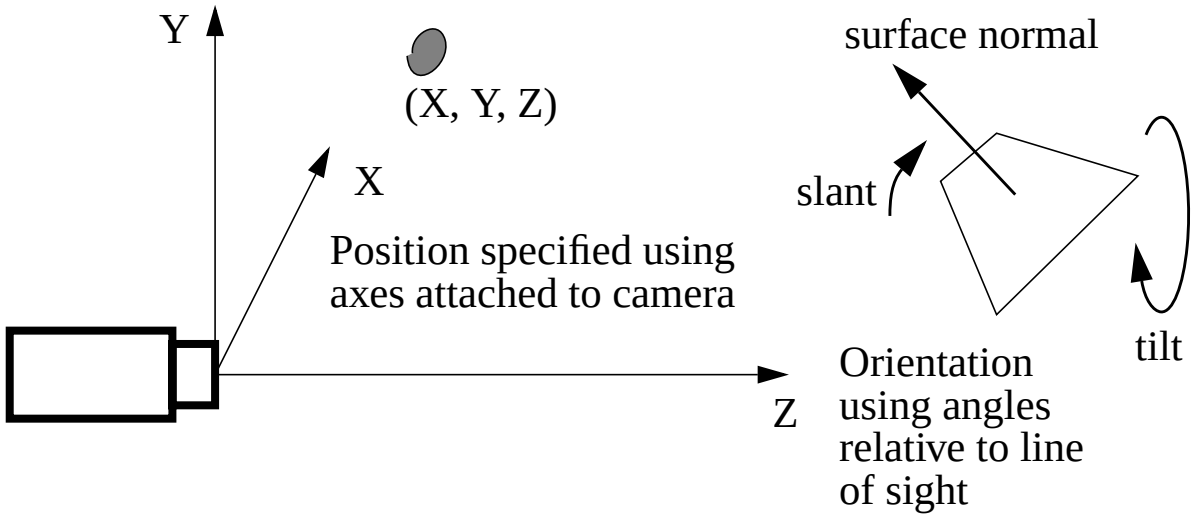
Lecture 8: Introduction to 3-D Vision

So far we have looked at methods for analysing 2-D images.

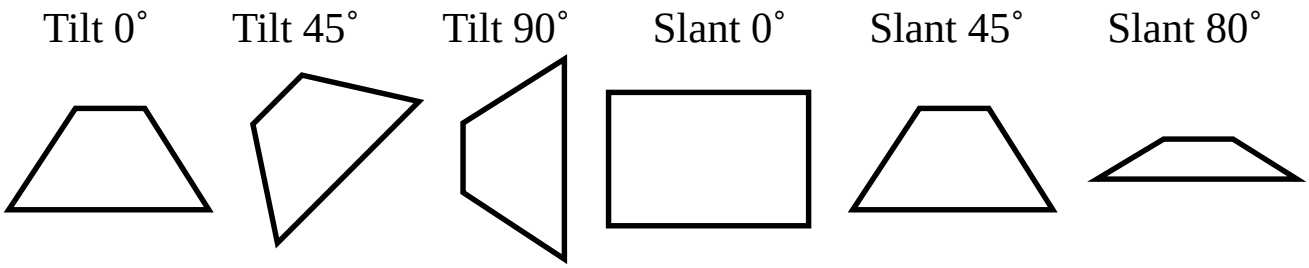
Vision is concerned with obtaining the information needed to interact successfully with a 3-D environment.

Layout

The layout of a scene means the positions and orientations of surfaces, usually described relative to the vision system's frame of reference.



Camera view:



The role of layout representations

Obtaining a good representation of layout used to be taken as the defining goal of intermediate-level vision (e.g. Marr's 2-1/2 D sketch).

Explicit representations of layout have been used very successfully with mobile robots.

But Marr also pointed to work by Reichardt and Poggio on the landing behaviour of flies:

> "... it is extremely unlikely that the fly has an explicit representation of the visual world around him — no true conception of a surface, for example, but just a few triggers and some specifically fly-centred parameters ..."

It remains unclear how much of the spatial layout of our own environment is explicitly represented by our own visuo-motor systems.

In the field of *ecological psychology* an important idea is that potential *actions* are used to represent our environment. For example, what matters in crossing the road is not the *(X,Y,Z)* coordinates of a car, but whether it will hit us before we reach the other side. This ties in with Gibson's concept of *affordances* as the central objects of perception.

Vision research has moved towards a task-oriented view, but has not abandoned the need to extract 3-D information — this has just become part of a larger picture. For a good discussion see the introduction to *Active Perception* edited by Y. Aloimonos (in bibliography on web page).

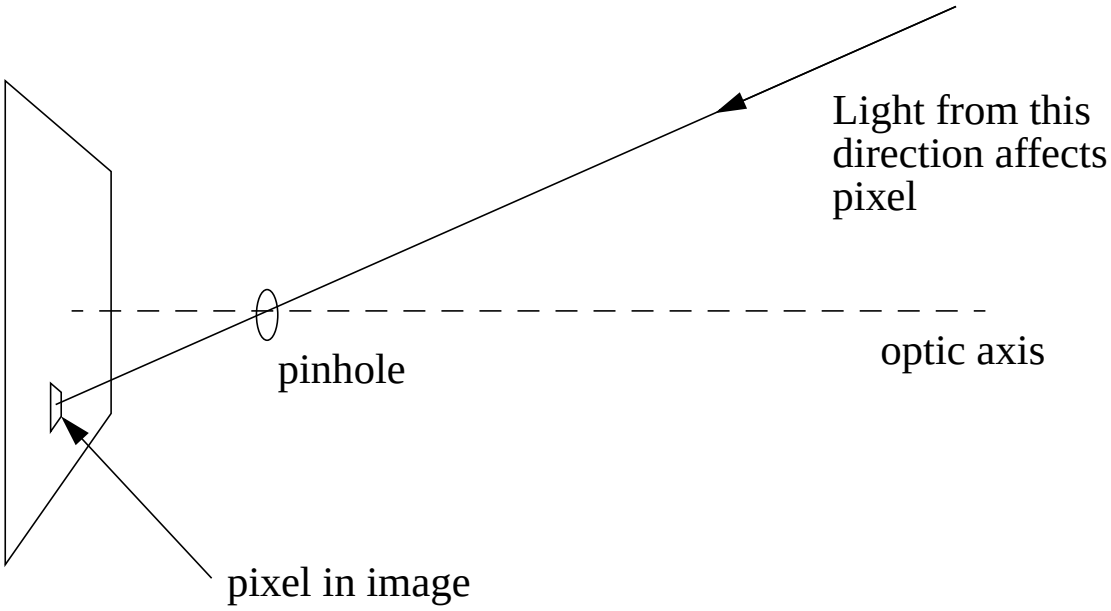
Without prejudice to our overall view, we will look at ways of obtaining 3-D information from images.

The geometry of images and objects

Lecture 2 revisited. Remember that the *optic array* is the fundamental description of the input to the visual system.

An image pixel corresponds to a *direction in space*.

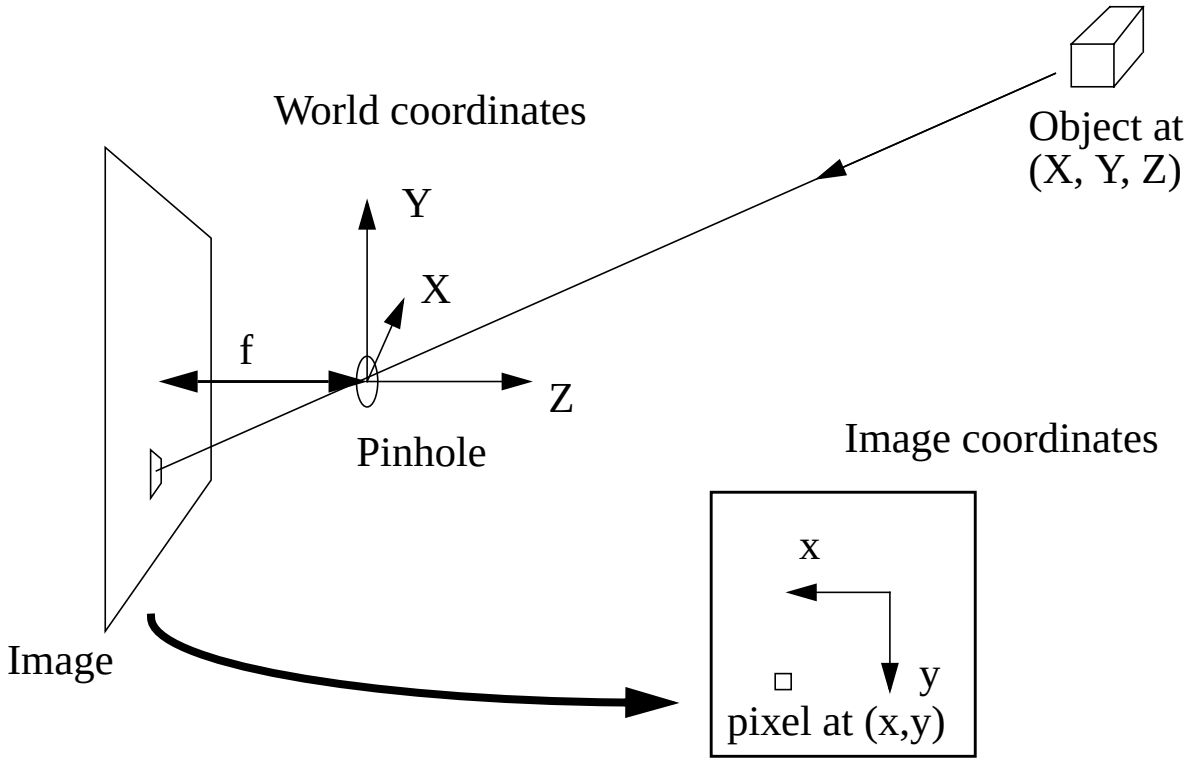
A *pinhole camera* demonstrates this.



If the camera is *calibrated* we know the mapping from each pair of pixels coordinates to the corresponding direction.

Perspective projection

The pinhole camera is quite a good model of real cameras used in computer vision. (The biggest difference is that real cameras can be out of focus. They may also distort the image in various ways, but corrections can be made.)



The equations of perspective projection are those for a pinhole camera:

$$x = \frac{Xf}{Z}$$

$$y = \frac{Yf}{Z}$$

Here x and y are not the same as column and row — it may be necessary to change sign and apply an offset.

Reverse projection

For the pinhole camera, *calibration* means knowing where in the image the origin of x and y is, and the value of f.

The direction in space is given by

$$\frac{X}{Z} = \frac{x}{f}$$

$$\frac{Y}{Z} = \frac{y}{f}$$

General cameras

For real cameras, the sequence is still

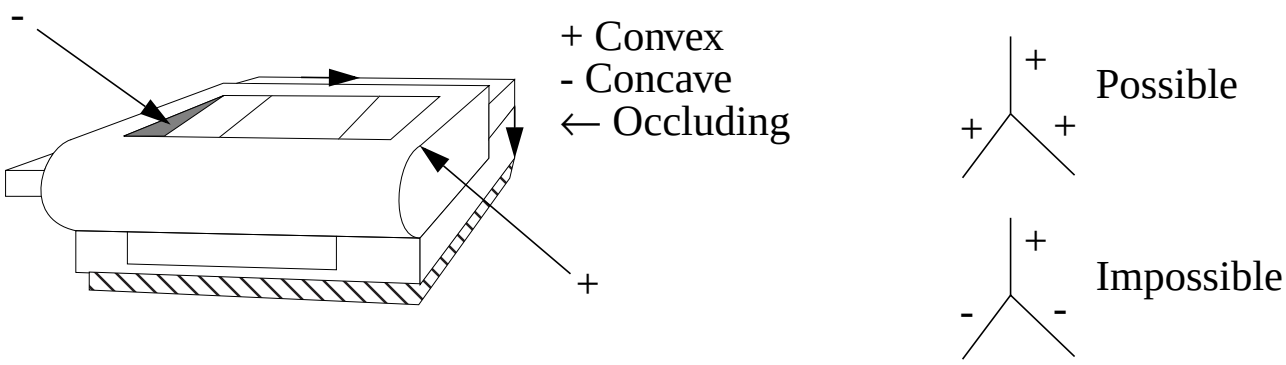
Row and column of pixel → x and y image coordinates → Direction in space

with the first step using camera calibration and the second using projection equations.

Monocular methods for depth

Line labelling

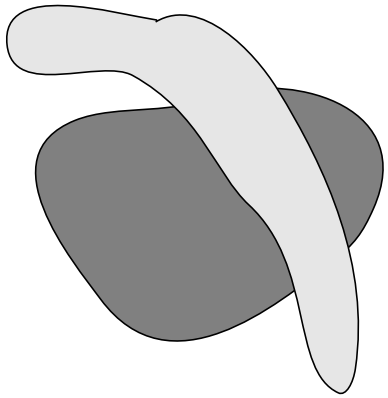
Classify edges into convex, concave, occluding to get *qualitative* shape information.



Assumes very high quality image analysis.

Has contributed to AI via the search technique *Waltz filtering*, which finds consistent interpretations.

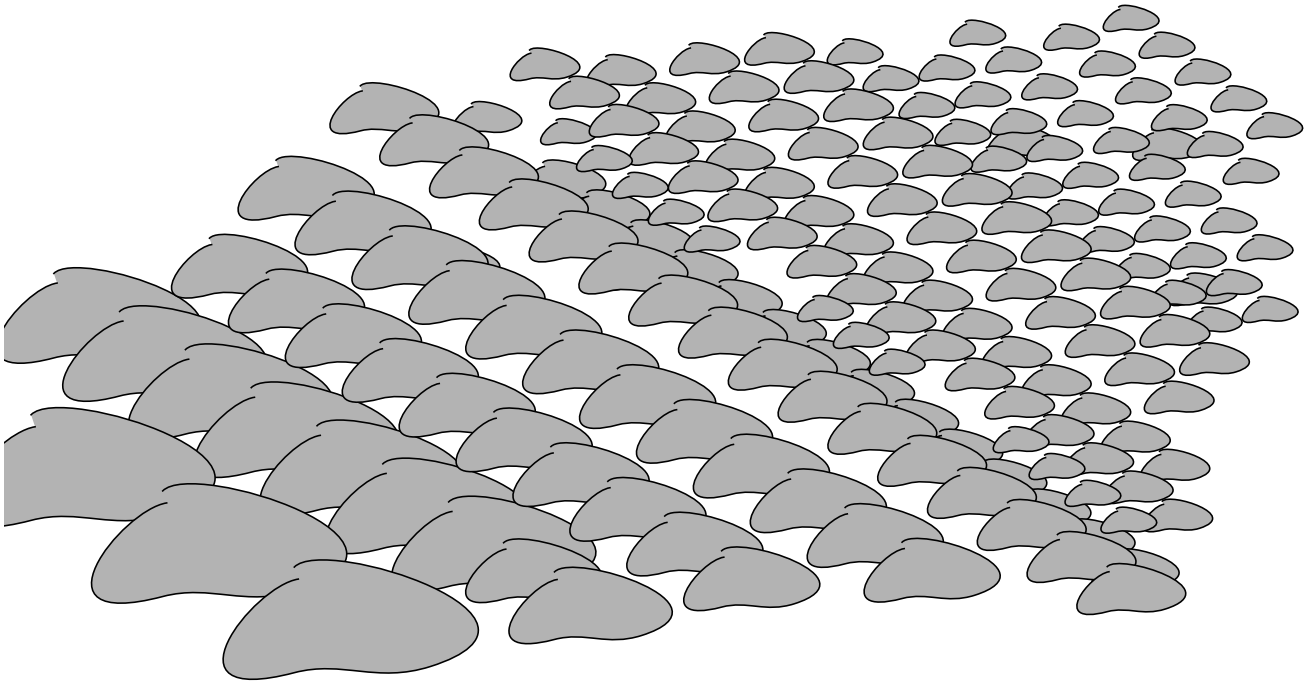
The important cue of *occlusion* is closely related.



We assume that we are viewing from a *general viewpoint* — not accidentally lining objects up.

Texture gradients

Changes in texture *density* can be powerful cues.

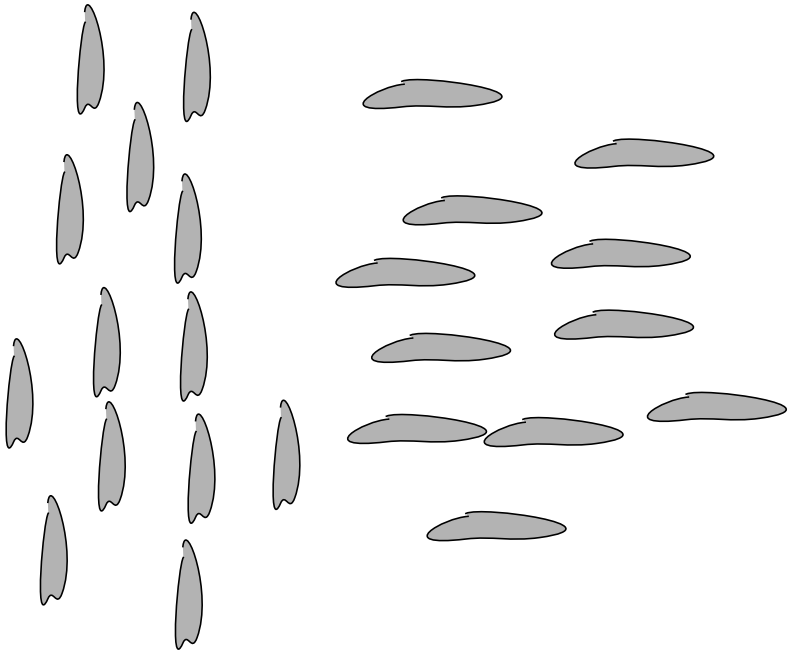


Estimating texture density is difficult: filtering to detect edges can destroy small scale texture.

We assume that the physical source of the texture elements is uniform.

Texture orientation

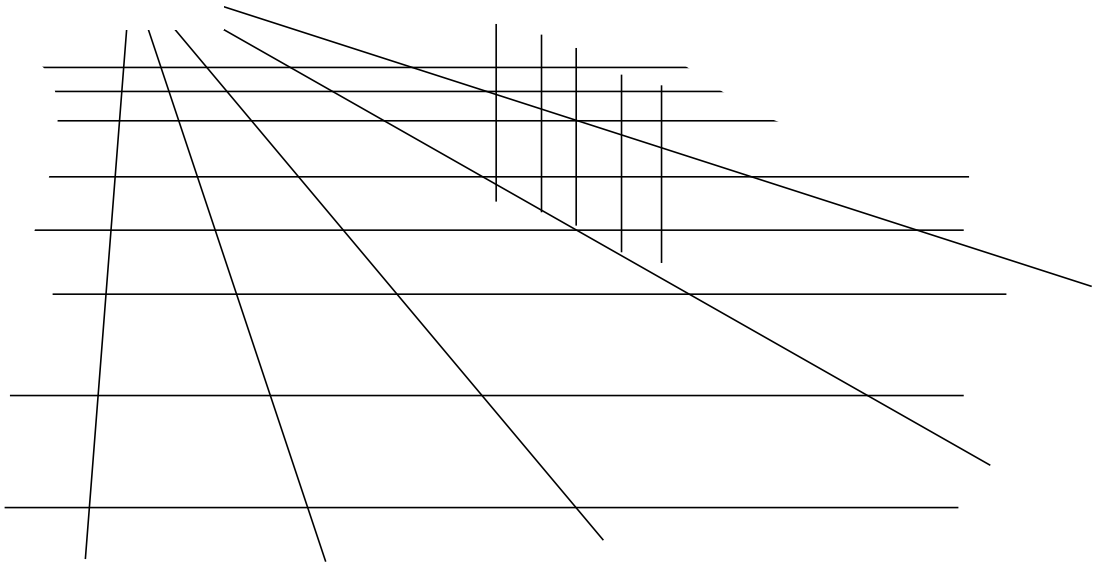
The way texture elements are oriented can give information about surface slant.



We assume that the physical source of the texture elements has no preferred orientation, and that foreshortening has produced the strongly linear textures.

Linear perspective

Parallel lines in space project to fans of lines in the image.

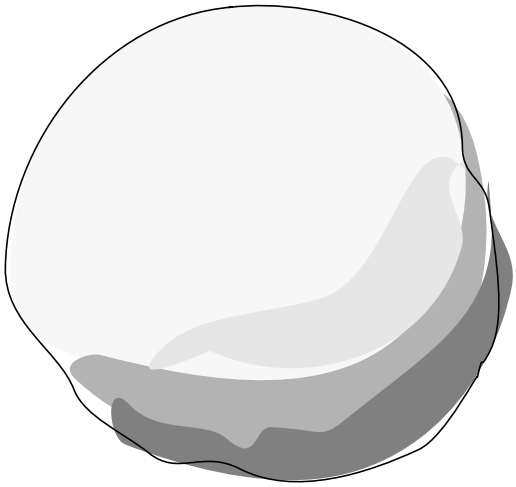


Such families of lines diverge from a *vanishing point*. Computer vision systems can use, for example, the Hough transform to detect vanishing points and to infer 3-D structure, assuming that the lines really are parallel in space.

Shading

The distribution of light and dark can be used to infer shape. Often the amount of light reaching the camera is greatest when a surface is oriented towards the light source (though the details are complex and depend on the surface).

It is necessary to assume that the variation of light and dark in the image is caused by the orientation variation and not by the markings on the surface.

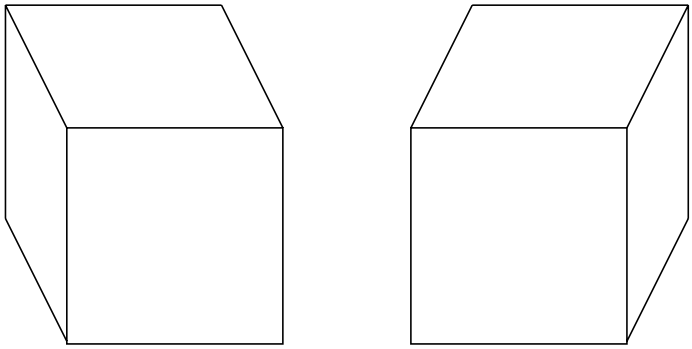


Using multiple viewpoints for layout

Using more than one camera position gives powerful geometrical information about position.

Methods divide into two closely related families:

Stereo. This is short for stereopsis or stereoscopic vision. It involves using more than one camera (or eye) simultaneously.



Motion. One or more cameras capture views from different positions as they move about.