Computer Vision

Lecture 9: Stereoscopic vision

Last lecture

- Layout, camera coordinates, slant and tilt
- The role of layout representations
- The geometry of image formation: perspective projection
- Monocular methods for depth

This lecture

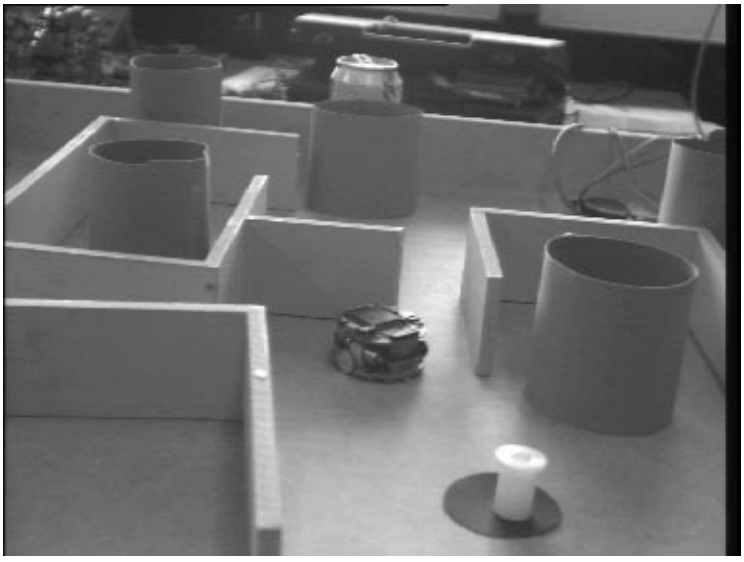
- Seeing with two eyes
- Stereo geometry
- The correspondence problem

The basis of stereo

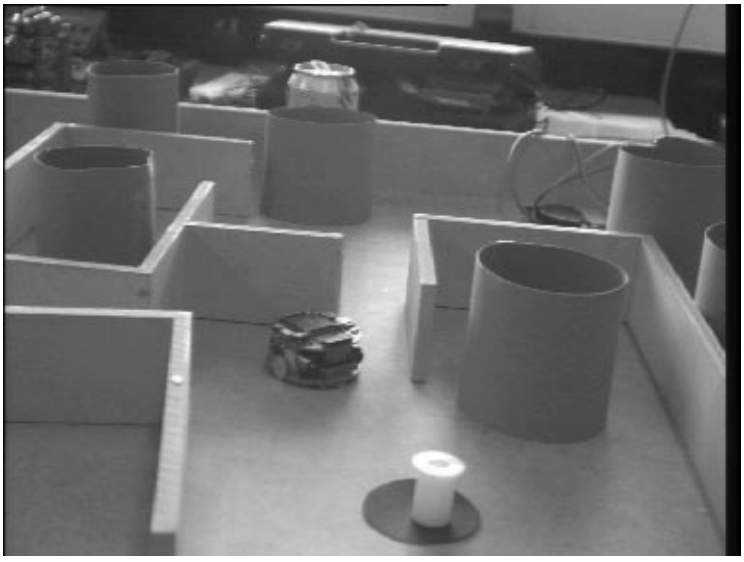
Most of us fuse the images from our eyes in order to obtain depth information.

This is exploited in the stereo viewer, where photographs taken from two positions are presented one to each eye, and a subjective sense of depth is produced.

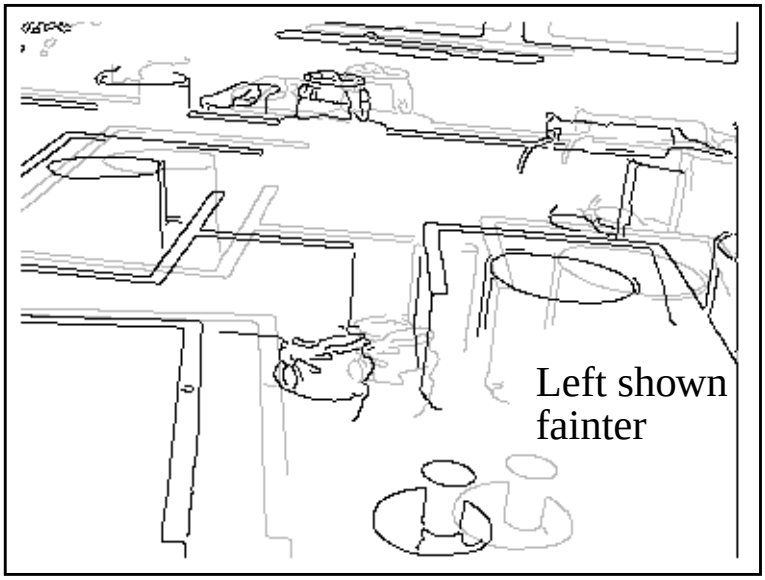
Left



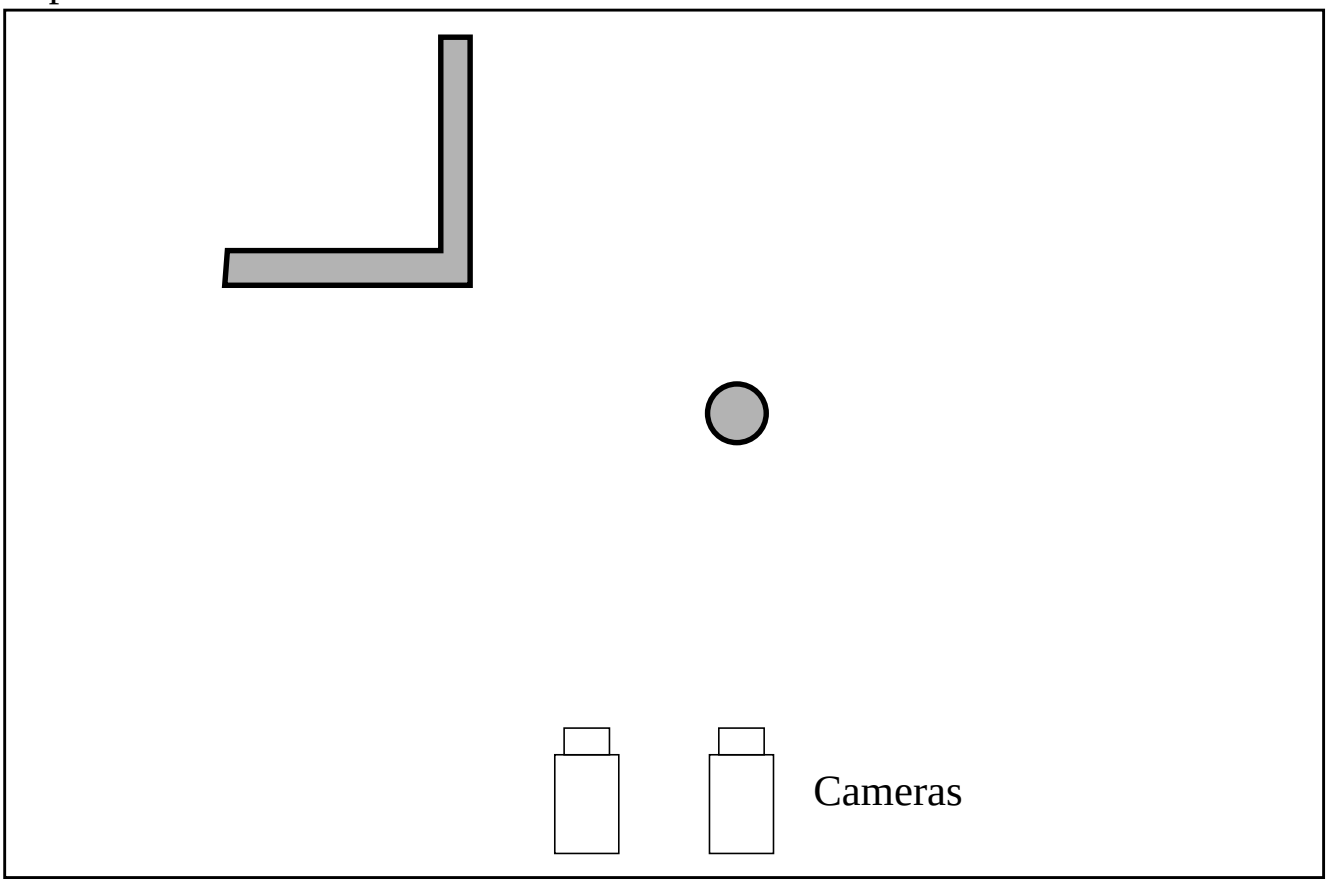
Right



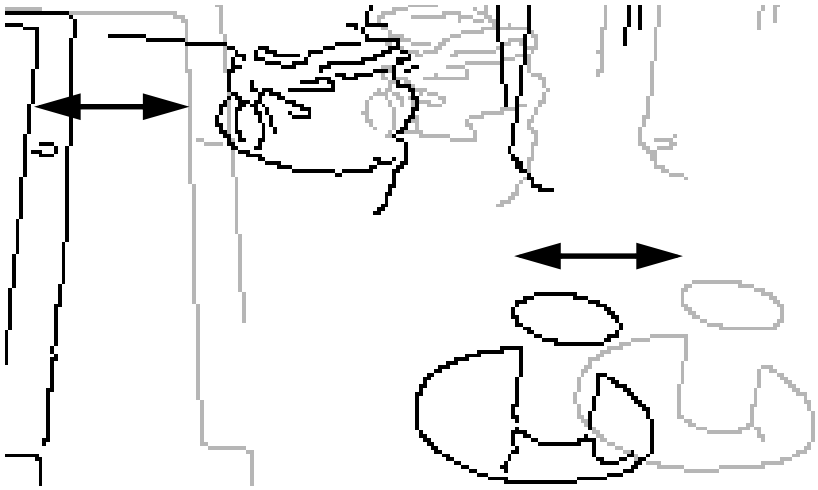
Superimposing the edges from the two views shows what has happened:



In plan view:



The separation between two matching objects is called the *stereo disparity.*

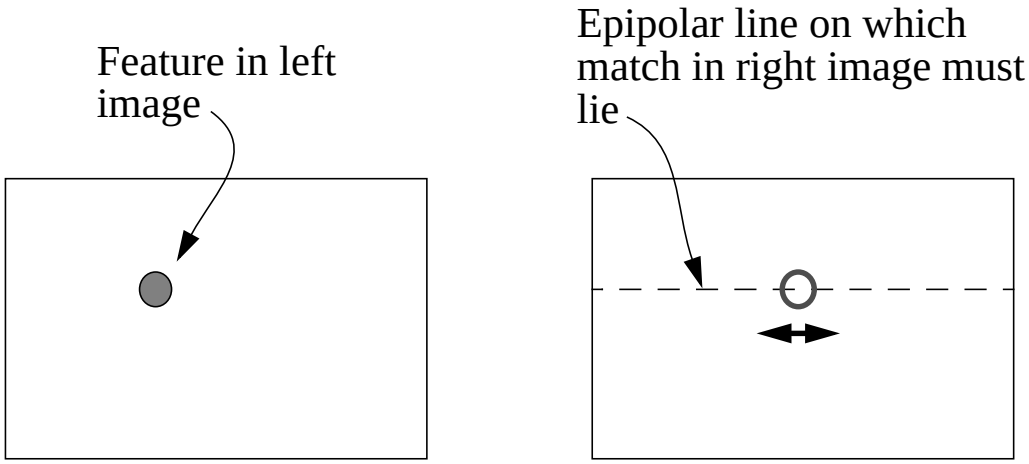


Disparity is measured in pixels and can be positive or negative (conventions differ). It will vary across the image.

For two close cameras, side by side and pointing in the same direction, matching points are roughly on the same row in each image — the disparity is *horizontal*.

The line on which a match must lie is called an *epipolar line*.

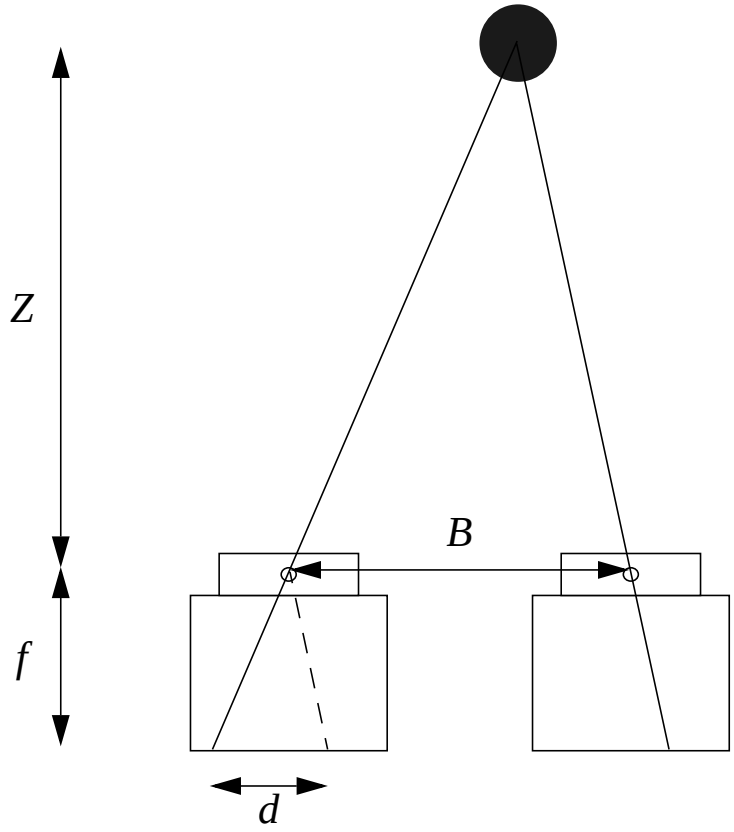
In general, it will be not be exactly horizontal.



Disparity and depth

Parallel cameras

If the cameras are pointing in the same direction, the geometry is simple.



B is the *baseline* of the camera system, *Z* is the *depth* of the object, *d* is the disparity (left *x* minus right *x*) and *f* is the focal length of the cameras. Then the unknown depth is given by

$$Z = \frac{f\ B}{d}$$

For parallel cameras:

- disparity is inversely proportional to depth — so stereo is most accurate for close objects;
- once we have found depth, the other coordinates in 3-D follow easily — e.g. taking either one of the images,

$$X = \frac{xZ}{f} = \frac{xB}{d}$$

where x is the image coordinate, and likewise for Y.

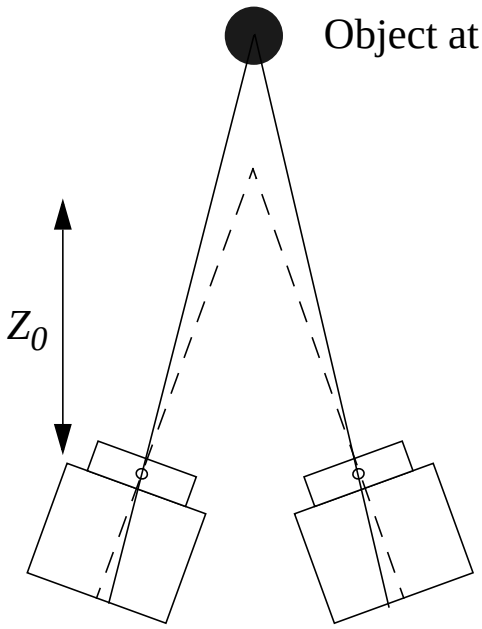
Converging cameras

This is the more realistic case. Our own convergence system is highly developed. Computer systems generally use converged cameras, and *active* stereo heads are becoming more usual.

The depth at which the cameras converge, Z_0, is the depth at which objects have zero disparity.

Closer objects have convergent disparity (numerically positive) and further objects have divergent disparity (numerically negative).

Finding Z_0 is part of stereo *calibration*.



$$Z \approx \frac{f\,B}{d + \frac{f\,B}{Z_0}}$$

The correspondence problem

To measure disparity, we first have to find corresponding points in the two images.

This turns out not to be easy.

Our own visual systems can match at a low level, as shown by *random-dot stereograms*, in which the individual images have no structure above pixel scale, but which when fused show a clear 3-D shape.

Stereo matchers need to start from some assumptions.

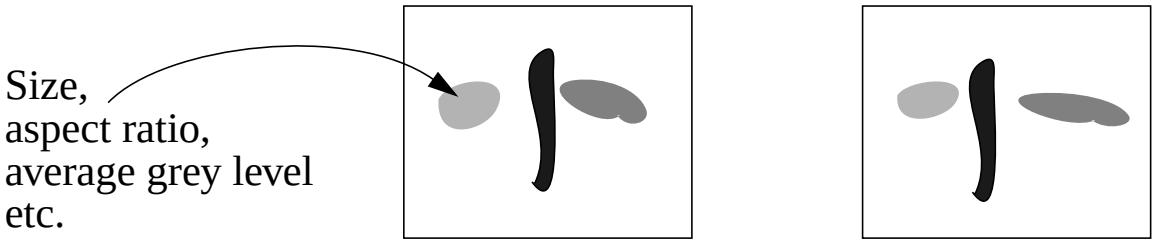
- Corresponding image regions are similar.
- A point in one image may only match a single point in the other image.
- If two matched features are close together in the images, then in most cases their disparities will be similar, because the environment is made of continuous surfaces separated by boundaries.

Many matching methods exist. The basic distinction is between

- feature-based methods which start from image structure extracted by preprocessing; and
- correlation-based methods which start from individual grey-levels.

Feature-based methods

1. Extract feature descriptions



2. Pick a feature in the left image.
3. Take each feature in the right image in turn (or just those close to the epipolar line), and measure how different it is from the original feature:

$$S = \frac{1}{w_0(l_r - l_l)^2 + w_1(\theta_r - \theta_l)^2 + w_2(g_r - g_l)^2 + \ldots}$$

   S is a measure of similarity, w_0 etc. are weights, and the other symbols are different measures of the feature in the right and left images, such as length, orientation, average grey level and so on. (You have to be careful with orientation, as 359° is close to 1°.)
4. Choose the right-image feature with the largest value of S as the best match for the original left-image feature.
5. Repeat starting from the matched feature in the right image, to see if we achieve consistency.

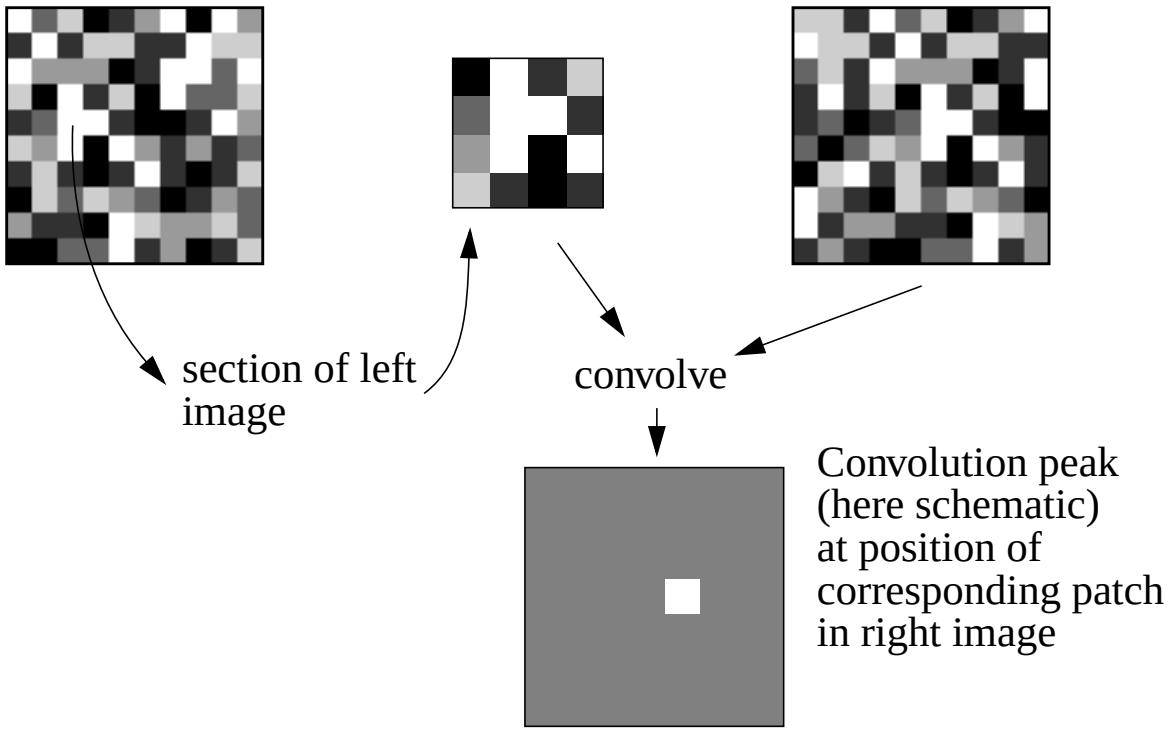
It is possible to use very simple features (just points, in effect) if the constraint that the disparity should vary smoothly is taken into account.

Feature-based methods give a *sparse* set of disparities — disparities are only found at feature positions.

Correlation-based methods

Imagine taking a small patch of the left image as a mask (row and column numbering reversed as necessary) and convolving it with the part of the right image close to the epipolar line.

The peak of the convolution output gives the position of the matching area of the right image, and hence the disparity of the best match.



Convolution-based methods can give a *dense* set of disparities — disparities are found for every pixel.

These methods can be very computationally intensive, but can be done efficiently on parallel hardware.

Variations on the basic method

Scale-space. Methods that exploit scale-space can be very powerful.

- Smooth the image so that the only features detected are separated by more than the likely disparity.
- Match each such feature with the nearest one in the other image — this should be reliable but will approximate because of blurring.
- Use the disparities found to guide the search for matches in a less smoothed image.

A good example is the work by Nishihara reprinted in *Readings in Computer Vision*, p. 63 (see bibliography on web page).

Relaxation. This is important in the context of neural modelling.

- Set up possible feature matches in a network.
- Construct an energy function that captures the constraints of the problem.
- Make incremental changes to the matches so that the energy is steadily reduced.

An early example of relaxation matching may be found in section 3.3 of Marr's book *Vision* (see bibliography on web page).

Other aspects of stereo

Very precise stereo systems can be made to estimate disparity at *sub-pixel* accuracies. This is important for industrial inspection and mapping from aerial and satellite images.

In controlled environments, *structured light* (e.g. stripes of light) can be used to provide easy matches for a stereo system.