Vision in a 3D world

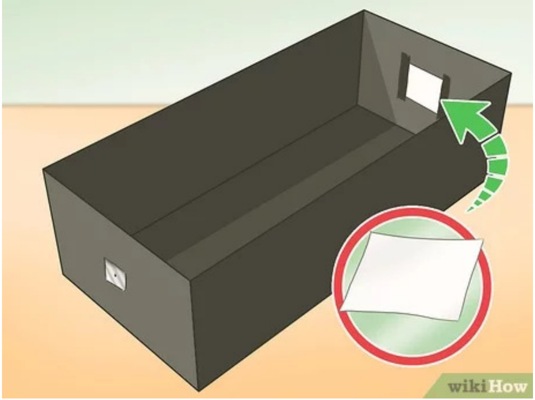
Adrian F. Clark
CSEE, University of Essex
alien@essex.ac.uk

We live in a 3D world

The world around us is 3D and we make use of visual cues such as the sizes of familiar objects to understand how far away they are

Here, we shall explore how to mimic some of these capabilities using computer vision

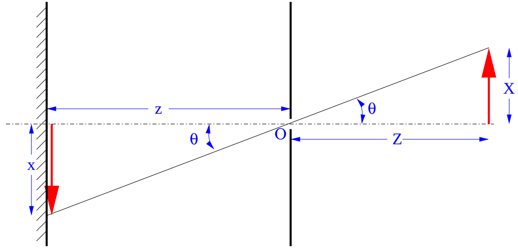
Let’s start with a pinhole camera



Traditionally made from a shoebox

A higher-tech approach

How does a camera work?

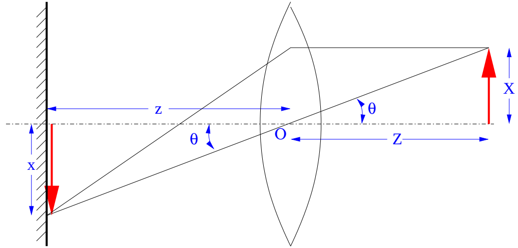


The pin-hole camera model

$$\tan \frac{x}{z} = \tan \frac{X}{Z} \implies \frac{x}{z} = \frac{X}{Z}$$

This means that if we know x, z and X, we can calculate Z

What about a real camera?

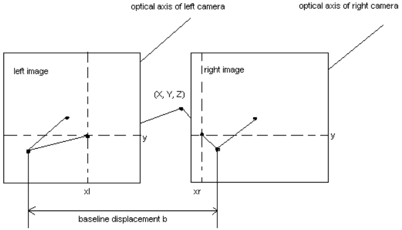


A real camera model

A real camera works in much the same way as a pinhole one, it's just that the lens focuses incoming parallel light onto the back focal plane, where our sensor lies

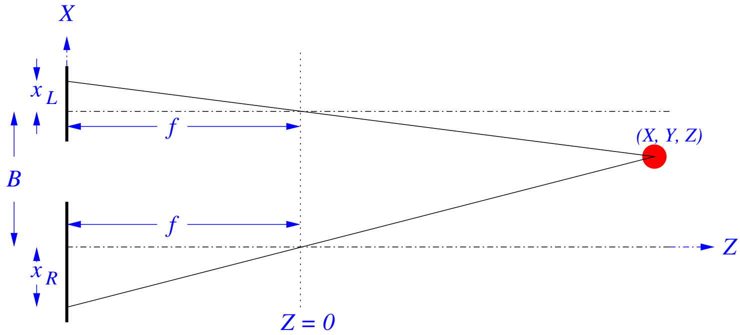
A two-camera configuration

Just like the eyes in your head, we can arrange a pair of cameras side by side



A pair of cameras

The distance apart of the optical centres is known as the *baseline*



A plan view of stereo cameras

In the right-hand camera, we have

$$\frac{X}{Z} = -\frac{x_r}{f}$$

and in the left-hand camera we have

$$\frac{B - X}{Z} = \frac{x_L}{f}$$

Adding them together, we obtain

$$\frac{x_l - x_R}{f} = \frac{B - X + X}{Z} = \frac{B}{Z}$$

Re-arranging, we obtain

$$Z = \frac{fB}{x_L - x_R}$$

This means that if we know

- f, the focal length, which is usually written on a lens but can easily be measured
- B, the distance between the optical axes of the lenses
- x_L and x_R, the positions of the same feature in the left and right images

we can calculate how far away that image feature is

Use the same units

f, B, x_L and x_R all need to be in the same units

x_L and x_R are normally measured in *pixels* so we need to know the size of the pixels or how far apart they are spaced on the sensor

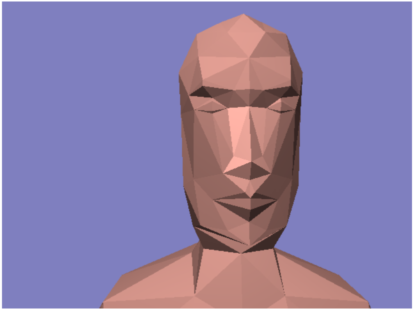
```
+---+---+---+---+---+---+
|   |   |   |   |   |   |
+---+---+---+---+---+---+
|   |   |   |   |   |   |
+---+---+---+---+---+---+
|   |   |   |   |   |   |
+---+---+---+---+---+---+
|   |   |   |   |   |   |
+---+---+---+---+---+---+
  ->|   |<-
```

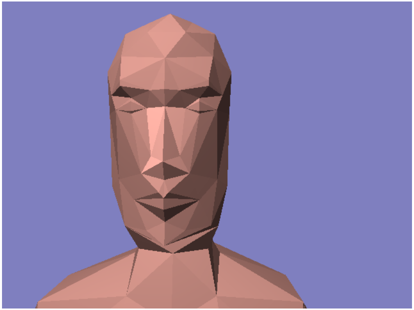
Image and world coordinate systems

```
    0   1   2   3 --> x
  +---+---+---+---+---+---+
0 |   |   |   |   |   |   |
  +---+---+
1 |   |   |     y ^
  +---+---+       |
2 |   |   |       +---> x
| +---+---+
v |   |   |
y
```

The world coordinate system is in the centre of the sensor (and so the centre of the image) and y increases in the opposite direction; see (9.5) and (9.6) of the lecture notes

Candide





Where is the tip of Candide's nose?

If, as in the notes, we have:

- $x_L = 69$ mm
- $x_R = -68$ mm
- $B = 150$ mm
- $f = 477.35$ mm

then

$$Z = \frac{fB}{x_L - x_R} = \frac{477.36 \times 150}{69 + 68} = 522.6\,\text{mm}$$

We know the true value is Z is 520 mm so the calculated value is reasonably close

Why isn't the calculation exactly right?

```
   |          |          |          |
---+----------+----------+----------+---
   |          |          |          |
   |          |          |          |
   |          |  x       |          |
   |          |          |          |
   |          |          |          |
---+----------+----------+----------+---
   |          |          |          |
```

The precise location of the tip of Candide's nose probably isn't in the exact centre of a pixel

We **cannot** measure the location more accurately that the nearest pixel

Uncertainty in measurements

When we have a measurement, we normally write it something like

$$(69 \pm 0.5)\,\text{mm}$$

if our pixels were 1 mm square

This uncertainty is usually known as *experimental error*, though *error* means uncertainty rather than a mistake

Propagating uncertainty

There are two rules:

- when adding or subtracting quantities, we always *add* their uncertainties
- when multiplying or dividing quantities, we always *add* their fractional uncertainties

When the uncertainty is not given, *you have to estimate it*

Where is the tip of Candide's nose (with uncertainty)?

Imagine we have:

- $x_L = (69 \pm 0.5)$ mm
- $x_R = (-68 \pm 0.5)$ mm
- $B = (150 \pm 2)$ mm
- $f = (477.35 \pm 0.1)$ mm

Then $D = x_L - x_R = (137 \pm 1)$ mm

We have

$$Z = \frac{fB}{D}$$

If we write δB as the uncertainty in B *etc*, we have

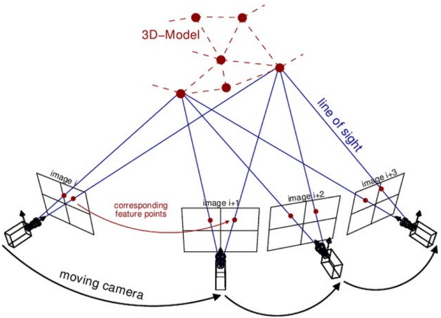
$$\frac{\delta Z}{Z} = \frac{\delta B}{B} + \frac{\delta f}{f} + \frac{\delta D}{D}$$

$$= \frac{0.1}{477.35} + \frac{2}{150} + \frac{1}{137} = 0.020842$$

so $\delta Z = 522.6 \times 0.020842 = 10.9$ mm

$Z - \delta Z$ easily encompasses the true value of Z

Visual Structure From Motion (SFM)



We must be able to find the same feature in several images

We must capture the scene from different orientations

Capturing the Fenwick Treasure

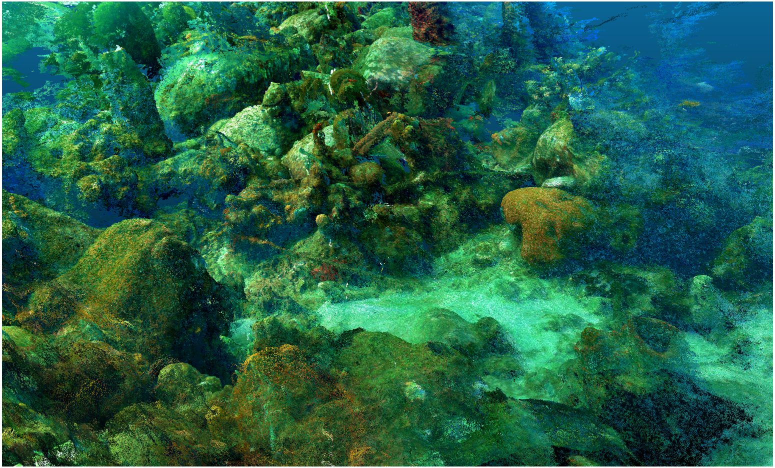
Some of the images we took. . .

Reconstructed Fenwick Treasure

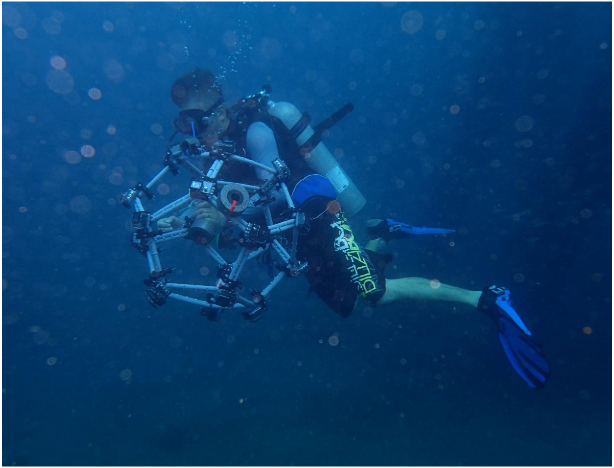
Reconstructed square 2

Elmstead's church: 332 million 3D points, two weeks of calculation

Reconstructing coral reefs

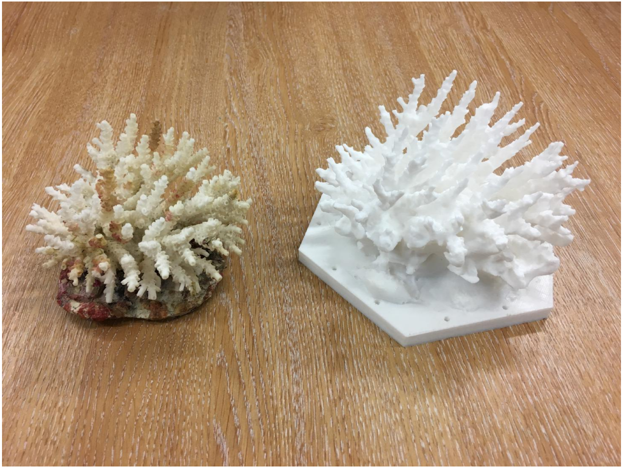


Our capture rig



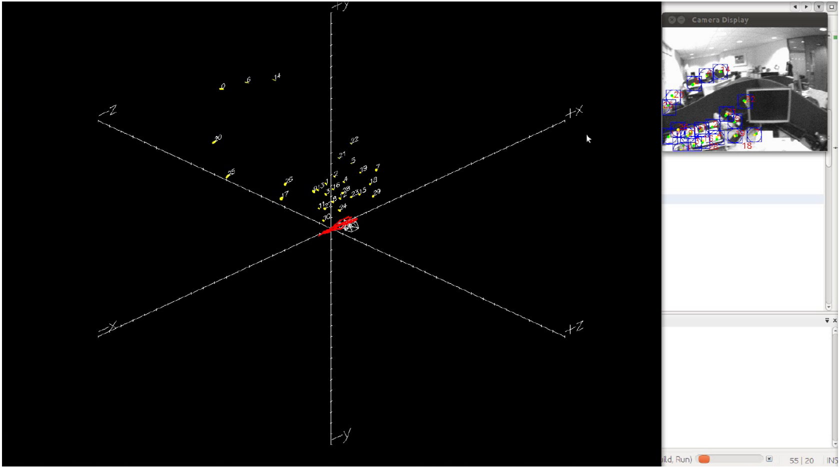
Jon Chamberlain had to go and capture the data, poor bloke

Models are good enough to be printed



Real and reconstructed acropora coral

Visual SLAM



The same maths but using few corners and ORB allows us compute sparse models in real time