

Computer Vision CITS4240

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

Stereo and Structured Light

We have two eyes, and precisely because of the way the world is projected differently onto our eyes, we are able to compute the relative distances of objects. Nevertheless, about 10% of us do not have true stereo vision—in this case depth information is retrieved from motion cues.

In Figure 1 below, we see that close objects are more widely separated on our retinas, and far objects project more closely together.

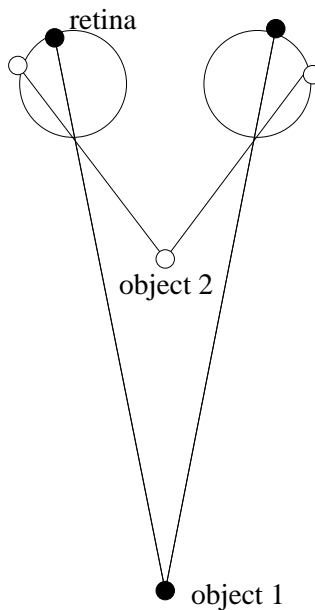


Figure 1: The projection of objects at different distances onto the retinas.

In 1960 Bela Julesz invented random dot stereograms. These consist of two images of random dots; the second has a portion of the first shifted relative to its original position, and the space thus created is filled in with more random dots. When the views are merged, one achieves the perception of depth, depending on the amount of disparity between the central portions. Moreover, depending on whether the central portion is moved to the

left or the right, the depth impression is behind or in front of the border. Note that it only makes sense to move the central portion horizontally, as our eyes are horizontally displaced.

Thus Julesz determined

- stereo is done at a very low level—we don't need to “interpret” the scene before perceiving depth.
- the main problem is clearly that of matching the dots.

The problem

Given two images formed in two image planes P and P' :

- For a point $m \in P$, which point $m' \in P'$ corresponds to m ? This is called the *correspondence problem*.
- Given two corresponding points m and m' , compute the 3D coordinates of the point M from which they came.

The simplest set-up (see Figure 2) to solve has

- two cameras, parallel and separated horizontally by a baseline distance b ;
- the cameras have the same focal length.

Definition 1 A *conjugate pair* is two points in different images that are the projections of the same point in the scene.

Definition 2 *Disparity* is the distance between points of a conjugate pair when the two images are superimposed.

If we take the left camera focal point, C , as the origin and use similar triangles, we have

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

and

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

implying

$$Z = \frac{bf}{x_l - x_r}.$$

We can then solve for X as

$$X = \frac{x_l Z}{f}.$$

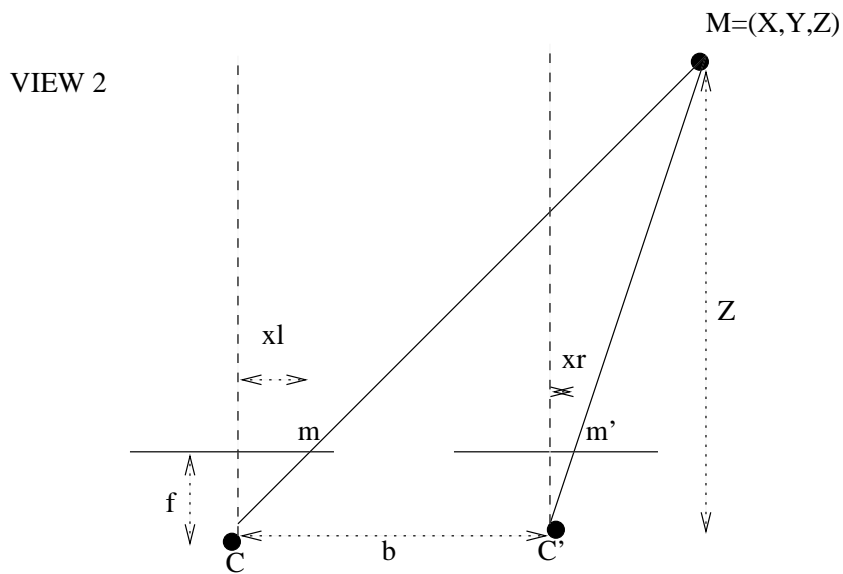
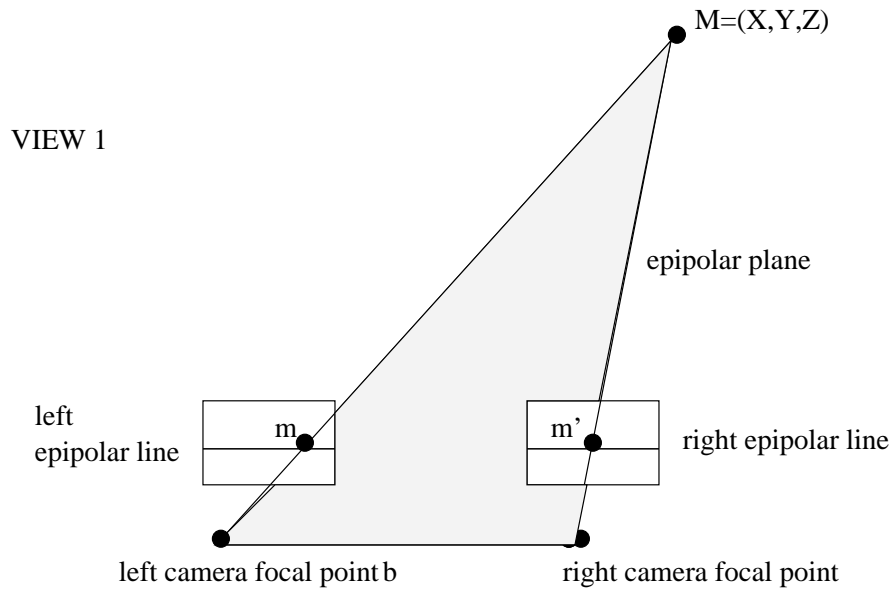


Figure 2: The geometry of the simple set-up.

As the cameras are parallel, and the baseline is aligned with the X axis, the projected y coordinate of the point will be identical in both images:

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

hence

$$Y = \frac{yZ}{f}.$$

Cameras in arbitrary position

Now suppose the cameras have arbitrary position and orientation with respect to each other as shown in Figure 3. With a verging camera geometry, the cameras' optical axes

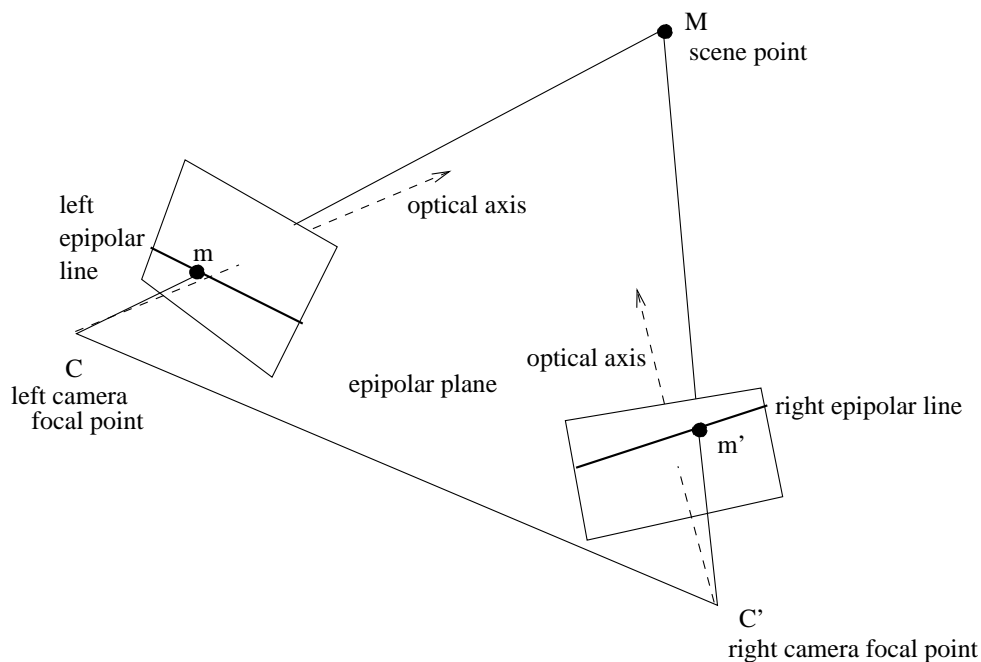


Figure 3: The geometry of the arbitrary set-up.

may or may not intersect at a point in space. However, the difference in orientations of the two optical axes gives an angle known as the *vergence angle*. For the simple set-up shown in Figure 2, the vergence angle is 0° and the corresponding image points projected by 3D points at infinity have zero disparity. For a verging camera set-up, corresponding image points that have zero disparity values are projected by 3D points at a *finite* distance from the cameras. This is illustrated in Figure 4.

Points to note when setting up your cameras:

Matching corresponding points is easy if the difference in position and orientation of the stereo views is small. Matching is difficult if the difference is large.

Accuracy of the 3D reconstruction is poor if the difference in position and orientation of stereo views is small. The accuracy is improved if the difference is large.

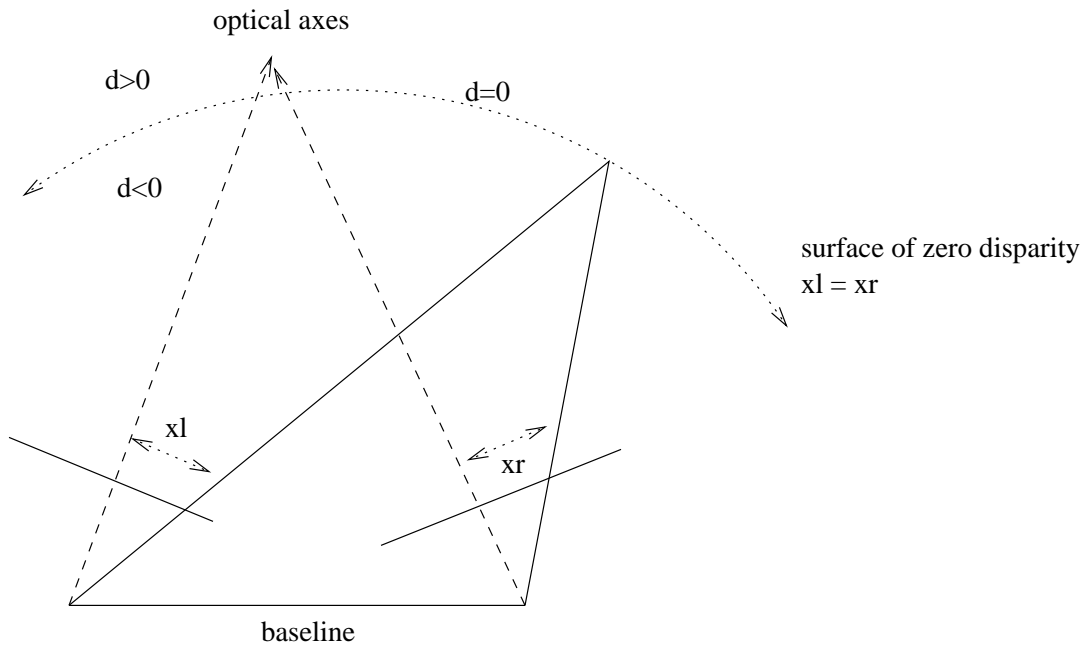


Figure 4: The disparity zones of a verging camera set-up.

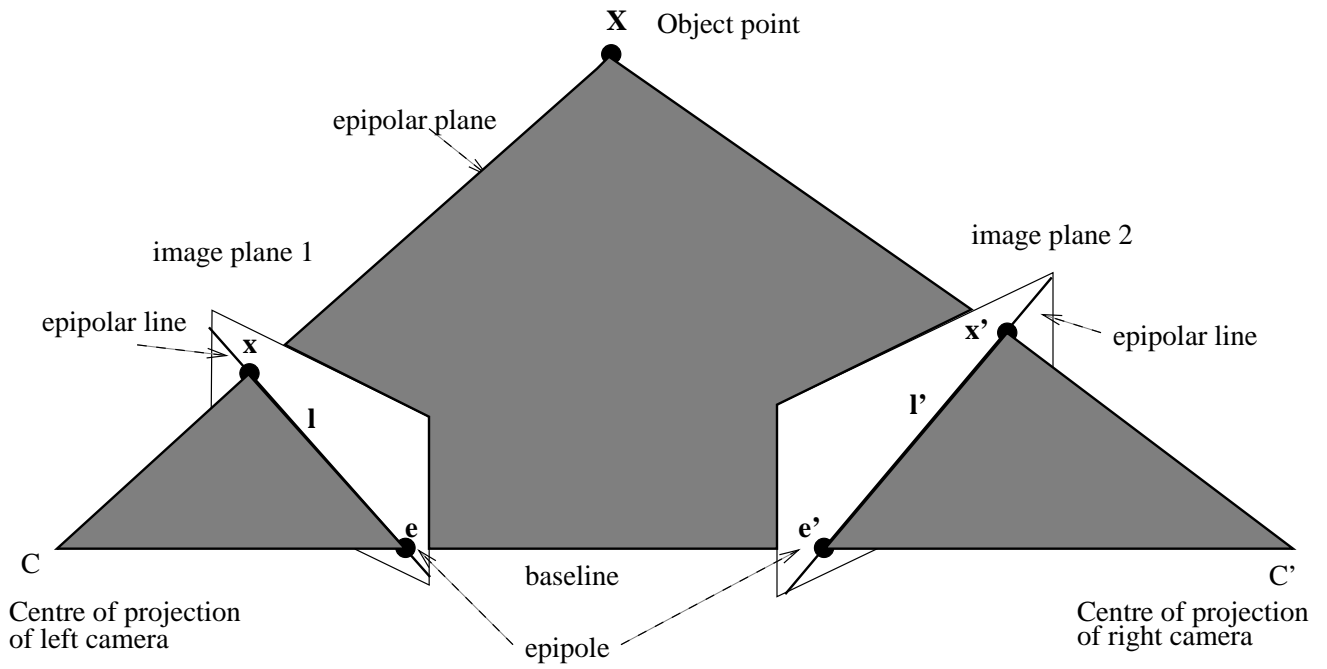


Figure 5: Epipolar Geometry.

The matching problem

For each point in the left image, we need to find the corresponding point in the right image. This is a very difficult problem. It helps if the point to be matched is clearly different from its surrounding pixels. This implies that we first need to find matchable features. Researchers have used both edges and regions for matching.

An important tool in the matching process are epipolar lines. When we see a feature in one image we know that it lies anywhere along the viewing ray. We can project this viewing ray into the other image. This forms a line (an epipolar line) in the second image on which the feature we are trying to match must lie.

All epipolar lines pass through the projection of the other image's projection centre in the current image. This point is known as the *epipole*.

Edge matching

Edge matching only gives depth values at the edge points, implying a sparse depth map. The standard algorithm implements a multiscale approach, and assumes a parallel geometry, that is, it assumes that the epipolar lines are the rows of the image.

1. Filter each image with Gaussian filters at four filter widths such the each filter is twice as wide as the next. (This can be done by repeated convolution with the smallest filter.)
2. Compute edge positions on each row.
3. Match edges in corresponding rows at the coarse resolutions by comparing orientations and strengths (horizontal edges can't be matched).
4. Refine the disparity estimates by matching at finer scales.

Region matching

Edge based matching gives only sparse depth information, and meaningless information along occluding edges.

To find regions of points that are interesting for matching (thus, clearly not all the same), we look for windows of high variance.

- We compute the directional variances in a window S_1 centred at (x_c, y_c) in image 1 as follows:

$$I_1 = \sum_{(x,y) \in S_1} [f(x, y) - f(x, y + 1)]^2$$
$$I_2 = \sum_{(x,y) \in S_1} [f(x, y) - f(x + 1, y)]^2$$
$$I_3 = \sum_{(x,y) \in S_1} [f(x, y) - f(x + 1, y + 1)]^2$$

$$I_4 = \sum_{(x,y) \in S_1} [f(x,y) - f(x+1, y-1)]^2$$

The interest value at the central pixel (x_c, y_c) is given by

$$I(x_c, y_c) = \min(I_1, I_2, I_3, I_4).$$

- If $I(x_c, y_c) > \text{threshold}$, then it indicates that the region encloses some interest point (e.g., a corner) and region correlation between window S_1 in image 1 and a window in image 2 should be carried out. We assume feature points to be matched (along epipolar lines) within S_1 have disparity (d_x, d_y) . Then a measure of similarity is given by the correlation coefficient between two regions of the same size centred around the features in the two images f_1 and f_2 :

$$r(d_x, d_y) = \frac{\sum_{(x,y) \in S_1} (f_1(x,y) - \bar{f}_1) (f_2(x+d_x, y+d_y) - \bar{f}_2)}{\sqrt{\sum (f_1(x,y) - \bar{f}_1)^2 \sum (f_2(x+d_x, y+d_y) - \bar{f}_2)^2}}, \quad (1)$$

where \bar{f}_1 is the mean grey value within S_1 in image 1, and \bar{f}_2 is the mean grey value within the window S_2 centred at $(x_c + d_x, y_c + d_y)$ in image 2.

Note that in (1) the denominator term, which is the product of the standard deviations of intensity variations within windows S_1 and S_2 , normalises the result relative to the variances within the windows S_1 and S_2 . Thus, invariance to brightness differences in the two images is obtained by working with grey values relative to the mean value within the window. The range of values of the correlation coefficient computed using the formula above is $[-1, 1]$, with value 1 indicates a perfect match between the two windows.

Constraints used for matching

Similarity black dots only match black dots.

Uniqueness almost always, one black dot can match no more than one black dot.

Continuity disparity values vary smoothly almost everywhere.

Ordering if $m \leftrightarrow m'$ and $n \leftrightarrow n'$ then if m is to the left of n then m' is to the left of n' .

Matching problems

A common problem for correlation based approaches is that an inclined surface will appear with different degrees of foreshortening in the two images. Thus the pattern of grey levels in one image may not be easily matched to the corresponding region in the other image.

Surface discontinuities (edges) in the scene imply that some points seen in one image will be occluded in the other. Thus some features are unmatchable.

Reconstruction of 3-D coordinates

Given that the cameras have been calibrated and the two perspective projection matrices $\mathbf{C} = [q_{ij}]$ and $\mathbf{C}' = [q'_{ij}]$ are known, then for any scene point M with unknown 3-D coordinates (X, Y, Z) , that projects onto the two image planes at (u, v) and (u', v') , we have

$$\mathbf{C} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} = \begin{bmatrix} su \\ sv \\ s \end{bmatrix} \quad \text{and} \quad \mathbf{C}' \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} = \begin{bmatrix} s'u' \\ s'v' \\ s' \end{bmatrix}.$$

Eliminating s and s' and combining the two equations into matrix form gives

$$\begin{bmatrix} q_{11} - uq_{31} & q_{12} - uq_{32} & q_{13} - uq_{33} \\ q_{21} - vq_{31} & q_{22} - vq_{32} & q_{23} - vq_{33} \\ q'_{11} - u'q'_{31} & q'_{12} - u'q'_{32} & q'_{13} - u'q'_{33} \\ q'_{21} - v'q'_{31} & q'_{22} - v'q'_{32} & q'_{23} - v'q'_{33} \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} uq_{34} - q_{14} \\ vq_{34} - q_{24} \\ u'q'_{34} - q'_{14} \\ v'q'_{34} - q'_{24} \end{bmatrix}. \quad (2)$$

(Note that q_{34} is typically fixed at 1, as mentioned in a previous lecture.) This is a linear system in (X, Y, Z) . The 3-D coordinates of M can be easily computed.

Range imaging using structured light

There is a difference between active stereo and passive stereo.

In an active stereo vision system, one of the cameras is replaced with a projector or a laser unit, which projects onto the object of interest a sheet of light (or multiple sheets of light simultaneously). Figure 6 illustrates an example of an active stereo vision system. The idea is that once the perspective projection matrix of the camera and the equations of the planes containing the sheets of light relative to a global coordinate frame are computed from calibration, the triangulation for computing the 3-D coordinates of object points simply involves finding the intersection of a ray (from the camera) and a plane (from the sheet of light of the projector or laser).

The geometry of the set-up is given in Figure 7. From this idealised set-up we can solve for X, Y and Z

$$(X, Y, Z) = \frac{b}{f \cot \theta - x}(x, y, f).$$

In general we will work in terms of a camera calibration matrix C and an equation of the plane of light that has been calibrated in terms of the global reference frame.

For any image point on the stripe of light we can write two equations from the image u, v coordinates of that point using the calibration matrix

$$(q_{11} - uq_{31})X + (q_{12} - uq_{32})Y + (q_{13} - uq_{33})Z + (q_{14} - uq_{34}) = 0$$

and

$$(q_{21} - vq_{31})X + (q_{22} - vq_{32})Y + (q_{23} - vq_{33})Z + (q_{24} - vq_{34}) = 0.$$

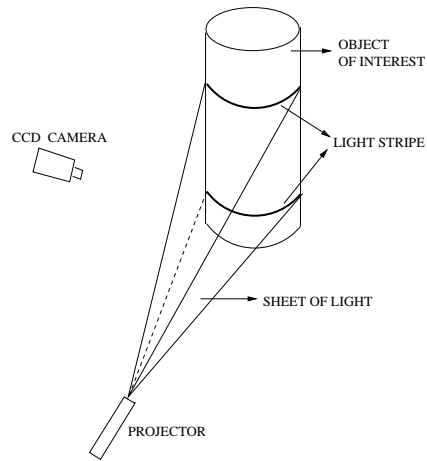


Figure 6: Example of an active stereo vision system.

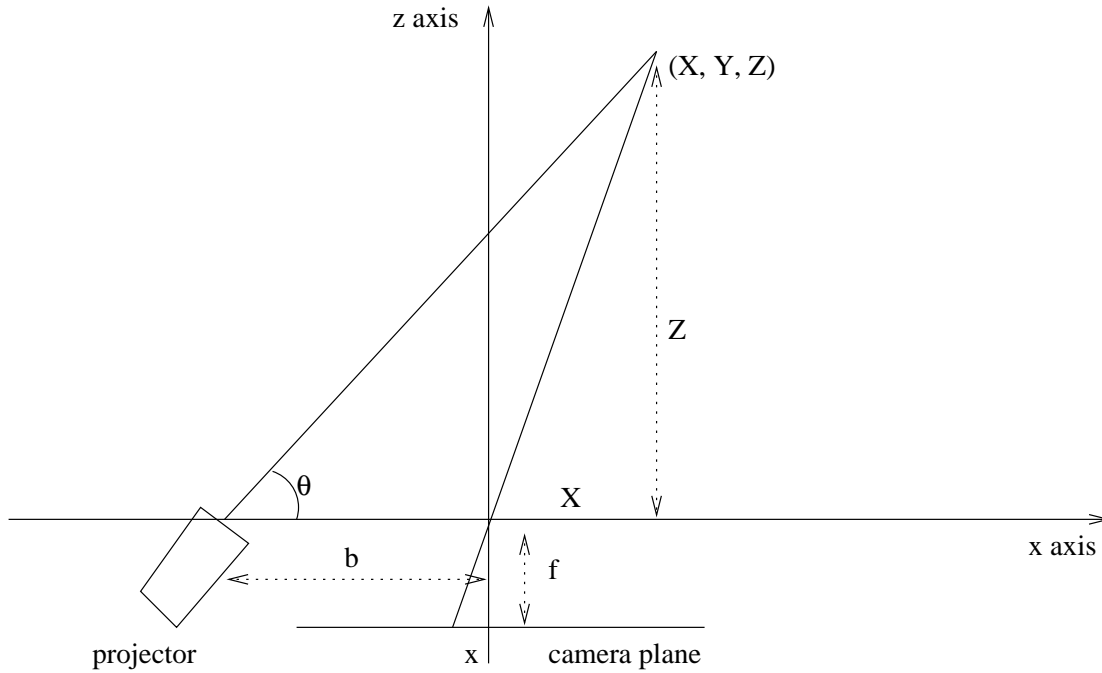


Figure 7: The geometry of the structured light set-up.

(Note that these two equations are equivalent to the first two rows of Equation (2).)
 If the equation of the light stripe plane relative to the same global reference frame is

$$aX + bY + cZ + d = 0,$$

then we can set up a matrix equation in the three unknowns X, Y and Z :

$$\begin{bmatrix} q_{11} - uq_{31} & q_{12} - uq_{32} & q_{13} - uq_{33} \\ q_{21} - vq_{31} & q_{22} - vq_{32} & q_{23} - vq_{33} \\ a & b & c \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} uq_{34} - q_{14} \\ vq_{34} - q_{24} \\ -d \end{bmatrix}.$$

It is possible to invert the matrix above symbolically so that the X, Y and Z coordinates can be calculated directly from the image coordinates of the light stripe by a matrix multiplication.

The main points to note about active stereo are:

- it is ideal for scenes that do not contain sufficient features (for matching).
- there is no correspondence problem.
- it requires good lighting control, and thus is restricted to indoor environments.
- both camera and projector must be pre-calibrated.
- it gives a dense range map.