Chapter 11

3D data extraction

Chapter 10 already considered the use of several images as input, rather than analysing a single image in isolation. The reason there was that we were interested in the motions of objects. A second good reason is when we want to reconstruct a scene or object in 3D, i.e. if we want to build three-dimensional models. This typically entails taking several views as well, or at least combine special illumination from directions different from the viewing direction. This chapter describes way to get distances to objects, so-called 'range data'.

11.1 I - A taxonomy of methods

A 3-D acquisition taxonomy is given in figure 11.1. It distinguishes *active* from *passive methods*. With active techniques the light sources are specially controlled to give quantitative information about the scene. Active lighting incorporates some form of temporal or spatial modulation used to extract the three-dimensional information. With passive techniques, on the other hand, light is not controlled or only with respect to image quality. With *uni-directional methods* all components of the image acquisition system are aligned along one direction. With *multi-directional systems*, several viewpoints or controlled illumination directions are involved. From a computational viewpoint, active methods are least demanding. Their applicability is restricted to environments where the special illumination techniques can be applied. The methods mentioned in the taxonomy will be discussed in the next sections. Only two of them will be treated in some detail, however. One is stereo and the second active triangulation.

11.2 I - Stereo with calibrated cameras

11.2.1 A simple configuration

Suppose we have two images, taken at the same time and from different viewpoints. The situation is illustrated in fig. 11.3. The principle behind stereo-based 3D reconstruction is simple: given two projections of the same point, its 3D position is found as the intersection

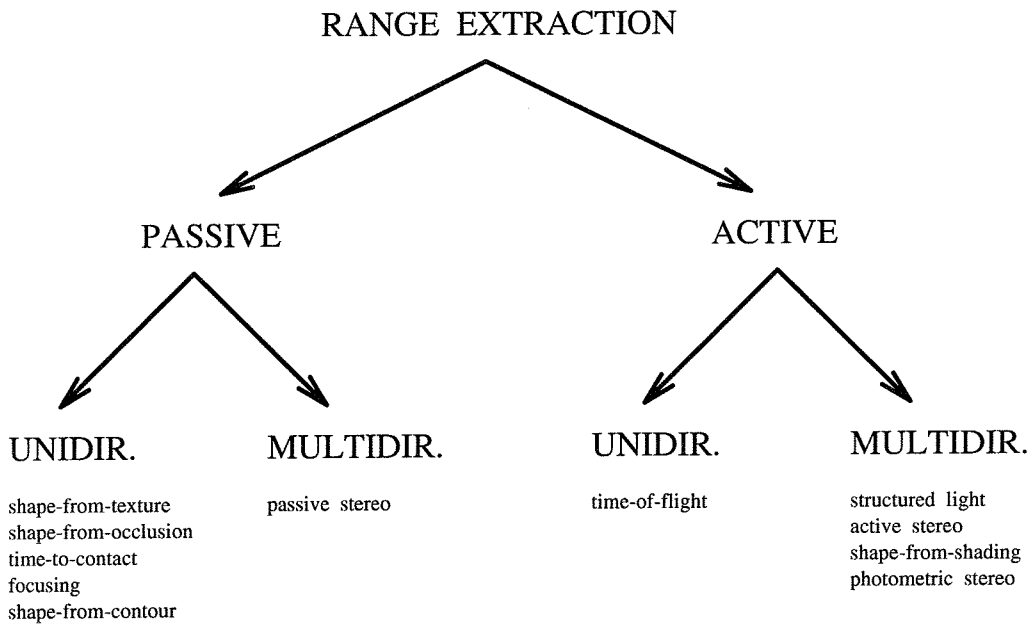

Figure 11.1: *Taxonomy of methods for the extraction of information on position and/or orientation in three-dimensional space.*

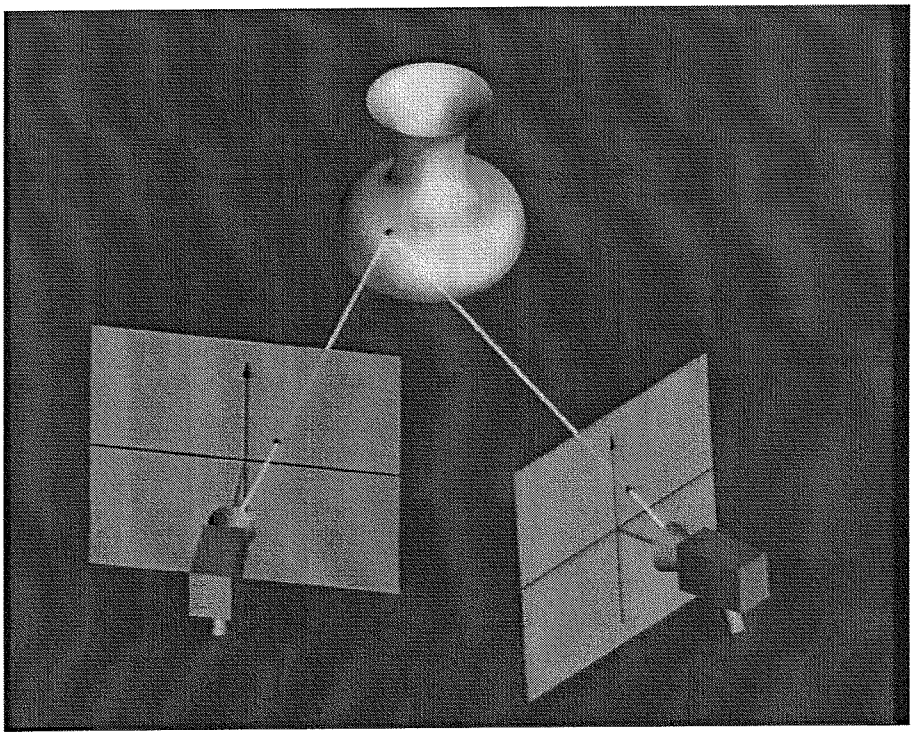

Figure 11.2: *The principle behind stereo-based 3D reconstruction if very simple: given two images of a point, the point's position in space is found as the intersection of the two projection rays.*

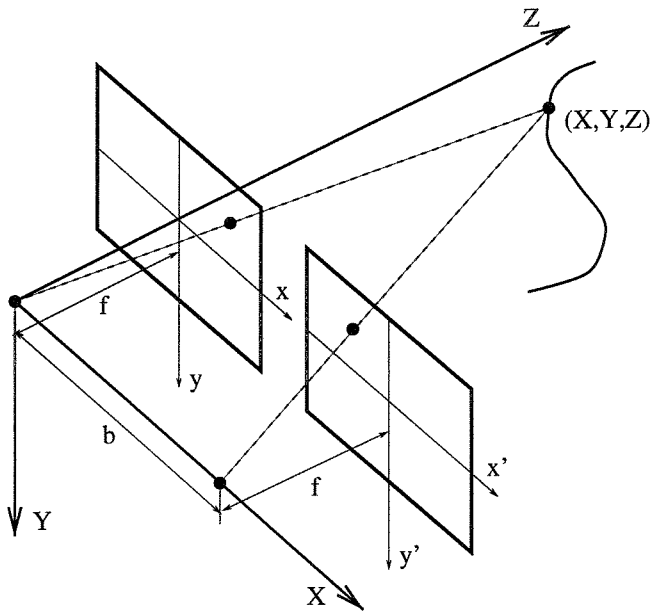

Figure 11.3: *Simple stereo setup with cameras having coplanar image planes and parallel axes. Moreover, they have aligned x axes.*

of the two projection rays through its two images. Repeating such a projection for several points yields the 3D shape and configuation of the objects in the scene. Note that this construction – sometimes referred to as *triangulation* requires complete knowledge of both the internal and the external parameters of the cameras.

As a first step towards a mathematical analysis, a very simple but often used configuration is discussed first. Suppose that the two cameras have identical internal parameters and that $s = x_0 = y_0 = 0$. Moreover, the optical axes of the cameras are parallel and their image planes are coplanar, with coincident x-axes. This special configuration is shown in fig. 11.3). The *baseline* is the distance between the two camera centers. As announced in chapter 4, the image planes are placed in front of the optical centers of the cameras. Also the situation for the external parameters is simplified. We let the world coordinate frame coincide with the camera frame of the left camera.

The images of a point with coordinates $(X, Y, Z)^t$ are then

$$\rho \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = K \begin{pmatrix} X \\ Y \\ Z \end{pmatrix}$$

for the first camera (left image) and

$$\rho' \begin{pmatrix} x' \\ y' \\ 1 \end{pmatrix} = K \begin{pmatrix} X - b \\ Y \\ Z \end{pmatrix}$$

for the second camera (right image), where

$$K = \begin{pmatrix} fk_x & 0 & 0 \\ 0 & fk_y & 0 \\ 0 & 0 & 1 \end{pmatrix} .$$

Hence,

$$\begin{cases} x = \dfrac{fk_x X}{Z} , \\ y = \dfrac{fk_y Y}{Z} , \end{cases} \quad \text{and} \quad \begin{cases} x' = \dfrac{fk_x (X-b)}{Z} , \\ y' = \dfrac{fk_y Y}{Z} , \end{cases}$$

Note that $y = y'$ for all point projections. Solving these equations for X, Y, and Z yields

$$\begin{aligned} X &= b\frac{x}{(x-x')} , \\ Y &= b\frac{k_x}{k_y}\frac{y}{(x-x')} , \\ Z &= bk_x\frac{f}{(x-x')} . \end{aligned}$$

The difference $(x - x')$ is the so-called *disparity*. The distance Z, also referred to as the *range*, is inversely proportional to disparity. Beyond a certain distance, depth measurement will therefore become very coarse. Human stereo depth perception is limited to distances of about 10 meters. Beyond, depth impressions arise from other cues. Note also that the disparity is directly proportional to b and f. Depth resolution can be increased by increasing one or both of these variables. Upon such a change, the same distance will correspond to a larger disparity and therefore distance sampling gets finer. It should be noted, however, that one should strike a balance between increasing resolution and keeping visible to both cameras as much of the scene as possible. Indeed, when disparities get larger, chances of finding both projections within the images diminish. Fig 11.4 shows planes of equal disparity for two focal lengths and two baseline distances. The same distance is seen to be sampled finer after the focal length or the baseline have been increased. A smaller part of the scene is visible to both cameras upon such a change, certainly for those parts which are close.

The whole method is seen to be based on finding the projection of points in both images. More specifically, if a point in the left image is selected, the problem is to find the corresponding point in the right image, assuming one exists (the point could be invisible to the right camera). Solving this *correspondence problem* is the hard part of stereo (see section 11.2.4).

An example stereo image pair, taken with a simple camera configuration as discussed in this section, is shown in fig. 11.5. Two views of the three-dimensional reconstruction, obtained by the calculation of stereo disparities, is shown in fig. 11.6. The results are quite convincing for the torsos, but the homogeneous wall in the background could not be reconstructed. The difference lies in the presence or absence of texture, resp., and the related ease or difficulty of finding corresponding points.

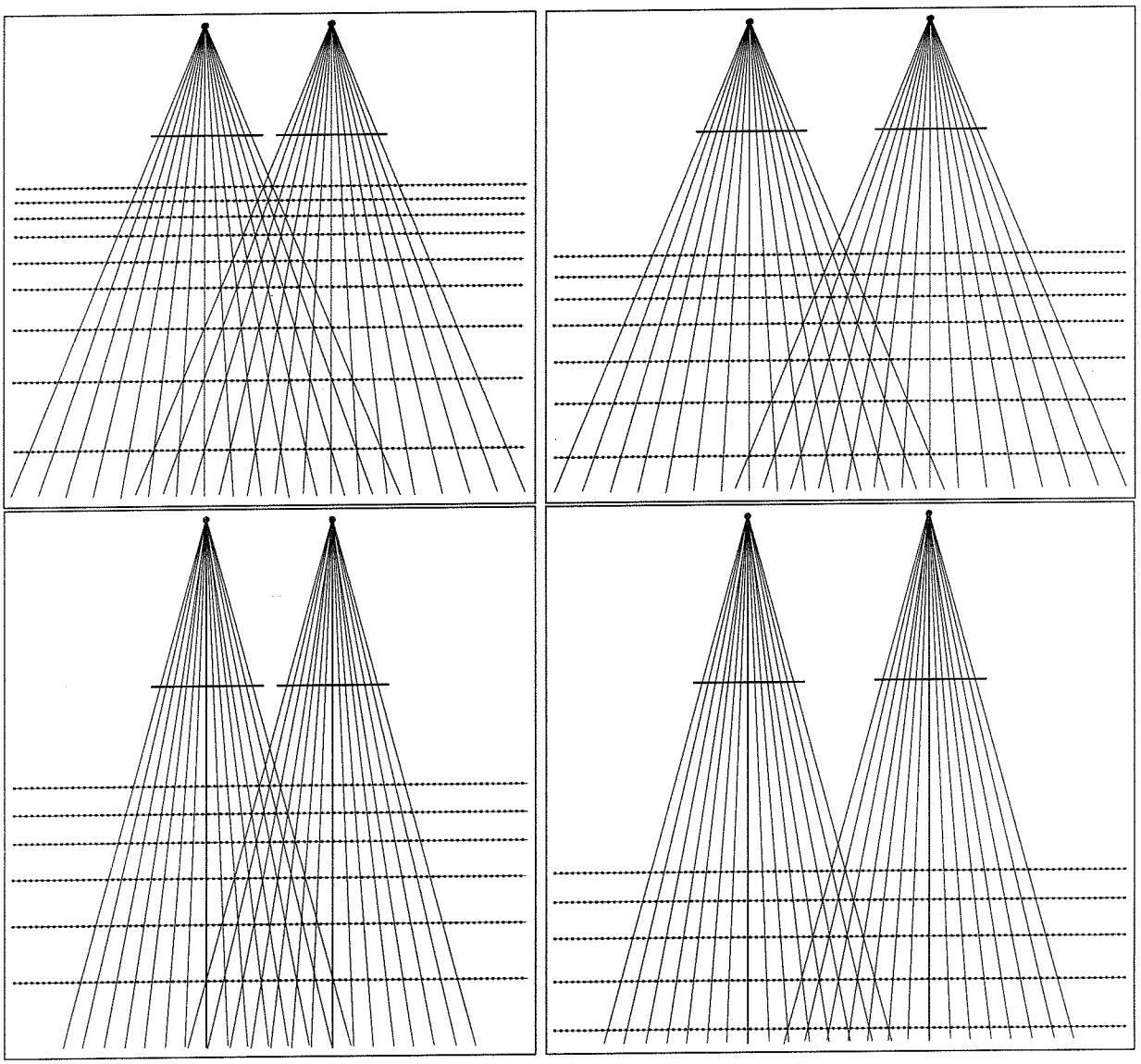

Figure 11.4: *Equal disparities correspond to points of equal distance: rays that project onto points with a fixed disparity intersect in planes of constant range. At a given distance these planes get closer (i.e. sample distance more densely) if the baseline is increased (top-left to top-right), the focal length is increased (top-left to bottom-left), or both (top-left to bottom-right).*

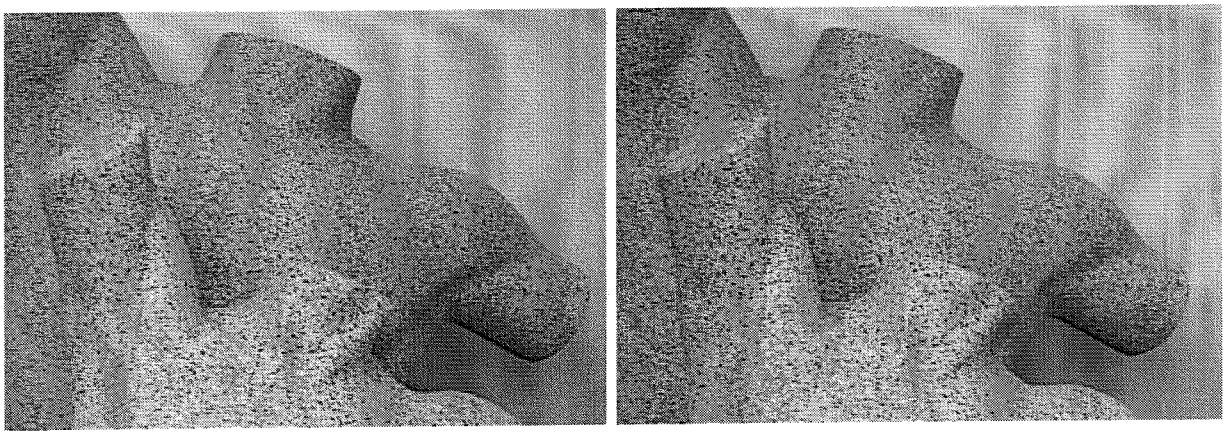

Figure 11.5: *Pair of stereo images for a scene with two torsos.*

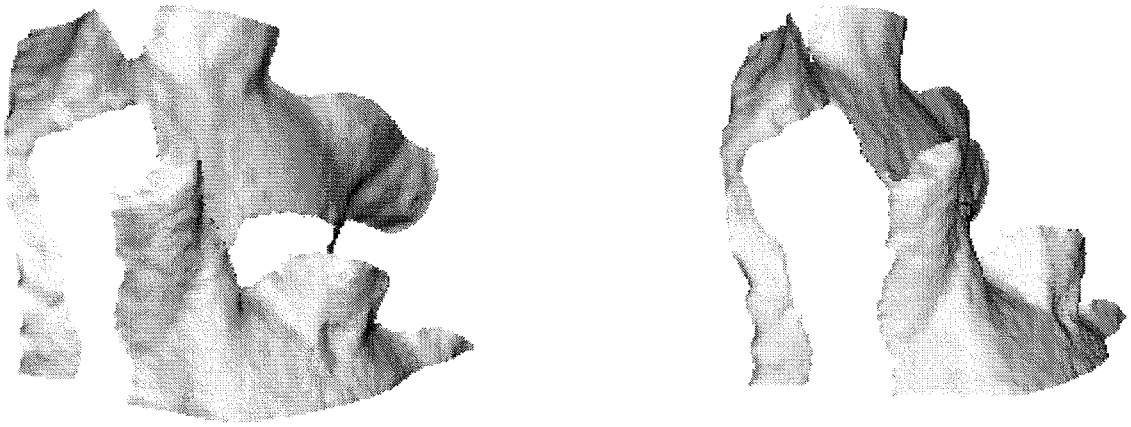

Figure 11.6: *Two views of the three-dimensional reconstruction obtained for the stereo image pair of fig. 11.5.*

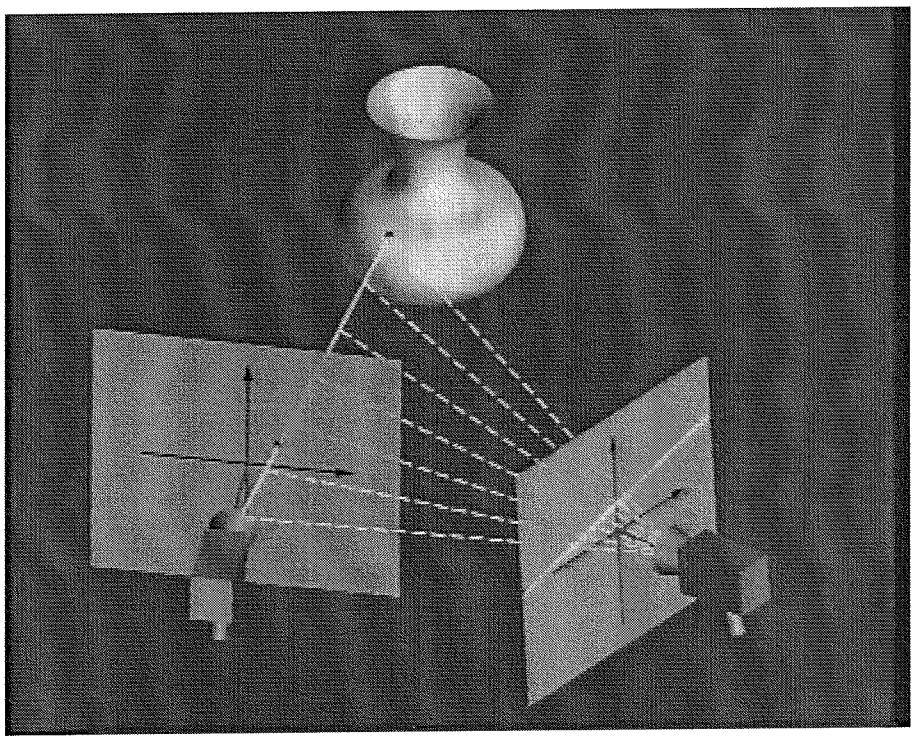

Figure 11.7: *The point in the second image corresponding to a point in the first image lies somewhere on the line which is the projection in the second image of the ray of sight through the point's projection in the first camera.*

11.2.2 General configurations - the epipolar constraint

The foregoing analysis can be generalized towards arbitrary camera configurations. We start this analysis by deriving a relation between the two projections of a point. Indeed, if we know the projection of a point in the first image, then there are restrictions on where its projection in the second point can be, and v.v. Knowledge about these restrictions will be important to ease the correspondence problem, which gets more difficult for general camera configurations.

Given the projection $\mathbf{p}$ in the first image of a point $\mathbf{P}$, it's position in the scene could be anywhere along the ray of projection connecting the camera center $\mathbf{C}$ with $\mathbf{p}$. The projection $\mathbf{p}'$ in the second image must ly on the projection $\mathbf{l}'$ of this ray in the second view, as depicted in Figure 11.7. To turn this into a mathematical formula, suppose for a moment that the camera parameters of both cameras are known. The direction of the ray of sight creating the image point $\mathbf{p}$ in the first camera is given by the 3-vector $K^{-1}\mathbf{p}$ in the camera-centered reference frame of the first camera, where K is the calibration matrix of the first camera. With respect to the world frame, this direction vector is $RK^{-1}\mathbf{p}$, where R is the rotation matrix expressing the orientation of the first camera in the world frame. As the position of the first camera in the world frame is given by $\mathbf{C}$, the parameter equations

of the ray of sight in the world frame are:

$$\mathbf{P} = \mathbf{C} + \mu\, RK^{-1}\,\mathbf{p} \qquad \text{for some } \mu \in I\!R. \tag{11.1}$$

Every scene point $\mathbf{P}$ satisfying equation (11.1) for some real number μ projects onto $\mathbf{p}$ in the first image. If K', R' and $\mathbf{C}'$ are the camera parameters of the second camera, then the projection $\mathbf{p}'$ of $\mathbf{P}$ in the second image is given by

$$\rho'\,\mathbf{p}' = K'R'^{t}\,(\mathbf{P} - \mathbf{C}') \tag{11.2}$$

for some non-zero $\rho' \in I\!R$. Substituting expression (11.1) for $\mathbf{P}$ into this formula, yields

$$\rho'\,\mathbf{p}' = \mu\, K'R'^{t}RK^{-1}\,\mathbf{p} + K'R'^{t}\,(\mathbf{C} - \mathbf{C}') \ . \tag{11.3}$$

The last term in this equation corresponds to the projection $\mathbf{e}'$ of the position $\mathbf{C}$ of the first camera in the second image:

$$\rho'_e\,\mathbf{e}' = K'R'^{t}\,(\mathbf{C} - \mathbf{C}') \ . \tag{11.4}$$

$\mathbf{e}'$ is called the *epipole* of the first camera in the second image. The first term in the righthand side of the equality, on the other hand, indicates the direction of the ray of sight (11.1) in the second image. Indeed, recall that $RK^{-1}\,\mathbf{p}$ is the direction vector of the ray of sight in the world frame. In the camera-centered reference frame of the second camera, this vector is $R'^{t}RK^{-1}\,\mathbf{p}$. The corresponding position in the second image is then expressed (in homogeneous coordinates) by $K'R'^{t}RK^{-1}\,\mathbf{p}$. Put differently, $K'R'^{t}RK^{-1}\,\mathbf{p}$ *are the homogeneous coordinates of the vanishing point of the ray of sight (11.1) in the second image.* Indeed, the first term specifies the projection $\mathbf{p}'$ if μ evolves to infinity.

To simplify the notation, put $A = \frac{1}{\rho'_e}K'R'^{t}RK^{-1}$. The factor $1/\rho'_e$ is included because it simplifies notations that are to come. It doesn't change the effect of the transformation, as the matrix is $K'R'^{t}RK^{-1}$ can be multiplied by an arbitrary scalar. Then A is an invertible 3×3-matrix which, for every point $\mathbf{p}$ in the first view, gives the homogeneous coordinates $A\mathbf{p}$ of the vanishing point in the second view of the ray of sight observing $\mathbf{p}$ in the first camera. In terms of projective geometry, A is a matrix of the homography that maps the first image onto the second one via the plane at infinity of the scene. Formula (11.3) can now be rewritten as

$$\rho'\,\mathbf{p}' = \rho'_e\,(\mu\, A\,\mathbf{p} + \mathbf{e}') \ . \tag{11.5}$$

Observe that equation (11.5) algebraically expresses that, *for a given point* $\mathbf{p}$ *in one view, the corresponding point* $\mathbf{p}'$ *in the other view lies on the line* $\mathbf{l}'$ *through the epipole* $\mathbf{e}'$ *and the vanishing point* $A\mathbf{p}$ *of the ray of sight of* $\mathbf{p}$ (see Figure 11.7). This line $\mathbf{l}'$ is called the *epipolar line* corresponding to $\mathbf{p}$ in the second image. Relation (11.5) is more efficiently expressed in another way, as derived now. First, note that it is equivalent to

$$|\,\mathbf{p}'\ \mathbf{e}'\ A\mathbf{p}\,| = \mathbf{p}'^{t}\,(\mathbf{e}' \times A\mathbf{p}) = 0 \ ; \tag{11.6}$$

Note that a consequence of this is that all epipolar lines intersect in the epipole, a fact that is illustrated in fig. 11.8

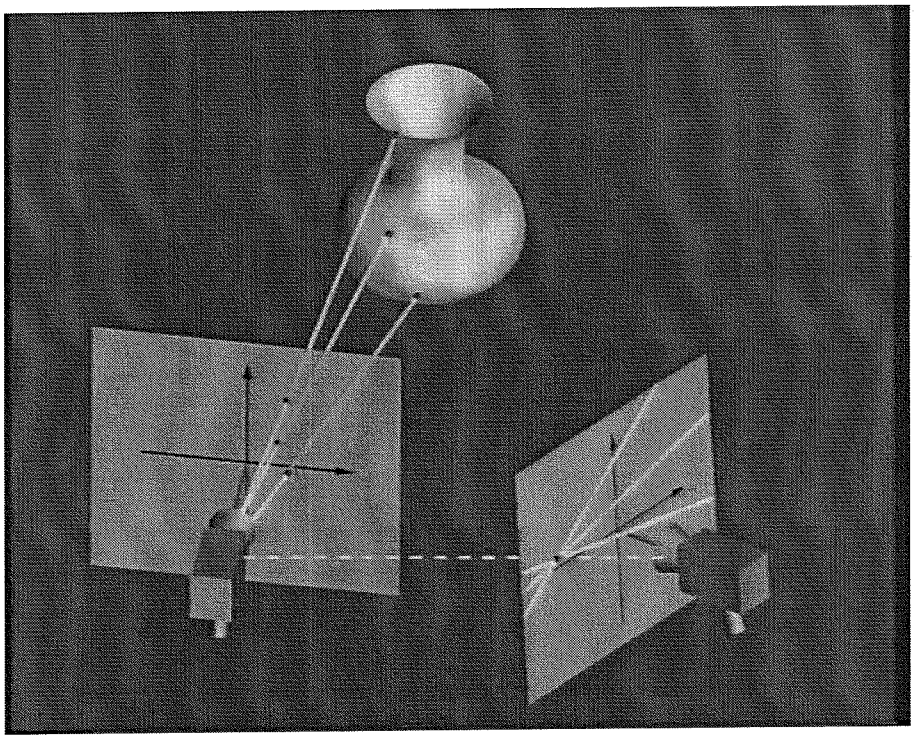

Figure 11.8: *The epipole, i.e. the projection of the center of projection of the left camera in the image of the right camera lies on all the epipolar lines.*

Next, we introduce for a 3-vector $\mathbf{a} = (a_1, a_2 a_3)^t \in I\!R^3$ the skew-symmetric 3×3-matrix

$$[\mathbf{a}]_\times = \begin{pmatrix} 0 & -a_3 & a_2 \\ a_3 & 0 & -a_1 \\ -a_2 & a_1 & 0 \end{pmatrix} , \tag{11.7}$$

which represents the cross product with $\mathbf{a}$; i.e. $[\mathbf{a}]_\times \mathbf{v} = \mathbf{a} \times \mathbf{v}$ for all $\mathbf{v} \in I\!R^3$. Observe that $[\mathbf{a}]_\times$ has rank 2 if $\mathbf{a}$ is non-zero. Now, eq. (11.6) can be rewritten as

$$|\mathbf{p}' \; \mathbf{e}' \; A\mathbf{p}| = \mathbf{p}'^t [\mathbf{e}']_\times A \, \mathbf{p} = 0 \; . \tag{11.8}$$

The 3×3-matrix $F = [\mathbf{e}']_\times A$ is called the *fundamental matrix.* It has rank 2 because $[\mathbf{e}']_\times$ has rank 2.

Remark. By definition of the fundamental matrix, $\mathbf{e}'^t F = 0$. So, the epipole $\mathbf{e}'$ in I' can be computed if F is known. Also, as $\mathbf{e}' = A\mathbf{e}$ one has $F\mathbf{e} = [\mathbf{e}']_\times A \, \mathbf{e} = [\mathbf{e}']_\times \mathbf{e}' = 0$ and the epipole $\mathbf{e}$ in the first image can be computed as well. Moreover, the epipolar constraint (11.8) brings, for each pair of corresponding image points $\mathbf{p}$ and $\mathbf{p}'$ in I and I' respectively, one homogeneous equation $\mathbf{p}'^t F \, \mathbf{p} = 0$ that is linear in the entries of the fundamental matrix F. Hence, *F can be computed linearly, up to a non-zero scalar factor, from (at least) 8 point correspondences between the two images.* Due to the presence of noise in the images, the matrix F computed from these point correspondences generally will not be of rank 2. Imposing the rank 2 constraint for the computation of F, however, results in a non-linear criterium.

The epipolar relations between corresponding point projections are an important tool for the practical realisation of stereo based 3D reconstruction. Again, once corresponding points have been identified, further 3D reconstruction follows from triangulation. Knowing all internal and external camera parameters as well as the corresponding point pair $(\mathbf{p}, \mathbf{p}')$ one has 6 linear equations

$$\begin{aligned} \mathbf{P} &= \mathbf{C} + \mu R K^{-1} \mathbf{p} \\ \mathbf{P}' &= \mathbf{C}' + \mu' R' K'^{-1} \mathbf{p}' \end{aligned}$$

in 5 unknowns (the 3 coordinates X, Y, and Z of the point and μ and μ'.

A complicating factor is that these equations may have no solution because of image noise and discretisation errors. Such deviations from the ideal case are the cause that the computed projection rays may not intersect. One usually takes the point in the middle of the the points of the rays where they are closest to eachother.

11.2.3 * Relations between three views: the trifocal constraints

Traditional stereo uses two images to reconstruct a scene in 3D. Obviously, one could use more than just two images. Suppose that three images of the same scene are given. In the case of two images, the epipolar constraint yields restrictions on where corresponding points may lie. Are there similar relations between the position of a point's projections in three images?

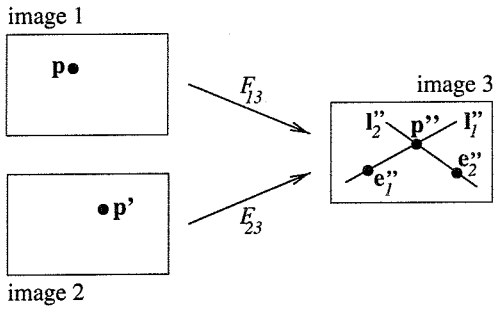

Figure 11.9: *The point* $\mathbf{p}''$ *in the third image corresponding to the points* $\mathbf{p}$ *in the first and* $\mathbf{p}'$ *in the second view is the intersection of the epipolar lines* $\mathbf{l}_1''$ *and* $\mathbf{l}_2''$ *corresponding to* $\mathbf{p}$ *and* $\mathbf{p}'$ *respectively.*

Suppose a point $\mathbf{P}$ projects to the points $\mathbf{p}$, $\mathbf{p}'$, and $\mathbf{p}''$ in the three images, resp. If the fundamental matrices F_{13} and F_{23} between the first and the third, respectively the second and the third, view are known, then the position of the point $\mathbf{p}''$ is at the intersection of the epipolar lines $\mathbf{l}_1''$ of $\mathbf{p}$ and $\mathbf{l}_2''$ of $\mathbf{p}'$ in the third image. This is depicted in Figure 11.9. Unfortunately, this construction breaks down when the epipolar lines $\mathbf{l}_1''$ and $\mathbf{l}_2''$ coincide. This happens for scene points belonging to the plane defined by the three centers of projection. Moreover, the construction is also poorly conditioned for scene points that are close to this plane. The construction even fails to work for all points if the three centers are collinear. Fortunately, there are constraints between the projections in three views that are algebraically independent of the epipolar constraints, i.e. essentially different constraints. These are the so-called "trifocal constraints" and they are the topic of this section.

First, a geometrical interpretation is given. The trifocal constraints describe the incidence relations between image points and lines in three views. The fundamental relationship is depicted in Figure 11.10. Suppose $\mathbf{p}$, $\mathbf{p}'$ and $\mathbf{p}''$ are corresponding projections ons scene point $\mathbf{P}$ in three views I, I' and I'', as before. Consider an arbitrary line $\mathbf{l}''$ in the third image. The rays of sight in the world, creating the line $\mathbf{l}''$ in I'', rule out a world plane π'' containing the center of projection $\mathbf{C}''$ of the third camera. Moreover, if the line $\mathbf{l}''$ passes through $\mathbf{p}''$, then this projecting plane π'' must also contain the scene point $\mathbf{P}$. Similarly, every line $\mathbf{l}'$ through the point $\mathbf{p}'$ in the second view defines a projecting plane π' in the world which also contains $\mathbf{P}$. Consequently, the projecting planes π' and π'' intersect in a line $\mathbf{L}$ through the point $\mathbf{P}$ in the scene. The projection $\mathbf{l}$ of $\mathbf{L}$ in the first image therefore must contain the image point $\mathbf{p}$ in I. The trifocal constraints express this incidence relation algebraically. But first, we fix the following notation: An image line $\mathbf{l}$ with equation $a\,x + b\,y + c = 0$ is represented in the sequel by the column vector $(a, b, c)^t \in I\!R^3$. As the triple $(\lambda a, \lambda b, \lambda c)^t \in I\!R^3$ represents the same line for every non-zero $\lambda \in I\!R$, these column vectors have to be interpreted as the *homogeneous coordinates* of the given line.

To derive the algebraic formulation of the trifocal constraints, the first step is to express

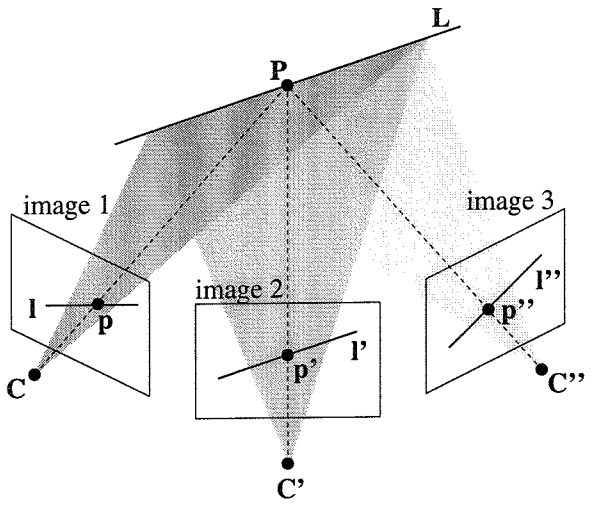

Figure 11.10: *Two arbitrary lines* $\mathbf{l}'$ *and* $\mathbf{l}''$ *through corresponding points* $\mathbf{p}'$ *and* $\mathbf{p}''$ *in respectively the second and the third image define a line* $\mathbf{L}$ *in the scene whose projection* $\mathbf{l}$ *in the first image contains* $\mathbf{p}$.

the equation for the projecting plane of a line in the image. This plane contains all the points with image projections on the line and hence

$$\mathbf{l}^t K R^t (\mathbf{P} - \mathbf{C}) = 0 \tag{11.9}$$

Now, let $\mathbf{P}$ be the scene point of which $\mathbf{p}$, $\mathbf{p}'$ and $\mathbf{p}''$ are the projections in the different images. According to eq. 11.9 $\mathbf{P}$ belongs to the projecting plane π' and π'' generating the lines $\mathbf{l}'$ and $\mathbf{l}'$, resp., if and only if

$$\begin{array}{rcl} \mathbf{l}'^t K' R'^t (\mathbf{P} - \mathbf{C}') & = & 0 \ , \\ \mathbf{l}''^t K'' R''^t (\mathbf{P} - \mathbf{C}'') & = & 0 \ . \end{array} \tag{11.10}$$

$\mathbf{P}$ must also lie on the ray of sight observing $\mathbf{p}$ in the first camera; i.e.

$$\mathbf{P} = \mathbf{C} + \mu R K^{-1} \mathbf{p} \qquad \text{for some } \mu \in I\!R, \tag{11.11}$$

Substituting this expression in the equations (11.10) of the planes yields

$$\left\{ \begin{array}{rcl} \mathbf{l}'^t K' R'^t (\mathbf{C} - \mathbf{C}') + \mu\, \mathbf{l}'^t K' R'^t R K^{-1} \mathbf{p} & = & 0 \\ \mathbf{l}''^t K'' R''^t (\mathbf{C} - \mathbf{C}'') + \mu\, \mathbf{l}''^t K'' R''^t R K^{-1} \mathbf{p} & = & 0 \end{array} \right. \tag{11.12}$$

Remember from formula (11.4) that $K'R'^t(\mathbf{C} - \mathbf{C}') = \rho'_e \mathbf{e}'$ gives the epipole $\mathbf{e}'$ of the first camera in the second view; and that $K''R''^t(\mathbf{C} - \mathbf{C}'') = \rho''_e \mathbf{e}''$ gives the epipole $\mathbf{e}''$ of this camera in the third image. We Define $A = \frac{1}{\rho'_e} K'R'^t R K^{-1}$ as before (section 11.2.2) and similarly $B = \frac{1}{\rho''_e} K''R''^t R K^{-1}$, Then, after dividing by the nonzero scalars ρ'_e and ρ''_e respectively, system (11.12) becomes

$$\left\{ \begin{array}{rcl} \mathbf{l}'^t \mathbf{e}' + \mu\, \mathbf{l}'^t A \mathbf{p} & = & 0 \\ \mathbf{l}''^t \mathbf{e}'' + \mu\, \mathbf{l}''^t B \mathbf{p} & = & 0 \end{array} \right. \tag{11.13}$$

Eliminating the unknown scalar $\mu \in \mathbb{R}$ in the previous two equations, gives

$$\left(\mathbf{l}'^t A\mathbf{p}\right)\left(\mathbf{l}''^t\mathbf{e}''\right) - \left(\mathbf{l}'^t\mathbf{e}'\right)\left(\mathbf{l}''^t B\mathbf{p}\right) = 0 \ . \tag{11.14}$$

As $\mathbf{l}'^t\mathbf{e}' = \mathbf{e}'^t\mathbf{l}'$ and $\mathbf{l}''^t(B\mathbf{p}) = (B\mathbf{p})^t\,\mathbf{l}''$, equation (11.14) can be rewritten as

$$\mathbf{l}'^t\left[(A\mathbf{p})\ \mathbf{e}''^t - \mathbf{e}'\ (B\mathbf{p})^t\right]\mathbf{l}'' = 0 \ . \tag{11.15}$$

This is a relation between the point $\mathbf{p}$ and the lines $\mathbf{l}'$ and $\mathbf{l}''$. The expression between square brackets is a 3×3-matrix $[T\mathbf{p}]$ with (i,j)th entry

$$[T\mathbf{p}]_{ij} = (A\mathbf{p})_i\,(\mathbf{e}'')_j - (\mathbf{e}')_i\,(B\mathbf{p})_j \ . \tag{11.16}$$

With $\mathbf{p} = (x, y, 1)^t$ one can deduce that

$$T_k^{ij} = a_{ik}(\mathbf{e}'')_j - b_{jk}(\mathbf{e}')_i \qquad \text{for } 1 \le i, j, k \le 3. \tag{11.17}$$

In summary, for every three views I, I' and I'' of a static scene, there exists a $3\times3\times3$-tensor $T = \left(T_k^{ij}\right)_{1\le i,j,k\le 3}$, called the *trifocal tensor* of the image triple (I, I', I''), with the following property: If $\mathbf{p} \in I$, $\mathbf{p}' \in I'$ and $\mathbf{p}'' \in I''$ are corresponding points in the images, then for every line $\mathbf{l}'$ through $\mathbf{p}'$ in I' and for every line $\mathbf{l}''$ through $\mathbf{p}''$ in I'',

$$\mathbf{l}'^t\,[T\mathbf{p}]\,\mathbf{l}'' = 0 \ , \tag{11.18}$$

where $[T\mathbf{p}]$ is the 3×3-matrix whose (i,j)th entry is

$$[T\mathbf{p}]_{ij} = T_1^{ij}x + T_2^{ij}y + T_3^{ij} \tag{11.19}$$

and with $\mathbf{p} = (x, y, 1)^t$. Moreover, if, with the notations as before, $A = \frac{1}{\rho'_e}K'R'^tRK^{-1}$ and $B = \frac{1}{\rho''_e}K''R''^tRK^{-1}$, then

$$[T\mathbf{p}] = (\,A\mathbf{p}\,)\,\mathbf{e}''^t - \mathbf{e}'\,(\,B\mathbf{p}\,)^t \ , \tag{11.20}$$

where $\mathbf{e}'$ and $\mathbf{e}''$ are the epipoles of the first camera in the second and the third view respectively; and, the entries T_k^{ij} of the trifocal tensor T are given by

$$T_k^{ij} = a_{ik}(\mathbf{e}'')_j - b_{jk}(\mathbf{e}')_i \qquad \text{for } 1 \le i, j, k \le 3, \tag{11.21}$$

where $(\mathbf{e}')_i$ and $(\mathbf{e}'')_j$ are the ith and jth entry of the epipoles $\mathbf{e}'$ and $\mathbf{e}''$, and with a_{ij} and b_{ij} being the (i,j)th entry of the matrices A and B respectively.

The trifocal constraint (11.18) expresses a trilinear algebraic relation between (the homogeneous coordinates of) 1 image point and 2 image lines. In tensor terminology, this is expressed by stating that the trifocal tensor has one *covariant* and two *contravariant* indices, viz. k and ij respectively. Furthermore, it should also be emphasized that the lines $\mathbf{l}'$ and $\mathbf{l}''$ in the theorem do *not* need to be *corresponding* lines in the images.

11.2.4 The correspondence problem

Finally, we have to discuss ways of solving the correspondence problem, i.e. choosing a point in the second image as corresponding to a specified point in the first image or vice versa. We already derived the epipolar constraint. If one has more information on the possible range values, one can often restrict the analysis to a segment of the epipolar line. The further selection of a point along the epipolar line is typically carried out following one of two strategies. The approach can be *correlation-based* or *feature-based* and may apply *regularization.*

The correlation method considers a window around the point in the first image. This window is used as a template to be shifted over the second image. Where the correlation between this template and the underlying intensity pattern of the second image is maximal, the corresponding point is assumed to lie. Besides the problem of computational cost, this method may fail because of distortions and changes in intensity when changing viewpoint. Small windows can be handled faster, but the result can be highly noisy and several wrong correspondences might result. On the other hand, larger windows less accurately localize the corresponding point, due to the correlation output having broad peaks, but better withstand noise.

A second method is based on matching features in both images. One could e.g. look for corners or sharp edges in both images and try to retrieve correspondence only for these features. Not so many of such features will occur typically, but still there is the danger of finding several corners or edges along the epipolar line. In that case we have to further disambiguate the correspondence. This can be done by looking at additional attributes such as edge contrast and polarity. Another useful constraint might be that most of the time, the ordering between edges in both views will not change, a constraint which has to be handled carefully though. The advantage of having not too many candidates in the other image (not so many corners or edges) also brings the disadvantage of only obtaining distance at so few a number of points. Distances at other points will then have to be interpolated one way or another.

Finally, these two methods are not the only ones. Some methods use a regularization type of approach, e.g. suggesting matches for points with almost identical intensity. Together with a smoothness assumption on depth this can be enough to get a dense range image. Smoothness can also be used as an additional consideration with the two methods described here of course and we will return to this issue in chapter 14.

11.2.5 II - Reconstruction from uncalibrated image sequences

full text Not available yet

In order to put the Longuet-Higgins equation to work and obtain the epipolar line in the second image for a given point in the first, the relative position and orientation of the cameras has to be known, as well as their focal lengths and pixel sizes. Procedures to extract these parameters are referred to as *calibration techniques.* Calibration is a painstaking process and quite some recent research was devoted to bypass it. By now, it is

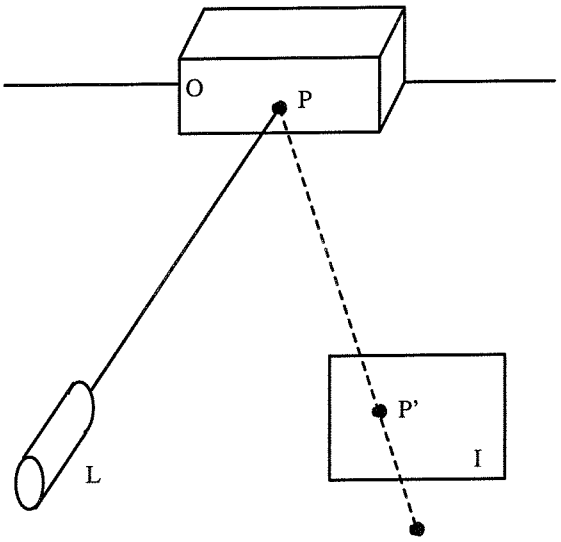

Figure 11.11: *The triangulation principle used already with stereo, can also be used in an active configuration. The laser L projects a ray of light on the object O. The intersection point P with the object is viewed by a camera and forms the spot P' on its image plane I. This information suffices for the calculation of the three-dimensional coordinates of P, assuming that the laser-camera configuration is known.*

known that some information can be retrieved from uncalibrated stereo pairs. In particular, three-dimensional projective reconstructions (rather than Euclidean reconstructions) can be made from arbitrary stereo pairs and more precise, affine reconstructions are possible from a pair of cameras where one camera is translated with respect to the other. The concepts of plane projective and affine transformations are explained in chapter 15. The Longuet-Higgins equation being of no use with uncalibrated stereo, correspondence search can turn out a hard problem.

11.3 I - Active triangulation

The correspondence problem with stereo vision can be solved by replacing one of the cameras by a projection devise. Such sensor configuration is therefore a member of the active class (cf. the 3D taxonomy). If the projector is e.g. a laser, then a single spot is projected onto the scene. The spot formed by reflection on the first surface it encounters will be easily detectable in the image. If we know the position and orientation of both the laser ray and the camera's projecting ray, then the corresponding three-dimensional point is known to be their intersection. Its three-dimensional position can therefore be calculated. The principle is illustrated in fig. 11.11.

The problem is that knowledge of one point is hardly sufficient in most cases. By scanning the environment with the laser, requiring quite precise mechanical apparatus (e.g. by steering rotating mirrors that reflect the laser light into controlled directions), a complete

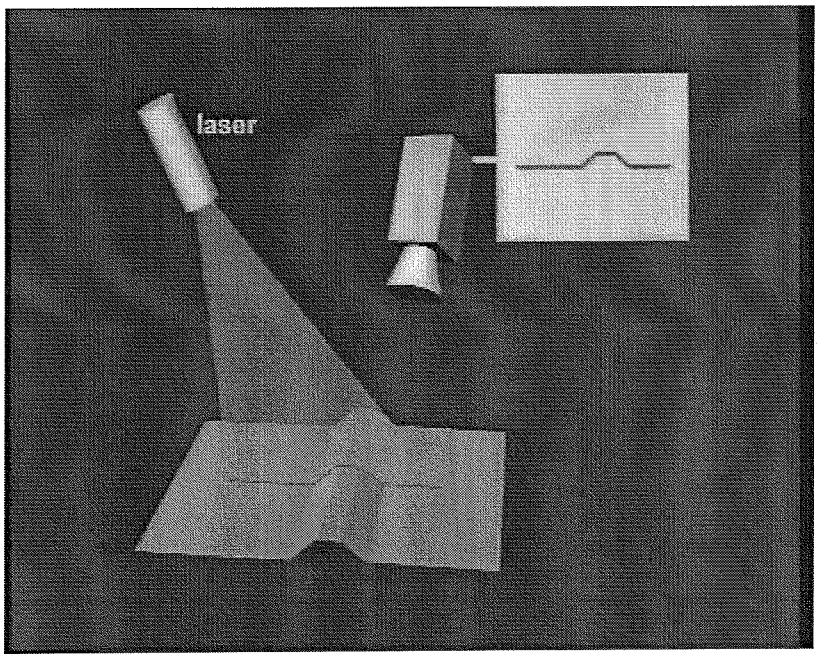

Figure 11.12: *If the triangulation configuration is somewhat altered, i.e. if the laser spot is turned into a line (e.g. by the use of a cylindrical lens), then scanning can be restricted to a one-directional motion, transversal to the line.*

three-dimensional view is generated. For each position of the spot a new image has to be taken. However, the spot can without much difficulty be extended to a line. Scanning has then only to be performed in one direction and the number of images can be reduced. The configuration is depicted in fig. 11.12. Further extension to two dimensional patterns is less straightforward as the points cannot be identified from a uniform illumination pattern. Encoding point positions within one pattern has already been proposed for 64 x 63 range images. An example is shown in fig. 11.13. The figure shows a face on which the special pattern is projected. The pattern is such that each vertical column has its own disctinctive signature. Three-dimensional reconstructions obtained with this technique are shown in figure 11.14. Other solutions will use a limited set of masks which are projected subsequently and together build the position codes. Most of such patterns are binary for reasons of robustness. Fig. 11.15 gives an example.

Triangulation with line projection has found wide industrial application, especially with seam welding. The line is projected perpendicular to the seam trajectory. The deformation of the line as viewed by the sensor gives the necessary information about the position of the seam and its form. Both the vision module and the welding system are mounted on the robot arm. In order to prevent the welding arc from adversely affecting image quality, it is advisable to use intense monochromatic light in combination with a narrow bandpass filter for the same frequency in front of the camera. Furthermore, the camera should best look ahead of the arc. The detection of the seam can be carried out using template matching (see chapter 16) or line fitting. Information from several lines can be combined, so as to

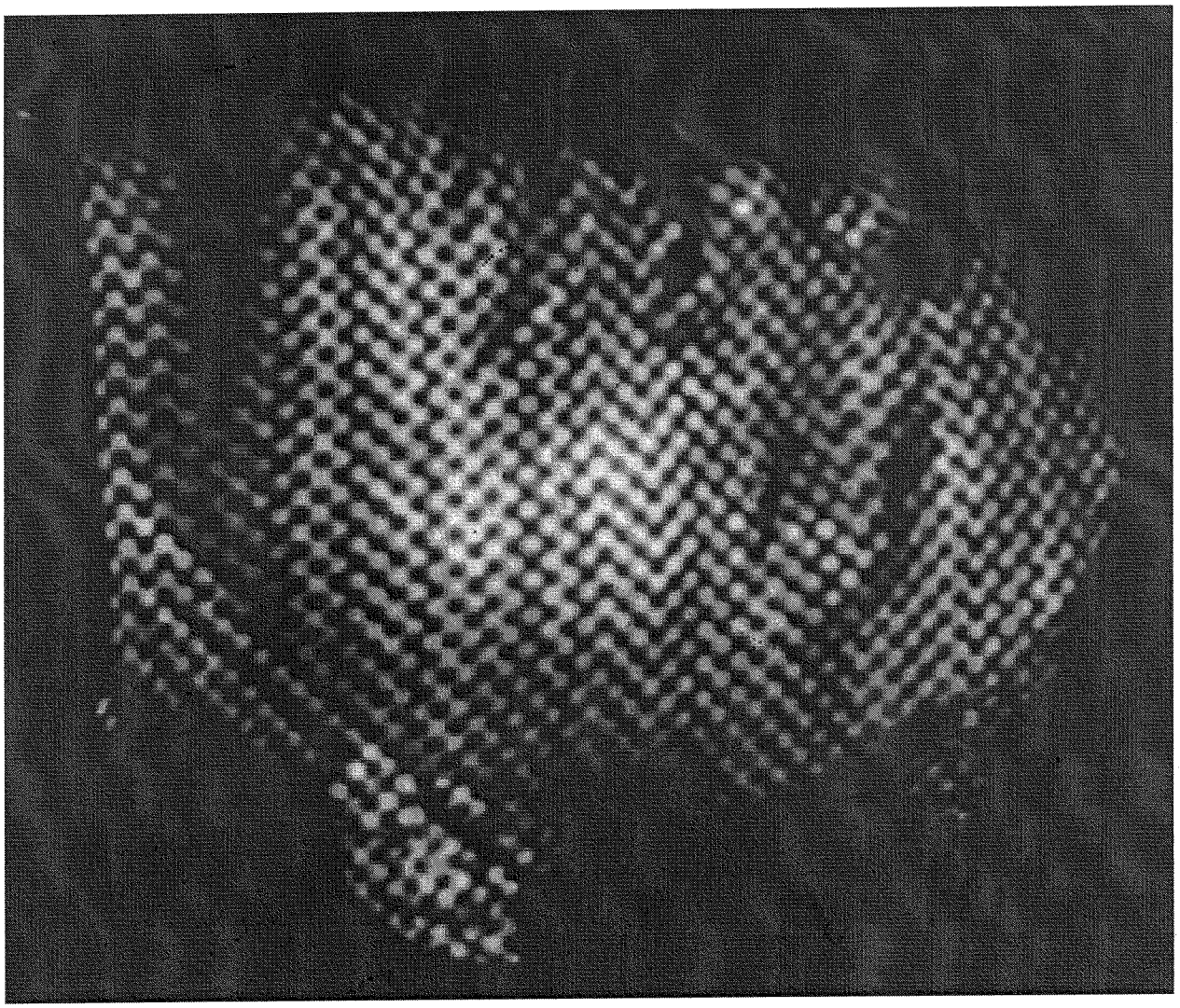

Figure 11.13: *Example of one-shot active range technique. A special pattern allowing disambiguition of its different vertical stripes is here projected on a face.*

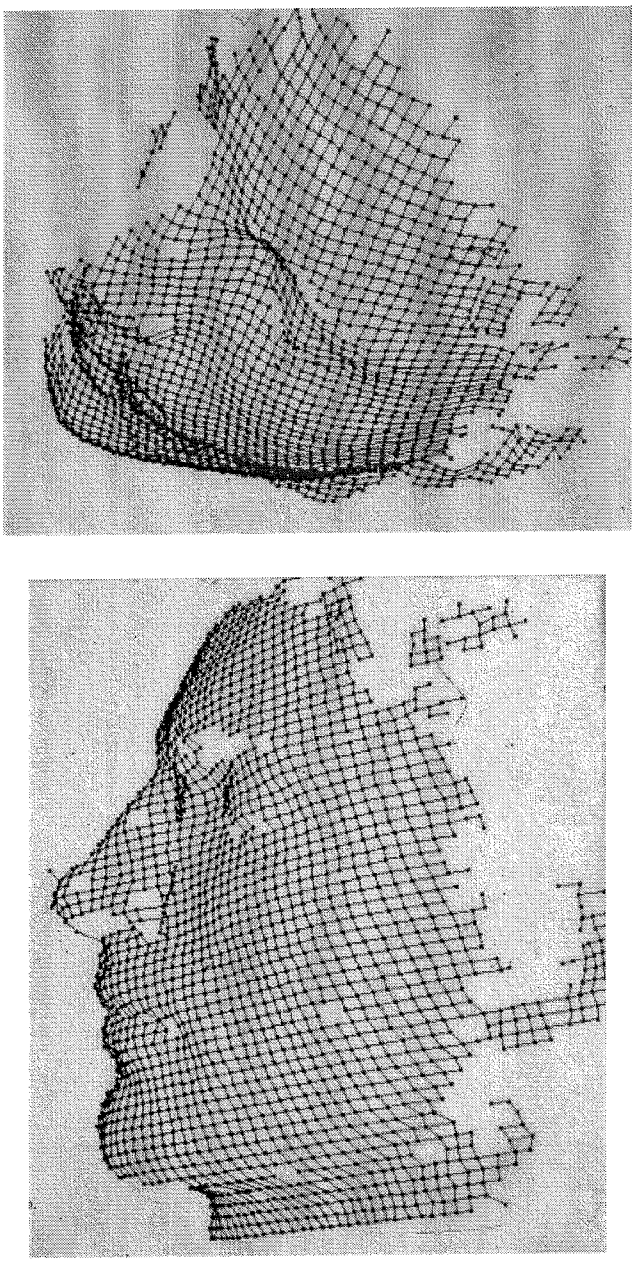

Figure 11.14: *Two views of the 3D description obtained with the active method of fig. 11.13.*

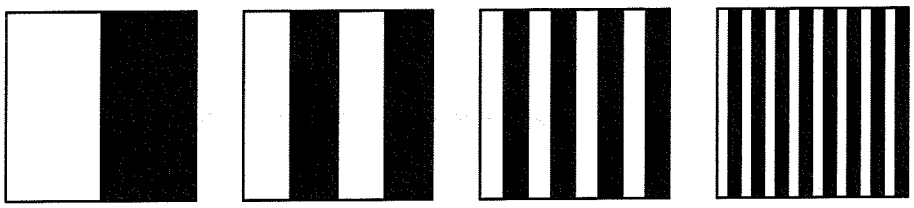

Figure 11.15: *Series of masks that can be projected for active stereo applications. Subsequent masks contain ever finer stripes. Each of the masks is projected and for a point in the scene the sequence of black/white values is recorded. The subsequent bits obtained that way characterize the vertical position of the points. The accuracy that is required imposes the number of such masks that has to be used.*

describe the weld area with planes or other primitives. The methods based on thresholding to find the laser line are vulnerable to multiple reflections, which often occur in industrial settings. This problem can be attenuated by convolution based line detection (see chapter 8).

Sometimes one uses a camera stereo pair in combination with a laser. The laser position and orientation of projection, usually varied rapidly in time, then no longer are to be known. The correspondence problem is also largely reduced as long as the laser spot or line stands out in both images. This technique is thus a kind of feature-based stereo, where the features to be detected are artificially created by projection.

11.4 I - Other methods

A detailed discussion of all methods used for range extraction is out of the scope of this introductory course. Nevertheless, we will succinctly describe the different methods mentioned in the 3D taxonomy of fig. 11.1. We start off with some active techniques.

The basic principle of time-of-flight sensors is the measurement of the time a modulated light signal - usually from a laser - needs to travel before returning to the sensor. This time is proportional to the distance from the object. This is an active, uni-directional approach. Depending on the type of waves used, one calls such devices *radar* (electromagnetic waves of low frequency), *sonar* (acoustic waves), or *optical radar* (optical waves). A first category uses pulsed waves and measures the delay between the transmitted and the received pulse. A second category is used for smaller distances and measures phase shifts. The low level of the returning signal and the high bandwith required for detection lead to a small signal to noise ratio. Measurement problems and health hazards with lasers can be alleviated by the use of ultrasound. The bundle has a much larger opening angle then, and resolution decreases.

"Shape from shading" uses directional lighting with known direction and intensity. Gray levels of object surface patches then convey information on their three-dimensional orientation. This process does not only require information on the sensor-illumination configuration, but also on the reflection characteristics of the surface. The complex re-

lationship between gray levels and surface orientation can theoretically be calculated in some cases, but is usually derived from experiments and then stored in "reflectance maps" for table-lookup. The orientation of an isolated patch cannot be derived uniquely from its intensity. Local information from different patches is combined with extra assumptions on surface smoothness. With smooth surfaces the surface normals at the occluding contours can easily be retrieved from the image contour normals, yielding accurate boundary conditions.

Constraint propagation can be made unnecessary by using different positions of the light source ("photometric stereo"). This eliminates the summation of errors typical for methods working their way inwards from the rim, where the boundary constraints described earlier serve as initial data to be propagated via continuity and restrictions on local orientation. We therefore classified this approach as multi-directional. But as we have also mentioned, single source shape-from-shading techniques exist and are still the subject of intensive research, the rationale being that human vision seems to be quite successful in deriving shape from intensity variations. It is also interesting to note that when the light sources are given different colours the differently directioned sources can be applied at the same time. The interpretation of shadows has been a subject in its own right as well.

Passive unidirectional methods include shape from texture, shape from contour, and shape from occlusion. Also these methods don't yield true range data. The former two allow the determination of surface orientation, not distance, except for exceptional cases where pattern scale can be used to estimate it. The latter only yields information such as 'object 1 is closer than object 2'.

Shape from texture assumes a slanted and tilted surface to have a homogeneous texture. Systematic inhomogeneity of the imaged texture (e.g. anisotropy in the statistics of edge orientations) is regarded as the result of projection. Surface orientations which allow the original texture to be maximally isotropic or homogeneous are selected. Figure 11.16 shows an example of a textured scene. The impression of an undulating surface is immediate. The right hand side of the figure shows the results for a shape-from-texture algorithm that uses the regularity of the pattern for the estimation of the local surface orientation. Actually, what is assumed here is a square shape of the pattern's period (i.e. a kind of discrete isotropy). This assumption suffices to calculate the local surface orientation. The ellipses represent circles with such calculated orientation of the local surface. The small stick at their center shows the normal to the surface.

Shape from contour makes similar assumptions about the true shape on the basis of a contour's projection. An example is the maximization of area over perimeter squared over all possible deprojections. An ellipse would be deprojected to a circle for this measure, consistent with human vision. Humans also tend to see ellipses as tilted circles. The deviation from the selected "true" shape then yields information on three-dimensional orientation. Several other selection principles have been put forward, but their discussion would lead us too far. In fact, regular figures with at least a 3-fold rotational symmetry and that are known (or assumed) to be regular by the observer all yield a single orientation that could make up for the deformation in the image, except for the mirror reversal with respect to the image plane (assuming that perspective distortions are too small to be picked up).

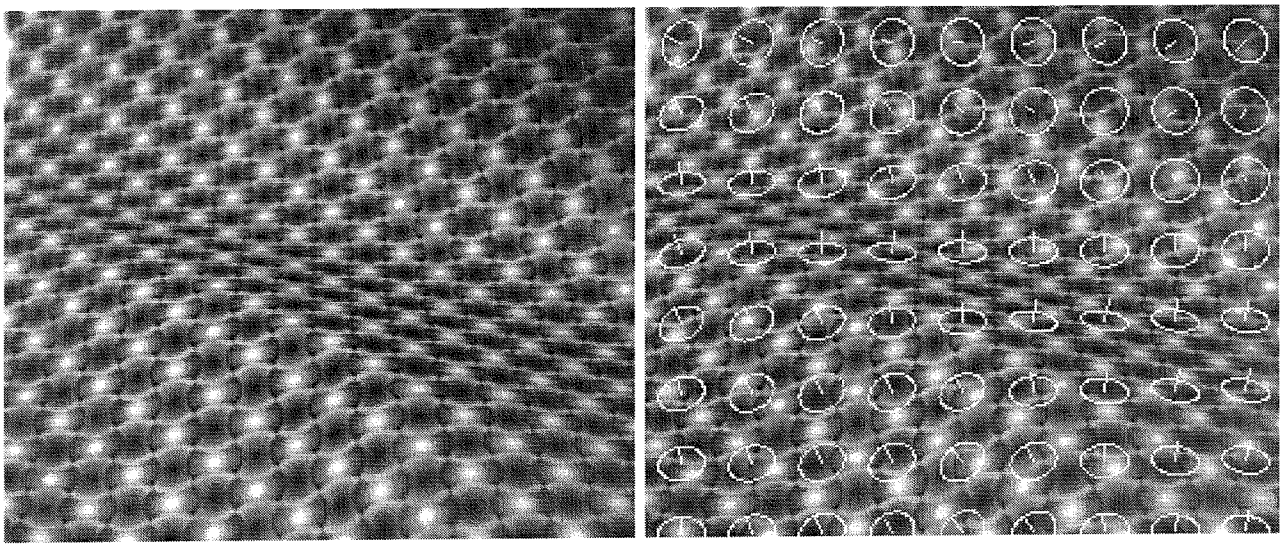

Figure 11.16: *Left: the regular texture yields a clear perception of a curved surface. Right: the result of a shape-from-texture algorithm.*

This is but a special case of the more general result, that a unique orientation (up to the mirror reflection) also results when two copies of a shape are observed in the same plane (with the exception where their orientation differs by 0° or 180° in which case nothing can be said on the mere assumption that both shapes are identical).

In shape from occlusion relationships for the positions of objects like "behind" and "in-front-of" can be deduced in the sense that, if occlusion occurs, it is the nearest object that will occlude the others.