PROF R CIPOLLA

Engineering Tripos Part IIB FOURTH YEAR

Module 412: Computer Vision and Robotics

Solutions to 2010 Tripos Paper

1. **Feature detection**

(a) $I(x, y)$ is a function of many variables, including the position of the camera; the properties of the lens and the CCD; the shape of the structures in the scene; the nature and distribution of light sources; and the reflectance properties of the visible surfaces. [10%]

(b) (i) To preserve the mean intensity of the image, a should be set to 1+4+6+4+1 = 16. [10%]

(ii) The smoothed pixels $s(x)$ are obtained by discrete convolution:

$$s(x) = \sum_{i=-2}^{2} g(i)I(x-i)$$

where $g(-2) = 1/16$, $g(-1) = 1/4$, $g(0) = 3/8$, $g(1) = 1/4$ and $g(2) = 1/16$. Intensity discontinuities are localised by differentiating the smoothed pixels to obtain the gradient $d(x)$. This can be achieved by convolution with the kernel $[\,1 \quad -1\,]$:

$$d(x) = s(x) - s(x-1)$$

Intensity discontinuites are then localised at local maxima of $d(x)$. [20%]

(iii) Averaging neighbouring pixels once is equivalent to convolution with the kernel $[\,\frac{1}{2} \quad \frac{1}{2}\,]$. Averaging twice is equivalent to convolution with the kernel

$$\left[\begin{array}{cc} \frac{1}{2} & \frac{1}{2} \end{array}\right] * \left[\begin{array}{cc} \frac{1}{2} & \frac{1}{2} \end{array}\right] = \left[\begin{array}{ccc} \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{array}\right]$$

Averaging three times is equivalent to convolution with the kernel

$$\left[\begin{array}{cc} \frac{1}{2} & \frac{1}{2} \end{array}\right] * \left[\begin{array}{ccc} \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{array}\right] = \left[\begin{array}{cccc} \frac{1}{8} & \frac{3}{8} & \frac{3}{8} & \frac{1}{8} \end{array}\right]$$

Averaging four times is equivalent to convolution with the kernel

$$\left[\begin{array}{cc} \frac{1}{2} & \frac{1}{2} \end{array}\right] * \left[\begin{array}{cccc} \frac{1}{8} & \frac{3}{8} & \frac{3}{8} & \frac{1}{8} \end{array}\right] = \left[\begin{array}{ccccc} \frac{1}{16} & \frac{1}{4} & \frac{3}{8} & \frac{1}{4} & \frac{1}{16} \end{array}\right]$$

which is the kernel in (b). So 1D smoothing with the kernel in (b) is equivalent to successively averaging neighbouring pixels. Four averaging operations are required. [30%]

(c) [Book work] [30%]

Assessor's comment:

Good understanding and well-attempted by most candidates.

2. **Camera calibration and vanishing points**

(a) The relationship is valid under the assumption that the image is formed by a "pinhole camera", such that rays pass through a single point (the optical centre) before striking the image plane. The relationship does not account for nonlinear distortion, which affects all real cameras to some extent.

Geometrically, s can be thought of as a scale factor, controlling the size of the image formed by an object in the world. It depends on the distance Z_c of the object from the camera. The elements p_{ij} describe an isometry (a rotation and translation), followed by perspective projection onto an image plane and sampling of the plane by a CCD array.

Algebraically, s and the elements p_{ij} allow the imaging process to be expressed as a linear relationship in homogeneous coordinates. In Cartesian coordinates the perspective image formation process cannot be expressed linearly, requiring a division by Z_c. [20%]

(b) The process of estimating the elements p_{ij} is known as *camera calibration.* It is necessary to observe a scene where the world positions of several distinguished features are known. For example, we might set up the camera to view a calibrated grid of some sort.

P can be estimated by observing the images of known 3D points. Each point we observe gives us a pair of equations:

$$u = \frac{su}{s} = \frac{p_{11}X + p_{12}Y + p_{13}Z + p_{14}}{p_{31}X + p_{32}Y + p_{33}Z + p_{34}}$$

$$v = \frac{sv}{s} = \frac{p_{21}X + p_{22}Y + p_{23}Z + p_{24}}{p_{31}X + p_{32}Y + p_{33}Z + p_{34}}$$

Since we are observing a known scene, we know X, Y, and Z, and we observe the pixel coordinates u and v in the image. So we have two linear equations in the unknown camera parameters. Since there are 11 unknowns (the overall scale of P does not matter), we need to observe at least 6 points to calibrate the camera.

The equations can be solved using orthogonal least squares. First, we write the equations in matrix form:

$$\mathbf{Ap} = \mathbf{0}$$

where $\mathbf{p}$ is the 12×1 vector of unknowns (the twelve elements of P), A is the $2n \times 12$ matrix of coefficients and n is the number of observed calibration points. The orthogonal least squares solution corresponds to the eigenvector of $\mathrm{A}^T\mathrm{A}$ with the smallest corresponding eigenvalue.

It is essential that the calibration points are not coplanar, since otherwise we are not exercising all the degrees of freedom of the camera model and the set of linear

equations will not be independent. Consequently, the least squares procedure will not find a unique solution (there will be a degenerate zero eigenvalue).

Given the projective camera matrix, we can attempt to recover the intrinsic and extrinsic parameters using QR decomposition. Writing

$$\begin{aligned} \mathrm{P} &= \begin{bmatrix} fk_u & 0 & u_0 & 0 \\ 0 & fk_v & v_0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \left[\begin{array}{ccc|c} & \mathrm{R} & & \mathbf{T} \\ \hline 0 & 0 & 0 & 1 \end{array}\right] = \left[\begin{array}{ccc|c} fk_u & 0 & u_0 & \\ 0 & fk_v & v_0 & \mathrm{R} \quad \mathbf{T} \\ 0 & 0 & 1 & \end{array}\right] \\ &= \mathrm{C}\left[\begin{array}{c|c} \mathrm{R} & \mathbf{T} \end{array}\right] = \left[\begin{array}{c|c} \mathrm{CR} & \mathrm{C}\mathbf{T} \end{array}\right] \end{aligned}$$

it is apparent that we need to decompose the left 3×3 sub-matrix of P into an upper triangular matrix C and an orthogonal (rotation) matrix R. This can be achieved using QR decomposition. $\mathbf{T}$ can then be recovered using

$$\mathbf{T} = \mathrm{C}^{-1}\left[\begin{array}{cccc} p_{14} & p_{24} & p_{34} \end{array}\right]^T$$

If the camera we're calibrating is high quality (so it does something approaching a perspective projection onto a well mounted CCD array) and the calibration has been properly performed, we should find that the recovered intrinsic matrix C has a zero in the middle of its top row, as expected. If we scale the matrix C so that it has a 1 in its lower right hand corner (this is acceptable, since the overall scale of P does not matter), then we can recover the principle point (u_0, v_0) by looking at c_{13} and c_{23}, and the products fk_u and fk_v by looking at c_{11} and c_{22}. It is not possible to decouple the focal length from the pixel scaling factors. [50%]

(c) Distant points on lines parallel to the world X-axis can be represented in homogeneous coordinates as

$$\tilde{\mathbf{X}} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ 0 \end{bmatrix}$$

Applying the projection matrix P, we find the image of these points is

$$\begin{bmatrix} su \\ sv \\ s \end{bmatrix} = \begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ p_{31} & p_{32} & p_{33} & p_{34} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ 0 \end{bmatrix}$$

The vanishing point of lines which are parallel to the world X-axis for example is therefore $(u, v) = (p_{11}/p_{31}, p_{21}/p_{31})$.

[30%]

Assessor's comment:

Part (a) and (b) were attempted. Most candidates failed to answer the component on the projection of points at infinity and vanishing points.

3. **Planar Homography**

(a) (i) When the camera is viewing a plane, the relationship between pixels and world positions is given by

$$\begin{bmatrix} su \\ sv \\ s \end{bmatrix} = \begin{bmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{bmatrix} \begin{bmatrix} X \\ Y \\ 1 \end{bmatrix}$$

or $\tilde{\mathbf{w}} = \mathrm{P}\tilde{\mathbf{X}}^p$ for short. For a second image of the same point, we have $\tilde{\mathbf{w}}' = \mathrm{P}'\tilde{\mathbf{X}}^p$. It follows that $\tilde{\mathbf{w}}' = \mathrm{P}'\mathrm{P}^{-1}\tilde{\mathbf{w}} = \tilde{\mathbf{w}}$, where $\equiv \mathrm{P}'\mathrm{P}^{-1}$ is a 3×3 matrix. Hence the relationship between points in the original image and corresponding points in the second image is a 2D projective transformation. [15%]

(ii) Assume, without loss of generality, that before the camera is rotated, the camera is aligned with the world coordinate system and hence

$$\tilde{\mathbf{w}} = \left[\mathrm{I} \mid \mathbf{O} \right] \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} = \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \mathbf{X}$$

where K is the 3×3 matrix of intrinsic camera parameters:

$$= \begin{bmatrix} fk_u & 0 & u_0 \\ 0 & fk_v & v_0 \\ 0 & 0 & 1 \end{bmatrix}$$

It follows that

$$\mathbf{X} =^{-1} \tilde{\mathbf{w}}$$

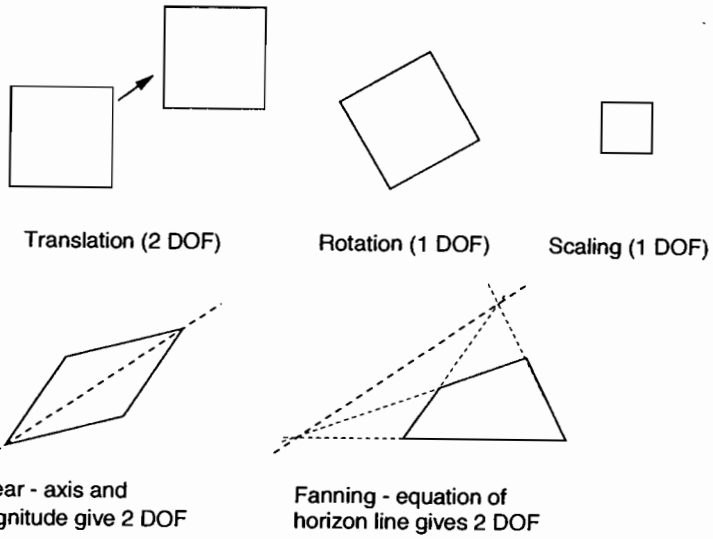
After rotating by R about the optical centre, the same world point $\mathbf{X}$ projects to a different image point $\tilde{\mathbf{w}}'$ as follows:

$$\tilde{\mathbf{w}}' = \left[\mathrm{R} \mid \mathbf{O} \right] \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} = \mathrm{R} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \mathrm{R}\mathbf{X} = \mathrm{R}^{-1}\tilde{\mathbf{w}} = \tilde{\mathbf{w}}$$

where $\equiv \mathrm{R}^{-1}$. Hence the relationship between points in the original image and corresponding points in the second image is a 2D projective transformation. [15%]

(b) Since the transformation operates on homogeneous coordinates, the overall scale of the transformation matrix does not matter and we could, for instance, set t_{33} to 1. The transformation therefore has 8 degrees of freedom.

The image of a square could take any of the forms shown on the next page.



[30%]

Assessor's comment:

The most disappointing question with a large number of students not knowing all the 8 DOFs of a 2D projective transformation. Part (c) was very variable in quality even though it was bookwork.

4. **Stereo vision**

(b) The fundamental matrix F relates points in the left and right images of a stereo pair:

$$\tilde{\mathbf{w}}'^T \mathrm{F} \tilde{\mathbf{w}} = 0$$

where $\tilde{\mathbf{w}} = (u, v, 1)$ are the point's pixel coordinates in the left image, and $\tilde{\mathbf{w}}'$ are the coordinates of the corresponding point in the right image. The constraint arises from the requirement that the rays from the two cameras' optical centres through $\tilde{\mathbf{w}}$ and $\tilde{\mathbf{w}}'$ must intersect at a point in space. F has zero determinant and can be determined only up to scale.

F can be estimated from point correspondences. Each point correspondence $\tilde{\mathbf{w}} \leftrightarrow \tilde{\mathbf{w}}'$ generates one constraint on F:

$$\begin{bmatrix} u' & v' & 1 \end{bmatrix} \begin{bmatrix} f_{11} & f_{12} & f_{13} \\ f_{21} & f_{22} & f_{23} \\ f_{31} & f_{32} & f_{33} \end{bmatrix} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = 0$$

This is a linear equation in the unknown elements of F. Given eight or more perfect correspondences (image points in *general* position, no noise), F can be determined uniquely up to scale by solving the simultaneous linear equations. In practice, there may be more than eight correspondences and the image measurements will be noisy. The system of equations can then be solved by least squares, or using a robust regression scheme to reject outliers.

The linear technique does not enforce the constraint that det F = 0. If the eight image points are noisy, then the linear estimate of F will *not* necessarily have zero determinant and the epipolar lines will not meet at a point. Nonlinear techniques exist to estimate F from 7 point correspondences, enforcing the rank 2 constraint.

Assessor's comment:

This question tested the candidates understanding of stereo vision. Very good answers to most components except the equation of an epipolar line.