4 F12 (2004) Solutions (draft)

(a) (i) seperation of $G_\sigma(x,y) = G_\sigma(x)\,G_\sigma(y)$. [10]

(ii) The intensity of a smoothed pixel is computed by discrete convolution:

$$S(x,y) = \sum_{i=-n}^{n} \sum_{j=-n}^{n} G_\sigma(i,j) I(x-i, y-j)$$

The 2D convolution can be decomposed into two 1D convolutions as follows:

$$G_\sigma(x,y) * I(x,y) = \sum_{i=-n}^{n} \sum_{j=-n}^{n} g_\sigma(i)\, g_\sigma(j)\, I(x-i, y-j) = g_\sigma(x) * [g_\sigma(y) * I(x,y)]$$

where $g_\sigma(x)$ is a 1D discrete approximation to the Gaussian kernel:

$$g_\sigma(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\sigma^2}\right)$$ [20%]

(b) By examining first order Taylor expansions, we find that

$$\left.\frac{d^2 I}{dx^2}\right|_{(x,y)} \approx I_{(x-1,y)} - 2I_{(x,y)} + I_{(x+1,y)}$$

and

$$\left.\frac{d^2 I}{dy^2}\right|_{(x,y)} \approx I_{(x,y-1)} - 2I_{(x,y)} + I_{(x,y+1)}$$

It follows that the Laplacian can be estimated as follows:

$$\nabla^2 I\big|_{(x,y)} = \left.\frac{d^2 I}{dx^2}\right|_{(x,y)} + \left.\frac{d^2 I}{dy^2}\right|_{(x,y)} \approx I_{(x-1,y)} + I_{(x+1,y)} + I_{(x,y-1)} + I_{(x,y+1)} - 4I_{(x,y)}$$

This estimate can be computed by convolving with the kernel

0	1	0
1	-4	1
0	1	0

[30]

(c) [Book work] The principle advantage of the Marr-Hildreth operator is computational efficiency: edge detection requires only a single convolution and the detection of zero-crossings. Conversely, the Canny operator requires an additional, costly search for a local maximum normal to the gradient direction. The advantage of the Canny operator is enhanced robustness to noise. Any differential operator amplifies noise. The Canny operator computes only first derivatives and then searches for a local maximum (which is equivalent to a zero-crossing of the second derivative) normal to the gradient. The Marr-Hildreth operator computes second derivatives both along and normal to the edge. Computation of the second derivative along the edge emphasizes noise in that direction while serving no purpose in edge detection. [30]

2. **Perspective and weak perspective projection**

(a) [Book work] The mapping from camera-centered coordinates (X_c, Y_c, Z_c) to pixel coordinates (u, v) involves a perspective projection onto the image plane (x, y) followed by an anisotropic scaling and translation in the image plane to account for the dimensions and positioning of the CCD array.

The perspective projection is a non-linear operation in Cartesian coordinates:

$$x = \frac{fX_c}{Z_c}\ ,\quad y = \frac{fY_c}{Z_c}$$

where f is the focal length of the camera. This can be rewritten as a linear operation in homogeneous coordinates:

$$\begin{bmatrix} sx \\ sy \\ s \end{bmatrix} = \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} X_c \\ Y_c \\ Z_c \\ 1 \end{bmatrix}$$

The mapping from image plane coordinates (x, y) to pixel coordinates (u, v) is given by:

$$u = u_0 + k_u x\ ,\quad v = v_0 + k_v y$$

where the optical axis intersects the CCD array at the pixel with coordinates $(u_0.v_0)$ and there are k_u pixels per unit length in the u direction and k_v in the v direction. In homogeneous coordinates, this becomes

$$\begin{bmatrix} su \\ sv \\ s \end{bmatrix} = \begin{bmatrix} k_u & 0 & u_0 \\ 0 & k_v & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} sx \\ sy \\ s \end{bmatrix}$$

Concatenating the two transformations, we obtain

$$\begin{bmatrix} su \\ sv \\ s \end{bmatrix} = \begin{bmatrix} k_u f & 0 & u_0 & 0 \\ 0 & k_v f & v_0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} X_c \\ Y_c \\ Z_c \\ 1 \end{bmatrix}$$

(2)

(b) (i) Under weak perspective projection, we assume that all points lie at approximately the same depth Z_A from the camera. This allows the projection to be re-written as follows:

$$\begin{bmatrix} su_A \\ sv_A \\ s \end{bmatrix} = \begin{bmatrix} k_u f & 0 & 0 & u_0 Z_A \\ 0 & k_v f & 0 & v_0 Z_A \\ 0 & 0 & 0 & Z_A \end{bmatrix} \begin{bmatrix} X_c \\ Y_c \\ Z_c \\ 1 \end{bmatrix}$$

✓

(ii) Under full perspective we have

$$u = \frac{k_u f X_c + u_0 Z_c}{Z_c}$$

Under weak perspective we have

$$\begin{aligned} u_A &= \frac{k_u f X_c + u_0 Z_A}{Z_A} = \left(\frac{k_u f X_c + u_0 Z_A}{Z_c}\right)\left(\frac{Z_c}{Z_A}\right) \\ &= \left(\frac{k_u f X_c + u_0 Z_c + u_0(Z_A - Z_c)}{Z_c}\right)\left(\frac{Z_c}{Z_A}\right) = \left(u + \frac{u_0 \Delta Z}{Z_c}\right)\left(\frac{Z_c}{Z_A}\right) \end{aligned}$$

where $\Delta Z \equiv Z_A - Z_c$. So

$$\begin{aligned} u - u_A &= u - \left(\frac{uZ_c + u_0 \Delta Z}{Z_c}\right)\left(\frac{Z_c}{Z_A}\right) = \left(\frac{uZ_A}{Z_c} - \frac{uZ_c + u_0 \Delta Z}{Z_c}\right)\left(\frac{Z_c}{Z_A}\right) \\ &= \left(\frac{u(Z_A - Z_c) - u_0 \Delta Z}{Z_c}\right)\left(\frac{Z_c}{Z_A}\right) = (u - u_0)\frac{\Delta Z}{Z_A} \end{aligned}$$

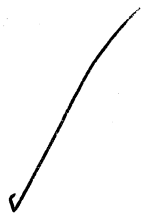
Similarly for $(v - v_A)$, we find that

$$v - v_A = (v - v_0)\frac{\Delta Z}{Z_A}$$

So the weak perspective approximation is perfect at the centre of the image, but gets progressively worse away from the centre.

(iii) [Book work] Weak perspective is a good approximation when the depth range of objects in the scene is small compared with the viewing distance. A good rule of thumb is that the viewing distance should be at least ten times the depth range.

The main advantage of the weak perspective model is that it is easier to calibrate than the full perspective model. The calibration requires fewer points with known world position, and, since the model is linear, the calibration process is also better conditioned (less sensitive to noise) than the nonlinear full perspective calibration.



3. **Planar projective transformations**

(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 $\tilde{\mathbf{w}}' = \mathrm{P}'\mathrm{P}^{-1}\tilde{\mathbf{w}} = \mathrm{T}\tilde{\mathbf{w}}$, where $\mathrm{T} \equiv \mathrm{P}'\mathrm{P}^{-1}$ 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. [20%]

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

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

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

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

It follows that

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

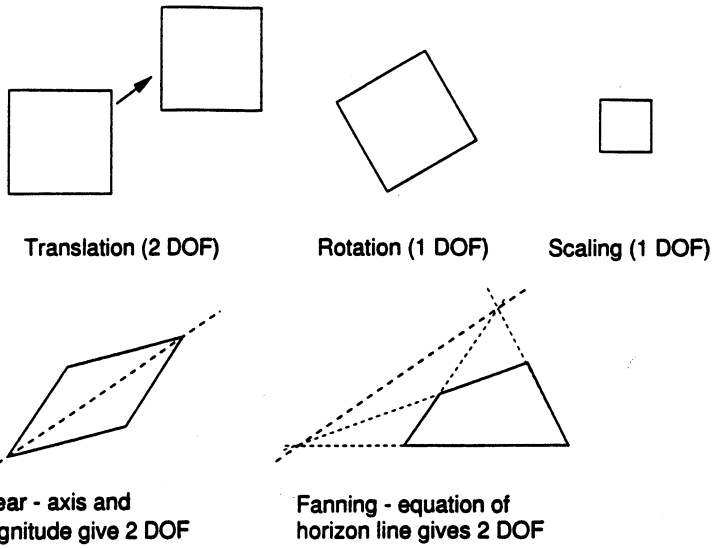
After rotating by R about the optical centre, the same world point $\mathbf{X}$ projects to a different image point $\tilde{\mathbf{w}}'$ as follows:

$$\tilde{\mathbf{w}}' = \mathrm{K}\left[\,\mathrm{R}\,|\,\mathbf{O}\,\right]\begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} = \mathrm{KR}\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \mathrm{KR}\mathbf{X} = \mathrm{KRK}^{-1}\tilde{\mathbf{w}} = \mathrm{T}\tilde{\mathbf{w}}$$

where $\mathrm{T} \equiv \mathrm{KRK}^{-1}$. Hence the relationship between points in the original image and corresponding points in the second image is a 2D projective transformation. [20%]

(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.



[20%]

(c) The equation of the line in the first image is $\mathbf{l}^T \tilde{\mathbf{w}} = 0$, where $\mathbf{l} = [\ l_1 \ \ l_2 \ \ l_3\ ]^T$. Since $\tilde{\mathbf{w}} = \mathrm{T}^{-1}\,\tilde{\mathbf{w}}'$, it follows that the equation of the line in the second image is $\mathbf{l}^T\,\mathrm{T}^{-1}\,\tilde{\mathbf{w}}' = 0$, or simply $\mathbf{l}' = \mathrm{T}^{-T}\,\mathbf{l}$. [20%]

(d) The equation of the conic in the first image is $\tilde{\mathbf{w}}^T\,\mathrm{C}\,\tilde{\mathbf{w}} = 0$, where

$$\mathrm{C} \equiv \begin{bmatrix} a & b/2 & d/2 \\ b/2 & c & e/2 \\ d/2 & e/2 & f \end{bmatrix}$$

Using again the relationship $\tilde{\mathbf{w}} = \mathrm{T}^{-1}\,\tilde{\mathbf{w}}'$, we find the equation of the corresponding conic in the second image as follows:

$$(\mathrm{T}^{-1}\,\tilde{\mathbf{w}}')^T\,\mathrm{C}\,\mathrm{T}^{-1}\,\tilde{\mathbf{w}}' = 0 \quad \Leftrightarrow \quad \tilde{\mathbf{w}}'^T\,\mathrm{T}^{-T}\,\mathrm{C}\,\mathrm{T}^{-1}\,\tilde{\mathbf{w}}' = 0$$

Alternatively, the conic in the second image can be expressed simply as $\mathrm{C}' = \mathrm{T}^{-T}\,\mathrm{C}\,\mathrm{T}^{-1}$. [20%]

4. **Stereo vision**

(a) The stereo camera geometry constrains each point feature identified in one image to lie on a corresponding *epipolar line* in the other image. If the cameras are calibrated, then the equation of the epipolar line can be derived from the essential matrix. For uncalibrated cameras, it is possible to estimate the fundamental matrix from point correspondences and derive epipolar lines from the fundamental matrix. Epipolar lines meet at the *epipole*: this is the image of one camera's optical centre in the other camera's image plane. There are two epipoles, one for each image.

[3₀]

(b) The essential matrix E describes the epipolar geometry of a stereo rig in terms of rays $\mathbf{p} = [\ x\ \ y\ \ f\ ]^T$, where (x, y) are the metric image plane coordinates of an observed point and f is the camera's focal length.

To derive the essential matrix in terms of R and $\mathbf{T}$, we start with the equation relating the two coordinate systems:

$$\begin{aligned}
\mathbf{X'_C} &= \mathrm{R}\mathbf{X_C} + \mathbf{T} \Rightarrow \mathbf{T} \times \mathbf{X'_C} = \mathbf{T} \times \mathrm{R}\mathbf{X_C} \\
\Rightarrow \mathbf{X'_C}.(\mathbf{T} \times \mathrm{R}\mathbf{X_C}) &= 0 \Leftrightarrow \mathbf{X'_C}.(\mathrm{T_\times}\mathrm{R}\mathbf{X_C}) = 0\ , \quad \text{where } \mathrm{T_\times} = \begin{bmatrix} 0 & -T_z & T_y \\ T_z & 0 & -T_x \\ -T_y & T_x & 0 \end{bmatrix} \\
\Leftrightarrow \mathbf{p'}^T[\mathrm{T_\times}\mathrm{R}]\mathbf{p} &= 0\ , \quad \text{since rays and camera-centered positions are parallel.}
\end{aligned}$$

The essential matrix is therefore given by $\mathrm{E} = \mathrm{T_\times R}$.

Epipolar geometry can be expressed in pixel coordinates and the epipolar constraint leads to the fundamental matrix. The essential and fundamental matrices are related by the internal calibration matrices K and of the left and right cameras, where

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

f is the focal length, k_u and k_v the pixel scale factors and (u_0, v_0) the point where the optical axis intersects the image plane. F= $'^{-T}\mathrm{T_\times}\mathrm{R}^{-1}$

(c) 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.

(d) The epipoles lie in the null spaces of F and F^T. So, for the left epipole we have:

$$\mathrm{F}\tilde{\mathbf{w}}_\mathbf{e} = \mathbf{0}$$

If F were invertible, we would be able to write