2008 IIB 4F12 COMPUTER VISION AND ROBOTICS PROF R CIPOLLA

Module 4F12: Computer Vision and Robotics

Solutions to 2008 Tripos Paper

1. (a) Consider smoothing an image, first with a Gaussian of standard deviation σ_1, then with a Gaussian of standard deviation σ_2:

$$s(x) = g_{\sigma 2}(x) * (g_{\sigma 1}(x) * I(x))$$

Since convolution is associative, we can write this as the convolution of the image with the kernel $g_{\sigma 2}(x) * g_{\sigma 1}(x)$:

$$s(x) = (g_{\sigma 2}(x) * g_{\sigma 1}(x)) * I(x)$$

The easiest way to evaluate the convolution of two Gaussians is to find their Fourier transforms and then multiply them in the frequency domain. If $g_\sigma(x) \leftrightarrow G_\sigma(\omega)$, then:

$$\begin{aligned}
G_\sigma(\omega) &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{x^2}{2\sigma^2}\right) e^{-j\omega x} dx \\
&= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left[-\left(\frac{x^2}{2\sigma^2} + j\omega x\right)\right] dx \\
&= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2\sigma^2}\left(x^2 + 2j\omega\sigma^2 x\right)\right] dx \\
&= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2\sigma^2}\left((x + j\omega\sigma^2)^2 - j^2\omega^2\sigma^4\right)\right] dx \\
&= \exp\left(-\frac{\omega^2\sigma^2}{2}\right) \times \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{(x + j\omega\sigma^2)^2}{2\sigma^2}\right) dx \\
&= \exp\left(-\frac{\omega^2\sigma^2}{2}\right) \text{ (since the integral is a standard Gaussian)}
\end{aligned}$$

Hence

$$g_{\sigma 2}(x) * g_{\sigma 1}(x) \leftrightarrow G_{\sigma 2}(\omega) \times G_{\sigma 1}(\omega) = \exp\left(-\frac{\omega^2\sigma_2^2}{2}\right) \times \exp\left(-\frac{\omega^2\sigma_1^2}{2}\right)$$

or

$$g_{\sigma 2}(x) * g_{\sigma 1}(x) \leftrightarrow G_{\sigma 2}(\omega) \times G_{\sigma 1}(\omega) = \exp\left(-\frac{\omega^2(\sigma_2^2+\sigma_1^2)}{2}\right)$$

The expression on the right is the Fourier transform of a Gaussian with standard deviation $\sqrt{\sigma_2^2+\sigma_1^2}$. Hence, the convolution of two Gaussians with variances σ_1^2 and σ_2^2 is a Gaussian with variance $\sigma_1^2+\sigma_2^2$. It follows that consecutive smoothing with a series of 1D Gaussians, each with a particular standard deviation σ_i^2, is equivalent to a single convolution with a Gaussian of variance $\sum_i \sigma_i^2$. [6]

(b) In practice, only a discrete set of scales can be considered, giving rise to an *image pyramid*. For a given image size, we an *octave* of scales is examined, corresponding to Gaussians with standard deviations from σ to 2σ. The image is then subsampled by a factor of 2 and the process is repeated for the next octave.

To improve efficiency, the above can be performed in an incremental manner. Within the octave σ to 2σ, for i-th interval (our of s) of the pyramid, we want $\sigma_i = 2^{i/s}\sigma$. To achieve this incrementally:

$$G(\sigma_{i+1}) = G(\sigma_i) * G(\hat{\sigma})$$

Then from (a):

$$\hat{\sigma} = \sqrt{\sigma_{i+1}^2 - \sigma_i^2}$$

and since

$$\sigma_{i+1} = 2^{(i+1)/s}$$

we have:

$$\hat{\sigma} = \sigma_i\sqrt{2^{2/s}-1}$$

The separability property of the Gaussian function should also be used to reduce the computational cost.

[8]

2. (a) The mapping from camera-centred 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:

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

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:

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

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

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

(b) 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 rewritten as follows:

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

(c) Weak perspective is a good approximation when the depth range of objects in the scene is small compared to the viewing distance. A good rule of thumb is that the viewing distance should be at least ten times the depth range.

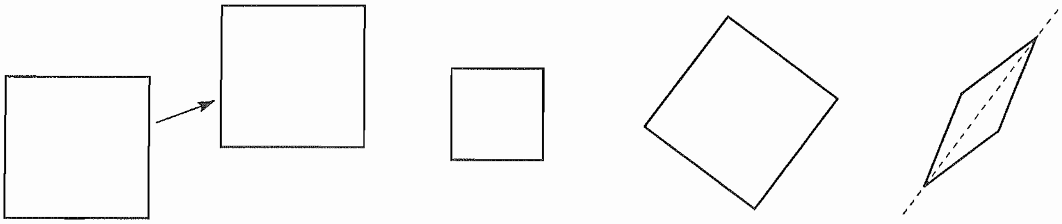
The main advantage of the weak perspective model is that it is easier to calibrate than the full perspective model. The calibration requires fewer points with known world position, and, since the model is linear, the calibration process is also better conditioned (less sensitive to noise) than the nonlinear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

3. (a) The mapping from world plane to image plane is an affine transformation under the weak perspective model, which is a good approximation when the depth range of objects in the scene is small compared to the viewing distance.

Projective transformation is linear in homogeneous coordinates:

$$\mathrm{w}' = \begin{bmatrix} t_{11} & t_{12} & t_{13} \\ t_{21} & t_{22} & t_{23} \\ t_{31} & t_{32} & t_{33} \end{bmatrix} \mathrm{w}'$$

Affine transformation is its special case $t_{31} = t_{32} = 0$ and thus has 6 DOF as the transformation matrix is defined up to a scale factor.

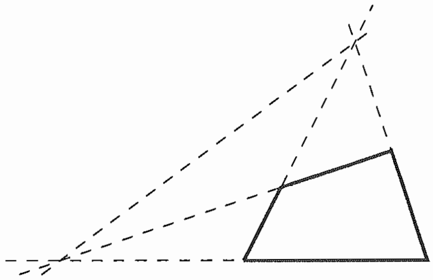


Translation (2 DOF) Scale (1 DOF) Rotation (1 DOF) Shear (2 DOF)

[4]

(b) The additional 2 DOF define fanning by specifying the horizon line:

[4]



Fanning (2 DOF)

4. (a) The task of matching features between left and right images can be simplified using several constraints:

Epipolar: 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 cameras optical centre in the other cameras image plane. There are two epipoles, one for each image.

Uniqueness: For scenes containing only opaque objects, each point in the left image has at most one match in the right image.

Ordering: Corresponding points lying on the surface of an opaque object will be ordered identically in left and right images. The ordering constraint will not necessarily hold if the points do not lie on the surface of the same opaque object.

Figural continuity: When distinguished points lie on image contours, we can sometimes use figural continuity as a matching constraint.

Disparity gradient: If surfaces are smooth, then point disparities (differences in location between left and right images) must be locally smooth. So a further constraint comes from imposing a limit on the allowable spatial derivatives of disparity.

[6]

(b) See (a) under *Epipolar*. [4]

(c) The weak perspective camera model is $\tilde{\mathrm{w}} = \mathrm{P}_{wp}\tilde{\mathrm{X}}$, where

$$\mathrm{P}_{wp} = \mathrm{P}_c\mathrm{P}_{pll}\mathrm{P}_r = \begin{bmatrix} k_u & 0 & u_0 \\ 0 & k_v & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 0 & Z_c^{av} \end{bmatrix} \begin{bmatrix} & \mathrm{R} & & \mathrm{T} \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} fk_ur_{11} & fk_ur_{12} & fk_ur_{13} & fk_uT_x + u_0Z_c^{av} \\ fk_vr_{21} & fk_vr_{22} & fk_vr_{23} & fk_vT_y + v_0Z_c^{av} \\ 0 & 0 & 0 & Z_c^{av} \end{bmatrix}$$

If we assume, without loss of generality, that the left camera is aligned with the world coordinate system (so that $\mathrm{R} = \mathrm{I}$), then the camera matrix reduces to

$$\begin{bmatrix} fk_ur_{11} & 0 & 0 & fk_uT_x + u_0Z_c^{av} \\ 0 & fk_vr_{22} & 0 & fk_vT_y + v_0Z_c^{av} \\ 0 & 0 & 0 & Z_c^{av} \end{bmatrix}$$

Discarding the nonlinear constraints, we obtain affine models for the left and right cameras:

Left:

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} p_{11} & 0 & 0 & p_{14} \\ 0 & p_{22} & 0 & p_{24} \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}$$

Right:

$$\begin{bmatrix} u' \\ v' \end{bmatrix} = \begin{bmatrix} p'_{11} & p'_{12} & p'_{13} & p_{14} \\ p'_{21} & p'_{22} & p'_{23} & p_{24} \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}$$

Eliminating X and Y from the above equations gives

$$u' = p'_{11}\frac{u - p_{14}}{p_{11}} + p'_{12}\frac{v - p_{24}}{p_{22}} + p'_{13}Z + p'_{14}$$
$$v' = p'_{21}\frac{u - p_{14}}{p_{11}} + p'_{22}\frac{v - p_{24}}{p_{22}} + p'_{23}Z + p'_{24}$$

Eliminating Z we obtain

$$u' = p'_{11}\frac{u - p_{14}}{p_{11}} + p'_{12}\frac{v - p_{24}}{p_{22}} + p'_{14} + \frac{p'_{13}}{p'_{23}}\left(v' - p'_{21}\frac{u - p_{14}}{p_{11}} - p'_{22}\frac{v - p_{24}}{p_{22}} - p'_{24}\right)$$

[10]

or, alternatively

$$au' + bv' + cu + dv + 1 = 0$$

We can rewrite this in matrix form:

$$\begin{bmatrix} u' & v' & 1 \end{bmatrix} \mathrm{F}_A \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = 0$$

where F_A is the affine fundamental matrix:

$$\mathrm{F}_A = \begin{bmatrix} 0 & 0 & a \\ 0 & 0 & b \\ c & d & 1 \end{bmatrix} = 0$$

By inspection, F_A has zero determinant and therefore maximum rank 2.

5. (a) Given 8 or more perfect correspondences (image points in general position, noiseless), F can be determined uniquely up to scale. In practice, we may have more than 8 correspondences and the image measurements will be noisy. The system can then be solved by least squares or using a robust regression scheme to reject outliers. The linear constraint does not enforce that $\det \mathrm{F} = 0$ and the epipolar lines do not meet at a point. Nonlinear techniques exist to estimate F from 7 point correspondences, enforcing the rank 2 constraint.

Given F, we can establish correspondences with relative ease. If we know the intrinsic camera parameters K, we can also find the essential matrix, decompose E into T and R, and recover metric structure by triangulation. Without K we can only recover structure up to a 3D projective transformation, which can later be disambiguated using further constraints.

(b) [This is bookwork] One answer can include interest point detection and their robust matching as in 3(d). Alternatively, a boosted classifier cascade or a support vector machine can be used.