Numerisk analys och datalogi, KTH
Mårten Björkman

EXAM

Bildbehandling och datorseende 2D1421
Wednesday, March 9th 2005, 14.00-19.00

Allowed material: Calculator, mathematics handbook (e.g. Beta) and a hand-written (not copied) sheet of paper in A4 format with your own personal notes. These notes have to be handed in together with your answers and will be returned after answers have been corrected.
Language: Answers can be given in either English or Swedish.
General: The examination consists of **six** exercises that can give at most 50 credits. To pass the examination you need about half of all credits. The bonus credits (at most 5) will be added to the total sum of your credits, given that you passed the laboratory exercises on time during this year. The results will be announced within three weeks.
Course evaluation: We would appreciate if you fill in the evaluation form available on the website.

The grades were set as follows: 3 (credits ≥ 25), 4 (credits ≥ 32) and 5 (credits ≥ 41).

Exercise 1 (5*2=10 credits)

Answer *five* out of the following *seven* short questions. If you respond to more than *five* questions, only the first *five* will be corrected and counted.

(a) In what sense is vision an “active process”?

The vision system is active in that it can control its environment, either by moving itself or adaptively adjust its parameters. It is a processes since new images are coming all the time and the system peforms operations in time.

(b) Mention at least two differences between “cones” and “rods”.

Cones are more sensitive to colours, while rods are only sensitive to luminance. Cones are located in the center of the retina, while rods cover the whole retina.

(c) What is a “neighbour” and what is a “connected component”? Show with an illustration.

In vision the concept of neighbours determines whether two nearby pixels are to be considered as connected. A connected component it a group of pixels, such that each pair of two pixels can be connected by a path.

(d) What is an “ideal” low-pass filter and why is such a filter not suitable for computer vision?

An ideal low-pass filter is a filter that removes all frequencies above a particular cut-off frequency. Unfortunately, it results in ringing effects in the spatial domain.

(e) What is the difference between "optical flow" and "motion field"?

Optical flow the movement of image data, that does not necessarily correspond to the projected motion of objects in the scene, such as in the case for texture-less objects. This projected motion is called the motion field.

(f) What are "epipolar lines" and why are these important in stereo vision?

Given a point in one particular camera, the corresponding point in the other can be found along a one-dimensional line, the epipolar line. Thus searches for matching points between cameras can be done in one-dimension, instead of two.

(g) What is "entropy" and why is it of interest in image compression?

Entropy is a statistical measure of the average number of bits required to code a symbol. It can be used to determine a theoretical limit on the compression rate.

Exercise 2 (5+2=7 credits)

We have a set of five 2D points; $\mathbf{p_1} = (-2,-1)$, $\mathbf{p_2} = (-1,-2)$, $\mathbf{p_3} = (0,0)$, $\mathbf{p_4} = (1,1)$ and $\mathbf{p_5} = (2,1)$.

(a) Assume the points originate from a two-dimensional distribution. Based on the few points given, characterize this distribution by an ellipse. Compute and draw this ellipse.

The center of the ellipse will be in the center of gravity; $\overline{\mathbf{p}} = (0,-0.2)$. *We then compute the covariance matrix using the points.*

$$\mathbf{C_{pp}} = \frac{1}{5}\begin{pmatrix} 10.0 & 7.0 \\ 7.0 & 6.8 \end{pmatrix}$$

The major axis will then be given by the eigenvector corresponding to the largest eigenvalue; $\mu_1 = (0.78, 0.62)$. *The lengths of the axis will then be the square root of the eigenvalues, i.e.* $\sqrt{\lambda_1} = 1.77$ *and* $\sqrt{\lambda_2} = 0.49$. *Alternatively, the ellipse can be expressed in terms of* (x,y) *using* $\mathbf{C_{pp}^{-1}}$, *without the eigenvalues.*

(b) Alternatively, assume that the points come from a one-dimensional distribution, i.e. they are placed along a line. Find this line and compute the mean square distance of the points to the line.

We solve this exactly like the previous exercise. The direction of the line is determines by the major axis of the ellipse. The mean square error equals the least eigenvalue. It is just a different way of interpreting the results. The computations are exactly the same.

Exercise 3 (2+2+3=7 credits)

Common for many operations in computer vision is the reduction of data. This is true for both object recognition and image compression.

(a) Why do we like (or need) to reduce data in object recognition and image compression respectively? What is the difference in the data we ignore?

In object recognition we like to remove image data that depends on e.g. orientation, position, illumination and deformation, in order for comparisons to models to be invariant to these changes. In compression we like to remove data that can not be observed by humans in order to make the representation as small as possible.

(b) What kind of redundancies can be exploited in image compression? Mention at least two kinds of redundancies and explain them.

Coding redundancy: Some symbols are more common than others and should thus be coded with a fewer number of bits.
Spatial redundancy: locally in the images many values are the same, which means that pixels should not necessarily be coded as if they were independent.

(c) Matching of features for recognition has to be made invariant to, among other things, 1) scale, 2) illumination and 3) rotation. Describe how this can be done in practice. Mention at least one typical example for each kind.

One could e.g. 1) use a feature detector that simultaneously estimates the scale of detected features, 2) use a representation based on gradient information, where the scale of gradient values has been normalized, and 3) reorient the representation based on the dominating gradient direction.

Exercise 4 (2+2+4=8 credits)

Assume we have a camera defined by a perspective projection, $\mathbf{x} = \mathbf{MX}$, with the projection matrix

$$\mathbf{M} = \begin{pmatrix} m_{11} & m_{12} & m_{13} & m_{14} \\ m_{21} & m_{22} & m_{23} & m_{24} \\ m_{31} & m_{32} & m_{33} & m_{34} \end{pmatrix}$$

and the homogeneous world and image coordinates given by $\mathbf{X} = (X,Y,Z,1)^T$ and $\mathbf{x} = (x,y,1)^T$.

(a) The matrix $\mathbf{M}$ can be expressed in terms of intrinsic and extrinsic camera parameters. Give at least two parameters of each kind.

Intrinsic: projection centre, aspect ratio, skew and focal length.
Extrinsic: relative rotation and translation.

(b) How would $\mathbf{M}$ look like if it were an "affine" projection matrix? Under what assumptions is an affine projection feasible?

The last row of $\mathbf{M}$ *would be* $(0,0,0,m_{34})$*, which means that parallel lines are projected to parallel lines in the image. The scale due to differences in depth does not matter. This is a feasible approximation of a perspective transformation, if observed objects are placed far from the observer in relation to variations in depth, and the field of view is small.*

(c) Assume we have a plane in 3D space defined by the equation

$$2X - Z + 2 = 0.$$

Introduce two coordinate axes in this plane, so that a point on the plane is given by the coordinates $\mathbf{x}' = (x',y',1)^T$. Show that the projection from the 3D plane to the image plane can be expressed

as $\mathbf{x} = \mathbf{A}\mathbf{x}'$, where A is a 3×3 projective transformation. What is A for your choice of coordinate axes? Hint: Use for example $(1,0,2,0)^T$ and $(0,1,0,0)^T$ as axes with $(0,0,2,1)^T$ as origin.

The two coordinate axes do not necessaily have to be orthogonal and normalized. Given the suggested coordinate axes a point on the plane can be given by

$$\begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix} = x' \begin{pmatrix} 1 \\ 0 \\ 2 \\ 0 \end{pmatrix} + y' \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 2 \\ 1 \end{pmatrix}$$

If you combine this with $\mathbf{M}$ *we get*

$$\mathbf{A} = \begin{pmatrix} m_{11} & m_{12} & m_{13} & m_{14} \\ m_{21} & m_{22} & m_{23} & m_{24} \\ m_{31} & m_{32} & m_{33} & m_{34} \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 2 & 0 & 2 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} m_{11}+2m_{13} & m_{12} & 2m_{13}+m_{14} \\ m_{21}+2m_{23} & m_{22} & 2m_{23}+m_{24} \\ m_{31}+2m_{33} & m_{32} & 2m_{33}+m_{34} \end{pmatrix}$$

Exercise 5 (2+2+5=9 credits)

Assume we have a robot with cameras placed $h = 120$ cm above an ugly coloured plastic floor bought by your grandmother in the late 60s. The texture of the floor is orange with lots of large brown circles on.

(a) The cameras are tilted towards the floor such that a particular circle is projected in the center of the left camera. Measured in the image the projection is $w_c = 40$ pixels in width and $h_c = 15$ pixels in height. How far away from the left camera is the circle on the floor?

The foreshortening angle θ *can either be determined from* $\cos(\theta) = h/Z$*, where* Z *is the distance to the circle on the floor, or from measurements* $\cos(\theta) = h_c/w_c$*. Thus the distance is given by*

$$Z = \frac{h \times w_c}{h_c} = \frac{120 \times 40}{15}\ cm = 320\ cm$$

(b) Assume that the two cameras are placed perfectly in parallel and have the same intrinsic parameters, with focal lengths equal to $f = 600$ pixels. How large is the baseline between the cameras, if the disparity in the center of the projected circle is $d = 24$ pixels? What is the diameter of the circle measured in centimetres?

The distance to a point in the scene from a stereo pair like this is $Z = bf/d$*, where* b *is the baseline. Thus the baseline can be given by*

$$b = \frac{Z \times d}{f} = \frac{320 \times 24}{600}\ cm = 12.8\ cm$$

The projection of a point is given by

$$\frac{x}{f} = \frac{X}{Z} \Rightarrow X = \frac{x \times Z}{f}$$

Thus the diameter can be written as

$$\delta X = \frac{\delta x \times Z}{f} = \frac{w_c \times Z}{f} = \frac{40 \times 320}{600}\ cm = 21.3\ cm$$

(c) Grandmother's walls are neatly coloured blue. To separate images into wall and floor pixels we apply pixel classification, using a single blue-yellow colour channel, $z \in [\,0, 255\,]$. We assume the classes of pixels to be described by normal distributions

$$p(z \mid C_k) = \frac{1}{\sqrt{2\pi}\sigma_k} e^{-(z-m_k)^2/(2\sigma_k^2)},$$

with standard deviations $\sigma_W = 80$ and $\sigma_F = 40$, and means $m_W = 60$ and $m_F = 180$. If the prior probabilities are $p(C_F) = 0.7$ and $p(C_W) = 0.3$, between which colour values will a pixel z be classified as a floor pixel?

The decision boundary is given by the z that satisfies $p(z \mid C_W)p(C_W) = p(z \mid C_F)p(C_F)$, i.e. it's equally possible that it's a wall and floor pixel. This means that

$$\frac{0.3}{80\sqrt{2\pi}}\, e^{-(z-60)^2/(2\times 80^2)} = \frac{0.7}{40\sqrt{2\pi}}\, e^{-(z-180)^2/(2\times 40^2)}$$

We simplify by removing some factor and taking the logarithm.

$$\ln(3/14) - (z-60)^2/(2\times 80^2) = -(z-180)^2/(2\times 40^2) \Rightarrow$$

$$3z^2 - 1320z + 129600 - 3600 + 12800\ln(3/14) = 0 \Rightarrow$$

$$z^2 - 440z + 42000 + \frac{12800}{3}\ln(3/14) = 0 \Rightarrow$$

$$z_{1,2} = 220 \pm 113.897... \Rightarrow z_1 = 106.103...$$

Given that $m_F > m_W$ and $z \in [\,0, 255\,]$, a pixel should be regarded as a floor pixel is $z > 106.1$, and a wall pixel otherwise.

Exercise 6 (3+4+2=9 credits)

The primary reason for expressing discrete filters in frequency space is to understand their behaviours, in particular in relation to their equivalents in continuous space.

(a) Show using the definitions of convolutions and Fourier transforms that the Fourier transform of a convolution of two kernels is the same as the product of the Fourier transforms of each kernel, i.e. $\mathcal{F}(h*g) = \mathcal{F}(h)\mathcal{F}(g)$.

Proof:

$$\mathcal{F}(h*g) = \int_{x\in\mathbb{R}^n} \Big(\int_{\eta\in\mathbb{R}^n} h(x-\eta)g(\eta)d\eta\Big)\, e^{-i\omega^T x} dx \qquad \{\text{rewrite}\}$$

$$= \int_{\eta\in\mathbb{R}^n} \Big(\int_{x\in\mathbb{R}^n} h(x-\eta)e^{-i\omega^T(x-\eta)}dx\Big)\, g(\eta)e^{-i\omega^T\eta}d\eta \quad \{\text{with } (x-\eta)=\zeta\}$$

$$= \int_{\eta\in\mathbb{R}^n} \Big(\int_{\zeta\in\mathbb{R}^n} h(\zeta)e^{-i\omega^T\zeta}d\zeta\Big)\, g(\eta)e^{-i\omega^T\eta}d\eta \quad \{\text{separate}\}$$

$$= \Big(\int_{z\eta\in\mathbb{R}^n} h(\zeta)e^{-i\omega^T\zeta}d\zeta\Big)\, \Big(\int_{\eta\in\mathbb{R}^n} g(\eta)e^{-i\omega^T\eta}d\eta\Big) = \mathcal{F}(h)\mathcal{F}(g)$$

(b) The Fourier transform of a continuous second order derivative is $\mathcal{F}(\delta_x^2) = -\omega^2$. Unfortunately, on discrete data we have to approximate these derivatives. Assume that we twice apply a first order differentiation kernel $h_x = \frac{1}{2}[-1,0,1]$. What is the frequency response of this approximation? What is the frequency response if we instead apply the second order differentiation $h_{xx} = [-1,2,-1]$?

The frequency characteristic of h_x is given by

$$\hat{h}_x(\omega) = (e^{i\omega} - e^{-i\omega})/2 = ((\cos\omega + i\sin\omega) - (\cos\omega - i\sin\omega))/2 = i\sin\omega$$

and if it's applied twice

$$\hat{h}_x^2(\omega) = -\sin^2\omega$$

The frequency characteristic of h_{xx} is given by

$$\hat{h}_{xx}(\omega) = (-e^{i\omega} + 2 - e^{-i\omega}) = 2 - (\cos\omega + i\sin\omega) - (\cos\omega - i\sin\omega) = 2(1-\cos\omega)$$

(c) Draw the frequency responses of the approximations in (b) as functions of ω. Which alternative is preferable, if we are looking for an approximation to a continuous second order derivative?

If we disregard the incorrect signs, the frequency characteristic $\hat{h}_{xx}(\omega)$ is closer to $F(\delta_x^2) = -\omega^2$ than $\hat{h}_x^2(\omega)$, at least for low frequencies. Thus the latter filter ought to be used.

Note: If you like to know whether you passed or failed this exam before the next exam, which is given on **March 18th, 08.00-13.00**, write your email address on top of the hand-ins.

Good luck!