School of Computer Science and Communication, KTH
Lecturer: Mårten Björkman

EXAM

Image Analysis and Computer Vision, DD2423

Tuesday, 14th of January 2014, 14.00–19.00

Allowed helping material: Calculator, the mathematics handbook Beta (or similar).

Language: The answers can be given either in English or Swedish.

General: The examination consists of Part A and Part B. For the passing grade E, you have to answer correctly at least 80% of Part A. If your score is less than 80%, the rest of the exam will not be corrected. Part B of the exam consists of **six** exercises that can give at most 50 points.

The results will be announced within three weeks.

Part A

Provide short answers to the questions! Each answer is worth maximum one point.

1. What factors determine the intensity of a pixel when measured by a camera sensor?

   Answer: exposure time, pixel size, aperture of lens, etc.

2. What kind of errors affect the quality of pixels in the sampling process?

   Answer: discretization (number of pixels), quantization (number of grey-levels), image noise (exposure time), etc.

3. Why is a definition of neighbourhood system necessary for connected components?

   Answer: Because the neighbourhood system defines connectivity.

4. Why is it enough to know the impulse response of a linear shift-invariant filter to know the effect it has on all signals (images)?

   Answer: Because all responses can be given by a shifted and weighted sum of impulse responses.

5. What is the benefit of Fourier transforms for image filtering?

   Answer: It may be easier to define and understand filters, and faster to compute.

6. What does it mean that a Fourier transform is conjugate symmetric?

   Answer: The Fourier transform is symmetric around zero frequency in magnitude, but anti-symmetric for the phase.

7. In what sense is an ideal low-pass filter not really ideal?

   Answer: It often generates ringing effects around edges that can be more annoying than the image noise one tries to remove.

8. Why is the notion of scale important in image analysis?

   Answer: Because things often appear in different scales due to the varying distances to objects in the scene.

9. What characterizes an image feature that is good for stereo matching?

   Answer: It is distinct and has structure in at least one dimension.

10. What is the similarity and difference between K-means and Mean-shift for image segmentation?

    Answer: They are both iterative methods often used for segmentation, but the optimize different things.

11. What do graph based segmentation methods normally try to optimize?

    Answer: They search for segmentations with maximum similarity within segments and minimum similarity between segments.

12. What invariances should an object recognition method typically be able to handle?

    Answer: They should be invariant to changes in view point, illumination, scale, noise levels, etc.

13. What components do most feature based recognition methods consist of?

    Answer: Feature based recognition methods consist of a feature detector for finding features, a descriptor to describe them and a metric for comparing them.

14. Derive the relation between the disparity and the depth of a 3D point for parallel cameras.

    A point in one image $P_1 = (X,Y,Z)^T$ is shifted along the x-axis in the other image, $P_2 = (X + T,Y,Z)$. The difference in projection is the disparity $d = f(X+T)/Z - fX/Z = fT/Z$, where f is the focal length and Z is the depth.

15. Why can it be hard to separate rotations from translations when computing the motion of a camera using corner features tracked over time?

    Answer: Because the optical flow may look very similar, especially if the scene is flat.

Part B

Exercise 1 (1+3+3=7 points) 4.5 hours

1. Assuming that you have two perspective cameras with different focal lengths over-looking a scene, for which camera are parallel 3D lines closest to being parallel also in the 2D projection?

   Answer: For the camera with the largest focal length the lines will look most parallel. The field of view is smaller, which makes the projection closer to an orthographic model for which parallel 3D lines are projected to parallel lines in the image.

2. Compute the projections of the 3D points with homogeneous coordinates

$$\mathbf{X}_1 = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 1 \end{pmatrix}, \ \mathbf{X}_2 = \begin{pmatrix} 2 \\ 1 \\ 1 \\ 1 \end{pmatrix} \text{ and } \ \mathbf{X}_3 = \begin{pmatrix} 2 \\ 3 \\ -1 \\ 1 \end{pmatrix}$$

   in the camera with projection matrix

$$\mathbf{P} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix}.$$

   What is the interpretation of the projection of $\mathbf{X}_3$?

   Answer: The projections will be $\mathbf{PX}_1 = (1,2,4)^T$, $\mathbf{PX}_2 = (2,1,2)^T$ *and* $\mathbf{PX}_3 = (2,3,0)^T$. *The last one can be interpreted as a point at infinity in the image plane, with direction given by* $(2,3)$.

3. Let $P_i = (X_i, Y_i, Z_i)^T$, $i = 1,...,N$, be a set of 3D points with image projections $p_i = (X_i/Z_i, Y_i/Z_i)^T$. Assuming a rigid camera motion, the transformed point coordinates are given by

$$P_i' = RP_i + T,$$

   where R is the rotation and T is the translation of the motion, with projections $p_i' = (X_i'/Z_i', Y_i'/Z_i')^T$. Show that in the case of pure rotation ($T = 0$), it is not possible to recover the depths Z_i given any number of matched pairs (p_i, p_i').

   Answer: The rotation in 3D space, $P_i' = RP_i$, *can be written in the projection as*

$$\omega_i' \begin{pmatrix} x_i' \\ y_i' \\ 1 \end{pmatrix} = R \begin{pmatrix} x_i \\ y_i \\ 1 \end{pmatrix} = \begin{pmatrix} r_1^T \\ r_2^T \\ r_3^T \end{pmatrix} \begin{pmatrix} x_i \\ y_i \\ 1 \end{pmatrix},$$

   which means that (x_i', y_i') *can be expressed as a function of* (x_i, y_i) *through*

$$x_i' = \frac{(x_i, y_i, 1)r_1}{(x_i, y_i, 1)r_3}, \ \textit{and} \ \ y_i' = \frac{(x_i, y_i, 1)r_2}{(x_i, y_i, 1)r_3}.$$

   Since these relations do not involve Z_i *and* Z_i', *two matched points could be placed on any depths as long as the projections are the same. Thus one cannot be recover the depths through matching of feature points.*

Exercise 2 (3+3+2=8 points) 3.5 hours

1. What properties do Gaussian filters possess that make them suitable for scale-space representation?

   Answer: Linear, shift-invariant, isotropic, scale covariance under scaling, semi-group structure, non-creation of structure with increasing scales.

2. Show through derivation that the Gaussian function

$$g(x,y;\,t) = \frac{1}{2\pi t} e^{-(x^2+y^2)/2t}$$

   satisfies the heat equation

$$L_t = \frac{1}{2}\nabla L, \;\; \text{where} \;\; \nabla L = L_{xx} + L_{yy}.$$

   Answer: First compute the derivatives with respect to x, y and t.

$$g_t = \frac{1}{2\pi}(-\frac{1}{t^2} + \frac{1}{t}\frac{x^2+y^2}{2t^2}) e^{-(x^2+y^2)/2t} = \frac{1}{2\pi t^2}(\frac{x^2+y^2}{2t} - 1) e^{-(x^2+y^2)/2t}$$

$$g_x = -\frac{1}{2\pi t}\frac{x}{t} e^{-(x^2+y^2)/2t}$$

$$g_{xx} = \frac{1}{2\pi t}(\frac{x^2}{t^2} - \frac{1}{t}) e^{-(x^2+y^2)/2t} = \frac{1}{2\pi t^2}(\frac{x^2}{t} - 1) e^{-(x^2+y^2)/2t}$$

$$g_{yy} = [\ \textit{the same way}\ ] = \frac{1}{2\pi t^2}(\frac{y^2}{t} - 1) e^{-(x^2+y^2)/2t}$$

   Thus

$$\frac{1}{2}(g_{xx} + g_{yy}) = \frac{1}{2\pi t^2}(\frac{x^2}{2t} - \frac{1}{2} + \frac{y^2}{2t} - \frac{1}{2}) e^{-(x^2+y^2)/2t} = \frac{1}{2\pi t^2}(\frac{x^2+y^2}{2t} - 1) e^{-(x^2+y^2)/2t} = g_t$$

3. From the above, show that the scale-space representation of an arbitrary function f, that is

$$L(x,y;\,t) = g(x,y;\,t) * f(x,y)$$

   also satisfies the heat equation.

   Answer: Since derivations and convolutions are linear

$$L_t = \frac{\partial (g*f)}{\partial t} = g_t * f,\; L_{xx} = g_{xx} * f \;\; \textit{and} \;\; L_{yy} = g_{yy} * f$$

   Thus

$$L_t = g_t * f = \frac{1}{2}(g_{xx} + g_{yy}) * f = \frac{1}{2}(g_{xx} * f + g_{yy} * f) = \frac{1}{2}(L_{xx} + L_{yy})$$

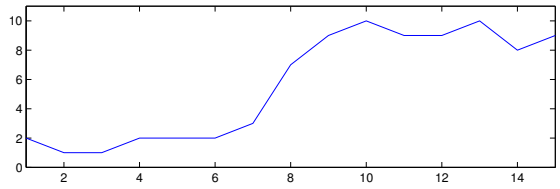
Exercise 3 (2+2+3+3=10 points) 5.5 hours

1. To the 1D image $f = [\,2,1,1,2,2,2,3,7,9,10,9,9,10,8,9\,]$ (see below) apply the two filter kernels

$$g_1 = [\,1,2,1\,] \text{ and } g_2 = [\,1,0,-1\,].$$

Answer:

$$f * g_1 = [\,5,5,7,8,9,15,26,35,38,37,37,37,35\,]$$
$$f * g_2 = [\,-1,1,1,0,1,5,6,3,0,-1,1,-1,-1\,]$$



2. The image obviously includes an edge at about $x = 8$. Propose a sharping method to enhance this edge and apply the method to the image.

Answer: The easiest approach is unsharp masking using $g = (1+k)f - k(f * g_1)/4$, where k is some constant and g_1 comes from above. For example, if $k = 1$ then

$$g = [\,3,3,9,8,7,9,30,35,42,35,35,43,29\,]/4.$$

3. Consider a 3×3 spatial filter mask

$$h = \frac{1}{8}\begin{pmatrix} 0 & 1 & 0 \\ 1 & 4 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

that computes the weighted average of the center point and its four closest neighbours. Find the corresponding frequency space representation $H(u,v)$ and show that the filter is a lowpass filter.

Answer: Assume (u,v) is the angular frequency, the Fourier transform is given by

$$H(u,v) = \int_x \int_y h(x,y) e^{-i(ux+vy)}\, dx\, dy = \frac{1}{8}(4 + e^{-iu} + e^{+iu} + e^{-iv} + e^{+iv}) =$$
$$= \frac{1}{4}(2 + \cos(u) + \cos(v)) \approx \frac{1}{4}(2 + 1 - \frac{1}{2}u^2 + 1 - \frac{1}{2}v^2) = 1 - \frac{1}{8}(u^2 + v^2).$$

Since it decays for increasing frequencies, it has to be a lowpass filter.

4. Propose a suitable discrete filter that approximates a Laplacian and show using Taylor Series expansion how it compares to a true Laplacian with transfer function $G(u,v) = -(2\pi)^2(u^2+v^2)$.

Answer: A suitable filter is (probably)

$$h = \begin{pmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

The Fourier transform is

$$H(u,v) = \int_x \int_y h(x,y) e^{-2\pi i(ux+vy)}\, dx\, dy = -4 + e^{-2\pi iu} + e^{+2\pi iu} + e^{-2\pi iv} + e^{+2\pi iv} =$$

$$= -4 + 2\cos(2\pi u) + 2\cos(2\pi v) = -4 + 2(1 - \frac{(2\pi u)^2}{2!}) + 2(1 - \frac{(2\pi v)^2}{2!}) + O(u^4, v^4) =$$

$$= -(2\pi)^2(u^2 + v^2) + O(u^4, v^4)$$

The suggested filter is equal to a true Laplacian up to the third power of the frequency.

Exercise 4 (3+1+2+3=9 points) 4.5 hours

1. Assume you have the image below. If you where to perform histogram equalization on this image, how would the resulting image look like? Explain the individual steps of the computation.

0	0	0	0	0	0	0	0
0	1	1	1	1	1	1	1
0	1	2	2	2	2	2	2
0	1	2	3	3	3	3	3
0	1	2	3	4	4	4	4
0	1	2	3	4	5	5	5
0	1	2	3	4	5	6	6
0	1	2	3	4	5	6	7

Answer: The histogram is $h(x) = [\ 15,13,11,9,7,5,3,1\ ]$*, which leads to a transformation* $y = T(x) = \lfloor 7 \cdot [\ 15,28,39,48,55,60,63,64\ ]/64 \rfloor = [\ 1,3,4,5,6,6,6,7\ ]$*. The final image would be*

1	1	1	1	1	1	1	1
1	3	3	3	3	3	3	3
1	3	4	4	4	4	4	4
1	3	4	5	5	5	5	5
1	3	4	5	6	6	6	6
1	3	4	5	6	6	6	6
1	3	4	5	6	6	6	6
1	3	4	5	6	6	6	7

with a corresponding histogram $h(y) = [\ 0,15,0,13,11,9,15,1\ ]$.

2. Why does discrete histogram equalization rarely lead to uniform histograms?

   Answer: Because we have a number of different grey-level values to begin with and two pixels with the same original value cannot be given different values afterwards.

3. Suggest two 3×3 differential operators, one for x-wise derivatives and one for y-wise derivatives. Why are these just approximations of derivatives?

   Answer: One suggestion is $f_x(x,y) = (f(x+1,y) - f(x-1,y))/2$ *and* $f_y(x,y) = (f(x,y+1) - f(x,y-1))/2$*. In theory the spacing inbetween compared pixels should be* $h \to 0$*, but since we are working with discrete pixels it can never be less than* $h = 1$.

4. Apply the two operators to the image above and compute a second moment matrix for the whole image, using a uniform window function. What can the second moment matrix be used for?

   Answer: With the above operators applied to the image (ignoring the boundaries), you will get

$$f_x = \frac{1}{2}\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 & 0 & 0 \\ 2 & 2 & 1 & 0 & 0 & 0 \\ 2 & 2 & 2 & 1 & 0 & 0 \\ 2 & 2 & 2 & 2 & 1 & 0 \\ 2 & 2 & 2 & 2 & 2 & 1 \end{bmatrix}, \quad f_y = \frac{1}{2}\begin{bmatrix} 1 & 2 & 2 & 2 & 2 & 2 \\ 0 & 1 & 2 & 2 & 2 & 2 \\ 0 & 0 & 1 & 2 & 2 & 2 \\ 0 & 0 & 0 & 1 & 2 & 2 \\ 0 & 0 & 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

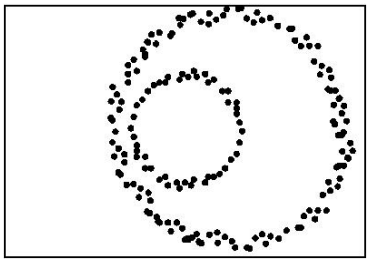
   This would result in a second moment matrix

$$S = \sum_{(x,y)} \begin{pmatrix} f_x^2 & f_x f_y \\ f_x f_y & f_y^2 \end{pmatrix} = \frac{1}{4}\begin{pmatrix} 66 & 6 \\ 6 & 66 \end{pmatrix}$$

   The second moment matrix captures the amount of structure the image has in two orthogonal directions and is usefully for corner detection, as well as for estimation of optical flow. If both eigenvalues are large (such as in this case) you have corner-like structures, which means that optical flow can be reliable computed.

Exercise 5 (2+3+3=8 points) 4.0 hours

Consider an image that consists of two sets of points. For the first set, points are spread around a circle of radius R centered at point C_1, which is near the center of the image. The other set of points is spread on a circle of radius $2R$, which is centered at point C_2 located inside the other circle. Assume that the points in each set are distributed densely enough so that the distances between points on the same circle are smaller than the distances between points on different circles.



1. Assume that we want to divide the points into two clusters corresponding to the circles. Describe what result can be expected when using the K-means clustering algorithm on the point positions.

   Answer: K-means should divide the points into two clusters, splitting the points along a straight line. The exact division depends on the starting condition, but in most cases this line would be vertical in the image.

2. Explain how the circles could be found using RANSAC. What would the necessary steps be?

   Answer: Select a minimum set (3) of points randomly and fit these to a circle model. Count how many other points are close enough (given some threshold) to the circle. Do this multiple times and keep the circle that most points agree with. Then you have found one circle. Repeat this process to find also the second circle.

3. Assume that you instead try to use a Hough transform to find the circles. How could the accumulator space be set up and what steps would you need to perform?

   Answer: Define a 3D acculator space using for example x-position, y-position and radius as coordinates. For each point in the image, place votes in this space. The votes are 2D surfaces with one coordinate (e.g. radius) given as a function of the others, e.g. $r = r(x,y) = \sqrt{(x-x_i)^2+(y-y_i)^2}$. *Find the two points in the accumulator space with argest number of votes.*

Exercise 6 (1+3+3+1=8 points) 4.0 hours

1. Why is the epipolar geometry constraint relevant to stereo matching?

   Answer: Stereo matching can be terribly noise sensitive and slow, if it's not constrained somehow. Using the epipolar geometry constraint you can constrain stereo matching to searches along (epipolar) lines.

2. Assume you have two cameras, c_1 and c_2, where c_2 is placed two units to the right of c_1 and one unit forward in the coordinate system of c_1. Also assume that c_2 is rotated 30° around the y-axis (see figure below). Compute the essential matrix that relates image points between the cameras.

   Answer: The essential matrix can be written as $E = RT_\times$, *where R is the relative rotation between the cameras and* $T_\times$ *is a skew-symmetric matrix with the relative translation* $t = (2,0,1)^T$ *as null space. Here these are*

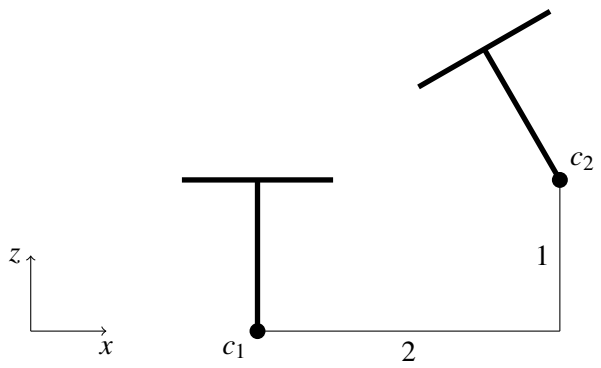
$$R = \frac{1}{2}\begin{pmatrix} \sqrt{3} & 0 & 1 \\ 0 & 2 & 0 \\ -1 & 0 & \sqrt{3} \end{pmatrix} \text{ and } T_\times = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & -2 \\ 0 & 2 & 0 \end{pmatrix} \Rightarrow E = \begin{pmatrix} 0 & -\frac{\sqrt{3}}{2}+1 & 0 \\ 1 & 0 & -2 \\ 0 & \sqrt{3}+\frac{1}{2} & 0 \end{pmatrix}$$

3. Where are the epipoles in the images of c_1 and c_2? How are these related to the epipolar lines?

   Answer: The epipoles are the left and right null spaces of E. One epipole is simply $e_1 = t = (2,0,1)^T$, *while the other is given by* $e_2 = (2\sqrt{3}+1,0,\sqrt{3}-2)^T \simeq (-(5\sqrt{3}+8),0,1)^T$. *All epipolar lines will pass through the epipoles that represent the projections of each camera center on the image of the other camera.*

4. Assume that you don't know the position and orientation of c_2 with respect to c_1. How many image point matches between the cameras would you theoretically need to find the essential matrix?

   Answer: Five points are needed, since there are six degrees of freedom (three for rotation and three for translation), but one degree cannot be determined (translation magnitude), since we are using homogeneous coordinates.


c_2
1
z
x
c_1
2