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

EXAM

Image Analysis and Computer Vision, DD2423

Friday, 16th of December 2011, 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 70% of Part A. If your score is less than 70%, 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 bonus credits from the labs will be added to Part A if you do not reach 70% - otherwise they will be added to Part B.

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 is the characteristic of a 'pinhole camera' and why is it important in computer vision?
2. If you are expressing perspective projections using homogeneous coordinates, why doesn't a scaling of the homogeneous coordinates matter?
3. If you look at an image of a city scene, how could you determine if an affine projection matrix would be reasonable to describe the image projection of the scene, instead of using a perspective projection matrix?
4. What is a 'neighborhood system' and a 'connected component', and how are the two concepts related?
5. Give an example of a non-linear filter that you have seen in the course. Why are linear filters often preferable from non-linear filters?
6. What does it mean that a 2D kernel is separable? Is a Fourier Transform separable?
7. What happens to the Fourier domain representation of an image, if the image is translated?
8. Mention a segmentation method that doesn't take spatial coherence into consideration. Why is it often a bad idea to ignore spatial coherence for segmentation?
9. What kind of image features are preferable for matching in stereo or motion? Why are they preferable?
10. Describe shortly how one can create a multi-scale (scale-space) representation of an image. Why are such representations useful in computer vision?
11. Given a set of edge points, a Hough Transform can be used to find straight lines. How is this done in practice? Shortly explain the steps involved.
12. There are many kinds of shape descriptions, and they differ in different ways. Give two examples of aspects for which shape descriptions may differ.
13. Give two reasons why robust object recognition is so hard in practice.
14. What is a 'vergence angle' and what is a 'gaze direction'?
15. What are the effects of morphological 'opening' and 'closing' operations?

Part B

Exercise 1 (1+2+3+2=8 points)

Grey-level transformations is a convenient tool to enhance details in images, without doing any spatial filtering, by simply changing the grey-level values.

1. Assume that you apply a grey-level transformation $s = T(r)$ to an image. For which parts of the image do the contrasts increase, and for which do they decrease?
2. What is the goal of 'histogram equalization' and why would you be interested in applying it to an image?
3. Assume you have an image with a histogram of grey-level values given by the distribution $p_R(r) = \frac{3}{5}(4r - 4r^2 + 1)$, $r \in [0,1]$. Determine a transformation $s = T(r)$, such that the histogram after the transformation becomes $p_S(s) = 1$, $s \in [0,1]$. For which values of r do you get a stretching of grey-level values, and for which do you get a compression?
4. How do you derive the transformation in the general case, when $p_S(s)$ is not necessarily a uniform transformation, such as in question 3? Give a short sketch of a solution, without doing any calculations.

Exercise 2 (2+3+2+2=9 points)

Most operations in image processing involve filtering in one way or the other. Depending on the image quality and your objective, it is essential to know what kind of filter to choose.

1. What are the properties of linear shift invariant filters? Mention at least two properties and explain what they mean.
2. Assume you have a 1D image given by

   $$F(x) = \{1,6,8,11,7,3,1\}.$$

   Using a convolution, apply to $F(x)$ each of the following three filters:

   $$G_1(x) = \{1,0,-1\},$$
   $$G_2(x) = \{1,2,1\} \text{ and}$$
   $$G_3(x) = \{1,-2,1\}.$$

   Answer with only five values per filter, disregarding the remaining undefined values.
3. Which one of the three filters above respectively corresponds to a smoothing operation, an approximate 1st order derivative and a 2nd order derivative?
4. For the filter you believe is a 2nd order derivative, compute its Fourier Transform. How well does this correspond to the frequency response of a real 2nd order derivative, that is

   $$G(\omega) = -\omega^2?$$

   Illustrate with a simple drawing, if it helps you.

Exercise 3 (1+1+1+3+2=8 points)

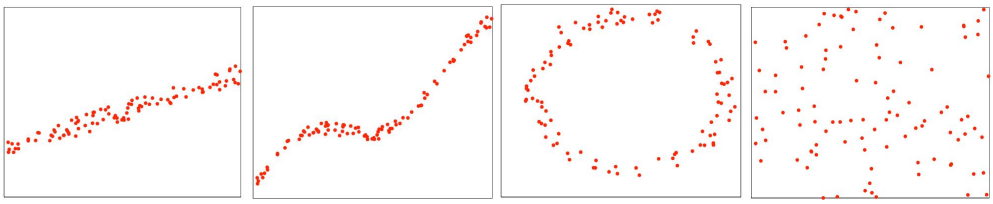
In order to limit the amount of data to process and concentrate on the most relevant parts of an image, most computer vision methods rely on the extractions of features, such as edges, corners and regions.

1. Why is edge detection important, if you want to tell something about the 3D world?
2. What is an image gradient magnitude and how can you compute it?
3. Why does edge detection normally include image smoothing as a first stage?
4. In Lab2 we found edges in images by asserting that two different conditions (on the image derivatives) have to be true for a pixel to be part of an edge. What are these two conditions and why do we need each of them?
5. Why would we often like to avoid using a 2nd order image derivative for edge detection? How does the Canny Edge detector solve the same problem, without using a 2nd order derivative?

Exercise 4 (2+2+2+4=10 points)

Grouping of data (pixels or feature data) is essential for both image segmentation and classification. Even if the purpose varies, many of the methods used share similarities.

1. What are the similarities and differences between segmentation methods based on either 'K-means clustering' or 'Mean-shift'? Respond with at least one similarity and one difference.
2. Which of the four point clouds below apparently come from one-dimensional distributions? For which of these would Principal Component Analysis (PCA) work for dimensionality reduction? Why doesn't it work for the remaining case(s)?



a) b) c) d)

3. You are given a set of seven points placed on a 2D plane

   $$\mathbf{p_1} = (8,6),\ \mathbf{p_2} = (1,5),\ \mathbf{p_3} = (5,5),\ \mathbf{p_4} = (0,4),\ \mathbf{p_5} = (7,3),\ \mathbf{p_6} = (5,3) \text{ and } \mathbf{p_7} = (2,2).$$

   Apply K-means clustering with $\mathbf{c_1} = (0,0)$ and $\mathbf{c_2} = (7,7)$ as the two initial cluster centers, and show how the division of points will be.

4. Assume that the seven points $\mathbf{p_i}$ mentioned in question 3 actually come from a one-dimensional distribution, i.e. without noise they would be placed along a line. Find the line that best fits the seven points, if you apply Principal Component Analysis (PCA).

Exercise 5 (2+2+4=8 points)

Instead of using pixel data directly, object recognition is usually performed using some low-dimensional representation of images. Part of the problem is solved by finding the best possible such representation.

1. If you have a set of images of some object classes, how can you tell whether a particular representation would be good for recognition of these classes?

2. Explain shortly how 'nearest neighbour' classification works. What makes nearest neighbour different from statistical classification methods?

3. We want to find a classifier to separate two classes, class A and class B. The classes are both assumed to have normal distributions, i.e. the probability density functions are of the form

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

where z is a one-dimensional measurement and k a class. The classes are further assumed to have means $m_A = -4$ and $m_B = 2$, and variances $\sigma_A^2 = 18$ and $\sigma_B^2 = 2$. Estimate the optimal classification boundaries and decision rules, if the prior probabilities are $p(k = A) = 2/3$ and $p(k = B) = 1/3$.

Exercise 6 (3+2+2=7 points)

Stereo is useful in order to determine the distance to objects seen in an image. However, even in cases with only one camera, distances can often be computed, if you add some assumptions.

1. Assume you are moving towards a person seen in an image. At time t_0 the height of the person in the image is $h_0 = 200$ pixels and at time t_1 it is $h_1 = 240$ pixels. If you have moved 1.00 meter between t_0 and t_1, what was the distance Z_0 to the person at time t_0? If we assume the person is 1.80 meters in height, what does the focal length f of the camera have to be (measured in pixels)?

2. Assume that you have another camera placed in parallel with and next to the first camera in the question 1 and the distance between the two cameras is $b = 0.10$ meters. If both these cameras have the same focal length f, what is the stereo disparity at the region corresponding to the person in the image at time t_0?

3. What is an epipolar plane and an epipolar line? How do these concepts facilitate stereo matching? Make a drawing, if it helps you explain.