102470 - Computer Vision Course
Institut für Informatik
Universität Bern

Exam

13/02/2018

- **You can use one A4 sized hand-written sheet of paper.**
- **No books, notes, computers, calculators and cellular phones are allowed.**
- **The number of points of the exam is 100. The questions are divided into 4 groups of 25 points each.**

Multiple-Choice Questions (10 Points)

Correct answer: +1 Point, Wrong answer: -1 Point, No answer: 0 Points.
Negative total points will be elevated to 0.

1. **True** **False** The values of the smoothing filter can be negative.
   Solution.
   No.

2. **True** **False** The values of the derivative filters sum to 1.
   Solution.
   Zero: No response in constant regions.

3. **True** **False** Gaussian filter is high-pass filter.
   Solution.
   False.

4. **True** **False** Median filter is more effective to remove ”pepper and salt” noise (impulsive noise) than Gaussian filter.
   Solution.
   True.

5. **True** **False** RANSAC does not need to fit the model to all samples to find the global optimum.

   Solution.
   TRUE. When optimum is found early, the rest of the trials don't matter.

6. **True** **False** RANSAC can only be used when the number of outliers is less than 50%.

   Solution.
   FALSE. There is no limit on the number of outliers.

7. **True** **False** The rank of the fundamental matrix is 3.

   Solution.
   False. 2

8. **True** **False** Structure from motion can recover the absolute scale of the scene.

   Solution.
   False.

9. **True** **False** The mean-shift algorithm is suitable for multiple segmentations.

   Solution.
   True.

10. **True** **False** The SIFT feature descriptor is robust to any shift over sub-patches in the image because it doesn't preserve the spacial information.

    Solution.
    False. Robust to small shift and still preserve the spacial information

Photometry, Features & Filters [21 points total]

1. Shrinking the lens aperture of a camera can make the captured image sharper. Why do we not make the aperture as small as possible? [**2 points**]

   Solution.

   Less light gets through. [**1 points**] Diffraction effects. [**1 points**]

2. Why are 2D separable kernels (e.g., the Gaussian filter) useful? [**2 points**]

   Solution.

   A 2D convolution can be reduced to two 1D convolutions. [**2 points**]

3. Let us consider an image x. Let also $p, q \in \mathbf{R}^2$ be two pixels (represented as two 2D vectors) in x. We define *self-similarity* as the property that

$$x[q] = x[\alpha(q-p)+p] \qquad \forall q : |q-p| \leq \rho \tag{1}$$

   where $\rho > 0$ is the radius of a ball around p. See Fig. 1. In other words, rays originating at p should have constant image intensity. Notice that, in particular, the self-similarity property is satisfied at corners and at edges. If x satisfies eq. (1) at a ball around p, what orientation will the gradient of x have at q? [**6 points**]

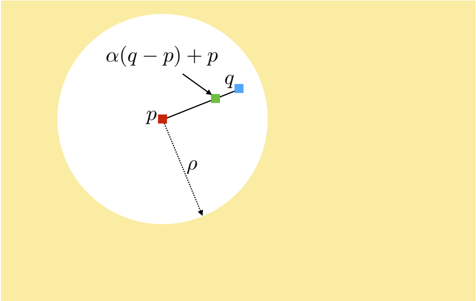

Figure 1: Gradients and self-similarity.

Solution The self-similarity property is equivalent to

$$x[q] - x[\alpha(q-p)+p] = 0 \qquad \forall q : |q-p| \leq \rho \tag{2}$$

[**3 points**] Then, we obtain

$$\lim_{\alpha \to 1} \frac{x[q] - x[\alpha(q-p)+p]}{1-\alpha} = \nabla x[q]^T (q-p) = 0. \tag{3}$$

[**2 points**] That is, the gradient in the patch should align with the vector originating at p. [**1 points**]

4. The relationship between a 3D point at world coordinates (X, Y, Z) and its corresponding 2D pixel at image coordinates (u, v) can be defined as a projective transformation, i.e. a 3×4 camera projection matrix P. How many degrees of freedom does the projection matrix P have in the most general case? Briefly justify your answer. [**4 points**]

   Solution. 11 degrees of freedom. [**2 points**]

   Any multiplication by a non-zero scalar results in an equivalent camera matrix. Since P maps vectors that are represented in homogeneous coordinates, so $P(aX)$ (a is a non-zero scalar) represents the same point as PX. [**2 points**]

5. Suppose that the normal map $\mathbf{n}$ of a depth map d is given. Write the equation that relates the depth map to the normal map. [**7 points**]

   Solution If we write the 3D point on the surface corresponding to a pixel (x, y) as [**1 points**]

$$P = \begin{bmatrix} x \\ y \\ d(x,y) \end{bmatrix} \tag{4}$$

then we can write the tangent vectors by taking the first order derivatives with respect to x and y [**2 points**]

$$T_x = \begin{bmatrix} 1 \\ 0 \\ d_x(x,y) \end{bmatrix} \qquad T_y = \begin{bmatrix} 0 \\ 1 \\ d_y(x,y) \end{bmatrix}. \tag{5}$$

Since the normal to the surface must be orthogonal to both tangent vectors, we have [**1 points**]

$$\mathbf{n} = \begin{bmatrix} n_1 \\ n_2 \\ n_3 \end{bmatrix} \propto \begin{bmatrix} \nabla d \\ -1 \end{bmatrix}. \tag{6}$$

Then we can write [**2 points**]

$$d_x = D_x d = -\frac{n_1}{n_3} \tag{7}$$

$$d_y = D_y d = -\frac{n_2}{n_3} \tag{8}$$

where D_x and D_y are the matrices denoting derivatives with respect to x and y. Thus we have a linear system

$$\begin{bmatrix} D_x \\ D_y \end{bmatrix} d = -\begin{bmatrix} \frac{n_1}{n_3} \\ \frac{n_2}{n_3} \end{bmatrix}. \tag{9}$$

[**1 points**]

Optical Flow, Tracking, Registration, Fitting & Recognition [26 points total]

1. Briefly describe a general way to track over many frames in a video. [**4 points**]

   Solution.
   Select features in first frame.[**2 points**]

   For each frame:

   - Update positions of tracked features. [**1 points**]
   - Terminate inconsistent tracks. [**1 points**]
   - Find more features to track. [**1 points**]

2. Does the K-Means Algorithm always converge to the same solution when run multiple times on the same data? Justify your answer. [**4 points**]

   Solution.
   No. [**1 points**] The K-Means algorithms is a non-convex optimization problem, therefore it can end to different local minima when run multiple times. [**3 points**]

3. Briefly describe a way to align two images when the correspondences of the two images are not given. [**4 points**]

   Solution.
   Extract features. [**1 points**] Compute putative matches.[**1 points**] Loop:

   - Hypothesize transformation T. [**1 points**]
   - Verify transformation.[**1 points**]

4. The task of optical flow is to find the motion field u and v by minimizing the functional

$$E[u, v] = |I_x^{t-1} u + I_y^{t-1} v + I^{t-1} - I^t|^2 + \lambda(|\nabla u| + |\nabla v|), \tag{10}$$

where I is the grayscale image, $|\nabla u|$ and $|\nabla v|$ are the total variation on u and v, and t is the index of the video frame. This is derived from the Taylor series expansion (up to the first order) of the brightness constancy equation

$$I(x - u(x, y), y - v(x, y), t - 1) = I(x, y, t). \tag{11}$$

Give three scenarios (based on motion and brightness) where optical flow fails. Justify your answer by using the formulas above. [**6 points**]

Solution.
1. The motion is considerably larger than 1 pixel. In this case the Taylor approximation is not accurate. **2 p**
2. The motion is not uniform locally. TV prior favors locally uniform solutions. **2 p**
3. Brightness changes rapidly. The brightness constancy assumption (and the equation derived from it) does not hold. **2 p**

5. Consider the equation of a parabola $y = ax^2 + x$. Compute the parameter a that best fits the points (1, 3), (-1, 1) and (2, 0.5) with the least squares method. Write the least squares objective and show all your calculations. [**8 points**]

Solution.
The least squares objective is $L(a) = \sum_i (y_i - ax_i^2 - x_i)^2$ **2 p**. This is convex in a, the minimum can be found where $\frac{\partial L}{\partial a} = 0$ **2 p**. Because $\frac{\partial L}{\partial a} = -2\sum_i (y - ax_i^2 - x_i)x_i^2$, we get **2 p**

$$a = \frac{\sum_i (y_i x_i^2 - x_i^3)}{\sum_i x_i^4} = \frac{2 + 0 - 6}{1 + 1 + 16} = -\frac{4}{18}. \tag{12}$$

2 p for final answer.

Epipolar Geometry, Multiple Views & Motion [16 points total]

1. Why is image rectification useful in stereo matching? [**4 points**]

   Solution
   All epipolar lines are perfectly horizontal. [**2 points**] Image rectification makes the correspondence problem easier and reduces computation time. [**2 points**]

2. Epipolar geometry is the intrinsic projective geometry between two views I and I'. It depends only on the camera intrinsic parameters and their relative pose (rotation and translation between the camera centers). The Fundamental Matrix F is a 3×3 matrix.

   (a) How are the fundamental matrices F, going from I to I', and F' going from I' to I, related? [**2 points**]

   Solution
   $F^T = F'$. [**2 points**]

   (b) What is the geometric meaning of the epipoles **e** and **e**$'$? How are they (algebraically) related to the fundamental matrix? [**4 points**]

   Solution
   Geometrically the epipoles are the points of the intersection of the line joining the camera centers (the baseline) with the image planes. Equivalently, an epipole is the image in one view of the camera center of the other view. Algebraically, **e** is the right null space of **F**, i.e. **Fe** $= 0$. Similarly, **e'** is the left null space of **F**, i.e. **e'**^{T}**F** $= 0$. [**4 points**]

   (c) What is the effect of applying the fundamental matrix F to a point x? [**2 points**]

   Solution
   The fundamental matrix maps a point in one image to a line in the second image. $l' = Fx$ and $l = F^T x'$. [**2 points**]

3. Briefly describe one way to improve window-based stereo matching. [**4 points**]

   Solution
   The similarity constraint is local (each reference window is matched independently) [**2 points**]

   Enforce non-local correspondence constraints [**2 points**]

Energy minimization & Bayesian estimation [27 points total]

1. Find the solution to the following energy minimization problem **[6 points]**

$$\arg\min_u |Au - f|^2 + \lambda |u - f|^2 \tag{13}$$

where $A \in \mathbf{R}^{n\times n}$ and $u, f \in \mathbf{R}^n$.

Solution To find the solution we need to compute the gradient and set it to zero **[3 points]**

$$\nabla E = 2A^\top (Au - f) + 2\lambda(u - f) = 0. \tag{14}$$

By rearranging the equation we obtain **[3 points]**

$$A^\top Au - A^\top f + \lambda(u - f) = 0 \tag{15}$$

$$\left(A^\top A + \lambda I\right) u = (A^\top + \lambda I) f \tag{16}$$

$$u = \left(A^\top A + \lambda I\right)^{-1} (A^\top + \lambda I) f. \tag{17}$$

2. Infer the probability that a coin shows up heads, given a series of observed coin tosses. Suppose $X_i \sim$ Ber(θ), where X_i = 1 represents "heads", X_i = 0 represents "tails", and $\theta = p(X_i \equiv$ "head") is the probability of heads. Note: Assume the data samples are iid (independent and identically distributed). **[10 points]**

Solution

Suppose the data are iid, the likelihood has the form **[2 points]**

$$p(D|\theta) = \theta^{N_1}(1-\theta)^{N_0},$$

where we have $N_1 = \sum_{i=1}^N \mathbb{1}(x_i = 1)$ heads and $N_0 = \sum_{i=1}^N \mathbb{1}(x_i = 0)$ tails and $N = N_0 + N_1$. **[2 points]**

Maximizing the likelihood is equivalent to maximizing the log-likelihood: **[2 points]**

$$\log p(D|\theta) = N_1 \log\theta + N_0 \log(1-\theta)$$

Calculating the derivative with respect to θ: **[2 points]**

$$\frac{d \log p(D|\theta)}{d\theta} = \frac{N_1}{\theta} - \frac{N_0}{1-\theta}$$

Setting the derivative to 0 and solving for θ yields: **[2 points]**

$$\frac{N_1}{\theta} - \frac{N_0}{1-\theta} = 0$$

$$\theta_{MLE} = \frac{N_1}{N_0 + N_1} = \frac{N_1}{N}$$

3. Suppose we are given a task of fitting the parameters of a Gaussian Mixture Model (GMM) $p(x, z)$ to the data $\{x^{(1)}, \ldots, x^{(m)}\}$ consisting of m independent samples, where z denotes discrete latent variable. Each $z^{(i)}$ identifies the Gaussian from which the sample $x^{(i)}$ was generated.

   (a) Write the data log-likelihood under a Gaussian Mixture Model. [**3 points**]
   Solution The data likelihood can be expressed as

$$\begin{aligned}\ell(\theta) &= \sum_{i=1}^{m} \log p(x;\theta) \\ &= \sum_{i=1}^{m} \log \sum_{z} p(x, z; \theta).\end{aligned}$$

   (b) Why do we need the EM algorithm to fit the parameters of GMM? Why do we not simply maximize the likelihood by setting $\nabla_\theta \ell(\theta)$ to 0? [**4 points**]
   Solution The main difficulty comes from the fact that variable $z^{(i)}$ are not observed and the posterior does not factorize, making it much harder to compute. [**2 points**] This also complicates the computation of MAP and ML estimates of the parameters. [**2 points**]

   (c) Describe the two main steps of the EM algorithm. [**4 points**]

Solution

4. Given the data likelihood $\ell(\theta)$ in E-step we construct a lower bound on it. [**2 points**] In the M-step of the algorithm, we then maximize our lower-bound with respect to the parameters to obtain a new setting of the θ's. Repeatedly carrying out these two steps gives us the EM algorithm. [**2 points**]