Final Exam

Computer Vision 2011-2012, Utrecht University

Duration: 17.00 - 20.00

Instructions:

1. *Do not look at or read the questions before you are asked to do so.*
2. *Write your name and student number on every separate answer sheet (only when you use more than one sheet).*
3. *Write your answers as complete as possible. Adding correct information to your answers will increase your score (of course, adding wrong information will decrease your score).*
4. *During the exam, you are not allowed to speak with other students. If you fail to comply, you will be asked to leave the exam room and your final-exam grade is zero.*
5. *Ensure that your handwriting is readable. If I cannot read your handwriting, you will be asked to copy your answers by typing them into a computer. If necessary, use capital letters.*
6. *You do not need to write your answers in the same order as the questions, however you have to write the question numbers clearly (failing to do so will cost you points).*
7. *You are allowed to leave the room anytime after 17.30, by first handing in both your answer and question sheets.*
8. *Cheating or other misconduct is not tolerated. In doing so, you will automatically fail to pass the course and will be reported to the exam commission.*

1. **Geometric Image Formation:**

   (a) Write in a matrix form, the intrinsic and extrinsic parameters of a camera, and explain the meaning of the variables in those matrices as detailed as possible.

   (b) Write a step-by-step procedure to do the 3D-to-2D projection.

   (c) Write a step-by-step procedure to do the 2D-to-3D back projection.

   Note: For the last two questions (b and c), you have to mention the inputs, the step-by-step process, and the outputs.

2. **Voxel Reconstruction and Tracking:**

   (a) Write the pseudocode of the voxel-based volume reconstruction that employs a look-up table. Note: you have to include the pseudocode of creating the look-up table, and to mention or indicate the inputs and the outputs of each of them clearly.

   (b) Given voxel data and its 2D images (from 3 cameras), write the algorithm of a two-person tracking using voxel data and MAP (maximum a posteriori) by considering the estimation in the previous time step can be possibly incorrect. To answer this question, you have to include the following mathematical definition of MAP and discuss the meaning of *every mathematical term, variable, and notation* in the final MAP model (Eq.(2)), when $t = 0$, $t = 1$, and $t = 2$:

$$l_t^* = \arg\max_{\{l_t\}} p(l_t|\{d\}_t) \quad (1)$$

$$= \arg\max_{\{l_t\}} p(d_t|l_t) \sum_{\{l_{t-1}\}} p(l_t|l_{t-1}) p(l_{t-1}|\{d\}_{t-1}) \quad (2)$$

   where l is the estimated 3D location, t is the time step, d is an observation, $\{d\}_t$ is the set of observations from time step 0 to t.

3. **Markov Random Field:** Given the image shown in the left figure of Fig.1, our goal is to clean up the noise and have the output as shown in the right figure.

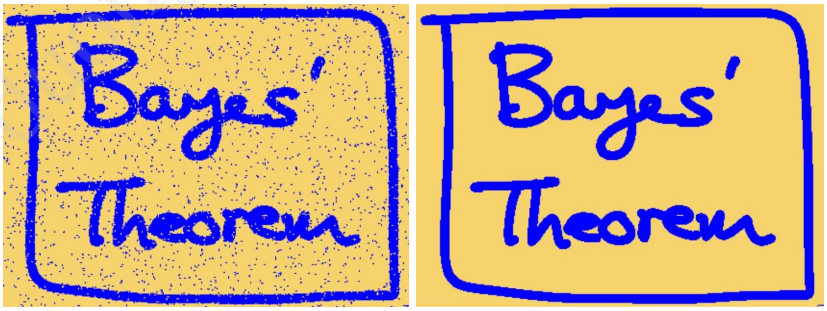


Figure 1: Left: input image. Right: output

To achieve the goal:

   (a) Write the formula describing the mathematical model of Markov random field (which consists of a data and prior term) to obtain the optimum configuration of the labels of hidden variables (=$\{z^*\}$).

   (b) Draw the graphical model of MRF by stating the hidden variables, the observation, the correlation between a hidden variable and its observed data, and the correlation between neighboring hidden variables.

(c) Explain how to model the data term, and the meaning of the model.

(d) Explain how to model the prior term, and the meaning of the model.

4. **Object Recognition:** The questions below are related to the bags of visual words.

   (a) Given images from all classes, write the flowchart of creating the visual words.

   (b) Why do we need all images of all classes to create the visual words?

   (c) Given the generated visual words and the images of a certain class, write the flowchart of creating weak classifiers.

   (d) Why do we need all images of a certain class to create those classifiers?

5. **SIFT (Scale Invariant Feature Transform):**

   (a) A SIFT descriptor consists of 128 numbers. What is the meaning of the formula to generate each of these numbers? If you apply this descriptor to an image, what does the descriptor imply? In other words, what information will you obtain? Hint: every 8 consecutive numbers are taken from 4 by 4 pixels.

   (b) Mention at least 3 main reasons why SIFT can be better for image matching or object recognition compared with other features such as: colors, edges, 2D/3D shapes.

   (c) Consider the image in Fig.2. If we cut the image into two images with respect to the horizontal line in the image (the line separating the person and its reflection), will SIFT be able to match those two images (the image of the person and the image of the person's reflection)? Whether your answer is yes or no, you must provide the reasons.

Figure 2: Mirror reflection

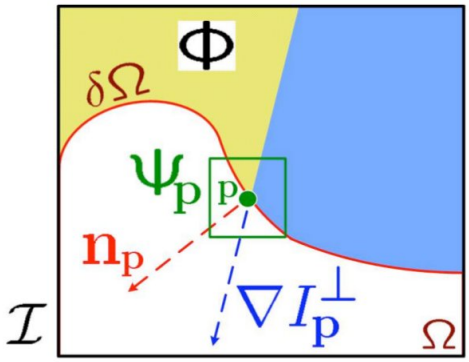
6. **Image Inpainting:**

   (a) In one of the techniques of image inpainting, we should first find a patch that has the highest priority. Given a patch Ψ_p centred at the point p for some $\mathbf{p} \in \delta\Omega$ (see Fig.3), the priority $P(\mathbf{p})$ is defined as:

$$P(p) = C(p)D(p) \tag{3}$$

$$C(p) = \frac{\sum_{q \in \Psi_p \cap (I-\Omega)} C(q)}{|\Psi_p|} \tag{4}$$

$$D(p) = \frac{|\nabla I_p^{\perp} \cdot \mathbf{n}_p|}{\alpha} \tag{5}$$



Notation diagram. Given the patch $\Psi_{\mathbf{p}}$, $\mathbf{n_p}$ is the normal to the contour $\delta\Omega$ of the target region Ω and $\nabla I_{\mathbf{p}}^{\perp}$ is the isophote (direction and intensity) at point $\mathbf{p}$. The entire image is denoted with $\mathcal{I}$.

Figure 3: Inpainting mathematical notations

where during the initialization step, the function $C(\mathbf{p})$ is set to $C(\mathbf{p}) = 0, \forall_{\mathbf{p}} \in \Omega$, and $C(\mathbf{p}) = 1, \forall_{\mathbf{p}} \in I - \Omega$.
What does each of the above equations (i.e., Eq.(3), (4), (5)) intuitively mean?

(b) Write the mathematical definition of $\nabla I_p^{\perp}$ and the function to calculate it in C/C++ (given the inputs of the function are the patch Ψ_p and the location p).

7. **Gradient Space Manipulation:** Given two images, we want to merge them seamlessly. However, a direct copy-and-paste operation will generate discontinuities (boundaries) between those two images. One of the solutions is to use the gradient manipulation technique.

   (a) Mention (at least) 3 basic steps using gradient space manipulation to merge two images seamlessly, and give an example in one dimensional data for each step.

   (b) Write the discrete version of:

$$\frac{\delta}{\delta x}\left(\frac{\delta I}{\delta y}\right) \quad \text{and} \quad \frac{\delta^2 I}{\delta x^2} + \frac{\delta^2 I}{\delta y^2} \tag{6}$$

   where I is a 2D image.

8. **Particle Filtering:**

   (a) The weights in particle filters can be calculated using:

$$w_t^m = w_{t-1}^m \frac{p(d_t|l_t^m)p(l_t^m|l_{t-1}^m)}{q(l_t^m|l_{t-1}^m, d_t)} \tag{7}$$

   where q is the proposal distribution, from which random values are sampled; l is the location to estimate, d is the observation, t is the time step, and m is the index of the particle. Regarding the last equation, what is the meaning of $p(d_t|l_t^m)$, and $p(l_t^m|l_{t-1}^m)$? How to calculate them?

   (b) In tracking, what are the weights w_t^m for?

   (c) Describe the basic idea of solving the degeneracy problem by using resampling. Note, in your answer, discuss the degeneracy problem first, and include the analysis of $p(l_t|\{d\}_t)$ when the time step is t and $t-1$.