ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Computer Vision
AS 2021

Prof. Marc Pollefeys *Prof. Siyu Tang* *Prof. Fisher Yu*

Final Exam

January, 2022

General Remarks

- The following materials are allowed for this exam:
  - exam questionnaire & blank paper (both provided by us)
  - ruler / square & pen
- Remove all material from your desk which is not allowed by examination regulations.
- Wearing headphones is not allowed.
- Please **do not use a pencil or red color pen** to write your answers.
- Place your **student ID** and the **3G certificate** in front of you.
- You have **2 hours** for this exam.
- Fill in your first and last name and your ETH number and sign the exam.
- Put your name and ETH number on top of each sheet.
- Check that your exam questionnaire is complete.
- You may provide at most one valid answer per question. Invalid solutions must be clearly crossed out.
- Write down your answers directly on the exam sheets. You can use both sides.

First and Last name: _____

ETH number: _____

Signature: _____

	Topic	Maximum Points	Points Achieved	Points Check
1	Projective Geometry	10		
2	Deep Learning	10		
3	Model Fitting	10		
4	Stereo / MVS	10		
5	Object Recognition	20		
6	Tracking	20		
7	Epipolar Geometry	20		
Total		100		

Grade: .....

Question 1: Projective Geometry (10 pts.)

For this part, we consider the following 2D lines defined in parametric form as follows:

$$L_1 \begin{cases} x = 5m + 4 \\ y = 2m + 2 \end{cases} \qquad L_2 \begin{cases} x = 3m - 1 \\ y = -m + 2 \end{cases} .$$

a) What is the line equation associated to L_1 (equality with right-hand side 0)? What is the homogeneous representation l_1 associated to L_1? Same questions for L_2. **3 pts.**

b) Given a 2D point (x, y), what is one possible homogeneous representation p? Given the homogeneous representation l of a 2D line, what condition needs to be satisfied by a point in homogeneous representation p laying on the line? Does the point $(6.5, 3)$ lie on L_1? What about $(-2, 5)$ on L_2? **3 pts.**

c) Using their associated homogeneous representations, compute the point of intersection between L_1 and L_2. **2 pts.**

d) What is the equation of a 3D plane (equality with right-hand side 0)? Given a 3D point (x, y, z), what is one possible homogeneous representation p? **2 pts.**

Question 2: Deep Learning (10 pts.)

a) Let us consider a linear layer with a 10-dimensional input $i \in \mathbb{R}^{10}$ and 5-dimensional output $o \in \mathbb{R}^5$. Write the formula relating the output to the input using the weights matrix W and the bias vector b. What is the size of the weights matrix? What is the dimension of the bias vector? What is the total number of trainable parameters? **3 pts.**

b) Name two well-known non-linear activation functions. Give their mathematical definition with respect to the input $x \in \mathbb{R}$ and plot their rough shape between -10 and 10. **3 pts.**

c) Consider the following 4×4 image: $I = \begin{bmatrix} 1 & 5 & 3 & 2 \\ 9 & 0 & 1 & 4 \\ 5 & 1 & 7 & 3 \\ 3 & 4 & 3 & 2 \end{bmatrix}$. What is the output of a max pooling layer with stride 2, kernel size 2×2, and no padding? **1 pt.**

d) Consider the following 3×3 image: $I = \begin{bmatrix} 1 & 5 & 3 \\ 9 & 0 & 1 \\ 5 & 1 & 7 \end{bmatrix}$. What is the output of a convolutional layer with stride 1, padding with 0, and kernel $k = \begin{bmatrix} 1 & 0 & 1 \\ 0 & -1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$? **1 pt.**

e) Another common type of pooling is average (or mean) pooling which returns the average inside its kernel. Unlike min and max pooling, average pooling can be expressed as a convolutional layer. What would be the parameters of the equivalent convolution layer to an average pooling layer with stride 2, kernel size 2×2, and no padding? **2 pts.**

Question 3: Model Fitting (10 pts.)

a) We use RANSAC to robustly fit a model to a set of observations. The model fitting requires a subset of size N. Given the inlier ratio e, write down the formula to compute the maximum number of iterations K such that with probability at least p (e.g., $p = 0.001$), there exist outlier(s) in all the random subsets sampled so far. You can suppose that the sampling is done with replacement. Explain all intermediate steps. **5 pts.**

b) Based on the previous question, write down the main steps of the adaptive RANSAC algorithm that determines online the required number of iterations based on the best inlier ratio e found so far. **5 pts.**

Hint: The inlier ratio e is no longer known, to start with, we can simply initialize it as 0. The probability p and size of random subset N are the same as previous question. k is the number of the current iteration. Based on the previous question we compute online K to represent the maximum number of iterations such that with probability at least p, there exist outlier(s) in each random subset with size N sampled so far.

1. Let $K = \infty$, $e = 0$, and $k = 0$.

Question 4: Stereo / MVS (10 pts.)

a) For stereo, what are the problems when the baseline is too large or too small respectively? **2 pts.**

b) Write down the relationship between Δd (change of disparity d) and ΔZ (change of depth Z) for the standard stereo setup shown in Fig. 1. **2 pts.**
Hint: take the derivative of disparity with respect to the depth $\frac{\partial d}{\partial Z}$.

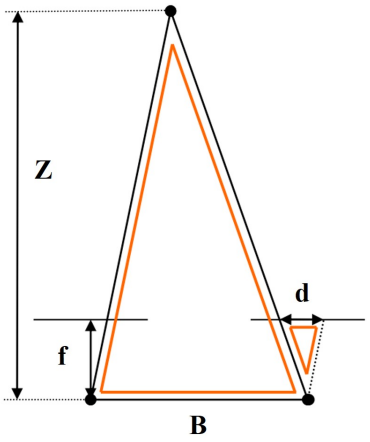

Figure 1: Stereo geometry. Z: depth; f: focal length; B: baseline; d: disparity.

c) What is the main benefit of image pair rectification for stereo matching? Briefly explain how should we do the rectification? **3 pts.**

d) Name two traditional matching criteria for stereo matching. **1 pt.**

e) Given a rectified pair of images, for each pixel $\mathbf{p}$ in the left image, we search for its correspondence in the right image along the scanline. We measure the similarity with the matching criteria within a certain window. What's the problem if the window size is too small? For textureless regions that are not distinct and of high ambiguity, what can we do with the window? **2 pts.**

Question 5: Object Recognition (20 pts.)

a) Given a representative set of training images, explain the process of dictionary construction for the Bag-of-Words (BoW) approach. Given a reference dictionary and a query image, explain how to obtain its BoW representation. **4 pts.**

b) In above steps, how are the 'visual words' represented with regard to the clustering? How is the vocabulary size represented with regard to the clustering? **2 pts.**

c) Given 6 points: (3,1), (3,2), (4,1), (4,2), (1,3), (1,4), and 2 initial centroids: (0,4) and (3,3), write down the process of the first iteration for a K-means clustering procedure (where k=2). **4 pts.**

d) The metrics of Precision and Recall are often used to evaluate BoW classification results. How to compute them? Please write down the definitions of Precision and Recall with TP, TN, FP, FN. **2 pts.**

TP: true positive, TN: true negative, FP: false positive, FN: false negative.

e) Judge the correctness of the following statements and **select** the corresponding box (☑). For each statement, 1 pt, 0 pt and -**1 pt** are given for a correct answer, both empty/selected boxes, and an incorrect answer, respectively. The minimum number of points is 0. **6 pts.**

		True	**False**
1)	Given the same data, the results of K-Means clustering method are the same under different trials.	☐	☐
2)	The initial weight of each classifier in AdaBoost is chosen at random.	☐	☐
3)	Each classifier in AdaBoost is learned independently from each other.	☐	☐
4)	The sliding window approach in object detection has high computational complexity.	☐	☐
5)	R-CNN is a sliding-window approach.	☐	☐
6)	The receptive fields get larger for deeper / further layers in AlexNet.	☐	☐

f) Let us consider a binary classification problem. After the final iteration in AdaBoost, assume that each classifier is noted as $h_m(x)$, and the weight for each classifier is noted as α_m ($m = 1, 2, ..., M$), write down the final prediction function. **2 pts.**

Question 6: Tracking (20 pts.)

a) What are the two main problems with tracking a point with the energy term $E(h) = [I_0(x+h) - I_1(x)]^2$ (2 pts.)? Describe the scenario when each problem occurs (1 pt each). What is the characteristic of a good point to track (1 pt)? **5 pts.**

b) To track an object that might change its appearance over time, we could learn the appearance model online. However, this method could gradually change the target object that we are tracking. How can we prevent the model from tracking the wrong object? **2 pts.**

c) Describe the overview of each step in the Kalman filter. Let $\hat{x}_{k-1|k-1}$ denotes the system state estimation before the k-th measurement y_k. $P_{k|k-1}$ is the corresponding uncertainty. The steps include prior knowledge, prediction, update, estimation, and progressing time step. You can omit the details of the prediction step and update step. Explain the values used during each step in term of $\hat{x}$, P, k, and y if applicable.

5 pts.

d) In a tracking algorithm, we sometimes use the constant-velocity heuristic to predict the potential locations of the object. By doing so, what assumption do we make about the camera? In particular, what information do we need (about the camera) for the velocity heuristic to work properly? **2 pts.**

e) For similarity learning, what is a binary loss? What is a triplet loss? **3 pts.**

f) Judge the correctness of the following statements and **select** the corresponding box (☑). For each statement, 1 pt, 0 pt and **-1 pt** are given for a correct answer, both empty/selected boxes, and an incorrect answer, respectively. The minimum number of points is 0. **3 pts.**

		True	**False**
1)	Histogram of Oriented Gradients is a rotation invariant feature descriptor	☐	☐
2)	Hungarian Algorithm can be used to match candidate regions between two frames based on similarity scores between each pair.	☐	☐
3)	To track by detection, we do not need to know the type of the object to be tracked beforehand.	☐	☐

Question 7: Epipolar Geometry (20 pts.)

You have two images I_1, I_2 of the same scene with camera poses $\mathbf{C}_1$, $\mathbf{C}_2$, respectively. The images contain two corresponding keypoints $\mathbf{x}_1 = (400, 200)^\intercal$ (in I_1) and $\mathbf{x}_2 = (480, 260)^\intercal$ (in I_2).

$$\mathbf{C}_1 = [I|0] = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

$$\mathbf{C}_2 = [R|t] = \begin{bmatrix} 0 & 0 & 1 & -2 \\ 0 & 1 & 0 & 1 \\ -1 & 0 & 0 & 2 \end{bmatrix}$$

(given as transformations from world space to camera space)

The images are of size $(w, h) = (800, 600)$. For both images the focal length is 200 pixels and the principal point is at $(400, 300)$.

a) Compute the Essential Matrix E so that, for a perfect correspondence $(\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2)$, $\hat{\mathbf{x}}_2^\intercal E \hat{\mathbf{x}}_1 = 0$. Scale your solution so that the value in the bottom right corner is 1. **5 pts.**

b) Write down the intrinsic parameter matrix K. **2 pts.**

c) How is the Fundamental Matrix F related to the Essential Matrix? Give the formula and explain the difference between the two matrices. **3 pts.**

d) Write down the projection matrices for the two images. **2 pts.**

e) Compute the residual of the Fundamental Matrix constraint for the correspondence ($\mathbf{x}_1$, $\mathbf{x}_2$). **4 pts.**

Hint: You don't need to explicitly assemble the Fundamental Matrix.

f) Assume you know that the 3D point $\mathbf{X}$ lies in the 3D plane $z = 3$ (i.e., $\mathbf{X}_z = 3$) and its observation $\mathbf{x}_1$ in image I_1 is noise-free. Compute the 3D point position and its 2D reprojection residual in pixels compared to $\mathbf{x}_2$ in image I_2. (You don't need to compute the norm) **2 pts.**

g) What is the main disadvantage of algebraic errors compared to geometric errors? **2 pts.**