Module: **M4. 3D Vision** **Final exam**
Date: February 25, 2021
Teachers: Antonio Agudo, Josep Ramon Casas, Pedro Cavestany, Gloria Haro, Javier Ruiz, Federico Sukno
Time: 2h

Problem 1 *1.8 Points*

Geometric transformations and rectification

(a) Suppose an affinely rectified image. Let $\mathbf{l} = (l_1, l_2, l_3)$ and $\mathbf{m} = (m_1, m_2, m_3)$ be the image of two lines that are orthogonal in the world. Describe the steps of the algorithm that performs the metric rectification of the image from $\mathbf{l}$ and $\mathbf{m}$.

(b) List the 3D transformations that an object can undergo. For each transformation you should explain the anatomy of their operator, their degrees of freedoms and their invariants.

Problem 2 *1.2 Points*

Calibration and pose estimation

(a) Describe the camera resectioning problem, clearly state which are the unknowns and what is the available data.

(b) Derive the expression of the matrix that describes the linear system of equations in the algebraic method for resectioning (DLT method). Justify what is the minimum amount of data we need to solve the problem.

(c) Enumerate two practical scenarios where we can use camera resectioning methods.

(d) What is the main difference between camera resectioning and pose estimation?

Problem 3 *1.8 Points*

We plan to estimate the fundamental matrix F between two images I and I' taken by the same camera using the 8-point algorithm.

a) Given two correspondences $p_1 = (1, 2)$, $p'_1 = (3, 4)$ and $p_2 = (5, 6)$, $p'_2 = (7, 8)$ write the first 2 rows of the matrix W that allows us to estimate the fundamental matrix F (expressed as a vector column f) with a homogeneous system ($Wf = 0$).

Suppose that the singular value decomposition of the matrix W above can be expressed as:

$$W = \begin{bmatrix} 0 & 0 & -1 & 1 & -1 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 1 & 1 & 1 \\ -1 & 1 & 0 & 1 & -1 & 1 & 1 & -1 & 0 \\ 1 & 0 & -1 & 0 & 0 & 0 & -1 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ -1 & 1 & 0 & 0 & -1 & 1 & 1 & -1 & 1 \\ 1 & -1 & -1 & 0 & 0 & 0 & -1 & 1 & -1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & -1 & 1 & -1 & 1 \end{bmatrix} D \begin{bmatrix} 1 & 0 & -1 & 1 & -1 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 & 1 & 1 & 1 \\ -1 & 1 & 0 & 1 & -1 & 1 & 1 & -1 & 1 \\ 1 & 0 & -1 & 0 & 0 & 0 & -1 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ -1 & 1 & 0 & 0 & 0 & -1 & 1 & -1 & 0 \\ 1 & 0 & -1 & 0 & 0 & 0 & -1 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ -1 & 1 & 0 & 0 & 0 & -1 & 1 & -1 & 1 \end{bmatrix}^T$$

b) Obtain a first approximation of the fundamental matrix.

Suposse now than the singular value decomposition of the fundamental matrix obtained in the previous question is:

$$F_b = \begin{bmatrix} 0 & -1 & 0 \\ \frac{\sqrt{2}}{2} & 0 & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & 0 & \frac{\sqrt{2}}{2} \end{bmatrix} \begin{bmatrix} \sqrt{2} & 0 & 0 \\ 0 & \sqrt{2} & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} -1 & 0 & 0 \\ 0 & -\frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ 0 & -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{bmatrix}^T$$

c) Obtain a second approximation of the fundamental matrix that ensures all the epipolar lines cross at the same point (epipole).

d) Is $e = (0, -1)$ the epipole in Image I?

e) Taken into account the fundamental matrix obtained above, are the two images I and I' rectified?

f) Compute the epipolar line l' in image I' for point $p_1 = (1, 2)$.

Problem 4 *1.5 Points*

Depth estimation and new view synthesis.

(a) Describe the main steps for estimating the depth from a pair of stereo rectified images with a local method. How is disparity related to depth?

(b) Define the Sum of Squares Differences and Normalized Cross Correlation. Compare their advantages/disadvantages.

(c) Describe different possible failure cases when estimating the disparity through local methods and two views. What are possible solutions to these problems?

(d) Explain what are the inputs and outputs of the Neural Radiance Field (NeRF) method. Which is the loss used to train the network?

Problem 5 *1.4 Points*

Formulate the structure-from-motion problem by assuming P 2D point tracks in a monocular video with I images (indicating the corresponding size of every matrix) in terms of a measurement matrix **W**, assuming: 1) a perspective camera and a rigid shape, 2) an orthographic camera and a non-rigid shape (you can assume a linear low-rank shape model of rank K). You should provide the entries in every matrix in the most representative way as possible.

To compute shape and motion by factorization, which rank do we have to enforce in every case and how can this be done?

Problem 6 *1.4 Point*

In 2022, a production company in Barcelona named ReVi commits to creating contents for *"Real Virtuality"*, a new line of VR series promoted in a joint initiative of HBO and Netflix media platforms. ReVi will **capture real world** scenes with **2D and 3D sensor**s to be later rendered for virtual reality glasses. The *real* look of the 3D rendered scenes is expected to beat contents from synthetic graphics and attract new audiences towards the VR sector, which has focused more in gaming than in content from real scenes until present.
ReVi decides to hire specialists from the Computer Vision Master and request from them to answer a few questions first. Could you provide short answers to their questions?

a) What kind of sensors should ReVi use to capture 3D outdoor scenes? Should they combine different sensor types, such as adding standard RGB cameras? If so, why would they need to combine such sensors and how can we achieve this "fusion" of 2D/3D captured data?

b) How many sensors should be used in order for the content to be freely navigated (Free Viewpoint) when displayed on VR glasses? Should ReVi care about the number of sensors capturing the scene?

c) Would it be useful to capture the background of an empty set separately, when talent (actors and actresses) are not performing? Which are the cases in which this would work best? How could the foreground (performers and foreground scene objects) added to the already captured background if captured separately?

d) Would we need to limit the possible points of view when rendering the captured scenes for the VR glasses of the users? Why? Discuss both the foreground and the background cases.

e) Propose 3D data processing tools that may be needed for the fusion of sensors in questions a+b, and for the separate capture and fusion of foreground with background in question d. Remember that the foreground sensors may be moving during foreground capture. Name two libraries for point cloud processing that might be useful for this.

f) Should HBO and Netflix be worried about compression schemes for streaming this type of "real virtuality" contents we are talking about? Why? Which tools are available at the moment?

g) How do you think the quality of the rendered scenes would compare to HD/4K video contents or even rendering quality of synthetic scenes in gaming environments?

Problem 7 *0.90 Points*

We wish to use deep learning to process point-cloud data.

a) Would you consider a good idea using a neural network consisting only on fully-connected layers? (Yes / No / Explain why)

b) Explain what is the strategy followed by PoinNet to address the difficulty to get a meaningful ordering of the input points.

c) Another option is to have the point cloud represetned as a graph and use Graph Convolutional Networks (GCNs) which, as their name indicate, allow constructing convolutional layers that can be applied to graphs. Given two function $\mathbf{f}$ and $\mathbf{g}$ defined over the vertices of a graph, explain in detail the *trick* used by GCNs to compute the convolution $\mathbf{f} * \mathbf{g}$ over the graph (this question refers to the work by Bruna et al.).