Module: **M4. 3D Vision** **Final exam**
Date: March 3, 2022
Teachers: Antonio Agudo, Josep Ramon Casas, Gloria Haro, Javier Ruiz, Federico Sukno
Time: 2h

Problem 1 *1.8 Points*

(a) Consider two lines expressed in homogeneous coordinates, demonstrate that if they are parallel, then they intersect at a point at infinity.

(b) Consider two lines which we know are parallel in the real World[1]. When we image them, we obtain lines $\mathbf{l} = (4, 2, 1)^T$ and $\mathbf{m} = (5, 2, 2)^T$.

(i) Can you tell what type of planar transformation there is between the real World and the image we obtained? Justify.

(ii) Based on the knowledge that the two lines $\mathbf{l}$ and $\mathbf{m}$ are parallel in the real World, would you be able to perform metric-rectification to the obtained image? Clearly justify your answer.

(c) Consider the following planar transformation matrix:

$$\mathbf{H} = \begin{bmatrix} 0 & -4 & 5 \\ 4 & 0 & -3 \\ 0 & 0 & 1 \end{bmatrix}$$

(i) Indicate the type of the above planar transformation.

(ii) Enumerate at least 3 invariants for this type of transformation

(iii) Describe with as much detail as possible what will be the effect of transformation $\mathbf{H}$ (e.g. if it were a rotation, by what angle, or if it will translate points, by how much, etc).

(a) Consider lines $\mathbf{l}$ and $\mathbf{m}$; if they are parallel, then we can write $\mathbf{l} = (\ell_1, \ell_2, \ell_3)^T$ and $\mathbf{m} = (\ell_1, \ell_2, m_3)^T$ because their slopes must be the same. These lines intersect at:

$$\begin{aligned} \mathbf{p} &= \mathbf{l} \cap \mathbf{m} \\ &= \mathbf{l} \times \mathbf{m} = \\ &= \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \ell_1 & \ell_2 & \ell_3 \\ \ell_1 & \ell_2 & m_3 \end{vmatrix} = \begin{pmatrix} \ell_2 m_3 - \ell_2 \ell_3 \\ \ell_1 \ell_3 - \ell_1 m_3 \\ \ell_1 \ell_2 - \ell_1 \ell_2 \end{pmatrix} \end{aligned}$$

We see that the third element of the resulting point is zero, which corresponds to a point at infinity in homogeneous coordinates (since the mapping to Cartesian coordinates would imply a division by the third element).

[1] These lines have **no relation** to the lines from the previous item

(b) Lines $\mathbf{l} = (4, 2, 1)^T$ and $\mathbf{m} = (5, 2, 2)^T$ are not parallel in the image, which is clear from the slope or also because they intersect at:

$$\mathbf{p} = \mathbf{l} \times \mathbf{m} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ 4 & 2 & 1 \\ 5 & 2 & 2 \end{vmatrix} = \begin{pmatrix} 2 \\ -3 \\ -2 \end{pmatrix}$$

which correspond to the finite point $(-1, 1.5)^T$ in Cartesian coordinates.

(i) Since parallel lines have been imaged to non-parallel lines, the transformation is a projectivity (8 d.o.f.)

(ii) No; we could at most perform affine-rectification by restoring parallelism between the two lines. This would allow us to remove 2 d.o.f. from $\mathbf{H}$ but we would still have 6 d.o.f. that cannot be recovered (affinity). To perform metric rectification we still need to remove two additional d.o.f. to arrive to a Similarity (4 d.o.f).

(c) Consider the following planar transformation matrix:

$$\mathbf{H} = \begin{bmatrix} 0 & -4 & 5 \\ 4 & 0 & -3 \\ 0 & 0 & 1 \end{bmatrix} \tag{1}$$

(i) The transformation above is a Similarity, since the 2×2 top-left corner is an orthogonal matrix that can be expressed as rotation and scaling, and the first two elements of the last row are zeros.

(ii) Ratio of lengths, ratio of areas, angles, circular points.

(iii) The 2×2 top-left corner is a similarity, which produces a 90 degree rotation and a scaling by a 4.0. On the other hand, the third row introduces a translation by 5 and -3 in the X and Y components, respectively.

Problem 2 *0.9 Points*

Consider two images that correspond to different views of the same planar object. We wish to relate the two images by some planar transformation $\mathbf{H}$

(a) How many degrees of freedom will $\mathbf{H}$ have?

(b) Given a pair of corresponding points $\mathbf{x}$ and $\mathbf{x}'$ (in the first and second image, respectively), how many independent equations can we derive from them (to solve for $\mathbf{H}$)? Write down those equations.

(c) How many correspondence pairs $\{\mathbf{x}; \mathbf{x}'\}$ we need in order to fully determine $\mathbf{H}$? Is there any constraint regarding their selection (i.e. beyond the fact that there cannot be repeated points)?

(a) The transformation is a projectivity, hence it has 8 d.o.f.

(b) We can derive two independent equations. Starting from the effect of the transformation on any point $\mathbf{x}$ to its corresponding point $\mathbf{x}'$ and expressing $\mathbf{H}$ in terms of its rows $\mathbf{h}_1^T$, $\mathbf{h}_2^T$ and $\mathbf{h}_3^T$

$$\mathbf{H}\,\mathbf{x} = \mathbf{x}' = \begin{pmatrix} \mathbf{h}_1^T \\ \mathbf{h}_2^T \\ \mathbf{h}_3^T \end{pmatrix} \begin{pmatrix} x \\ y \\ w \end{pmatrix} = \begin{pmatrix} x' \\ y' \\ w' \end{pmatrix} \tag{2}$$

Therefore:

$$\mathbf{h}_1^T \mathbf{x} = x'$$
$$\mathbf{h}_2^T \mathbf{x} = y'$$
$$\mathbf{h}_3^T \mathbf{x} = w'$$

Now we consider the ratios of x' and y' with respect to the third component and operate:

$$\frac{x'}{w'} = \frac{\mathbf{h}_1^T \mathbf{x}}{\mathbf{h}_3^T \mathbf{x}} \quad \Rightarrow \quad \mathbf{h}_1^T \mathbf{x}\, w' - \mathbf{h}_3^T \mathbf{x}\, x' = 0$$
$$\frac{y'}{w'} = \frac{\mathbf{h}_2^T \mathbf{x}}{\mathbf{h}_3^T \mathbf{x}} \quad \Rightarrow \quad \mathbf{h}_2^T \mathbf{x}\, w' - \mathbf{h}_3^T \mathbf{x}\, y' = 0$$

(c) We need at least 4 pairs $\{\mathbf{x}; \mathbf{x}'\}$ to fully determine $\mathbf{H}$ since each pair provides 2 independent equations (2 d.o.f.). And these 4 pairs must not be co-linear, since otherwise we would not get 8 independent equations.

Problem 3 *1.7 Points*

Camera calibration and pose estimation.

(a) Describe the camera projection matrix and the different elements that form it.

(b) Given an estimation of the camera projection matrix, how would you estimate the internal parameters from it by assuming that the camera has zero skew?

(c) Once you have the internal parameters, how would you estimate the external ones?

(d) In Zhang's algorithm for calibration we need to estimate homographies that relate a template with the different views of it taken by the camera. Consider a generic view of the template, derive the equations that relate the elements of the projection matrix with the columns of the homography associated to that view.

(e) What is the minimum amount of views we need in the Zhang's algorithm. Justify your answer.

(f) Describe the pose estimation problem. How many parameteres do we need to estimate? Which kind of data do the classical methods (non learning-based) use to solve that problem?

(a) The camera projection matrix, P, is a 3×4 matrix that can be written as follows:

$$P = K\,[R \,|\, \mathbf{t}]$$

in terms of the internal parameters, the elements of K, and the external parameters: the rotation matrix R and the translation vector $\mathbf{t}$. The matrix K has the following form:

$$K = \begin{bmatrix} f_x & s & x_0 \\ 0 & f_y & y_0 \\ 0 & 0 & 1 \end{bmatrix}$$

where f_x and f_y express the focal length (in pixels) in dimensions x and y respectively, s is the skew factor and (x_0, y_0) are the coordinates of the principal point.

(b) We can use the elements of the 3×3 submatrix of P which are related to the elements of K as follows:

$$A = P_{3\times3} P_{3\times3}^T = (KR)(KR)^T = KRR^TK = KK^T$$

$$\begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix} = \begin{bmatrix} f_x & 0 & x_0 \\ 0 & f_y & y_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} f_x & 0 & 0 \\ 0 & f_y & 0 \\ x_0 & y_0 & 1 \end{bmatrix} = \begin{bmatrix} f_x^2 + x_0^2 & x_0\, y_0 & x_0 \\ x_0\, y_0 & f_y^2 + y_0^2 & y_0 \\ x_0 & y_0 & 1 \end{bmatrix}$$

Then we normalize $(P_{3\times3} P_{3\times3}^T)_{33} = A_{33} = 1$ and establish the following equations:

$$x_0 = A_{13}, \quad y_0 = A_{23}$$

$$f_x = \sqrt{A_{11} - x_0^2}, \quad f_y = \sqrt{A_{22} - y_0^2}$$

(c) $[R\,|\,\mathbf{t}] = K^{-1}P$

(d) We know the relation of a flat object and its image projection, which is given by a homography H with columns $\mathbf{h}_1$, $\mathbf{h}_2$, $\mathbf{h}_3$:

$$\begin{bmatrix} \lambda u \\ \lambda v \\ \lambda \end{bmatrix} = \underbrace{\begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix}}_{H} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = [\mathbf{h}_1\,\mathbf{h}_2\,\mathbf{h}_3] \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}$$

On the other hand, we consider the flat object (the template pattern) to be at plane $Z = 0$. It is projected to the image with the following equation:

$$\begin{bmatrix} \alpha u \\ \alpha v \\ \alpha \end{bmatrix} = K \left[\underbrace{\mathbf{r}_1\,\mathbf{r}_2\,\mathbf{r}_3}_{R}\ \mathbf{t}\right] \begin{bmatrix} x \\ y \\ 0 \\ 1 \end{bmatrix} = [K\mathbf{r}_1\,K\mathbf{r}_2\,K\mathbf{t}] \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}$$

By comparing both equations and taking into account the two different scaling factors involved (α and λ) we have that:

$$[K\mathbf{r}_1\,K\mathbf{r}_2\,K\mathbf{t}] \sim [\mathbf{h}_1\,\mathbf{h}_2\,\mathbf{h}_3]$$

(e) The minimum number of views is three. We need to estimate the elements of the absolute conic (5 d.o.f.) and each view provides us with two equations, then we need a minimum of three views if no further assumptions are done.

(f) In pose estimation we know the internal parameters of the camera (intrinsics) and the goal is to estimate the position and orientation (extrinsics) of the camera, that is 6 d.o.f. Classical methods use 3D-to-2D point correspondences to solve the problem.

Problem 4 *1.8 Points*

Consider the two images I (left) and I' (right) with non-coincident optical centers. Answer the following questions:

a) What is the rank of the corresponding fundamental matrix F between the two images?

b) How many degrees of freedom does the fundamental matrix F have?

c) All corresponding points p and p' in the two images must satisfy the epipolar constraint. What is the mathematical expression for the epipolar constraint?

d) What is the mathematical expression in terms of F for the epipolar line l' in image I' (right) for the point p in image I (left)?

e) What is the mathematical expression in terms of F for the epipolar line l in image I (left) for the points p' in image I' (right)?

f) What is the mathematical relationship between the epipole e on image I (left) and F?

g) What is the mathematical relationship between the epipole e' on image I' (right) and F?

h) Why is the epipolar constraint useful for stereo matching (i.e. finding corresponding points between left and right images)?

i) What happens to the epipoles and the epipolar lines in the rectified images after applying image rectification?

j) What are the advantages of applying image rectification before we do stereo matching?

a) 2

b) 7

c) $\tilde{p}'^T F \tilde{p} = 0$

d) $l' = F\tilde{p}$

e) $l = F^T \tilde{p}'$

f) $F\tilde{e} = 0$

g) $F^T \tilde{e}' = 0$

h) only searches in 1D line.

i) They become paralel to X axes

j) only searches in X coordinates

Problem 5 *1 Point*

Triangulation and depth.

(a) We want to solve the triangulation problem with the geometric method, which is the minimization problem we need to solve? Define all the involved variables.

(b) Define the concept of signed distance function in the context of 3D reconstruction.

(c) How can we obtain the signed distance function of the surface associated to a certain depth map?

(d) Explain the main idea of estimating a 3D reconstruction by depth map fusion.

(a)

$$\min_{\widehat{\mathbf{x}},\widehat{\mathbf{x}}'} d^2(\mathbf{x},\widehat{\mathbf{x}}) + d^2(\mathbf{x}',\widehat{\mathbf{x}}') = \min_{\widehat{\mathbf{x}},\widehat{\mathbf{x}}'} \|[\mathbf{x}] - [\widehat{\mathbf{x}}]\|_2^2 + \|[\mathbf{x}'] - [\widehat{\mathbf{x}}']\|_2^2$$

$$\text{such that } \widehat{\mathbf{x}}'^T F \widehat{\mathbf{x}} = 0.$$

where $\mathbf{x}, \mathbf{x}', \widehat{\mathbf{x}}, \widehat{\mathbf{x}}'$ are 2D points in homogeneous coordinates, F is the fundamental matrix that relates the two images (3×3 matrix), and the operator $[\cdot]$ represents the conversion from homogeneous to cartesian coordinates.

(b) A signed distance function is an implicit representation of a surface as a scalar function of three variables. At every 3D point the value of the function indicates the signed distance of the point to the surface (being negative outside the surface and positive inside). The surface is represented as the zero level set of the signed distance function.

(c) Given a depth map d the signed distance function at a 3D point $\mathbf{z}$ ($\mathbf{X}$ in homogeneous coordinates) can be computed as:

$$sdf(\mathbf{z} = [\mathbf{X}]) = (P\mathbf{X})_3 - d([P\mathbf{X}])$$

where P is the projection matrix of the camera corresponding to the depth map. $[\cdot]$ represents the conversion from homogeneous to cartesian coordinates and $(\cdot)_3$ denotes the third coordinate.

(d) We consider different viewpoints and compute the depth map of every view. From every depth map d_i we compute its associated signed distance function, sdf_i. Then we average all signed distance functions. The resulting signed distance function encodes the average surface of the surfaces encoded by the different depth maps.

Problem 6 *1.4 Points*

Consider the structure-from-motion problem: 1) Explain briefly the alternatives that you know to solve the rigid case in terms of processing and algorithms. Provide the most important characteristics of every alternative. 2) Let us consider a toy problem composed of 50 points (full visibility) of a deformable object and 10 images, as well as a low-rank linear subspace model of rank 4 to encode the shape deformations. Obtain the number of unknowns and the equations to solve the problem by bundle adjustment. To do that, you only need to consider the data term.

1) The problem can be solved by factorisation or by non-linear optimisation. In the first case, a closed-form solution can be achieved by using SVD, enforcing a specific rank (3 for an orthographic camera, and 4 for a perspective one). In theory, it is hard to accurately enforce constraints. In non-linear optimisation the solution is achieved iteratively, the computational cost can be bigger but additional priors can be enforced accurately. On the other hand, in terms of processing, the problem can be solved in an offline or online fashion. In the first case, all the frames are processed at once, after video capture. In the second one, the frames are processed as the data arrive, frame by frame. More real applications can be considered but can become less accurate.

2) 1000 equations. 680 unknowns.

Problem 7 *1.0 Points*

Range sensors and 3D data.

(a) Describe in a few lines these two technologies for range imaging: structured light (SL) and time-of-flight (ToF).

(b) Mention one aspect that SL and ToF have in common and one main difference.

(c) Why is data acquired with range sensors sparse? Provide an example of dense (non-sparse) 3D data (not acquired with range sensors)

(a) **Structured Light** (SL) is the process of projecting a known pattern (often grids or horizontal bars) onto a scene. The deformation of the pattern when striking surfaces allows calculating the depth and surface information of the objects in the scene.
A **Time-of-Flight** sensor (ToF) is a range imaging system that resolves the distance by computing the time it takes for the light produced by the sensor laser or LED to travel its path along the round trip between the camera and the subject for each point of the image.

(b) As with all active optical methods, reflective or transparent surfaces raise difficulties for both SL and ToF. Reflections cause light to be reflected either away from the camera or right into its optics, which may exceed the dynamic range of the sensor in both cases. Transparent or semi-transparent surfaces also cause major difficulties, as light is refracted or not bouncing back to the sensor.
A main difference between SL and ToF is the strategy to compute range. Measuring time-of-flight is simpler and direct, but challenging due to the speed of light, whereas structured light patterns help solving the stereo correspondence problem by scanning multiple points or the entire field of view at once. Scanning an entire field of view in a fraction of a second may help reducing the problem of distortion from motion in sequential ToF scanners, but computation may require more time than the direct measure of time-of-flight.

(c) Range data is sparse because range sensors just sense the depth at the opaque interface in the field of view in front of the sensor, i. e. range sensor sample of a 2D surface in a 3D scene, which will be sparse by nature.
On the contrary, dense 3D volumetric representations are produced by sensing several depth levels for every sensing direction. This is the case in medical imaging, for both Computed Tomography (CT) and Magnetic Resonance Imaging (MRI), which generate 3D images of the inside of objects/tissues from a large series of 2D (X-ray or ultrasound) images at sequential depths.

Problem 8 *0.4 Points*

In pointcloud processing, organized data structures may help in vision analysis tasks. Graphs and trees have been proposed to analyze (segment, detect, classify) pointclouds.

a) Explain how adjacency, hierachy, captured primitives (visual or geometric features) and visual boundaries play a role for analysis tasks within the edge and node elements of graphs and trees.

b) What is the reason why graph and tree structures, specially for point clouds, are expected to perform better than raw data?

a) The graph representation simplifies the sparse and inhomogeneous input data by grouping homogeneous points on of the point cloud into nodes, while preserving the boundary information. Such representations are appropriate for building a hierarchical description and for performing scene analysis and segmentation. Supervoxel connectivity graphs, for instance, add adjacency relationships and homogenize the raw data in pointclouds. This can be used at the basis of a hierarchical tree segmentation structure in several levels, varying from coarse to fine. Such structures can represent, for instance, object segmentation at different scales of object-connectivity, from the super-voxel graph up to the scene level. We can use Graphs or Trees to define meaningful primitives for processing, detection and classification in the nodes, whereas edges will represent relationships and hierarchies. Data pooling procedures are then able to enrich edges and nodes with features and/or classes. Additionally, trees and graphs use to be well suited for parallel computing.

b) Pointclouds are a relatively sparse kind of data, unstructured and based on simple and numerous primitives (the points in the cloud). This is why graph structures have become an essential tool for modeling point clouds obtained from RGB-D sensors.