Module: **M6. 3D Vision** **Final exam**
Date: April 28, 2016
Teachers: Coloma Ballester, Josep Ramon Casas, Gloria Haro, Javier Ruiz **Time: 2h**

- Books, lecture notes, calculators, phones, etc. are not allowed.
- All sheets of paper should have your name.
- Answer each problem in a separate sheet of paper.
- All results should be demonstrated or justified.

Problem 1 *1 Point*

(a) *(0.25 points)* How do we represent a planar projective transformation in the projective space? How many degrees of freedom does it have?

A 2D projective transformation (2D homography) is represented by a 3×3 non-singular matrix. It has 8 degrees of freedom (9 elements - a scaling factor).

(b) *(0.5 points)* How does a planar projective transformation act on points and lines?

Let H be a 2D homography.

A point $\mathbf{x} \in \mathbb{P}^2$ is transformed by $\mathbf{x}' = H\mathbf{x}$.

A line $\mathbf{l} \in \mathbb{P}^2$ is transformed by $\mathbf{l}' = H^{-T}\mathbf{l}$.

(c) *(0.25 points)* Which are the geometric invariants for a planar projective transformation?

Collinearity, concurrency, cross ratio, and order of contact.

Problem 2 *1 Point*

Consider the problem of computing a 2D homography H between two image views of a plane objet. Let $\mathbf{x}_i$ in $\mathbb{P}^2$, $i = 1, \ldots, n$, be a set of points on the first image and let $\mathbf{x}'_i$ in $\mathbb{P}^2$, $i = 1, \ldots, n$, be a set of points on the second image such as, in pairs, they correspond: $\mathbf{x}_i \longleftrightarrow \mathbf{x}'_i$, $\forall i = 1, \ldots, n$.

(a) *(0.25 points)* What is the minimum value of n? More precisely, how many corresponding points in general position do you need to compute H such that $\mathbf{x}'_i = H\mathbf{x}_i$, $\forall i = 1, \ldots, n$?
(Recall that general position means that no three points are collinear).

The minimum number n of corresponding points in general position is four because the 2D homography H has nine entries to compute, minus one scale factor. That is, eight unknowns. On the other hand, each pair of corresponding points provides two equations.

(b) *(0.75 points)* Describe the Normalized Direct Linear Transformation (Normalized-DLT) algorithm to compute H.

The Normalized-DLT applies a normalization of the data consisting of translation and scaling of image coordinates before applying the DLT algorithm. Finally, an appropriate correction to the

result expresses the computed H with respect to the original coordinate system. More precisely, slides 14-15 of lecture2.pdf, where the usual DLT algorithm is summarized on slide 13.

Problem 3 *1 Point*

Consider an image of a 3D scene containing flat objects.

(a) *(0.5 points)* Explain the method of affine rectification via the vanishing line.

First, we compute the line at infinity ℓ on the image which has a projective distortion. To this goal, we take two sets of two parallel lines, be it $\ell^a, \ell^b, \ell^c, \ell^d$, and we compute the vanishing point of each pair of parallel lines, $v^{ab} = \ell^a \times \ell^b$ and $v^{cd} = \ell^c \times \ell^d$. Then, from these two points, which are on the vanishing line $\ell = (l_1, l_2, l_3)$, compute the vanishing line as $\ell = v^{ab} \times v^{cd}$.

Finally, the projective transformation of $\mathbb{P}^2$ given by $\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ l_1 & l_2 & l_3 \end{pmatrix}$ affinely rectifies the image.

Moreover, the family of projective transformations of $\mathbb{P}^2$ that map ℓ to $\ell_\infty = (0,0,1)^T$ can be written as

$$H_{a \leftarrow p} = H_a \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ l_1 & l_2 & l_3 \end{pmatrix},$$

where H_a is any affine transformation.

(b) *(0.5 points)* Explain the method of metric rectification via orthogonal lines.

The method of metric rectification via orthogonal lines is applied to an image which has been affine rectified (for instance, using the previous method (a) and uses two pairs of lines that are orthogonal in the world.

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. Then $\mathbf{l}^T M \mathbf{m} = 0$, where $M = \begin{pmatrix} S & \vec{0} \\ \vec{0}^T & 0 \end{pmatrix}$ and $S = \begin{pmatrix} s_1 & s_2 \\ s_2 & s_3 \end{pmatrix}$ is symmetric. This equation writes

$$(l_1 m_1, l_1 m_2 + l_2 m_1, l_2 m_2)\vec{s} = 0,$$

where $\vec{s}^T = (s_1, s_2, s_3)^T$ is the vector with the entries of S.

The image of two such orthogonal pairs provide two equations and they permit to compute (s_1, s_2, s_3) as its null vector (is an homogenous system of two equations with three unknowns).

The Cholesky decomposition of S allows to compute an upper triangular matrix K such that $S = KK^T$. The matrix K is a possible matrix A that can be used to metrically rectify the image.

Indeed, it suffices to define $H_{a \leftarrow s} = \begin{pmatrix} K & \vec{0} \\ \vec{0}^T & 1 \end{pmatrix}$ and $H = H_{\mathrm{a} \leftarrow \mathrm{s}}^{-1} H_{\mathrm{a}}$, where

$$H_{\mathrm{a} \leftarrow \mathrm{s}}^{-1} = H_{\mathrm{s} \leftarrow \mathrm{a}} = \begin{pmatrix} K^{-1} & \vec{0} \\ \vec{0}^T & 1 \end{pmatrix}$$

We know that H is a similarity. Let's denote it by H_{s}. That is, $H_{\mathrm{a}} = H_{\mathrm{a} \leftarrow \mathrm{s}} H_{\mathrm{s}}$. The rectified image can be defined by

$$u_{\mathrm{metrect}}(\vec{x}_s) = u_{\mathrm{affrect}}([H_{\mathrm{a} \leftarrow \mathrm{s}} \mathbf{x}_s]),$$

where $\mathbf{x}_s = (\vec{x}_s, 1)$ and $[(p_1, p_2, p_3)] = (p_1/p_3, p_2/p_3)$.

Problem 4 *0.5 Points*

What is the general form of a finite projective camera matrix P? Describe its internal and external parameters.

P decomposes in $P = K[R|\mathbf{t}]$, where K and R are 3×3 matrices and $\mathbf{t}$ is a 3×1 vector. K is the calibration matrix containing the internals parameters, and R, $\mathbf{t}$ represent the external parameters of the camera. In particular,

$$K = \begin{pmatrix} \alpha_x & s & x_0 \\ 0 & \alpha_y & y_0 \\ 0 & 0 & 1 \end{pmatrix}$$

and R and $\mathbf{t}$ give the position and orientation of the camera in the world coordinate system. See lecture3.pdf for the description of the internal parameters $x_0, y_0, \alpha_x, \alpha_y, s$.

Problem 5 *2 Points*

We are studying a system of two cameras where the intrinsic parameters of both cameras are $K = K' = \begin{bmatrix} 10 & 0 & 0 \\ 0 & 10 & 0 \\ 0 & 0 & 1 \end{bmatrix}$, the estimated fundamental matrix of the system is $F = \begin{bmatrix} 0 & 1 & 1 \\ -1 & 0 & 0 \\ -1 & 0 & 0 \end{bmatrix}$ and two possible pixel correspondences are $p_1 = (5, 0)$, $p'_1 = (10, 1)$ and $p_2 = (1, 1)$, $p'_2 = (1, 10)$. Answer the following questions:

(a) Explain briefly the main difference between the fundamental matrix F and the essential matrix E.

F expresses the relation in uncalibrated cameras (in pixel coordinates) while E expresses the relation in the calibrated case (camera coordinates).

(b) State very briefly the steps to estimate the fundamental matrix F using the 8-point algorithm.

1. Create matrix W from p_i and p'_i correspondences.
2. Compute the SVD of matrix $W = UDV^T$
3. Create vector f from last column of V.
4. Compose fundamental matrix F_{rank3}.
5. Compute the SVD of fundamental matrix $F_{rank3} = UDV^T$.
6. Remove last singular value of D to create D'.
7. Re-compute matrix $F = UD'V^T$ (rank 2).

(c) Justify if $e = (0, -1)$ is an epipole. What does it mean if one of the coordinates of the epipole is negative?

$Fe = 0$ so it is an epipole. Negative coordinates mean it is outside the image.

(d) Justify if any of the two correspondences p_1, p'_1 or p_2, p'_2 is an outlier.

$\tilde{p}'^T F \tilde{p} = 0$ is satisfied for the first correspondence and not for the second, thus the second correspondence is outlier.

(e) Find the essential matrix E of the system.

$E = K'^T F K$ so $E = \begin{bmatrix} 0 & 10 & 1 \\ -10 & 0 & 0 \\ -1 & 0 & 0 \end{bmatrix}$ (up to scale)

Let us suppose now that the matrix E of the system could be decomposed in:

$$E = UDV^T = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 & 1 \\ -1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}^T$$

(f) Obtain the two possible translation vectors of the system.

$T = \pm u_3$ (last column of U) $T = \pm(1, 0, 0)^T$.

(g) Obtain the two possible rotation matrices of the system (Note: you might need the matrices $W = \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ or $Z = \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$).

$R = UWU^T = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix}$ and $R = UW^TU^T = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$

(h) Explain briefly how would you decide the correct T and R from the 4 possible solutions obtained above.

Project a point X to determine in which of the 4 configurations it is in front of both cameras.

Problem 6 *0.75 Points*

Triangulation methods.

(a) *(0.25 points)* Describe the triangulation problem, what are the unknowns and the available data.

Given two corresponding points $\mathbf{x}$, $\mathbf{x}' \in \mathbb{P}^2$, the problem is to estimate a point $\widehat{\mathbf{X}} \in \mathbb{P}^3$ that satisfies $\widehat{\mathbf{x}} = P\widehat{\mathbf{X}}$, $\widehat{\mathbf{x}}' = P'\widehat{\mathbf{X}}$, for some points $\widehat{\mathbf{x}}$ and $\widehat{\mathbf{x}}'$ in the images near the corresponding points $\mathbf{x}$, $\mathbf{x}'$.

(b) *(0.25 points)* Which is the minimization problem we need to solve when we use the homogenoeus linear method and how its solution is computed?

The homogeneous linear method is based on the following minimization problem:

$$\min_{\mathbf{X}} \|\mathbf{A}\mathbf{X}\|_2$$

$$\text{such that } \|\mathbf{X}\|_2 = 1.$$

The solution is the singular vector associated to the minimum singular value of $\mathbf{A}$.

(c) *(0.25 points)* Which is the minimization problem we need to solve when we use the geometric method?

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

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

where F is the fundamental matrix that relates the two images and the operator [.] represents the conversion to cartesian coordinates.

Problem 7 *1.5 Point*

Consider the *structure from motion* problem.

(a) *(0.25 points)* Describe the problem, what are the unknowns and the available data.

The data available in the structure from motion problem are a set of images (views) and a set of point correspondences across different pairs of images; these points correspond to the 2D projections of a set of unknown 3D points. The goal is to find the 3D points that project to the given image points in the different images and the camera projection matrices associated to each view.

(b) *(0.25 points)* Why is there a projective ambiguity in the reconstruction if we don't assume any further information?
Let $\mathbf{x} \in \mathbb{P}^2$ and $\mathbf{X} \in \mathbb{P}^3$, then by projecting $\mathbf{X}$ to the image plane via the camera projection matrix P we get the 2D point $\mathbf{x} = P\mathbf{X}$. In the structure from motion problem $\mathbf{x}$ is known and P and $\mathbf{X}$ are the unknowns. If P and $\mathbf{X}$ are a possible solution, i.e. $\mathbf{x} = P\mathbf{X}$, then, given any 3D homography H we have that $\hat{P} = PH^{-1}$ and $\hat{\mathbf{X}} = H\mathbf{X}$ since $\mathbf{x} = \hat{P}\hat{\mathbf{X}}$.

(c) *(0.25 points)* What are the main steps of a stratified reconstruction method?
Projective reconstruction.
Affine reconstruction (optional).
Metric reconstruction.

(d) *(0.5 points)* Explain the main idea of the factorization method that allows us to estimate a solution of the structure from motion problem (just explain the essence of the algorithm, we are not asking for a detailed pseudo-code containing all the technical and numerical details).
Given the projective equations:

$$\mathbf{x}_j^i \equiv P^i \mathbf{X}_j \qquad \longrightarrow \qquad \lambda_j^i \mathbf{x}_j^i = P^i \mathbf{X}_j$$

where $j = 1, ...n$ denote the points, $i = 1, ...m$ denote the images (views), and λ_j^i are unknown scalar factors, called *projective depths*. We can collect all projective eq's into a matrix equation:

$$\begin{bmatrix} \lambda_1^1\mathbf{x}_1^1 & \lambda_2^1\mathbf{x}_2^1 & ... & \lambda_n^1\mathbf{x}_n^1 \\ \lambda_1^2\mathbf{x}_1^2 & \lambda_2^2\mathbf{x}_2^2 & ... & \lambda_n^2\mathbf{x}_n^2 \\ ... & ... & ... & ... \\ \lambda_1^m\mathbf{x}_1^m & \lambda_2^m\mathbf{x}_2^m & ... & \lambda_n^m\mathbf{x}_n^m \end{bmatrix} = M = \begin{bmatrix} P^1 \\ P^2 \\ ... \\ P^m \end{bmatrix} \begin{bmatrix} \mathbf{X}_1 & \mathbf{X}_2 & ... & \mathbf{X}_n \end{bmatrix}$$

We have

$$\underbrace{M}_{3m \times n} = \underbrace{\begin{bmatrix} P^1 \\ P^2 \\ ... \\ P^m \end{bmatrix}}_{3m \times 4} \underbrace{\begin{bmatrix} \mathbf{X}_1 & \mathbf{X}_2 & ... & \mathbf{X}_n \end{bmatrix}}_{4 \times n}$$

then M has at most rank 4 and a possible solution for the camera projection matrices and the 3D points is obtained by the SVD decomposition of M, $M = UDV^T$,

$$\begin{bmatrix} P^1 \\ P^2 \\ ... \\ P^m \end{bmatrix} = UD_4 \qquad \begin{bmatrix} \mathbf{X}_1 & \mathbf{X}_2 & ... & \mathbf{X}_n \end{bmatrix} = V_4^T$$

where D_4 is the $n \times 4$ submatrix of D, and V_4^T is the $4 \times n$ submatrix of V^T.

(e) *(0.25 points)* What is the principal limitation of the data we require in the factorization method?
The main limitation is that the set of 3D points must be visible in all the images and we need to locate them in all the views. That is, given a set of points in one image we need to find, for every point, its correspondences in the rest of the images.

Problem 8 *0.75 Points*

Describe which is the projective transformation we need so as to update a projective reconstruction to an affine one. Explain how we can estimate the elements of this transformation.

The projective transformation we need is represented by a 4×4 non-singular matrix which has the following form:

$$H_{a \leftarrow p} = \begin{pmatrix} I & \mathbf{0} \\ \Pi^T & 1 \end{pmatrix}$$

where Π represents the image of the plane at the infinity.

As every plane, the plane at infinity is determined by three points on it, let us denote them by $\mathbf{X}_i$, $i = 1, 2, 3$. Then we have that $\Pi^T \mathbf{X}_i = 0$, $i = 1, 2, 3$, which can be rewriten as:

$$\underbrace{\begin{pmatrix} \mathbf{X}_1^T \\ \mathbf{X}_2^T \\ \mathbf{X}_3^T \end{pmatrix}}_{A\,(3\times4 \text{ matrix})} \Pi = \mathbf{0}$$

If the three points are in general position (not on the same line), they provide linearly independent equations and the matrix A they form is rank 3. Thus Π is obtained uniquely (up to scale) as the 1-dimensional right null space of A.

There are different ways to compute the three points on the plane at infinity. One possibility is to estimate the 2D vanishing points on the images and triangulate them to obtain the points $\mathbf{X}_i$.

Problem 9 *0.5 Points*

Depth sensors

The first generation of depth sensors (Kinect1, Asus, Structure, Orbbec ...) can be classified as non-contact distance measurement methods. Among the following alternatives, could you choose the correct sub-classification for first generation depth sensors and, for each alternative, state a reason for your choice in a few words?

(a) non-optical/optical

(b) passive/active
Active, as depth sensors project infrared light into the scene

(c) stereo/structure from motion/shape from silhouette/light coding/time-of-flight
Light coding, as they exploit the setup of a stereo rig where one of the sensors has been replaced by a projector (light) and the projected patterns (structured light) contain code-words (coding) to ease the correspondence task

(d) triangulation/active stereo
Triangulation, as they employ a structured light to find projector-camera correspondences (whereas active stereo employs a structured light but performs stereo matching only for camera-camera correspondences, in the same way as passive stereo)

Problem 10 *0.5 Points*

Point clouds

A point cloud obtained from a depth sensor can be explained as a non-regular sampling of 3D space. Why? Define what is an organized point cloud and discuss the advantages of organized point clouds over unorganized point clouds in terms of 3D processing and analysis.

In terms of 3D sampling, even if the depth values are sampled in lines and columns at regular intervals in the image plane of the sensor, once the sampled points are back-projected to 3D space to form the point cloud, the 3D spatial intervals between points in the cloud are not regular, as inter-distances depend on the depth measured at each point.

An organized point cloud is arranged as a 2D array of points with the same properties of points obtained from a projective camera. We have X Y and Z for each point, but the memory layout is that of a 2D array and closely related to the spatial layout as represented by these XYZ values. In unorganized point clouds points are simply listed in a list without any specific order.

Problem 11 *0.5 Points*

Depth scans and meshing

Scanning a whole object with a single sensor requires either rotating it or moving the sensor around it. In either option, describe the additional computer vision techniques you need to apply to the depth images or point clouds resulting from the sequence of scanned frames to generate a full watertight mesh of the scanned object.

Geometric registration of subsequent scanned depth frames to build the 3D volume. This could be done, for instance with Iterative Closest Point to minimize the difference of the subsequent clouds of points. It could also be done adapting Structure from Motion as applied for image sequences.

Once a point cloud with all sequence depth data registered is obtained, the next step is obtaining a mesh from the set of surface samples. This step can be approached from computational geometry, surface fitting or implicit function fitting. Poisson surface reconstruction is one example of these techniques. The mesh will define the ordering of surface points for interpolating a continuous surface in intermediate positions.

Finally, in some cases, there is a requirement for the mesh to be watertight. This implies filling any hole in the surface by smooth surface interpolation.