Midterm 1: Advanced Topics in Computer Vision (580.464)

Instructor: René Vidal

March 30, 2005

HONOR SYSTEM: This examination is strictly individual. You are not allowed to talk, discuss, exchange solutions, etc., with other fellow students. Furthermore, you are not allowed to use the book or your class notes. You may only ask questions to the class instructor. Any violation of the honor system, or any of the ethic regulations, will be immediately reported according to JHU regulations.

NAME: _____ **Signature:**_____

CHEAT SHEET:

1. Brightness constancy constraint (BCC): $I(x+u, y+v, t+1) - I(x,y,t) = 0 \approx I_x u + I_y v + I_t$.
2. Two view geometry of a nonplanar scene: $\lambda_2 \boldsymbol{x}_2 = \lambda_1 R \boldsymbol{x}_1 + T \implies \boldsymbol{x}_2^T \widehat{T} R \boldsymbol{x}_1 = 0$.
3. Two view geometry of a planar scene: $\lambda_1 N^T \boldsymbol{x}_1 = d$ and $\lambda_2 \boldsymbol{x}_2 = \lambda_1 R \boldsymbol{x}_1 + T \implies \widehat{\boldsymbol{x}_2} H \boldsymbol{x}_1 = 0$, where $H = (R + T N^T/d)$ is the so-called homography.
4. **Lemma 1** *If A is positive definite, then*

$$\max_{Q \in SO(3)} trace(QA) = trace(A) \tag{1}$$

   Therefore, one solution to the optimization problem is $Q = I$.
5. Given $B = AA^T$, then A can be computed up to a rotation via QR decomposition. If in addition A is upper triangular, then A can be computed via Choleski decomposition.
6. Reprojection error: Let $\boldsymbol{x}_{fp} \in \mathbb{R}^2$ be the image of point $\boldsymbol{X}_p \in \mathbb{R}^3$ in frame f. Let $\pi : \mathbb{R}^3 \to \mathbb{R}^2$ be a projection map. The reprojection error is

$$\sum_{p=1}^{P} \sum_{f=1}^{F} \|\boldsymbol{x}_{fp} - \pi(R_f \boldsymbol{X}_p + T_f)\|^2. \tag{2}$$

1. **(10 Points) Optical flow with changes in illumination.** Let $I(x, y, t)$ be a video sequence taken by a rigidly moving camera observing a rigid, static and Lambertian scene. Assume that between two consecutive views there is an affine change in the image intensities, *i.e.* the brightness constancy constraint reads

$$I(x + u, y + v, t + 1) = aI(x, y, t) + b, \tag{3}$$

where $u(x, y)$ and $v(x, y)$ represent the optical flow (motion parameters) and $a(x, y)$ and $b(x, y)$ represent photometric parameters. Propose a linear algorithm for estimating (u, v, a, b) from the image brightness I and its spatio-temporal derivatives I_x, I_y, I_t. What is the minimum size of a window around each pixel that allows one to solve the problem?

After subtracting $I(x, y, t)$ on both sides, and applying the BCC, we obtain

$$I_x u + I_y v + I_t = (a - 1)I + b, \tag{4}$$

which reduces to the standard BCC when $a = 1$ and $b = 0$. This new BCC can be re-written as

$$I_x u + I_y v + (1 - a)I - b = -I_t \implies \begin{bmatrix} I_x & I_y & I & 1 \end{bmatrix} \begin{bmatrix} u \\ v \\ 1 - a \\ -b \end{bmatrix} = -I_t. \tag{5}$$

From this equation, we can solve for the parameters (u, v, a, b) in a least squares sense by assuming that such parameters are constant on a neighborhood Ω around each pixel. This leads to the following linear system of equations

$$\sum_{\Omega} \begin{bmatrix} I_x^2 & I_x I_y & I_x I & I_x \\ I_x I_y & I_y^2 & I_y I & I_y \\ I_x I & I_y I & I^2 & I \\ I_x & I_y & I & 1 \end{bmatrix} \begin{bmatrix} u \\ v \\ 1 - a \\ -b \end{bmatrix} = -\sum_{\Omega} \begin{bmatrix} I_t I_x \\ I_t I_y \\ I_t I \\ I_t \end{bmatrix}. \tag{6}$$

Since there are four unknowns, we need at least 4 pixels, e.g. a 2×2 window.

2. **(20 Points) Self-calibration and motion estimation for purely rotating cameras.**

   (a) **(10 Points)** Given two *calibrated* perspective images of a scene $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ related by a purely rotational motion $R \in SO(3)$, propose a linear algorithm for computing R from a set of P *noisy* point correspondences $(\boldsymbol{x}_1, \boldsymbol{x}_2)$. In your derivation of the algorithm, please write the linear system you need to solve explicitly, find the number of unknowns you need to solve for, find the minimum number of point correspondences needed, describe the way the linear system is solved using SVD, and derive a method for projecting a noisy matrix $\tilde{R} \in \mathbb{R}^{3\times 3}$ onto $SO(3)$ by minimizing the Frobenius norm of the error $\min_{R\in SO(3)} \|R - \tilde{R}\|_F^2 = trace((R - \tilde{R})(R - \tilde{R})^T)$.
   The relationship between the two camera matrices can be written as $\boldsymbol{X}_2 = R\boldsymbol{X}_1$, hence we have $\lambda_2 \boldsymbol{x}_2 = \lambda_1 R\boldsymbol{x}_1$. After taking the cross product with $\boldsymbol{x}_2$ we obtain

$$\widehat{\boldsymbol{x}_2} R \boldsymbol{x}_1 = 0. \tag{7}$$

   Notice that this equation is entirely analogous to the equation relating two views of a planar scene, $\widehat{\boldsymbol{x}_2} H \boldsymbol{x}_1 = 0$, where $H = R + TN^T/d$ is the homography of the plane $N^T \boldsymbol{X} = d$. In fact, when $T = 0$, H reduces to R.
   In order to solve for R from (7), notice that (7) gives 3 linear equations on the 9 entries of $R \in \mathbb{R}^{3\times 3}$. Out of these 3 equations, only two are linearly independent, because $\text{rank}(\widehat{\boldsymbol{x}_2}) = 2$. If we let $\boldsymbol{x}_2 = (y_1, y_2, 1)^T$, then two independent rows of $\widehat{\boldsymbol{x}_2}$ are $(1, 0, -y_1)$ and $(0, 1, -y_2)$. Therefore, we can write the two equations explicitly as

$$A\boldsymbol{r} = \begin{bmatrix} \vdots & & \\ \boldsymbol{x}_1^T & 0^T & -y_1 \boldsymbol{x}_1^T \\ 0^T & \boldsymbol{x}_1^T & -y_2 \boldsymbol{x}_1^T \\ \vdots & & \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{33} \end{bmatrix}^T = 0 \tag{8}$$

For P correspondences, we have $A \in \mathbb{R}^{2P\times 9}$, therefore the minimum number of correspondences needed to solve for $\boldsymbol{r}$ up to scale can be obtained from $2P \geq 9-1=8$, which gives $P=4$. In the presence of noise, we solve the linear system $A\boldsymbol{r}=0$ in a least squares sense, i.e. we let $\boldsymbol{r}=V_9$, where $A=U\Sigma V^T$ is the SVD of A, and V_9 is the 9th column of V.

Notice however that the so-obtained matrix $\tilde{R} \in \mathbb{R}^{3\times 3}$ is an arbitrary matrix, and not necessarily a rotation, thus we need to find a way of projecting this matrix onto $R \in SO(3)$. To this end, let $\tilde{R}=U\Sigma V^T$ be the SVD of $\tilde{R}$. We seek a matrix R such that minimizes

$$\|R-U\Sigma V^T\|_F^2 = 3-2trace(R^TU\Sigma V^T)+trace(\Sigma^2). \tag{9}$$

This is equivalent to maximizing $trace(R^TU\Sigma V^T)=trace(VR^TU\Sigma)$. Therefore, from Lemma 1 we have $VR^TU=I$, from which it follows that $R=UV^T$.

(b) **(10 Points)** Given two *uncalibrated* perspective images of a scene $(\boldsymbol{x}_1,\boldsymbol{x}_2)$ related by a purely rotational motion $R \in SO(3)$, propose a linear algorithm for computing the camera calibration matrix K and the camera rotation R from a set of P point correspondences. What is the minimum number of correspondences needed? *Hint: Show that one can linearly solve for the homography at infinity $H=KRK^{-1}$. Then show that if we let $S=KK^T$, then $HSH^T=S$.*

In the case of uncalibrated cameras, we replace $\boldsymbol{x}_i$ by $K^{-1}\boldsymbol{x}_i$ for $i=1,2$. Therefore, the equation relating the two views becomes $\lambda_2K^{-1}\boldsymbol{x}_2=\lambda_1RK^{-1}\boldsymbol{x}_1$, i.e. $\lambda_2\boldsymbol{x}_2=\lambda_1KRK^{-1}\boldsymbol{x}_1$. If we let $H=KRK^{-1}$, then we have $\widehat{\boldsymbol{x}_2}H\boldsymbol{x}_1=0$. We can solve for the homography at infinity H up to a scale factor using the same algorithm as in the previous problem, except that we do not need to apply the projection step, because H is not a rotation. In order to obtain the scale of H, notice that $\det(H)=\det(K)\det(R)\det(K^{-1})=1$. That is, once H has been computed, we divide the solution by the cubic root of its determinant, $H/\sqrt[3]{\det(H)}$ so that $\det(H)=1$ from now on.

Now, if we let $S=KK^T$, then we have that $HSH^T=KRK^{-1}KK^TK^{-T}R^TK^T=S$. This gives a set of 6 equations on the 6 unknowns in S, because S is symmetric. We can solve for S up to a scale factor from this linear system. Since $\det(S)=\det(K)^2>0$, we divide S by the sign of its determinant $\text{sign}(\det(S))$, so that $\det(S)>0$ from now on. Given such an S one can factor it as $S=KK^T$ using Choleski decomposition, because K is upper triangular. Notice that the scale of K is not correct, because S was computed up to scale only. So we obtain the correct K by dividing by $K_{3,3}$ so that $K_{3,3}=1$. Once K is known, we get $R=K^{-1}HK$.

3. **(20 points) 3-D Reconstruction from Multiple Calibrated Orthographic Views.** Let $\boldsymbol{x}_{fp} \in \mathbb{R}^2$ be the a *known* measurement for the *orthographic projection* of an *unknown* point $\boldsymbol{X}_p \in \mathbb{R}^3$ in frame $f=1,\ldots,F$, where $p=1,\ldots,P$. That is, $\boldsymbol{x}_{fp}=M_f\boldsymbol{X}_p+V_f$, where

$$\begin{bmatrix} M_f & V_f \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} R_f & T_f \end{bmatrix} \tag{10}$$

is the projection matrix associated with frame f and $(R_f,T_f) \in SE(3)$ is the *unknown* pose of the camera at frame f relative to some fixed world coordinate frame.

(a) **(4 points)** Show that the optimal solution for the 2-D translation $V_f \in \mathbb{R}^2$ in the sense of minimizing the reprojection error is

$$V_f = \bar{\boldsymbol{x}_f} = \frac{1}{P}\sum_{p=1}^{P}\boldsymbol{x}_{fp} \qquad f=1,\ldots,F. \tag{11}$$

Hint: show that one can assume that $\bar{\boldsymbol{X}}=\frac{1}{P}\sum_{p=1}^{P}\boldsymbol{X}_p=0$ without loss of generality.

The optimal solution for $\{V_f\}_{f=1}^F$ is obtained by minimizing the reprojection error

$$\sum_{p=1}^{P}\sum_{f=1}^{F}\|\boldsymbol{x}_{fp}-M_f\boldsymbol{X}_p-V_f\|^2. \tag{12}$$

After setting the partial derivative with respect to V_f to zero we get

$$\sum_{p=1}^{P}2(\boldsymbol{x}_{fp}-M_f\boldsymbol{X}_p-V_f)(-1) = -2P(\bar{\boldsymbol{x}_f}-M_f\bar{\boldsymbol{X}}-V_f)=0, \tag{13}$$

from which it follows that $V_f = \bar{\boldsymbol{x}_f} - M_f \bar{\boldsymbol{X}}$. Now, in solving the reconstruction problem we always have the choice of a reference frame. Typically, one chooses the reference frame to coincide with the coordinate frame of the first camera, *i.e.* $R_1 = I$ and $T_1 = 0$. Alternatively, we may choose the origin of the reference frame to coincide with the center of gravity of the points in 3D space, *i.e.* $\bar{\boldsymbol{X}} = 0$. This has the advantage of allowing us to estimate camera translation V_f *without* knowing camera rotation M_f.

(b) **(8 points)** Let $\boldsymbol{w}_{fp} = \boldsymbol{x}_{fp} - \bar{\boldsymbol{x}_f}$ be the mean subtracted point correspondences and define a data matrix

$$W = \begin{bmatrix} \boldsymbol{w}_{11} & \cdots & \boldsymbol{w}_{1P} \\ \vdots & & \vdots \\ \boldsymbol{w}_{F1} & \cdots & \boldsymbol{w}_{FP} \end{bmatrix} \in \mathbb{R}^{2F \times P}. \tag{14}$$

Show that the measurement matrix W factors as $W = MS$, where $M = \begin{bmatrix} M_1 \\ \vdots \\ M_F \end{bmatrix} \in \mathbb{R}^{2F \times 3}$ and $S = \begin{bmatrix} \boldsymbol{X}_1 & \boldsymbol{X}_2 & \cdots & \boldsymbol{X}_P \end{bmatrix} \in \mathbb{R}^{3 \times P}$ are the so-called *motion* and *structure* matrices, respectively. Show that $\text{rank}(W) \leq 3$ and $\text{rank}(M) \geq 2$ and derive conditions on the camera motion and the 3D structure such that $\text{rank}(W) = 3$. Under such conditions, propose an algorithm for computing the motion and structure matrices $M = \tilde{M}K$ and $S = K^{-1}\tilde{S}$ up to an unknown invertible matrix $K \in \mathbb{R}^{3\times 3}$.

Note that by construction $\boldsymbol{w}_{fp} = M_f S_p$, hence we immediately obtain that $W = MS$. Since M has 3 columns, it is obvious that $\text{rank}(W) \leq 3$. Since the rows of M_f are rows of a rotation matrix, such rows are orthogonal, hence it is obvious that $\text{rank}(M) \geq \text{rank}(M_f) \geq 2$. Furthermore, $\text{rank}(M) = 2$ if and only if the rows of M_f span the same subspace as the rows of M_1 for all $f = 2, \ldots, F$. Since the rows of M_f are rows of a rotation, notice that this only happens when all rotation matrices $\{R_f\}$ have a common third row. Also, notice $\text{rank}(S) = 1$ if the 3D structure is a line, $\text{rank}(S) = 2$ if the 3D structure is a plane, and $\text{rank}(S) = 3$ for a general 3D structure. Therefore, in order to have $\text{rank}(W) = 3$ we need the camera rotations to have different third rows, and the 3D structure not to be contained on a plane. Under such conditions, we can compute the SVD of $W = U\Sigma V^T$, and then set $\tilde{M} = U_{1:3}$ and $\tilde{S} = \Sigma V^T_{1:3}$. Notice that this factorization process does not give M and S exactly, because for any invertible 3×3 matrix K we have $W = MS = \tilde{M}KK^{-1}\tilde{S}$.

(c) **(8 points)** Let $Q = KK^T \in \mathbb{R}^{3\times 3}$. Show that the sub-matrix of $\tilde{M}$ consisting of rows $2f-1$ and $2f$, $\tilde{M}_{2f-1:2f} \in \mathbb{R}^{2\times 3}$, is such that

$$\tilde{M}_{2f-1:2f} Q \tilde{M}^T_{2f-1:2f} = I \qquad f = 1, \ldots, F. \tag{15}$$

Propose a linear algorithm to compute Q. What is the minimum number of frames needed? Given Q, show how to compute K up to a rotation. Given such a K show how to compute M, S, R_f and T_f. Is there any ambiguity in the reconstruction?

Recall that $M = \tilde{M}K$, thus $M_f = \tilde{M}_{2f-1:2f}K$. Since the two rows of M_f are rows of a rotation matrix, we have $M_f M_f^T = \tilde{M}_{2f-1:2f} Q \tilde{M}^T_{2f-1:2f} = I$. This gives 3 linearly independent equations per frame on the 6 unknowns in Q. From the equation $3F \geq 6$, the minimum number of frames needed to compute Q is $F = 2$. Given $Q = KK^T$, one can solve for $K = \tilde{K}\mathcal{R}$ using the QR decomposition up to an unknown rotation matrix $\mathcal{R}$. From the equation $M = \tilde{M}K$, note that $\mathcal{R}$ rotates all matrices M_f equally. Therefore, we can choose $\mathcal{R}$ arbitrarily, as it corresponds to the choice of the rotational part of the reference frame. Thus one way of fix $\mathcal{R}$ is by choosing $R_1 = I$. Given K, the motion and structure parameters are given by $M = \tilde{M}K$ and $S = K^{-1}\tilde{S}$. From M_f we immediately know the first two rows of each R_f, thus we can obtain the third row from the cross product of the first two. Finally, we already know the first two entries of T_f from V_f, which was computed in part a). Note that the third entry of each T_f does not show up in the equations, and so we can not recover it.