Handouts: Geometry

Ronen Basri

Introduction to Computer and Human Vision

1 Projection

_Projection_ is the process of forming a 2D image from a 3D scene. _Perspective projection_ is formed by rays reflected from the objects through the _optical center_ (or _focal center_) onto a capturing device (CCD or film). To avoid reversal of the image we usually assume that the capturing device lies in front of the optical center. We form a coordinate system by setting the origin \( O = (0, 0, 0) \) at the optical center, the X- and Y-axes parallel to the image plane, and the Z-axis along the _optical axis_. A scene point \( P = (X, Y, Z) \) then projects to an image point \( p = (x, y, f) \) according to the following formulae:

\[
p = \frac{f}{Z} P,
\]

where \( f \) denotes the _focal length_. More explicitly,

\[
x = \frac{fX}{Z}, \quad y = \frac{fY}{Z}.
\]

These formulae form the _pinhole camera model_. Denoting the image points as 3-vectors is referred to as _homogeneous coordinates_.

_Optographic projection_ is formed by parallel rays that travel from the object onto the capturing device. This approximation is useful when objects are relatively far from the camera. Often we assume in addition that the objects are scaled uniformly. (This model is called sometimes _weak perspective_.) In this case a scene point \( P = (X, Y, Z) \) projects to an image point \( p = (x, y, f) \) with

\[
x = sX, \quad y = sY,
\]

where \( s \) is typically set to \( s = f/Z_0 \) and \( Z_0 \) is the average distance to the object.

2 Projective Geometry

Consider a perspective projection of points that lie on one plane onto another plane (e.g. the CCD). Such a projection defines a one-to-one mapping. A finite sequence of such mappings is called a _projective transformation_ or _homography_. Homography is a mapping \( H : \mathbb{P}^2 \rightarrow \mathbb{P}^2 \) and can be represented by an invertible, \( 3 \times 3 \) matrix. \( \mathbb{P}^2 \), the _projective plane_, can be identified with the set of central rays (rays
though the origin) in $\mathbb{R}^3$. Consequently, a point in $\mathbb{P}^2$ can be written as $\alpha p$ with arbitrary scale factor $\alpha \neq 0$. If we identify $\mathbb{P}^2$ with the plane $Z = f$ then the rays parallel to the image plane form ideal points and their third coordinate vanishes. The collection of ideal points form the line at infinity.

A line in $\mathbb{P}^2$ is parameterized by a 3-vector $l$ satisfying $l^T p = 0$. The point of intersection between two lines $l$ and $l'$ is given by $p = l \times l'$, where `$\times$' represents the cross product between vectors. Similarly, given two points, $p$ and $p'$, the line through $p$ and $p'$ is given by $l = p \times p'$. A conic in $\mathbb{P}^2$ is parameterized by a $3 \times 3$ matrix $C$ and is expressed by a quadratic form $p^T C p = 0$. Homography maps a line to a line and a conic to a conic. Specifically, suppose $q = H p$. A line parameterized by $l$ (satisfying $l^T p = 0$) is mapped by $H$ to $H^{-T} l$ (since $l^T H^{-1} q = l^T p = 0$). Similarly, a conic parameterized by $C$ (satisfying $p^T C p = 0$) is mapped by $H$ to $H^{-T} C H^{-1}$ (since $q^T H^{-T} C H^{-1} q = p^T C p = 0$).

### 2.1 Hierarchy of transformations

Homography is represented by a general, invertible $3 \times 3$ matrix, denoted $H$. By restricting the components of $H$ we can obtain other familiar geometric transformations. For example,

$$H = \begin{pmatrix}
a_{11} & a_{12} & t_x \\
a_{21} & a_{22} & t_y \\
0 & 0 & 1
\end{pmatrix}$$

represents an affine transformation.

$$H = \begin{pmatrix}
a & -b & t_x \\
b & a & t_y \\
0 & 0 & 1
\end{pmatrix}$$

represents similarity, and

$$H = \begin{pmatrix}
cos \theta & -sin \theta & t_x \\
sin \theta & cos \theta & t_y \\
0 & 0 & 1
\end{pmatrix}$$

represents a rigid (Euclidean) transformation. Rigid transformations preserve distances and angles. Similarity transformations scale distances and preserve angles. Affine transformations maintain parallelism.

### 2.2 Two views related by homography

Two images of a rigid scene are related by homography in the following cases:


2. Planar scene: $P' = RP + t$ and $P$ satisfies $n^T P = d$. We can combine those to obtain

$$P' = \left( R + \frac{1}{d} ln^T \right) P$$
Given correspondences between points in two images related by homography $H$ we can determine the 9 components of $H$ (up to a scale factor) by using at least four pairs of corresponding points. Each pair supplies two equations.

3 Two-view geometry

3.1 Epipolar relations: the Essential Matrix

Given a stereo pair of images $I$ and $I'$, let $O$ and $O'$ denote the respective camera centers. Consider a point $P \in \mathbb{R}^3$. Clearly, the vectors connecting the three points, $O$ and $O'$ and $P$ lie on a plane, and so their triple product vanishes, i.e.,

$$OP \cdot (OO' \times O'P') = 0.$$

Let $p$ and $p'$ denote the projections of $P$ in $I$ and $I'$, then also

$$Op \cdot (OO' \times O'p') = 0.$$

Since $Op = p$, $OO' = t$ and $O'p' = Rp'$, where $R$ and $t$ denote the relative rotation and translation between the two cameras, then

$$p^T(t \times (Rp')) = 0.$$

Denote

$$E = [t] \times R,$$

where $[t] \times$ is the skew symmetric matrix expressing the cross product of $t$ with an argument vector, then we obtain

$$p^T E p' = 0.$$

$E = [t] \times R$ is called the Essential Matrix, and this bilinear form represents the epipolar line relation between $I$ and $I'$. In addition, $E$ is rank 2, and its (left and right) null spaces specify the location of the epipoles.

$E$ is $3 \times 3$. Given point correspondences it can be determined up to a scale factor. Since every pair of corresponding points supplies one equation (epipolar relation) then $E$ can be recovered using 8 points by a linear algorithm. However, since $E$ is determined by a 6 degree-of-freedom rigid transformation in 3D space, and since translation can be determined only up to a scalar factor, $E$ can be recovered by a non-linear algorithm with 5 points only.

3.2 Internal calibration: the Fundamental Matrix

Recovering the depth of a scene, or the relative position and orientation of cameras in different images, requires calibration of the cameras. This requires knowledge of the following parameters: the camera center $(c_x, c_y)$, the focal length $f$, the pixel size $(a_x, a_y)$ and the skew $s$. These are usually placed in a calibration matrix $K$ of the form:

$$K = \begin{pmatrix}
    fa_x & b & c_x \\
    0 & fa_y & c_y \\
    0 & 0 & f
\end{pmatrix}.$$
A pixel $p$ in a calibrated image appears at $q = Kp$ in its uncalibrated version. We can therefore write:

$$p^T E p' = (K^{-1}q)^T E (K^{-1}q') = q^T (K^{-T}EK^{-1})q.$$  

Consequently, we define the **Fundamental Matrix** to be

$$F = K^{-T}EK^{-1}.$$  

As with the Essential matrix, the bilinear relation $p^T F p' = 0$ defines an epipolar relation, $F$ is of rank 2, and its left and right null spaces define the epipoles.

Since the epipolar relation is homogeneous $F$ can be recovered using a linear algorithm with 8 pairs of corresponding points. An additional constraint is that $\text{det}(F) = 0$ (since it is of rank 2). This constraint can be used to reduce the required number of pairs to 7. The calibration matrix $K$ can be recovered by using a calibration pattern (a collection of points whose position in 3D is known) or by using auto-calibration methods.

<table>
<thead>
<tr>
<th></th>
<th>Homography</th>
<th>Perspective</th>
<th>Orthographic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relation</td>
<td>$q = \alpha H p$</td>
<td>$p^T E q = 0$</td>
<td>$p^T E q = 0$</td>
</tr>
<tr>
<td>DOF</td>
<td>Group transformation</td>
<td>Epipolar lines</td>
<td>Epipolar lines (parallel)</td>
</tr>
<tr>
<td>Eqs/pnt</td>
<td>8</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>Minimal config</td>
<td>4 points</td>
<td>5-8 points</td>
<td>4 points</td>
</tr>
<tr>
<td>Depth recovery</td>
<td>Not possible</td>
<td>Possible</td>
<td>A third image is required</td>
</tr>
</tbody>
</table>

## 4 Stereo

Assuming $I_l$ and $I_r$ constitute a **rectified** stereo pair of images, so that their epipolar lines are horizontal and of identical heights ($y$-coordinate). We set the origin $O = (0, 0, 0)$ to lie between the two centers of projection $O_l = (-b/2, 0, 0)$ and $O_r = (b/2, 0, 0)$ ($b$ denotes the **baseline distance**). A scene point $P = (X, Y, Z)$ is then projected to $(x_l, y_l)$ and $(x_r, y_r)$, where

$$x_l = \frac{f(X + b/2)}{Z} \quad y_l = \frac{fY}{Z},$$  

$$x_r = \frac{f(X - b/2)}{Z} \quad y_r = \frac{fY}{Z}.$$  

Consequently,

$$Z = \frac{fb}{x_l - x_r}.$$  

The difference $x_l - x_r$ is called **disparity**, and it is inversely related to depth. Since depth is directly proportional to $b$ reconstruction is generally more robust when $b$ is larger. Unfortunately, in that case the fields of views of the cameras are typically more disjoint.
To recover depth we need to solve a correspondence problem, for every scene point we need to identify both of its projections onto the two images. To that end we define a cost function:

\[
F = \sum_{p \in I_1} D_p(d_p) + \sum_{p,q \in N} V_{pq}(d_p, d_q),
\]

where the unary data term encourages similarity between the intensities of corresponding points, and the binary smoothness term encourages similar disparities for neighboring points. In class we discussed two algorithms to minimize such a cost function based on dynamic programming and iterated graph cuts (Ex. 3).

5 Multi-view geometry: SVD-factorization

Suppose we track the location of \( p \) points in \( f \) video frames (images). Assuming orthographic projection, the location of point \( j \) in frame \( i \) is given by:

\[
\begin{align*}
    x_{ij} &= r_{i1}X_j + r_{i2}Y_j + r_{i3}Z_j + t_{ix} \\
    y_{ij} &= s_{i1}X_j + s_{i2}Y_j + s_{i3}Z_j + t_{iy},
\end{align*}
\]

where \( \vec{r}_i = (r_{i1}, r_{i2}, r_{i3}) \) and \( \vec{s}_i = (s_{i1}, s_{i2}, s_{i3}) \) denote the upper two rows of the \( i \)th rotation matrix, \( (t_{ix}, t_{iy}) \) denote the two components of the translation vector, and \( (X_j, Y_j, Z_j) \) denote the 3D location of the \( j \)th point. Below we assume zero translation – the translation can be discarded by placing the centroid of the \( p \) points to be the origin in every image. Writing these equations in matrix forms yields:

\[
M = TS,
\]

where the \( 2f \times p \) measurement matrix \( M \) is defined by

\[
\begin{pmatrix}
    x_{11}, \ldots, x_{1p} \\
    \vdots \\
    x_{f1}, \ldots, x_{fp} \\
    y_{11}, \ldots, y_{1p} \\
    \vdots \\
    y_{f1}, \ldots, y_{fp}
\end{pmatrix},
\]

The \( 2f \times 3 \) transformation matrix \( T \) is defined by

\[
\begin{pmatrix}
    \vec{r}_1 \\
    \vec{r}_f \\
    \vec{s}_1 \\
    \vec{s}_f
\end{pmatrix},
\]
and the $3 \times p$ shape matrix $S$ is defined by

$$
\begin{pmatrix}
X_1, ..., X_p \\
Y_1, ..., Y_p \\
Z_1, ..., Z_p
\end{pmatrix}.
$$

Our objective is given $M$ to recover both $T$ and $S$.

Note that in the absence of noise $M$ should have rank 3. However, when noise is present we can eliminate noise by forcing $M$ to have rank 3 using SVD:

$$
M = U \Delta V^T,
$$

where $U$ and $V$ are orthonormal matrices and $\Delta$ is a diagonal matrix containing the non-negative singular values of $M$. Denote by $\Delta_3$ the diagonal matrix $\Delta$ in which all but the three largest singular values are replaced by zeros. Then $M_3 = U \Delta_3 V^T$ is the best rank 3 approximation to $M$ in a least squares sense. Next we define

$$
T = U \sqrt{\Delta_3}
$$

and

$$
S = \sqrt{\Delta_3} V^T.
$$

Now $M = TS$, but this solution is not unique since for any non-singular $3 \times 3$ matrix $A$ $M = (TA)(A^{-1}S)$ also forms a factorization solution. To resolve this ambiguity we enforce orthonormality of rows in $T$. Recall that $\vec{r}_i$ and $\vec{s}_i$ denote two rows of a rotation matrix, then

$$
||\vec{r}_i||^2 = ||\vec{s}_i||^2 = 1
$$

and

$$
\vec{r}_i \vec{s}_i^T = 0.
$$

Let $T_i$ denote the $i$'th row of $T$, then we expect $\vec{r}_i = T_i A$ and $\vec{s}_i = T_{i+f} A$. Plugging these into the orthonormality constraints we get

$$
T_i A A^T T_i^T = T_{i+f} A A^T T_{i+f}^T = 1
$$

and

$$
T_i A A^T T_{i+f}^T = 0.
$$

These quadratic constraints in $A$ are linear in $A A^T$ (which is symmetric $3 \times 3$ and so has 6 degrees of freedom). 3 such linear equations are obtained for every frame, providing an over-constrained linear system. Solving this system we can find $A A^T$, which then can be factored to obtain $A$. Once $A$ is found we declare the solution to include the transformation matrix $T A$ and the shape matrix $A^{-1} S$. This solution is unique up to a global rotation (which has no effect on the recovered shape).