H.I. Bozma

Electric Electronic Engineering
Bogazici University

May 25, 2017
Binocular Stereo

- Introduction
- Arbitrary Orientation
- Determining Depth
- Correspondence Problem
- Epipolar Geometry
- Matchable Features
- 3D Reconstruction
- Essential Matrix
- Fundamental Matrix
Biological Systems

- The world is projected differently onto our two eyes.
- Precisely due to this difference → Relative distances of objects.
- Close objects – more widely separated on our retinas while the reverse for far objects.
- 10% of people do not have true stereo vision, however they are still able to determine depth which is based on motion cues.
Stereo Image Pairs

**Figure:** Stereo image pairs.
Projection of close and far objects on the retina

Figure: Projection of close and far objects on the retina.
Random Dot Stereograms

- In 1960 – Bela Julesz
- These consist of two images of random dots where the second has a portion of the first shifted relative to its original position as seen in Fig. 5.
- Thus, this space thus created is filled in with more random dots.
- It only makes sense to move this portion horizontally, as our eyes are horizontally displaced.
Random Dot Stereograms – II

When viewed appropriately – the views are merged:

▶ A perception of depth which depends on the amount of shift between the central portions.
▶ Depending on whether the central portion is shifted to the left or the right, the depth perception is behind or in front of the border.
Random dot stereograms – III

Figure: Random dot stereograms [?].
Vergence Angle

- Cameras have arbitrary position and orientation with respect to each other.
- The cameras' optical axes may or may not intersect at a point in space.
- **Vergence angle**: The difference in orientations of the two optical axes
Binocular Vision

Given two corresponding points $x_l$ and $x_r$, it is possible to compute the 3D coordinates of the underlying point.

Consider two cameras, parallel and separated horizontally by a baseline distance $b$, and having the same focal length

- A conjugate pair: Two points in different images that are the projections of the same point in the scene.
- Disparity – The distance between points of a conjugate pair when the two images are superimposed.
Triangulation

The 3D position of a point must lie on the straight line through the centre of projection and the image point.
Stereo Geometry

\[ (x_1, x_2, x_3) \]

\[ X_1, X_2, X_3 \]

\[ f \]

\[ b \]
Binocular Vision Equations

\[ \frac{X_1}{X_3} = \frac{x_{l1}}{f} \frac{X_1 - b}{X_3} = \frac{x_{r1}}{f} \]

\[ \rightarrow \text{Depth} \quad X_3 = \frac{bf}{x_{l1} - x_{r1}} \]

\[ X_1 = \frac{x_{l1}X_3}{f} \quad X_2 = \frac{x_{l2}X_3}{f} \]
Julesz also found that stereo is done at a very low level –

⇒ Don’t ”‘interpret” the scene before perceiving depth.

The correspondence problem – namely matching the dots

Given two images formed in two image planes $X_L$ and $X_R$,

For a point $x_l$ in $X_L$, which point $x_r$ in $X_R$ corresponds to it?
Triangulation refers to the determination of the intersection of two such lines generated from two independent images.
Stereo Image Pairs

Figure: Stereo image pairs.
Correspondence Problem II

**Correspondence problem** – The problem of matching a point in one image to another point in the other image belonging to the same object.

- Pair up the points in the right and left images – that is for each point in the left image, what is the corresponding point in the right image?
The corresponding image points that have zero disparity values – Are projected by 3D points at a finite distance from the cameras.

If the difference in position and orientation of the stereo views is small, matching corresponding points is easy.

Matching is difficult if the difference is large.

Conversely, the accuracy of the 3D reconstruction is poor if the difference in position and orientation of stereo views is large.
Points are expressed either in Euclidean or projective coordinates.

- $X \in \mathbb{R}^3 \rightarrow x \in \mathbb{R}^2$.
- $x = [x_1 \ x_2]^T$ or $x = [x_1 \ x_2 \ 1]$.
- Note that $[x_1 \ x_2 \ 1]^T = [sx_1 \ sx_2 \ s]^T$, for any non-zero scalar $s$ represent the same image coordinates.
The correspondence does not require a search through the whole image – rather a single line.

- A feature in one image lies anywhere along the viewing ray.
- Project this viewing ray into the other image.
- A line (an epipolar line) in the second image on which the feature we are trying to match must lie.
- A point \( x \) in one image generates a line in the other on which its corresponding point \( x' \) must lie.
- Hence, the epipolar line is the straight line of intersection of the epipolar plane with the image plane. It is the image in one camera of a ray through the optical center and image point in the other camera.
Epipolar Line
Two Cameras: Epipolar Line
**Epipole** – The point of intersection of the line joining the optical centers, that is the baseline, with the image plane.

- Thus the epipole is the image, in one camera, of the optical centre of the other camera.
- All epipolar lines intersect at the epipole.
Epipole

Figure: Epipole.
Epipolar Plane

Epipolar plane – The plane defined by a 3D point X and the optical centers C and C’.

➤ Note that different 3D points give rise to different epipolar planes.

➤ All epipolar planes intersect at the baseline of the binocular system.
The vergence angle is $0^\circ$.
Features to Match

If the point to be matched is clearly different from its surrounding pixels, this is a simpler task. → Find matchable features.

- Edges
- Regions
- Features
Edge Matching

- Depth values at the edge points → A sparse depth map.
- Standard algorithm → Implements a multiscale approach, and assumes a parallel geometry, that is, it assumes that the epipolar lines are the rows of the image.
- Meaningless information along occlusions
Standard Algorithm

1. Filter each image with Gaussian filters at four filter widths such the each filter is twice as wide as the next via repeated convolution with the smallest filter.

2. Compute edge positions on each row.

3. Match edges in corresponding rows at the coarse resolutions via the comparison of orientations and strengths. Note that horizontal edges can’t be matched.

4. Improve the disparity estimates by matching at finer scales.
An alternative approach is based on region matching.

Matching can be based on either intensity or some other region features.

In intensity based matching, one approach to finding interesting regions is to find regions of high variance.
Consider a point $x$ and define $N_\epsilon(x) = \{y \mid \|y - x\| < \epsilon\}$

\[
F_1(x) = \sum_{y \in N_\epsilon(x)} \|F(y) - F(y')\| \quad y' = x + \begin{bmatrix} 0 \\ 1 \end{bmatrix}
\]

\[
F_2(x) = \sum_{y \in N_\epsilon(x)} \|F(y) - F(y'')\| \quad y'' = x + \begin{bmatrix} 1 \\ 0 \end{bmatrix}
\]

\[
F_3(x) = \sum_{y \in N_\epsilon(x)} \|F(y) - F(\hat{y})\| \quad \hat{y} = x + \begin{bmatrix} 1 \\ 1 \end{bmatrix}
\]

\[
F_4(x) = \sum_{y \in N_\epsilon(x)} \|F(y) - F(\hat{y})\| \quad \hat{y} = x + \begin{bmatrix} 1 \\ -1 \end{bmatrix}
\]
Region Optimization

\[ F(x) = \min_i F_i(x) \]
Region-Based Algorithm

The algorithm is as follows:
Consider \( x \in X_L \)
If \( F(x) > \tau \), \( N_\epsilon(x) \) is a region of interest
For each disparity \( d \), compute correlation as

\[
 r(d; x) = \sum_{y \in N_\epsilon(x)} (I_L(x) - \bar{I}_L(x))(I_R(x + d) - \bar{I}_R(x + d))
\]

Note \( r(d; x) \in [-1, 1] \)
Find \( d(x) = \arg \max_d r(d; x) \) -- If this value is close to 1, \( d(x) \) is the corresponding disparity
Matching Features

- Similarity – Black dots only match black dots.
- Uniqueness – One black dot can match no more than one black dot.
- Continuity – Disparity values vary smoothly almost everywhere.
- Ordering – If \( x \) is to the left \( y \) in the left image, then the corresponding points \( x' \) and \( y' \) are also located similarly in the right image.
Camera Setup

- For a verging angle = 0 camera set-up, corresponding image points that have zero disparity values are projected by 3D points at infinite distance have 0 disparity.
- For a verging camera set-up, the corresponding image points that have zero disparity values are projected by 3D points at a finite distance from the cameras.
Disparity Zones

**Figure:** Disparity zones in verging camera setup.

H.I. Bozma
Matching Complexity

- If the difference in position and orientation of the stereo views is small, matching corresponding points is easy.
- Matching is difficult if the difference is large.

The accuracy of the 3D reconstruction is poor if the difference in position and orientation of stereo views is small.
Deriving 3D Structure

▶ Consider a stereo camera system that is capable of establishing some correspondence between the left and right images.
▶ Each correspondence \((m_l, m_r)\) ⇒ Two image points \(m_l\) and \(m_r\) are very likely to be the images of the same world point \(M\).
▶ Two approaches to stereo-based 3D structure
  ▶ **Calibrated Approach** (Classical method) – Calibrate both cameras (or viewpoints) with respect to some world coordinate system, and from this compute the three-dimensional Euclidean structure of the imaged scene.
  ▶ **Uncalibrated Approach** – Fundamental matrix is calculated from image correspondences, and this is then used to determine the projective 3D structure of the imaged scene.
Calibration

**Uncalibrated system** – Intrinsic and extrinsic parameters not known!

**Self-calibration**: The process of calculating all the intrinsic parameters of the camera using only the information available in the images

- No calibration frame or known object is needed: the only requirement is that there is *at least one static object* in the scene,
- The camera moves around taking images,
- The actual camera movement itself does not need to be known,
- Hence, self-calibration is ideal for a mobile camera such as a...
Two view-centered coordinate systems
Notation

- $o_1, o_2$ - The center of the camera image plane,
- $K$ denotes the camera matrix containing all the intrinsic parameters of the camera,
- $I$ denotes the $3 \times 3$ identity matrix.
A scene point $X$ is projected onto the image plane:

\[
\begin{bmatrix}
  s x_1 \\
  s x_2 \\
  s
\end{bmatrix} =
\begin{bmatrix}
  f & 0 & o_1 & 0 \\
  0 & f & o_2 & 0 \\
  0 & 0 & 1 & 0
\end{bmatrix}
\begin{bmatrix}
  X_1 \\
  X_2 \\
  X_3 \\
  1
\end{bmatrix} =
\begin{bmatrix}
  f & 0 & o_1 \\
  0 & f & o_2 \\
  0 & 0 & 1
\end{bmatrix}
\begin{bmatrix}
  I \\
  0
\end{bmatrix}
\begin{bmatrix}
  X_1 \\
  X_2 \\
  X_3 \\
  1
\end{bmatrix} =
K
\begin{bmatrix}
  l & 0
\end{bmatrix}
\begin{bmatrix}
  X_1 \\
  X_2 \\
  X_3 \\
  1
\end{bmatrix}
\]
Thus, relative to the coordinate system fixed at $C$, the image point $x$ of $X$ has projective coordinates

$$\tilde{x} = \begin{bmatrix} \frac{x_1 - o_1}{f} \\ \frac{x_2 - o_2}{f} \\ 1 \end{bmatrix} = \begin{bmatrix} x_1 - o_1 \\ x_2 - o_2 \\ f \end{bmatrix}$$
Now, consider the right camera with coordinate system centered at $C'$.

\[
\tilde{x}' = \begin{bmatrix}
\frac{x'_1 - o'_1}{f'} \\
\frac{x'_2 - o'_2}{f'} \\
1
\end{bmatrix} = \begin{bmatrix}
x'_1 - o'_1 \\
x'_2 - o'_2 \\
f'
\end{bmatrix}
\]
Suppose that the global coordinate system is fixed at $C'$. Then

\[
\begin{bmatrix}
  x'_1 - o'_1 \\
  x'_2 - o'_2 \\
  f'
\end{bmatrix}
= R
\begin{bmatrix}
  x_1 - o_1 \\
  x_2 - o_2 \\
  f
\end{bmatrix}
+ t
\]

Equivalently

\[
\tilde{x}' = R\tilde{x} + t
\]
A line going through two points, \( m \) and \( n \) - represented by the cross product \( m \times n \) - namely \( \alpha_1 m + \alpha_2 n \).

If \( x \) is a point on this line, then \( \det(x, m, n) = 0 \).

Equivalent to saying that \( x^T (m \times n) = 0 \).

Note that

\[
m \times n = \begin{bmatrix}
0 & -m_3 & m_2 \\
m_3 & 0 & -m_1 \\
-m_2 & m_1 & 0
\end{bmatrix} n
\]
Essential Matrix $E$

\[ \tilde{x}' = R\tilde{x} + t \]

Using the result

\[ \tilde{x}'^T (R\tilde{x} \times t) = 0 \]
\[ \tilde{x}'^T (t \times R\tilde{x}) = 0 \]
\[ \tilde{x}'^T TR\tilde{x} = 0 \]
\[ \tilde{x}'^T E\tilde{x} = 0 \]

$T$ - Skew-symmetric matrix formed by $t$. 
Essential Matrix

- Cameras are partially calibrated — namely no knowledge of $R$ and $t$, but only image coordinates in the image plane wrt to center of projection. Hence $o_1, o_2, f$ are known.

- Essential matrix $E$ can be estimated from a small number of corresponding points.

- In $\tilde{x}'^T E \tilde{x} = 0$, image coordinates include offsets and normalization by focal lengths.
A $3 \times 3$ matrix with only 5 degrees of freedom. (In fact six parameters, three from rotation and three from translation, but only two of translation is recoverable without additional information!

To estimate it using corresponding image points, the intrinsic parameters of both cameras must be known.
Fundamental Matrix

- The fundamental matrix - A mathematical construct that encodes the geometric information that relates two different viewpoints of the same scene. The two viewpoints:
  - A pair of stereo images,
  - A temporal pair of images taken at different times with the camera moving between image acquisitions.
Conjugate Points

\[
\begin{bmatrix}
    x_1 - o_1 \\
    x_2 - o_2 \\
    f
\end{bmatrix}
\text{ and }
\begin{bmatrix}
    x'_1 - o'_1 \\
    x'_2 - o'_2 \\
    f'
\end{bmatrix}
\]

\[
\begin{bmatrix}
    x'_1 - o'_1 \\
    x'_2 - o'_2 \\
    f'
\end{bmatrix} =
\begin{bmatrix}
    1 & 0 & -o'_1 \\
    0 & 1 & -o'_2 \\
    0 & 0 & f'
\end{bmatrix}
\begin{bmatrix}
    x'_1 \\
    x'_2 \\
    1
\end{bmatrix}
\]
Conjugate Points – Homogeneous Coordinates

Let \( x' = \begin{bmatrix} x'_1 & x'_2 & 1 \end{bmatrix}^T \).

\[
\begin{bmatrix}
    x'_1 - o'_1 \\
    x'_2 - o'_2 \\
    f'
\end{bmatrix} \approx \frac{1}{f'} \begin{bmatrix}
    1 & 0 & -o'_1 \\
    0 & 1 & -o'_2 \\
    0 & 0 & f'
\end{bmatrix} \begin{bmatrix}
    x'_1 \\
    x'_2 \\
    1
\end{bmatrix}
\]

\[
\Rightarrow \tilde{x}' \approx \begin{bmatrix}
    \frac{1}{f'} & 0 & -\frac{o'_1}{f'} \\
    0 & 1 & -\frac{o'_2}{f'} \\
    0 & 0 & 1
\end{bmatrix} x'
\]

\[
\Rightarrow \tilde{x}' \approx K'^{-1} x'
\]
\[ \tilde{x} \approx K^{-1}x \]

Using results from Essential Matrix calculations,

\[
(\tilde{x}')^T TR\tilde{x} = (K^{-1}x')^T TRK^{-1}x = (x')^T (K'^{-1})^T TRK^{-1} x = 0
\]

\[
(x')^T Fx = 0
\]
Fundamental Matrix – F

- \((x')^T Fx = 0\) – Epipolar constraint
- Note \(F \approx (K^{-1})^T EK^{-1}\)
- In \((x')^T Fx = 0\), image coordinates are relative to any arbitrary local image coordinate system!
- The fundamental matrix F has 7 degrees of freedom. There are 9 matrix elements, but only their ratio is significant, which leaves 8 degrees of freedom. In addition, the constraint that \(\det(F) = 0\) leaves only 7.
- F maps image points to their corresponding epipolar lines, that is, \(Fx = l'\), since \((x')^T l' = (x')^T Fm = 0\). Similarly, \(F^T x' = l\).
- Geometrically, F maps epipoles to the origin of the corresponding image plane.
1. $E$ encapsulates only the extrinsic parameters while $F$ encodes both the intrinsic and the extrinsic parameters of the camera.

2. Both $F$ and $E$ are rank-2 matrices as $\text{rank}(T) = 2$. Thus, $\det(F) = 0$ and $\det(E) = 0$. 
Calculating the Fundamental Matrix

\[ a^T f = 0 \] (1)

where

\[
a = \begin{bmatrix} x_1 x'_1 & x_2 x'_1 & x'_1 & x_1 x'_2 & x_2 x'_2 & x'_2 & x_1 & x_2 & 1 \end{bmatrix}^T
\]

\[
f = \begin{bmatrix} f_{11} & f_{12} & f_{13} & f_{21} & f_{22} & f_{23} & f_{31} & f_{32} & f_{33} \end{bmatrix}^T
\]

If we have at least 8 matches, then we should be able to solve for the \( f \) vector up to a scale factor by solving for the null vector of the data matrix.