Stereo Wrap Up
and
Structure from Motion

Introduction to Computer Vision
CSE 152
Lecture 8

Announcements

• HW1 due tonight
Stereo Vision Outline

- Offline:
  - Calibrate cameras & determine epipolar geometry

- Online
  1. Acquire stereo images
  2. Rectify images to convenient epipolar geometry
  3. Establish correspondence
  4. **Estimate depth**

What if stereo geometry isn’t convenient?
Rectification: Given a pair of images, transform both images so that epipolar lines are image rows.
Under perspective projection, the mapping from a plane to a plane is given by a linear transformation of homogeneous coordinates (called a projective transformation or homography).

Question: Where are the epipoles in the rectified image?
What happens to image during rectification?

- Image mapped to a quadrilateral
- Image may get stretched.
  - Pronounced when epipole is near edge of image

Is rectification always possible?

How would these images look like if they’re rectified?
Two approaches to finding correspondence

1. Feature-Based (Sparse)
   - From each image, process “monocular” image to obtain image features (e.g., corners, lines).
   - Establish correspondence between features, using appearance info

2. Area-Based (Dense)
   - Directly compare image regions between the two images.

Random Dot Stereograms
Random Dot Stereograms

Finding Correspondences

$W(p_l)$

$W(p_r)$
Simple matching methods

- SSD (Sum of Squared Differences)
  \[ \sum_{x,y} |W_1(x,y) - W_2(x,y)|^2 \]

- NCC (Normalized Cross Correlation)
  \[ \frac{\sum_{x,y} (W_1(x,y) - \bar{W}_1)(W_2(x,y) - \bar{W}_2)}{\sigma_{W_1}\sigma_{W_2}} \]
  where \( \bar{W}_i = \frac{1}{n} \sum_{x,y} W_i \), \( \sigma_{W_i} = \sqrt{\frac{1}{n} \sum_{x,y} (W_i - \bar{W}_i)^2} \)

- What advantages might NCC have over SSD?

Some Issues

- Ambiguity
- Epipolar ordering
- Window size
- Window shape
- Lighting
- Half occluded regions
Some Issues

- Ambiguity
- **Epipolar ordering**
- Window size
- Window shape
- Lighting
- Half occluded regions
A challenge: Multiple Interpretations

Each feature on left epipolar line matches one and only one feature on right epipolar line.

Some Issues

- Ambiguity
- Epipolar ordering
- **Window size**
- Window shape
- Lighting
- Half occluded regions
Window size

\[ \text{Better results with adaptive window} \]


Some Issues

- Ambiguity
- Epipolar ordering
- Window size
- **Window shape**
- Lighting
- Half occluded regions
Window Shape and Forshortening

- Even though the red window is centered at the image of same scene point, the line is pointing in a different direction.
- Occurs when angle between surface normal and optical axis is large. Window-based stereo matching may fail in this case.

Some Issues

- Ambiguity
- Epipolar ordering
- Window size
- Window shape
- **Lighting**
- Half occluded regions
Lighting Conditions (Photometric Variations)

Some Issues

- Ambiguity
- Epipolar ordering
- Window size
- Window shape
- Lighting
- Half occluded regions
Half occluded regions

Stereo matching as an optimization problem

Similarity measure (SSD or NCC)

Optimal path (dynamic programming)

Constraints
- epipolar
- ordering
- uniqueness
- disparity limit
- disparity gradient limit

Trade-off
- Matching cost (data)
- Discontinuities (prior)

(Cox et al. CVGIP’96; Koch’96; Falkenhagen’97; Van Meerbergen, Vergauwen, Pollefeys, VanGool IJCV’02)
Variations on Binocular Stereo

1. Uncalibrated Stereo
2. Trinocular Stereopsis
3. Multiview stereo
4. Helmholtz Reciprocity Stereopsis

For the measured pixel coordinates $q$ in Camera 1 with intrinsic parameters $K_1$, the relationship between the calibrated image plane location and pixel coordinate is

$$1p = (K_1^{-1}) q$$

Likewise, for Camera 2 with intrinsic parameters $K_2$, we have

$$2p' = (K_2^{-1}) q'$$
The Fundamental Matrix

The epipolar constraint is given by: \( \mathbf{1}^T \mathbf{E} \mathbf{2} \mathbf{p}' = 0 \) with \( \mathbf{E} = \begin{bmatrix} (\mathbf{t}_2)_x \\ 1 \end{bmatrix} \mathbf{R} \)

where \( \mathbf{1} \mathbf{p} \) and \( \mathbf{2} \mathbf{p}' \) are calibrated coordinates in the two images.

The relationship between the calibrated coordinates (\( \mathbf{1} \mathbf{p}, \mathbf{2} \mathbf{p}' \)) and uncalibrated coordinates (\( \mathbf{q}, \mathbf{q}' \)) can be expressed as \( \mathbf{1} \mathbf{p} = (\mathbf{K}_1^{-1}) \mathbf{q} \) and \( \mathbf{2} \mathbf{p}' = (\mathbf{K}_2^{-1}) \mathbf{q}' \)

Therefore, we can express the epipolar constraint as:

\[
\mathbf{1}^T \mathbf{E} \mathbf{2} \mathbf{p}' = 0 = (\mathbf{K}_1^{-1} \mathbf{q})^T \mathbf{E} (\mathbf{K}_2^{-1} \mathbf{q}') = \mathbf{q}^T ((\mathbf{K}_1^{-1})^T \mathbf{E} \mathbf{K}_2^{-1}) \mathbf{q}' = \mathbf{q}^T \mathbf{F} \mathbf{q}' = 0
\]

where \( \mathbf{F} = (\mathbf{K}_1^{-1})^T \mathbf{E} \mathbf{K}_2^{-1} \) is a 3x3 matrix called the Fundamental Matrix.

This can be solved using 8 point algorithm WITHOUT CALIBRATION.

---

Epipolar Constraint for Calibrated and Uncalibrated Cameras

- **Calibrated Cameras:** \( \mathbf{1}^T \mathbf{E} \mathbf{2} \mathbf{p}' = 0 \)
  - \( \mathbf{E} \) is the essential matrix
  - \( \mathbf{1} \mathbf{p} = (\mathbf{K}_1^{-1}) \mathbf{q} \), \( \mathbf{2} \mathbf{p}' = (\mathbf{K}_2^{-1}) \mathbf{q}' \)
  - where \( \mathbf{K}_1, \mathbf{K}_2 \) are the intrinsic parameters of the two cameras and are determined by calibration.

- **Uncalibrated Cameras:** \( \mathbf{q}^T \mathbf{F} \mathbf{q}' = 0 \)
  - \( \mathbf{F} \) is the Fundamental matrix
  - \( \mathbf{q} \) and \( \mathbf{q}' \) are pixel coordinates

- Note similarity of equations.

- The fundamental matrix can be estimated directly from pixel coordinates using the 8 point algorithm.
Epipolar constraint for Uncalibrated Cameras

\[ q^T F q' = 0 \]

1. The epipolar constraint is homogenous in \( q, q' \) and \( F \)
2. It is bilinear in \( q \) and \( q' \). E.g., for a given value of \( q \), it is linear in \( q' \) and vice versa
3. Given pixel coordinates \( q' \) in \( \Pi' \), the equation of the epipolar line \( l \) in \( \Pi \) is \( a^T q = 0 \) where \( a = Fq' \)
4. Given pixel coordinates \( q \) in \( \Pi \), the equation of the epipolar line \( l' \) in \( \Pi' \) is \( b^T q' = 0 \) where \( b = F^T q \)

The Essential Matrix and Epipoles

\[ q^T F q' = 0 \]

5. The eigenvector of \( F \) corresponding to the zero eigenvalue is the epipole \( e' \)
6. The eigenvector of \( F^T \) corresponding to the zero eigenvalue is the epipole \( e \)
7. \( F \) is singular (determinant is zero & can’t be inverted)
8. \( F \) can be estimated from 8 corresponding points using the 8 point algorithm
Uncalibrated stereo

- Epipolar geometry can be determined without calibration.
- Images can be rectified so epipolar lines are rows of the rectified image.
- Matching can proceed in the same way.
- But you need calibration to estimate depth. However, if you arbitrarily make up intrinsic and extrinsic parameters and estimate depth. The estimated 3D point locations \( \hat{p} \) and true locations \( \hat{p} \) in homogeneous coordinates will differ by a linear transformation \( \hat{p} = A p \).

Trinocular Epipolar Constraints

These constraints are not independent!

\[
\begin{align*}
    p_1^T E_{12} p_2 &= 0 \\
    p_2^T E_{23} p_3 &= 0 \\
    p_3^T E_{31} p_1 &= 0
\end{align*}
\]
Ambiguity in binocular stereo

• Trinocular stereo can remove the binocular ambiguity.
• For each potential binocular match, the third image can be used to the match.
• There’s no need to search in the third image.
Structure from Motion

Also called

Visual SLAM

(Simultaneous Localization and Mapping)

Estimate 3D structure from images
Structure-from-Motion (SFM)

Given two or more images or video w/o any information on camera position/motion as input, estimate camera motion and 3-D structure of a scene.

Two Approaches

1. **Discrete motion (wide baseline)**
2. Continuous (Infinitesimal) motion usually from video
For $N$ points measured in $M$ images, estimate $(R_1,t_1),\ldots,(R_M,t_M)$ and $X_1,\ldots,X_N$.

It's only possible to estimate $t_1,\ldots,t_M$ and $X_1,\ldots,X_N$ up to a scale factor. Why?
How many views and how many points are needed to solve this?

Consider $M$ images of $N$ points, how many unknowns
1. Affix world coordinate system to location of first camera frame: $(M-1)\times 6$ unknowns for cameras
2. 3-D Structure: $3\times N$ unknowns for points
3. Can only recover structure and motion up to scale factor (one fewer unknown)

Total number of unknowns: $(M-1)\times 6 + 3\times N - 1$
Total number of measurements: $2\times M\times N$

Solution is possible when more measurements than unknowns: $(M-1)\times 6 + 3\times N - 1 \leq 2\times M\times N$

Some values of $N$ and $M$ satisfying this:
$M = 2, \ N = 5$
$M = 3, \ N = 4$