STEREO VISION

The slides are from several sources through James Hays (Brown); Srinivasa Narasimhan (CMU); Silvio Savarese (U. of Michigan); Bill Freeman and Antonio Torralba (MIT), including their own slides.
There are many cues in monocular vision which suggests that vision in stereo starts very early from two similar 2D images.

Let's see a few...
Depth cues: Linear perspective
Depth cues: Aerial perspective
Depth ordering cues: Occlusion

Source: J. Koenderink
Shape cues: Texture gradient
Why Stereo Vision?

\[ P = (X, Y, Z) \]

\[ P = (x, y, f) \]

\[
x = f \frac{X}{Z} = f \frac{kX}{kZ}
\]

\[
y = f \frac{Y}{Z} = f \frac{kY}{kZ}
\]

Fundamental Ambiguity:
Any point on the ray \( OP \) has image \( p \)
Why Stereo Vision?

A second camera can resolve the ambiguity, enabling measurement of depth via triangulation. The two cameras are calibrated.
Triangulation: the geometry behind

• Find shortest segment connecting the two viewing rays and let \( X \) be the midpoint of that segment

K known

K' known

\( O_1 \)

\( x_1 \)

\( O_2 \)

\( x_2 \)

cameras calibrated

R and t also known
Triangulation: is a linear approach

\[ \lambda_1 x_1 = P_1 X \quad x_1 \times P_1 X = 0 \quad [x_1] P_1 X = 0 \]
\[ \lambda_2 x_2 = P_2 X \quad x_2 \times P_2 X = 0 \quad [x_2] P_2 X = 0 \]

Vector product written as matrix multiplication.

The \( P_1 \) and \( P_2 \) (3x4) are known.

The \( x_1 \) and \( x_2 \) (3x1) are also known.

Two independent equations per point, in terms of three unknown entries of \( X \) (the 4th coord. = 1)
Stereo photography and stereo viewers

Take *two pictures* of the same subject from two slightly different viewpoints and display so that each eye sees *only one* of the images.

Invented by Sir Charles Wheatstone, 1838

blue/red stereo viewer

a few stereo images
two images of the Eiffel tower

superposed

http://www.johnsonshawmuseum.org
Random dot stereograms

• Julesz 1960. Can we identify *without an object* a 3D structure from two images?

• Experiment. Pair of synthetic identical images and moving a small rectangular part in the middle one pixel to the left.

  When viewed monocularly they appear random, when viewed stereoscopically, is a 3D structure.

  Human binocular fusion appears early in the central nervous system.
Depth without objects

Julesz, 1971

Bela Julesz
Stereovision

**Goal:** estimate the position of P given the observation of P from two viewpoints.

**Assumptions:** known camera parameters and position (K, R, T).
Stereo vision

Subgoals:
- Solve the correspondence problem between the images; p, p'.
- Use corresponding observations to triangulate in 3D.
If two images become parallel...

when views are parallel these two steps becomes much easier -- correspondence and triangulate.
E matrix with parallel images

- Parallel epipolar lines.
- Epipoles at infinity.
- \( v = v' \)

\[
p^T E p' = 0
\]

\[
\begin{bmatrix}
0 & 0 & 0 \\
0 & 0 & -T \\
0 & T & 0
\end{bmatrix}
\begin{bmatrix}
u' \\
v \\
1
\end{bmatrix} = 0
\]

\[
\begin{bmatrix}
0 \\
0 \\
Tv'
\end{bmatrix} = 0
\]

*Rectification*: making two images “parallel”. 
Making the two images parallel.

**GOAL:** Estimate the perspective transformations that makes images parallel.

- **Impose** $v' = v$
- Leaves degrees of freedom for determining $H$ and $H'$ !!!
- If not appropriate $H$ ($H'$) is chosen, severe projective distortions are present.
First frame will call the "camera", the second frame the "world".

$X_2$ is the inhomogeneous coordinates in the world coordinates
$X_1$ is the inhomogeneous coordinates in the camera coordinates

$R$ and $t$ take from world coordinate system to the camera c.s. describing the camera with respect the world coordinates.

By choice, the zero in the second coor. system is the first origin of the camera $O = [0 \ 0 \ 0 \ 1]^T$ and $e'$ is the epipole in second c.s.

The origin of the camera center $O$ in second c.s. is related by

$X_1 = R^T(X_2 - t)$ (inverse $R$, $t$) to $O'$ measured in the first c.s.

Thus, the coordinates of the world origin in the first coor. system is the point $O' = [-R\, t \ 1]^T$ and $e$ is the epipole of $O'$ in first c.s.

$X_2 = [0 \ 0 \ 0]^T \ X_1 \to O'$ first c.s.
Compute epipoles.

\[ e = K[I \ 0] [-R^T t \ 1]^T \]

\[ e' = K'[R \ t] [0 \ 1]^T \]

\[ e = K R^T T = [e_1 \ e_2 \ 1]^T \]

sign does not matter
Map $e$ to the $x$-axis at location $[1,0,1]^T$ (normalization).

$$e = [e_1, e_2, 1]^T \Rightarrow [1 \ 0 \ 1]^T$$

$$H_1 = R_H T_H$$

two 3x3 matrix
Send epipole to infinity.

\[ e = [1 \ 0 \ 1]^T \rightarrow [1 \ 0 \ 0]^T \]

\[ H_2 = \begin{bmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
-1 & 0 & 1 \\
\end{bmatrix} \]
\( H = H_2 H_1 \)

\[ p \leftrightarrow p' \quad \text{epipolar lines are} \]
\[ l = e \times p \quad l' = e' \times p' \]
\[ \text{can be computed.} \]
\[ \text{Parallel in each image separately.} \]

\[ \overline{H'}^{-T} l' = \overline{H}^{-T} l \]
\[ \text{line map} \]
\[ \text{Matched pair of transformation.} \]
\[ \text{To have } v = v' \text{ need to find the final transformations.} \]
Align the epipolar lines across the two images based on the matched transformation.

Described also in
[Hartley and Zisserman] Chapter 11 (Sec. 11.12)... more advanced.
IMAGE RECTIFICATION

Denote $E$ from image 2 to 1 and is given. If the images are in parallel, and the $t$ is known, the matrix $E$ becomes

$$E = [1 \ 0 \ 0]_x = [i]_x .$$

The two $3 \times 3$ homographies are $H$ and $H'$. Let $\tilde{p}$ and $\tilde{p}'$ be the rectified homogenous points

$$\tilde{p}'^\top [i]_x \tilde{p} = 0$$

related to the original points by

$$\tilde{p} = Hp \quad \tilde{p}' = H'p'. $$

The fundamental matrix $F$ is

$$F = H'^\top [i]_x H \quad p'^\top Fp = 0 .$$

The two homographies have more degrees of freedom then we need to solve the alignment.
The essential matrix is known. Denote $E$ from image 2 to 1.

$$p'^T H'^T [i]_x H p = 0 \quad -- \quad 8 \text{ DOF each homography } H, H'. $$

A homography is decomposed in

(shearing.similarity).projective -- order starts at right!

projective: epipolar lines become parallel \textit{in each image};

similarity: rotate the points into alignment with line maps \textit{in the two images} simultaneously. Its enough...

shearing: minimizing the horizontal distortion \textit{in each image}.

after projective and similarity

$$\begin{bmatrix} a & b & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

two intervals with known original aspect ratio
Rectification is useful.

Projective transformation.

(a) Original image pair overlayed with several epipolar lines.

(b) Image pair transformed by the specialized projective mapping $H_p$ and $H'_p$. Note that the epipolar lines are now parallel to each other in each image.
Rotate, translate, uniformly scale only.

Shearing with \( u(u') \) parameters only, reduces the projective distortion separately in both images. E.g., to preserve perpendicularity and aspect ratio of two line segments.

(c) Image pair transformed by the similarity \( H_r \) and \( H'_r \). Note that the image pair is now rectified (the epipolar lines are horizontally aligned).

(d) Final image rectification after shearing transform \( H_s \) and \( H'_s \). Note that the image pair remains rectified, but the horizontal distortion is reduced.

The two cameras have some distortion.
Geometry for a simple stereo system

- Assuming by now parallel optical axes, known camera parameters (i.e., calibrated cameras).
The two cameras at T distance along the x axis.
$T = X_r - X_l = (Z \times_r)/f - (Z \times_l)/f$

The disparity is:

$$Z = f \frac{T}{x_r - x_l}$$

Sometime disparity is $x_l - x_r$. 
Depth from disparity

\( (x', y') = (x + D(x, y), y) \)

So if we could find the corresponding points in two images, we could estimate relative depth…
Example: depth from disparity

Venus data pair

\[ B = T \text{ translation} \]

\[ x - x' = \frac{B \cdot f}{z} \]

Stereo pair

Disparity map / depth map

Disparity map with occlusions

http://vision.middlebury.edu/stereo/
Basic stereo matching algorithm

• Rectify the two stereo images to transform epipolar lines into scanlines
• *For each pixel* $x$ *in the first image*
  – Find corresponding epipolar scanline in the right image.
  – Examine all pixels on the scanline and pick the best match $x'$.
  – Compute disparity $x-x'$ and set depth($x$) = $fB/(x-x')$. 
The epipolar constraint helps, but much ambiguity remains.
Correlation Methods (from 1970's on)

- Pick up a window around $p(u,v)$.

...for normalized correlation

Two parallel image.
Correlation Methods

- Pick up a window around \( p(u,v) \).
- Build vector \( w \).
- Slide the window along \( v \) line in image 2 and compute \( w' \).
- Keep sliding until \( w \cdot w' \) (inner product) is maximized.
Correlation Methods

Normalized correlation has to maximize:

\[
\frac{(w - \bar{w})(w' - \bar{w}')}{\| (w - \bar{w})(w' - \bar{w}') \|}
\]
Correspondence search with... 

...maximization of the normalized correlation \textit{in the second image}. 

Norm. corr
Correspondence search with...  

...the minimum of the sum squared differences (SSD).
Example:

Minimization of the sum squared differences (SSD) in the second image.
Image block matching

How do we determine correspondences?

• *block matching*

\[
E(x, y; d) = \sum_{(x', y') \in N(x, y)} [I_L(x' + d, y') - I_R(x', y')]^2
\]

d is the *disparity* (horizontal motion)

How big should the neighborhood be?

Slide credit: Rick Szeliski
Neighborhood size

Smaller neighborhood: more details.
Larger neighborhood: fewer isolated mistakes.

\[ w = 3 \quad w = 20 \]

Slide credit: Rick Szeliski
Issues on stereo matching:

- Foreshortening effect
- Occlusions

T = B !

large B/z ratio
Small error in measurements implies large error in estimating depth.
• Homogeneous regions

Hard to match pixels in these regions.
• Repetitive patterns

Correspondence problem is difficult.
Apply nonlocal constraints to help enforce the correspondences.
Nonlocal constraints can have other problems

- **Uniqueness**
  - For any point in one image, there should be at most one matching point in the other image

- **Ordering**
  - Corresponding points should be in the same order in both views

Ordering constraint doesn’t hold
Stereo testing and comparisons


---

Scene
Tsukuba laboratory
left image

Ground truth

...154 submissions, July 2013
Stereo—best algorithms

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Avg. Rank</th>
<th>Error Threshold=1</th>
<th>Tsukuba ground truth</th>
<th>Venus ground truth</th>
<th>Teddy ground truth</th>
<th>Cones ground truth</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>nonocc</td>
<td>all</td>
<td>disc</td>
<td>nonocc</td>
<td>all</td>
</tr>
<tr>
<td>AdaptingBP [17]</td>
<td>2.8</td>
<td>1.11</td>
<td>1.37</td>
<td>5.79</td>
<td>0.10</td>
<td>0.21</td>
</tr>
<tr>
<td>DoubleBP2 [35]</td>
<td>2.9</td>
<td>0.88</td>
<td>1.29</td>
<td>4.76</td>
<td>0.13</td>
<td>0.45</td>
</tr>
<tr>
<td>DoubleBP [15]</td>
<td>4.9</td>
<td>0.88</td>
<td>1.29</td>
<td>4.76</td>
<td>0.14</td>
<td>0.60</td>
</tr>
<tr>
<td>SubPixDoubleBP [30]</td>
<td>5.6</td>
<td>1.24</td>
<td>1.76</td>
<td>5.98</td>
<td>0.12</td>
<td>0.46</td>
</tr>
<tr>
<td>AdaptOvrSeqBP [33]</td>
<td>9.9</td>
<td>1.69</td>
<td>2.04</td>
<td>5.68</td>
<td>0.14</td>
<td>0.20</td>
</tr>
<tr>
<td>SymbBP+occ [7]</td>
<td>10.8</td>
<td>0.97</td>
<td>1.75</td>
<td>5.09</td>
<td>0.16</td>
<td>0.33</td>
</tr>
<tr>
<td>PlaneFitBP [32]</td>
<td>10.8</td>
<td>0.97</td>
<td>1.83</td>
<td>5.26</td>
<td>0.15</td>
<td>0.51</td>
</tr>
<tr>
<td>AdaptDispCalib [36]</td>
<td>11.8</td>
<td>1.19</td>
<td>1.42</td>
<td>6.15</td>
<td>0.23</td>
<td>0.34</td>
</tr>
<tr>
<td>Segm+visib [4]</td>
<td>12.2</td>
<td>1.30</td>
<td>1.57</td>
<td>6.92</td>
<td>0.21</td>
<td>1.06</td>
</tr>
<tr>
<td>C-SemiGlob [19]</td>
<td>12.3</td>
<td>2.61</td>
<td>2.29</td>
<td>8.99</td>
<td>0.25</td>
<td>0.57</td>
</tr>
<tr>
<td>SO+borders [29]</td>
<td>12.8</td>
<td>1.28</td>
<td>1.71</td>
<td>6.83</td>
<td>0.25</td>
<td>0.53</td>
</tr>
<tr>
<td>DistinctSM [27]</td>
<td>14.1</td>
<td>1.21</td>
<td>1.75</td>
<td>6.39</td>
<td>0.35</td>
<td>0.69</td>
</tr>
</tbody>
</table>
Example: window correlation

Window-based matching
(best window size)

Ground truth

not the best method
A few of the other methods:
- based on edges, corners, gradients;
- rank statistics;
- dynamic programming;
- matching as energy minimization (graph cuts);
- multiple-baseline stereo;
- space carving stereo etc...

None of the methods is the ultimate solution. Stereo (like everything else) works only if the matching is robust.
Active stereo (point)

Replace one of the two cameras by a projector
- Single camera.
- Projector geometry calibrated too.
- What’s the advantage of having the projector? Correspondence problem solved!
If the projector and camera are parallel... correspondence problem solved. Otherwise, transform it into parallel first.
High precision: Laser scanning

Optical triangulation

- Project a single stripe of laser light.
- Scan it across the surface of the object.
- This is a very precise version of structured light scanning.

Digital Michelangelo Project
Levoy et al.
http://graphics.stanford.edu/projects/mich/

Source: S. Seitz
Example: laser scanning

The Digital Michelangelo Project, Levoy et al.

Source: S. Seitz
Active stereo (color-coded stripes)

- Dense reconstruction. Projector, image are parallel.
- Correspondence found by a technique based on the color code.

L. Zhang, B. Curless, and S. M. Seitz 2002
S. Rusinkiewicz & Levoy 2002
KINECT
Structured infrared light

Figure 1. Kinect consists of Infra-red (IR) projector, IR camera and RGB camera (illustration from [11]).