Object: Building, 45º pose, 8-10 meters away

Object: Person, back; 1-2 meters away

Object: Police car, side view, 4-5 m away
Object: Building’s name;  
What are the businesses inside?  
Street or intersection name
Visual technology

- Scene understanding
- Navigation
- Interaction
- Augmentation
- Manipulation

Object 1

- semantic
- geometry

Object N

- semantic
- geometry

Image/video

- Recognizing objects under arbitrary viewing conditions
- Recognize their pose
Car: front-right
Iron: top-rear-left

training

• Minimal supervision
Single 3D object recognition

- Ballard, '81
- Grimson & L. Perez, '87
- Lowe, '87
- Edelman et al., '91
- Ullman & Barsi, '91
- Rothwell '92
- Linderberg, '94
- Murase & Nayar '94
- Zhang et al.'95
- Schmid & Mohr, '96
- Schiele & Crowley, '96
- Lowe, '99
- Jacob & Barsi, '99
- Rothganger et al., '04
- Ferrari et al., '05
- Brown & Lowe '05
- Snavely et al. '06
- Yin & Collins, '07
Recognition for Virtual sightseeing
Object manipulations

Robot arm for automatic gas fill up
Object manipulations
Single view object categorization
Safe driving

Security

photography
Overview

- Single 3D object recognition
- Single view object categorization
- 3D object categorization
Single 3D object recognition

• Ballard, ‘81
• Grimson & L.-Perez, ‘87
• Lowe, ’87
• Ballan et al. ’91
• Ullman & Barsi, ’91
• Rothwell ‘92
• Linderberg, ’94
• Murase & Nayar ‘94
• Zhang et al. ’95
• Schmid & Mohr, ‘96
• Schiele & Crowley, ’96
• Lowe, ‘99
• Jacob & Barsi, ‘99
• Rothganger et al., ‘04
• Ferrari et al., ’05
• Brown & Lowe ’05
• Snavely et al ’06
• Yin & Collins, ‘07
Basic scheme

- Representation
  - Features
  - Descriptors
  - Model

- Model learning

- Recognition
  - Hypothesis generation
  - Model verification
Object representation: Collection of patches in 3D

Rothganger et al. ’06

x, y, z + h, v + descriptor

Courtesy of Rothganger et al.
Model learning

Build a 3D model:

- N images of object from N different view points
- Match key points between consecutive views [create sample set]
- Use affine structure from motion to compute 3D location and orientation + camera locations [RANSAC]
- Find connected components
- Use bundle adjustment to refine model
- Upgrade model to Euclidean assuming zero skew and square pixels
Affine Structure from Motion

Books:
- Faugeras, ‘95
- Zisserman & Hartley, ’00
- Ma, Soatto, et al. ‘05

Affine epipolar geometry between image pairs.
Fundamental matrix $F$ imposes:
$x'Fx = 0$ and $l' = Fx$
Learnt models

Rothganger et al. ‘03 ’06
Basic scheme

- Representation
  - Features
  - Descriptors
  - Model

- Model learning

- Recognition [object instance from a single image]
  - Hypothesis generation
  - Model verification
Recognition

[Rothganger et al. ‘03 ‘06]

1. Find matches between model and test image features
1. Find matches between model and test image features

2. Generate hypothesis:
   • Compute transformation $M$ from $N$ matches ($N=2$; affine camera; affine key points)

3. Model verification
   • Use $M$ to project other matched 3D model features into test image
Goal: Estimate (fit) the best M in presence of outliers
Line fitting with outliers

\[ \pi : I \to \{ P, O \} \]

such that:

\[ f(P, \beta) < \delta \]

\[ \min_{\pi} \left| \{ O \} \right| \]

Model parameters

\[ f(P, \beta) = \left\| \beta - \left( P^T P \right)^{-1} P^T \right\| \]
Fitting homographies for stitching panoramas

\[
\begin{align*}
\pi : I^h & \rightarrow \{ P^h, O^h \} \\
\tau : I^k & \rightarrow \{ P^k, O^k \}
\end{align*}
\]

such that:

\[
f(P^h, P^k, \beta) < \delta
\]

\[
\min |O^h \cup O^k|
\]

\[
f(P^h, P^k, \beta) = \| P^h - H P^k \|
\]
Fitting homographies for stitching panoramas
RANSAC – Basic philosophy

(RANdom SAmple Consensus) : Fischler & Bolles in ‘81.
Learning technique to estimate parameters of a model by random sampling of observed data

• Data elements are used to vote for one (or multiple) models

• Robust to outliers and missing data

• Assumption 1: Noise features will not vote consistently for any single model (“few” outliers)

• Assumption 2: there are enough features to agree on a good model (“few” missing data)
Algorithm:

1. Select random sample of minimum required size to fit model
2. Compute a putative model from sample set
3. Compute the set of inliers to this model from whole data set

Repeat 1–3 until model with the most inliers over all samples is found

Sample set = set of points in 2D
RANSAC

Algorithm:
1. Select random sample of minimum required size to fit model = [2]
2. Compute a putative model from sample set
3. Compute the set of inliers to this model from whole data set
Repeat 1–3 until model with the most inliers over all samples is found

Sample set = set of points in 2D
Algorithm:

1. Select random sample of minimum required size to fit model = [2]
2. Compute a putative model from sample set
3. Compute the set of inliers to this model from whole data set

Repeat 1–3 until model with the most inliers over all samples is found

Sample set = set of points in 2D
RANSAC

Sample set = set of points in 2D

$|O| = 14$

Algorithm:

1. Select random sample of minimum required size to fit model = [2]
2. Compute a putative model from sample set
3. Compute the set of inliers to this model from whole data set

Repeat 1–3 until model with the most inliers over all samples is found
RANSAC

*(RANdom SAmple Consensus):*

Fischler & Bolles in ‘81.

\[|\mathcal{O}| = 6\]

Algorithm:

1. Select random sample of minimum required size to fit model [?]
2. Compute a putative model from sample set
3. Compute the set of inliers to this model from whole data set

Repeat 1–3 until model with the most inliers over all samples is found
How many samples?

• Number of samples \( N \)
  
  - Choose \( N \) so that, with probability \( p \), at least one random sample is free from outliers (e.g. \( p=0.99 \)) (outlier ratio: \( e \))

• Initial number of points \( s \)
  
  - Typically minimum number needed to fit the model

• Distance threshold \( \delta \)
  
  - Choose \( \delta \) so probability for inlier is \( p \) (e.g. 0.95)
  
  - Zero-mean Gaussian noise with std. dev. \( \sigma \): \( t^2=3.84 \sigma^2 \)

\[
N = \log \left( \frac{1 - p}{\log \left( 1 - (1 - e)^s \right)} \right)
\]

<table>
<thead>
<tr>
<th>( s )</th>
<th>5%</th>
<th>10%</th>
<th>20%</th>
<th>25%</th>
<th>30%</th>
<th>40%</th>
<th>50%</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>11</td>
<td>17</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>4</td>
<td>7</td>
<td>9</td>
<td>11</td>
<td>19</td>
<td>35</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>4</td>
<td>7</td>
<td>9</td>
<td>13</td>
<td>17</td>
<td>34</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>6</td>
<td>12</td>
<td>17</td>
<td>26</td>
<td>57</td>
<td>146</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>7</td>
<td>16</td>
<td>24</td>
<td>37</td>
<td>97</td>
<td>293</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>6</td>
<td>20</td>
<td>33</td>
<td>54</td>
<td>163</td>
<td>588</td>
</tr>
<tr>
<td>8</td>
<td>5</td>
<td>9</td>
<td>26</td>
<td>44</td>
<td>78</td>
<td>272</td>
<td>1177</td>
</tr>
</tbody>
</table>

Source: M. Pollefeys
1. Find matches between model and test image features

2. Generate hypothesis:
   - Compute transformation $M$ from $N$ matches ($N=2$; affine camera; affine key points)

3. Model verification
   - Use $M$ to project other matched 3D model features into test image
Object to recognize

Initial matches based on appearance

Matches after pose verification

Recovered pose

Courtesy of Rothganger et al
3D Object Recognition results

• Handle severe clutter
Test dataset
A comparative experiment
Other multi-view matching algorithms

- Ferrari et al. ‘04, ‘06
- Lazebnick et al. ‘04
- Brown et al. ‘05
- Toshev, Shi, Daniilidis, 07
Overview

• Single 3D object recognition
• Single view object categorization
• 3D object categorization
3D Object Categorization

Mixture of 2D single view models

- Weber et al. ‘00
- Schneiderman et al. ’01
- Bart et al. ‘04

Full 3D models

- Bronstein et al, ‘03
- Ruiz-Correa et al. ’03
- Funkhouser et al ’03
- Capel et al ’02
- Johnson & Herbert ‘99

Multi-view models

- Thomas et al. ‘06
- Savarese et al, 07, 08
- Chiu et al. ‘07
- Hoiem, et al., ’07
- Yan, et al. ’07
- Kushal, et al., ’07
- Liebelt et al 08
- Sun et al 08
Mixture of single-view 2D models

• Weber et al. ‘00
• Schneiderman et al. ’01
Mixture of single-view 2D models

- Mixture of 2-D models
  - Weber, Welling and Perona CVPR ‘00

\[
p(X^o, x^m, h) = \sum_{\omega=1}^{\Omega} p(X^o, x^m, h | \omega)p(\omega).
\]
Mixture of single-view 2D models

Single view models are independent
• No information is shared
• No sense of correspondences of parts under 3D transformations

- Weber et al. ‘00
- Schneiderman et al. ’01

Single view Model

3D Category model
3D Object Categorization

Mixture of 2D single view models

- Weber et al. ‘00
- Schneiderman et al. ’01
- Bart et al. ‘04

Full 3D models

- Bronstein et al, ‘03
- Ruiz-Correa et al. ’03,
- Funkhouser et al ’03
- Capel et al ’02
- Johnson & Herbert ‘99

Multi-view models

- Thomas et al. ‘06
- Savarese et al, 07, 08
- Chiu et al. ‘07
- Hoiem, et al., ’07
- Yan, et al. ’07
- Kushal, et al., ’07
- Liebelt et al 08
- Sun et al 08
A 3D model category is built from a collection of 3D range data or CAD models.
Shape distributions  Osada et al  02

Spherical harmonics  Kazhdan et al.  03
A 3D model category is built from a collection of 3D range data or CAD models.

- Build a 3D model is expensive
- Difficult to incorporate appearance information
- Need to cope with 3D alignment (orientation, scale, etc...)

Full 3D models

3D model instance

3D category model

- Bronstein et al, ‘03
- Ruiz-Correa et al. ’03
- Funkhouser et al ’03
- Kazhdan et al.03
- Osada et al ‘02
- Capel et al ’02
- Johnson & Herbert ’99
- Amberg et al ‘08
3D Object Categorization

Mixture of 2D single view models

- Weber et al. ‘00
- Schneiderman et al. ’01
- Bart et al. ‘04

Full 3D models

- Bronstein et al, ‘03
- Ruiz-Correa et al. ’03
- Funkhouser et al ’03
- Capel et al ’02
- Johnson & Herbert ‘99

Multi-view models

- Thomas et al. ‘06
- Savarese et al, 07, 08
- Chiu et al. ‘07
- Hoiem, et al., ’07
- Yan, et al. ’07
- Kushal, et al., ’07
- Liebelt et al 08
- Sun et al 08
Sparse set of interest points or parts of the objects are linked across views.
Multi-view models by rough 3d shapes

Yan, et al. ’07
Multi-view models by rough 3d shapes

Hoiem, et al., ’07
Sparse set of interest points or parts of the objects are linked across views.

Multi-view models by ISM representations

[Thomas et al. '06]
Multi-view models by ISM representations

ISM representation
Leibe, Leonardis, and Schiele, ECCV Workshop on Statistical Learning in Computer Vision 2004

• Visual codebook is used to index votes for object position
• Generalized Hough transform
Combining multi-views and ISM models

Region tracks

[Ferrari et al. ’04, ‘06]

Set of region-tracks connecting model views
Each track is composed of image regions of a single physical surface patch
along the model views in which it is visible.

[Thomas et al. ’06]
<table>
<thead>
<tr>
<th></th>
<th>Single view/Mixture</th>
<th>Multi-view</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Properties</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>View point invariant</td>
<td>X</td>
<td>✓</td>
</tr>
<tr>
<td>No supervision</td>
<td>✓</td>
<td>X</td>
</tr>
<tr>
<td># Categories</td>
<td>~300</td>
<td>2</td>
</tr>
<tr>
<td>Share information across views</td>
<td>X</td>
<td>✓</td>
</tr>
<tr>
<td>View synthesis</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Pose estimation</td>
<td>X  → ✓</td>
<td>X</td>
</tr>
</tbody>
</table>

Category View point