Methods for Representing and Recognizing 3D objects

part 2

1st Sino-USA Summer School in Vision, Learning, and Pattern Recognition

VLPR 2009 • July 20-27, 2009 • Peking University, Beijing, China

Silvio Savarese
University of Michigan at Ann Arbor
Part-based Multi-view models

- Canonical parts captures diagnostic appearance information
- 2d ½ structure linking parts via weak geometry
Canonical parts

- If physical part is planar, canonical part is stable point on the manifold.
- Canonical part can be computed from connected component of parts.
Canonical parts

connected component of parts
Connected components of parts

Unlabeled mix of images:
- category labels
- no pose labels;
- images of same instance from multiple views
\[ I^h = [x_1, x_2, \ldots, x_M] \]
Let $I^h = [x_1, x_2, \cdots, x_M]$ and $I^k$.

The transformations are defined as

\[ \pi : I^h \rightarrow \{ P_1^h, P_2^h, \cdots, P_N^h, O^h \} \]

\[ \tau : I^k \rightarrow \{ P_1^k, P_2^k, \cdots, P_N^k, O^k \} \]

such that:

\[ f(P_i^h, P_j^k, \beta_{i,j}) < \delta \quad \forall i, j = 1 \cdots N \]

\[ f(P_i^h, P_j^k, \beta_{i,j}) = \| P_i^h - H_{i,j} P_j^k \| \]

subject to

\[ \min \left| O^h \cup O^k \right| + c \cdot N \]
\[ I^h = [x_1, x_2, \ldots, x_M] \]

\[
\pi : I^h \rightarrow \{ P_1^h, P_2^h, \ldots P_N^h, O^h \}
\]

\[
\tau : I^k \rightarrow \{ P_1^k, P_2^k, \ldots P_N^k, O^k \}
\]

such that:

\[ f(P_i^h, P_j^k, \beta_{i,j}) < \delta \]
\[ \forall i, j = 1 \ldots N \]

\[ f(P_i^h, P_j^k, \gamma) < \delta \]

\[ f(P_i^h, P_i^k, \gamma) = \left\| P_i^h - H_{i,j} P_j^k \right\| \]

\[ f(P_i^h, P_j^k, \beta_{i,j}) = \left\| P_i^h - H_{i,j} P_j^k \right\| \]

\[ f(P_i^h, P_i^k, \gamma) = \left\| [P_1^h, \ldots P_N^h] F [P_1^k, \ldots P_N^k]^T \right\| \]
$I^h = [x_1, x_2, \ldots, x_M]$

$\pi : I^h \rightarrow \{ P^h_1, P^h_2, \ldots P^h_N, O^h \}$

$\tau : I^k \rightarrow \{ P^k_1, P^k_2, \ldots P^k_N, O^k \}$

$$\min \left| O^h \cup O^k \right| + c \cdot N$$

**GOAL:**
- discover partition while fitting multiple homographies
- fit global constraint
- minimize outlier set

- Use sequential RANSAC or RANSAC & J-linkage [toldo, fusiello eccv 08]
\[ \pi : I^h \rightarrow \{ P_1^h, P_2^h, P_3^h, O^h \} \]

\[ \tau : I^k \rightarrow \{ P_1^k, P_2^k, P_3^k, O^k \} \]
Canonical parts
Linkage structure
Cost = \sum_{h,k \in L} \sum_{i,j \in L} G(i, j, h, k) \delta_{ij} \delta_{hk} + \sum_{i,j \in L} A(i, j) \delta_{ij}

\sum_{j \in L} \delta_{ij} = 1 \quad \delta_i = \{0,1\} \quad \text{IQP problem}

Maciel & Costeira ‘03
Berg et al ’05
Leordeanu & Hebert ‘05
Category Model

Canonical view

\[ \mathcal{H}_{i,j} = \begin{pmatrix} 1 & t_{i,j} \\ 0 & 1 \end{pmatrix} \]

2D single view model of the object

Aspect Graphs
- Koenderink & V. Doorn 76
- Bowyer & Dyer 90
- Cyr & Kimia 04
1. Find hypotheses of canonical parts consistent with a given pose
Object Recognition

Query image

Algorithm

1. Find hypotheses of canonical parts consistent with a given pose
2. Infer position and pose of other canonical parts

model
Algorithm

1. Find hypotheses of canonical parts consistent with a given pose
2. Infer position and pose of other canonical parts
Object Recognition

Query image

model

Algorithm

1. Find hypotheses of canonical parts consistent with a given pose
2. Infer position and pose of other canonical parts
3. Optimize over \( E, G \) and \( s \) to find best combination of hypothesis → error
3D object class dataset

- bicycle
- car
- cellphone
- iron
- mouse
- shoe
- toaster
- stapler
• 8 azimuth angles
• 3 zenith
• 3 distances

~ 7000 images!
Examples

Category: car
Azimuth = 45°
Zenith = 30°
Distance = close
Examples

Category: mouse
Azimuth = 315°
Zenith = 0°
Distance = medium
### Classification accuracy

**Average Perf. = 75.7%**

<table>
<thead>
<tr>
<th></th>
<th>c</th>
<th>b</th>
<th>i</th>
<th>m</th>
<th>s</th>
<th>s</th>
<th>t</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>cellphone</td>
<td>.76</td>
<td>.03</td>
<td>.03</td>
<td>.02</td>
<td>.10</td>
<td>.03</td>
<td>.03</td>
<td></td>
</tr>
<tr>
<td>bike</td>
<td>.02</td>
<td>.81</td>
<td>.07</td>
<td>.02</td>
<td>.03</td>
<td>.02</td>
<td>.03</td>
<td></td>
</tr>
<tr>
<td>iron</td>
<td></td>
<td>.77</td>
<td>.09</td>
<td>.06</td>
<td>.04</td>
<td>.04</td>
<td></td>
<td></td>
</tr>
<tr>
<td>mouse</td>
<td>.04</td>
<td>.04</td>
<td>.87</td>
<td>.02</td>
<td>.02</td>
<td>.02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>shoe</td>
<td>.04</td>
<td>.06</td>
<td>.04</td>
<td>.62</td>
<td>.12</td>
<td>.12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>stapler</td>
<td></td>
<td>.11</td>
<td>.04</td>
<td>.04</td>
<td>.77</td>
<td>.04</td>
<td></td>
<td></td>
</tr>
<tr>
<td>toaster</td>
<td>.08</td>
<td>.06</td>
<td>.03</td>
<td>.06</td>
<td>.75</td>
<td>.03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>car</td>
<td>.04</td>
<td>.04</td>
<td>.12</td>
<td>.04</td>
<td>.07</td>
<td>.70</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Average accuracy in classifying 1 out 8 categories
Random=12%
Failure example

Category: car
Azimuth = 225°
Zenith = 30°
Distance = close
Pose estimation accuracy

<table>
<thead>
<tr>
<th></th>
<th>0°</th>
<th>45°</th>
<th>90°</th>
<th>135°</th>
<th>180°</th>
<th>225°</th>
<th>270°</th>
<th>315°</th>
</tr>
</thead>
<tbody>
<tr>
<td>0°</td>
<td>.64</td>
<td>.08</td>
<td>.16</td>
<td>.12</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>45°</td>
<td>.02</td>
<td>.40</td>
<td>.06</td>
<td>.06</td>
<td>.33</td>
<td>.06</td>
<td>.06</td>
<td></td>
</tr>
<tr>
<td>90°</td>
<td>.02</td>
<td>.47</td>
<td>.16</td>
<td>.02</td>
<td>.04</td>
<td>.18</td>
<td>.11</td>
<td></td>
</tr>
<tr>
<td>135°</td>
<td>.05</td>
<td>.13</td>
<td>.62</td>
<td>.07</td>
<td>.13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>180°</td>
<td>.25</td>
<td>.03</td>
<td>.03</td>
<td>.06</td>
<td>.53</td>
<td>.03</td>
<td>.06</td>
<td></td>
</tr>
<tr>
<td>225°</td>
<td>.12</td>
<td>.05</td>
<td>.02</td>
<td>.71</td>
<td>.07</td>
<td>.02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>270°</td>
<td>.20</td>
<td>.17</td>
<td>.57</td>
<td>.06</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>315°</td>
<td>.06</td>
<td>.03</td>
<td>.09</td>
<td>.03</td>
<td>.12</td>
<td>.64</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

av. of 8 categ 57.2%
### Summary

<table>
<thead>
<tr>
<th></th>
<th>Single view</th>
<th>Mixture / Multi-view</th>
<th>Sav. et al, 07</th>
</tr>
</thead>
<tbody>
<tr>
<td>View point invariant</td>
<td>X</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>No supervision</td>
<td>✓</td>
<td>X</td>
<td>X → ✓</td>
</tr>
<tr>
<td># Categories</td>
<td>~300</td>
<td>2</td>
<td>8</td>
</tr>
</tbody>
</table>

- lack of a coherent methodology for learning parameters
- need of multi-view observations of the same object instance
- no generative model for robust learning and recognition
Part-based Multi-view Models

Savarese, Fei-Fei, ICCV 07

Savarese, Fei-Fei, ECCV 08

Sun, Su, Savarese, Fei-Fei, CVPR 09

Su, Sun, Fei-Fei, Savarese, ICCV 09

Min Sun
University of Michigan, USA

Hao Su
Beihang University, China

Fei-Fei Li.
Stanford U, USA
Part-based Multi-view Models

- Probabilistic generative part-based model
- Dense multi-view representation on the viewing sphere
Part-based generative model

Viewing Sphere

\( \pi \sim \text{Dir}(\alpha) \)

\( Y_n \sim \text{Mult}(\eta) \)

\( X_n \sim N(\theta) \)

\( Y = \text{Codeword}, \ X = \text{Location} \)
Dense representation on view-sphere

Constraints based on view morphing geometry

\[
m = \sum_{g=1}^{3} \hat{m}_{T_k}^g \cdot S_g
\]

\[
\theta = (m, \Sigma)
\]

Part Appearance

Part shape

Seitz & Dyer SIGGRAPH 96
Xiao & Shah CVIU ’04
Dense representation on view-sphere

- Pre-warping transformations $H$
- Post-warping transformation $A$

$\theta = (m, \Sigma)$
Dense representation on view-sphere

Joint probability of the model:

\[
P(X, Y, T, S, K, \pi) = P(T|\phi)P(\pi|\alpha_T)P(S|\beta) \prod_{n}^{N} \{ P(x_n|\theta_{TK_n}(S), A)P(y_n|\eta_{TK_n}(S))P(K_n|\pi) \}
\]

Observable variables: X, Y, T, S
Latent variables: K, \pi + relevant priors
Learning

Variational EM

E step:
• update hidden part assignments \( \pi \)

part proportion \( K \)

M step:
• update part proportion prior \( \alpha \)

appearance \( \eta \)

Location/shape \( \theta \)

• weakly supervised
• incremental (training image)
Weakly supervised

- Class label
- No pose label
Incremental learning
Incremental learning

No need for observations of the same object instance from multiple views.
Incremental learning

\[ \hat{\eta}_{tK} = \frac{N_{tK}^{gw}}{N_{tK}} \]

\[ N_{tK}^{gw} = \sum_{j \in (T_j=t, G(S_j)=g)} \sum_{n \in (y_{n_j}=w)} \rho_{n_j}^K \]

\[ N_{tK}^g = \sum_w N_{tK}^{gw} \]

\[ \sum_n (y_{n_j}=w) \rho_{n_j}^K \]

\[ N_{T_j K}^{G(S_j)w} \]
key ingredients for weakly supervised learning

Initialization

Constraints
  • Across Triangle
  • Within Triangle
Initialization
Initialization

\[ \pi : I^h \rightarrow \{ P^h_1, P^h_2, P^h_3, O^h \} \]

\[ \tau : I^k \rightarrow \{ P^k_1, P^k_2, P^k_3, O^k \} \]

Sequential ransac
J-linkage
key ingredients for weakly supervised learning

Initialization

Constraints
  • Across Triangle
  • Within Triangle
Encoded as a penalty term in the Variational EM algorithm.
3D object class dataset

bicycle

iron
cellphone

mouse

shoe
toaster

stapler

3D object class dataset
Detection

Car

Bicycle

Morphing model, 2009
Savarese, & Fei-Fei ICCV ’07
Viewpoint Classification: Car

Classification Accuracy

0°
45°
90°
135°
180°
225°
270°
315°

Classification Accuracy

Morphing model, 2009
Savarese, & Fei-Fei ICCV ’07
Detection: Pascal 2006 dataset
Detection: Pascal 2006 dataset

**Car**
- ICCV09 Morphing Model: 0.35 (average p)

**Bicycle**
- ICCV09 Morphing Model: 0.347 (average p)
Viewpoint Classification: Car–Pascal 2006 dataset

Classification Accuracy

- Discrete, CVPR ‘09
- Morphing, ICCV ‘09

Pascal ‘06
Household Item Dataset: Detection

<table>
<thead>
<tr>
<th>Object Class</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sewing Machine</td>
<td>98.1</td>
</tr>
<tr>
<td>Microscope</td>
<td>87.9</td>
</tr>
<tr>
<td>Travel Iron</td>
<td>88.1</td>
</tr>
<tr>
<td>Swivel Chair</td>
<td>91.2</td>
</tr>
<tr>
<td>Calculator</td>
<td>97.2</td>
</tr>
<tr>
<td>Flashlight</td>
<td>87.1</td>
</tr>
<tr>
<td>Teapot</td>
<td>86.4</td>
</tr>
<tr>
<td>Watch</td>
<td>84.9</td>
</tr>
<tr>
<td>All</td>
<td>90.1</td>
</tr>
</tbody>
</table>

ROC curve analysis:
- Average ROC curve
- Plus one standard deviation
- Minus one standard deviation

Detection Rate vs. False Alarm Rate graph.
## Household Item Dataset: 8-Viewpoint Classification

<table>
<thead>
<tr>
<th>Object Class</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sewing Machine</td>
<td>71.4</td>
</tr>
<tr>
<td>Microscope</td>
<td>63.9</td>
</tr>
<tr>
<td>Travel Iron</td>
<td>73.5</td>
</tr>
<tr>
<td>Swivel Chair</td>
<td>58.6</td>
</tr>
<tr>
<td>Calculator</td>
<td>69.2</td>
</tr>
<tr>
<td>Flashlight</td>
<td>68.4</td>
</tr>
<tr>
<td>Teapot</td>
<td>60.0</td>
</tr>
<tr>
<td>Watch</td>
<td>61.9</td>
</tr>
<tr>
<td>All</td>
<td>70.2</td>
</tr>
</tbody>
</table>
Typical Examples

Bicycle

Binocular Microscope

Blue arrows indicate the viewpoint for the detected object (in red bounding box).
Typical Examples

Car

Travel Iron

Blue arrows indicate the viewpoint for the detected object (in red bounding box.)
Novel view object synthesis from a single image

For the first time!

car calculator flashlight bike

For natural or artificial scenes, see hoeim 07; saxena
## Conclusions

<table>
<thead>
<tr>
<th></th>
<th>Single view</th>
<th>Mixture / Multi-view</th>
<th>Sav. et al, 07</th>
<th>Morphing model</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>View point invariant</strong></td>
<td>X</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td><strong>No supervision</strong></td>
<td>✓</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td><strong>Category ; all views</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong># Categories</strong></td>
<td>~300</td>
<td>2</td>
<td>8</td>
<td>16</td>
</tr>
<tr>
<td><strong>Share information across views</strong></td>
<td>X</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td><strong>View synthesis</strong></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td><strong>Pose estimation</strong></td>
<td>X →</td>
<td>X</td>
<td>X →</td>
<td>✓</td>
</tr>
</tbody>
</table>

- **No supervision**: View point invariant across all views.
- **View point invariant**: All views have a consistent view point.
- **Share information across views**: Information is shared between different views.
- **View synthesis**: Can synthesize views.
- **Pose estimation**: Estimation of pose possible.

---

*Note: The Morphing model column indicates that the model can handle all views and instances without supervision.*
Thank you!