Human Parsing using Stochastic And-Or Grammars and Rich Appearances

Brandon Rothrock and Song-Chun Zhu
UCLA Dept. of Computer Science
MOTIVATIONS

• Rich appearance representation
• Semantic parts
• Part sharing
• Reconfigurable parts
BODY REPRESENTATION

layer 1

layer 2

layer 3

layer 4
COMPOSITION RULES

and-node

or-nodes

Part form

Part types
AND-OR GRAPH GRAMMAR

- or-node group
- and-node
AND-OR GRAPH GRAMMAR

- or-node group
- and-node

Grammar

Derivation tree

Thursday, November 17, 11
AND-OR GRAPH GRAMMAR

Graphemes:
- or-node group
- and-node

Grammar

Parse Graph

Thursday, November 17, 11
\[ p(pg|I; \Theta) \propto p(I|pg; \Theta)p(pg; \Theta) \]

\[ = \frac{1}{Z(\Theta)} \exp \left\{ -E_a(pg|I) - E_d(pg) - E_g(pg) \right\} \]

appearance \hspace{2cm} \text{derivation (part forms)} \hspace{2cm} \text{geometry}
APPEARANCE MODEL

• Adapted from Active Basis model

(IJCV 2010) Learning Active Basis Model for Object Detection and Recognition (Wu, Si, Gong, Zhu)
(CVPR 2009) Learning Mixed Image Templates for Object Recognition (Si, Gong, Wu, Zhu)

Template learned by pursuit

\[ q \rightarrow p_1 \rightarrow \ldots \rightarrow p_k = p(I_{A_{v_i}}|v_i) \sim f \]

Model derived by minimizing KL

\[ p_k^* = \arg \min \text{KL}(p_k||p_{k-1}) \quad s.t. \quad E_{p_k}[r_k] = E_f[r_k] \]

Resulting in the following form:

\[ p_k(I_A) = q(I_A) \exp \left\{ \sum_{i=1}^{k} \lambda_i r_i(I_{A_i}) - \log z_i \right\} \]
APPEARANCE MODEL
APPEARANCE MODEL

upper body

lower body

leg

upper arm
APPEARANCE MODEL

\[
p(I|pg) = q(I_{\Lambda_{pg}}) \prod_{v_i \in pg} p(I_{\Lambda_{v_i}}|v_i)
= q(I) \prod_{v_i \in pg} \exp \left\{ \sum_{j=1}^{k} \lambda_j r_j(I_{\Lambda_j}) - \log z_j \right\}
\]
DERIVATION MODEL

SCFG case is indifferent to neighboring forms

\[ p(pg) = \prod_{v_i \in V_{pg}} p(\omega(v_i)) \]
Allow child forms to depend on parent form

\[ p_d(pg) = \prod_{v_i \in V_{pg}} p(\omega(C(v_i))|\omega(v_i)) \]

\[ \approx \prod_{v_i \in V_{pg}} \prod_{(j,k) \in E(v_i)} p(\omega(v_j), \omega(v_k)|\omega(v_i)) \]
GEOMETRY MODEL

\[ p(v_j, v_k | \omega(v_i)) = \mathcal{N}(T_{kj}(x_k) - T_{jk}(x_j), 0, \Sigma_{ij}) \]

\[ p_g(pg) \propto \prod_{v_i \in V_{pg}} \prod_{(jk) \in E(v_i)} p(v_j, v_k | \omega(v_i)) \]
SAMPLING

constrained samples

unconstrained samples

Thursday, November 17, 11
INFERENCE

Recursive scoring function

\[ s(\rho | I) = s_a(\rho | I) + s_g(\rho) + s_d(\rho) + \sum_{\rho_i \in C(\rho)} s(\rho_i | I) \]

\[ s_a(\rho | I) = \log p(I_{\Lambda_\rho} | \rho) = \sum_{i=1}^{n} [\lambda_i r_i(I_{\Lambda_i}) - \log z_i] \]

\[ s_d(\rho) = \log \frac{\prod_{(ik) \in E(\rho)} p(\omega(\rho_j), \omega(\rho_k) | \omega(\rho))}{\prod_{\rho_j \in V(\rho)} p(\omega(\rho_j) | \omega(\rho))^{d(\rho_j-1)}} \]

\[ s_g(\rho) = \sum_{(ij) \in E(v)} \log p(\rho_i, \rho_j) \]
INFERENCE

\[ s^*(\rho, I) = s(\rho, I) + \max_{x_i, \omega(\rho_i) \forall \rho_i \in C(\rho)} \left( \sum_{\rho_i \in C(\rho)} s^*(\rho_i, I) \right) \]

\[ \rho_S^{MAP} = \max_{x, \omega(\rho_S)} s^*(\rho_S, I) \]
\[
\begin{align*}
\max_{x_i, \omega(\rho_i) \forall \rho_i \in \rho_C} \left( \sum_{\rho_i \in \rho_C} s^*(\rho_i, I) \right) &= \\
B_{\rho_j}(\rho_i) &= \max_{\omega(\rho_j)} \left( \max_{x_j} \left( s^*(\rho_j) + \log \frac{p(\omega(\rho_i), \omega(\rho_j))}{p(\omega(\rho_j))d(p_j)-1} + \log p(\rho_i, \rho_j) + \sum_{\rho_k \in O(\rho_j)} B_{\rho_k}(\rho_j) \right) \right)
\end{align*}
\]
## PERFORMANCE

<table>
<thead>
<tr>
<th>Method</th>
<th>head</th>
<th>torso</th>
<th>u.leg</th>
<th>l.leg</th>
<th>u.arm</th>
<th>l.arm</th>
<th>hand</th>
<th>foot</th>
<th>avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method of Yang et al.</td>
<td>1.00</td>
<td>1.00</td>
<td>0.975</td>
<td>0.839</td>
<td>0.951</td>
<td>0.577</td>
<td></td>
<td></td>
<td>0.869</td>
</tr>
<tr>
<td>Our method</td>
<td>1.00</td>
<td>1.00</td>
<td>0.933</td>
<td>0.857</td>
<td>0.915</td>
<td>0.719</td>
<td>0.420</td>
<td>0.339</td>
<td>0.884</td>
</tr>
</tbody>
</table>

*Note: The table shows the PCP detection rate for different body parts under two methods.*

*Graph: PCP detection rate vs. detection threshold for various body parts under the method of Yang et al. (CVPR11) and our method.*
CONCLUSIONS

• Generative model for representing rich appearance of highly deformable objects

• Capture semantic relationships between neighboring part productions

• DP framework for computing exact inference
THANKS!