Building and Using a *Semantivisual* Image Hierarchy

Li-Jia Li, Chong Wang, Yongwhan Lim, David M. Blei, Li Fei-Fei

Computer Science Department, Stanford University
Computer Science Department, Princeton University

*Presented by Lingbo Li*

CVPR, 2010
Outline

Introduction

Building the Semantivisual Image Hierarchy
   A Hierarchical Model for both Image and Text
   Learning the Semantivisual Image Hierarchy
   A Semantivisual Image Hierarchy

Using the Semantivisual Image Hierarchy
   Hierarchical Annotation of Images
   Image Labeling
   Image Classification

Conclusion
Introduction

- two types of hierarchies:
  1) language-based hierarchy: no important visual information that connects images together
  2) low-level visual feature based hierarchy: no quantitative evaluation

Motivation

- Construct a semantivisual hierarchy built upon both semantic and visual information related to images
- Automatically organize images in a general-to-specific relationship
- Hierarchy is quantitatively evaluated by both human objects and end tasks (image classification and annotation)
A Hierarchical Model for both Image and Text

- Each image is decomposed into a set of over-segmented regions $R = [R_1, \ldots, R_r, \ldots, R_N]$
- Each of the $N$ regions is characterized by four features: color, texture, location and quantized SIFT histogram of the small patches within each region
- Each image is associated with its tags $W = [W_1, \ldots, W_\omega, \ldots, W_M]$
- Each image is associated with a path of the hierarchy
- Image regions can be assigned to different nodes of the path
Introduction

Schematic illustration
A Hierarchical Model for both Image and Text

Graphical model and the notations of the variables

Joint likelihood

\[
p(C, \theta, Z, R, S, W, \beta, \varphi | \alpha, \phi, \lambda) = \\
\prod_{c \in T} \prod_{j=1}^{4} p(\beta_{j,c} | \phi_{j}) p(\varphi_{c} | \lambda) \prod_{d=1}^{D} p(C_{d} | C_{1:d-1}) p(\theta_{d} | \alpha) \\
\prod_{r=1}^{N_{d}} p(Z_{d,r} | \theta_{d}) \prod_{j=1}^{4} p(R_{d,r,j} | C_{d}, Z_{d,r}, \beta) \\
\prod_{m=1}^{M_{d}} p(S_{d,w} | N_{d}) p(W_{w} | C_{d}, Z_{d}, S_{d,w}, \varphi),
\]
Learning the Semantivisual Image Hierarchy

▶ Goal of learning: the concept index \( Z \), the coupling variable \( S \) and the path \( C \)

Sampling concept index \( Z \)

▶ The conditional distribution of a concept index of a particular region depends on 1) the likelihood of the region appearance, 2) the likelihood of tags associated with this region and 3) the concept indices of the other regions in the same image-text pair

\[
p(Z_d,r = l | ) \propto p(Z_d,r = l | Z_d^{-r}, \alpha) \prod_{j=1}^{4} p(R_d,r,j | R_{-dr}, C_d, Z, \phi_j) \\
p(\{W_{d,\omega} : \omega \in S_r\} | W_{-d\omega:\omega \in S_r}, C_d, Z_d,r, \lambda) \\
= \prod_{j=1}^{4} \frac{n_{d,l,r,j}^{-dr} + \phi_j}{n_{C_d,l,r,j}^{-dr} + V_j \phi_j} \times \prod_{\omega \in S_r} \frac{n_{d,l,r,\omega}^{-d\omega} + \lambda}{n_{C_d,l,r,\omega}^{-d\omega} + U \lambda} \times \frac{n_{d,l}^{-r} + \alpha}{n_{d,r}^{-r} + L \alpha}
\]
Sampling coupling variable $S$

- The conditional distribution depends on the likelihood of tag
- The path assignment is fixed at this step

$$p(S_d, \omega = r | \cdot) \propto p(W_d, \omega | S_d, \omega = r, S^{-d\omega}, W^{-d\omega}, Z_d, C_d, \lambda)$$

$$= \frac{n^{-d\omega}_{C_d, Z_d, r, W_d, \omega} + \lambda}{n^{-d\omega}_{C_d, Z_d, r, \cdot, + U\lambda}}$$
Learning the Semantivisual Image Hierarchy

Sampling path C

- The path assignment of a new image-text pair depends on the previous arrangement of the hierarchy and the likelihood of the image-text pair

\[ p(C_d|\text{rest}) \propto p(R_d, W_d|R_{-d}, W_{-d}, Z, C, S)p(C_d|C_{-d}), \]

where \( p(C_d|C_{-d}) \) is the prior probability induced by nCRP and \( p(R_d, W_d|R_{-d}, W_{-d}, Z, C, S) \) is the likelihood,

\[ p(R_d, W_d|R_{-d}, W_{-d}, Z, C, S) \propto \prod_{w=1}^{M_d} \frac{n_{C_d,Z_d,S_d,w}^{-d} + W_{d,w} + \lambda}{n_{C_d,Z_d,S_d,w}^{-d} + U \lambda} \times \]

\[ \prod_{l=1}^{L} \prod_{j=1}^{4} \left( \frac{\Gamma(n_{C_d,l,j,v}^{-d} + V_j \phi_j)}{\prod_v \Gamma(n_{C_d,l,j,v}^{-d} + \phi_j)} \times \frac{\prod_v \Gamma(n_{C_d,l,j,v}^{-d} + n_{C_d,l,j,v}^d + \phi_j)}{\Gamma(n_{C_d,l,j,v}^{-d} + n_{C_d,l,j,v}^d + V_j \phi_j)} \right). \]

A Semantivisual Image Hierarchy

Implementation

- 4000 images and 538 tags across 40 image class from Flickr
- Each image is divided into small patches of $10 \times 10$ pixels, and a collection of over-segmented regions
- Each patch is assigned to one codeword in a codebook of 500 visual words obtained by K-means
- Obtain 4 region codebooks for color, location, texture and normalized SIFT histogram
- Initial the levels in a path according to ti-idf scores
- obtain a hierarchy of 121 nodes, 4 levels and 53 paths
A Semantivisual Image Hierarchy
"How meaningful is the path?" Exp.

<table>
<thead>
<tr>
<th>Choice</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>semantivisual hierarchy</strong></td>
<td>92 %</td>
</tr>
<tr>
<td>nCRP</td>
<td>70 %</td>
</tr>
</tbody>
</table>

"How meaningful is the hierarchy?" Exp.

<table>
<thead>
<tr>
<th>Choice</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>semantivisual hierarchy</strong></td>
<td>59 %</td>
</tr>
<tr>
<td>nCRP</td>
<td>50 %</td>
</tr>
<tr>
<td>Flickr</td>
<td>45 %</td>
</tr>
</tbody>
</table>
For an unannotated image $I$ with posterior path assignments samples $\mathcal{S}_C = \{C^1_I, C^2_I, \ldots, C^{|\mathcal{S}_C|}_I\}$, the probability of tag $W$ for level $l$:

$$p(W|I, \text{level} = l) \approx \frac{1}{|\mathcal{S}_C|} \sum_{i=1}^{|\mathcal{S}_C|} p(W|\hat{\phi}_i, C^i_I(l))$$
Using the Semantivisual Image Hierarchy

Image Labeling

- For a test image \( I \) and its posterior samples
  \( S_C = \{ C_1^I, C_2^I, \ldots, C_{|S_C|}^I \} \) and \( S_Z = \{ C_1^I, C_2^I, \ldots, C_{|S_Z|}^I \} \), the probability of tag \( W \) given the image \( I \):

  \[
p(W|I) \approx \frac{1}{|S_C|} \sum_{i=1}^{|S_C|} \sum_{l=1}^L p(W|\hat{\phi}_i, C_i^I(l)) p(l|Z_i^I)
  \]
Using the Semantivisual Image Hierarchy

Image Classification

<table>
<thead>
<tr>
<th>Model</th>
<th>Class 1</th>
<th>Class 2</th>
<th>Class 3</th>
<th>Class 4</th>
<th>Class 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours</td>
<td>Bride</td>
<td>Christmas</td>
<td>City</td>
<td>Dessert</td>
<td>Party</td>
</tr>
<tr>
<td>SPM</td>
<td>Building</td>
<td>Flower</td>
<td>Sushi</td>
<td>Child</td>
<td>Friends</td>
</tr>
<tr>
<td>SVM</td>
<td>London</td>
<td>Party</td>
<td>Sushi</td>
<td>Italy</td>
<td>Christmas</td>
</tr>
<tr>
<td>Bart et al.</td>
<td>Soccer</td>
<td>City</td>
<td>Present</td>
<td>Sun</td>
<td>Fruit</td>
</tr>
<tr>
<td>Corr-LDA</td>
<td>Friends</td>
<td>High School</td>
<td>Italia</td>
<td>Dinner</td>
<td>Cake</td>
</tr>
<tr>
<td>LDA</td>
<td>Italia</td>
<td>High School</td>
<td>Child</td>
<td>Party</td>
<td>Cake</td>
</tr>
</tbody>
</table>
Conclusion

- Construct the semantivisual hierarchy by using both images and their tags
- Quantitatively evaluate the quality of the hierarchy by human subjects
- Perform different applications