Object Co-occurrence Assisted Hierarchical Model for Scene Understanding

Xin Li and Yuhong Guo
Temple University
What is scene?

- A view of Real-world Environment
- Semantic Description of A Space
- Configuration of Objects
The Problem

- **Object Recognition**
  Assign a set of object labels (e.g. laptop, desk, cup) to an image.

- **Scene Recognition (Global Information)**
  Assign a set of semantic labels (e.g. office, street, coast) to an image.
Controversy of Cognition

• Low-level features capture the gist of a scene which is good enough

• Intermediate semantic representation is still necessary because of the gap between representation and goal
Existing Work

- **Direction I: Low-level Modeling (Change et. al. 03; Ulrich et. al. 00)**
  - Using low-level features extracted from images, e.g. color, texture, shape for representation
  - Classifier: SVM

- **Direction II: Semantic Modeling (Bosch et. al. 06; Xing et. al. 10; Li et. al. 05)**
  - Dealing with the gap between low-level representation and recognition goals.
  - Classifier: SVM, PGM
Our Solution

- Our proposed solution: Hierarchical Probabilistic Graphical Model
  - Preprocessing (Segmentation)
  - Low-level Representation (Color, Shape, Texture, Location)
  - High-level Representation (Contextual Information)
Framework

- 3-level Hierarchical Model: superpixel level, object level, and scene level
Framework

- 3-level Hierarchical Model: superpixel level, object level, and scene level
Framework

- 3-level Hierarchical Model: superpixel level, object level, and scene level
Framework

- 3-level Hierarchical Model: superpixel level, object level, and scene level
- The chain of binary node \( O \) captures relationship between different object categories
Framework

- 3-level Hierarchical Model: superpixel level, object level, and scene level
- The chain of binary node $O$ captures relationship between different object categories
Framework

• 3-level Hierarchical Model: superpixel level, object level, and scene level

• The chain of binary node O captures relationship between different object categories
Framework

- 3-level Hierarchical Model: superpixel level, object level, and scene level
- The chain of binary node $O$ captures relationship between different object categories
• 3-level Hierarchical Model: superpixel level, object level, and scene level

• The chain of binary node $O$ captures relationship between different object categories
Framework

- 3-level Hierarchical Model: superpixel level, object level, and scene level
- The chain of binary node $O$ captures relationship between different object categories

$O_1 \rightarrow O_2 \rightarrow \ldots \rightarrow O_{No}$
Framework

- 3-level Hierarchical Model: superpixel level, object level, and scene level

- The chain of binary node $O$ captures relationship between different object categories
• 3-level Hierarchical Model: superpixel level, object level, and scene level

• The chain of binary node $O$ captures relationship between different object categories

• Features:
  ‣ Region-based ($tR$ nodes): shape, location (T. Malisiewicz et. al., CVPR 2008), texture (average responses of filterbanks in each region), color (simply compute from histograms)
  ‣ Patch-based ($tX$ nodes): SIFT
Framework

- **Joint Probability**

$$P(C, O_1, O_2, \ldots, O_n, tO, tR, tX | \alpha, \beta, \gamma, \eta)$$

- **Supervised Learning on Scene Level**

$$P(C) \cdot P(O_1|C, \eta_1) \cdot \prod_{i=2}^{N_o} P(O_i|O_{i-1}, C, \eta_i)$$

- **Unsupervised Learning on Object Level**

$$\prod_{l=1}^{N_r} (P(tO_l|O_1, O_2, \ldots, O_n, \gamma) \cdot \prod_{k=1}^{N_r} p(tR_k|tO_l, \alpha_k) \cdot \prod_{m=1}^{N_p} P(tX_{lm}|tO_l, \beta))$$
Image Preprocessing

- Algorithm:

- Why do this?
  Scene content contains multiple objects which occupy multiple regions.
Training Method

- Collapsed Gibbs Sampling Algorithm

Sample latent variables $O_j$ and $tO$ nodes

- Objective function of low-level modeling ($tO$ nodes are sampled)

$$P(tR_{dn} | \overline{tR}_{dn}, tO_{dn}) \cdot P(tX_{dn} | \overline{tX}_{dn}, tO_{dn})$$

- Objective function high-level modeling ($O$ nodes are sampled)

$$P(O_{dj} | C_d, O_{d1}, O_{d2}, \ldots, O_{dN_o}, tO_d)$$
Inference

- For testing data only $tR$ and $tX$ nodes are available

$$P(C = c | tR, tX)$$

- Latent variables ($O_s, tO_s$) are integrated out

$$P(C = c | tR, tX) \propto \prod_{n=1}^{N_r} \sum_{\{O_1, O_2, \ldots, O_{N_o}\}} (P(O_1, O_2, \ldots, O_{N_o} | C = c) \cdot \sum_{o} P(tR_n | tO_n = o) \cdot P(tX_n | tO_n = o) \cdot P(tO_n = o | O_1, \ldots, O_{N_o}))$$

- Final Prediction:

$$c^* = \arg \max_{c \in C} P(C = c | tR, tX)$$
Dealing with Local Optima

- Gibbs Sampling cannot guarantee a global optima
- Training with different initial values of parameters will end up with different models
Dealing with Local Optima

- Gibbs Sampling cannot guarantee a global optima
- Training with different initial values of parameters will end up with different models

Solution:
Train 5 models with different random starts and output the prediction voted by most models
Experiment settings

• Dataset: LabelMe (Russell et. al. 08)
  ‣ 10 Scene Categories
  ‣ Number of Object categories ($N_o$) = 30

• 3 Baselines:
  ‣ our methods without ensemble procedure
  ‣ model without the chain structure
  ‣ Object Bank + SVM (Xing et. al. NIPS10) state of the art
Experiments & Results

- Annotation is achieved automatically
- Coarse object recognition improve the prediction performance
- Missing critical objects may lead to wrong scene classification
Experiment & Results

- Object detectors in OB are trained on LabelMe, ImageNet, Flickr, etc.
- Our average performance is as good as the state-of-the-art method.
- Chain Structure improve the accuracy by up to 0.3.
- Ensemble procedure increase the robustness of model.

<table>
<thead>
<tr>
<th>Method</th>
<th>Proposed</th>
<th>Proposed w/o Ensemble</th>
<th>Model w/o Chain</th>
<th>OB+SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>bathroom</td>
<td>0.765</td>
<td>0.604</td>
<td>0.565</td>
<td>0.938</td>
</tr>
<tr>
<td>bedroom</td>
<td>0.704</td>
<td>0.673</td>
<td>0.573</td>
<td>0.568</td>
</tr>
<tr>
<td>airport</td>
<td>0.676</td>
<td>0.638</td>
<td>0.584</td>
<td>0.459</td>
</tr>
<tr>
<td>coast</td>
<td>0.920</td>
<td>0.875</td>
<td>0.607</td>
<td>0.534</td>
</tr>
<tr>
<td>corridor</td>
<td>0.786</td>
<td>0.757</td>
<td>0.550</td>
<td>0.964</td>
</tr>
<tr>
<td>livingroom</td>
<td>0.471</td>
<td>0.464</td>
<td>0.447</td>
<td>0.765</td>
</tr>
<tr>
<td>office</td>
<td>0.938</td>
<td>0.822</td>
<td>0.675</td>
<td>0.938</td>
</tr>
<tr>
<td>park</td>
<td>0.660</td>
<td>0.749</td>
<td>0.630</td>
<td>0.849</td>
</tr>
<tr>
<td>speech</td>
<td>0.769</td>
<td>0.592</td>
<td>0.438</td>
<td>0.846</td>
</tr>
<tr>
<td>street</td>
<td>0.718</td>
<td>0.688</td>
<td>0.425</td>
<td>0.875</td>
</tr>
<tr>
<td>Average</td>
<td>0.741</td>
<td>0.686</td>
<td>0.549</td>
<td>0.774</td>
</tr>
</tbody>
</table>
A hierarchical PGM which can achieve automatic and implicit object annotation is presented

Object recognition is helpful for the task of scene recognition

Contextual information (i.e. object co-occurrence) encoded by probabilistic chain structure improve the performance of classifier

The issue of local optima is addressed effectively by simply using ensemble strategy
Thank You

Q & A