Move and the Robot will Learn: Vision-based Autonomous Learning of Object Models

Xiang Li and Mohan Sridharan

Stochastic Estimation and Autonomous Robotics Lab (SEARL)
Department of Computer Science
Texas Tech University

Nov 27, 2013
Object Recognition on Robots

Challenges:

• Difficult to obtain many labeled images for learning.
• Identify Region Of Interest (ROI) in image.
• Relevant features to extract from ROI to learn object model.
• Reliable and efficient implementation.

Learn Object Model
color, texture or shape

Test
Database for Learning

- Related Work: many images for learning
- Our Work: 3 – 8 images for learning

- Object Classification [Roman AR10] and Object Recognition [Luo ICRA11]
- A small number of images [Li CVIU07]
Object Model

• Existing algorithms use different visual cues:
  - Gradient Features (Texture) [Calonder ECCV10]
  - Color [Gevers PAMI04, Salas PR11]
  - Parts [Felzenszwalb PAMI10, Pedersoli CVPR11]
  - Context [Divvala CVPR09, Parikh PAMI12]

• Also use combinations of visual cues:
  [Li CVIU07, Gehler IJCV09, Ommer PAMI10]

• Computationally expensive, require many labeled samples, or do not exploit complementary strengths of visual cues.
Observations and Objectives

• Observations
  ➢ Many objects possess unique characteristics and motion patterns.
  ➢ Images encode information using appearance-based and contextual cues.
  ➢ Subset of domain objects relevant to any task, especially those that move.

• Assumptions
  ➢ Moving objects are interesting. Objects of interest are textured.
  ➢ Relative motion of objects is not fast.
  ➢ No sudden changes in viewpoint or scale.
Contributions

• Learn object models from a small (3 - 8) number of images; automatically extract labeled ROIs corresponding to moving objects.

• Build object models using the complementary strengths of different visual cues.

• Generative model for information fusion and energy minimization algorithm for reliable recognition.

• **Long term objective:** support incremental learning on robots, using sensor inputs and human feedback based on need and availability.
Identifying Moving Objects

- **Supervised Learning:**
  - Images with the labeled regions.
- **Unsupervised Learning:**
  - Images without any labeled regions.

- Track and cluster local image gradient features:
  - A short sequence of images (motion cue).

![Image of hallway with ROI at time t and t+1]
Object Model

- Consider a given ROI (detected automatically or provided manually).
- Use the complementary strengths of different visual cues.
SCV from Gradient Features

- The individual gradient features may not be unique.
- The spatial arrangement of local gradient features corresponding to a specific object is difficult to duplicate.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>-1</td>
<td>-</td>
</tr>
</tbody>
</table>

TABLE I: X-axis SCV

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>-1</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>-1</td>
<td>-1</td>
<td>-</td>
</tr>
</tbody>
</table>

TABLE II: Y-axis SCV
Connection Potentials

- Connection potential is computed as the color distribution of pixels between gradient features in the image ROI.

- Build an undirected graph of connection potentials to model the neighborhood relationships between gradient features.
Parts from Image Segments

- Graph-based segmentation of image ROI.

- Pixels within a part have similar values, while pixels in neighboring parts have dissimilar values.

- Considers the arrangement of object parts modeled as Gaussians.
Color-based Representation

- Computes color space distributions (PDFs) and model distance between every pair of PDFs as a Gaussian.

Second order color distribution statistics
Local Context from Image Segments

- Probabilistic (Gaussian) mixture models of pixels in regions neighboring the ROI.

- Model relative positions (on, under, beside) of these regions w.r.t. ROI.
Information Fusion For Recognition

• Recognize objects irrespective of whether they are stationary or moving. *Movement critical only during learning.*

• Candidate ROIs:
  ➢ Moving objects: same procedure used during learning.
  ➢ Stationary objects: match gradient features; use energy minimization to iteratively consider different ROI candidates.

• Generative Model:
  ➢ Probabilistically model dependencies between cues (see paper for details).
  ➢ Combines evidence provided by models based on individual visual cues.
Robot Platform and Datasets

- 1.6GHz Core2 Duo CPU. On-board computation.
- Two (640 x 480) cameras (monocular & stereo).
- Laser range finder. Wi-Fi.
- Use monocular camera for experiments.

- 20 object categories.
- Separate models for 60 subcategories, e.g., different books and boxes.

- Experiments used ~2000 images:
  - ~700 captured by robot; ~1300 from benchmark datasets (Pascal VOC2006, Caltech-256) to show applicability to different domains.
  - Images of stationary objects + sequences of objects in motion.

- Use 3-8 images for learning (~250 total), test on remaining images. Repeat experiment 10 times.

- For images from benchmark datasets, ROIs already available.
Classification Accuracy

- Computational efficiency: >=5 images per second.
- More efficient if optimization algorithms used for energy minimization and graph-based segmentation.
Comparison Results

• Our algorithm provides higher accuracy than any individual components or any four of the components.

• Provides better balance of reliability and computational efficiency in comparison with other algorithms for learning object models and recognizing objects.

• See paper for quantitative results of comparison with popular algorithms, including those based on gradient features.
Conclusions + Future Work

• Conclusions:
  – Automatically identified interesting objects based on motion cues.
  – Automatically and efficiently learned object models from small number of images, exploiting complementary strengths of different visual cues.
  – Used generative model and energy minimization algorithms to reliably and efficiently recognize the learned objects in images of novel scenes.

• Future Work:
  – Image sequences with multiple moving objects.
  – Include depth information to disambiguate occluded objects.
  – Explore efficient energy minimization and sampling-based algorithms.
  – Long-term goal: enable robots to make best use of human and sensor inputs based on need and availability.
That’s all folks 😊