Learning Visual Object Models on a Robot Using Context and Appearance Cues

(Extended Abstract)

Xiang Li
Dept. of Computer Science
Texas Tech University
Lubbock, TX
xiang.li@ttu.edu

Mohan Sridharan
Dept. of Computer Science
Texas Tech University
Lubbock, TX
mohan.sridharan@ttu.edu

Catie Meador
Dept. of Computer Science
Swarthmore College
Swarthmore, PA
catie.meador@gmail.com

ABSTRACT

Visual object recognition is a key challenge to the deployment of robots in domains characterized by partial observability and unforeseen changes. Sophisticated algorithms developed for modeling and recognizing objects using different visual cues [3, 4] are computationally expensive, sensitive to changes in object configurations and environmental factors, and require many training samples and accurate domain knowledge to learn object models, making it difficult for robots to reliably and efficiently model and recognize objects. These challenges are partially offset by the fact that many objects possess unique characteristics (e.g., color and shape) and motion patterns, although these characteristics and patterns are not known in advance and may change over time. Furthermore, only a subset of domain objects are relevant to any given task and a variety of cues can be extracted from images to represent objects. This paper presents an algorithm that enables robots to identify a set of *interesting* objects, using appearance-based and contextual cues extracted from a small number of images to efficiently learn models of these objects. Robots learn the domain map and consider objects that move to be interesting, using motion cues to identify the corresponding image regions. Object models learned automatically from these regions consist of spatial arrangement of gradient features, graph-based models of neighborhoods of gradient features, parts-based models of image segments, color distributions, and mixture models of local context. The learned models are used for object recognition in novel scenes based on energy minimization and a generative model for information fusion. All algorithms are evaluated on wheeled robots in indoor and outdoor domains.

Categories and Subject Descriptors

I.4.8 [Scene Analysis]: Object recognition; I.2.9 [Robotics]: Sensors; I.5.4 [Applications]: Computer vision.

Keywords

Visual learning; Object Recognition; Mobile robots

1. PROBLEM FORMULATION

Robots use range data to learn and revise the domain map. Based on the observation that characteristic features of an object have similar relative motion between consecutive images, our prior work enabled robots to track local gradient features in short image sequences, identifying regions of interest (ROIs) corresponding to moving objects by clustering features with similar relative motion [2]. Color distributions and gradients extracted from these ROIs were used to learn object models. In this paper, the object model has five components, as shown in Figure 1, that exploit the complementary properties of local, global, temporal and contextual visual cues.

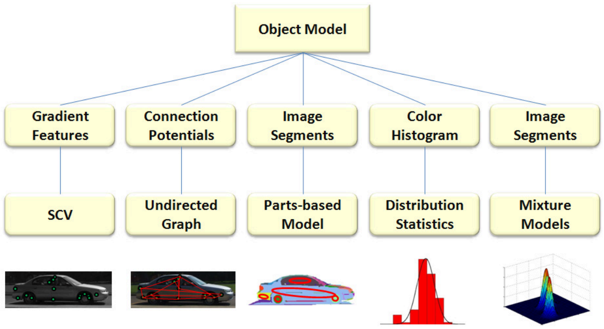

Figure 1: Learned model uses contextual and appearance-based cues to characterize objects.

Spatial Coherence of Gradient Features: To model an object in an image ROI, the first component includes local image gradient features extracted from the ROI and a *spatial coherence vector* that models the relative spatial arrangement of these features [2].

Graph-based Model of Connection Potentials: The *connection potential* between two gradient feature keypoints is computed as the color distribution of pixels on the line joining the keypoints. The second component of the object model includes connection potentials between gradient features in the ROI and a undirected graph of (local) neighborhood of these connection potentials.

Parts-based Models of Image Segments: A graph-based color segmentation algorithm [1] is used to extract segments from the ROI such that values of pixels within any segment are similar and significantly different from pixels in neighboring segments. The third component consists of these image segments, a parts-based model of these segments (with Gaussian parts) and measures of similarity (dissimilarity) within (between) parts.

Color Distribution Statistics: For any candidate ROI, the fourth component of the object model consists of color-space pdfs (extracted from images in the learning set) and a distribution of distances between these pdfs [2].

Gaussian Mixture Model of Context: The fifth component models the object's *local context*. For each image segment that shares a boundary with the ROI, relative spatial arrangement with respect to the ROI is used to assign labels "on", "under" and "beside". Pixels in segments with the same label are used to learn Gaussian mixture

Appears in: *Proceedings of the 12th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2013),* Ito, Jonker, Gini, and Shehory (eds.), May 6–10, 2013, Saint Paul, Minnesota, USA.

	Box	Car	Human	Robot	Book	Airplane	Bus	Motorbike
Box	**0.958**	0	0.017	0.025	0	0	0	0
Car	0.010	**0.927**	0	0.021	0	0	0	0.042
Human	0.080	0.024	**0.820**	0.060	0.016	0	0	0
Robot	0.027	0	0.042	**0.899**	0.027	0	0	0.005
Book	0.016	0	0	0.042	**0.942**	0	0	0
Airplane	0.029	0.051	0	0.023	0.009	**0.888**	0	0
Bus	0	0.072	0	0	0	0	**0.856**	0.072
Motorbike	0	0.073	0	0.010	0.016	0	0.062	**0.839**

Table 1: Object recognition accuracy averaged over different models (i.e., subcategories) in each object category.

models (GMMs) in normalized HSV color space. The object model includes the GMMs, and their relative positions and sizes.

The learned object models are used for object recognition in novel scenes, irrespective of whether the objects are stationary or moving. For any test image, robots perform object recognition by identifying ROI(s) and computing the probability of occurrence of each learned object in the ROI(s). For image sequences with moving objects, ROIs are identified using the same approach used to identify ROIs during learning. For individual snapshots of objects, the iterated conditional modes (ICM) energy minimization algorithm is used to iteratively select candidate ROIs. For any candidate ROI and each learned object model, the robot computes the probability of occurrence of the corresponding object in the ROI based on each component of the object model. Probabilistic generative models that capture the conditional relationships between components of the learned model are then used to merge these probabilities and compute the net probability of occurrence of the corresponding object in the ROI. Robots are thus able to exploit the complementary properties of different visual cues for reliable and efficient object recognition in different scenes.

2. EXPERIMENTAL RESULTS

The test platform was a wheeled robot equipped with cameras that provide 640 × 480 images. Data from range finders were used to learn the domain map. Although the robot has Wi-Fi capability, all experiments were performed using an on-board 2GHz processor and 1GB RAM. Experimental trials evaluated the robot's ability to learn models of interesting objects from a small set of images using appearance-based and contextual visual cues, using these models to reliably and efficiently recognize objects in novel scenes.

Robots learned 30 different object models over eight different object categories: human, box, airplane, book, car, motorbike, bus and humanoid robot. Since it is a challenge to obtain an image dataset of objects with well-defined motion, experiments were conducted over $\approx$ 1400 images, $\approx$ 700 of which were captured by the robot in indoor and outdoor environments. To establish applicability to different domains, images of airplanes, motorbikes and buses (and some cars) were chosen from the *Pascal VOC2006* benchmark dataset to obtain $\approx$ 700 images. To make learning challenging, each object model was learned autonomously using $\approx 3-5$ images, with $\approx$ 150 images used for learning all object models; the remaining images were used for evaluation. The images used for learning and recognition were chosen randomly in repeated trials. The robot is able to process $3-5$ frames/second to identify moving objects, learn models and recognize objects in novel scenes, while performing other tasks such as path planning for navigation.

Table 1 shows the object recognition accuracy for different object categories, averaged over different subcategories in each category. Accurate recognition requires an object in a test image to be matched with the correct subcategory. The robot exploits complementary properties of visual cues to reliably learn object models and recognize objects in novel scenes—the combination of cues provides higher accuracy than any single component. Most classification errors occur when an insufficient number of (test) image features are matched with the learned object models as a result of motion blur or a large difference in scale or viewpoint. Incremental revision of the learned object models helps eliminate many of these errors. Figure 2 shows examples of object recognition in test images—robots can reliably and efficiently recognize objects in cluttered backgrounds, and recognize multiple objects or multiple instances of the same object.

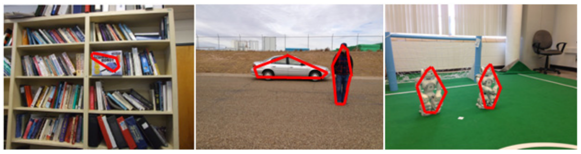

Figure 2: Illustrative examples of using the proposed algorithm to recognize one or more objects in test images.

3. CONCLUSIONS

This paper described an approach that enables mobile robots to use the complementary properties of appearance-based and contextual visual cues to identify interesting objects and learn representative models from a small set of images. These object models are used for reliable and efficient object recognition in novel scenes. Future work will consider additional visual cues and fully integrate learning with planning and collaboration [5].

ACKNOWLEDGMENT

This work was supported in part by the ONR Science of Autonomy award N00014-09-1-0658.

4. REFERENCES

[1] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient Graph-Based Image Segmentation. *International Journal of Computer Vision*, 59(2):167–181, 2004.

[2] Xiang Li and Mohan Sridharan. Autonomous Learning of Object Models on a Mobile Robot using Visual Cues. In *International Conference on Robotics and Automation*, 2011.

[3] K. Mikolajczyk and C. Schmid. Scale and Affine Invariant Interest Point Detectors. *International Journal of Computer Vision*, 60(1):63–86, 2004.

[4] J. Porway and S. C. Zhu. C4: Computing Multiple Solutions in Graphical Models by Cluster Sampling. *Pattern Analysis and Machine Intelligence*, 33(9):1713–1727, 2011.

[5] Mohan Sridharan. Integrating Visual Learning and Hierarchical Planning for Autonomy in Human-Robot Collaboration. In *AAAI Spring Symposium Series, Designing Intelligent Robots: Reintegrating AI II*, March 2013.