Recognition by Association via Learning Per-exemplar Distances

Tomasz Malisiewicz and Alexei A. Efros

Carnegie Mellon University

THE ROBOTICS INSTITUTE CMU

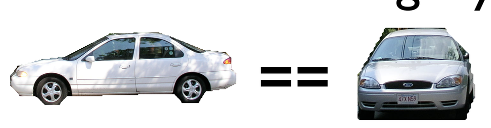
Association vs. Categorization

Goal: Given unlabeled image, recognize objects inside the image by associating generated segments with previously seen object exemplars (see last Figure)

vs.

Categorization

Observation: Categorization is difficult since visually dissimilar inputs need to be mapped to the same category



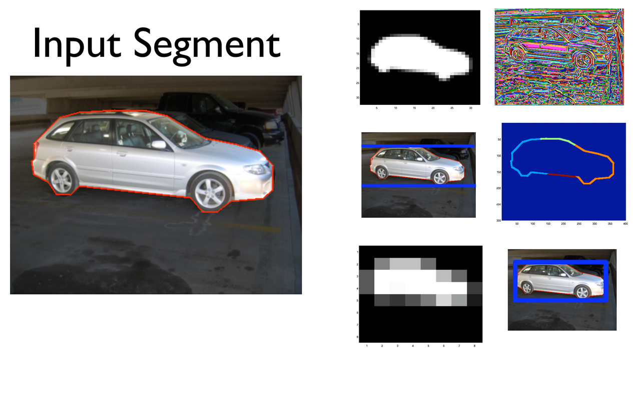
Exemplar representation

Background: Exemplar Theory from Psychology (Medin & Schaffer 1978, Nosofsky 1986, Krushke 1992) states that categories are represented in terms of remembered objects. When looking at new object, **similarity between all exemplars** is computed.

13,905 objects from LabelMe w/ 171 unique 'labels'

Similarity Occurs at Different Levels

Idea: Represent each exemplar with features that encode shape, color, texture, and absolute position



Feature Type	Feature Name	Dimension
Shape	Centered Mask BB Extent Pixel Area	32x32=1024 2 1
Texture	Right Boundary Tex-Hist Top Boundary Tex-Hist Left Boundary Tex-Hist Bot Boundary Tex-Hist Interior Tex-Hist	100 100 100 100 100
Color	Mean Color Color std Color Histogram	3 3 33
Position	Absolute Mask Top Pixel Height Bottom Pixel Height	8x8=64 1 1

Measuring Object Similarity

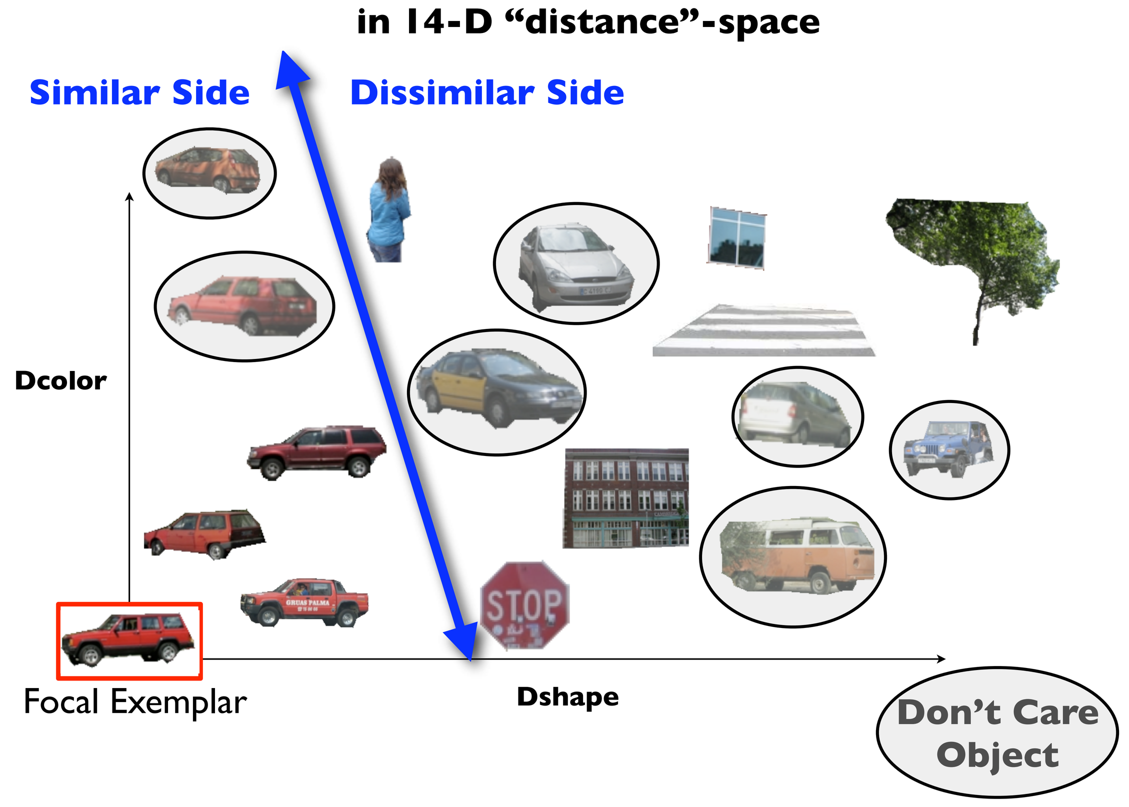
Approach: Measure L_2 distance between corresponding features to obtain **Elementary Distances**, then combine them using positive weights (a.k.a distance function)

$$D_e(z) = \mathbf{w}_e \cdot \mathbf{d}_{ez}$$

Distance Function Learning

Goal: Learn a different distance function **per-exemplar**; distance functions are learned **independently**

Distance function == linear decision boundary in 14-D "distance"-space



Approach: Iterative Learning Algorithm

$$\{\mathbf{w}^*, \boldsymbol{\alpha}^*\} = \operatorname{argmin}_{\mathbf{w},\boldsymbol{\alpha}} f(\mathbf{w}, \boldsymbol{\alpha})$$

$$f(\mathbf{w}, \boldsymbol{\alpha}) = \sum_{i \in C} \alpha_i L(-\mathbf{w} \cdot \mathbf{d}_i) + \sum_{i \notin C} L(\mathbf{w} \cdot \mathbf{d}_i)$$

Start with initial distance function (Tex-hist distance)

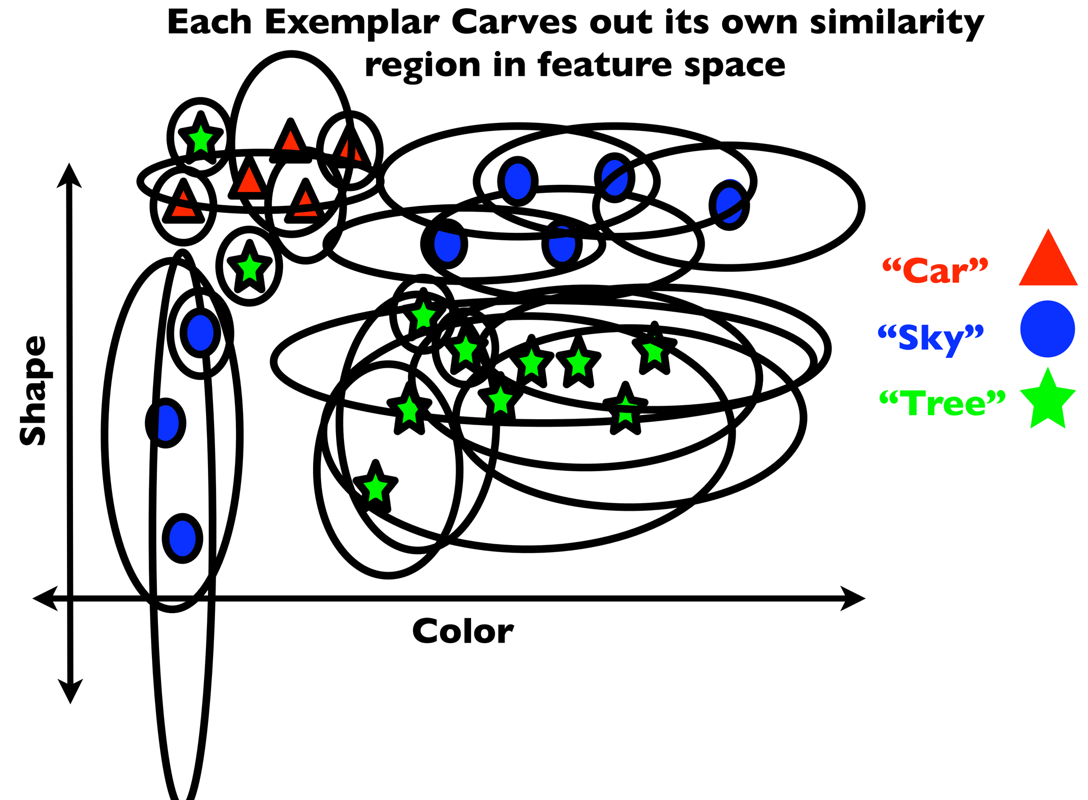
1.) Set K=10 closest exemplars with same label as "similar," other exemplars with same label as "don't care" and all other exemplars as "dissimilar"

2.) Learn new Distance Function by learning a **linear SVM** (Frome 2006)

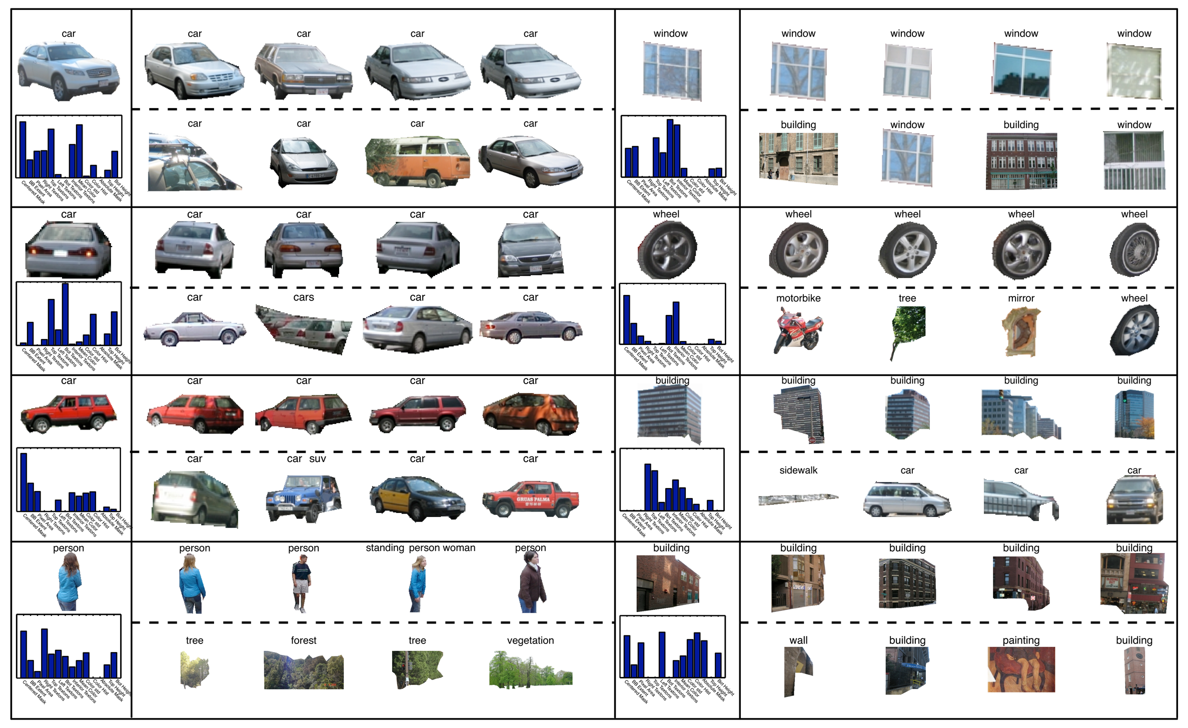
3.) If distance function changed, go to step 1

4.) Scale Distance Function so **D<1** means similar

Each Exemplar Carves out its own similarity region in feature space



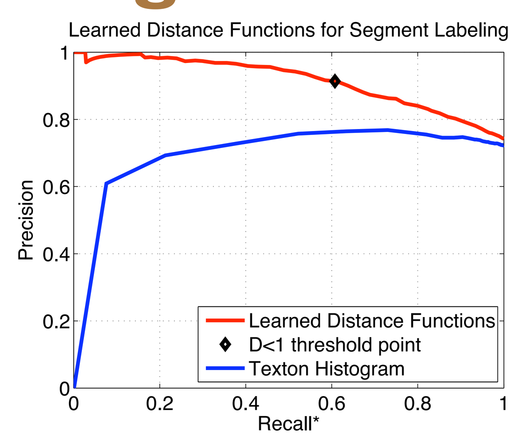
Visualizing Distance Functions



For each exemplar: top row shows 4 most similar exemplars after learning, bottom row shows 4 most similar exemplars w.r.t. tex-hist

Segment Labeling Task

Evaluate: Given perfect segment , determine object identity with single nearest neighbor

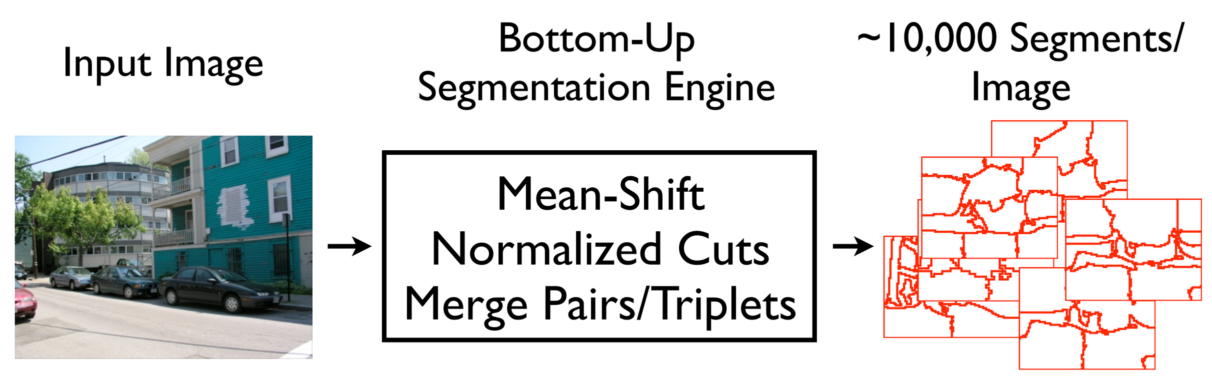


Recognition in Real Images

Problem: Objects are never presented one at a time, they are embedded inside images! If we only knew which pixels belonged to separate objects...

Multiple Segmentations

Approach: Generate **multiple segmentations** per image (Hoiem 2005, Russell 2006) and also consider pairs/triplets of contiguous segments (Malisiewicz 2007)

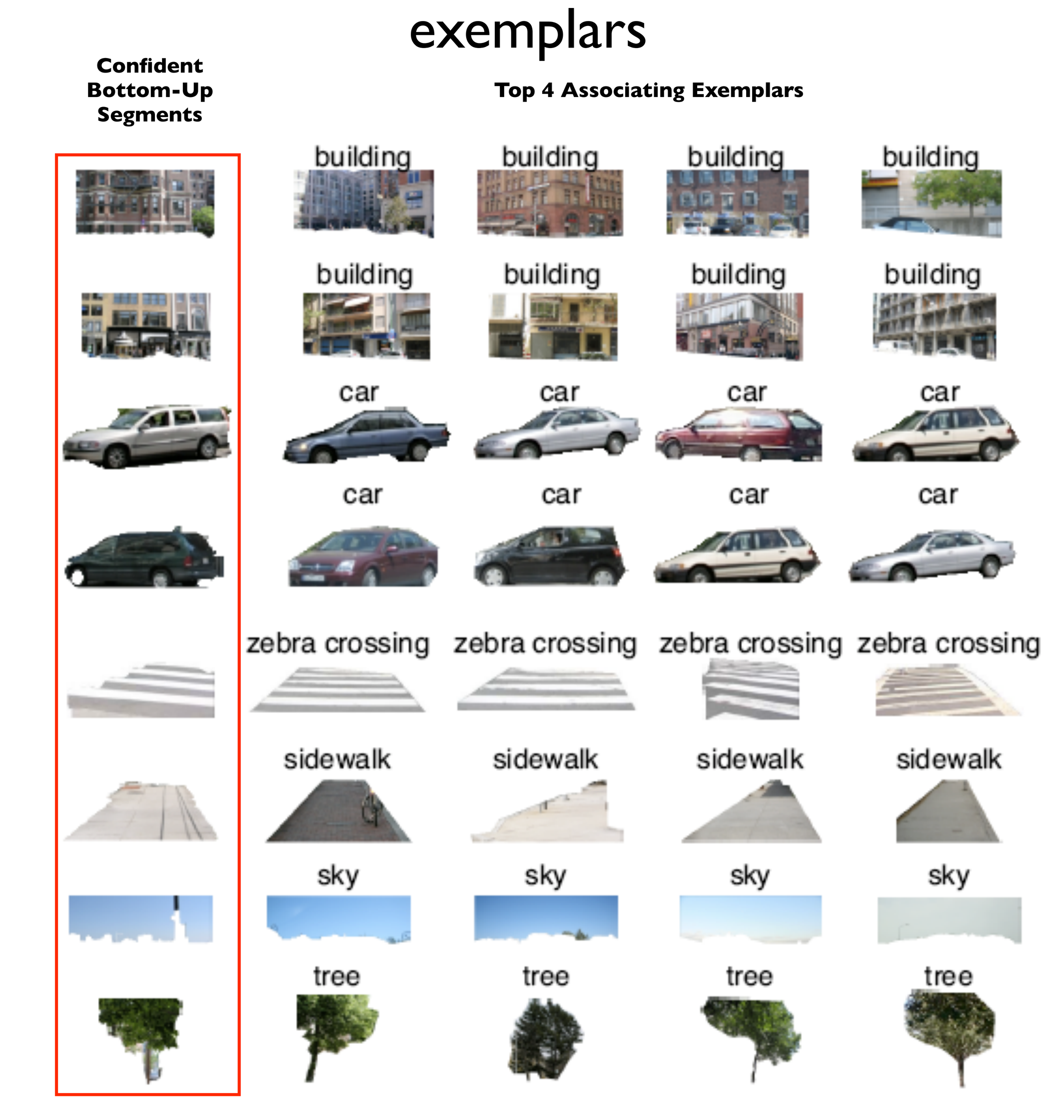


Approach: Create **associations** between bottom-up segments and object exemplars using distance functions; each distance function makes a separate binary "similar" or "dissimilar" decision for each input segment

Results

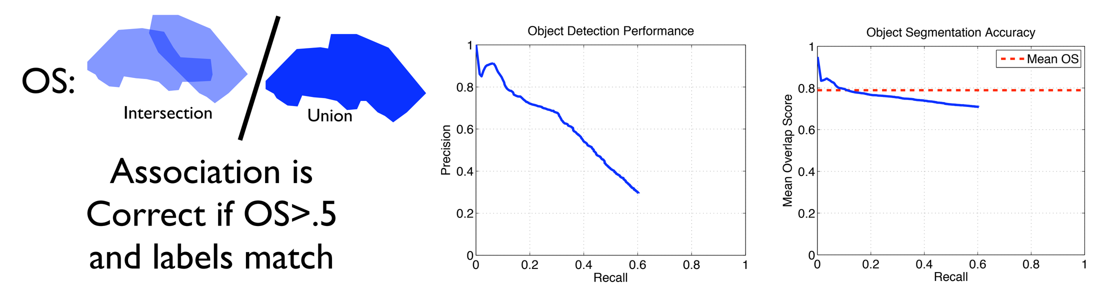
Test-set: 159 Outdoor Images from single folder of LabelMe

Evaluate: Recognition-Based Object Segmentation; each generated object "hypothesis" is a bottom-up segment and its list of associating exemplars



Idea: Association confidence score favors more associations and smaller distances; we vary this threshold to look at precision-recall

$$s(S,E) = 1/\sum_{e \in E} \frac{1}{D_e(S)}$$



Our Contributions

1.) Posing Recognition as Association

2.) Learning Object Similarity Per Exemplar

3.) Recognition-Based Object Segmentation

Toward Image Parsing

Greedily add most confident association while removing inconsistent (OS>.5) associations