Context in Recognition

Adrian Quark

March 27, 2008

Questions to Answer

This is a very broad topic.

- What is context?
- How do humans use context for recognition?
- How can computers use context for recognition?

Context in Recognition
Adrian Quark
March 27, 2008

1. Note pages are interleaved with slides. These notes cover some of the verbal content of the talk.

Questions to Answer
This is a very broad topic.
- What is context?
- How do humans use context for recognition?
- How can computers use context for recognition?

...

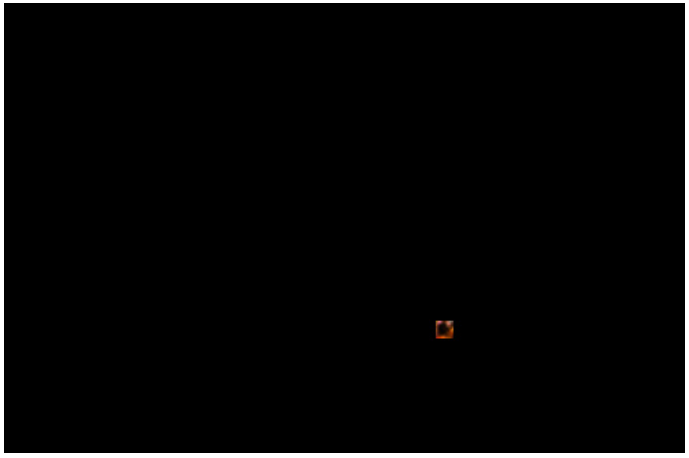
Outline

Outline
1. Introduction
2. Humans Use Context
3. Spatial Context
   Contextual Priming
   Spatial Hierarchies
   Scene Geometry
4. Temporal Context
   Place Recognition
5. Semantic Context
   Semantic Hierarchical Classifier
   Semantic Segmentation
   Semantic Agreement
6. Conclusion

...

What's the problem?

Most object recognition approaches are local.



"Kowloon", by * Toshio * on Flickr.com

What's the problem?
Most object recognition approaches are local.

...

What's the problem?

See how much information we threw away? That's context.

"Kowloon", by * Toshio * on Flickr.com

What's the problem?
See how much information we threw away? That's context.

...

What is visual context?

Approximate definition: any information not directly attributable to the foreground object. [Hoiem, 2004]
What can we infer from this definition?

- Context is open-ended
- Context is probabilistic
- Contextual relationships are learned
- Context is recursive

What is visual context?
Approximate definition: any information not directly attributable to the foreground object. [Hoiem, 2004]
What can we infer from this definition?
- Context is open-ended
- Context is probabilistic
- Contextual relationships are learned
- Context is recursive

1. Foreground object = object of interest
2. Anything can be context, so we have to choose wisely.
3. Usually context only *implies* something about the foreground object.
4. Learned assumptions and relationships are how we make use of context.
5. Elements of a scene can act both as background (context) and foreground (objects), so that as objects are recognized they can provide further context to recognize other objects, thus allowing our knowledge of a scene to reinforce itself.

What is context good for?

All aspects of recognition:

- *Identity: what is it?*
- *Location: where can I look to find it?*
- Relevance: how important is it?
- Role: what does it mean?

Focus on the first two.

What is context good for?
All aspects of recognition:
- *Identity: what is it?*
- *Location: where can I look to find it?*
- Relevance: how important is it?
- Role: what does it mean?

Focus on the first two.

…

Types of context

In order of sophistication.

- spatial
- temporal
- semantic

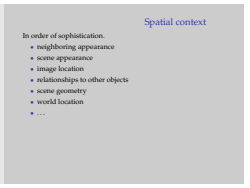
Types of context
In order of sophistication.
- spatial
- temporal
- semantic

1. Spatial = relationships in the image or 3D space, such as objects that tend to occur together at certain relative scales and positions.
2. Temporal = relationships in time, including knowledge about historical events and user behavioural patterns.
3. Semantic = Everything else.

Spatial context

In order of sophistication.

- neighboring appearance
- scene appearance
- image location
- relationships to other objects
- scene geometry
- world location
- …



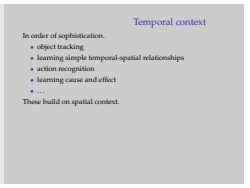
1. It might help to think of these in terms of absolute and relative relationships, but that's mostly a question of frame of reference.
2. Nearby appearance = Localized but still contextual information: faces are usually above bodies.
3. Scene appearance = the forest is usually green, the city is usually gray. Cars are found in the city, not the forest.
4. Image location = the sky is almost always towards the top of the image.
5. Surrounding objects = silverware is found near a plate; a computer is found on a desk.
6. Geometric location = people are on the sidewalk; this is more reliable than image location, but also harder to infer.
7. World location = certain objects may be in certain rooms, or certain landmarks at certain addresses; this is the hardest to infer.
8. Three broad categories: 2D appearance relationships, 2D object relationships, and 3D scene structure

Temporal context

In order of sophistication.

- object tracking
- learning simple temporal-spatial relationships
- action recognition
- learning cause and effect
- …

These build on spatial context.



1. This area has been explored but is not usually thought of in terms of context.
2. Ex: Face tracking to recover hard-to-detect views
3. Ex: place recognition combined with model of motion
4. Ex: abnormal event recognition
5. Maybe cause-and-effect is semantic context.

Semantic context

Everything else!

- associated text
- general concept associations
- model of user
- domain knowledge
- cultural knowledge
- …

These build on spatial and temporal context.

Semantic context
Everything else!
• associated text
• general concept associations
• model of user
• domain knowledge
• cultural knowledge
• …
These build on spatial and temporal context.

1. Ex: Names and Faces in the News
2. Ex: semantic hierarchies, semantic distance
3. Ex: Amazon book recommendations
4. Are flowers a symbol of romance (at a wedding) or grief (at a funeral).

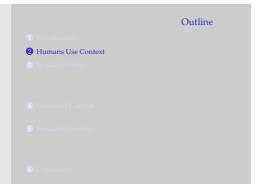
References

- Human Use of Context: The Role of Context in Object Recognition [Oliva and Torralba, 2007]
- Spatial Context:
  - Contextual Priming for Object Detection [Torralba, 2003]
  - Unsupervised Learning of Hierarchical Semantics of Objects (HSOs) [Parikh and Chen, 2007]
  - Putting Objects in Perspective [Hoiem et al, 2006]
- Temporal Context: Context-based vision system for place and object recognition [Torralba et al, 2003]
- Semantic Context
  - Semantic Hierarchies for Visual Object Recognition [Marszałek and Schmid, 2007]
  - Object Boundary Detection in Images using a Semantic Ontology [Hoogs and Collins, 2006]
  - Objects in Context [Rabinovich et al, 2007]

References
• Human Use of Context: The Role of Context in Object Recognition [Oliva and Torralba, 2007]
• Spatial Context:
  • Contextual Priming for Object Detection [Torralba, 2003]
  • Unsupervised Learning of Hierarchical Semantics of Objects (HSOs) [Parikh and Chen, 2007]
  • Putting Objects in Perspective [Hoiem et al, 2006]
• Temporal Context: Context-based vision system for place and object recognition [Torralba et al, 2003]
• Semantic Context
  • Semantic Hierarchies for Visual Object Recognition [Marszałek and Schmid, 2007]
  • Object Boundary Detection in Images using a Semantic Ontology [Hoogs and Collins, 2006]
  • Objects in Context [Rabinovich et al, 2007]

1. Main references for this talk, others are included in the appendix.

Outline



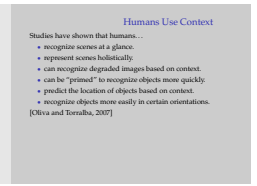
...

Humans Use Context

Studies have shown that humans...

- recognize scenes at a glance.
- represent scenes holistically.
- can recognize degraded images based on context.
- can be "primed" to recognize objects more quickly.
- predict the location of objects based on context.
- recognize objects more easily in certain orientations.

[Oliva and Torralba, 2007]



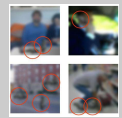
...

1. Scenes can be recognized without eye scanning or using foveal (detailed) vision
2. We remember statistical properties of scenes and object groups better than details
3. Priming = showing picture of related scene first
4. We can quickly learn spatial relationships between arbitrary shapes
5. When recognizing letter forms there is evidence that people mentally rotate them; it takes 2x as long to recognize an upside-down L as a sideways one

Example: Disambiguation

Context helps us disambiguate in presence of noise.

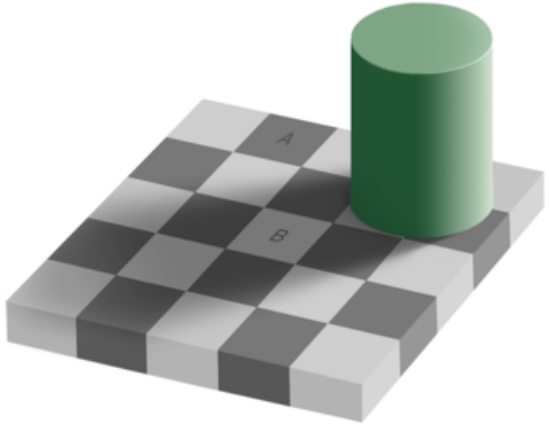
[Murphy et al, 2005]



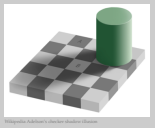
1. The circled objects all have the same appearance.

Example: Disambiguation

. . . or other sources of appearance variation: is square B white?



Wikipedia Adelson's checker shadow illusion

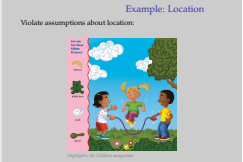


1. Squares A and B are the same color

Example: Location

Violate assumptions about location:

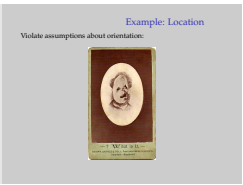
Highlights: for Children magazine



...

Example: Location

Violate assumptions about orientation:



1. It's a puppy (rotated 90 degrees).

Example: Scale

Violate assumptions about scene geometry:

Unknown source

1. Ames' Room

Outline

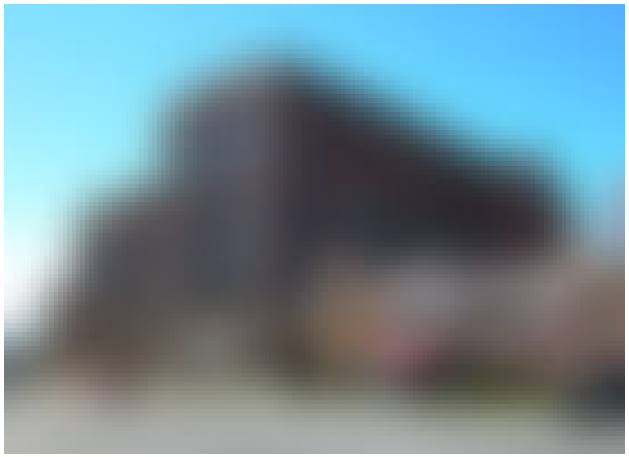
...

Is there a car?

Kerry Kelly 2006, “Beech-Maple Forest on Pierce Stocking Drive”

1. I blurred this image so it couldn’t be recognized
2. Color cues still suggest no car

Where are the cars?

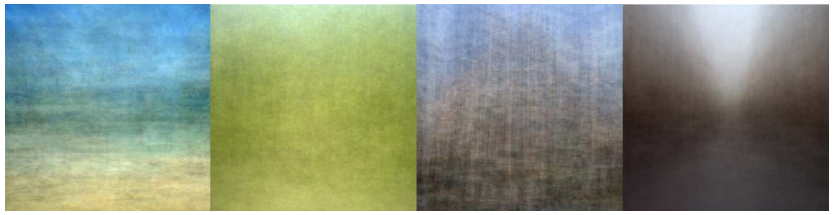


Nebraska State Historical Society, “K Street Facility”

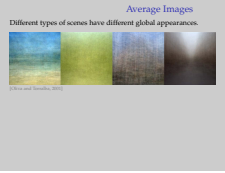
1. People predict cars between buildings and the street.

Average Images

Different types of scenes have different global appearances.



[Oliva and Torralba, 2001]

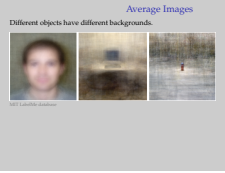


1. Scenes, not aligned or scaled: beach, forest, buildings, street.

Average Images

Different objects have different backgrounds.

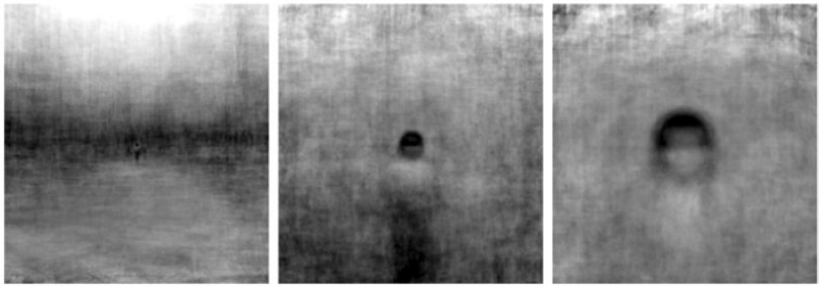
MIT LabelMe database



1. Objects from the LabelMe database aligned and scaled: face, computer, fire hydrant.

Average Images

Different object scales have different backgrounds.



[Torralba, 2003]

Average Images
Different object scales have different backgrounds.

1. Faces aligned at three different scales: small, medium, large.

Context Challenge

How far can you go before running an object detector?

- Object detection is hard.
- Chicken-and-egg problem: context recognition needs to be simpler than object recognition.
- Global scene information is useful.

Context Challenge
How far can you go before running an object detector?
- Object detection is hard.
- Chicken-and-egg problem: context recognition needs to be simpler than object recognition.
- Global scene information is useful.

1. Challenge set by Torralba to motivate Contextual Priming.

Contextual Priming [Torralba, 2003]

- Intuition: holistic image features are predictive of object identity, location, and scale.
- Probabilistic model: $P(o, \mathbf{x}, \sigma|\mathbf{v})$: the probability of an object o at position $\mathbf{x}$ and scale σ given image features $\mathbf{v}$.
- Local evidence: $P(o, \mathbf{x}, \sigma|\mathbf{v_L})$
- Contextual evidence: $P(o, \mathbf{x}, \sigma|\mathbf{v_C})$
- Bayes' rule lets us treat these separately: $P(o, \mathbf{x}, \sigma|\mathbf{v_C}) = P(\sigma|\mathbf{x}, o, \mathbf{v_C})P(\mathbf{x}|o, \mathbf{v_C})P(o|\mathbf{v_C})$
- ...and learn them from examples: $P(o|\mathbf{v_C}) = \frac{P(\mathbf{v_C}|o)P(o)}{P(\mathbf{v_C})}$

Contextual Priming [Torralba, 2003]
- Intuition: holistic image features are predictive of object identity, location, and scale.
- Probabilistic model: $P(o, \mathbf{x}, \sigma|\mathbf{v})$: the probability of an object o at position $\mathbf{x}$ and scale σ given image features $\mathbf{v}$.
- Local evidence: $P(o, \mathbf{x}, \sigma|\mathbf{v_L})$
- Contextual evidence: $P(o, \mathbf{x}, \sigma|\mathbf{v_C})$
- Bayes' rule lets us treat these separately: $P(o, \mathbf{x}, \sigma|\mathbf{v_C}) = P(\sigma|\mathbf{x}, o, \mathbf{v_C})P(\mathbf{x}|o, \mathbf{v_C})P(o|\mathbf{v_C})$
- ...and learn them from examples: $P(o|\mathbf{v_C}) = \frac{P(\mathbf{v_C}|o)P(o)}{P(\mathbf{v_C})}$

1. Most recognition approaches use only local evidence.
2. Torralba's contribution is incorporating contextual evidence.

Context representation

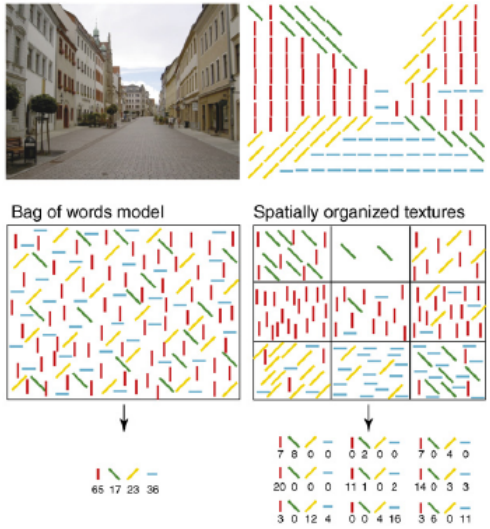
What background information is relevant?

- Statistics of structural elements
- Spatial organization
- Color distribution

Context representation
What background information is relevant?
- Statistics of structural elements
- Spatial organization
- Color distribution

1. Previous studies have shown that these properties are relevant for discrimination

Scene “gist”



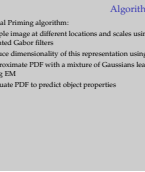
[Oliva and Torralba, 2007]

1. A compromise between bag-of-words and part-based models.

Algorithm

Contextual Priming algorithm:

1. Sample image at different locations and scales using oriented Gabor filters
2. Reduce dimensionality of this representation using PCA
3. Approximate PDF with a mixture of Gaussians learned using EM
4. Evaluate PDF to predict object properties



1. Joseph will cover this algorithm in more detail.
2. Essentially these same steps are used to learn identity, location, and scale.

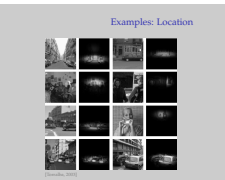
Examples: Identity

[Torralba, 2003]

1. Images ordered by probability that they contain an object (people versus chairs)

Examples: Location

[Torralba, 2003]



1. Locations likely to contain heads

Examples: Scale

[Torralba, 2003]

1. Scale is conditioned on global features and object type (but *not* object location)
2. Top row is heads, bottom row cars

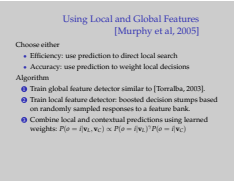
Using Local and Global Features [Murphy et al, 2005]

Choose either

- Efficiency: use prediction to direct local search
- Accuracy: use prediction to weight local decisions

Algorithm

1. Train global feature detector similar to [Torralba, 2003].
2. Train local feature detector: boosted decision stumps based on randomly sampled responses to a feature bank.
3. Combine local and contextual predictions using learned weights: $P(o = i|\mathbf{v}_L, \mathbf{v}_C) \propto P(o = i|\mathbf{v}_L)^{\gamma} P(o = i|\mathbf{v}_C)$



1. The real benefit of global context comes in supplementing local detection.
2. Murphy et al's global feature model is not identical to Torralba's but very similar. It uses steerable pyramids instead of Gabor filters and mixture density networks instead of gaussian mixtures.
3. Discriminative model does not require that local and contextual features are independent, but combination weight γ is fixed and must be learned offline.

Demo Movie

Demo movie…

Demo Movie
Demo movie…

The movie is not provided with the printable version of the presentation.

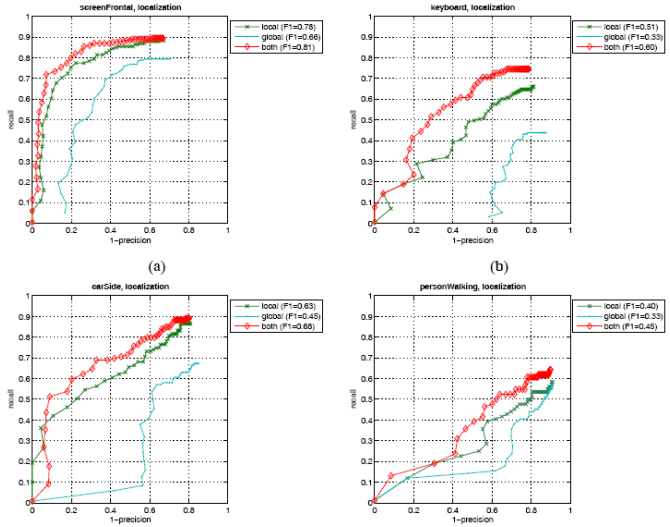
Examples: Identity and Location

[Murphy et al, 2005]

Examples: Identity and Location

1. The authors did not provide any examples of false positives.

Results: Identity and Location



[Murphy et al, 2005]

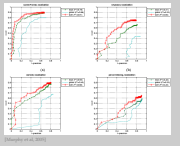
Contextual Priming: Conclusion

Good

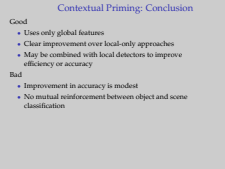
- Uses only global features
- Clear improvement over local-only approaches
- May be combined with local detectors to improve efficiency or accuracy

Bad

- Improvement in accuracy is modest
- No mutual reinforcement between object and scene classification

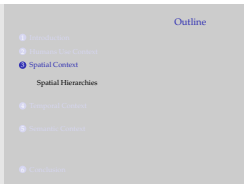


...



...

Outline



...

Object Relationships

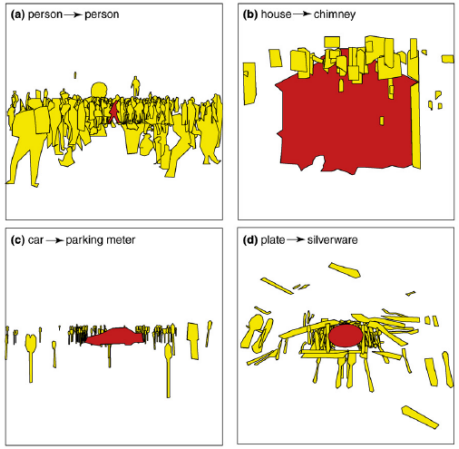
What is missing?

Valorem Furniture Plus Corner Office Desk

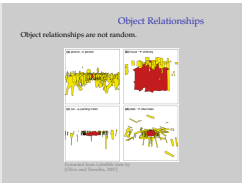
...

Object Relationships

Object relationships are not random.



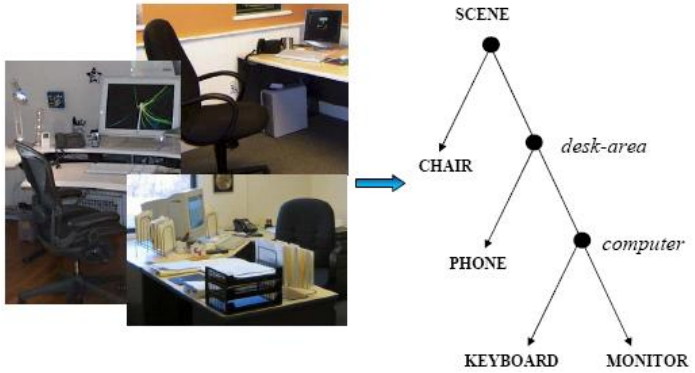
Extracted from LabelMe data by
[Oliva and Torralba, 2007]



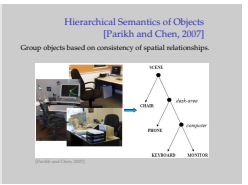
1. Constellation and part-based models work well for object recognition, why not scene understanding?

Hierarchical Semantics of Objects [Parikh and Chen, 2007]

Group objects based on consistency of spatial relationships.



[Parikh and Chen, 2007]



1. Avoids chicken-and-egg problem by organizing features bottom-up.
2. Operates on object instances, not categories
3. Completely unsupervised
4. Learns number of objects, object features, and object relationships simultaneously

Algorithm

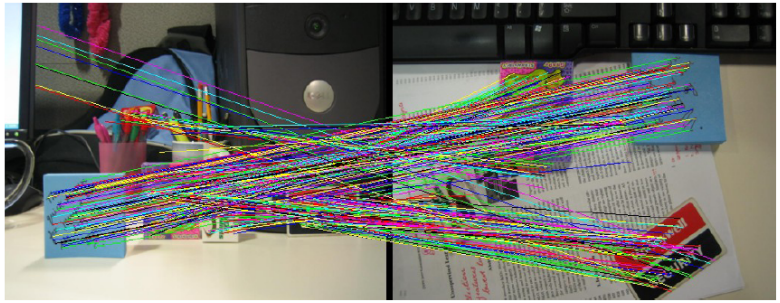
1. Extract features
2. Establish correspondences between features
3. Discard geometrically-inconsistent correspondences
4. Calculate correlation between pairs of feature correspondences
5. Hierarchically cluster features based on correlation
6. Merge nodes with a high geometric consistency

Algorithm
1. Extract features
2. Establish correspondences between features
3. Discard geometrically-inconsistent correspondences
4. Calculate correlation between pairs of feature correspondences
5. Hierarchically cluster features based on correlation
6. Merge nodes with a high geometric consistency

1. Details to follow

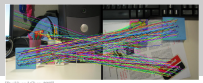
Feature Correspondences

1. Extract features: derivative of Gaussian, SIFT representation
2. Establish correspondences between features: k-nearest neighbors
3. Measure geometric consistency: use SIFT orientation and scale of one feature to predict relative location of another
4. Use spectral technique (Leodeanu and Hebert, 2005) to discard features with no geometrically-consistent support



[Parikh and Chen, 2007]

Feature Correspondences
1. Extract features: derivative of Gaussian, SIFT representation
2. Establish correspondences between features: k-nearest neighbors
3. Measure geometric consistency: use SIFT orientation and scale of one feature to predict relative location of another
4. Use spectral technique (Leodeanu and Hebert, 2005) to discard features with no geometrically-consistent support

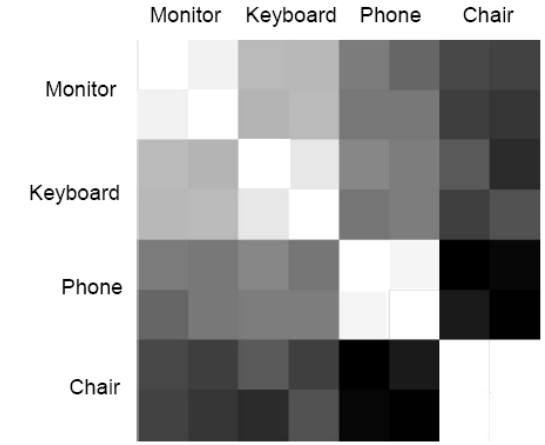


[Parikh and Chen, 2007]

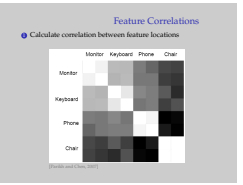
1. This step establishes feature correspondences and eliminates background clutter
2. It's very reliable, no false correspondences in the author's tests

Feature Correlations

1. Calculate correlation between feature locations



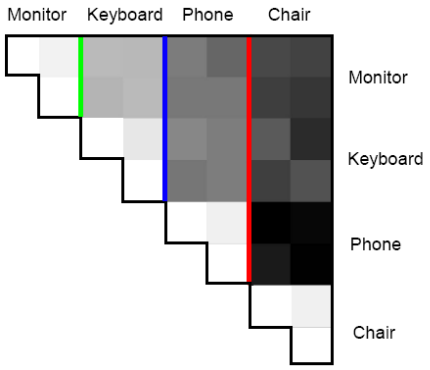
[Parikh and Chen, 2007]

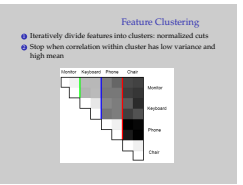


1. Correlation is based on covariance of x and y locations of features across all images
2. In this chart white is high correlation, and features have been sorted to show object structure

Feature Clustering

1. Iteratively divide features into clusters: normalized cuts
2. Stop when correlation within cluster has low variance and high mean

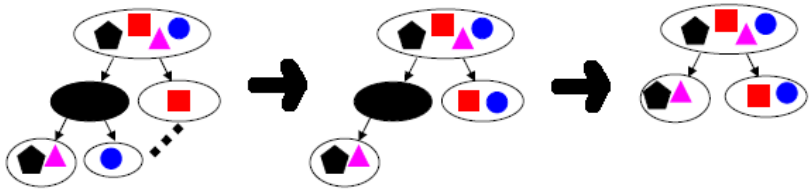




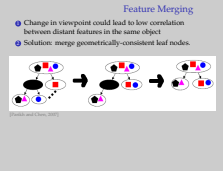
1. Start with a fully-connected graph of features, weighted by correlation
2. Normalized cuts separate groups of features with low correlation

Feature Merging

1. Change in viewpoint could lead to low correlation between distant features in the same object
2. Solution: merge geometrically-consistent leaf nodes.

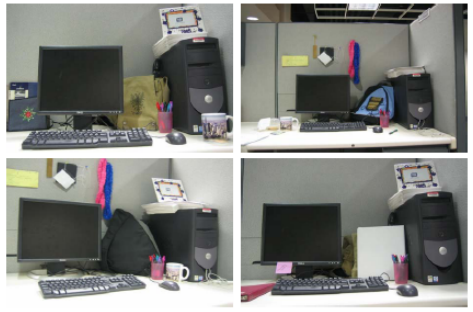


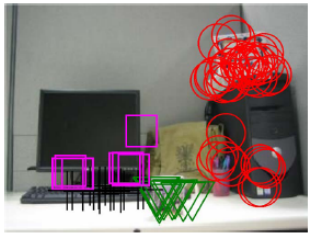
[Parikh and Chen, 2007]

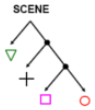


1. All pairs of clusters are examined. Those with high average geometric consistency are merged.
2. Clusters are merged into the lowest (most specific) level.
3. This corrects for the fact that the clusters were split prematurely.

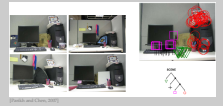
Example





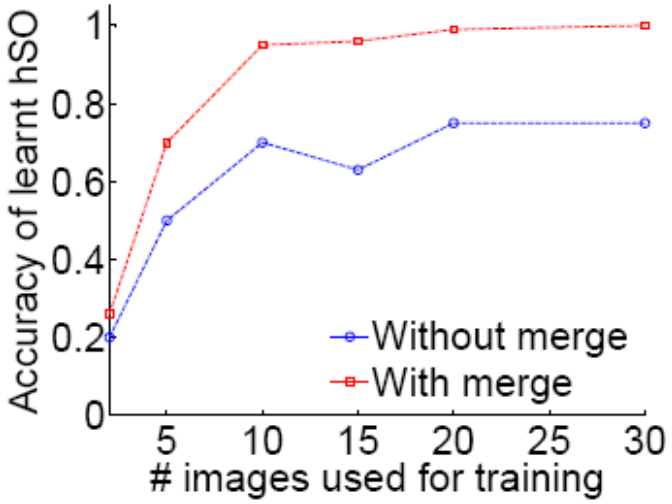


[Parikh and Chen, 2007]

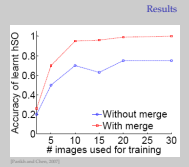


...

Results



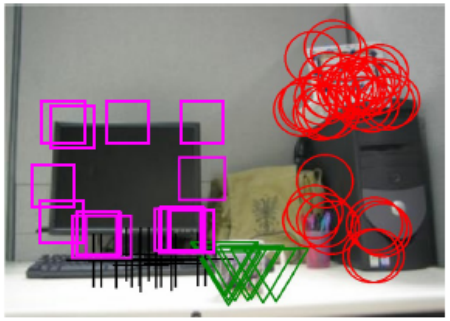
[Parikh and Chen, 2007]

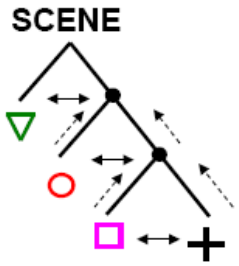


1. Compared against ground truth for a set of manually-chosen images

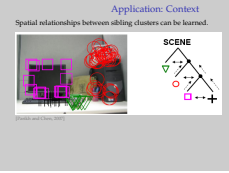
Application: Context

Spatial relationships between sibling clusters can be learned.





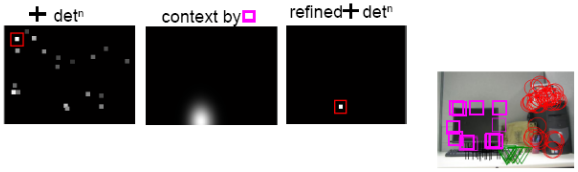
[Parikh and Chen, 2007]



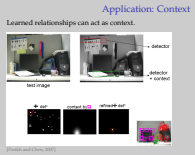
1. For example, we can estimate the relative position of cluster centers of gravity as mixture of Gaussians.

Application: Context

Learned relationships can act as context.



[Parikh and Chen, 2007]



...

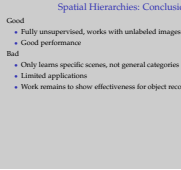
Spatial Hierarchies: Conclusion

Good

- Fully unsupervised, works with unlabeled images
- Good performance

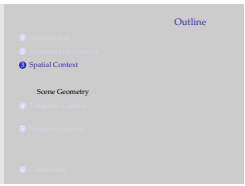
Bad

- Only learns specific scenes, not general categories
- Limited applications
- Work remains to show effectiveness for object recognition



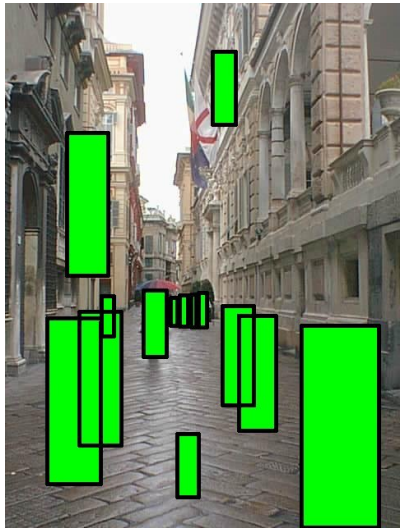
...

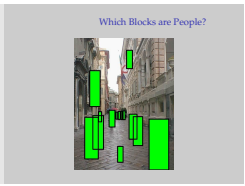
Outline



...

Which Blocks are People?





...

Biederman's Relations (1981)

Objects in a well-formed scene have stereotypical relationships

- *Support*
- *Size*
- *Position*
- Interposition
- Likelihood of appearance

These properties are mediated by semantics, 3D structure, and camera position.

Biederman's Relations (1981)
Objects in a well-formed scene have stereotypical relationships
- *Support*
- *Size*
- *Position*
- Interposition
- Likelihood of appearance

These properties are mediated by semantics, 3D structure, and camera position.

1. The next paper focuses on the first three

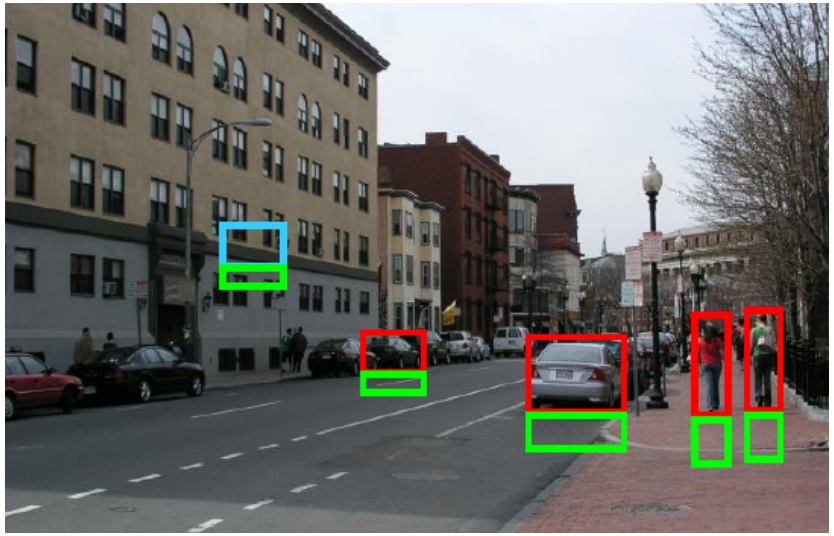
Putting Objects in Perspective [Hoiem et al, 2006]

1. Estimate object locations and sizes using local detector
2. Estimate support from 3D structure
3. Estimate camera properties from detected objects
4. Combine estimations, refine, and repeat

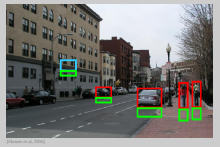
Putting Objects in Perspective [Hoiem et al, 2006]
1. Estimate object locations and sizes using local detector
2. Estimate support from 3D structure
3. Estimate camera properties from detected objects
4. Combine estimations, refine, and repeat

...

Object Support

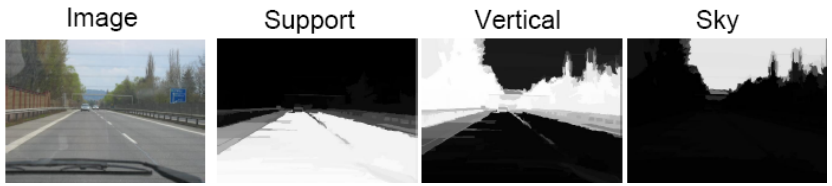


[Hoiem et al, 2006]



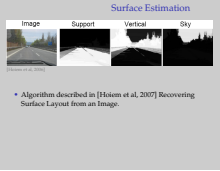
...

Surface Estimation



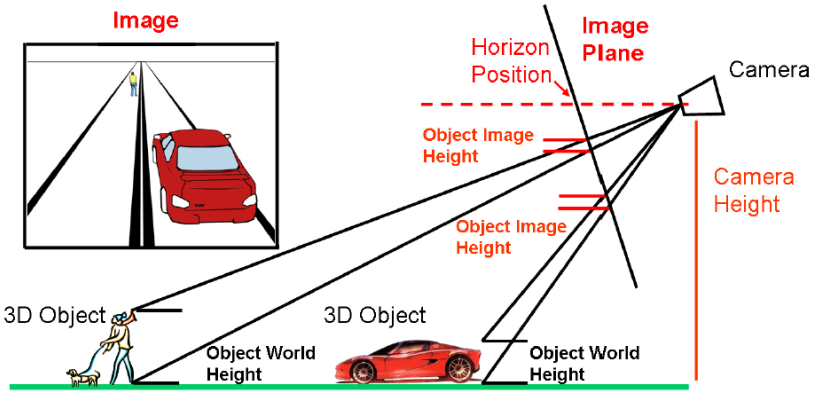
[Hoiem et al, 2006]

- Algorithm described in [Hoiem et al, 2007] Recovering Surface Layout from an Image.

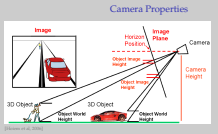


1. This is essentially the same algorithm as “Pop-up Photos”, discussed last class.

Camera Properties



[Hoiem et al, 2006]



...

Size and Horizon

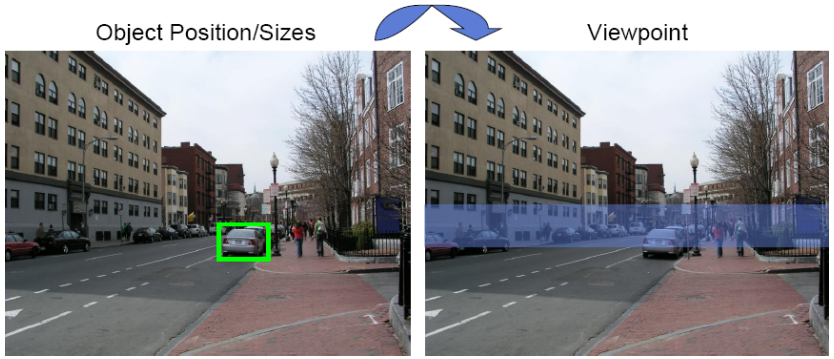
Initial estimate of object size and horizon

[Hoiem et al, 2006]

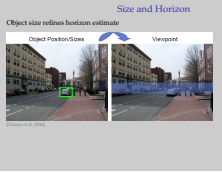
...

Size and Horizon

Object size refines horizon estimate



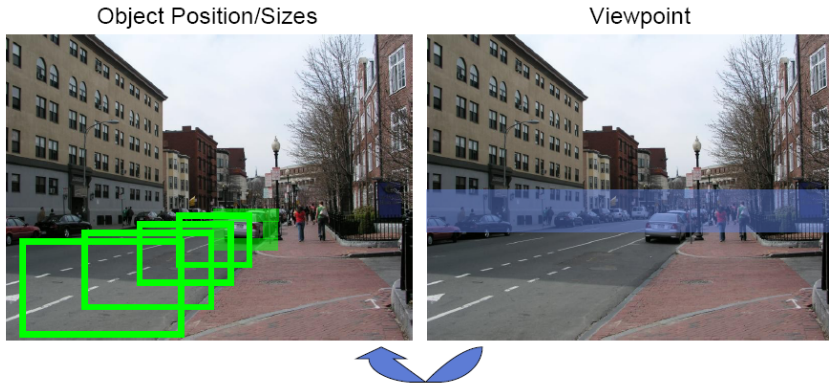
[Hoiem et al, 2006]



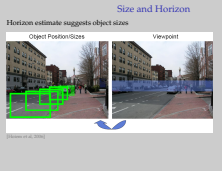
...

Size and Horizon

Horizon estimate suggests object sizes



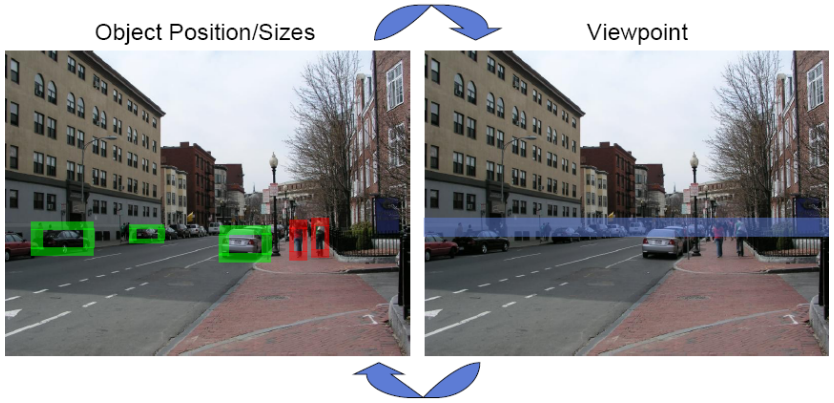
[Hoiem et al, 2006]



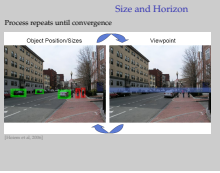
...

Size and Horizon

Process repeats until convergence

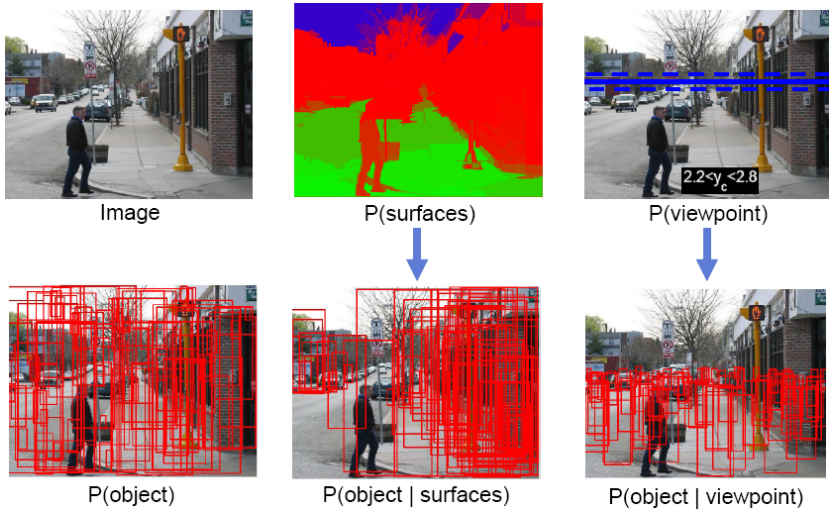


[Hoiem et al, 2006]

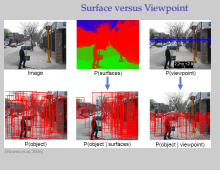


...

Surface versus Viewpoint

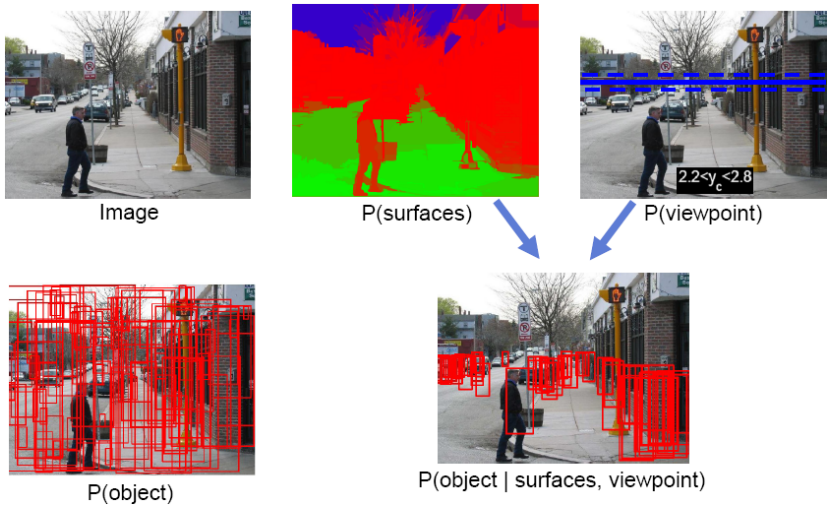


[Hoiem et al, 2006]

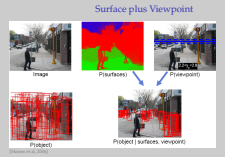


...

Surface plus Viewpoint

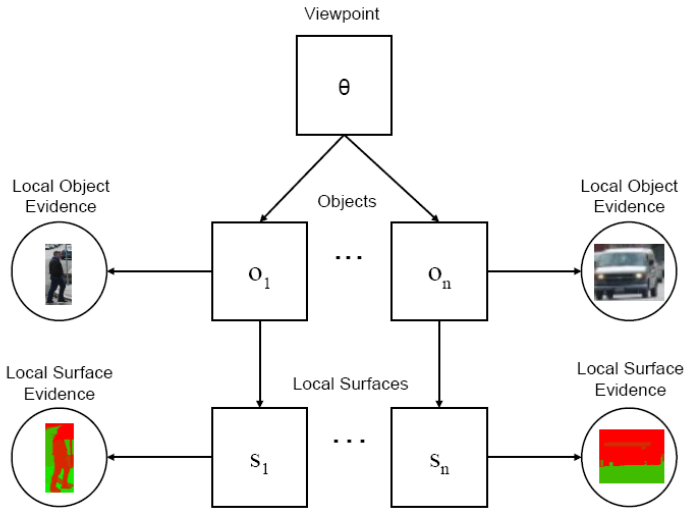


[Hoiem et al, 2006]

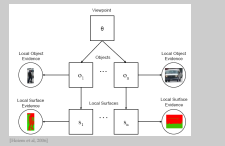


...

Bayesian Network

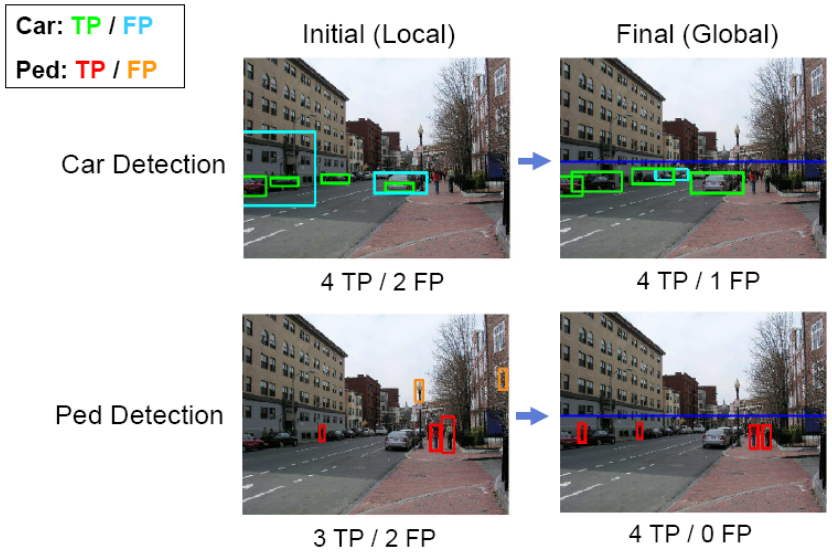


[Hoiem et al, 2006]

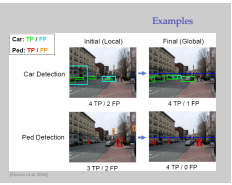


1. Only most significant dependencies are modeled, to simplify computation
2. Pearl's belief propogation algorithm used to find most probable explanation for the scene

Examples

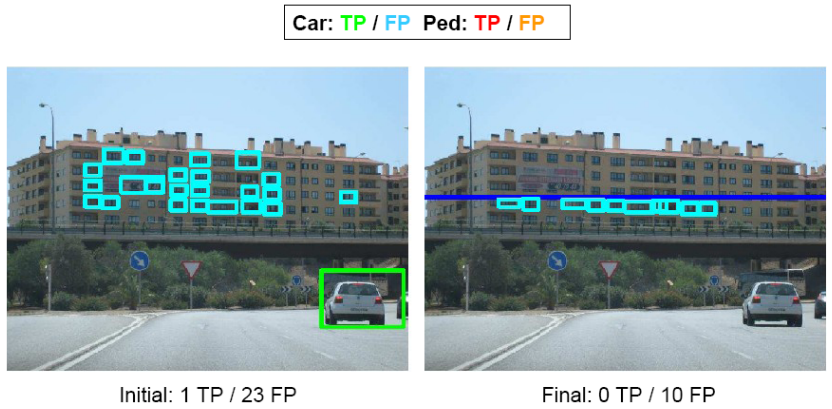


[Hoiem et al, 2006]

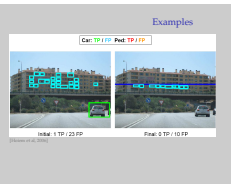


1. Examples of good results

Examples

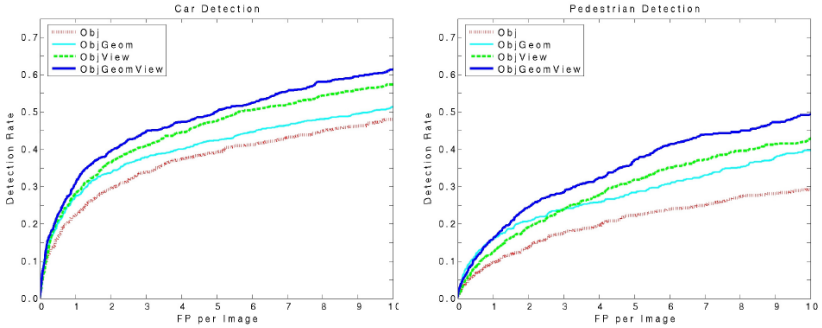


[Hoiem et al, 2006]

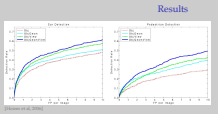


1. If false positives dominate the image, they can force true positives to be discarded

Results



[Hoiem et al, 2006]



1. ROC for cars and pedestrians
2. Both viewpoint and surface estimation improve results, and both combined show significant improvement.

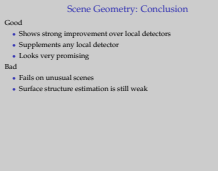
Scene Geometry: Conclusion

Good

- Shows strong improvement over local detectors
- Supplements any local detector
- Looks very promising

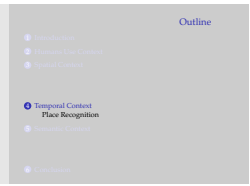
Bad

- Fails on unusual scenes
- Surface structure estimation is still weak

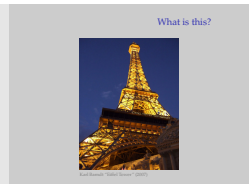


...

Outline

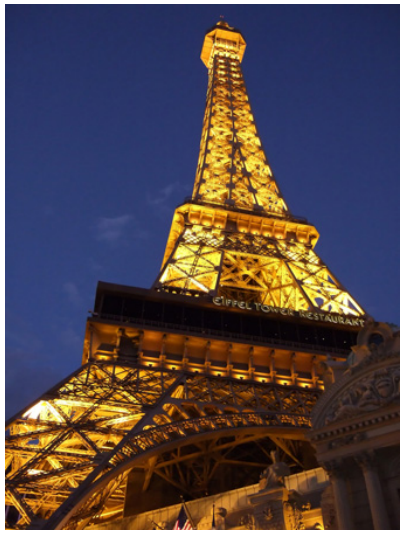


...



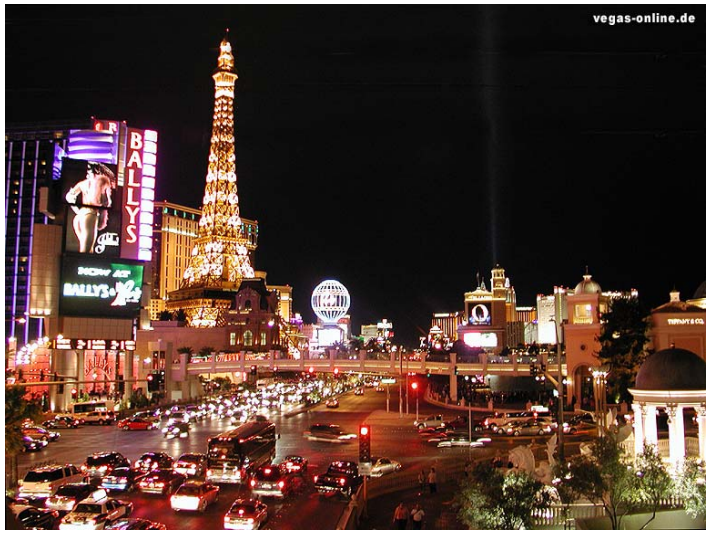
The Eiffel Tower restaurant in Las Vegas

What is this?



Karl Barndt "Eiffel Tower" (2007)

A few minutes earlier…



vegas-online.de “Vegas Strip South”

A few minutes earlier…

1. This view makes it clear this is Las Vegas, not Paris
2. History provides access to a broader spatial context

Place Recognition Task

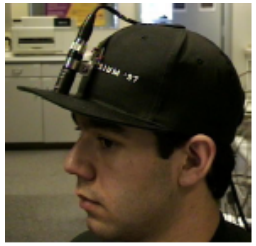
- Location is an important type of context
- Single image may be insufficient to establish location
- Can we use historical information to predict location?

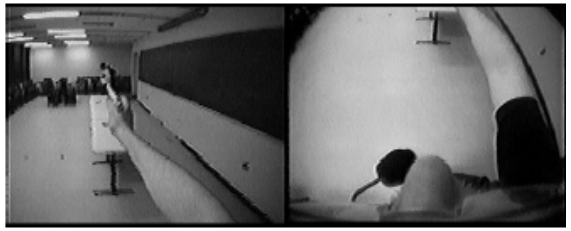
Place Recognition Task
- Location is an important type of context
- Single image may be insufficient to establish location
- Can we use historical information to predict location?

…

Context Awareness in Wearable Computing [Starner et al, 1998]

- Observe players in an indoor paintball-like game
- Very rudimentary global features (3 color samples)
- Multiple small HMMs combined with statistical bigram predict movement between rooms
- 84% accuracy at place recognition
- Also experimented with action and object recognition



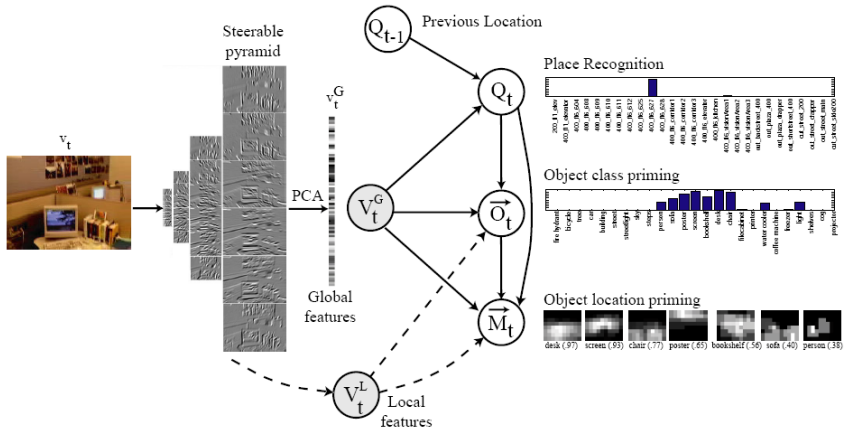


[Starner et al, 1998]

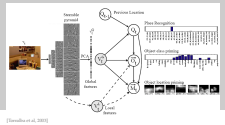
1. Probably the first example of this approach
2. Paper is vague on technical details and performance is not great
3. "Patrol" game played at MIT
4. Color samples from ahead, nose, and floor

Place and Object Recognition [Torralba et al, 2003]

- More recently: supplement contextual priming with HMM to model movement



[Torralba et al, 2003]



1. Extract global features similarly to Contextual Priming
2. Global features and history predict location
3. Global features and location predict object identity
4. Global features, location, and object identity predict object location

Modeling Locations

- Global features similar to [Torralba, 2003]. Three features tested:
  - Monochrome filter responses to steerable pyramid with 6 orientations and 4 scales
  - Color downsampled
  - Monochrome downsampled
- Choose K prototype views uniformly from training data
- Model location as mixture of K spherical gaussians based on prototypes
- σ and K chosen by cross-validation
- Better approaches possible

Modeling Locations
- Global features similar to [Torralba, 2003]. Three features tested:
  - Monochrome filter responses to steerable pyramid with 6 orientations and 4 scales
  - Color downsampled
  - Monochrome downsampled
- Choose K prototype views uniformly from training data
- Model location as mixture of K spherical gaussians based on prototypes
- σ and K chosen by cross-validation
- Better approaches possible

1. Prototypes and weights could be chosen by EM for better results
2. There are much better approaches to location recognition (we'll study them later)
3. But the point of this exercise is to show how much can be done solely with global features

Modeling Time

HMM used to compute probability distribution over locations:

$$\begin{aligned} P(Q_t = q|v_{1:t}) &\propto p(v_t|Q_t = q)P(Q_t = q|v_{1:t-1}) \\ &= p(v_t|Q_t = q)\sum_{q'} A(q', q)P(Q_{t-1} = q'|v_{1:t-1}) \end{aligned}$$

- $A(q', q)$ is transition matrix learned from training data
- $p(v_t|Q_t = q)$ is observation likelihood

Modeling Time
HMM used to compute probability distribution over locations:
$$\begin{aligned} P(Q_t = q|v_{1:t}) &\propto p(v_t|Q_t = q)P(Q_t = q|v_{1:t-1}) \\ &= p(v_t|Q_t = q)\sum_{q'} A(q', q)P(Q_{t-1} = q'|v_{1:t-1}) \end{aligned}$$
- $A(q', q)$ is transition matrix learned from training data
- $p(v_t|Q_t = q)$ is observation likelihood

1. Lack of training data is always a problem
2. Transition matrix is smoothed with Dirichlet prior so no transition is excluded
3. This should help generalize slightly from weak training data

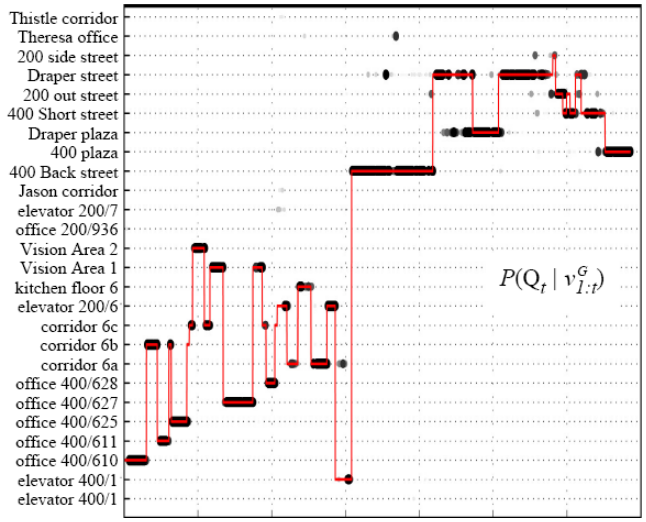
Demo Movie

Demo movie…

Demo Movie
Demo movie…

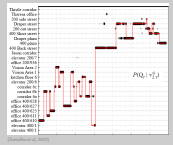
The movie is not provided with the printable version of the presentation.

Example



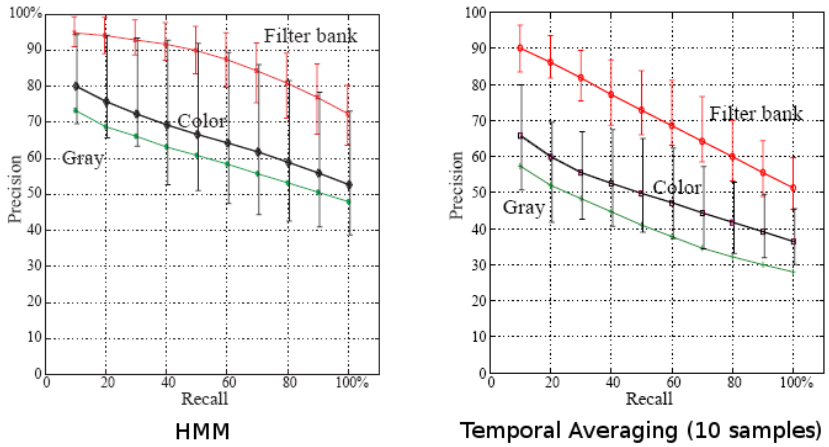
[Torralba et al, 2003]

Example

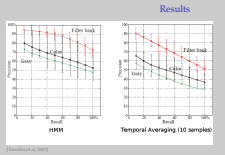


1. Red line is ground truth

Results



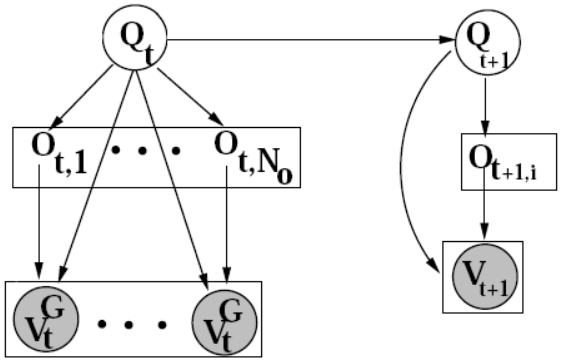
[Torralba et al, 2003]



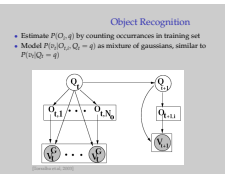
1. Median performance computed using leave-one-out cross-validation on 17 sequences
2. Error bars indicate 80% probability region
3. Non-HMM performance without averaging significantly worse

Object Recognition

- Estimate $P(O_i, q)$ by counting occurrances in training set
- Model $P(v_t|O_{t,i}, Q_t = q)$ as mixture of gaussians, similar to $P(v_t|Q_t = q)$



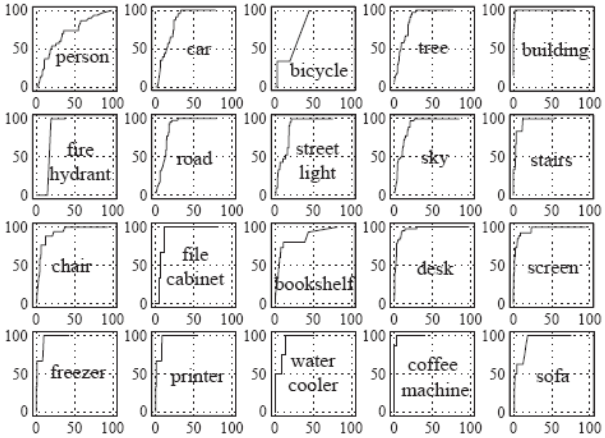
[Torralba et al, 2003]



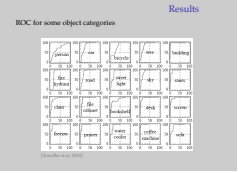
...

Results

ROC for some object categories



[Torralba et al, 2003]



...

Way cooler than GPS

[Torralba et al, 2003]

...

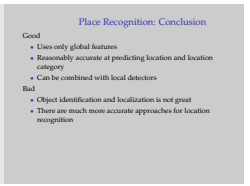
Place Recognition: Conclusion

Good

- Uses only global features
- Reasonably accurate at predicting location and location category
- Can be combined with local detectors

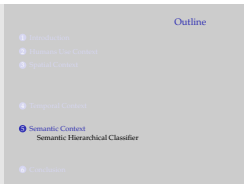
Bad

- Object identification and localization is not great
- There are much more accurate approaches for location recognition



...

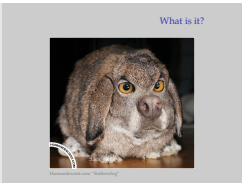
Outline



...

What is it?

Humandescent.com “Rabbowlog”



1. It’s not a rabbit, but it is an animal.
2. Object classifiers should degrade gracefully, like humans.

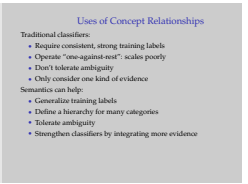
Uses of Concept Relationships

Traditional classifiers:

- Require consistent, strong training labels
- Operate “one-against-rest”: scales poorly
- Don’t tolerate ambiguity
- Only consider one kind of evidence

Semantics can help:

- Generalize training labels
- Define a hierarchy for many categories
- Tolerate ambiguity
- Strengthen classifiers by integrating more evidence



…

WordNet

A good source of semantic relationships.

- synonym = same
- antonym = opposite
- hypernym / hyponym = class
- holonym / meronym = part

For object recognition we can use:

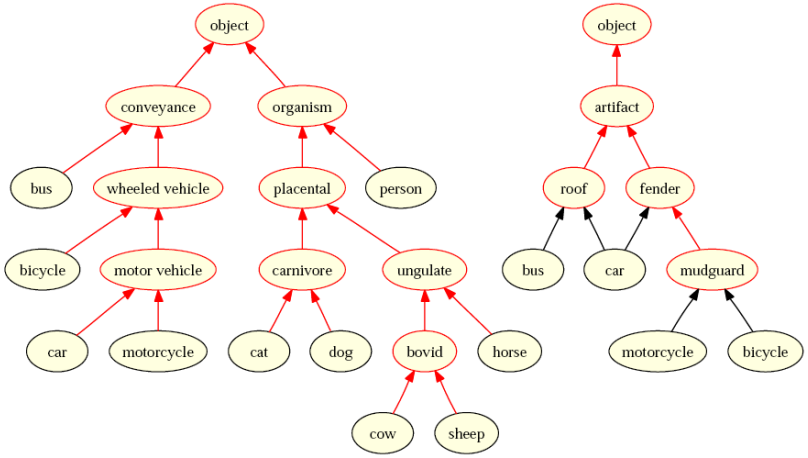
- hypernym and meronym for detection
- antonym for classification

WordNet
A good source of semantic relationships.
- synonym = same
- antonym = opposite
- hypernym / hyponym = class
- holonym / meronym = part

For object recognition we can use:
- hypernym and meronym for detection
- antonym for classification

1. Hypernym detection: if there is a Ford, there is a car
2. Meronym detection: if there is a car, there is a fender
3. Antonym classification: if this picture is a man, it is not a woman

WordNet
Some hypernyms and meronyms



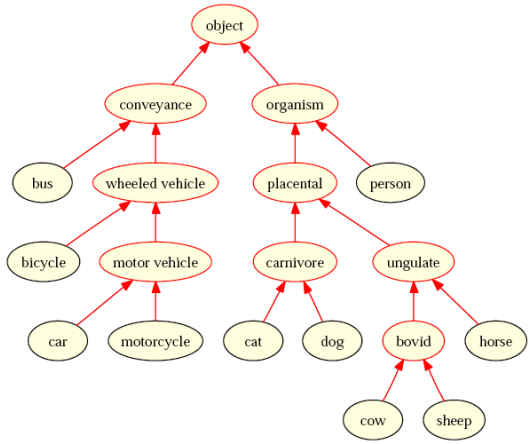
…

WordNet

Some hypernyms and meronyms

Semantic Hierarchies [Marszałek and Schmid, 2007]

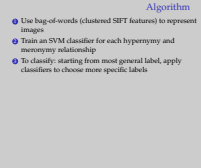
Organize classifiers into a cascade based on semantic concepts



...

Algorithm

1. Use bag-of-words (clustered SIFT features) to represent images
2. Train an SVM classifier for each hypernymy and meronymy relationship
3. To classify: starting from most general label, apply classifiers to choose more specific labels



1. Approach reminiscent of classifier cascade used for faces
2. Each SVM classifier discriminates only within a category
3. Drawback: it's possible to choose the wrong path early

Examples

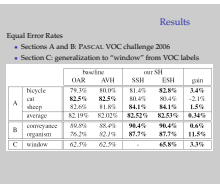
1. Note that false positives are still closely related to the query

Results

Equal Error Rates

- Sections A and B: PASCAL VOC challenge 2006
- Section C: generalization to "window" from VOC labels

		baseline		our SH		
		OAR	AVH	SSH	ESH	gain
A	bicycle	79.3%	80.0%	81.4%	**82.8%**	**3.4%**
	cat	**82.5%**	**82.5%**	80.4%	80.4%	-2.1%
	sheep	82.6%	81.8%	**84.1%**	**84.1%**	**1.5%**
	average	82.19%	82.02%	**82.52%**	**82.53%**	**0.34%**
B	conveyance	*89.8%*	*88.4%*	**90.4%**	**90.4%**	**0.6%**
	organism	*76.2%*	*82.1%*	**87.7%**	**87.7%**	**11.5%**
C	window	*62.5%*	*62.5%*	-	**65.8%**	**3.3%**



1. EER = point where precision equals recall
2. OAR is standard one-against-rest classifier
3. AVH is visual hierarchical classifier, obtained through iterative merging of classes with smallest χ^2
4. SSH uses only hyponymy, ESH uses meronymy also
5. OAR and AVH use post-labeling inference for generalization, while SSH and ESH do generalization automatically
6. Gains are fairly small

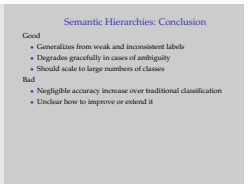
Semantic Hierarchies: Conclusion

Good

- Generalizes from weak and inconsistent labels
- Degrades gracefully in cases of ambiguity
- Should scale to large numbers of classes

Bad

- Negligible accuracy increase over traditional classification
- Unclear how to improve or extend it



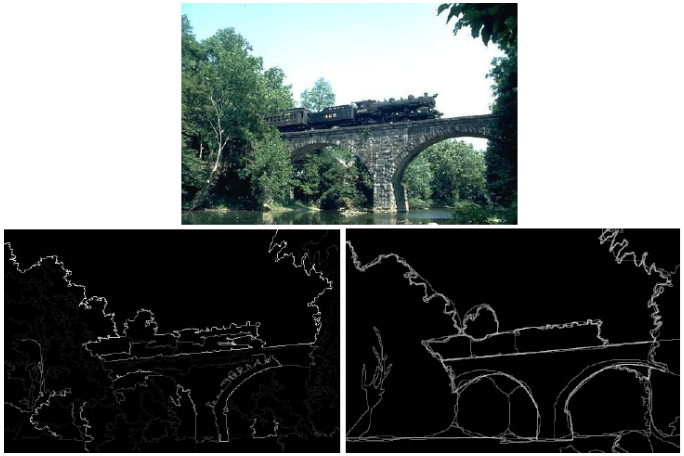
...

Outline

...

Segmentation Problem

Goal: segment an image into objects.



[Hoogs and Collins, 2006]

...

Semantic Segmentation [Hoogs and Collins, 2006]

- Parts of an object are semantically related.
- Semantic relationships can resolve appearance ambiguity.
- Use “semantic distance” to compare features instead of appearance distance.

Semantic Segmentation
[Hoogs and Collins, 2006]

- Parts of an object are semantically related.
- Semantic relationships can resolve appearance ambiguity.
- Use “semantic distance” to compare features instead of appearance distance.

...

Semantic Ontology

Building a semantic appearance model:

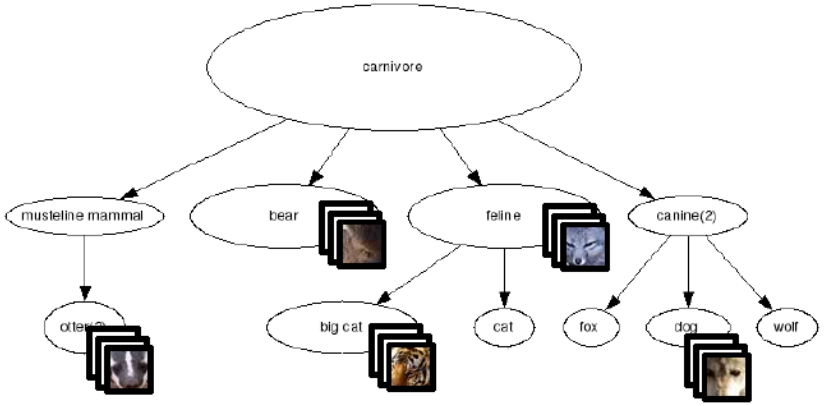
1. Start with manually segmented and labeled images
2. Segment further based on appearance (mean-shift)
3. Compute feature vectors for segments (textons)
4. Associate feature vectors with labels in semantic network
5. Compute probability of semantic labels

Semantic Ontology
Building a semantic appearance model:
1. Start with manually segmented and labeled images
2. Segment further based on appearance (mean-shift)
3. Compute feature vectors for segments (textons)
4. Associate feature vectors with labels in semantic network
5. Compute probability of semantic labels

...

Semantic Ontology

Augment WordNet with appearance exemplars and probabilities



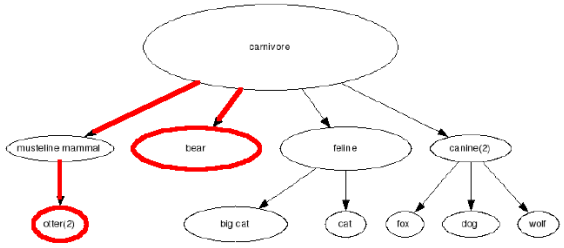
Semantic Ontology
Augment WordNet with appearance exemplars and probabilities

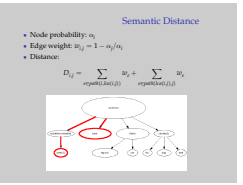
...

Semantic Distance

- Node probability: α_i
- Edge weight: $w_{i,j} = 1 - \alpha_j/\alpha_i$
- Distance:

$$D_{i,j} = \sum_{e \in path(i,lca(i,j))} w_e + \sum_{e \in path(lca(i,j),j)} w_e$$



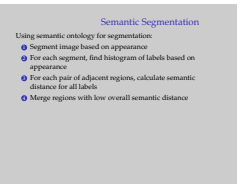


1. α computed from training statistics
2. w penalizes crossing a low-probability (distinctive) node
3. D is weight of shortest path between nodes

Semantic Segmentation

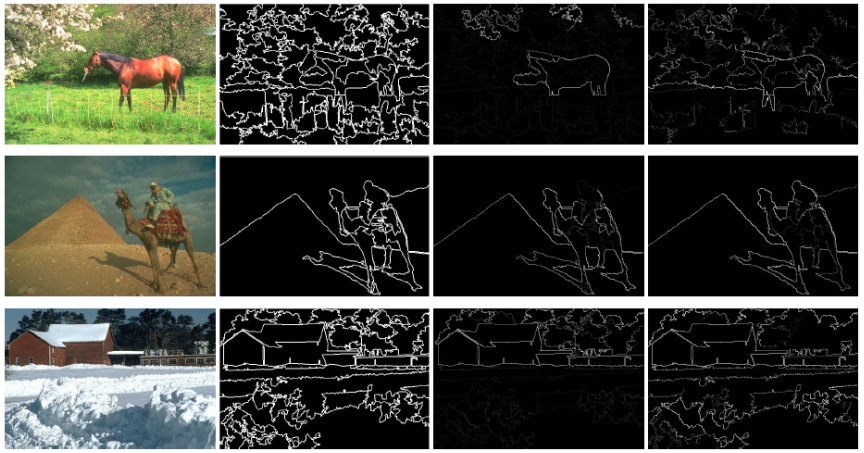
Using semantic ontology for segmentation:

1. Segment image based on appearance
2. For each segment, find histogram of labels based on appearance
3. For each pair of adjacent regions, calculate semantic distance for all labels
4. Merge regions with low overall semantic distance



...

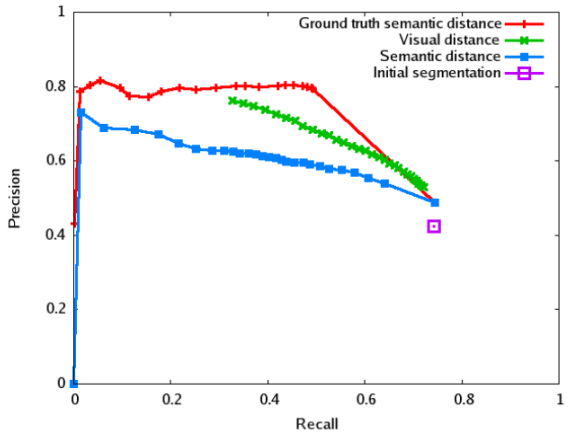
Examples



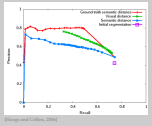
[Hoogs and Collins, 2006]

...

Results



[Hoogs and Collins, 2006]



...

Results

UC Berkeley segmentation benchmark:

Method	F-score
Human	0.79
Ground-truth SD	0.63
Visual Distance	0.62
Semantic Distance	0.59
Initial Segmentation	0.54
Random	0.43

- Semantic grouping performs well in theory
- Visual grouping performs well in practice

Results
UC Berkeley segmentation benchmark:
Method | F-score
Human | 0.79
Ground-truth SD | 0.63
Visual Distance | 0.62
Semantic Distance | 0.59
Initial Segmentation | 0.54
Random | 0.43
• Semantic grouping performs well in theory
• Visual grouping performs well in practice

...

Why so bad?

- Training data too sparse to capture appearance variation (average of 34 exemplars per node)
- Semantic model too restricted (no meronymy)
- Same-class object boundaries lost

[Hoogs and Collins, 2006]

- Poor initial segmentation?

Why so bad?
• Training data too sparse to capture appearance variation (average of 34 exemplars per node)
• Semantic model too restricted (no meronymy)
• Same-class object boundaries lost
[Hoogs and Collins, 2006]
• Poor initial segmentation?

...

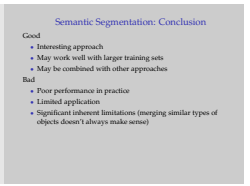
Semantic Segmentation: Conclusion

Good

- Interesting approach
- May work well with larger training sets
- May be combined with other approaches

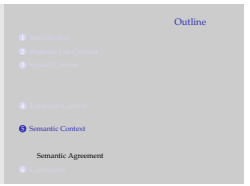
Bad

- Poor performance in practice
- Limited application
- Significant inherent limitations (merging similar types of objects doesn't always make sense)



...

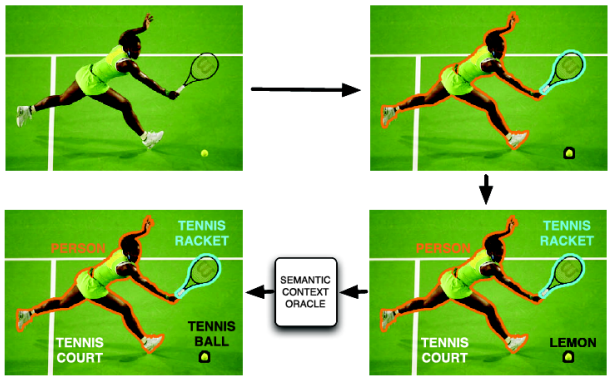
Outline



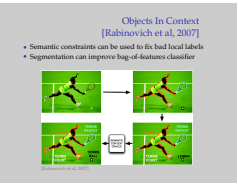
...

Objects In Context [Rabinovich et al, 2007]

- Semantic constraints can be used to fix bad local labels
- Segmentation can improve bag-of-features classifier



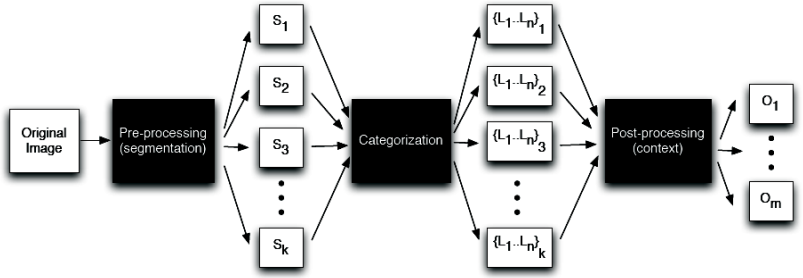
[Rabinovich et al, 2007]



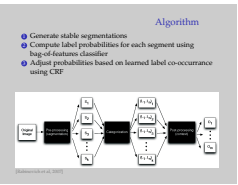
1. Rabinovich has a Google Tech Talk online which covers stable segmentations and "Objects in Context"
2. Arguably this is a spatial, not semantic technique
3. When co-occurance is learned from training, it is purely spatial
4. When co-occurance is learned from Google, it is semantic

Algorithm

1. Generate stable segmentations
2. Compute label probabilities for each segment using bag-of-features classifier
3. Adjust probabilities based on learned label co-occurrance using CRF



[Rabinovich et al, 2007]



...

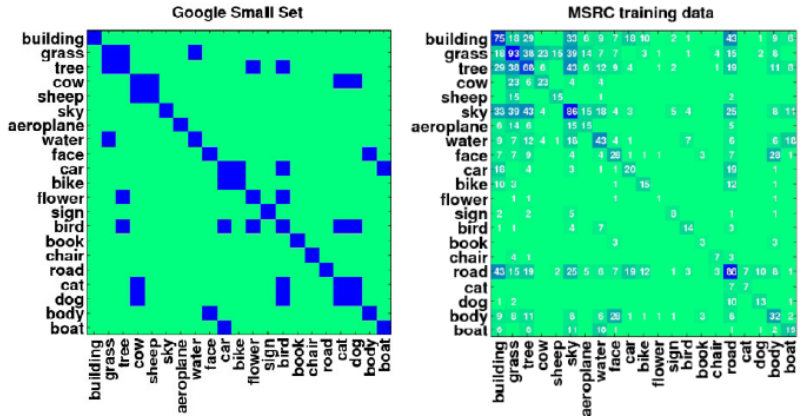
Learning co-occurrance

- From training data
- Google Sets (small)
- Google Sets (large)

Learning co-occurrance
- From training data
- Google Sets (small)
- Google Sets (large)

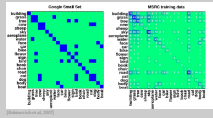
...

Co-occurrance: Examples



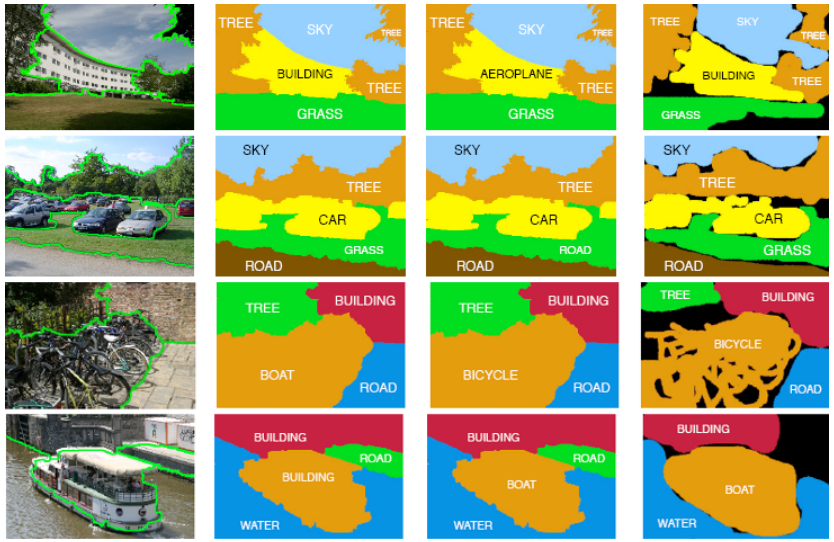
[Rabinovich et al, 2007]

Co-occurrance: Examples

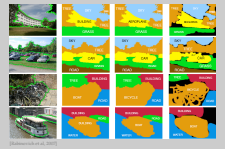


...

Examples



[Rabinovich et al, 2007]



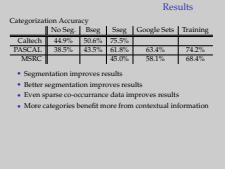
...

Results

Categorization Accuracy

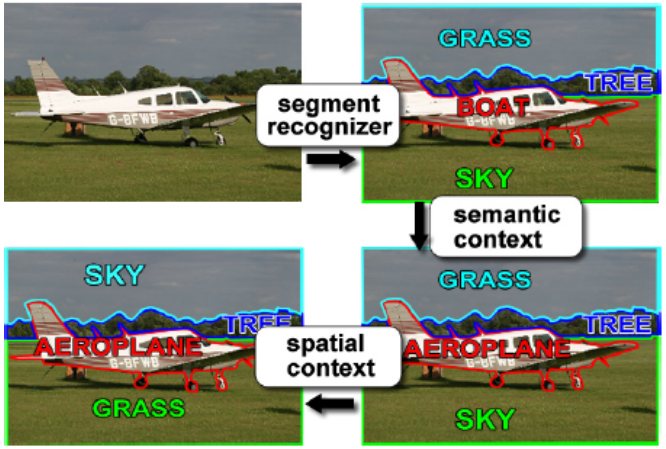
	No Seg.	Bseg	Sseg	Google Sets	Training
Caltech	44.9%	50.6%	75.5%		
PASCAL	38.5%	43.5%	61.8%	63.4%	74.2%
MSRC			45.0%	58.1%	68.4%

- Segmentation improves results
- Better segmentation improves results
- Even sparse co-occurrance data improves results
- More categories benefit more from contextual information

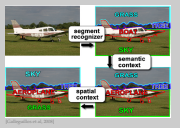


...

Adding Spatial Context [Galleguillos et al, 2008]



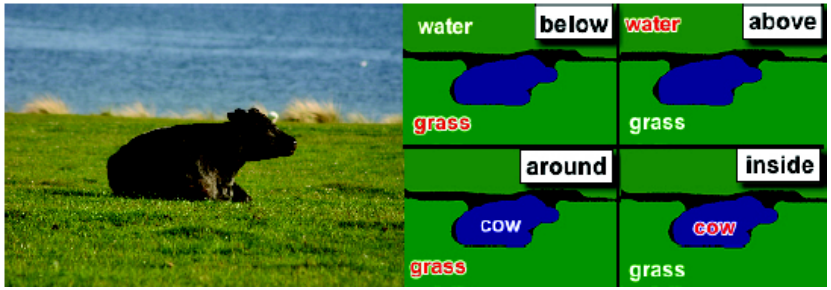
[Galleguillos et al, 2008]



...

Spatial Relationships

- Spatial relationships: vertical offset + percentage of bounding box overlap
- Vector quantized into 4 dimensions

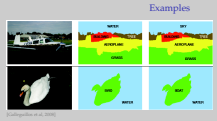


[Galleguillos et al, 2008]

...

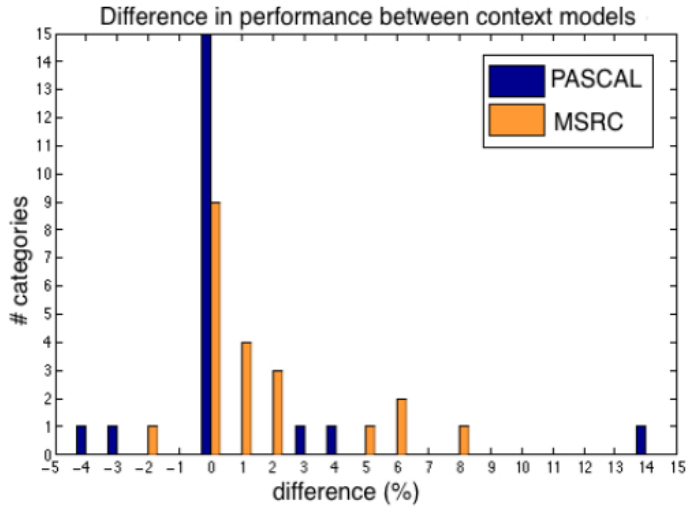
Examples

[Galleguillos et al, 2008]

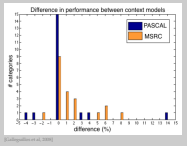


...

Results



[Galleguillos et al, 2008]



...

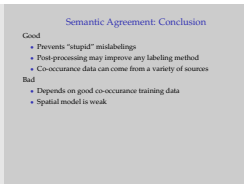
Semantic Agreement: Conclusion

Good

- Prevents “stupid” mislabelings
- Post-processing may improve any labeling method
- Co-occurance data can come from a variety of sources

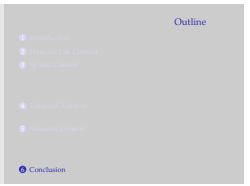
Bad

- Depends on good co-occurance training data
- Spatial model is weak



...

Outline



...

Review: Spatial Context

Context based on static 2D and 3D relationships.

- Contextual Priming: use global image appearance to predict object properties
- Hierarchical Semantics: cluster features based on consistent image relationships
- Objects in Perspective: use simple geometric constraints (horizon and object height) to improve local detection

Review: Spatial Context
Context based on static 2D and 3D relationships.
- Contextual Priming: use global image appearance to predict object properties
- Hierarchical Semantics: cluster features based on consistent image relationships
- Objects in Perspective: use simple geometric constraints (horizon and object height) to improve local detection

...

Review: Temporal Context

Context incorporating time dimension.

- Context-based Place Recognition
  - Use HMM to model motion and predict location
  - Use location to predict presence and location of objects
- Field of "action recognition" provides many other examples which we will study later.

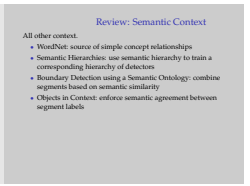
Review: Temporal Context
Context incorporating time dimension.
- Context-based Place Recognition
  - Use HMM to model motion and predict location
  - Use location to predict presence and location of objects
- Field of "action recognition" provides many other examples which we will study later.

...

Review: Semantic Context

All other context.

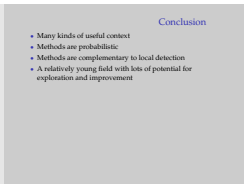
- WordNet: source of simple concept relationships
- Semantic Hierarchies: use semantic hierarchy to train a corresponding hierarchy of detectors
- Boundary Detection using a Semantic Ontology: combine segments based on semantic similarity
- Objects in Context: enforce semantic agreement between segment labels



...

Conclusion

- Many kinds of useful context
- Methods are probabilistic
- Methods are complementary to local detection
- A relatively young field with lots of potential for exploration and improvement



...

Discussion Questions

- What kinds of context are most useful?
- How can we capture the dual foreground/background roles of objects?
- When is it better to ignore context? How can we do this selectively?
- Does context enable new applications for recognition?
- Can the approaches discussed be combined? How?
- Could we have a single framework for combining all kinds of local and global detectors?
- Every method makes significant simplifying assumptions; can we avoid this? Does it matter?
- …

Discussion Questions
- What kinds of context are most useful?
- How can we capture the dual foreground/background roles of objects?
- When is it better to ignore context? How can we do this selectively?
- Does context enable new applications for recognition?
- Can the approaches discussed be combined? How?
- Could we have a single framework for combining all kinds of local and global detectors?
- Every method makes significant simplifying assumptions; can we avoid this? Does it matter?
- …

…

General References

D. Hoiem.
Putting Context Into Vision.
PowerPoint slides for CMU reading group, 2004.

A. Oliva and A. Torralba.
The Role of Context in Object Recognition.
TRENDS in Cognitive Sciences, 11(12), 2007.

Global Appearance

A. Oliva and A. Torralba.
Modeling the shape of the scene: a holistic representation of the spatial envelope.
International Journal of Computer Vision, 42(3):145–175, 2001.

A. Torralba.
Contextual Priming for Object Detection.
International Journal of Computer Vision, 2003.

K. Murphy, A. Torralba, D. Eaton and W. Freeman.
Object detection and localization using local and global features.
Lecture Notes in Computer Science, at Sicily workshop on object recognition, 2005.

Other Spatial Context

D. Hoiem, A. A. Efros, and M. Hebert.
Putting Objects in Perspective.
In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2006.

D. Hoiem, A. A. Efros, and M. Hebert.
Recovering Surface Layout from an Image
International Journal of Computer Vision 75(1), 2007.

D. Parikh and T. Chen.
Unsupervised Learning of Hierarchical Semantics of Objects (hSOs).
In *Proceedings of the International Conference on Computer Vision*, 2007.

Temporal Context

T. Starner, B. Schiele, and A. Pentland.
Visual Contextual Awareness in Wearable Computing.
In *Proceedings of Visual Contextual Awareness in Wearable Computing*, 1998.

A. Torralba, K. Murphy, W. Freeman, M. Rubin.
Context-based vision system for place and object recognition.
In *Proceedings of the IEEE Intl. Conference on Computer Vision*, 2003.

Semantic Context 1

M. Marszałek and C. Schmid.
Semantic Hierarchies for Visual Object Recognition.
In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

A. Hoogs and R. Collins.
Object Boundary Detection in Images using a Semantic Ontology.
In *Proceedings of the Association for the Advancement of Artificial Intelligence*, 2006.

Semantic Context 2

A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, S. Belongie.
Objects in Context.
In *Proceedings of the IEEE Intl. Conference on Computer Vision*, 2007.

C. Galleguillos, A. Rabinovich and S. Belongie
COLA: Co-Ocurrence, Location and Appearance for Object Categorization
UCSD Tech Report, 2008.