Overcoming Ambiguity in Visual Object Recognition

Prof. Trevor Darrell
UC Berkeley EECS Dept. &
Intl. Computer Science Inst. (ICSI)

Sources of Ambiguity

- Cue saliency varies across categories


069.fighter-jet


070.fire-extinguisher


071.fire-hydrant


072.fire-truck

- Individual categories have multiple senses


vs



vs


- Multiple surfaces confuse local features

Sources of Ambiguity

- Cue saliency varies across categories


069.fighter-jet


070.fire-extinguisher


071.fire-hydrant


072.fire-truck

- Individual categories have multiple senses


vs



vs


- Multiple surfaces confuse local features

Sources of Ambiguity

- Cue saliency varies across categories
  - *Probabilistic multi-kernel fusion*… [Christhoudias]
  - *Joint regularization across categories*… [Quattoni]
- Individual categories have multiple senses
  - *Dictionary grounded visual models*… [Saenko]
- Multiple surfaces confuse local features
  - *Local feature models for transparent objects [Fritz]*

Today: Snapshots

- Probabilistic multi-kernel fusion
- Joint regularization across categories
- Multimodal sense grounding
- Local feature models for transparent objects

Today: Snapshots

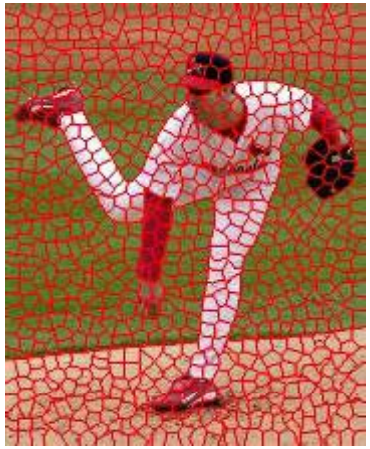
- Probabilistic multi-kernel fusion
- Joint regularization across categories
- Multimodal sense grounding
- Local feature models for transparent objects

Local Representations

Wide variety of proposed local feature representations:

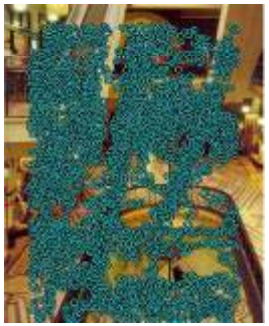
SIFT [Lowe]

Shape context
[Belongie et al.]

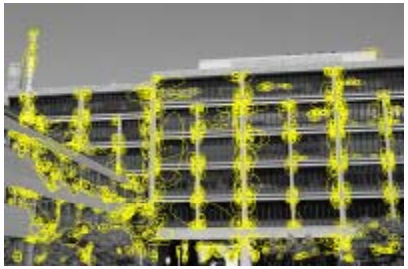


Superpixels
[Ren et al.]

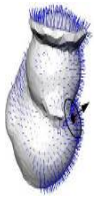
Maximally Stable Extremal
Regions [Matas et al.]



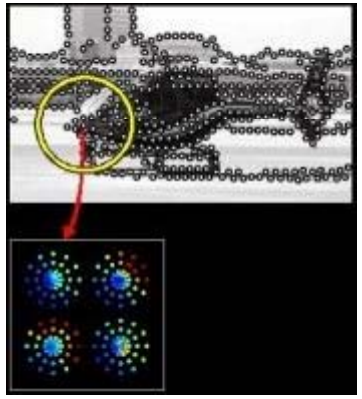
Salient regions
[Kadir et al.]



Harris-Affine
[Schmid et al.]



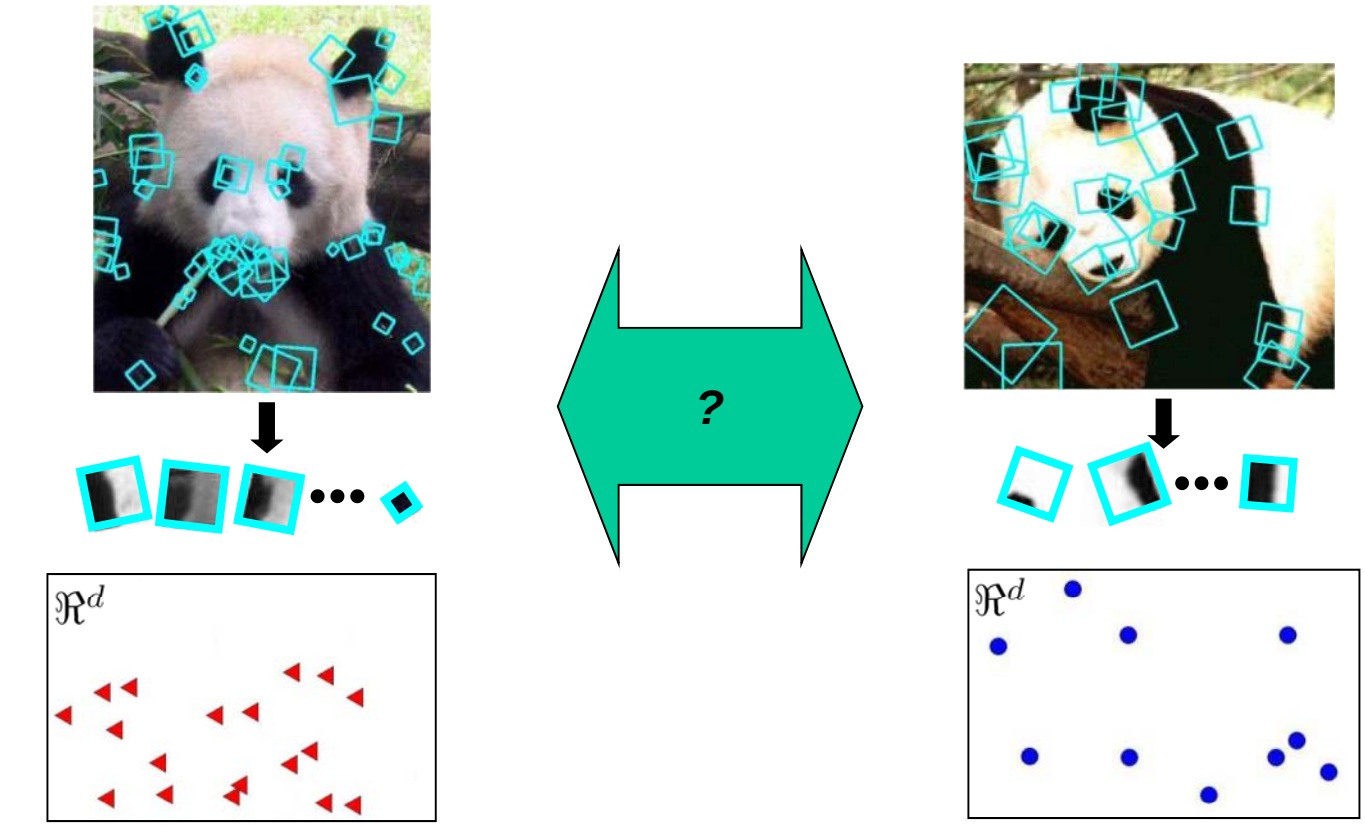
Spin images
[Johnson
and Hebert]



Geometric Blur
[Berg et al.]

How to Compare Sets of Features?

- Each instance is unordered set of vectors
- Varying number of vectors per instance



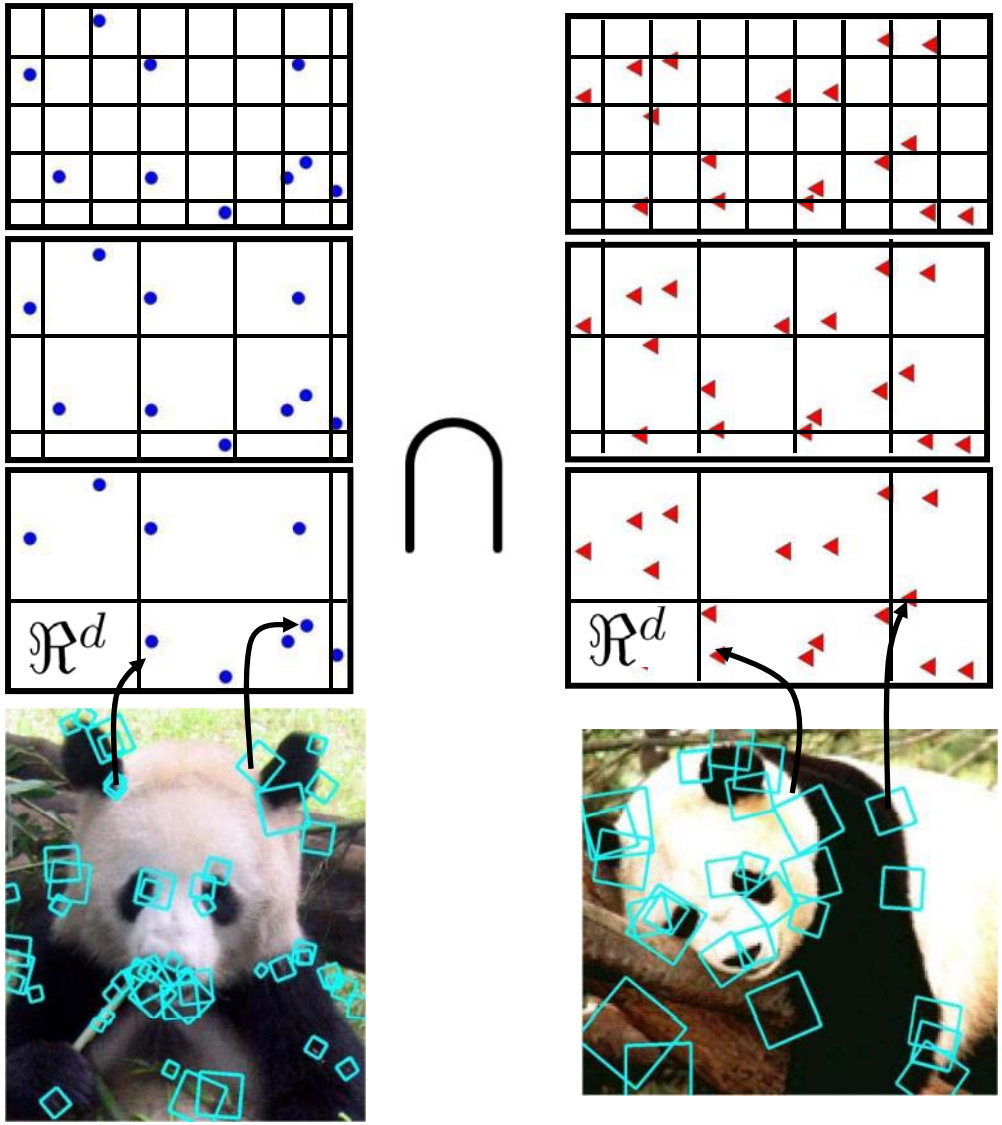
$$\mathbf{X} = \{\vec{\mathbf{x}}_1, \ldots, \vec{\mathbf{x}}_m\}$$

$$\mathbf{Y} = \{\vec{\mathbf{y}}_1, \ldots, \vec{\mathbf{y}}_n\}$$

Pyramid Match

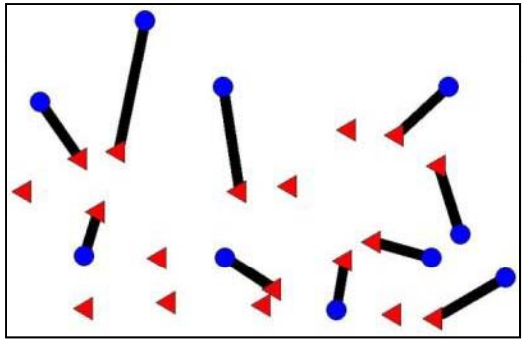
- Optimal matching $O(dm^3)$
- Greedy matching $O(dm^2 \log m)$
- Pyramid match $O(dmL)$

for sets with $O(m)$ features of dimension d



$\cap$

$\approx$



optimal partial matching

$$\max_{\pi:\mathbf{X}\rightarrow\mathbf{Y}} \sum_{\mathbf{x}_i \in \mathbf{X}} \mathcal{S}(\mathbf{x}_i, \pi(\mathbf{x}_i))$$

$$\mathbf{X} = \{\vec{\mathbf{x}}_1, \ldots, \vec{\mathbf{x}}_m\} \quad \mathbf{Y} = \{\vec{\mathbf{y}}_1, \ldots, \vec{\mathbf{y}}_n\}$$

[Grauman and Darrell, ICCV 2005, JMLR 2007]

Gaussian Process PMK

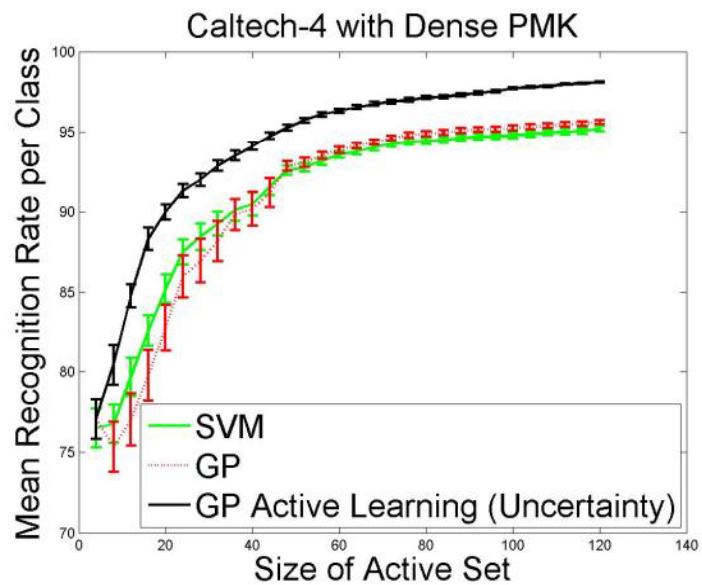
- The Pyramid Match defines a Mercer Kernel suitable for SVM and Gaussian Process based regression and classification

$$K_\Delta\left(\Psi(\mathbf{X}),\Psi(\mathbf{Y})\right) = \sum_{i=0}^{L}\frac{1}{2^i}\Big(\mathcal{I}\left(H_i(\mathbf{X}),H_i(\mathbf{Y})\right)-\mathcal{I}(H_{i-1}(\mathbf{X}),H_{i-1}(\mathbf{Y}))\Big)$$

- GP-based classification offers a natural paradigm for Active Learning:

Method	Criteria
Distance from Boundary (SVM)	$\mathbf{x}^* = \arg\min_{\mathbf{x}_u \in \mathbf{x}_U} \lvert\bar{y}_u\rvert$
Variance	$\mathbf{x}^* = \arg\max_{\mathbf{x}_u \in \mathbf{x}_U} \Sigma_u$
Uncertainty (GP)	$\mathbf{x}^* = \arg\min_{\mathbf{x}_u \in \mathbf{x}_U} \frac{\lvert\bar{y}_u\rvert}{\sqrt{\Sigma_u+\sigma^2}}$

Active Learning Criteria



[Kapoor, Grauman, Urtasun and Darrell, ICCV 2007, IJCV 2009]

Multiple Kernel Learning (MKL)

- Most approaches assume a global weighting over views [Bach '04][Varma '07]
- Global approaches have difficulty learning from insufficient views corrupted by heteroscedastic noise, containing missing data, or whose discriminative properties vary across the input space
- Localized approaches have been proposed within SVM learning frameworks [Lin et al '07][Gonan & Alpaydin '08]

Our Approach: Localized MKL

- Present a Bayesian framework for localized MKL with GPs
- Assume a GP prior with covariance

$$\bar{\mathbf{K}} = \sum_{v}^{V} \mathbf{K}^{(v)},$$

where $\mathbf{K}^{(v)}$ is computed over the v-th view and is defined by the product of two kernels

$$K_{ij}^{(v)} = k_{np}^{(v)}(i,j) \cdot k_{p}^{(v)}(\mathbf{x}_i^{(v)}, \mathbf{x}_j^{(v)})$$

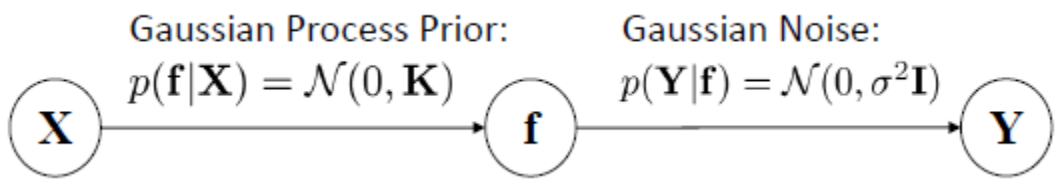
- Assume a rank-constrained non-parameteric covariance matrix

$$\mathbf{K}_{np}^{(v)} = (\mathbf{g}^{(v)})^T \mathbf{g}^{(v)},$$

[See http://www.eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-96.html for details…]

Gaussian Process Regression

- A Bayesian approach for regression that assumes a GP prior over the space of functions
- Training data: $D = \{(\mathbf{x}_i, \mathbf{y}_i), i = 1, \cdots, N\}$



- Marginalize latent $\mathbf{f}$: can be done in closed form to yield a Gaussian predictive distribution, $p(\mathbf{f}_*|\mathbf{Y}) = \mathcal{N}(\mathbf{M}, \mathbf{C})$, with

$$\mathbf{M} = \mathbf{K}_{*,f}(\mathbf{K}_{f,f} + \sigma^2_{noise}\mathbf{I})^{-1}\mathbf{Y}$$
$$\mathbf{C} = \mathbf{K}_{*,*} - \mathbf{K}_{*,f}(\mathbf{K}_{f,f} + \sigma^2_{noise}\mathbf{I})^{-1}\mathbf{K}_{f,*}$$

Non-Parameteric Covariance Representation

- Assume a rank-constrained non-parameteric covariance matrix

$$\mathbf{K}_{np}^{(v)} = (\mathbf{g}^{(v)})^T \mathbf{g}^{(v)},$$

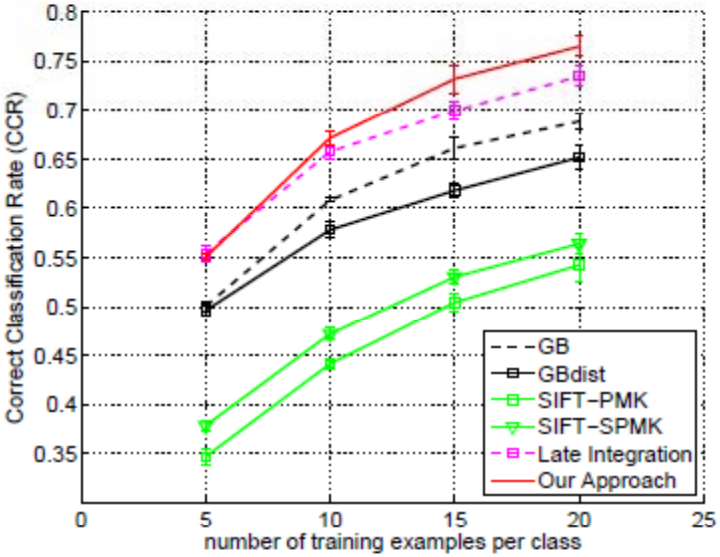
  where $\mathbf{g} = [\mathbf{g}_1, \cdots, \mathbf{g}_N]^T \in \mathcal{R}^{m \times N}$, and $m \ll N$.
- With $m = 1$ $g_j^{(v)}$ is a scalar value that measures sample confidence
- Assuming piecewise smoothness over input space weighting, and cluster space can further reduce number of model parameters,

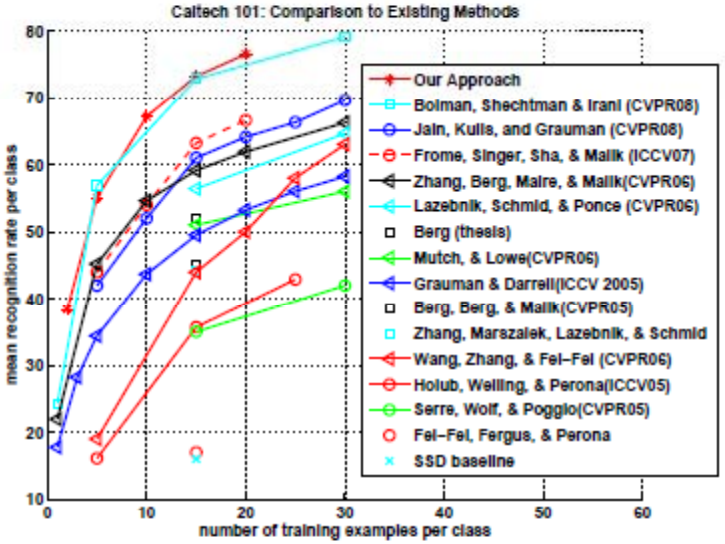
$$g_j^{(v)} = \boldsymbol{\alpha}^{(v)} \cdot \mathbf{e}_j$$

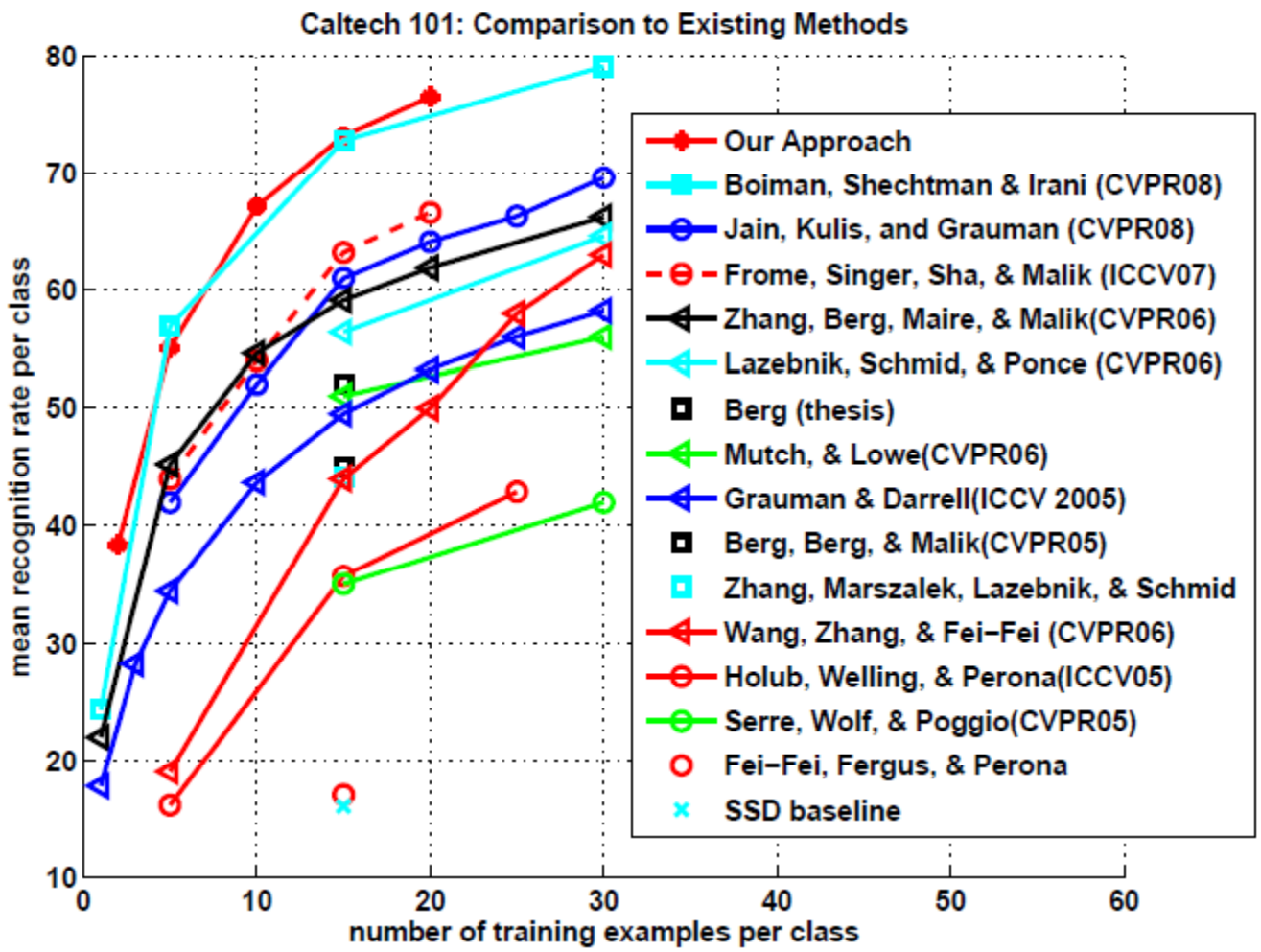
  where $\mathbf{e}_j \in 0, 1^{P \times 1}$ is an indicator of the cluster example j belongs to, $\boldsymbol{\alpha}^{(v)} \in \mathcal{R}^{1 \times P}$, P is the number of clusters.

Results: Object Category Classification

- Caltech-101 benchmark dataset
- Use four input views: geometric blur with and without distortion term [Zhang et al '06], and dense SIFT features with PMK [Grauman & Darrell '05] and spatial PMK [Lazebnik et al '06] similarity measures

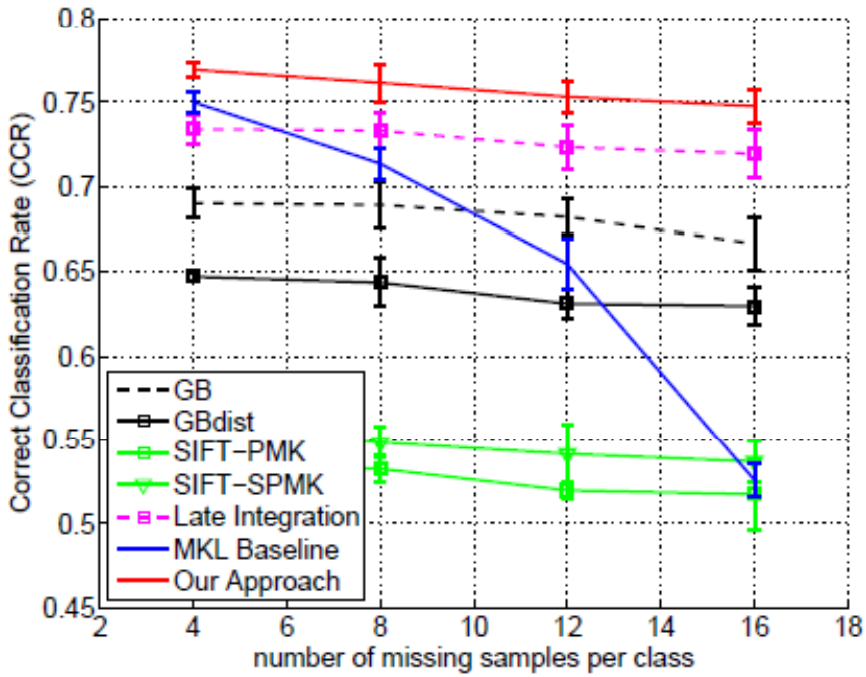





Caltech 101: Comparison to Existing Methods
mean recognition rate per class
number of training examples per class
80
70
60
50
40
30
20
10
0
10
20
30
40
50
60
Our Approach
Boiman, Shechtman & Irani (CVPR08)
Jain, Kulis, and Grauman (CVPR08)
Frome, Singer, Sha, & Malik (ICCV07)
Zhang, Berg, Maire, & Malik(CVPR06)
Lazebnik, Schmid, & Ponce (CVPR06)
Berg (thesis)
Mutch, & Lowe(CVPR06)
Grauman & Darrell(ICCV 2005)
Berg, Berg, & Malik(CVPR05)
Zhang, Marszalek, Lazebnik, & Schmid
Wang, Zhang, & Fei−Fei (CVPR06)
Holub, Welling, & Perona(ICCV05)
Serre, Wolf, & Poggio(CVPR05)
Fei−Fei, Fergus, & Perona
SSD baseline

Results: Missing Data

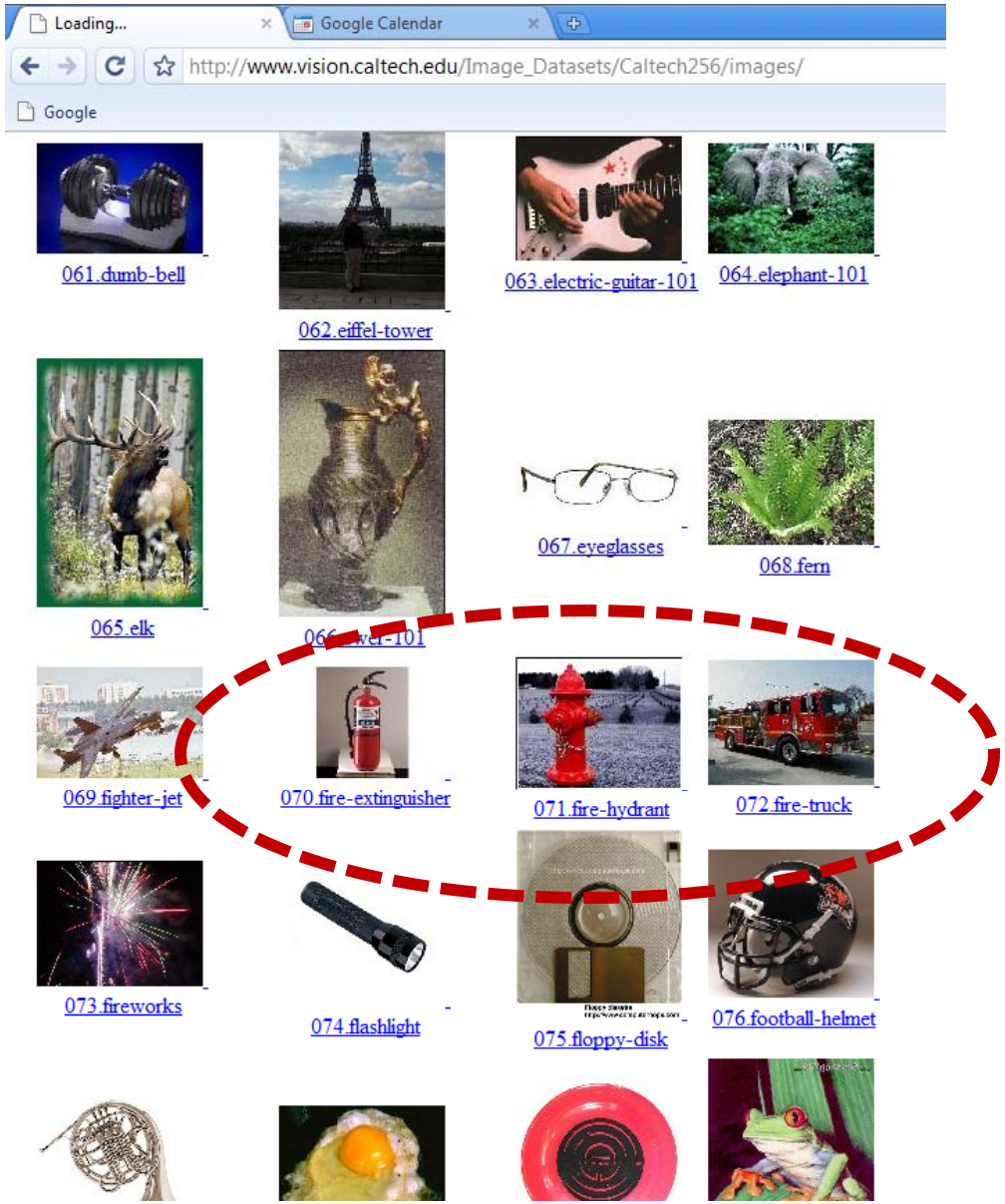
- Simulated missing data with Caltech-101, where we discard at most one view per sample in the training set



Today: Snapshots

- ✓ Probabilistic multi-kernel fusion
- Joint regularization across categories
- Multimodal sense grounding
- Local feature models for transparent objects

Standard “1 vs. all” paradigm….



SVM/GPC – Category 1

SVM/GPC – Category 2

SVM/GPC – Category 3

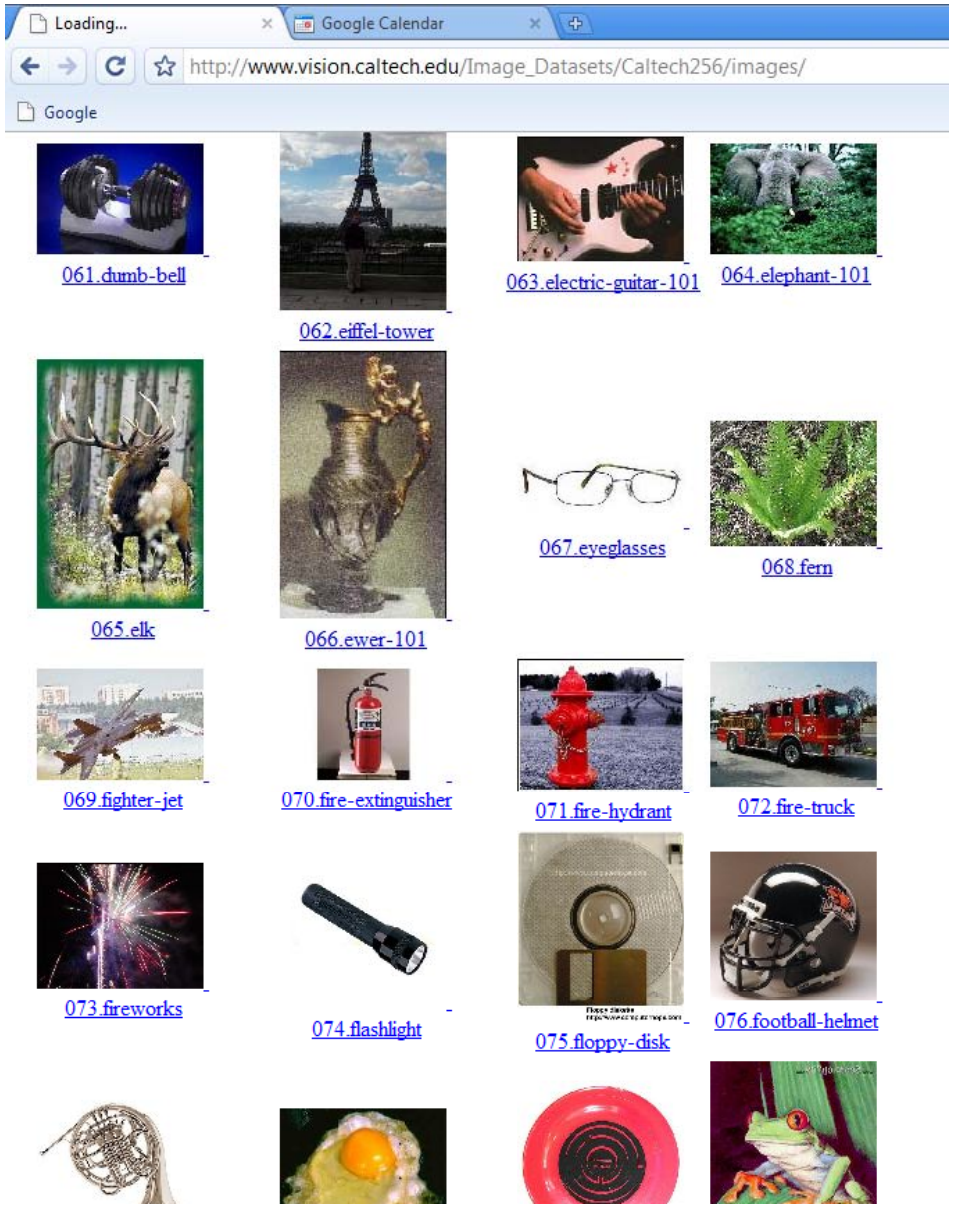
SVM/GPC – Category 4

⋮

SVM/GPC – Category 10,000?

How to exploit shared structure?

Consider ensemble of classifiers



classifier weights

SVM/GPC – Category 1 w_1

SVM/GPC – Category 2 w_2

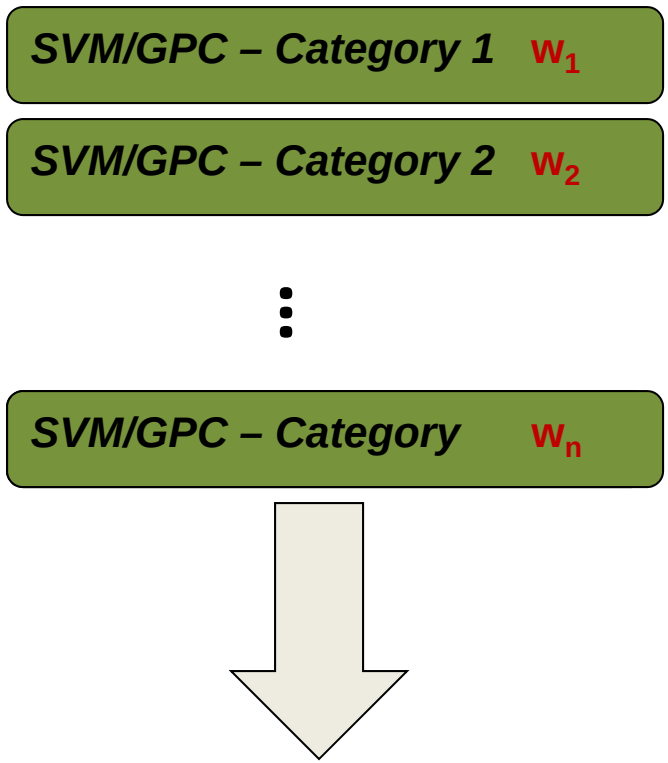
SVM/GPC – Category 3 w_3

SVM/GPC – Category 4 w_4

⋮

SVM/GPC – Category $w_{10,000}$

Consider ensemble of classifiers



$$W = [\ w_1\ w_2 \ldots w_n\ ]$$

Related tasks and/or object part structure will lead to correlated patterns in W…

[Quattoni, Collins, Darrell, CVPR 2007] explore Ando+Zhang style structure learning for scene recognition tasks.

Learn W jointly?

[Quattoni, Collins, Darrell, CVPR 2008] explore joint spare optimization via matrix norm penalty.

[Quattoni, Carreras, Collins, Darrell, ICML 2009] report an efficient learning scheme for this approach…

Joint Sparse Approximation

- Consider learning a single sparse linear classifier of the form:

$$f(x) = w \cdot x$$

That is, we want only a few features with non-zero coefficients

- L_1 regularization well-known to yield sparse solutions:

$$\min_{\mathbf{w}} \underbrace{\sum_{(x,y)\in D} l(f(x), y)}_{\text{Classification error}} + \underbrace{C\sum_{j=1}^{d} |w_j|}_{L_1 \text{ penalizes non-sparse solutions}}$$

Joint Sparse Approximation

Optimization over several tasks ***jointly:***

$$f_k(x) = \mathbf{w}_k \cdot x$$

$$\min_{\mathbf{w_1},\mathbf{w_2},\ldots,\mathbf{w_m}} \underbrace{\sum_{k=1}^{m} \frac{1}{|D_k|} \sum_{(x,y)\in D_k} l\,(f_k(x), y)}_{\text{Average Loss on training set k}} + \underbrace{C\,\mathrm{R}(\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_m)}_{\text{penalizes solutions that utilize too many features}}$$

Key idea: use a matrix norm…

[Obozinski et al. 2006, Argyriou et al. 2006, Amit et al. 2007]

Joint Regularization Penalty

How do we penalize solutions that use too many features?

$$\begin{bmatrix} W_{1,1} & W_{1,2} & \cdots & W_{1,m} \\ W_{2,1} & W_{2,2} & \cdots & W_{2,m} \\ \vdots & \ddots & \ddots & \vdots \\ W_{d,1} & W_{d,2} & \cdots & W_{d,m} \end{bmatrix}$$

Coefficients for for feature 2

Coefficients for classifier 2

$$\mathrm{R}(W) = \# \, non - zero - rows$$

Would lead to a hard combinatorial problem .

Joint Regularization Penalty

We use a $L_{1\text{-}\infty}$ norm [Tropp 2006]

$$R(W) = \sum_{i=1}^{d} \max_{k}(|W_{ik}|)$$

This norm combines:

The L_∞ norm on each row promotes non-sparsity on the rows. → Share features

An L_1 norm on the maximum absolute values of the coefficients across tasks promotes sparsity. → Use few features

The combination of the two norms results in a solution where only a few features are used but the features used will contribute in solving many classification problems.

Joint Sparse Approximation

Using the $L_{1\text{-}\infty}$ norm we can rewrite our objective function as:

$$\min_{\mathbf{W}} \sum_{k=1}^{m} \frac{1}{|D_k|} \sum_{(x,y)\in D_k} l\,(f_k(x), y) + C \sum_{i=1}^{d} \max_k (|W_{ik}|)$$

For any convex loss this is a convex objective.

For the hinge loss the optimization problem can be expressed as a linear program. [Quattoni et al. CVPR 2008]

See also [Quattoni et al ICML 2009] for efficient large scale solutions.

News Image Classification Experiments

SuperBowl

Sharon

Danish Cartoons

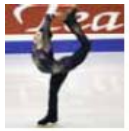
Australian open

Trapped coal miners

Golden globes

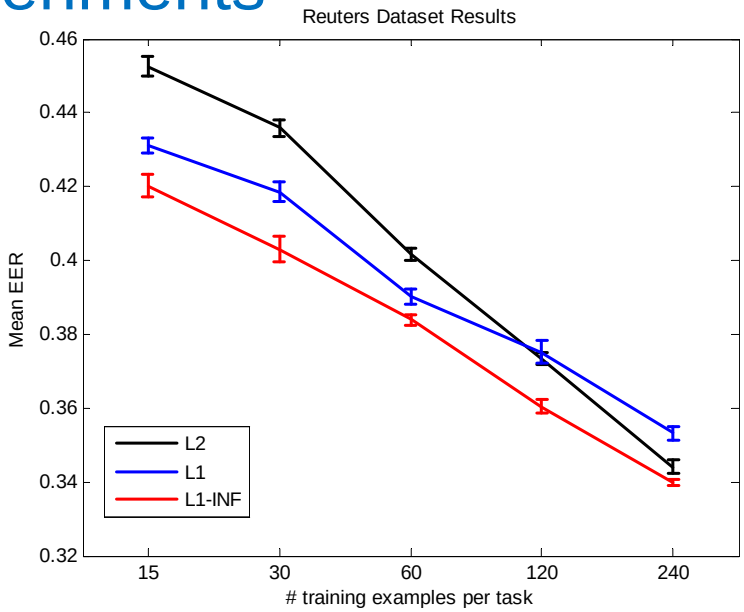
Grammys

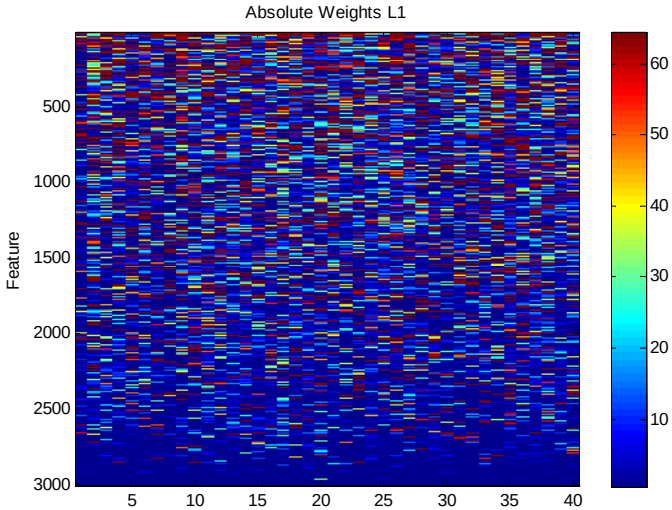
Figure skating



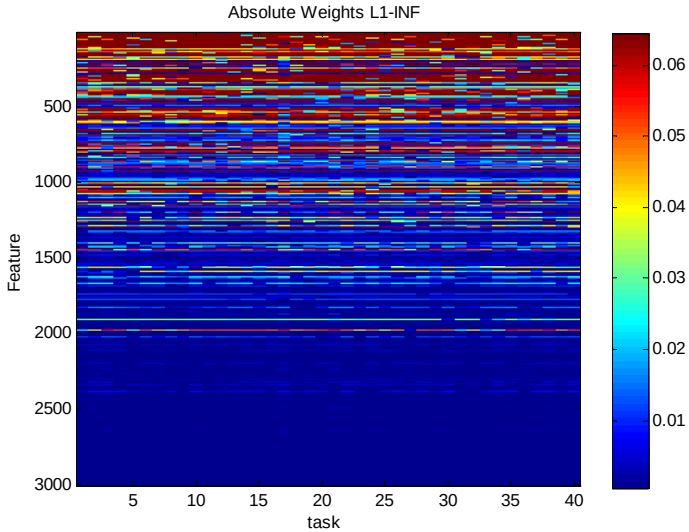
Academy Awards

Iraq





L_1



$L_{1,\infty}$

Today: Snapshots

- ✓ Probabilistic multi-kernel fusion
- ✓ Joint regularization across categories
- Multimodal sense grounding
- Local feature models for transparent objects

Goal: Object recognition in situated environments

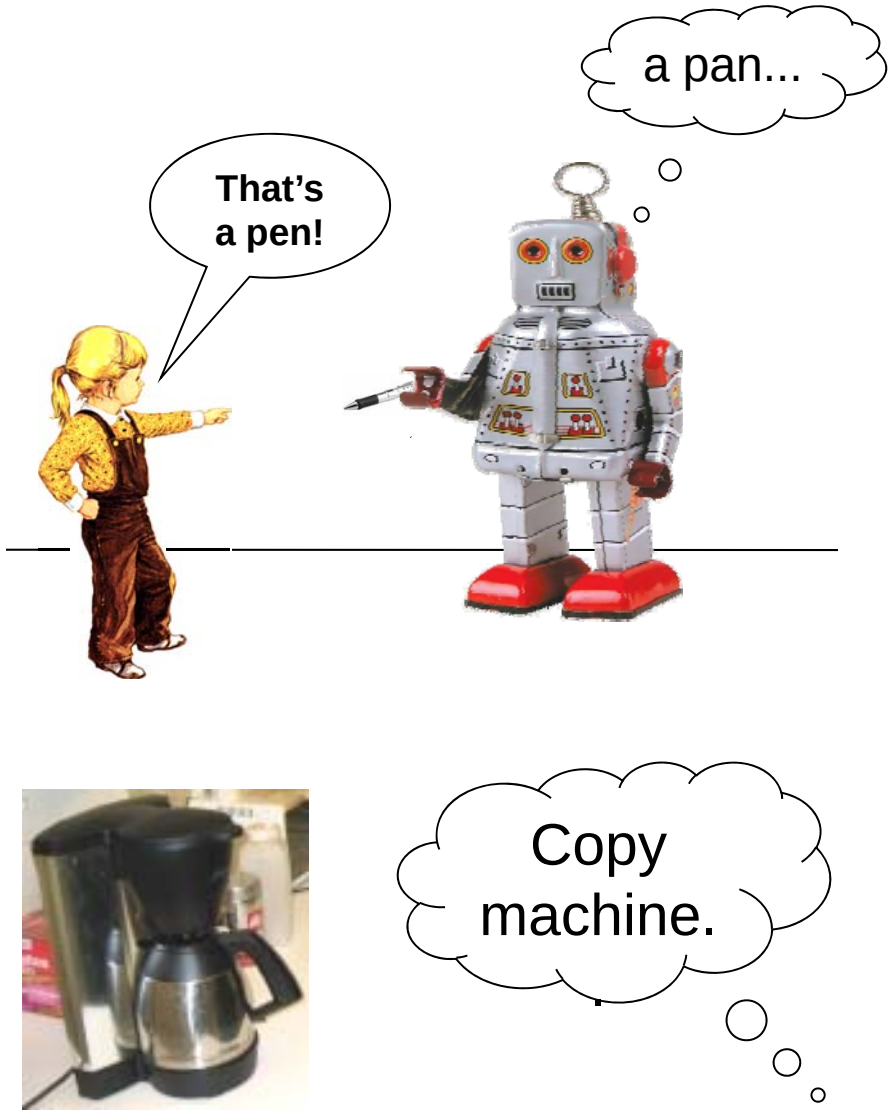
- Imagine using natural dialogue to instantiate object models in a robot

That's a cat over there…

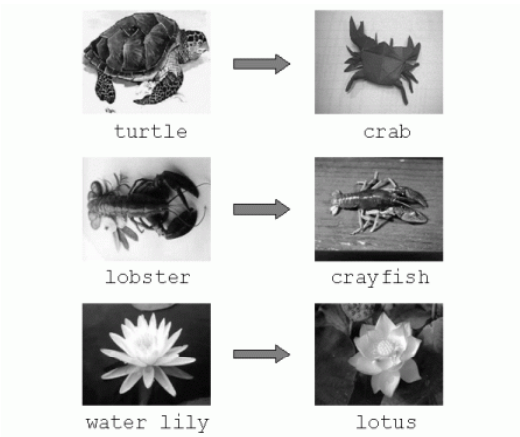
There's a lamp…

This is one of my purses.

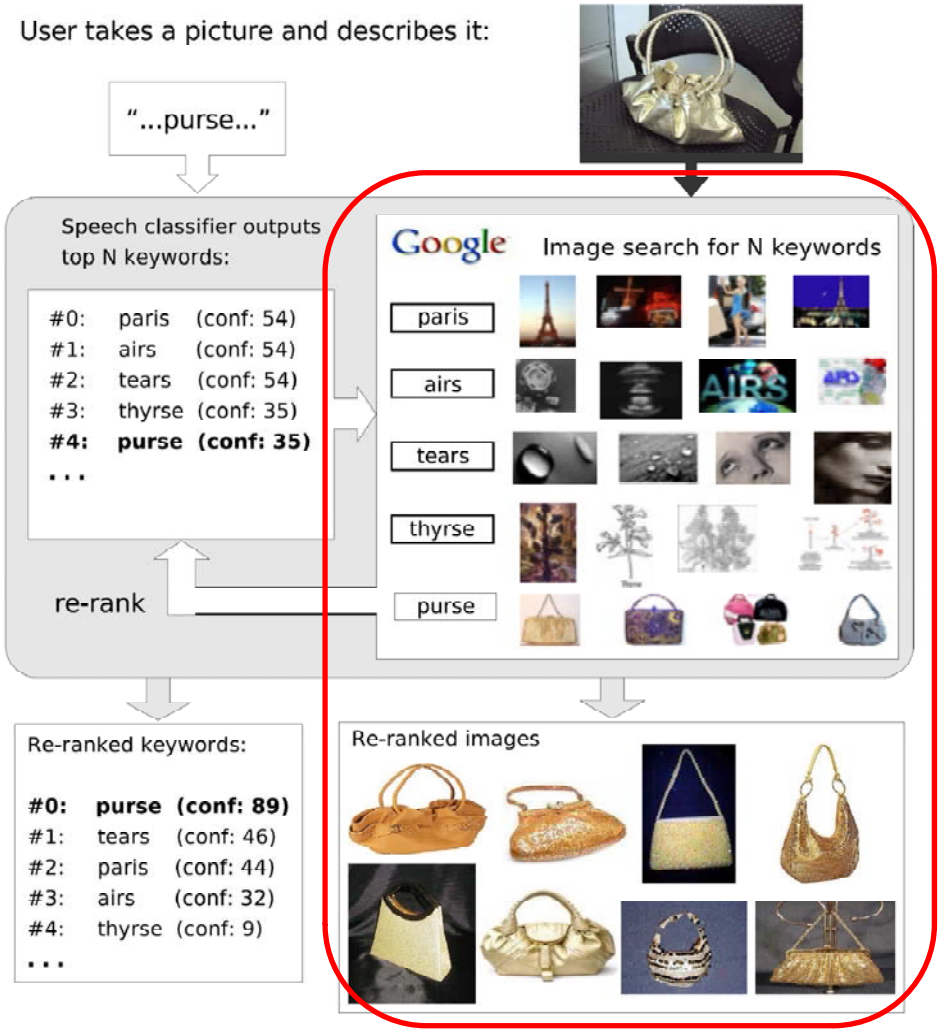
Speech, image can be complementary…



ant → fan
face → bass
piano → cannon

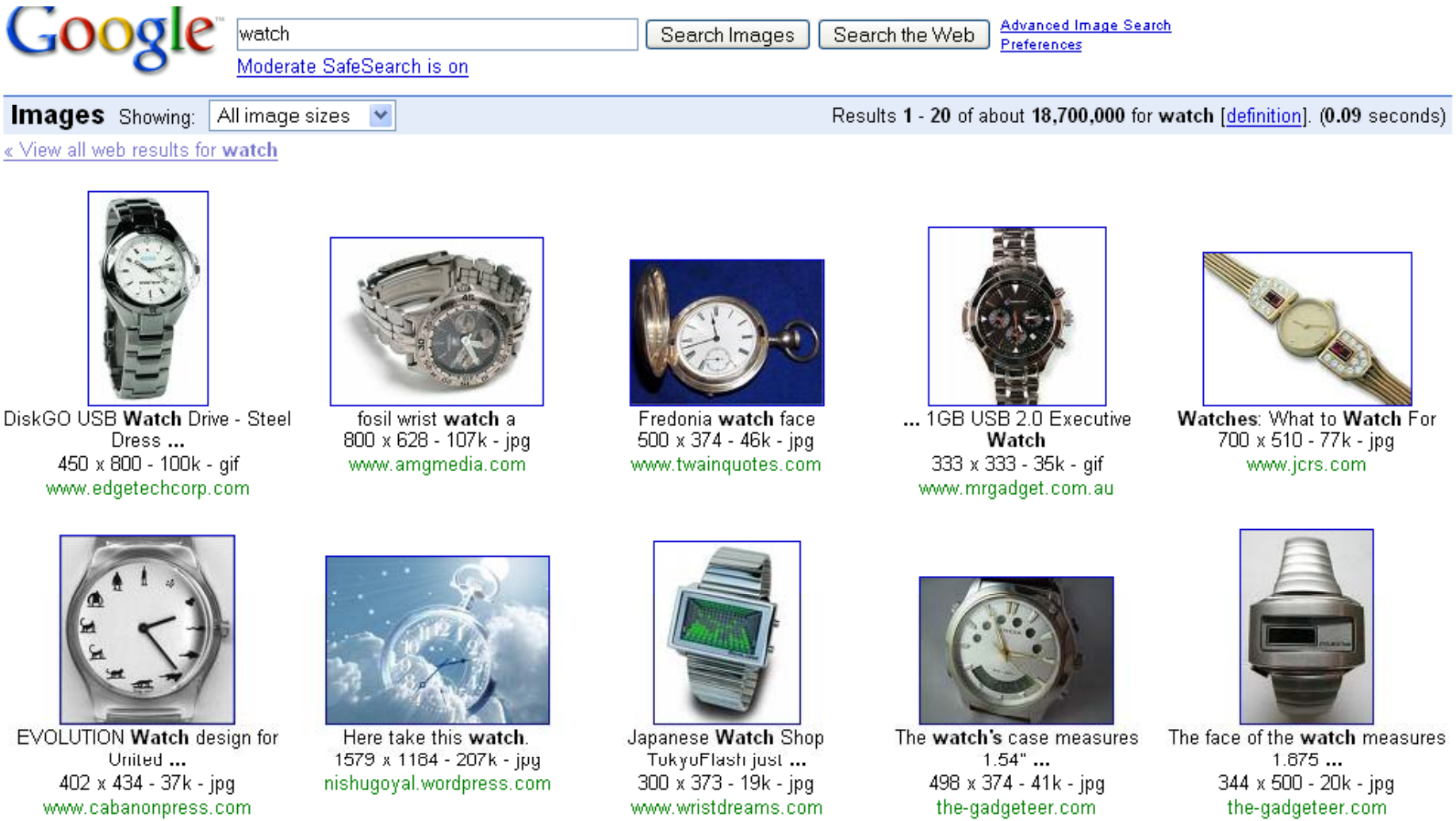


Towards very large object vocabularies…

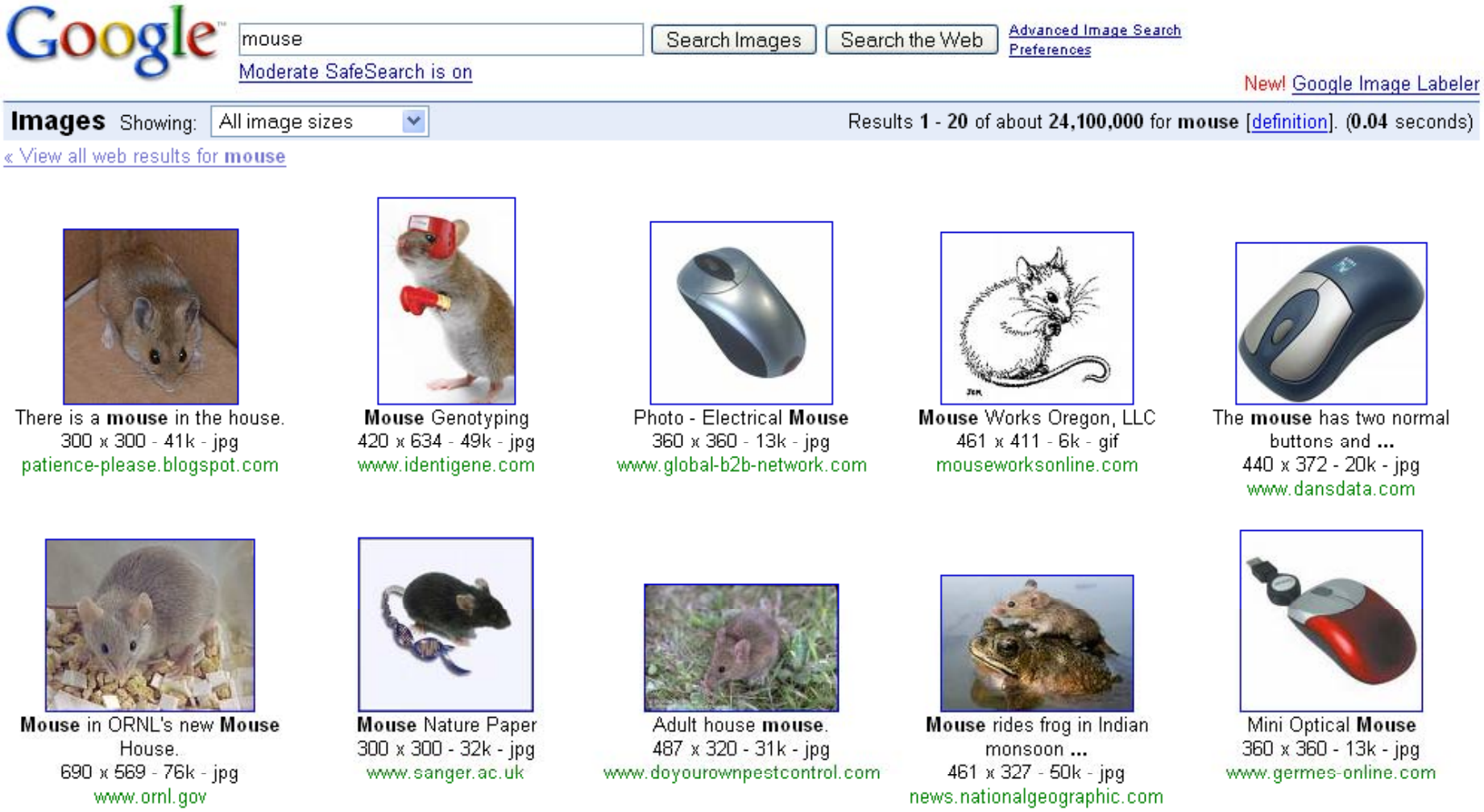


Learn visual models on the fly for N-best audio candidates…

Training images from online image search…



Problem: visual polysemy

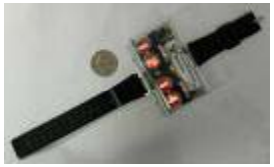


Sources of visual polysemy

Hurricane, tornado watch

Celebrity watch

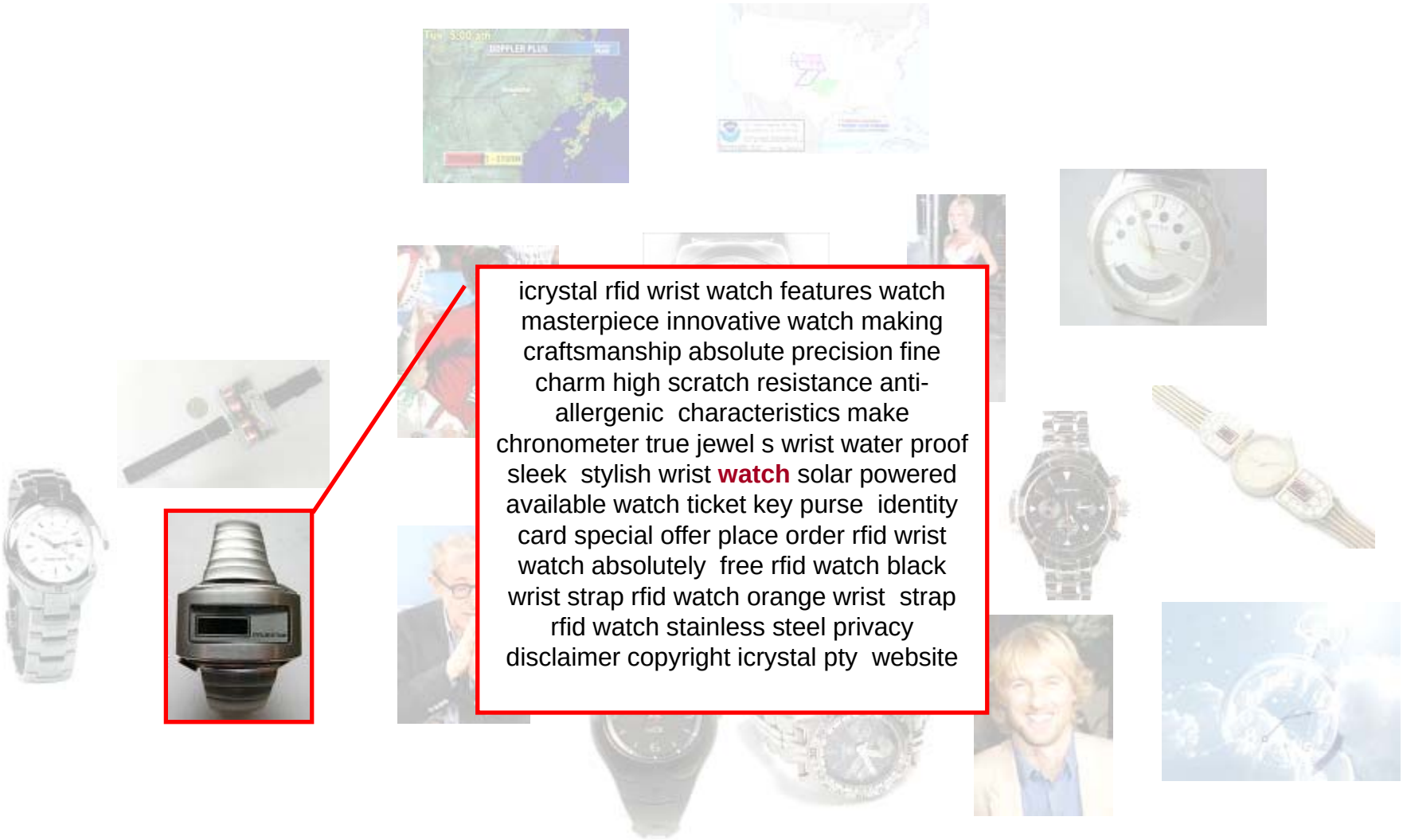
Watch out!



Would rather watch…

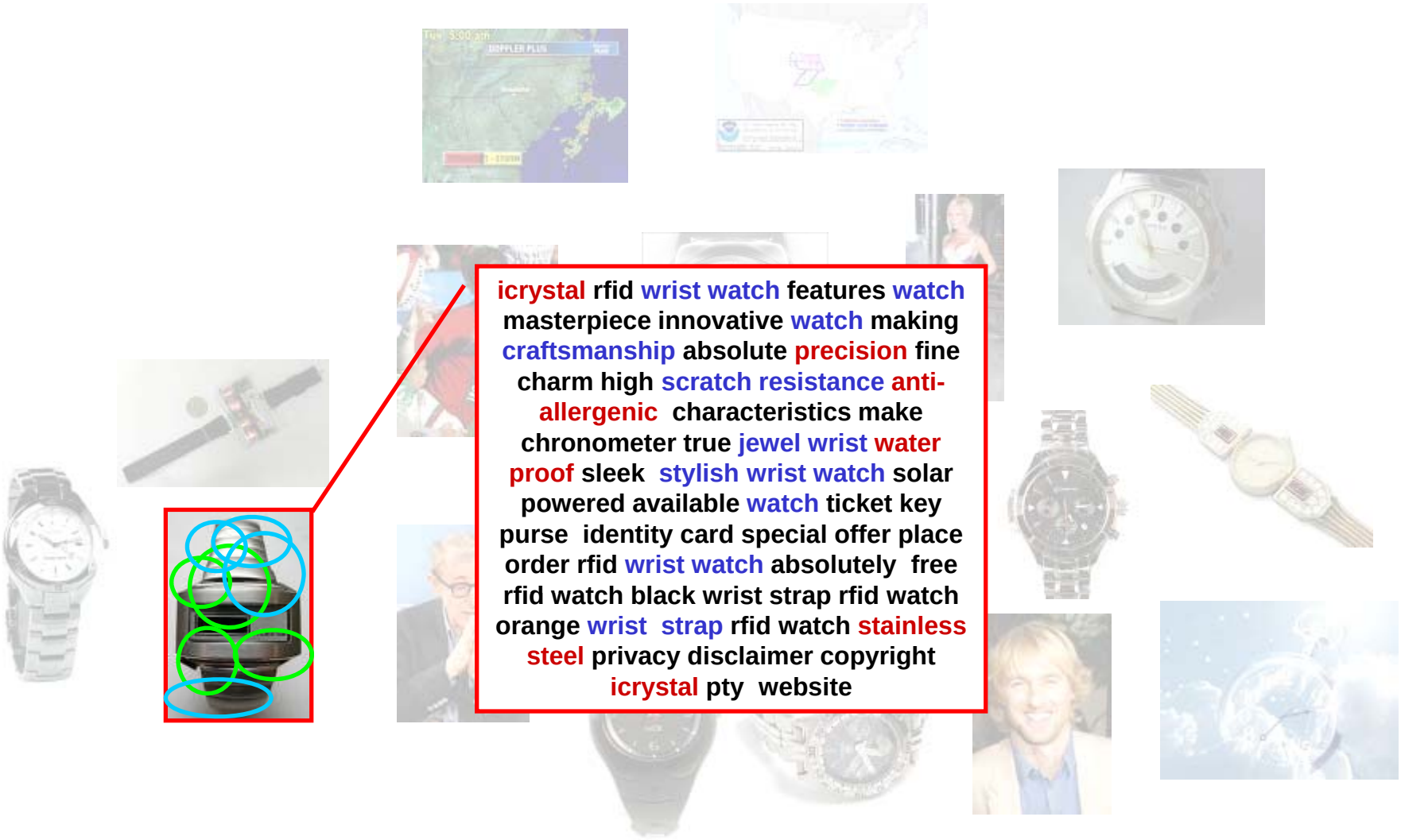
Suicide watch

Take advantage of text contexts



icrystal rfid wrist watch features watch masterpiece innovative watch making craftsmanship absolute precision fine charm high scratch resistance anti-allergenic characteristics make chronometer true jewel s wrist water proof sleek stylish wrist **watch** solar powered available watch ticket key purse identity card special offer place order rfid wrist watch absolutely free rfid watch black wrist strap rfid watch orange wrist strap rfid watch stainless steel privacy disclaimer copyright icrystal pty website

Latent Topics



Overview of approaches to web-based object model learning

- Some learn only from image features
  - (Li et al.07) bootstrap from labeled images
  - (Fergus et al.05) select correct image topic
- Some incorporate text features
  - (Schroff et al.07) use a category-independent text classifier
  - (Berg and Forsyth 06) ask user to sort text topics
- None address polysemy directly
  - (Loeff et al.06) do image sense discrimination, not identification
- All rely on labeled images of correct sense

WISDOM: Using dictionary entries to ground senses

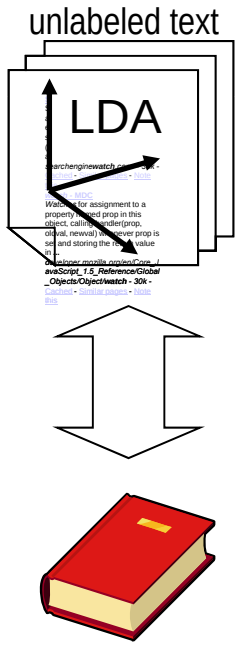
- Use entry text to learn a probability distribution over words for that sense
- Problem: entries contain very little text
  - Expand by adding synonyms, example sentences, etc.
  - Still, very few words are covered!

•S: (n) **mouse** (any of numerous small rodents typically resembling diminutive rats having pointed snouts and small ears on elongated bodies with slender usually hairless tails)
•*direct hyponym* / *full hyponym*
  •S: (n) house mouse, Mus musculus (brownish-grey Old World mouse now a common household pest worldwide)
  •S: (n) harvest mouse, Micromyx minutus (small reddish-brown Eurasian mouse inhabiting e.g. cornfields)
  •S: (n) field mouse, fieldmouse (any nocturnal Old World mouse of the genus Apodemus inhabiting woods and fields and gardens)
  •S: (n) nude mouse (a mouse with a genetic defect that prevents them from growing hair and also prevents them from immunologically rejecting human cells and tissues; widely used in preclinical trials)
  •S: (n) wood mouse (any of various New World woodland mice)
•***direct hypernym*** / *inherited hypernym* / *sister term*
  •S: (n) rodent, gnawer (relatively small placental mammals having a single pair of constantly growing incisor teeth specialized for gnawing)

WISDOM: Probabilistic dictionary-based model

Main idea:

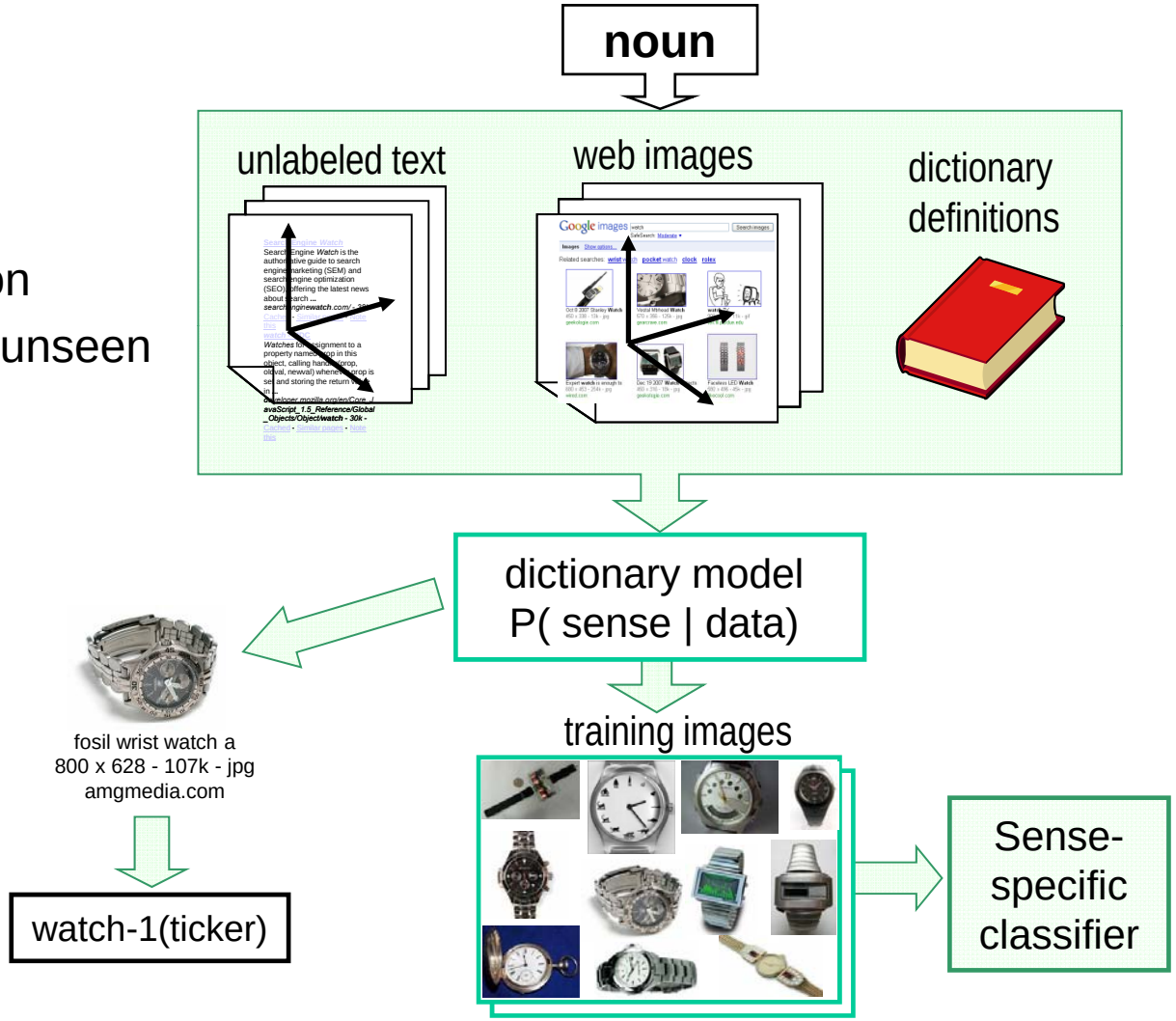
- Using LDA, learn latent sense-like dimensions on large amount of related text,
- Model dictionary senses in LDA space:
  - Map image contexts to topics
  - Map topics to senses



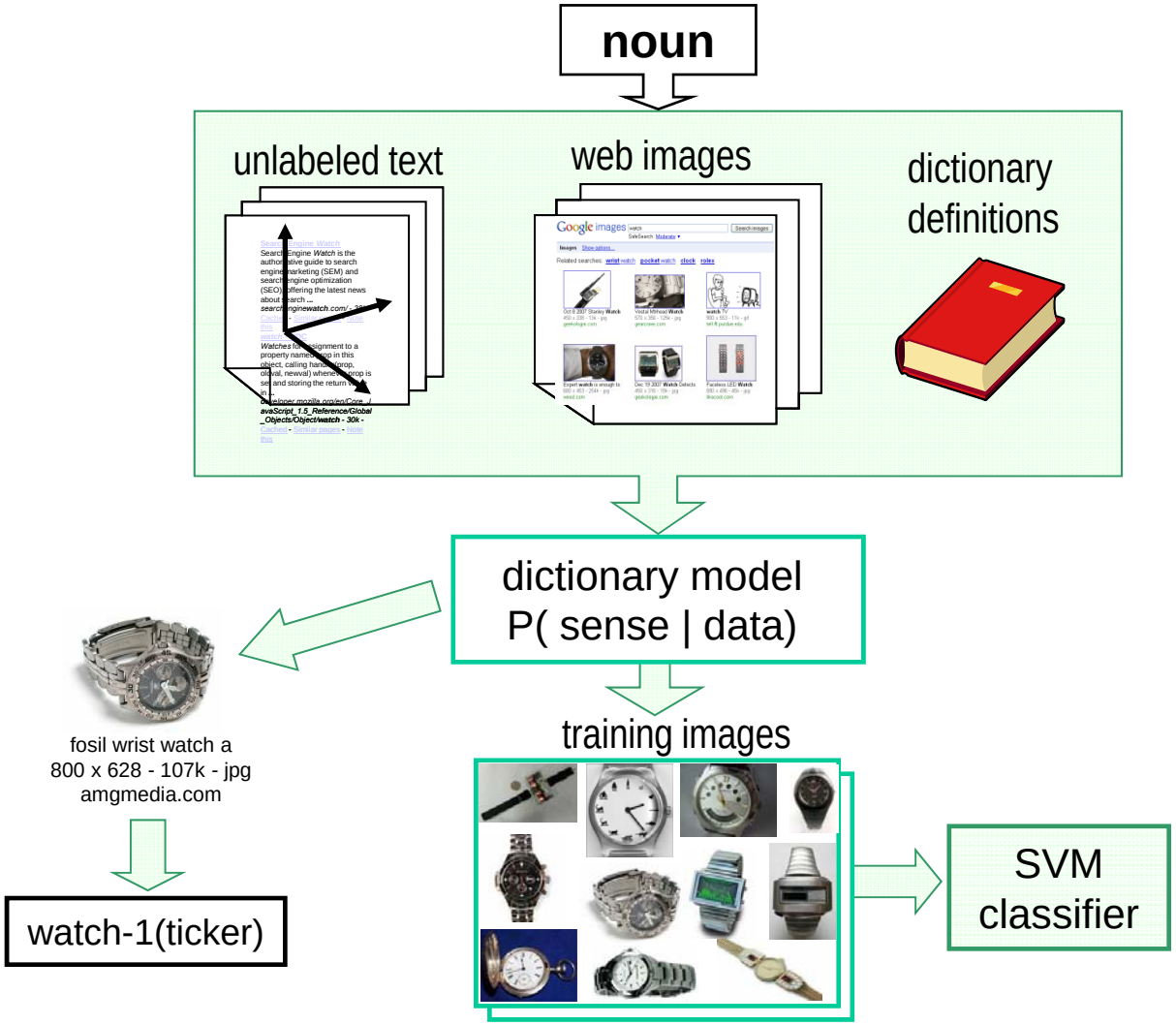
Web Image Sense DictiOnary Model

WISDOM does:

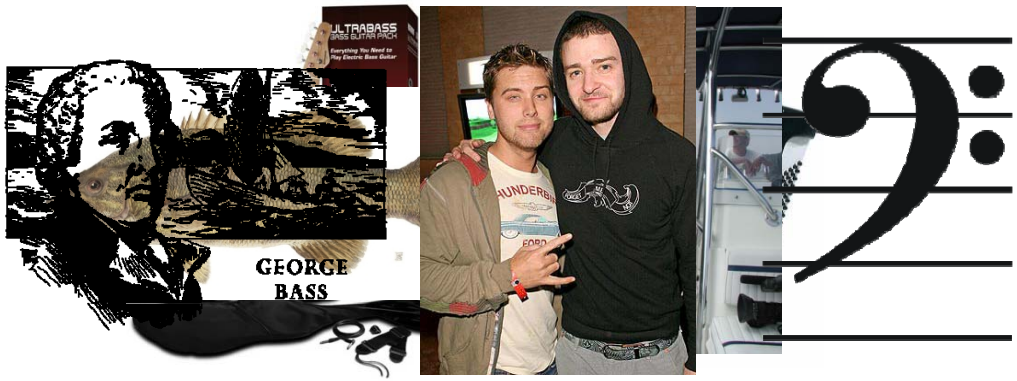
1. image sense disambiguation
2. dataset collection
3. classification of unseen images



WISDOM classifier



Evaluation datasets



unrelated

???

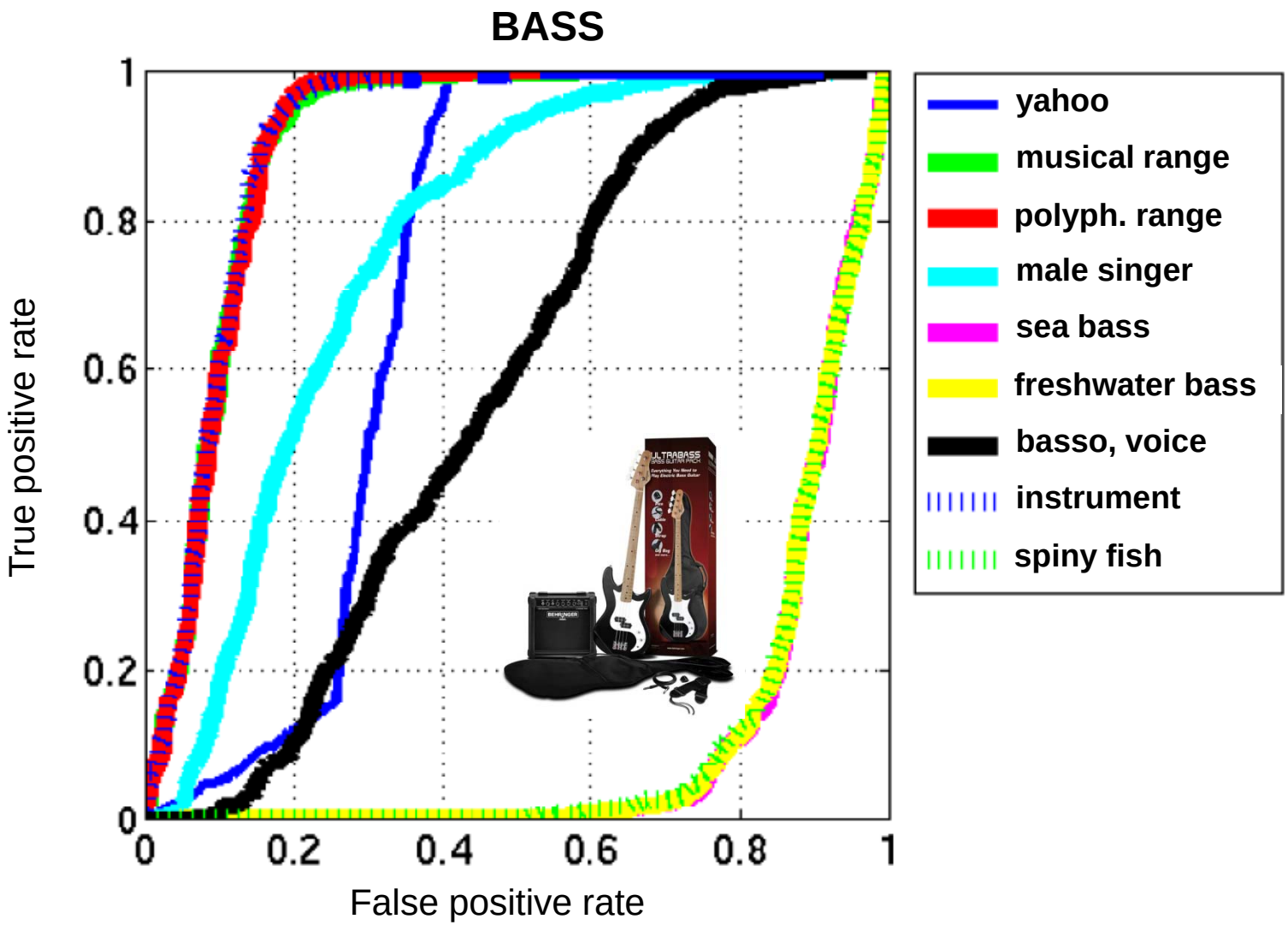
- Collected by querying YAHOO! Image Search
  - MIT-ISD: bass, face, mouse, speaker, watch
  - MIT-OFFICE: cellphone, fork, hammer, keyboard, mug, pliers, scissors, stapler, telephone, watch
  - UIUC-ISD: bass, crane, squash

Experimental Setup

- Task: Image sense disambiguation (ISD) in search results
  - Separate images according to visual sense
  - “core” labels are positive class, “related” and “unrelated” negative
  - Metrics: true positives vs. false positives (ROC), recall-precision curve (RPC)

- Task: object classification in a novel image
  - Classify image as having correct object category or not
  - “core” labels are positive class, other keyword’s “core” senses are negative class

ISD Results: ROC using each WordNet sense for BASS



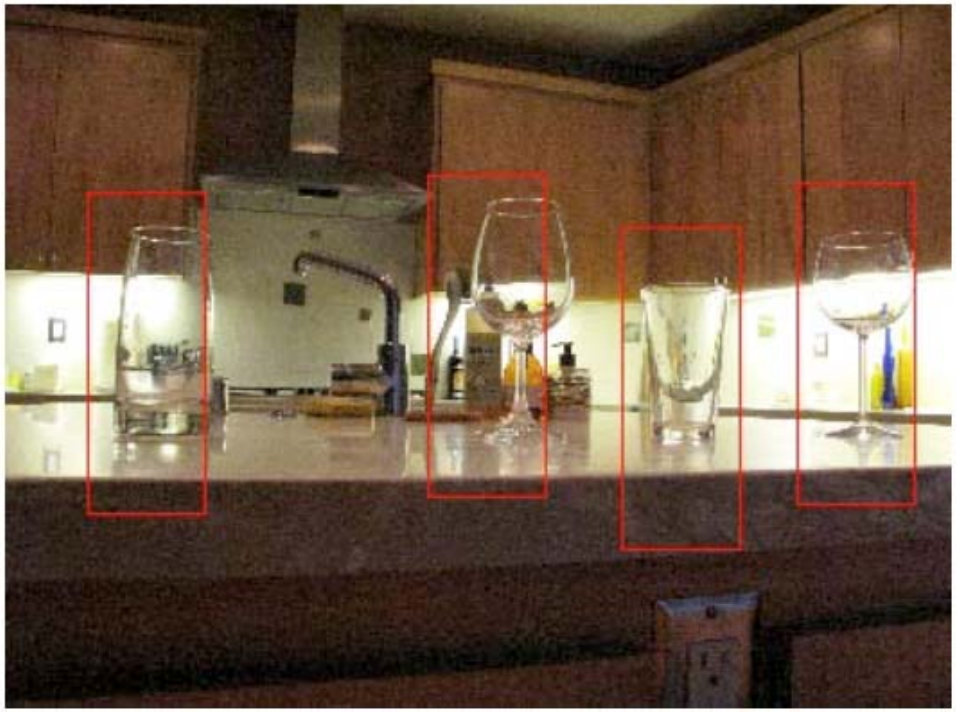
Today: Snapshots

- ✓ Probabilistic multi-kernel fusion
- ✓ Joint regularization across categories
- ✓ Multimodal sense grounding
- Local feature models for transparent objects


Recognizing
Transparent Objects

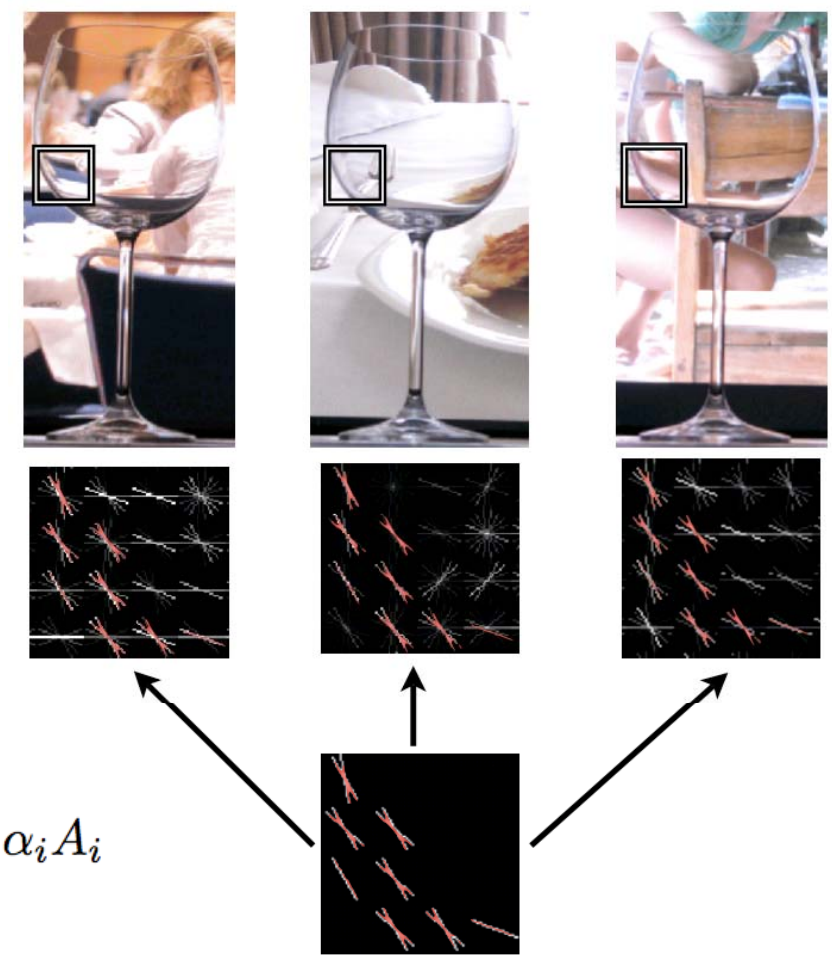
Motivation

- Transparent objects made out of glass or plastic are ubiquitous in domestic environments
- Traditional local feature approach inappropriate
- Full physical model intractable



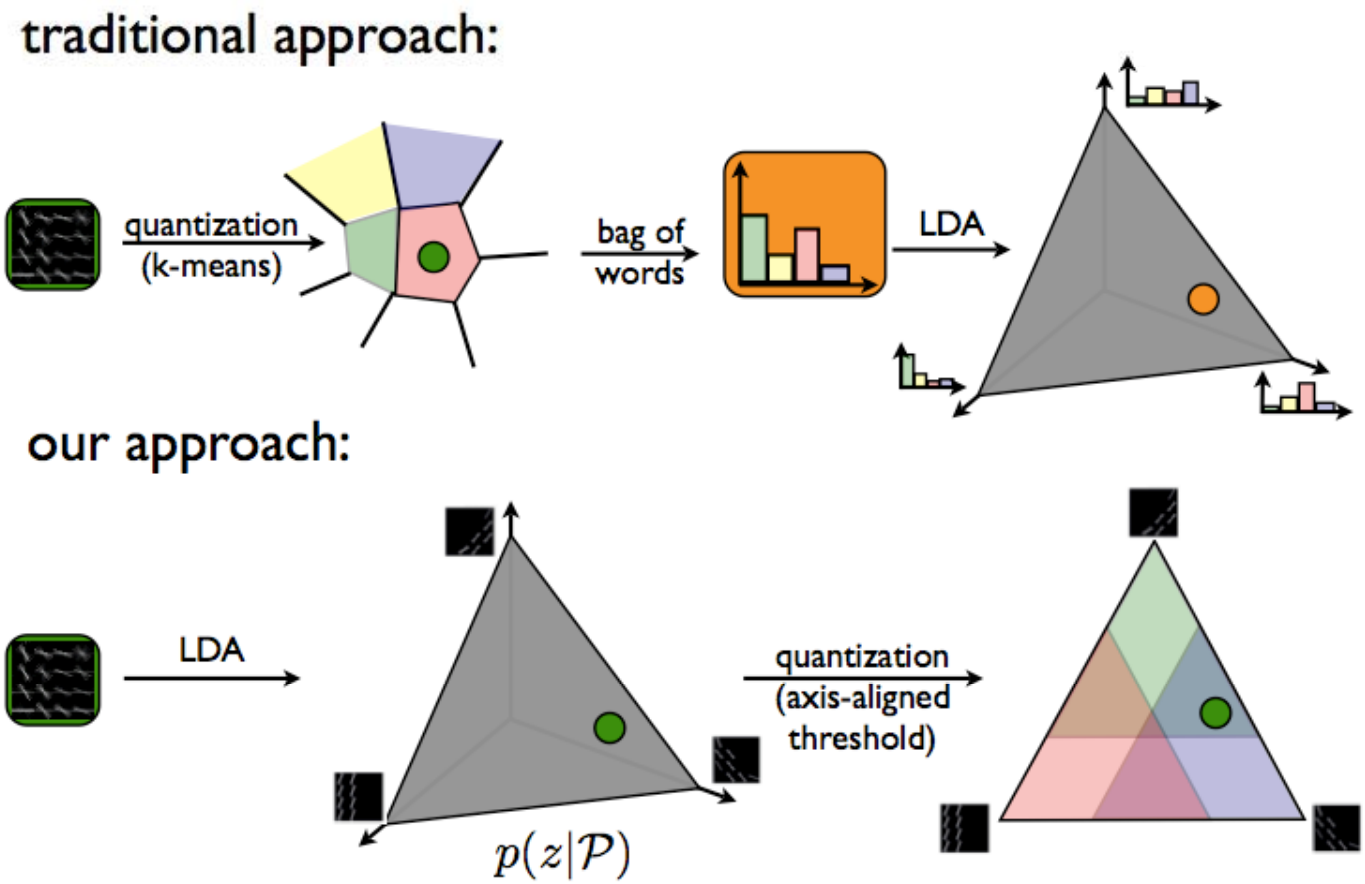
Local Additive Feature Model

- Significant variation in patch appearance
- ... but common latent structure

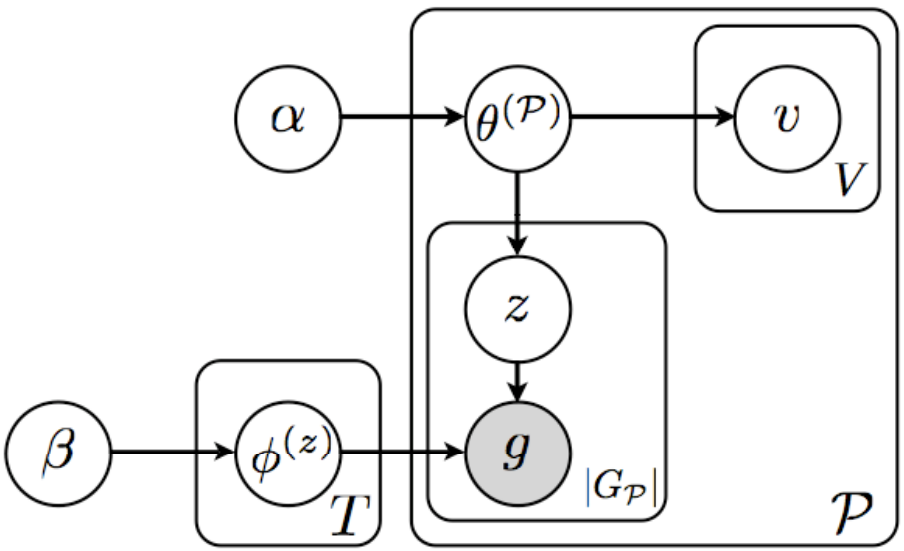


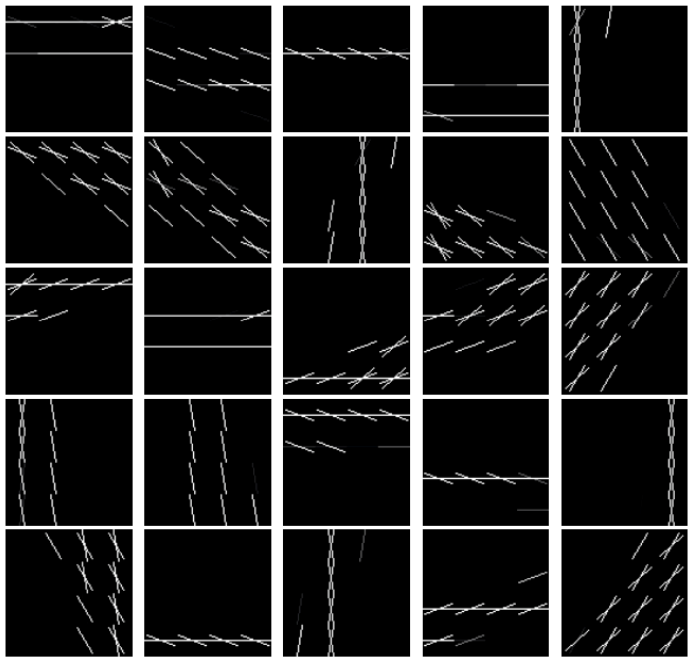
$$G_{\mathcal{P}} = [\, g_{\mathcal{P}}(0,0,0), \ldots, g_{\mathcal{P}}(M,N,T) \,] = \sum_i \alpha_i A_i$$

new LDA-SIFT model

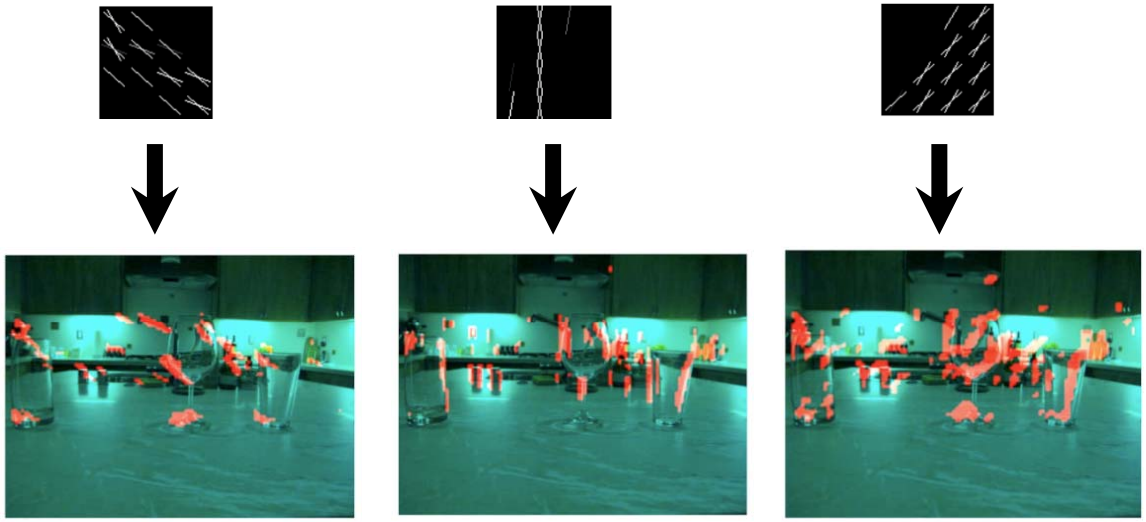


LDA-SIFT





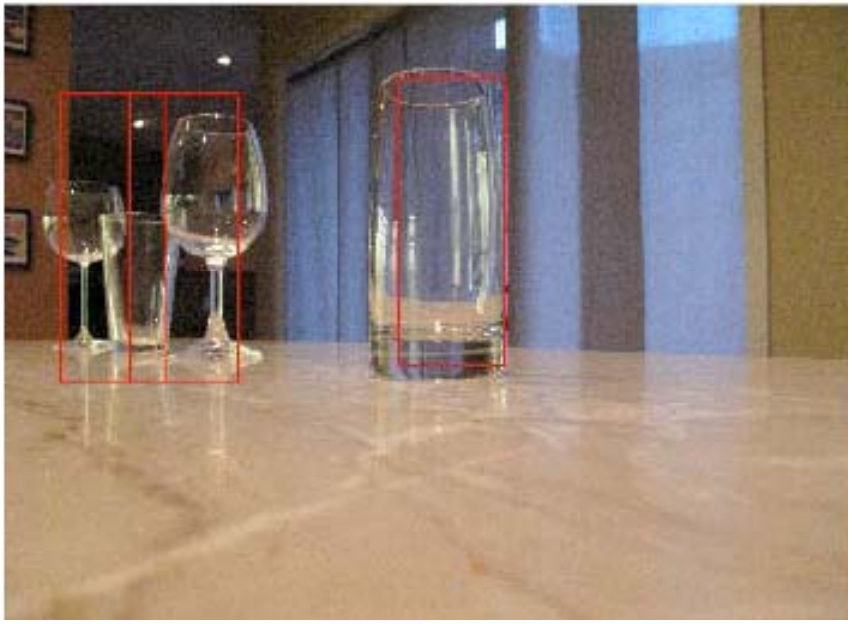
Transparent Visual Words

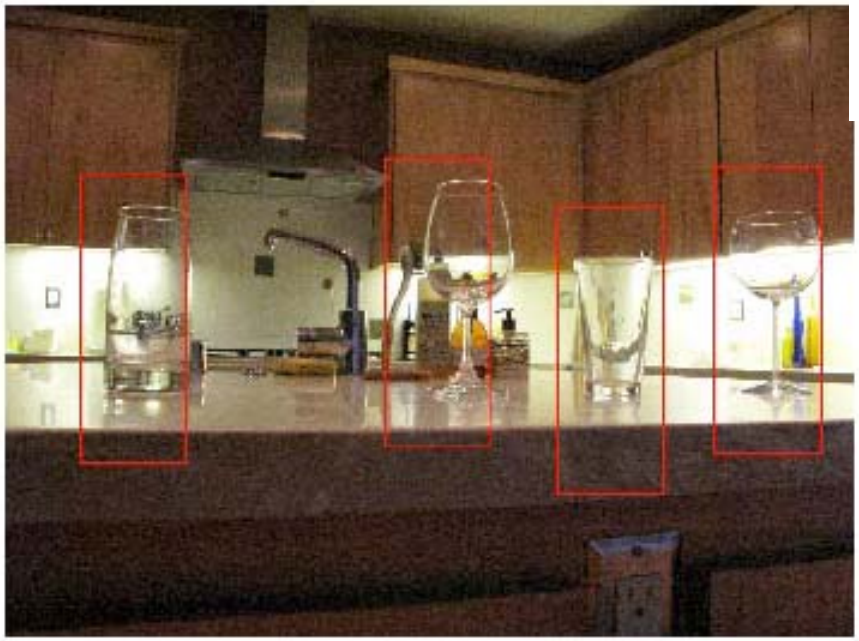


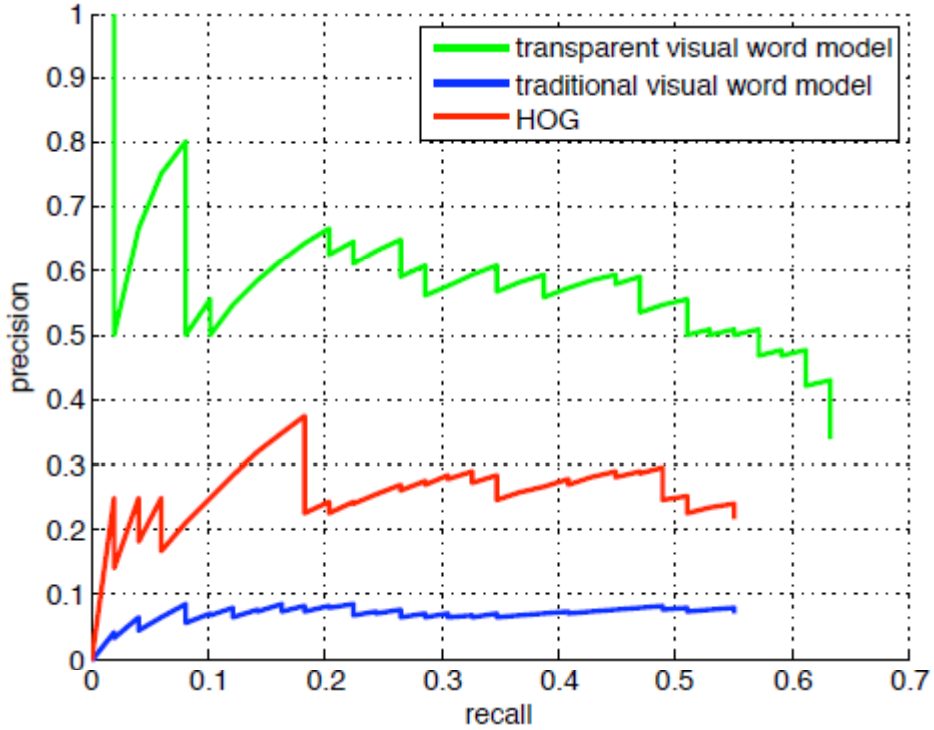
- For each patch we infer the latent mixture activations that characterize the additive structure
- We model the glass by learning a spatial layout of discrete “transparent local feature” activations

Training Data

Example Results







- Training on 4 different glasses in front of screen
- Testing on 49 glass instances in home environment
- Sliding window linear SVM-BOW detection

Overcoming Ambiguity

- Cue saliency varies across categories

- Individual categories have multiple senses

- Multiple surfaces confuse local features

For more information…

- Probabilistic multi-kernel fusion
  - *Christhoudias, Urtasun, Darrell, CVPR 2009*
- Joint regularization across categories
  - *Quattoni, Carreras, Collins, Darrell, ICML 2009.*
- Multimodal sense grounding
  - *Saenko and Darrell, NIPS 2008*
- Local feature models for transparent objects
  - *Fritz, Bradski, Black, and Darrell, in review…*

New ICSI/UCB Vision Group

Prof. Trevor Darrell

Research Scientist

Raquel Urtasun *(→ TTI-C)*

Postdocs

Mario Fritz

Brian Kulis

Mathieu Salzmann

Mario Christhoudias

Kate Saenko (Boston)

Graduate Students

Ashley Eden

Alex Shyr

Trevor Owens

Dave Golland

Carl Ek (Visiting)

Sergey Karayev ('09/'10)

Next BAVM: Berkeley, late Jan. 2010…..date conflicts?