CS294-43: Visual Object and Activity Recognition

Prof. Trevor Darrell

Feb 17th: Generative Object Models

Today

Sudderth guest lecture:

- Constellation Models (Fergus)
- Unsupervised Object Discovery with pLSA (Sivic)
- Scene Models (Li)
- Transformed Models (Sudderth)

Daphna B. student presentation:

- pLSA models of activity (Neibles)

Moreels guest lecture:

- A probabilistic formulation of voting / SIFT (Moreels)

Today

Sudderth guest lecture:

- **Constellation Models (Fergus)**
- Unsupervised Object Discovery with pLSA (Sivic)
- Scene Models (Li)
- Transformed Models (Sudderth)

Daphna B. student presentation:

- pLSA models of activity (Neibles)

Moreels guest lecture:

- A probabilistic formulation of voting / SIFT (Moreels)

Object class recognition using unsupervised scale-invariant learning

Rob Fergus
Pietro Perona
Andrew Zisserman

Oxford University
California Institute of Technology

Slide credit: Fergus

Goal

- Recognition of object categories
- Unassisted learning

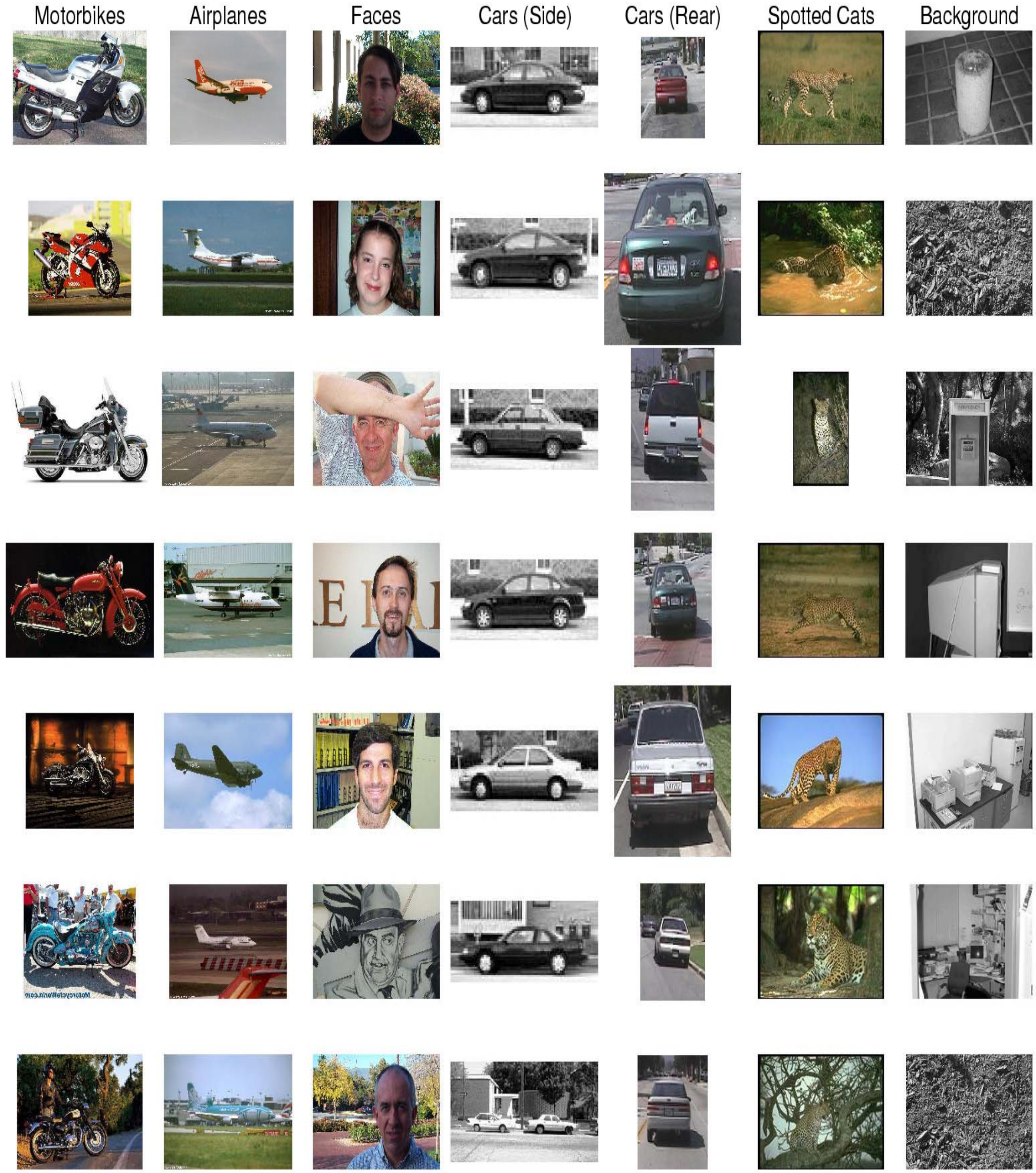
Slide credit: Fergus

Some object categories

Learn from examples

Difficulties:

- Size variation
- Background clutter
- Occlusion
- Intra-class variation



Main issues

- **Representation**
- Learning
- Recognition

Slide credit: Fergus

Representation

- Object as set of parts
  - Generative representation

- Model:
  - Relative locations between parts
  - Appearance of part

- Issues:
  - How to model location
  - How to represent appearance
  - Sparse or dense (pixels or regions)
  - How to handle occlusion/clutter

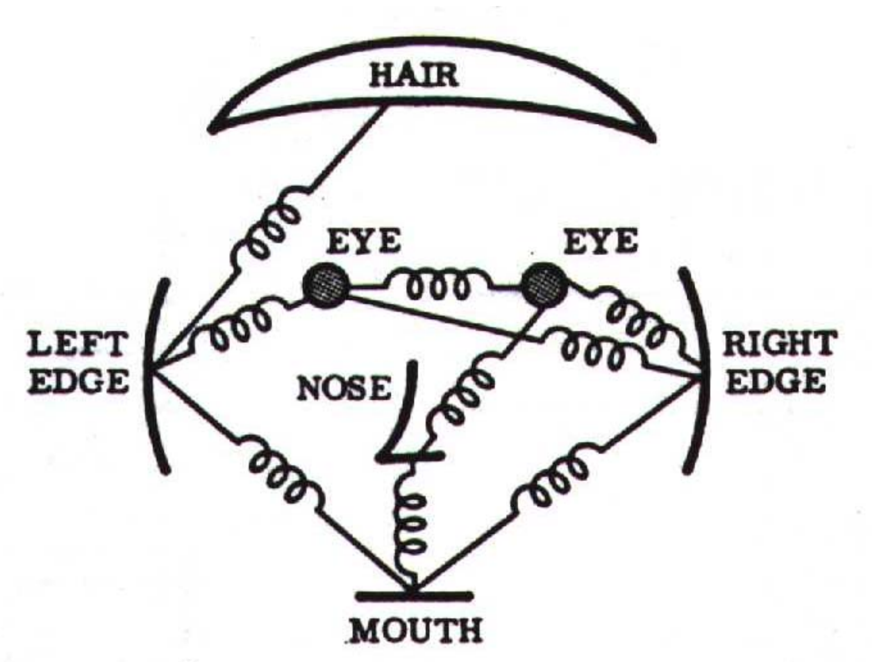
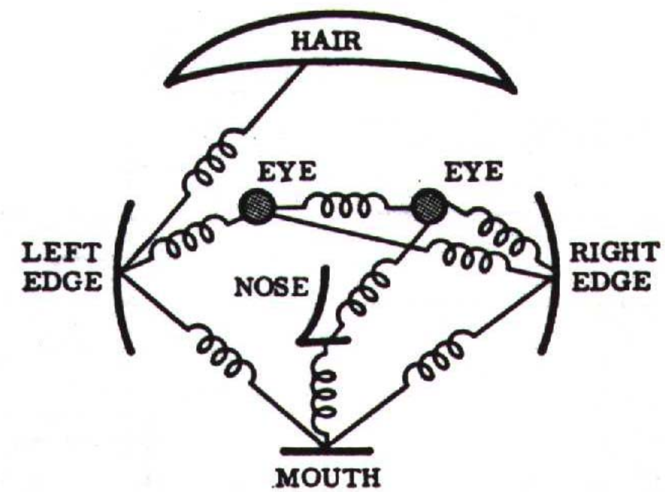


Figure from [Fischler & Elschlager 73]

Slide credit: Fergus

History of Parts and Structure approaches

- Fischler & Elschlager 1973



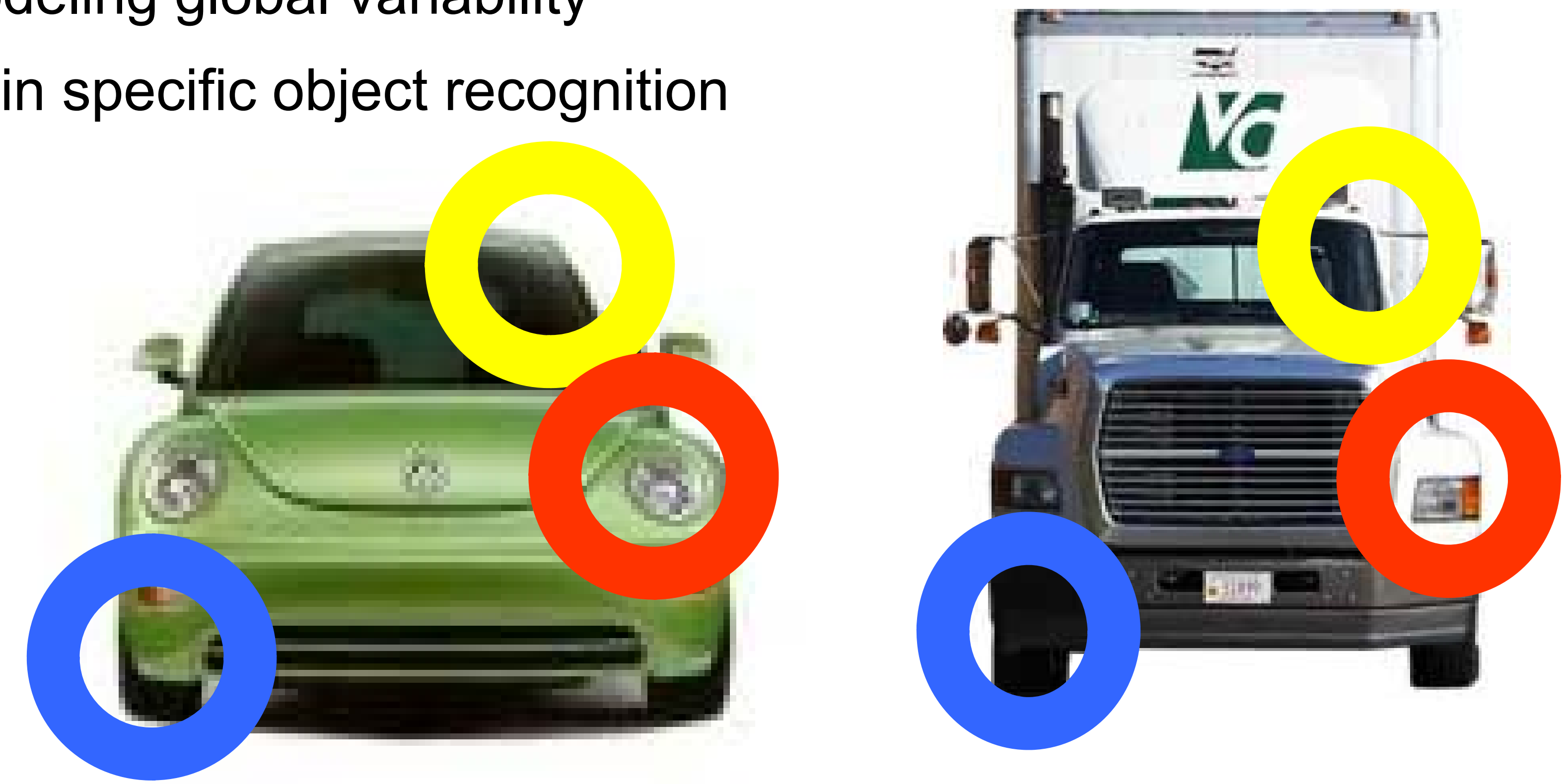
- Yuille ‘91
- Brunelli & Poggio ‘93
- Lades, v.d. Malsburg et al. ‘93
- Cootes, Lanitis, Taylor et al. ‘95
- Amit & Geman ‘95, ‘99
- Perona et al. ‘95, ‘96, ’98, ’00, ’03, ‘04, ‘05
- Felzenszwalb & Huttenlocher ’00, ’04
- Crandall & Huttenlocher ’05, ’06
- Leibe & Schiele ’03, ’04

- Many papers since 2000

Slide credit: Fergus

Sparse representation

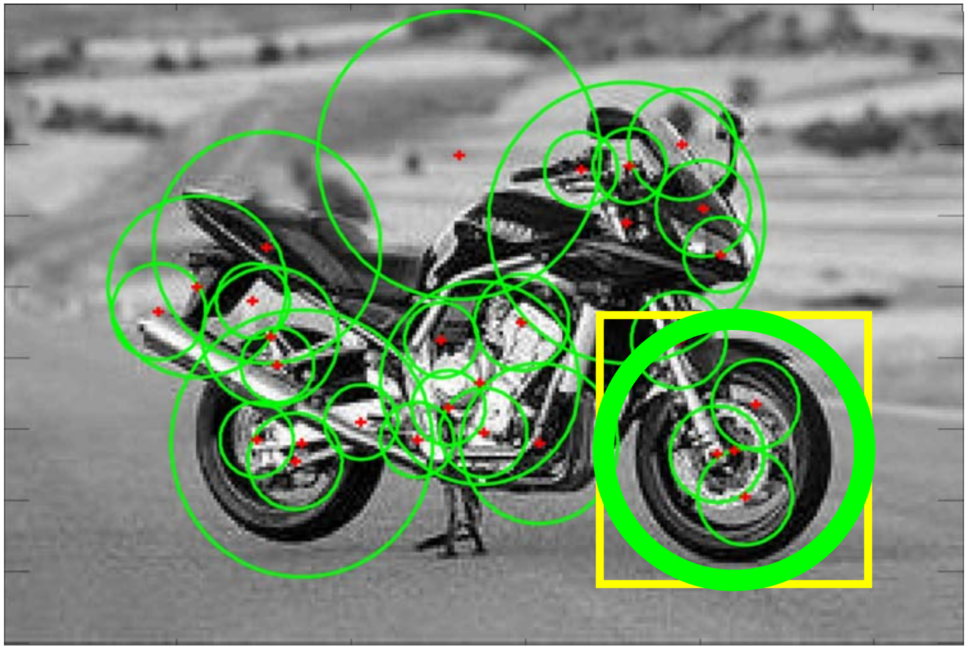
+ Computationally tractable (10^5 pixels → 10^1 -- 10^2 parts)
+ Generative representation of class
+ Avoid modeling global variability
+ Success in specific object recognition



- Throw away most image information
- Parts need to be distinctive to separate from other classes

Slide credit: Fergus

Detection & Representation of regions



- Find regions within image
- Use Kadir and Brady's salient region operator [IJCV '01]

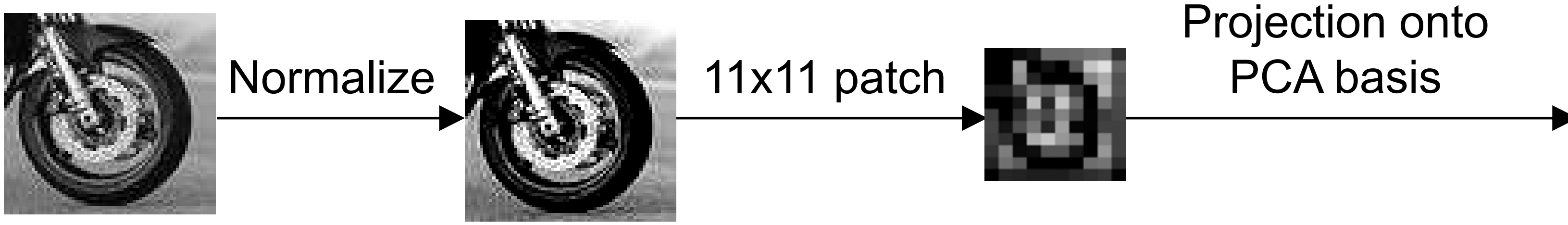
Location

(x,y) coords. of region center

Scale

Diameter of region (pixels)

Appearance



$$\begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_{15} \end{pmatrix}$$

Gives representation of appearance in low-dimensional vector space

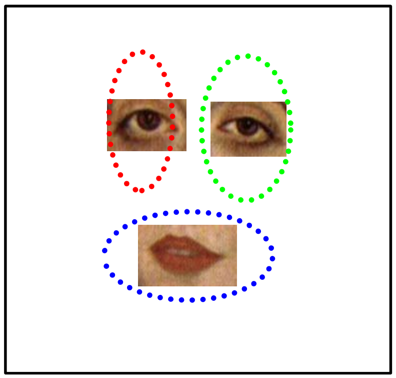
Slide credit: Fergus

Generative probabilistic model

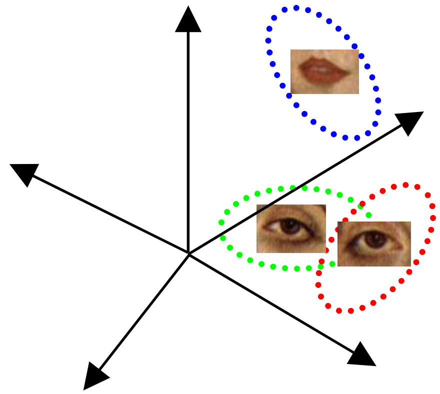
Foreground model

based on Burl, Weber et al. [ECCV '98, '00]

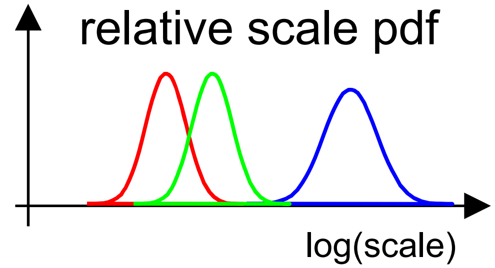
Gaussian shape pdf



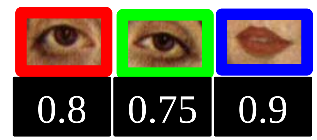
Gaussian part appearance pdf



Gaussian relative scale pdf

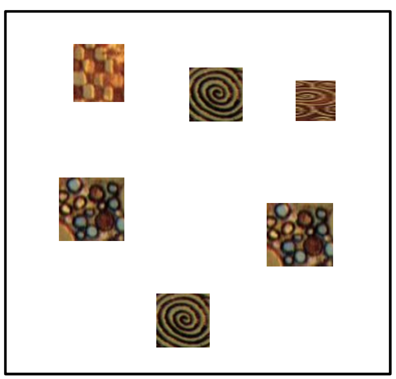


Prob. of detection

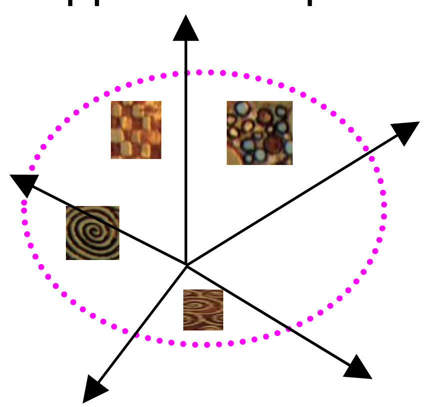


Clutter model

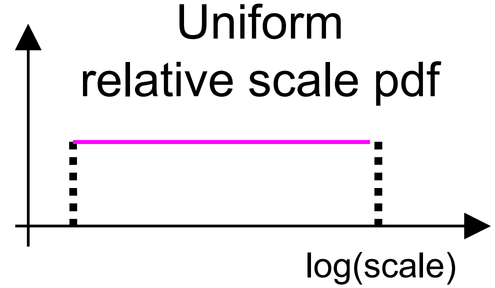
Uniform shape pdf



Gaussian background appearance pdf



Uniform relative scale pdf



Poission pdf on # detections

Slide credit: Fergus

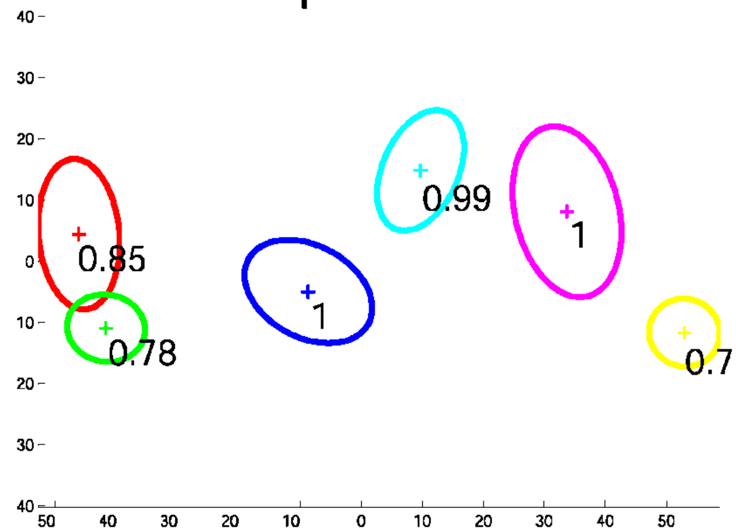
Recognition

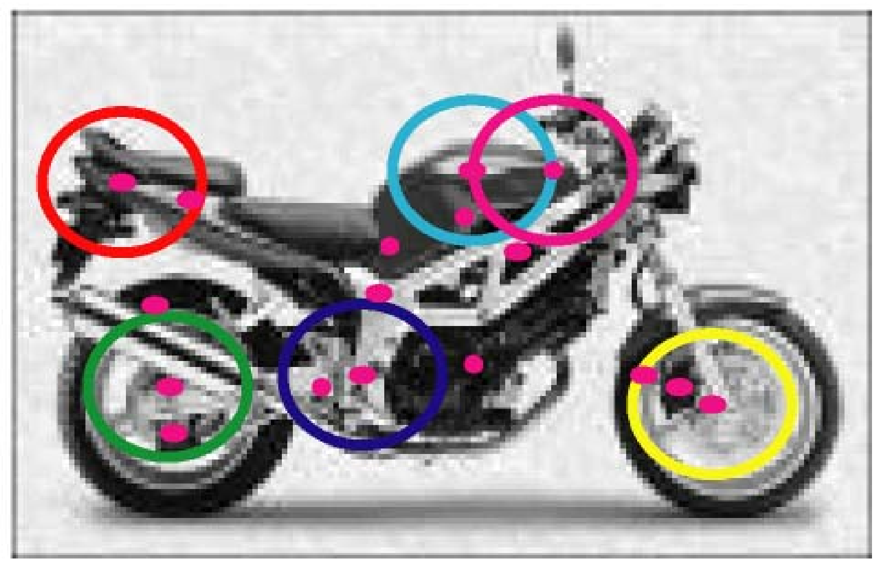
Slide credit: Fergus

Motorbikes

Samples from appearance model

Shape model

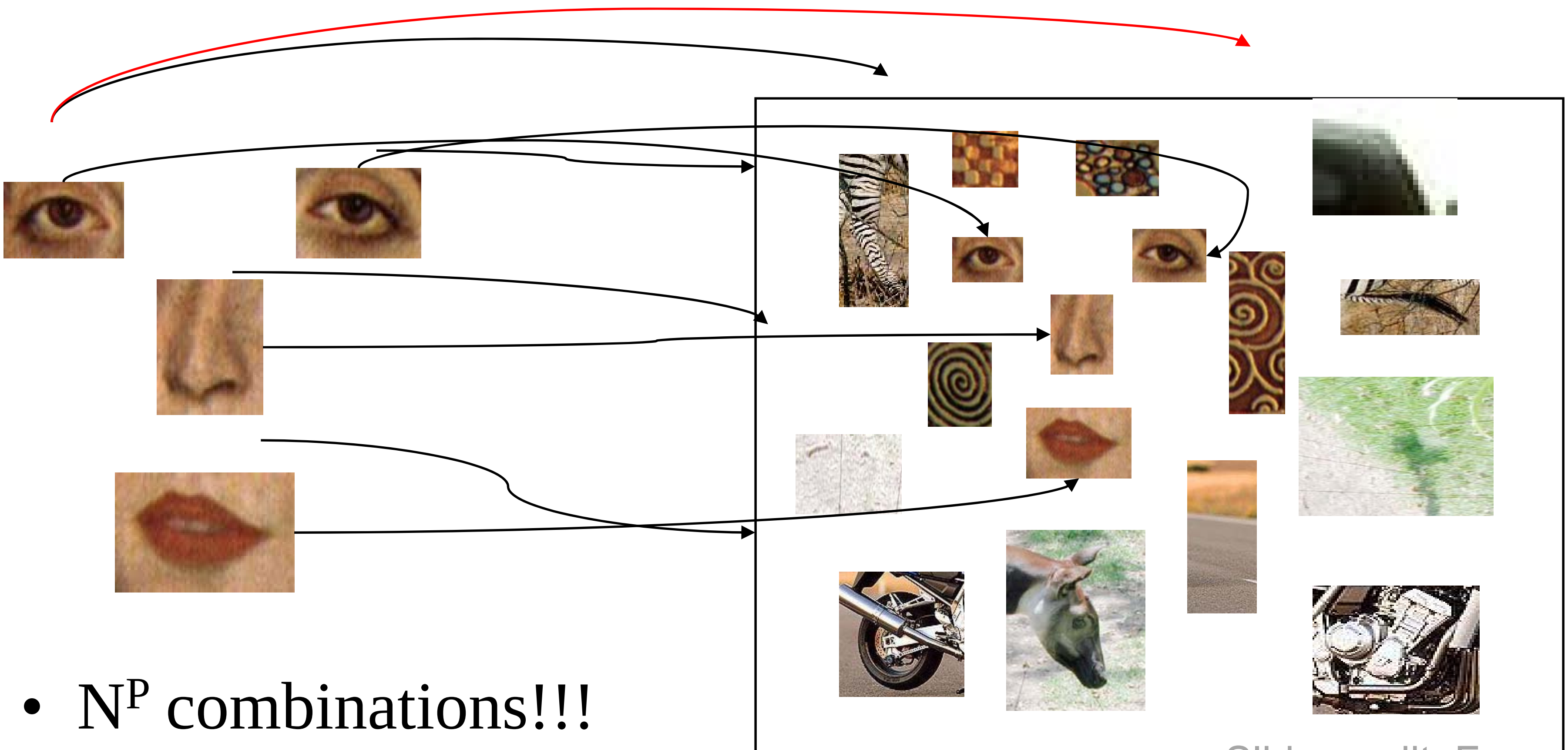




Slide credit: Fergus

The correspondence problem

- Model with P parts
- Image with N possible assignments for each part
- Consider mapping to be 1-1



- N^P combinations!!!

Slide credit: Fergus

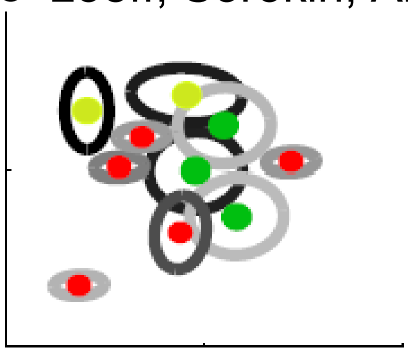
The correspondence problem

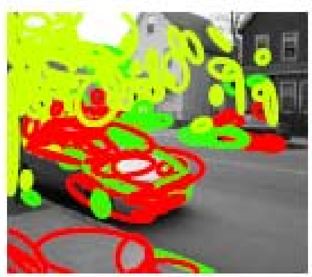
- 1 – 1 mapping
  - Each part assigned to unique feature

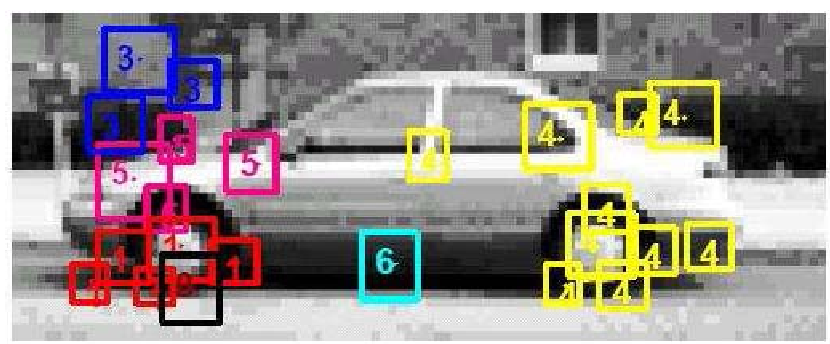
As opposed to:

- 1 – Many
  - Bag of words approaches
  - Sudderth, Torralba, Freeman '05
  - Loeff, Sorokin, Arora and Forsyth '05

- Many – 1
  - Quattoni, Collins and Darrell, 04





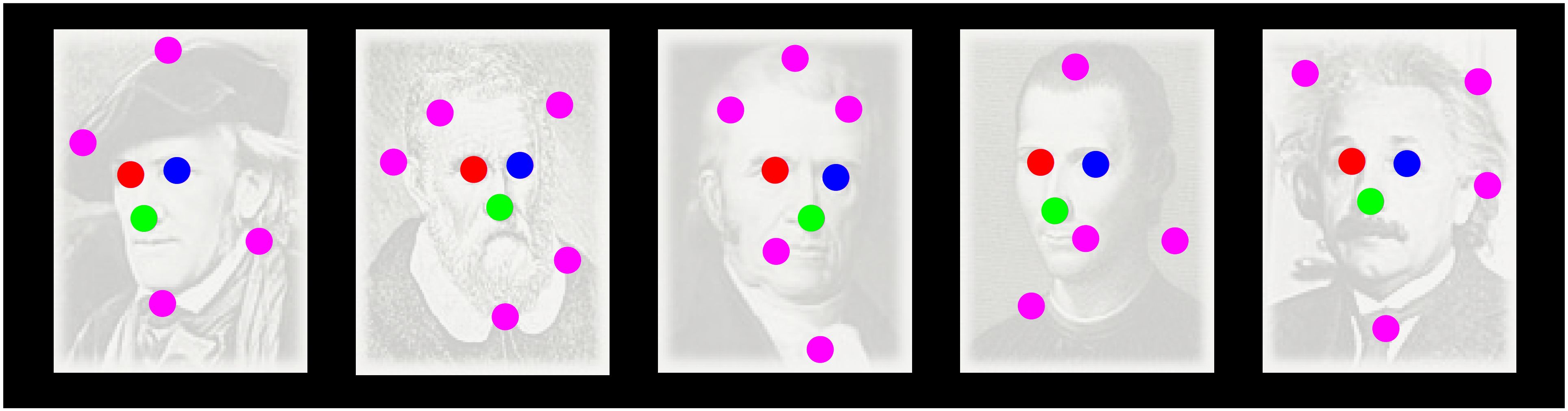


Learning

Slide credit: Fergus

Learning

- Task: Estimation of model parameters
- Chicken and Egg type problem, since we initially know neither:
  - Model parameters
  - Assignment of regions to foreground / background
- Let the assignments be a hidden variable and use EM algorithm to learn them and the model parameters



Slide credit: Fergus

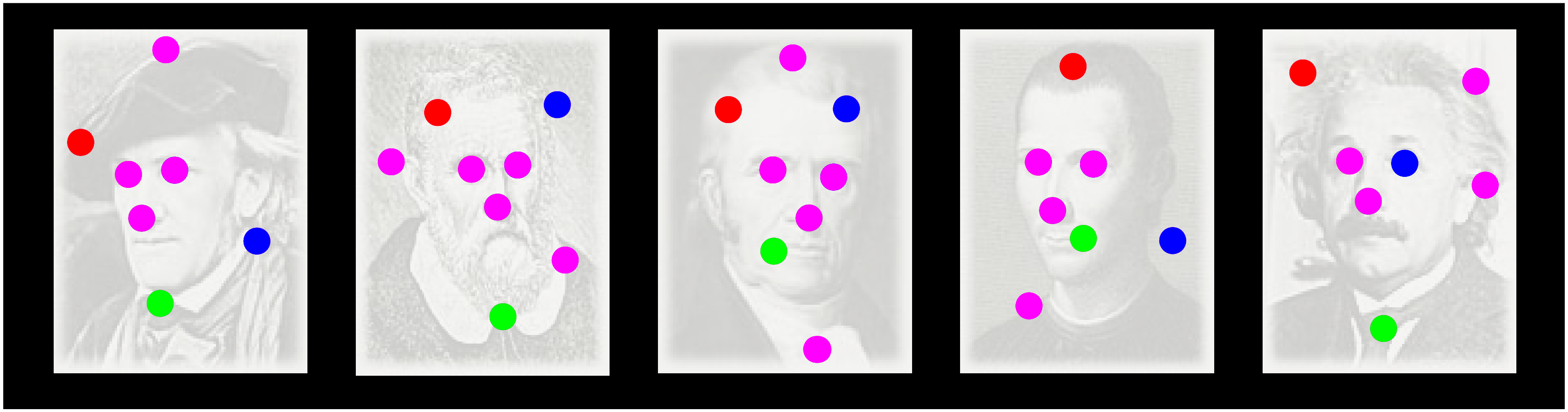
Learning procedure

- Find regions & their location, scale & appearance
- Initialize model parameters
- Use EM and iterate to convergence:

  E-step: Compute assignments for which regions are foreground / background

  M-step: Update model parameters

- Trying to maximize likelihood – consistency in shape & appearance



Slide credit: Fergus

Experiments

Slide credit: Fergus

Experimental procedure

Two series of experiments:

- Fixed-scale model - Objects the same size (manual normalization)
- Scale-invariant model - Objects between 100 and 550 pixels in width

Datasets

Training
- 50% images
- No identifcation of object within image

Motorbikes

Airplanes

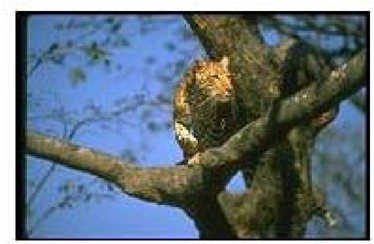
Frontal Faces

Testing
- 50% images
- Simple object present/absent test

Cars (Side)

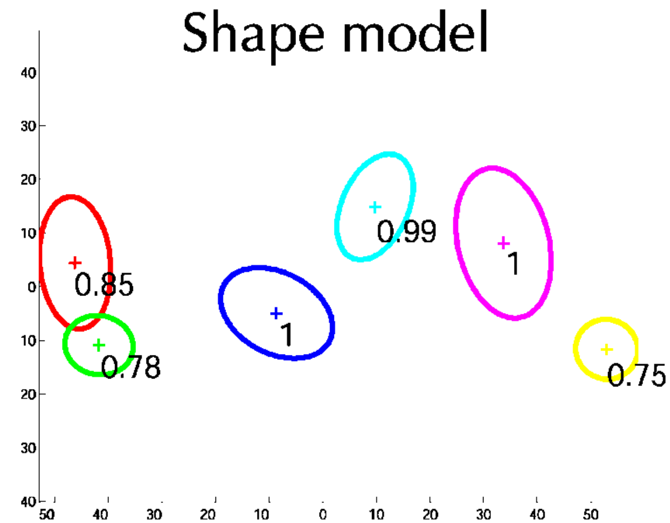
Cars (Rear)

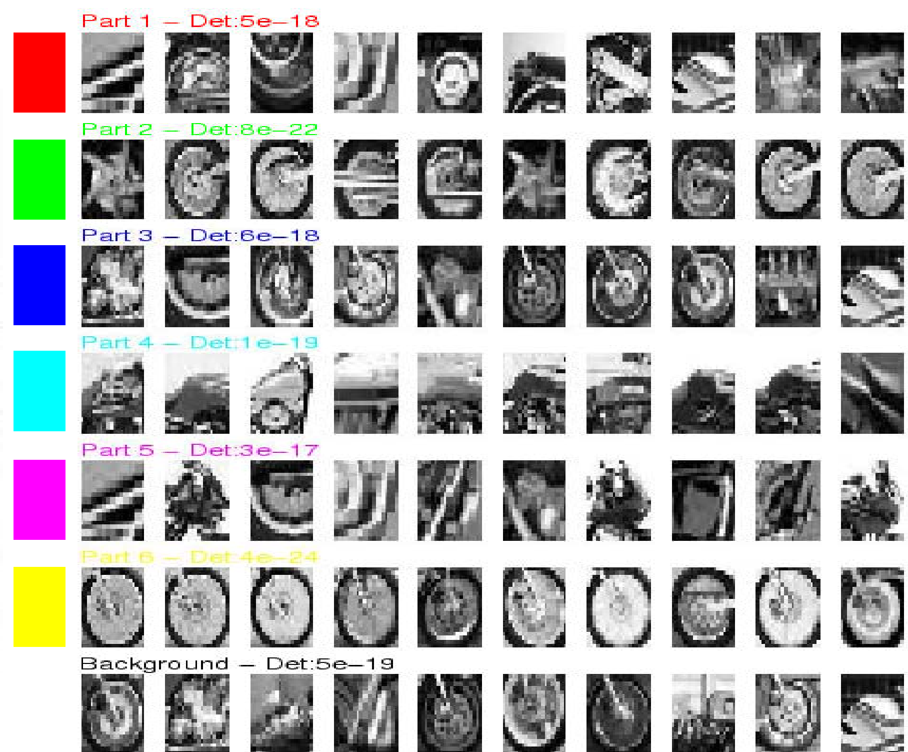
Spotted cats



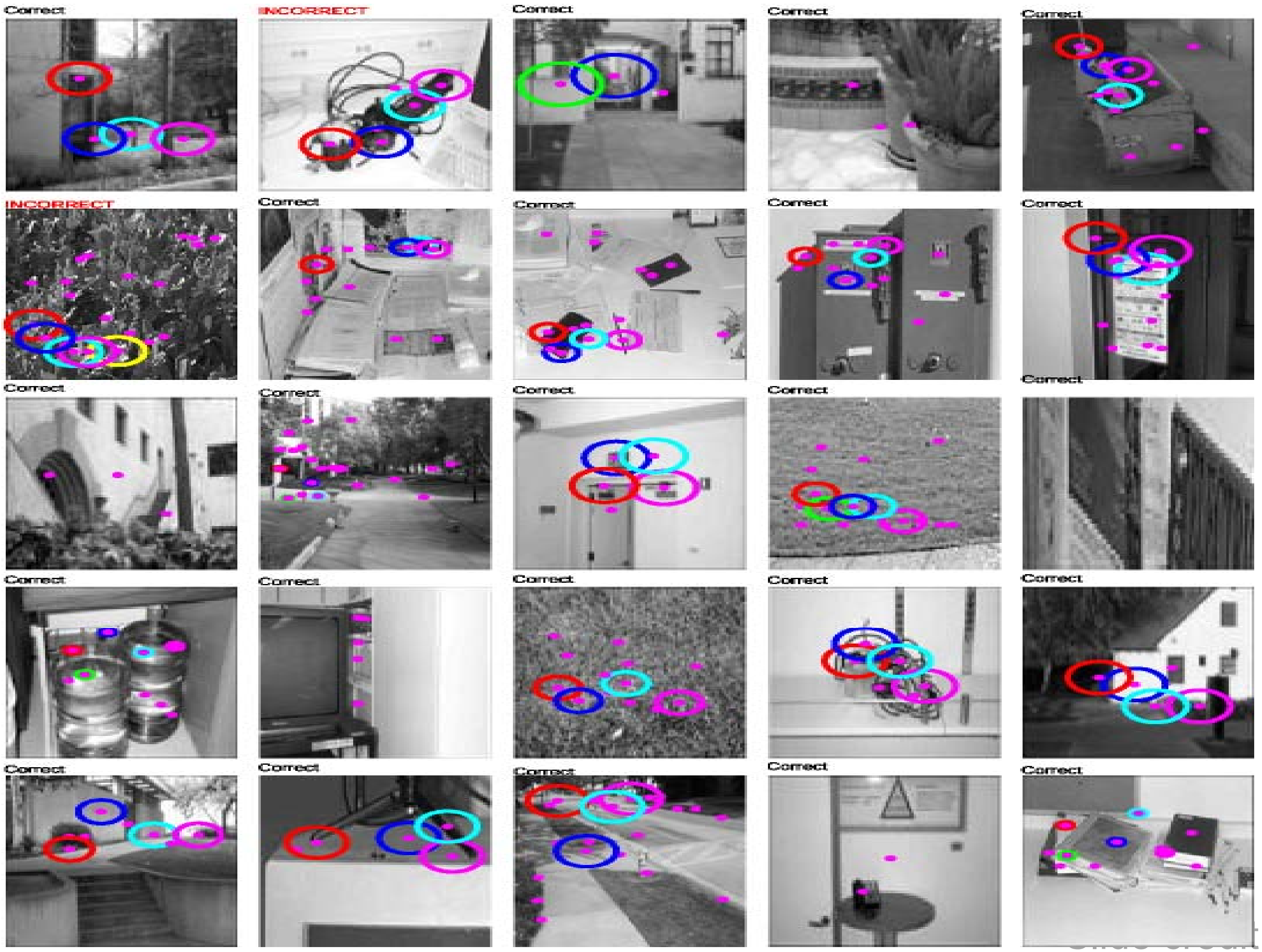
Slide credit: Fergus

Motorbikes


Shape model
0.85
0.78
1
0.99
1
0.75

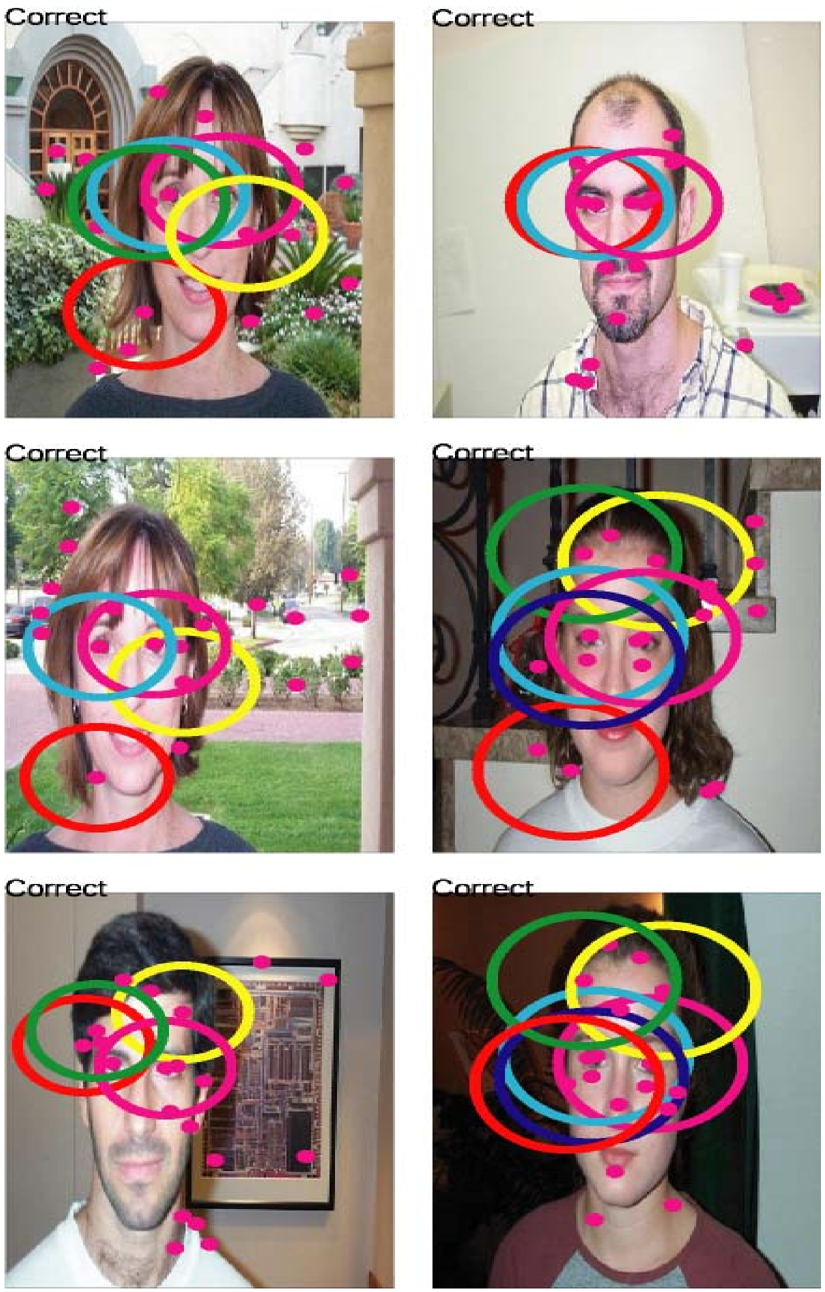

Part 1 – Det:5e–18
Part 2 – Det:8e–22
Part 3 – Det:6e–18
Part 4 – Det:1e–19
Part 5 – Det:3e–17
Part 6 – Det:4e–24
Background – Det:5e–19

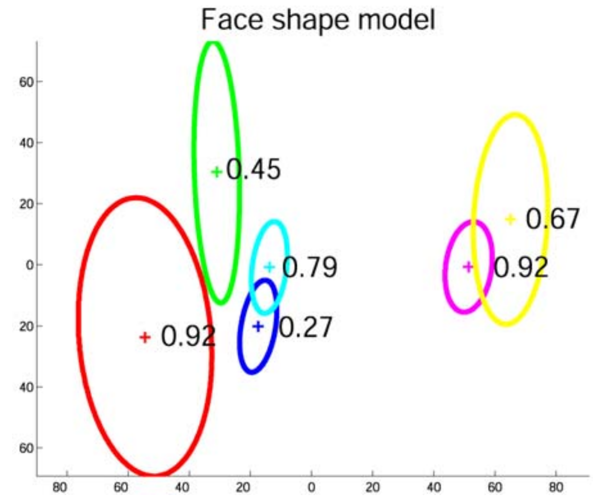
Background images evaluated with motorbike model



Slide credit: Fergus

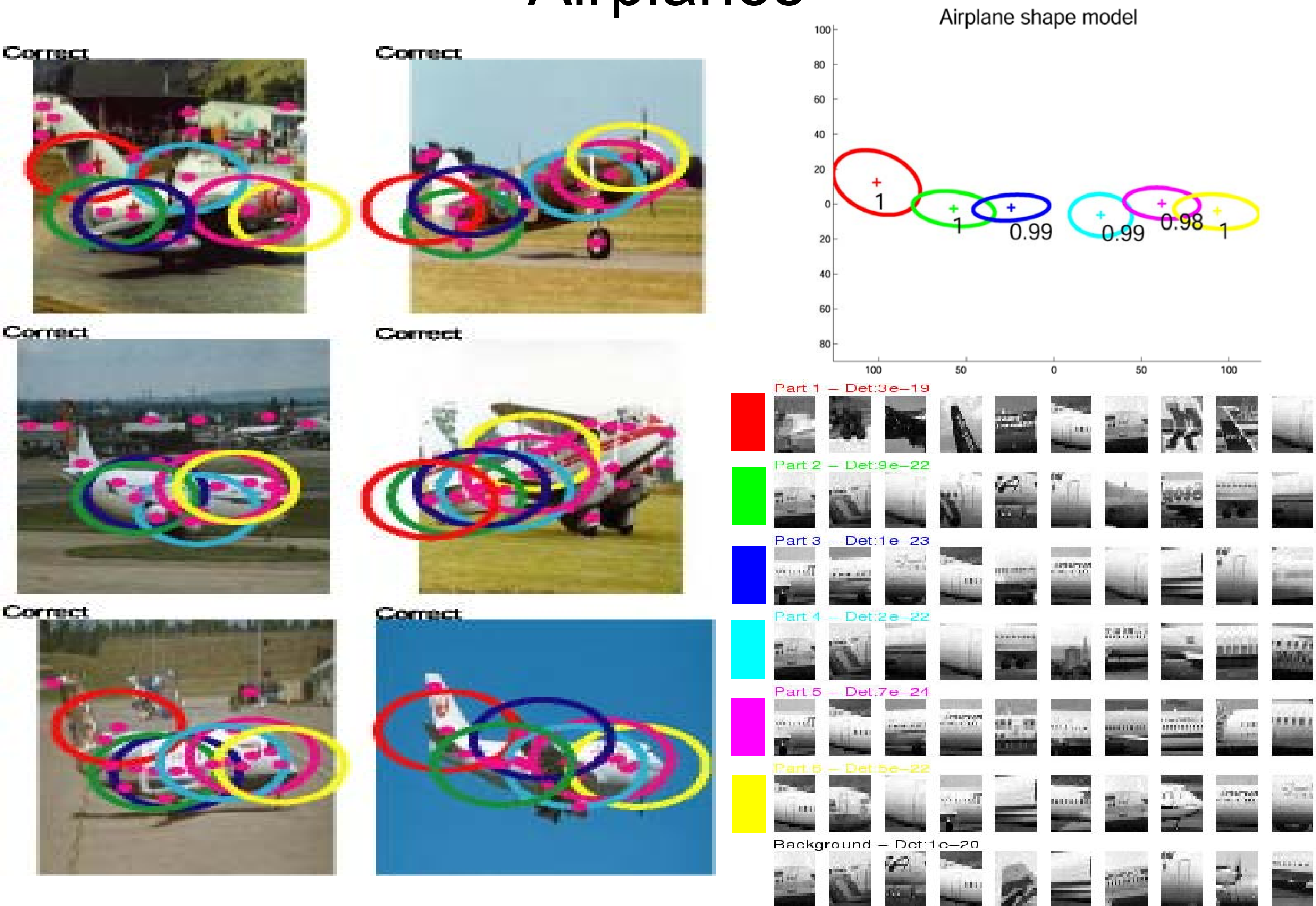
Frontal faces





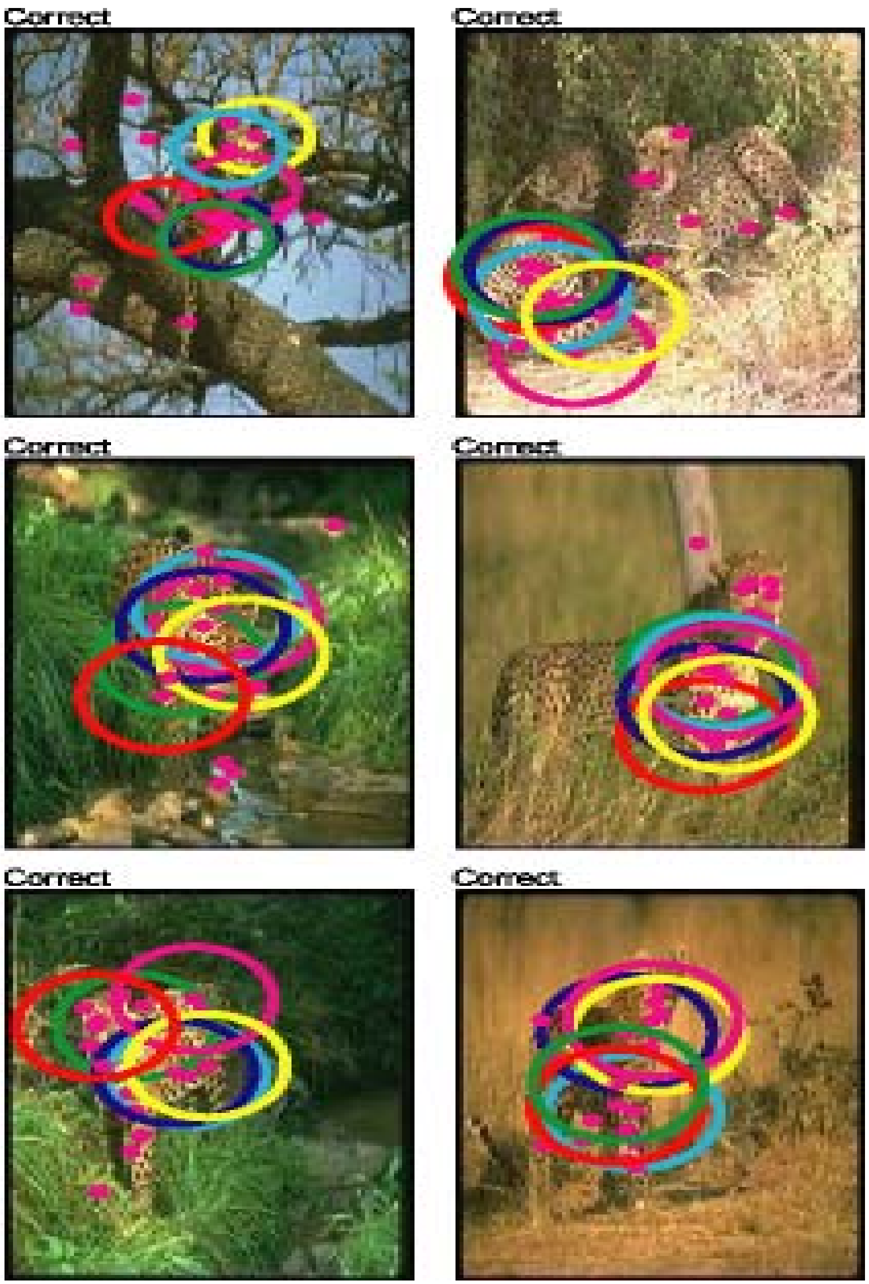
Slide credit: Fergus

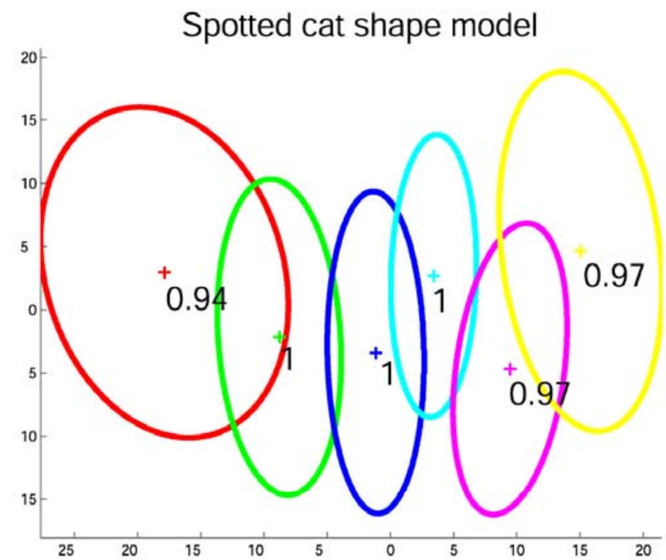
Airplanes

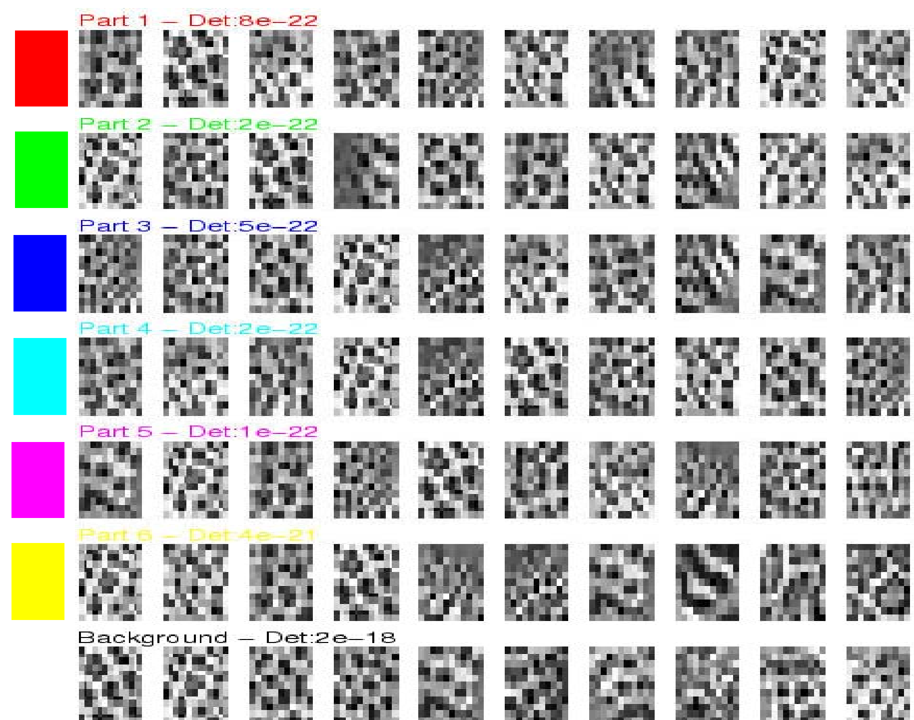


Correct
Correct
Correct
Correct
Correct
Correct
Airplane shape model
1
1
0.99
0.99
0.98
1
Part 1 – Det:3e–19
Part 2 – Det:9e–22
Part 3 – Det:1e–23
Part 4 – Det:2e–22
Part 5 – Det:7e–24
Part 6 – Det:5e–22
Background – Det:1e–20

Spotted cats







Slide credit: Fergus

Summary of results

Dataset	Fixed scale experiment	Scale invariant experiment
Motorbikes	7.5	6.7
Faces	4.6	4.6
Airplanes	9.8	7.0
Cars (Rear)	15.2	9.7
Spotted cats	10.0	10.0

% equal error rate

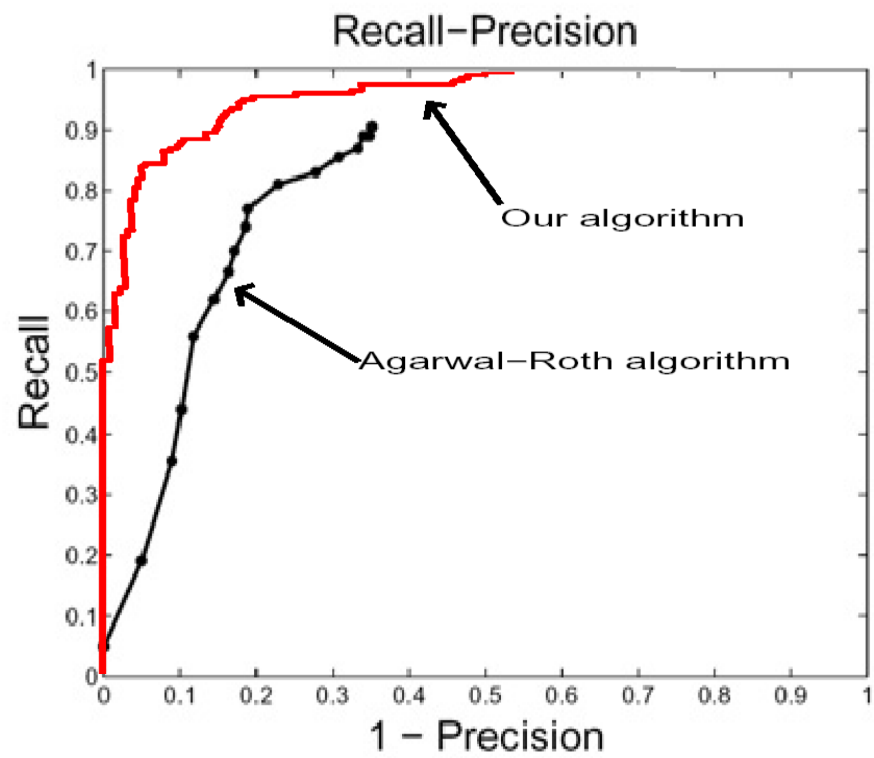
Note: Within each series, same settings used for all datasets

Slide credit: Fergus

Comparison to other methods

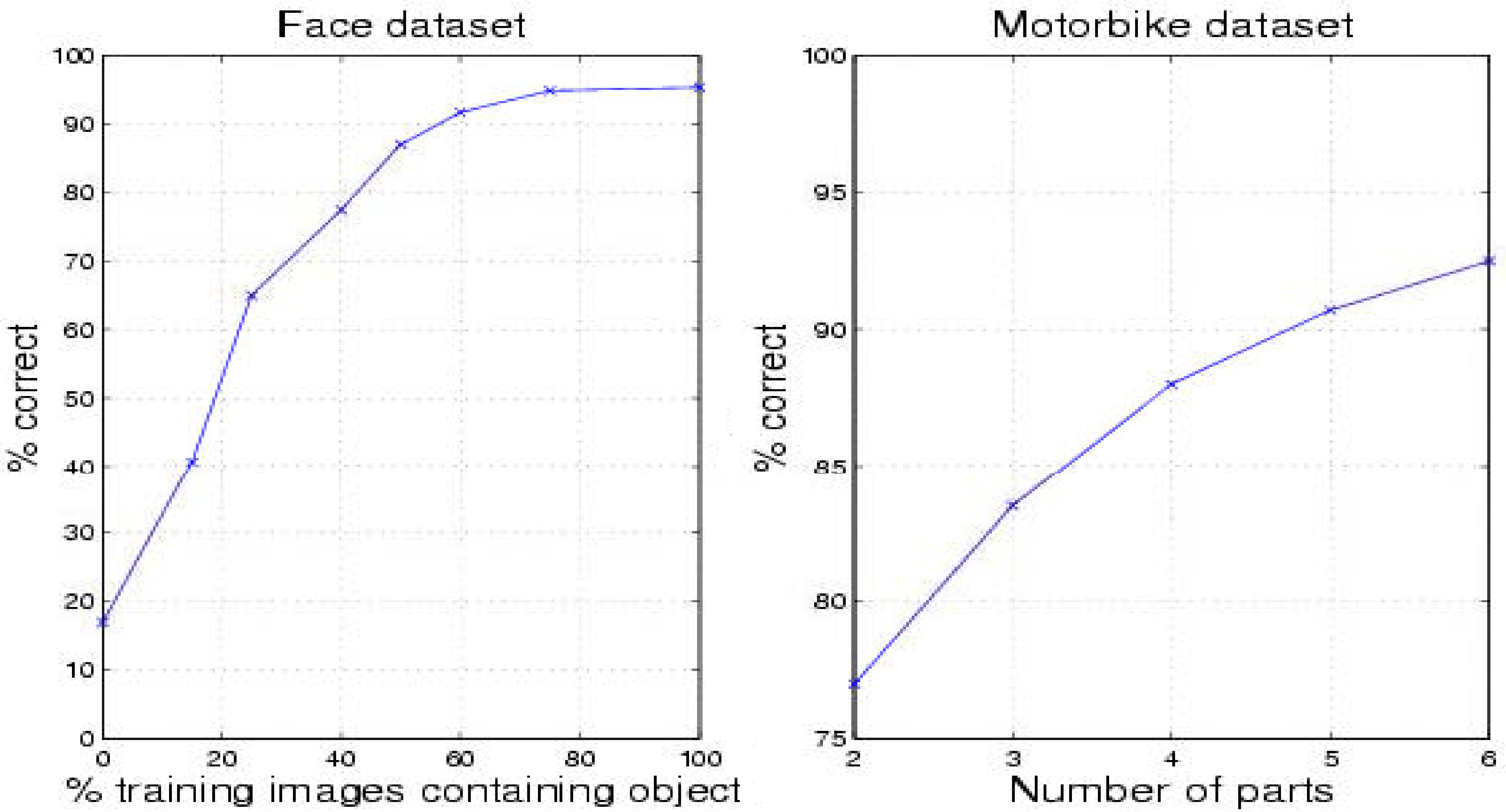
Dataset	Ours	Others	
Motorbikes	7.5	16.0	Weber et al. [ECCV '00]
Faces	4.6	6.0	Weber
Airplanes	9.8	32.0	Weber
Cars (Side)	11.5	21.0	Agarwal Roth [ECCV '02]

% equal error rate



Slide credit: Fergus

Robustness of Algorithm



Slide credit: Fergus

Summary

- Comprehensive probabilistic model for object classes
- Learn appearance, shape, relative scale, occlusion etc. simultaneously in scale and translation invariant manner
- Same algorithm gives <= 10% error across 5 diverse datasets with identical settings

Limitations $\rightarrow$ future work

- Very reliant on region detector
  Different part types (e.g. edgel curves)
- Only learns a single viewpoint
  Use mixture models
- Need lots of images to learn
  Bayesian learning - fewer images [ICCV '03 (Fei Fei, Fergus, Perona)]
- Need more through testing
  Looking towards testing 100's of datasets

Datasets available from:
http://www.robots.ox.ac.uk/~vgg/data

Slide credit: Fergus

Today

Sudderth guest lecture:

- Constellation Models (Fergus)
- **Unsupervised Object Discovery with pLSA (Sivic)**
- Scene Models (Li)
- Transformed Models (Sudderth)

Daphna B. student presentation:

- pLSA models of activity (Neibles)

Moreels guest lecture:

- A probabilistic formulation of voting / SIFT (Moreels)

Discovering Objects and Their Location in Images

J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, W. T. Freeman.
Presented at the International Conference on Computer Vision, 2005.

Slide credit: Sivic

How much supervision do you need to learn models of objects?

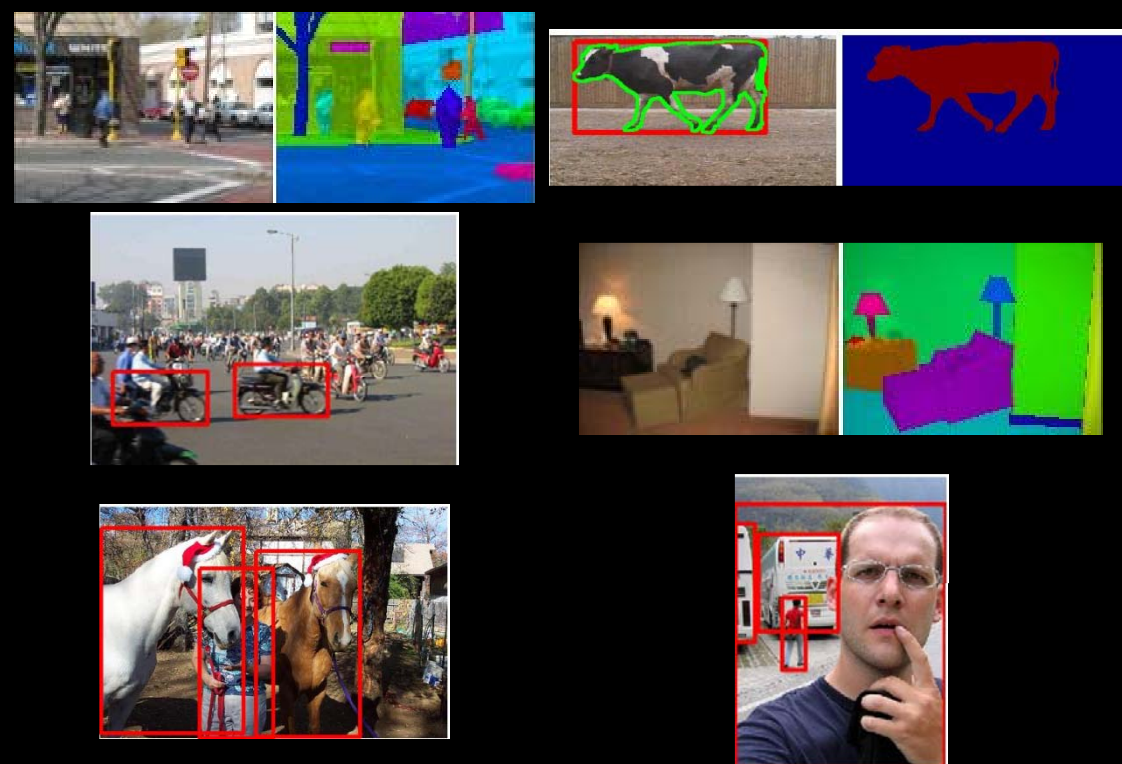
Slide credit: Sivic

Object label + segmentation

MIT+CMU frontal faces

Viola & Jones ’01
Rowley et al. ’98

LabelMe, PASCAL, TU Darmstadt,
MIT scenes and objects



Agarwal & Roth ’02, Leibe &
Schiele ’03, Torralba et al. ’05

Slide credit: Sivic

Object appears somewhere in the image

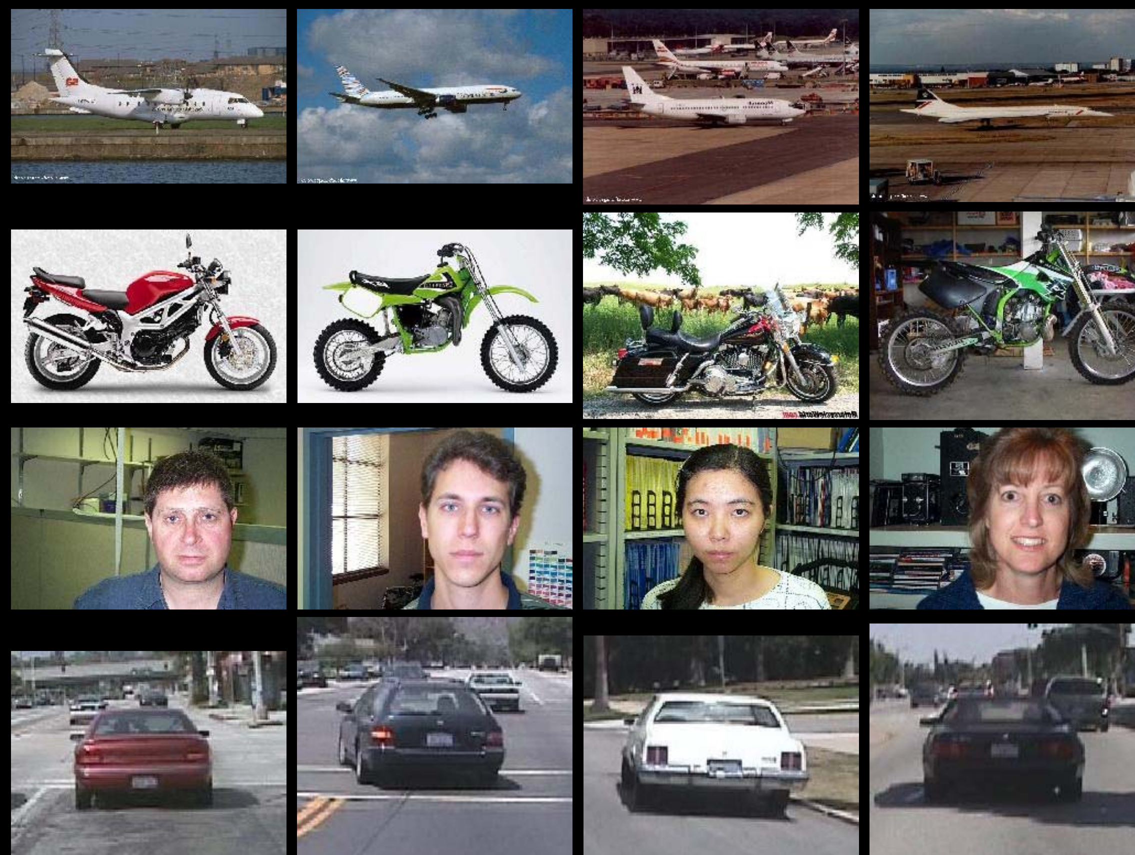
Caltech 101, PASCAL, MSRC

airplane

motorbike

face

car



Fergus et al. ’03, Csurka et al. ’04, Dorko & Schmid ’05

Slide credit: Sivic

Image + text caption

Corel, Flickr, Names+faces, ESP game

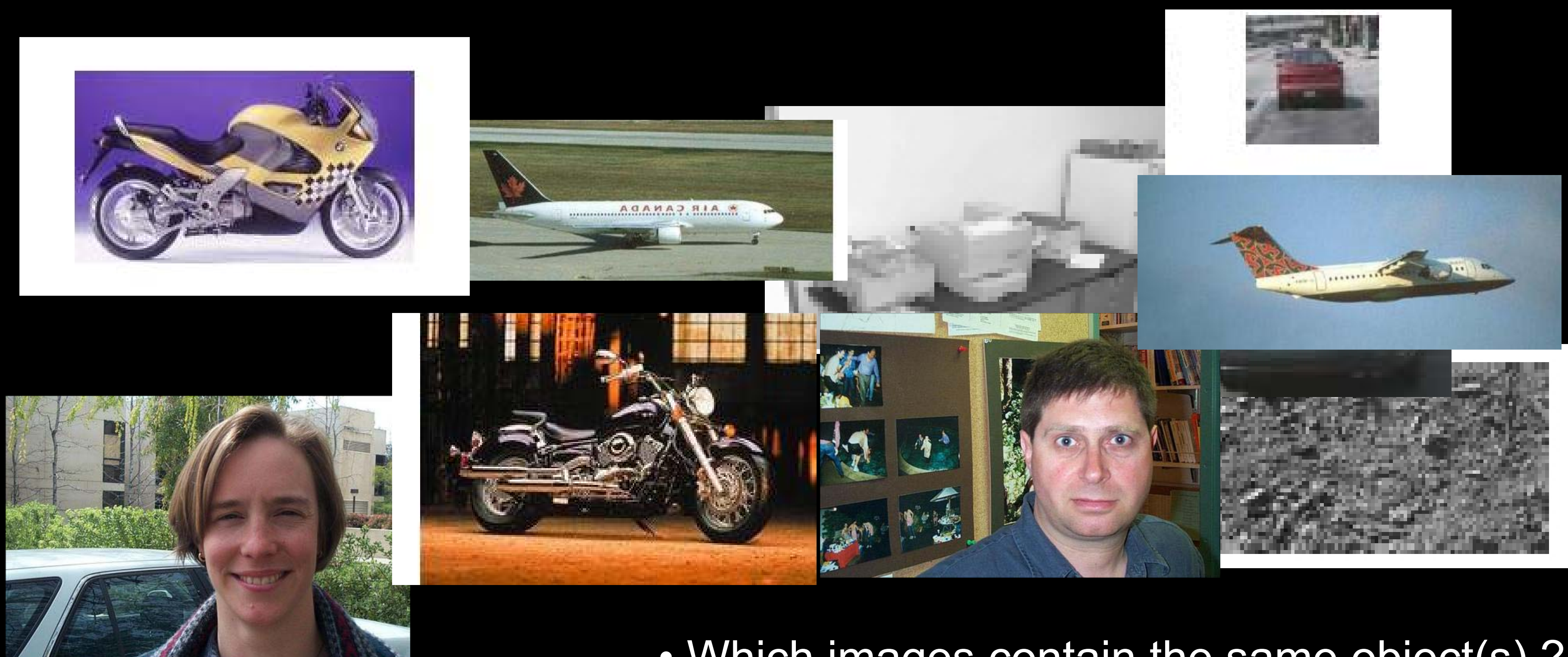
British director **Sam Mendes** and his partner actress **Kate Winslet** arrive at the London premiere of 'The Road to Perdition', September 18, 2002. The films stars **Tom Hanks** as a Chicago hit man who has a separate family life and co-stars **Paul Newman** and Jude Law. REUTERS/Dan Chung

Barnard et al. '03, Berg et al. '04

Slide credit: Sivic

Images only

Given a collection of unlabeled images, discover visual object categories and their segmentation



- Which images contain the same object(s) ?
- Where is the object in the image?

Slide credit: Sivic

Analogy: Discovering topics in text collections

Text document

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Discovered topics

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

Blei, et al. 2003

Slide credit: Sivic

Visual analogy

document	-	image
word	-	visual word
topics	-	objects

Slide credit: Sivic

System overview

Input image

Compute visual words

Discover visual topics

Slide credit: Sivic

System overview

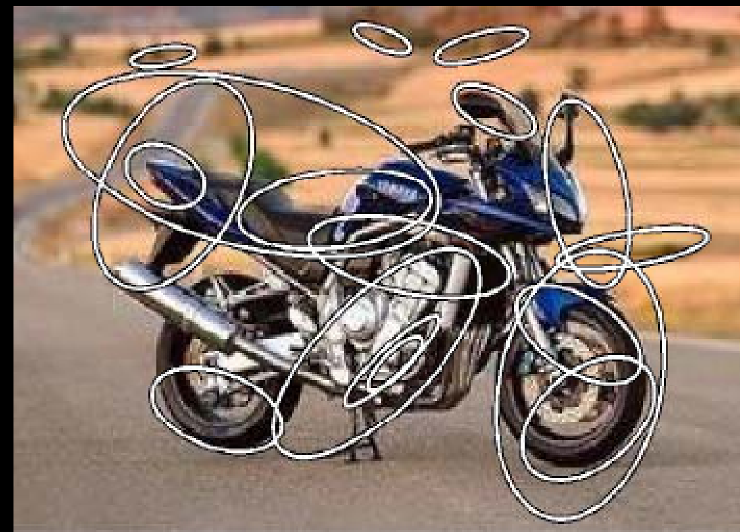
Input image

Compute visual words

Discover visual topics

Slide credit: Sivic

Finding and describing interest regions



Detect affine covariant regions:
- Multi-scale affine Harris [Mikolajczyk & Schmid ’02, Schaffalitzky & Zisserman’02]
- Maximally stable extremal regions [Matas et al. ’02]

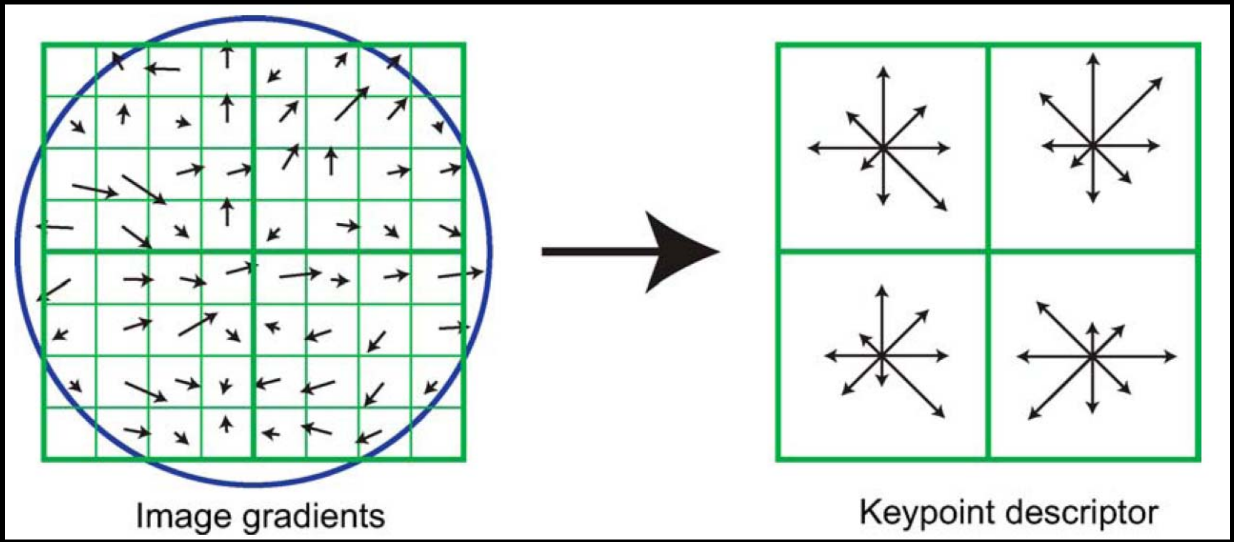
Detects corner regions and small blobs

Describe regions with SIFT descriptor [Lowe 1999]

Slide credit: Sivic

SIFT descriptor

Lowe 1999



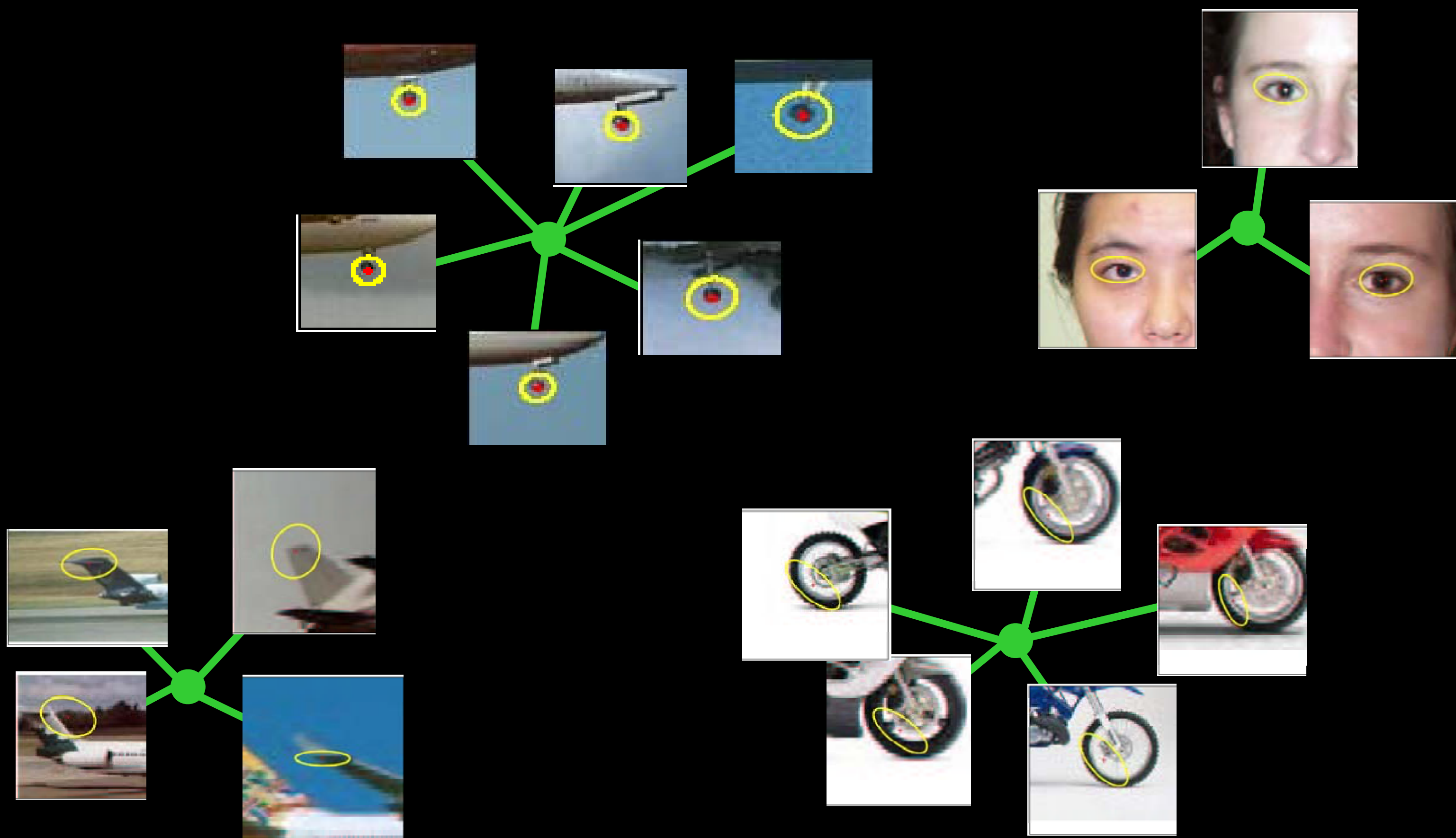
8 orientations
x 16 bins
128 dimensions

- Thresholded image gradients are sampled over 16x16 array of locations in scale space
- Create an array of oriented histograms

Slide credit: Sivic

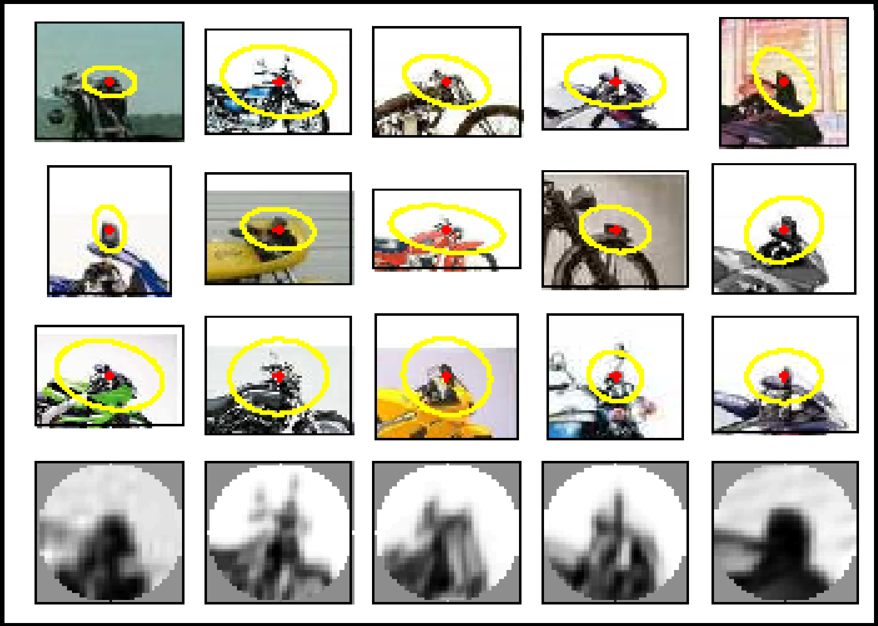
Form dictionary

Build visual vocabulary by k-means clustering
SIFT descriptors (K~2,000)

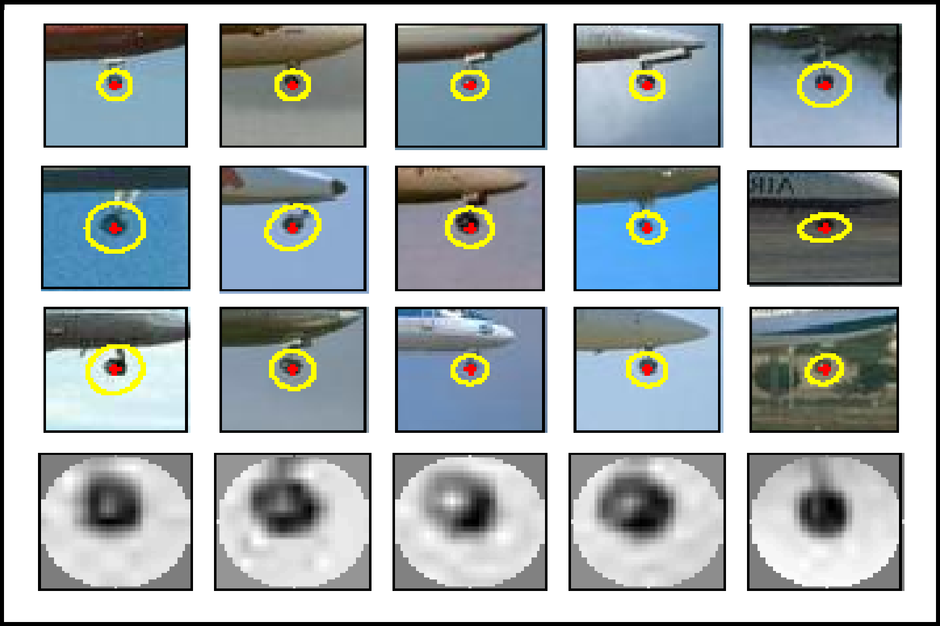


Slide credit: Sivic

Example regions assigned to the same dictionary cluster



Cluster 1



Cluster 2

Slide credit: Sivic

Polysemy

In English, “bank” refers to:

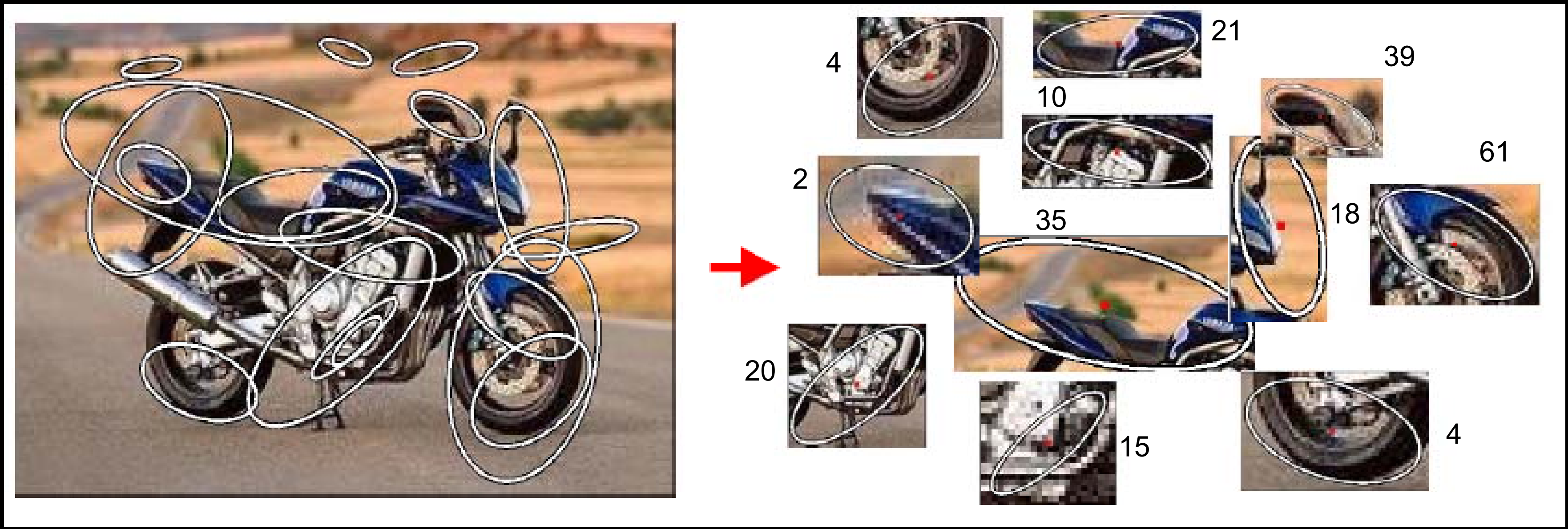
1. a institution that handle money
2. the side of a river

Regions that map to the same visual word:

Slide credit: Sivic

Representing an image with visual words

Sivic & Zisserman '03



Interest regions

Visual words

Slide credit: Sivic

System overview

Input image

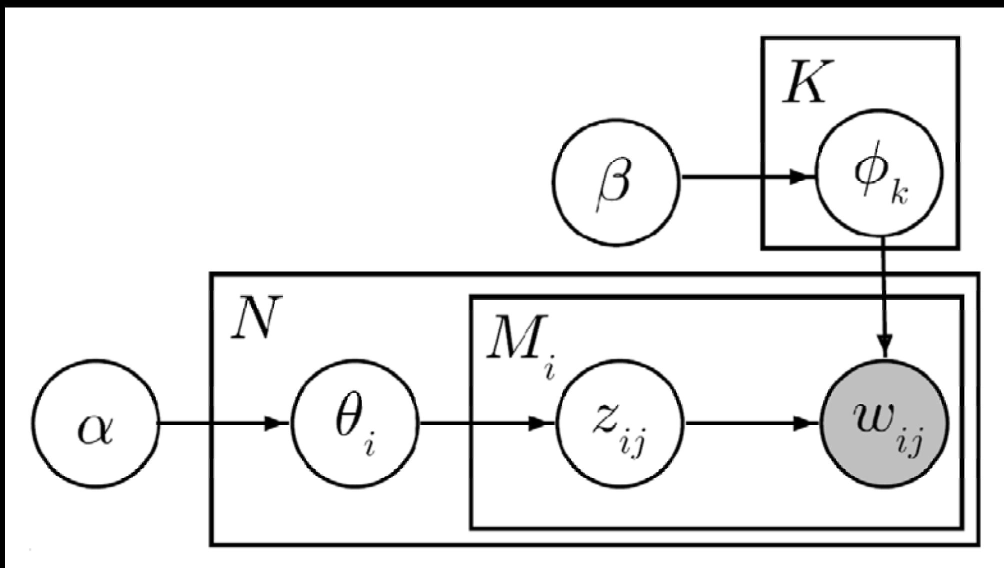
Compute visual words

Discover visual topics

Slide credit: Sivic

Latent Dirichlet Allocation (LDA)

Blei, et al. 2003



w_{ij} - words

z_{ij} - topic assignments

μ$_i$ - topic mixing weights

Á$_k$ - word mixing weights

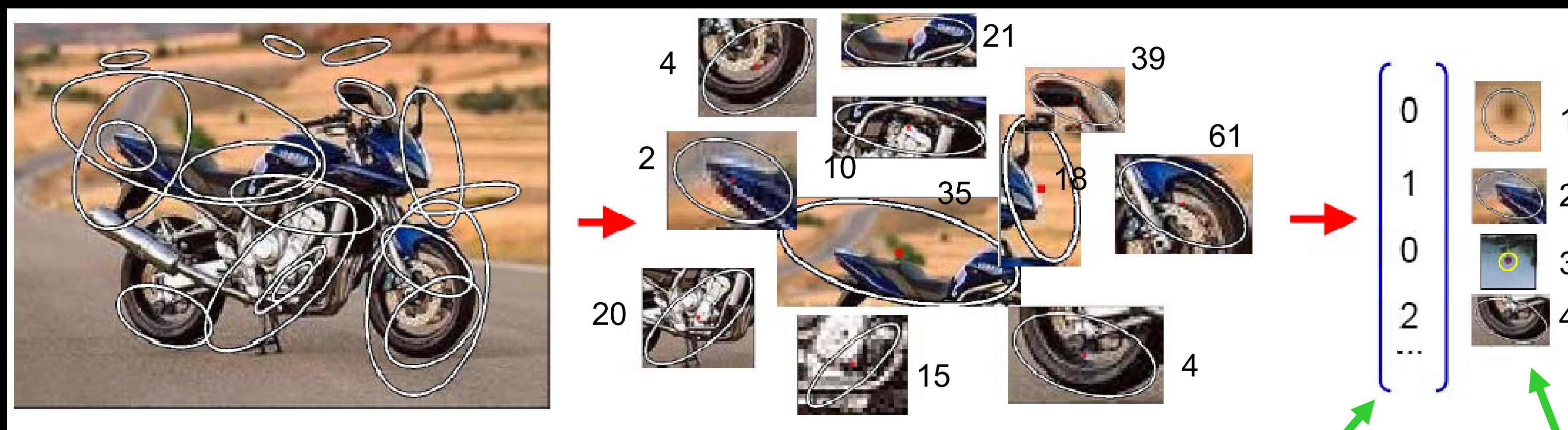
$$\begin{aligned} z_{ij}|\theta_i &\sim \theta_i \\ w_{ij}|z_{ij}=k,\phi &\sim \phi_k \end{aligned}$$

$$\begin{aligned} \theta_i|\alpha &\sim Dirichlet(\alpha) \\ \phi_k|\beta &\sim Dirichlet(\beta) \end{aligned}$$

Slide credit: Sivic

Bag of words

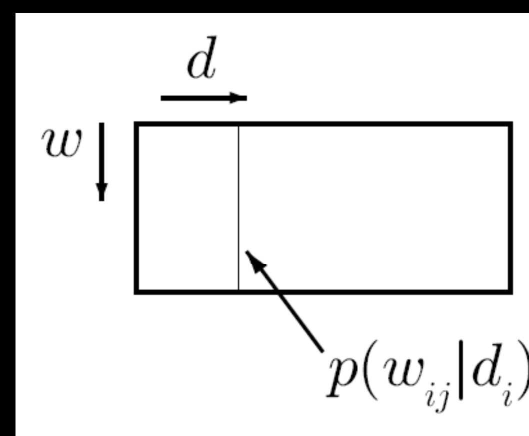
- LDA model assumes exchangeability
- Order of words does not matter



Interest regions Visual words Histogram Dictionary

Stack visual word histograms as columns in matrix

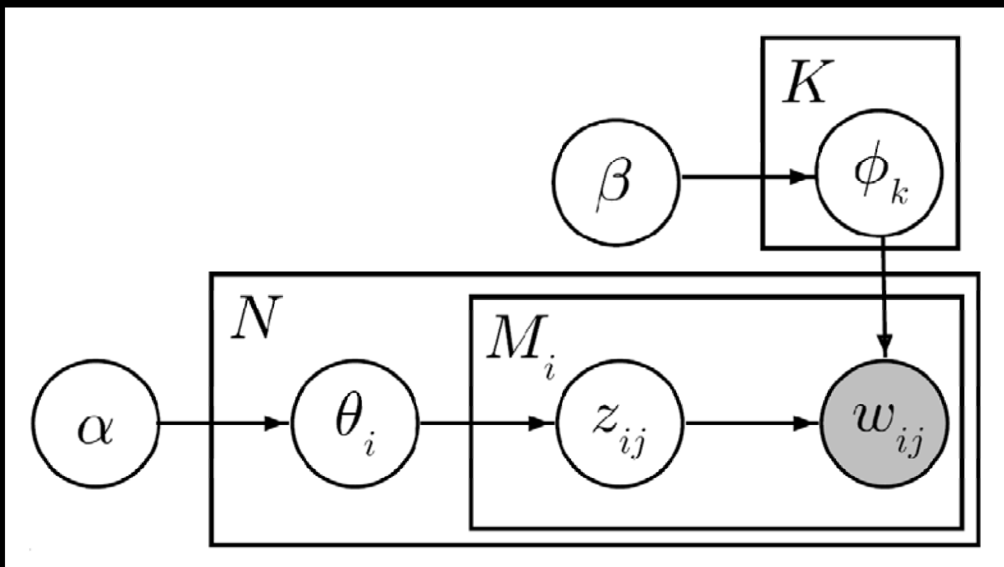
Throw away spatial information!



Slide credit: Sivic

Latent Dirichlet Allocation (LDA)

Blei, et al. 2003



w_{ij} - words

z_{ij} - topic assignments

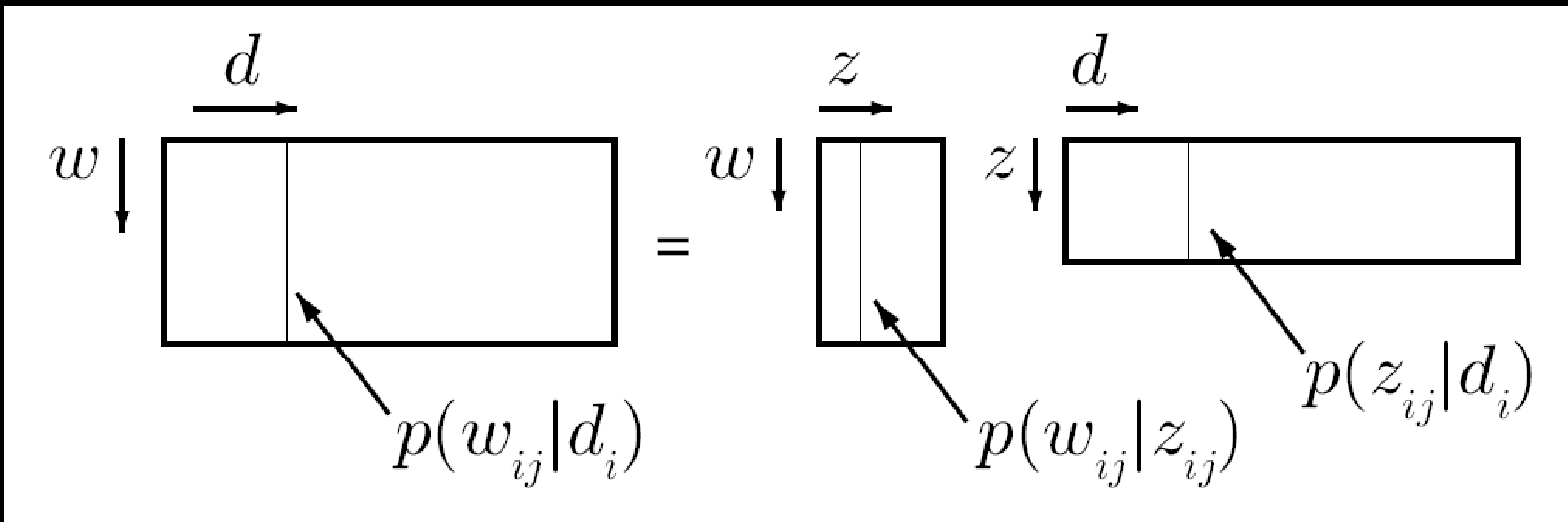
$µ_i$ - topic mixing weights

$Á_k$ - word mixing weights

$$p(w_{ij}) \propto \sum_{k=1}^{K} p(w_{ij}|z_{ij} = k, \phi_k)\ p(z_{ij} = k|\theta_i)$$

Slide credit: Sivic

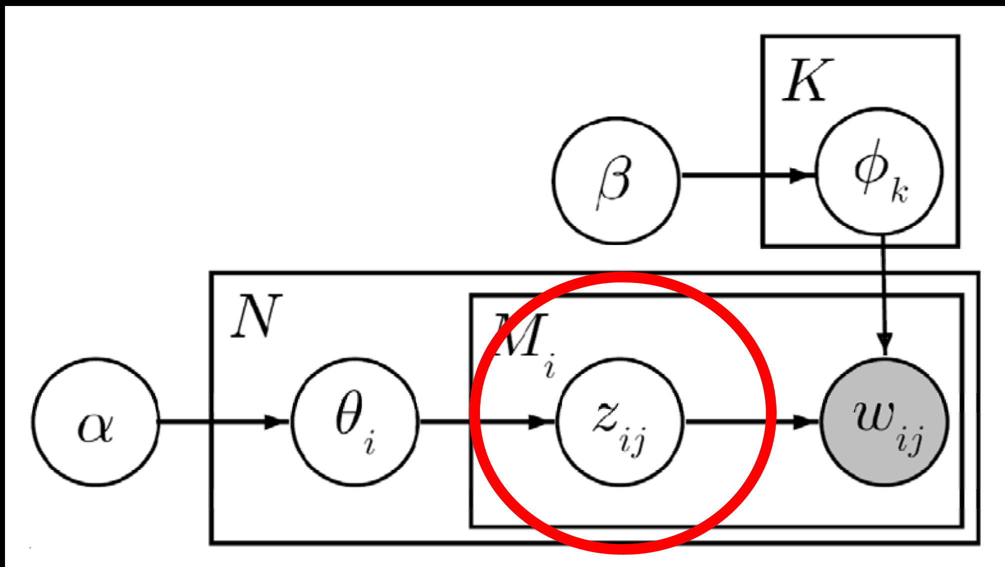
Low-rank matrix factorization



- Latent Semantic Analysis (Deerwester, et al. 1990)
- Probabilistic Latent Semantic Analysis (Hofmann 2001)

Slide credit: Sivic

Inference



w_{ij} - words

z_{ij} - topic assignments

μ_i - topic mixing weights

Á$_k$ - word mixing weights

Use Gibbs sampler to sample topic assignments

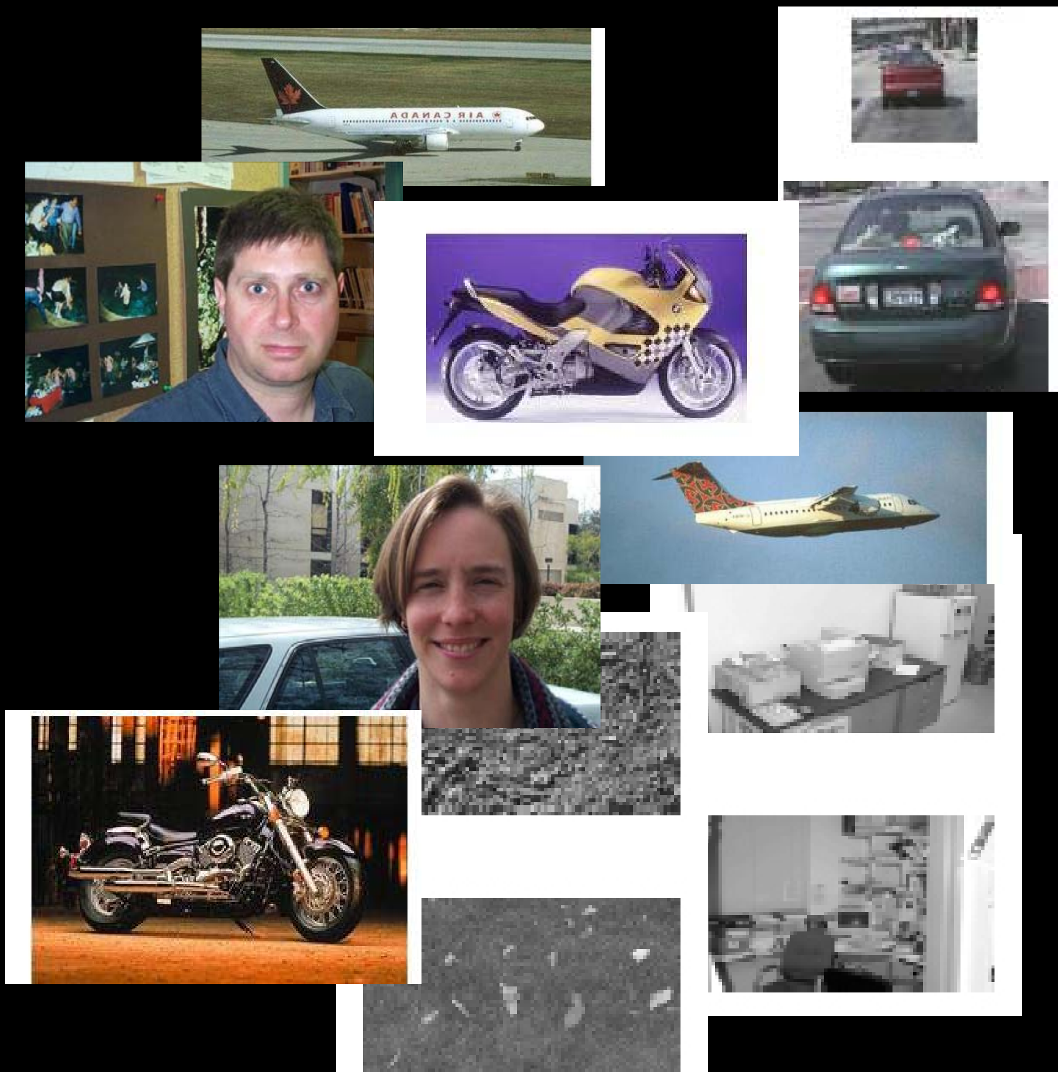
[Griffiths & Steyvers 2004]

$$z_{ij} \sim p(z_{ij} = k | w_{ij} = v, w_{\backslash(ij)}, z_{\backslash(ij)}, \alpha, \beta)$$

- Only need to maintain counts of topic assignments
- Sampler typically converges in less than 50 iterations
- Run time is less than an hour

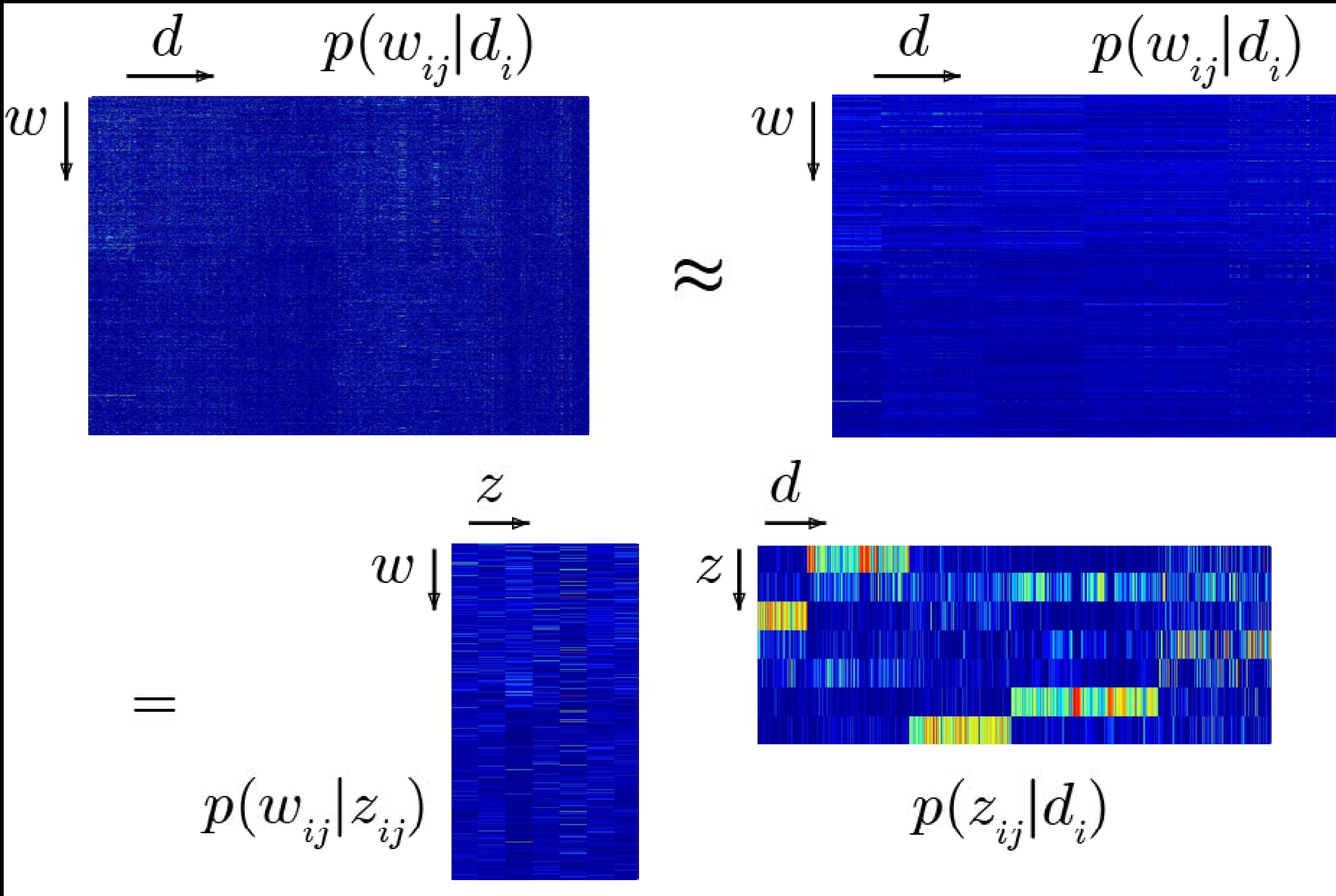
Slide credit: Sivic

Apply to Caltech 4 + background images

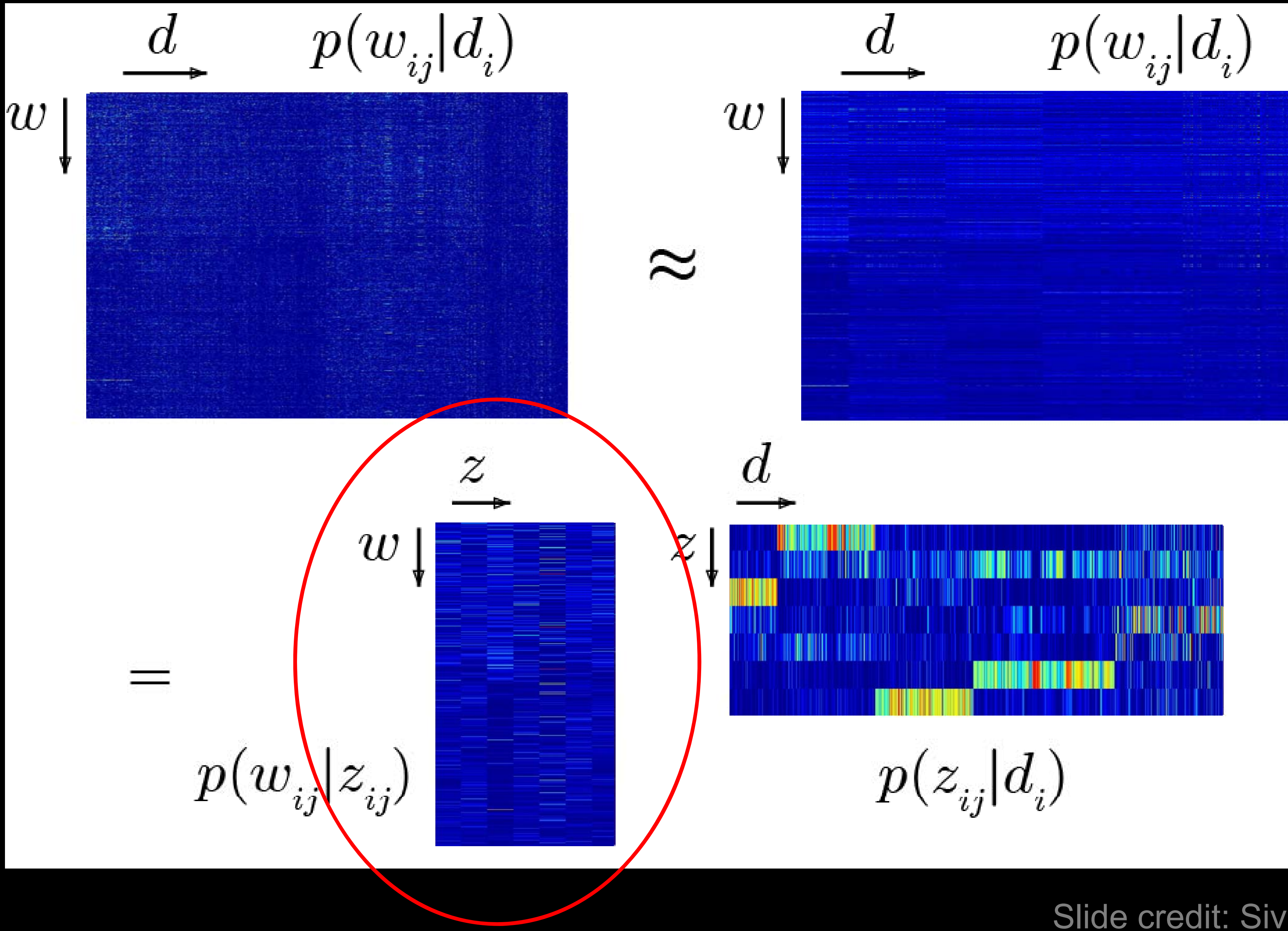


Faces	435
Motorbikes	800
Airplanes	800
Cars (rear)	1155
Background	900
Total:	**4090**

Slide credit: Sivic



Slide credit: Sivic



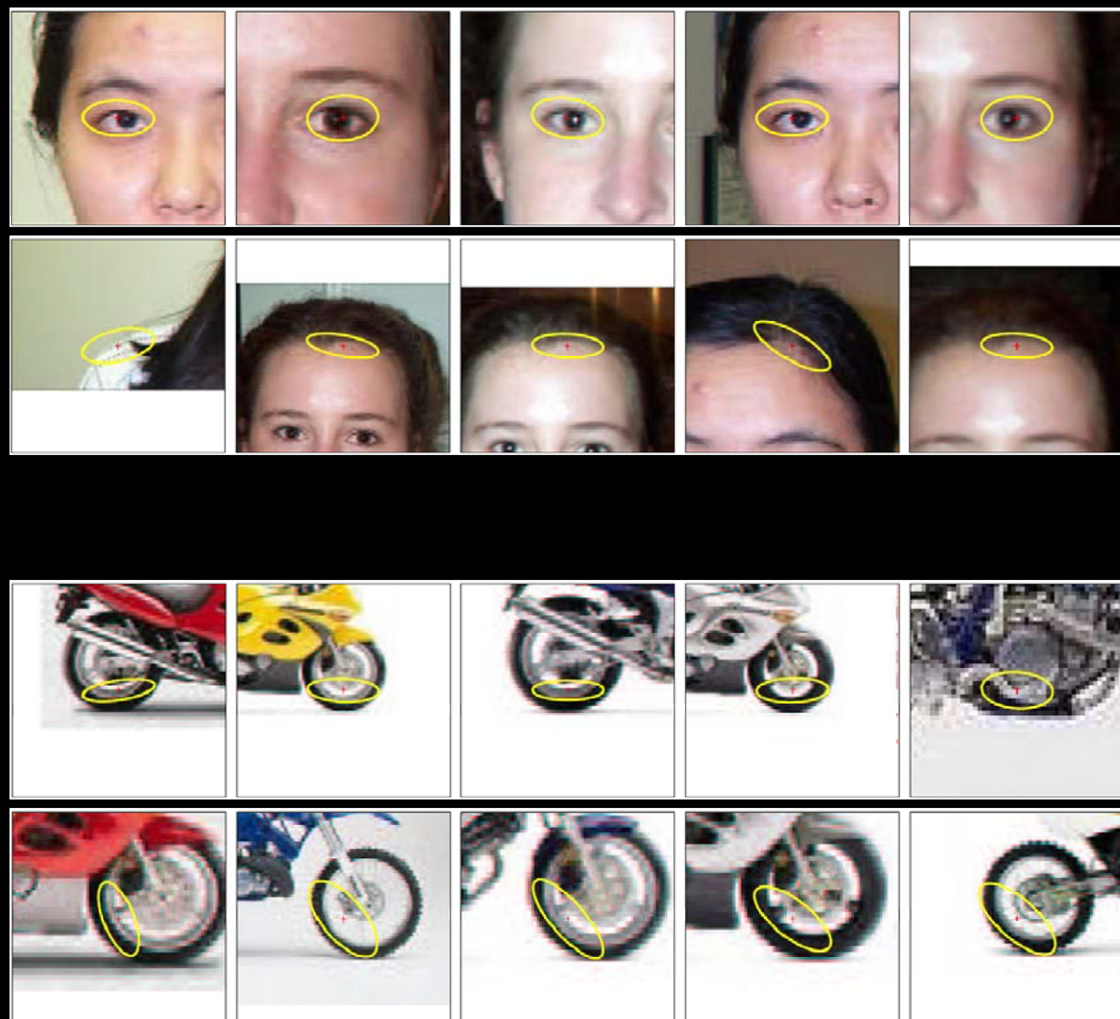
Slide credit: Sivic

Most likely words given topic

Topic 1

Word 1

Word 2



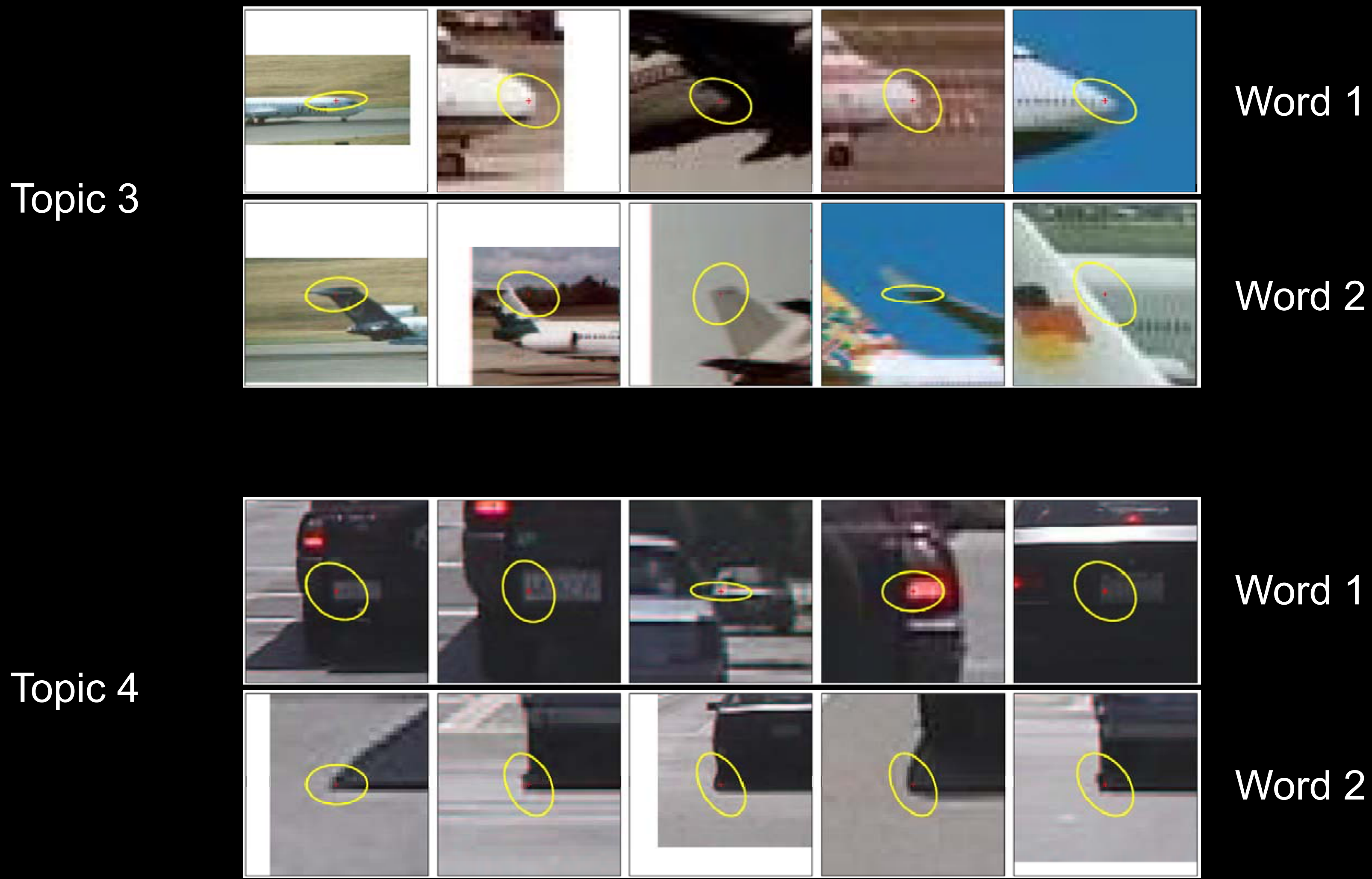
Topic 2

Word 1

Word 2

Slide credit: Sivic

Most likely words given topic



Slide credit: Sivic

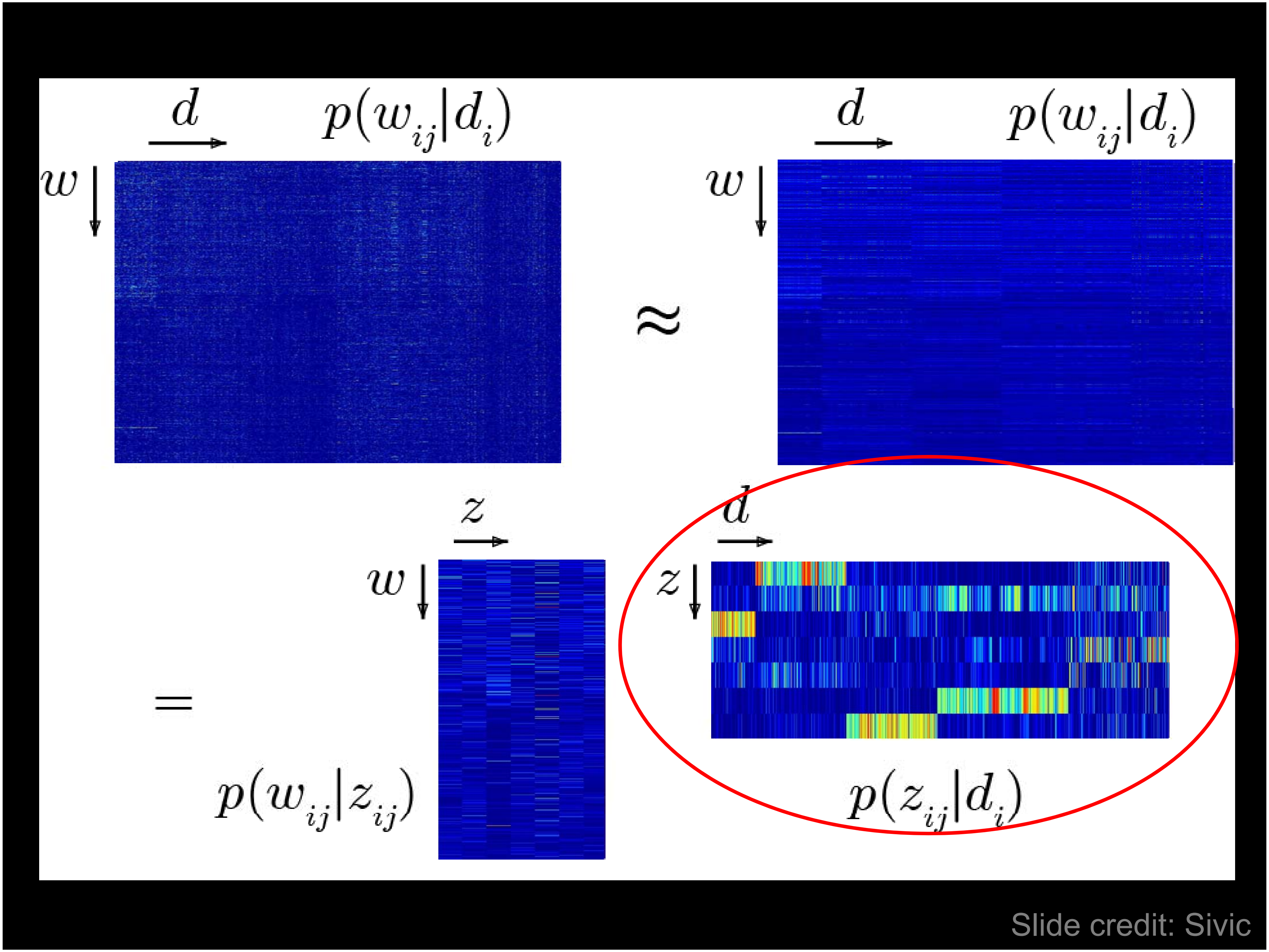
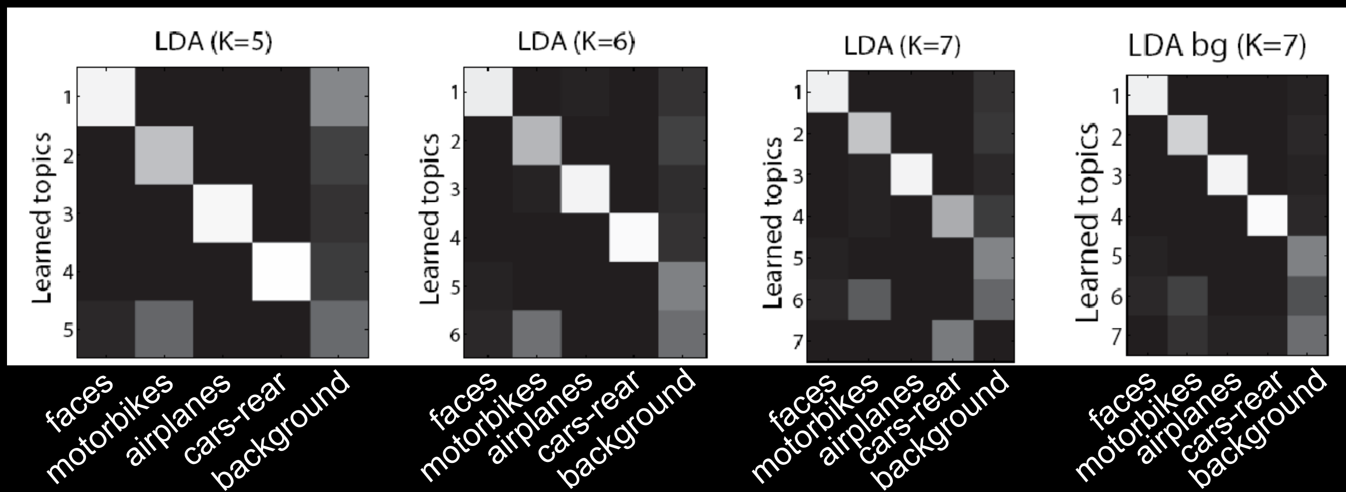

d
$p(w_{ij}|d_i)$
w
$\approx$
d
$p(w_{ij}|d_i)$
w
$=$
z
w
$p(w_{ij}|z_{ij})$
d
z
$p(z_{ij}|d_i)$
Slide credit: Sivic

Image clustering

Confusion matrices:



Average confusion:

Expt.	Categories	T	LDA %	LDA #	pLSA %	pLSA #	KM baseline %	KM baseline #
(1)	4	4	97	86	98	70	72	908
(2)	4 + bg	5	78	931	78	931	56	1820
(2)*	4 + bg	6	84	656	76	1072	–	–
(2)*	4 + bg	7	78	1007	83	768	–	–
(2)*	4 + bg-fxd	7	90	330	93	238	–	–

edit: Sivic

Comparison with supervised model

Percent ROC equal error rate

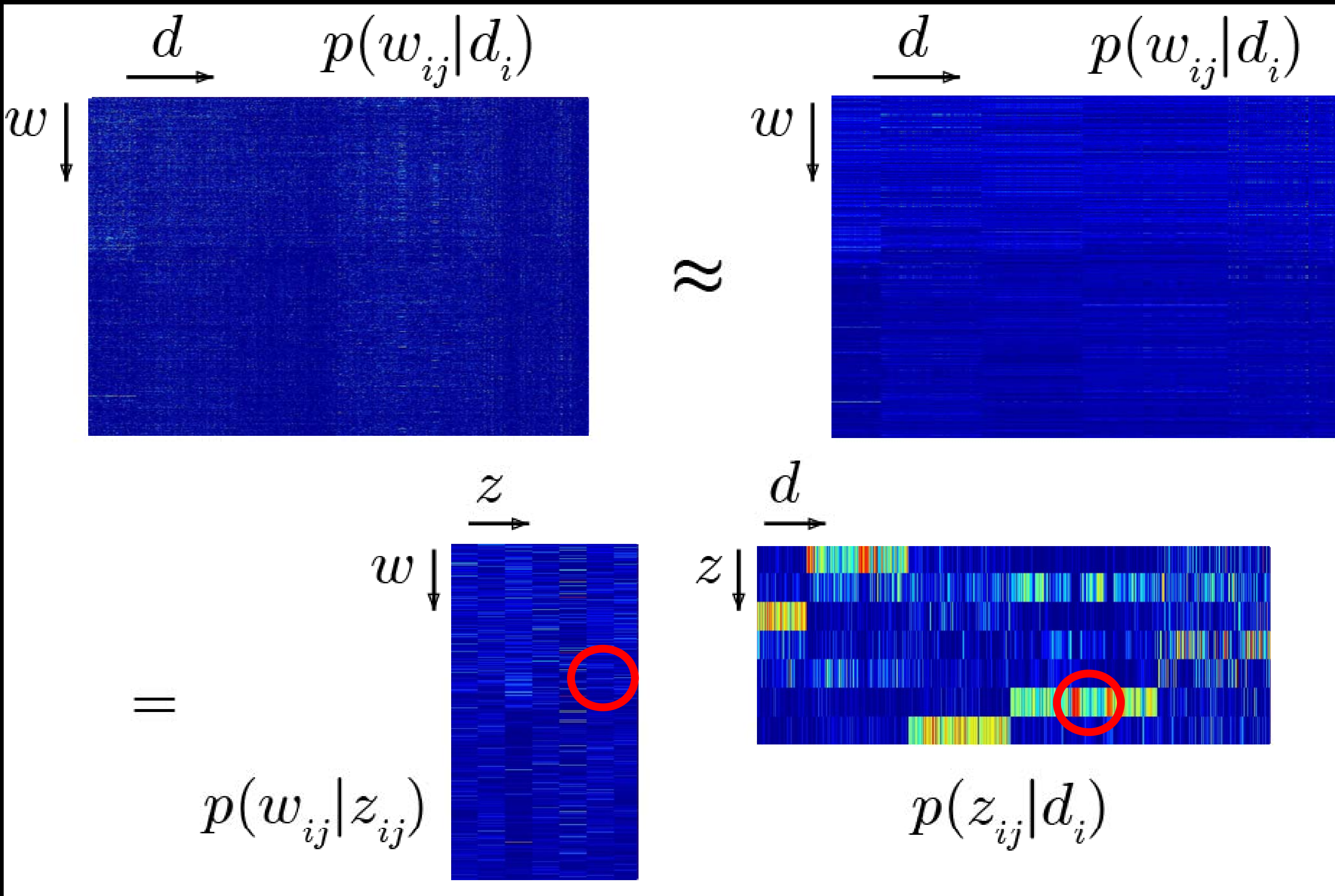
	LDA	Constellation model [Fergus et al. '03]
Faces	7.8	3.6
Motorbikes	9.9	6.7
Airplanes	2.5	7.0
Cars rear	8.5	9.7

- Comparable performance to constellation model
- Level of supervision:

  LDA: one number (of topics)

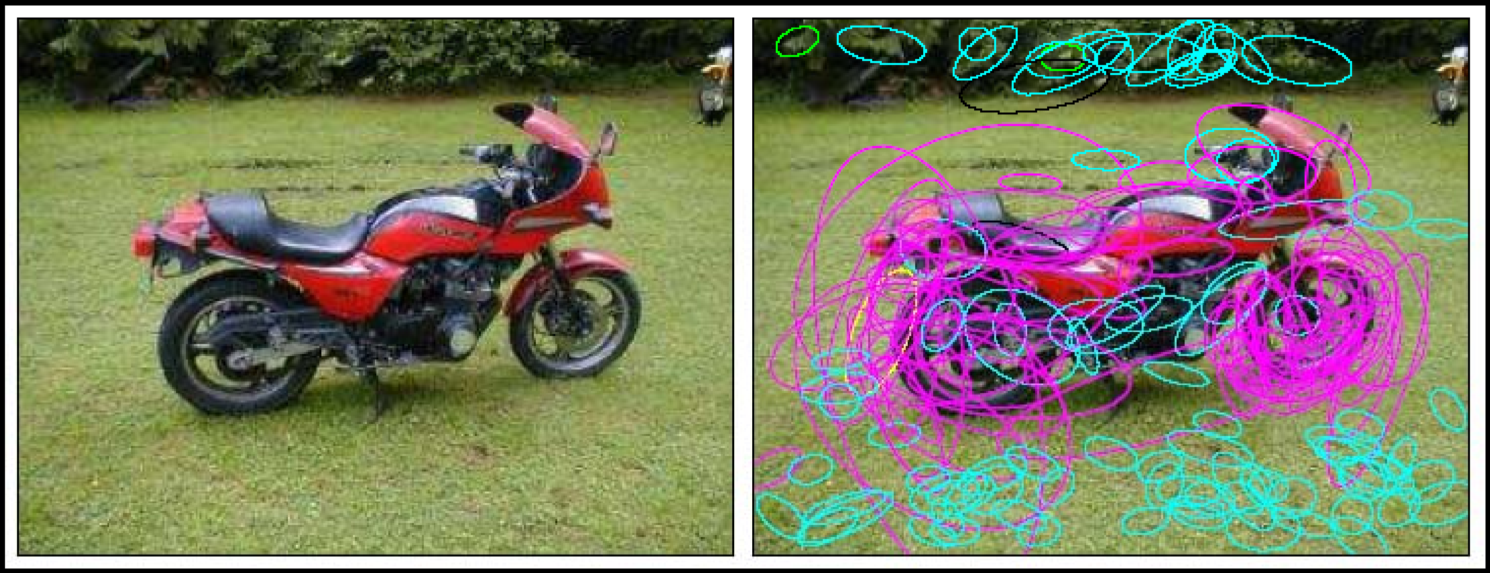
  Constellation model: 400 labels for each category
- Also an indication of the level of difficulty of the Caltech 5 dataset

Slide credit: Sivic

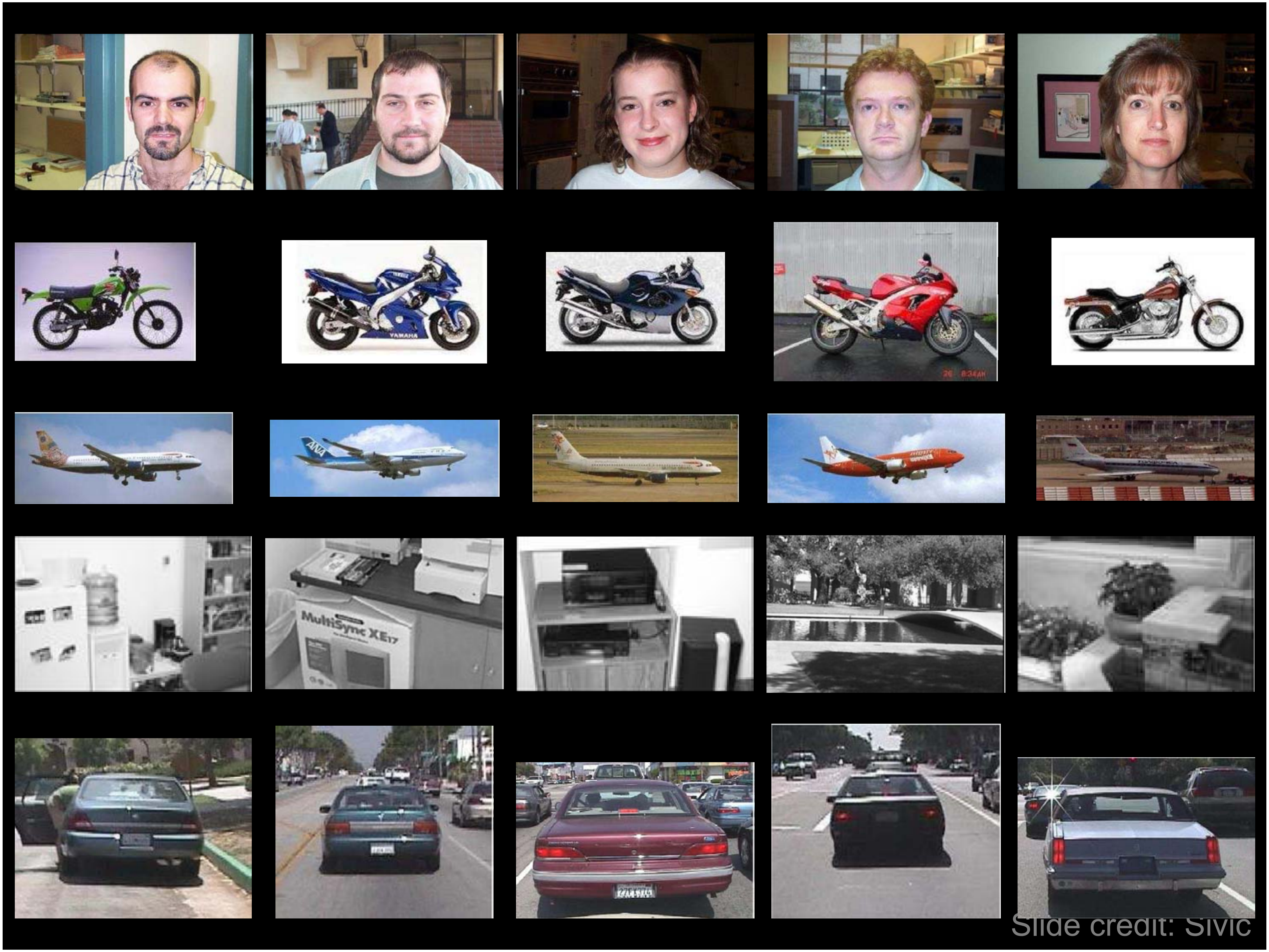


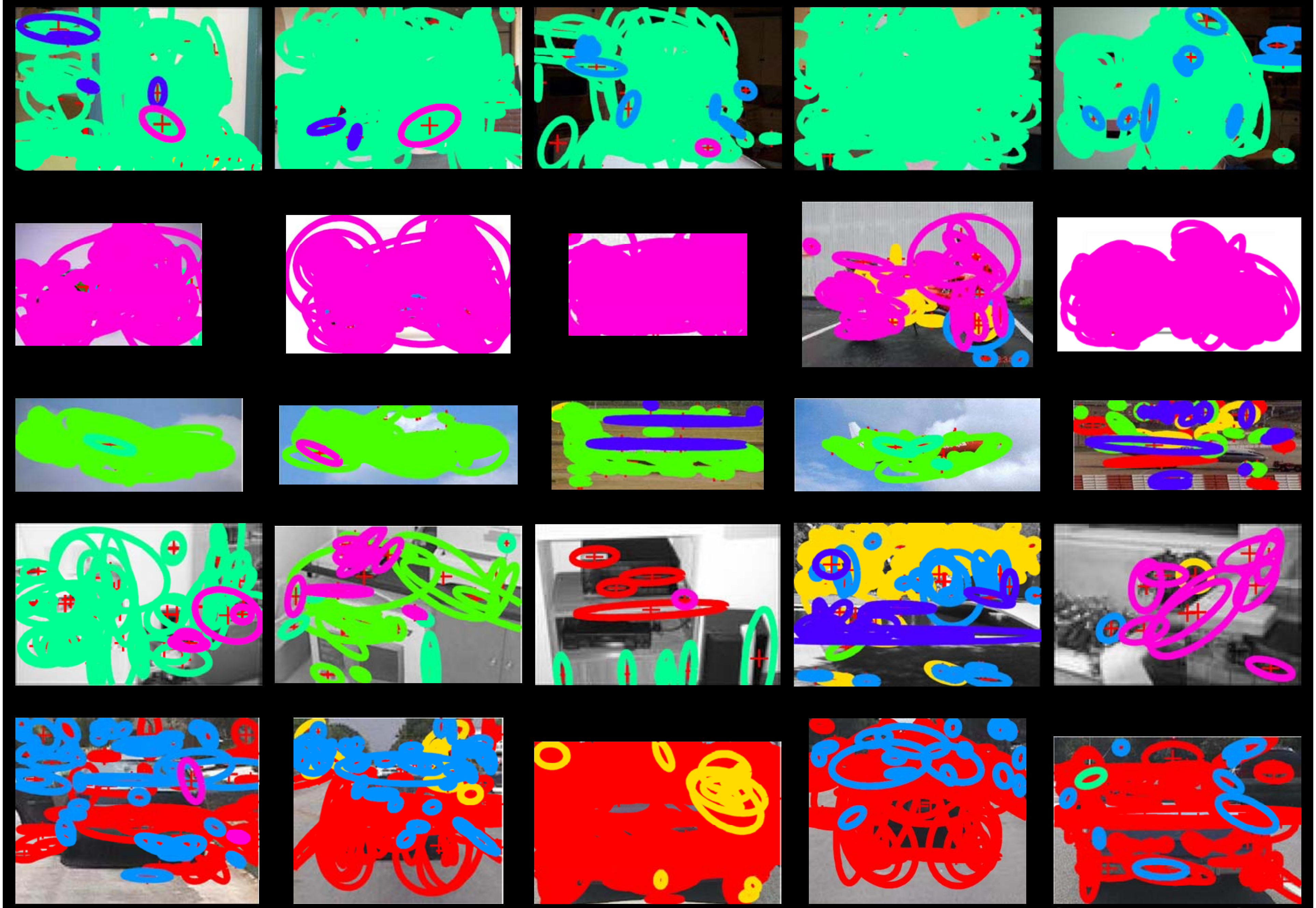
Slide credit: Sivic

Image as a mixture of topics (objects)



Slide credit: Sivic


ANA
MultiSync XE17
Slide credit: Sivic



Slide credit: Sivic

Summary

- Discovered visual topics corresponding to object categories from a corpus of unlabeled images
- Used visual words representation and topic discovery models from the text understanding community
- Classification on unseen images is comparable to supervised methods on Caltech 5 dataset
- The discovered categories can be localized within an image

Slide credit: Sivic

Today

Sudderth guest lecture:

- Constellation Models (Fergus)
- Unsupervised Object Discovery with pLSA (Sivic)
- **Scene Models (Li)**
- Transformed Models (Sudderth)

Daphna B. student presentation:

- pLSA models of activity (Neibles)

Moreels guest lecture:

- A probabilistic formulation of voting / SIFT (Moreels)

A Bayesian Hierarchical Model for Learning Natural Scene Categories

or

“LDA for Scene Recognition”

Fei-Fei Li & Pietro Perona

CVPR 2005

Slide credit: Li

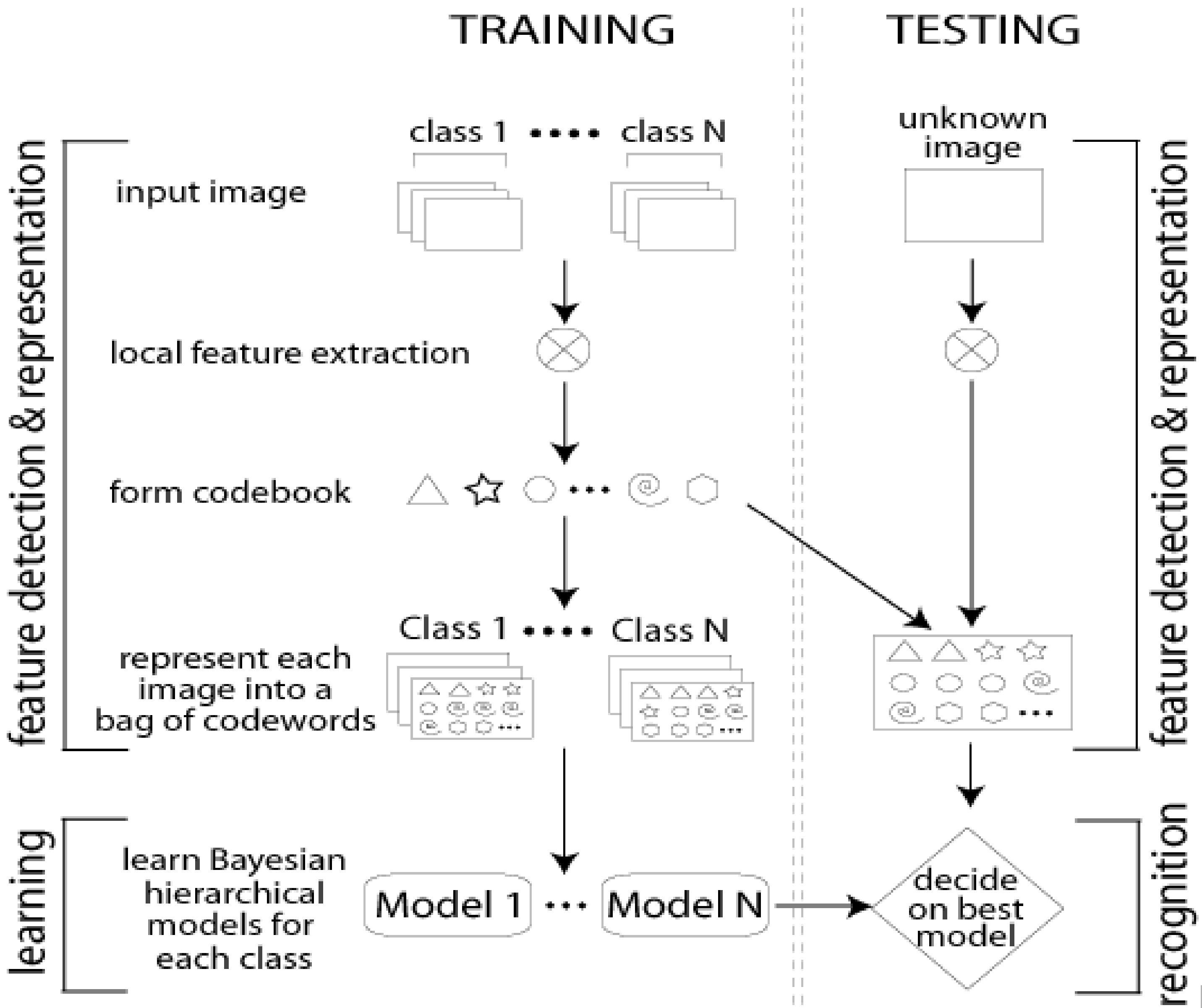
Scene Recognition

Outdoor Scenes

Indoor Scenes

Slide credit: Li

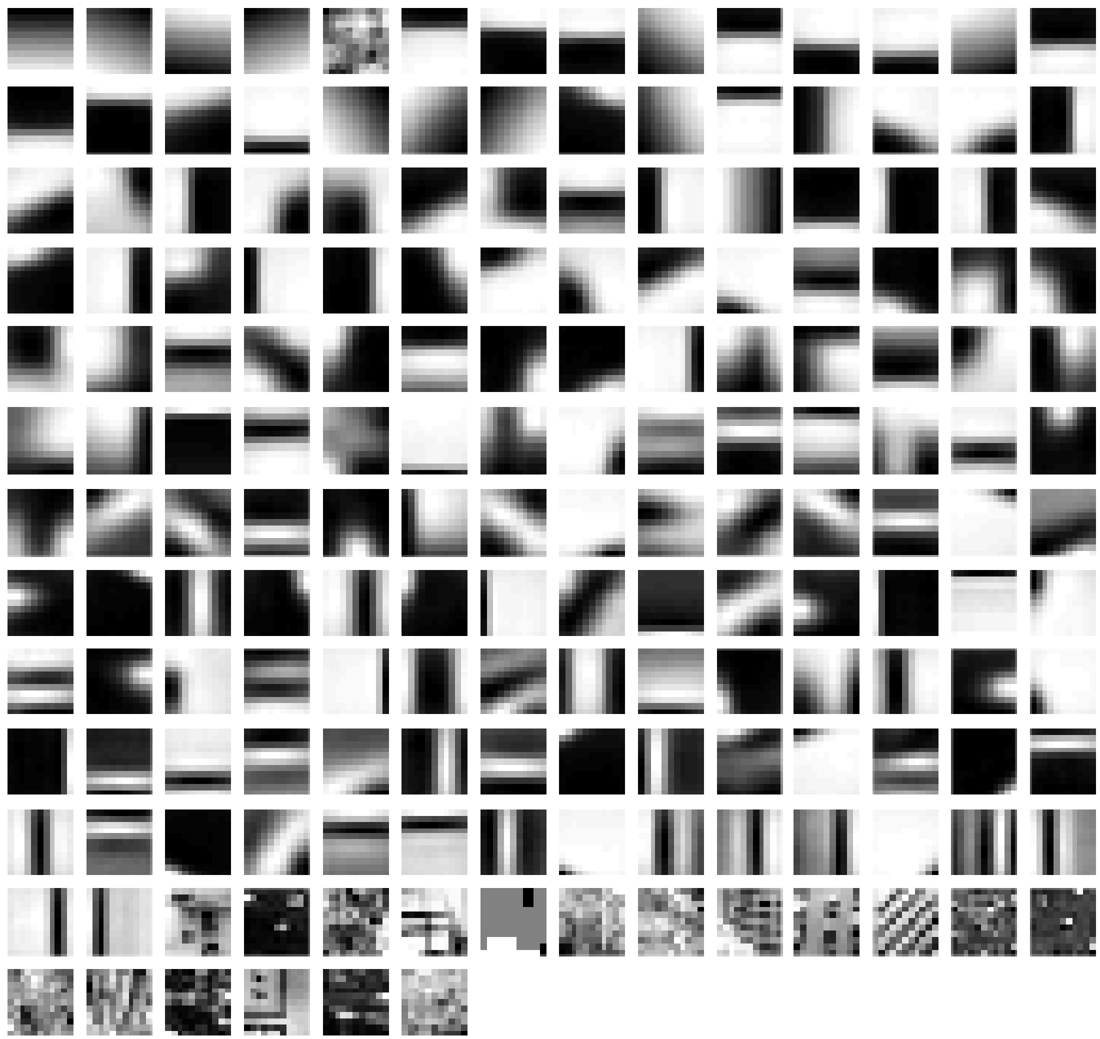
System Block Diagram



Slide credit: Li

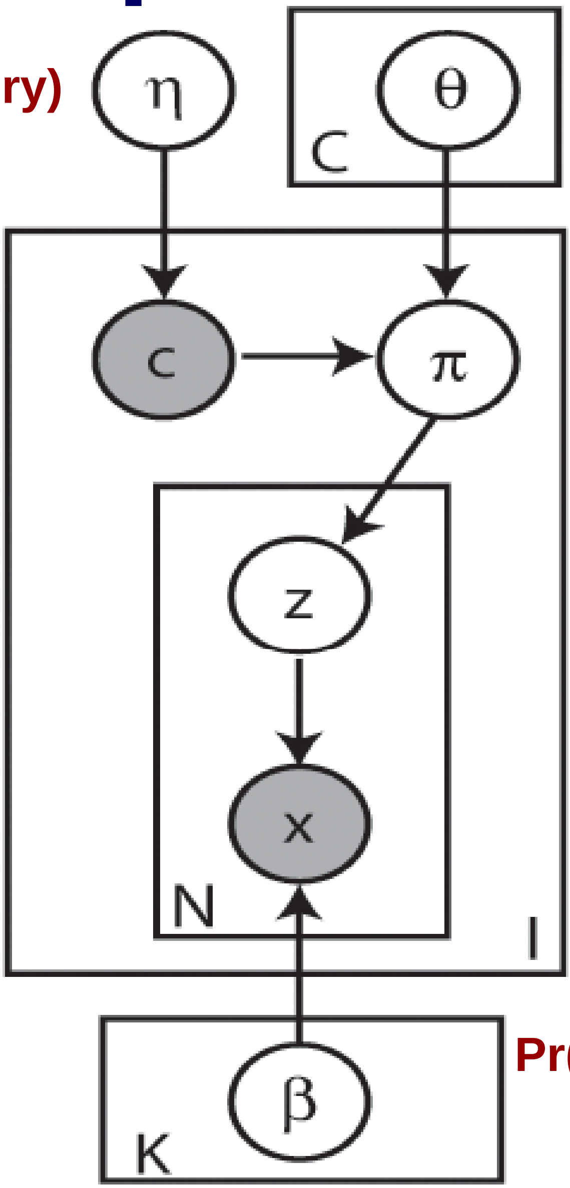
Visual Words

Descriptor	Grid	Random	Saliency [4]	DoG [7]
11 × 11 Pixel	64.0%	47.5%	45.5%	N/A
128-dim Sift	65.2%	60.7%	53.1%	52.5%

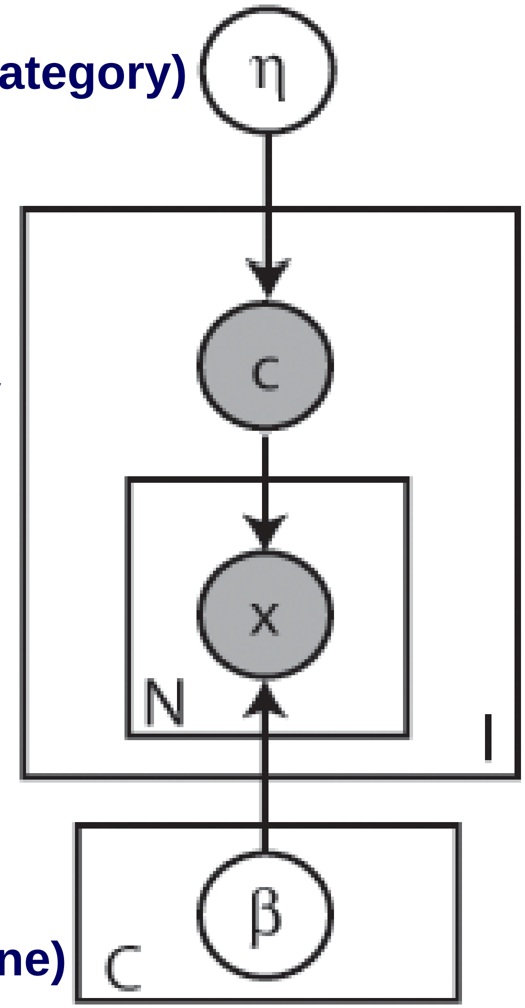


Slide credit: Li

Graphical Models for Scenes



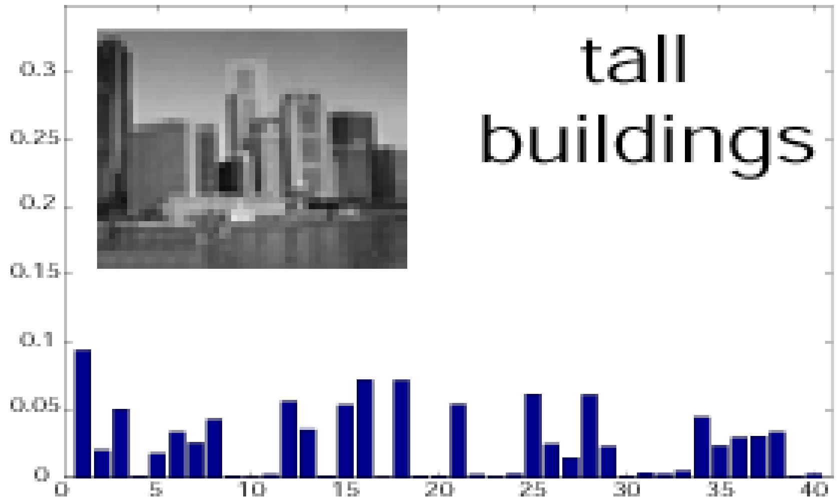
Latent Dirichlet Allocation



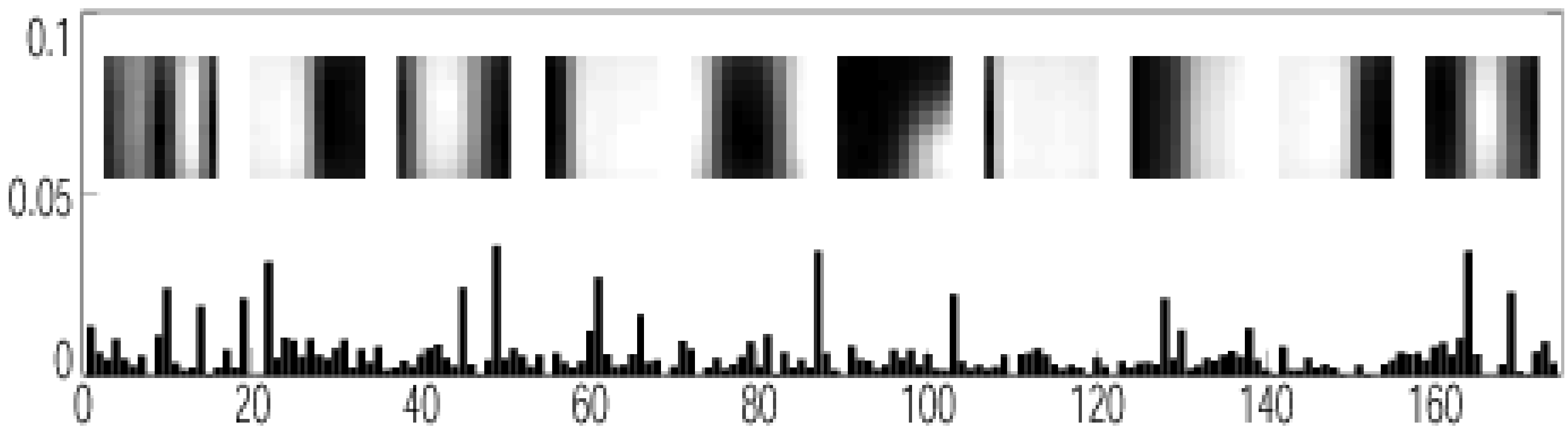
Texton Model

Slide credit: Li

Model for Tall Buildings



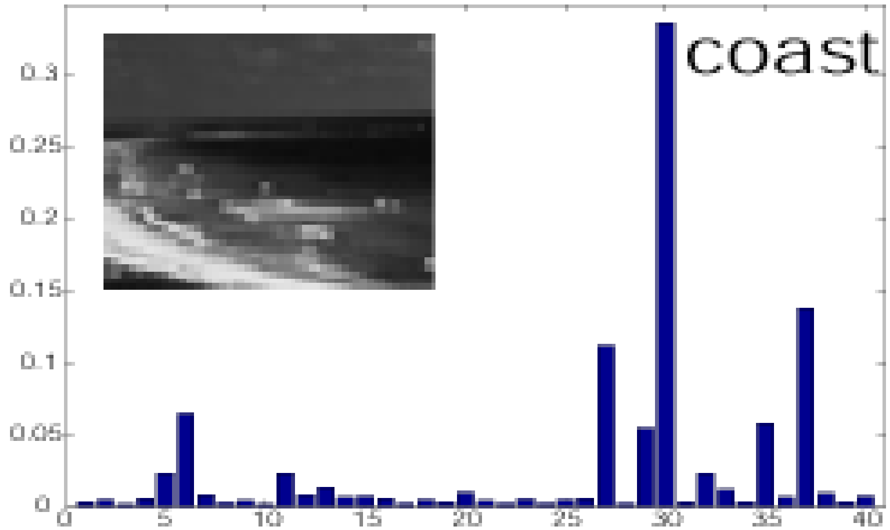
E[Pr(topic | scene)]



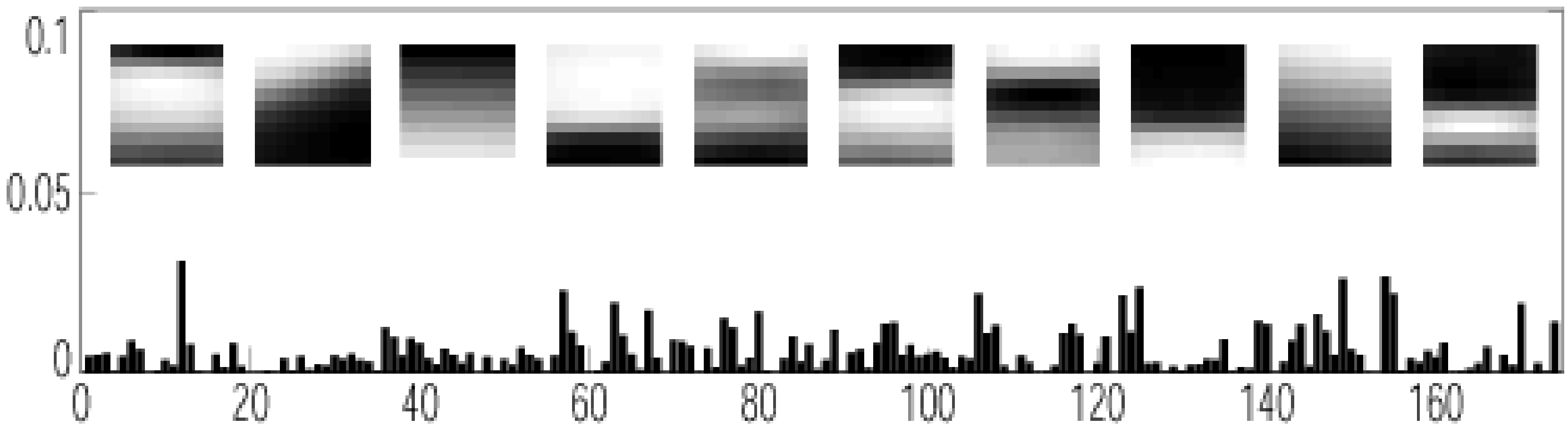
E[Pr(word | scene)]

Slide credit: Li

Model for Coasts



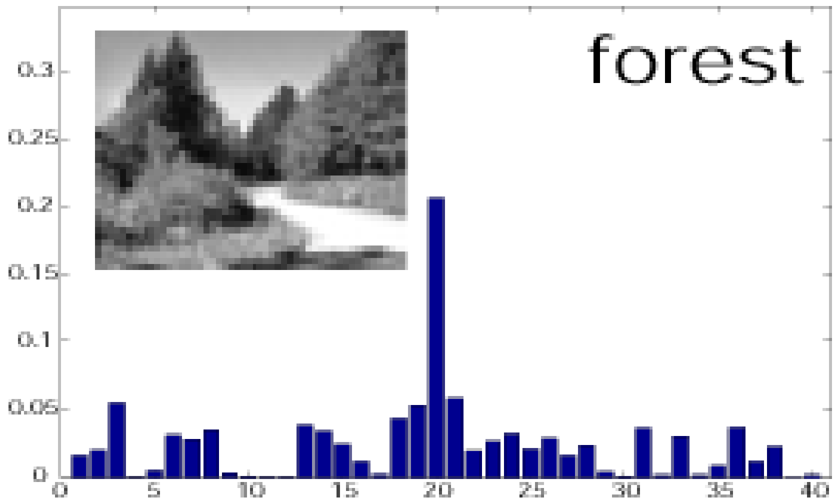
E[Pr(topic | scene)]



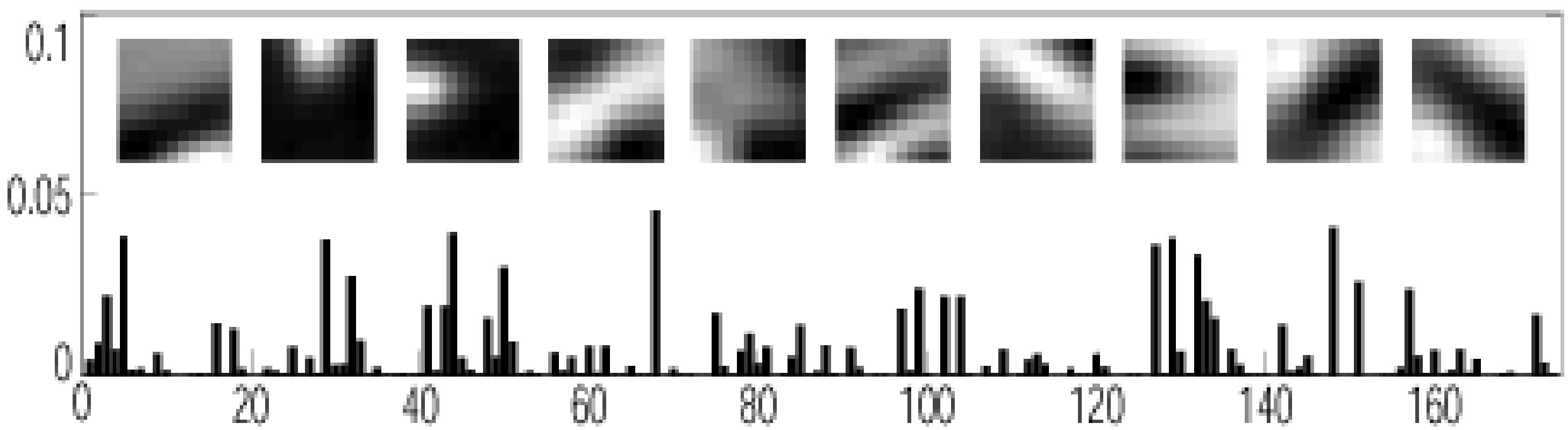
E[Pr(word | scene)]

Slide credit: Li

Model for Forests



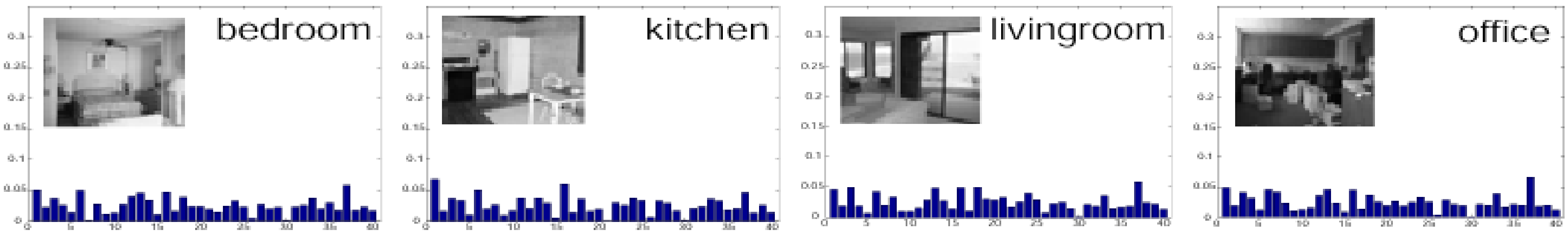
E[Pr(topic | scene)]



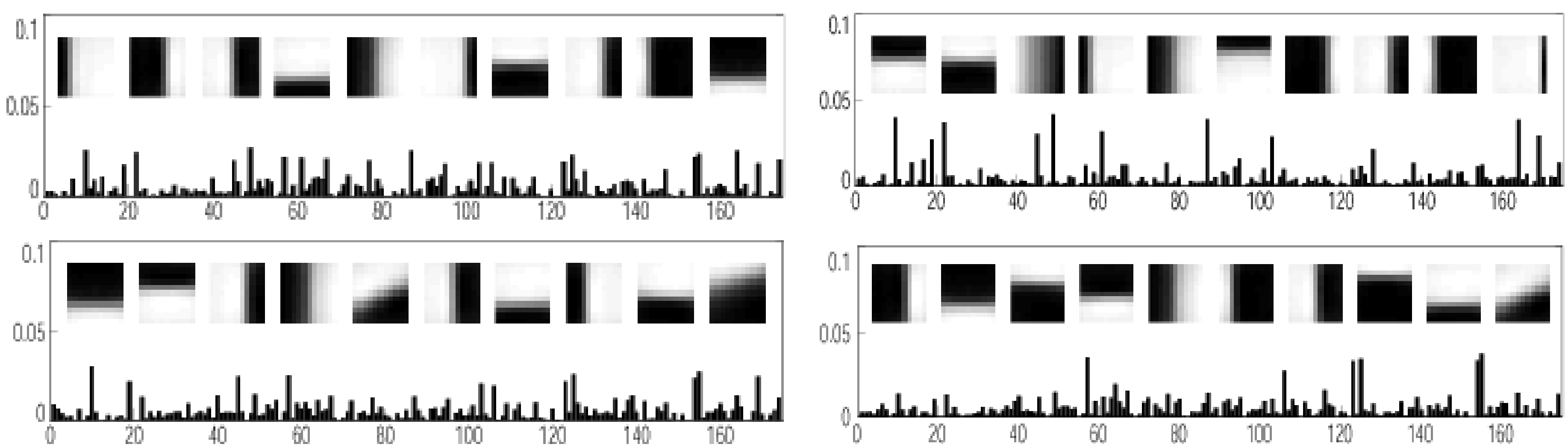
E[Pr(word | scene)]

Slide credit: Li

Models for Indoor Scenes



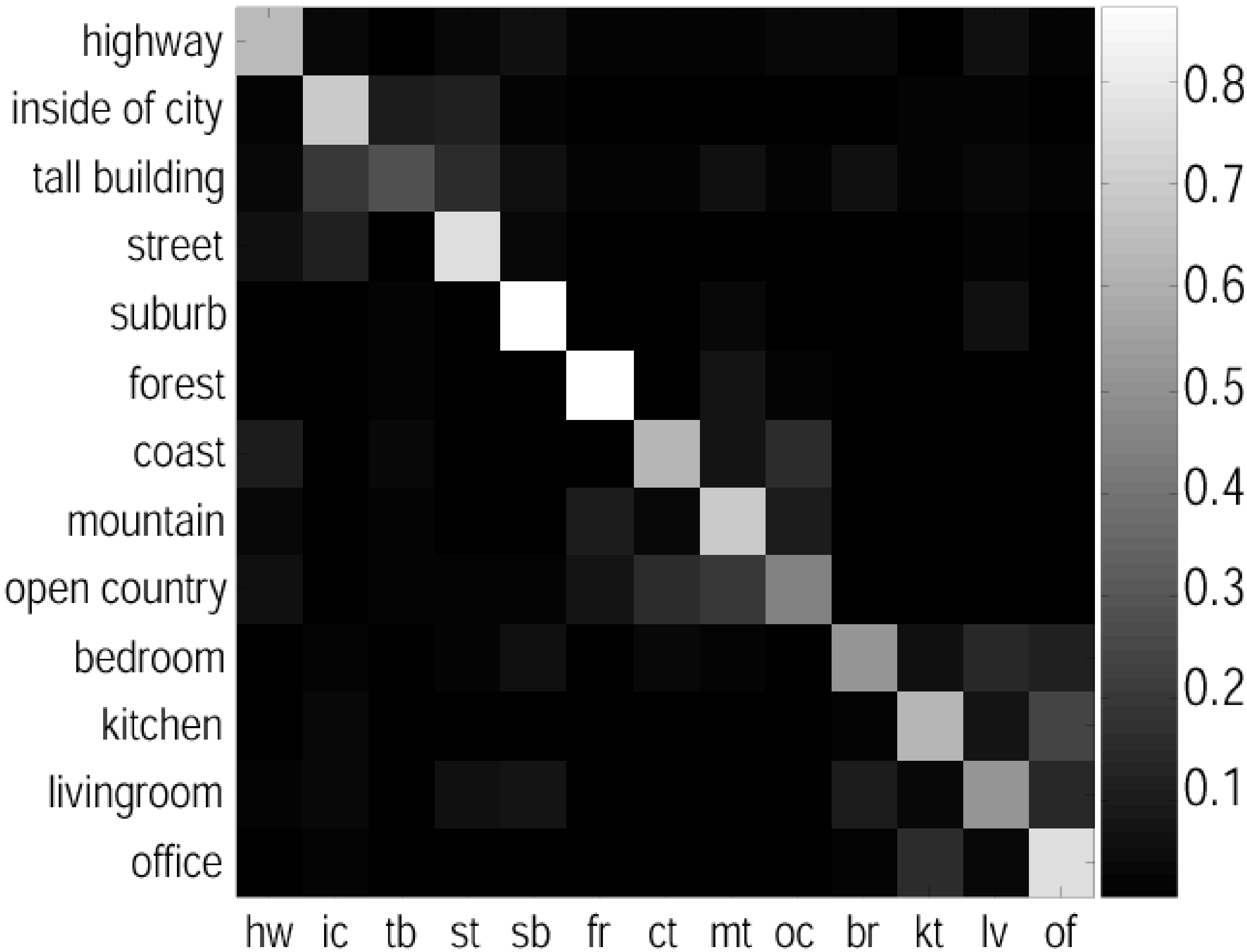
E[Pr(topic | scene)]



E[Pr(word | scene)]

Slide credit: Li

Scene Recognition Performance



Slide credit: Li

Scene Relationships

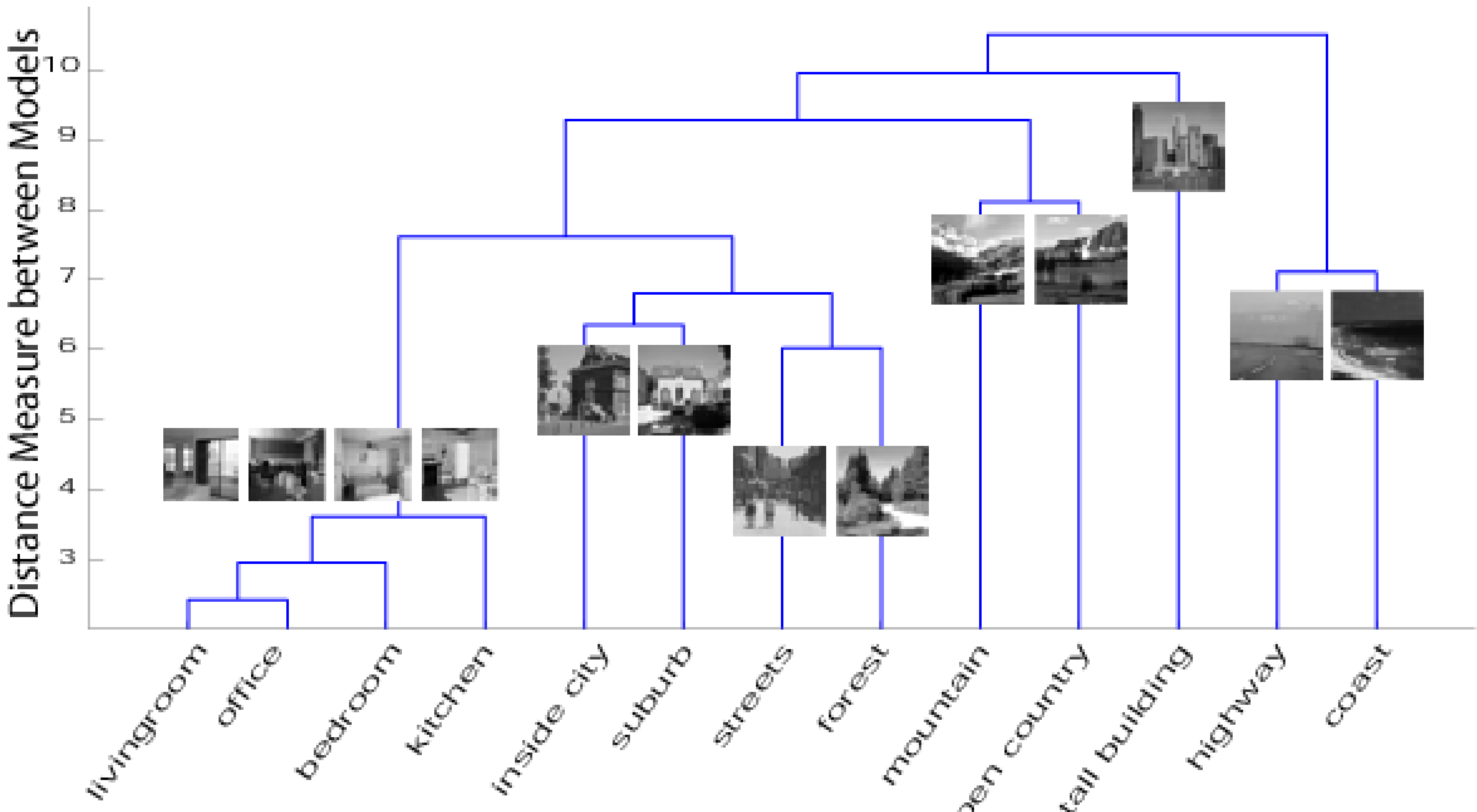
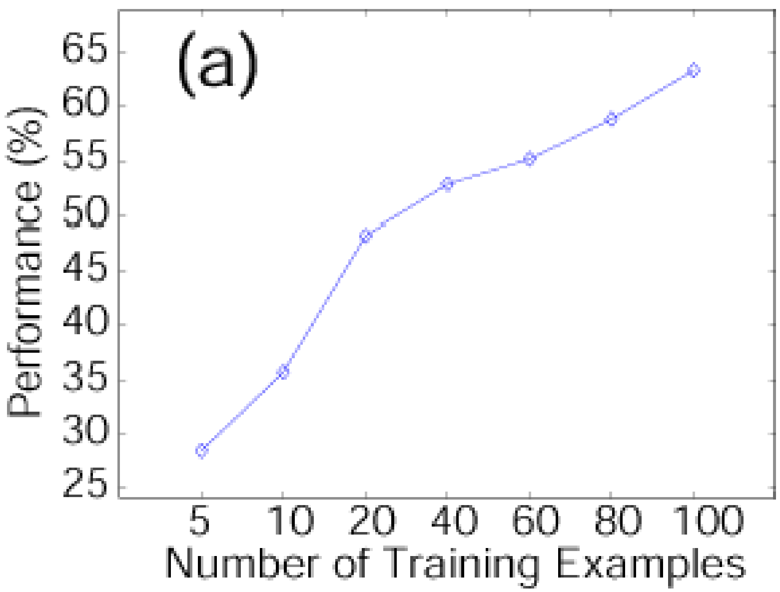
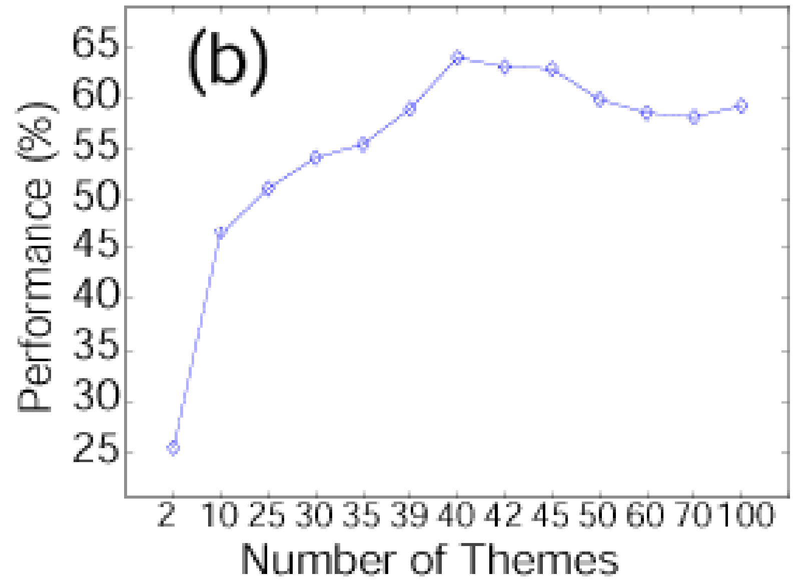


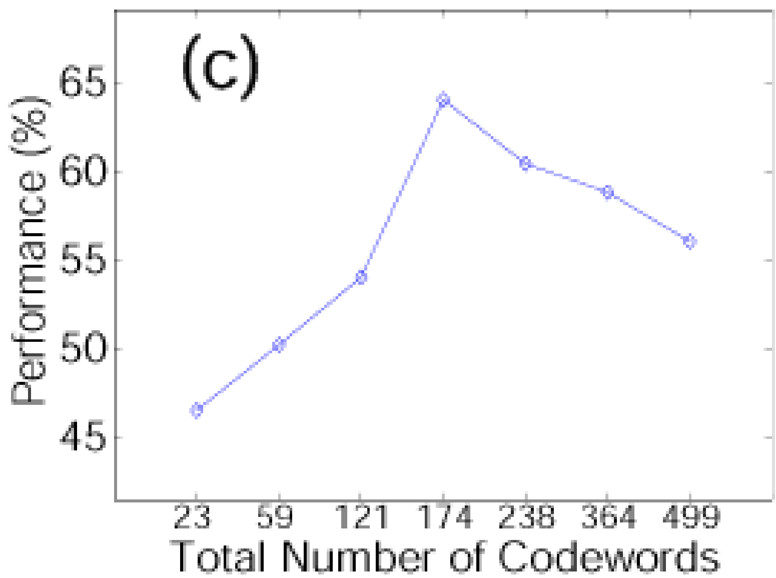
Figure 9. Dendrogram of the relationship of the 13 category

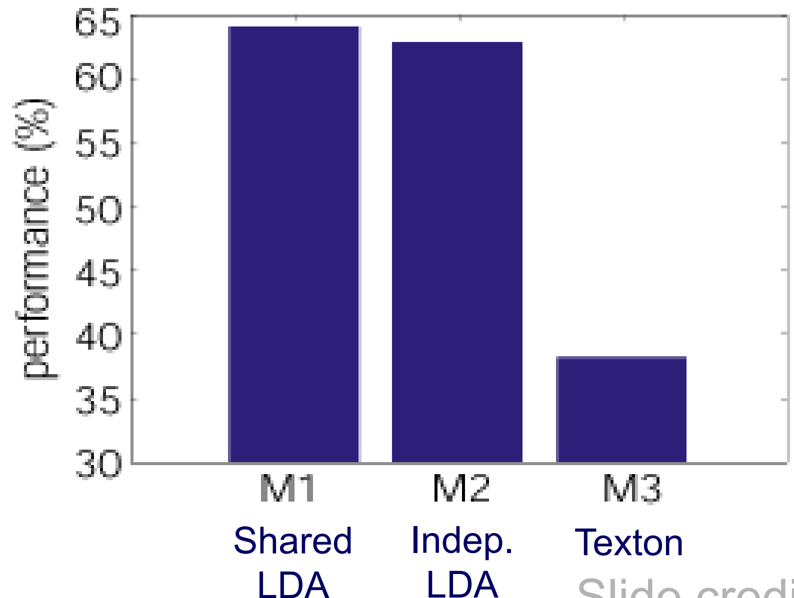
Slide credit: Li

Model Parameters









Slide credit: Li

Today

Sudderth guest lecture:

- Constellation Models (Fergus)
- Unsupervised Object Discovery with pLSA (Sivic)
- Scene Models (Li)
- **Transformed Models (Sudderth)**

Daphna B. student presentation:

- pLSA models of activity (Neibles)

Moreels guest lecture:

- A probabilistic formulation of voting / SIFT (Moreels)

Learning Object Appearance Models via Transformed Dirichlet Processes

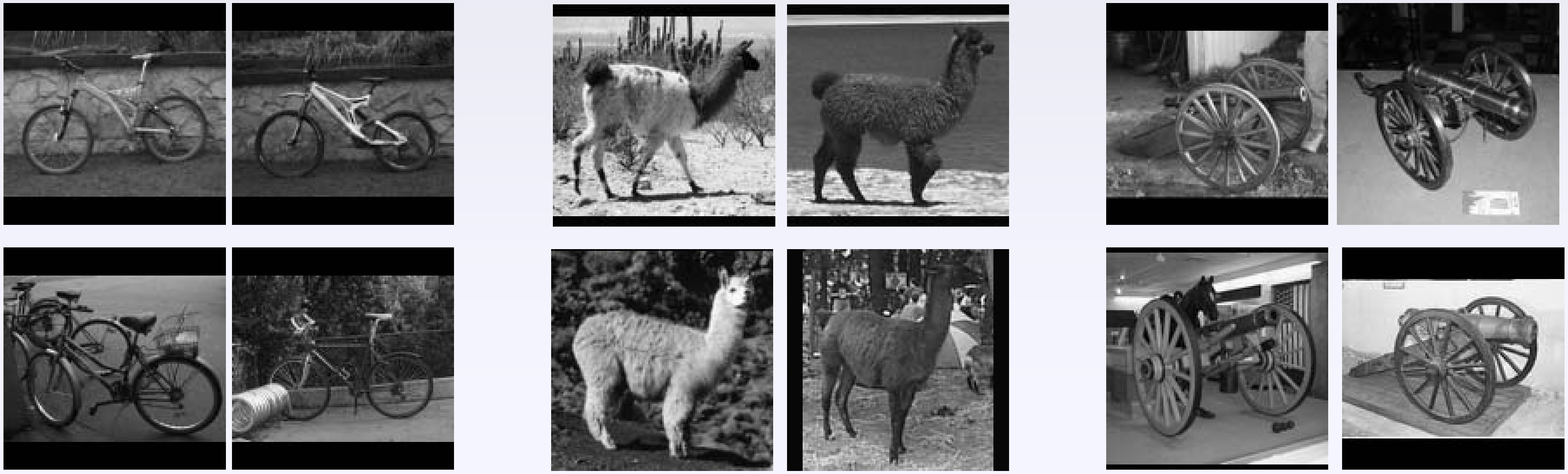
Erik Sudderth

University of California, Berkeley

Joint work with

Antonio Torralba
William Freeman
Alan Willsky

Visual Object Categorization



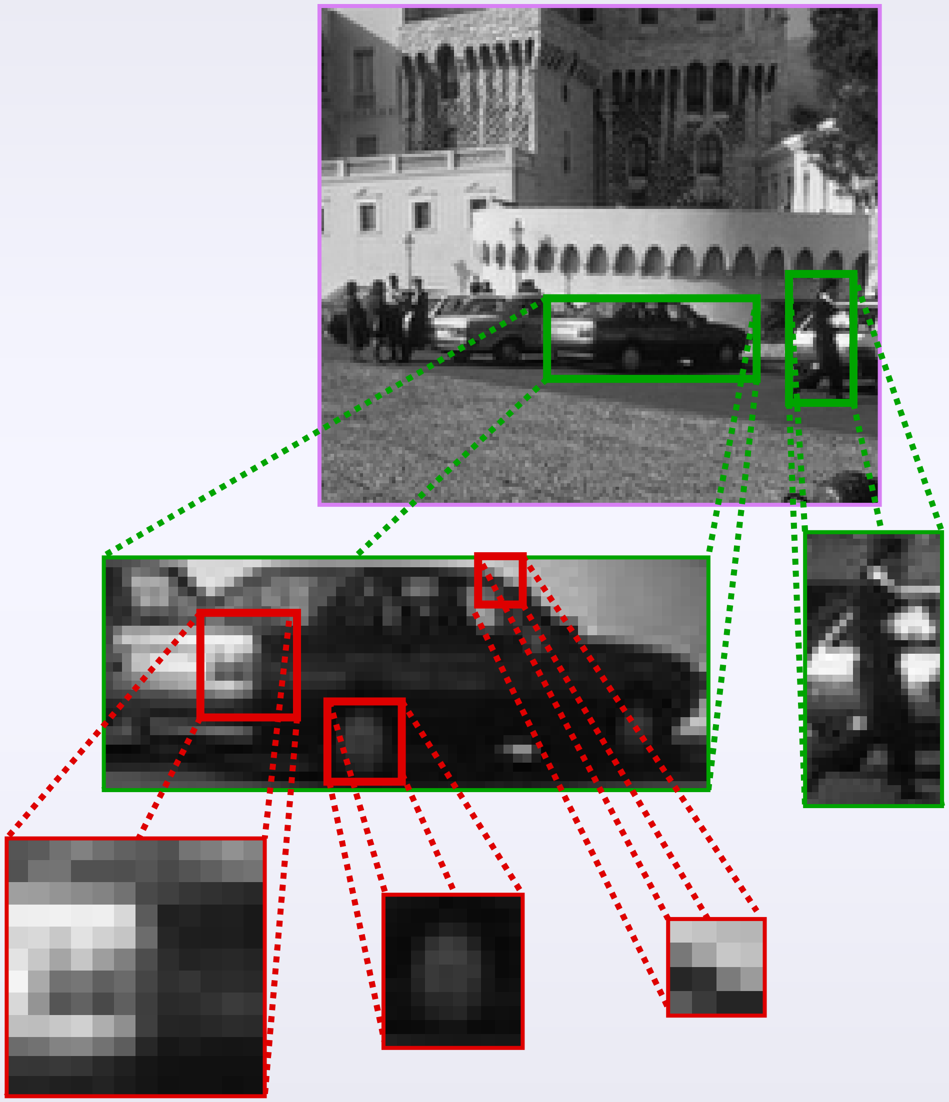
- **GOAL:** Visually *recognize* and *localize* object categories
- Robustly *learn* appearance models from few examples
  - ➢ Hierarchical model *transfers* knowledge among categories
  - ➢ Nonparametric, *Dirichlet process* prior gives flexibility

Detecting Objects in Scenes

Sliding Window Approach

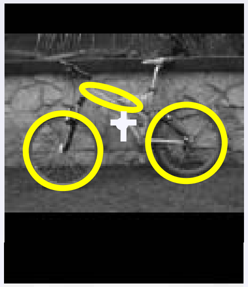
Greedy Feature Extraction Approach

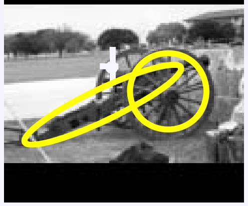
Scenes, Objects, and Parts



Outline

Object Recognition with Shared Parts

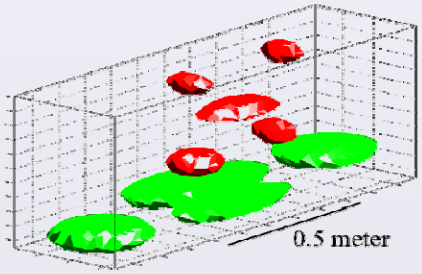




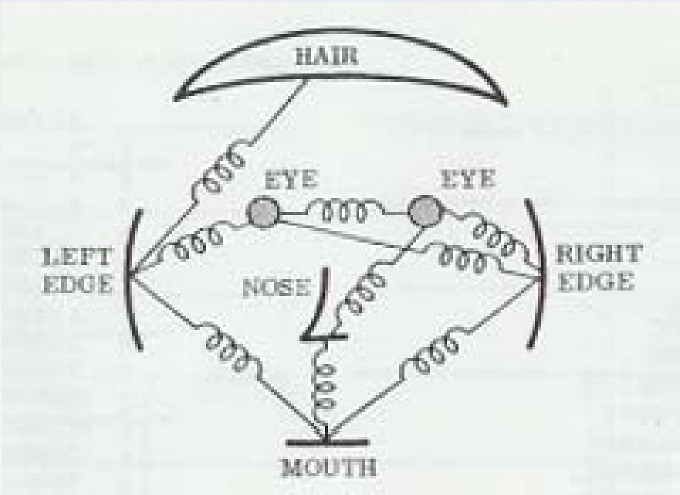
- Learning parts via Dirichlet processes
- Hierarchical DP model for 16 object categories

Multiple Object Scenes

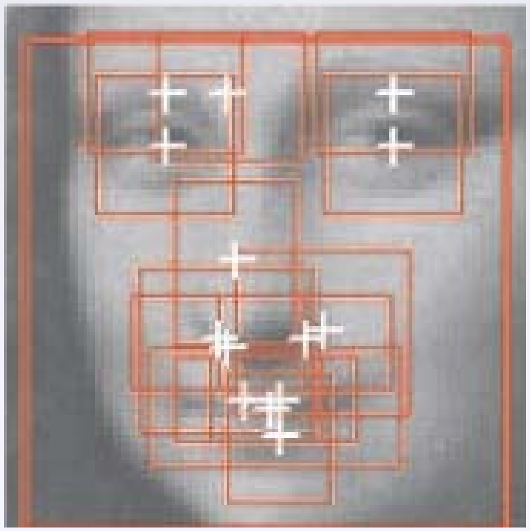
- Transformed Dirichlet processes
- Part-based models for 2D scenes
- Joint object detection & 3D reconstruction



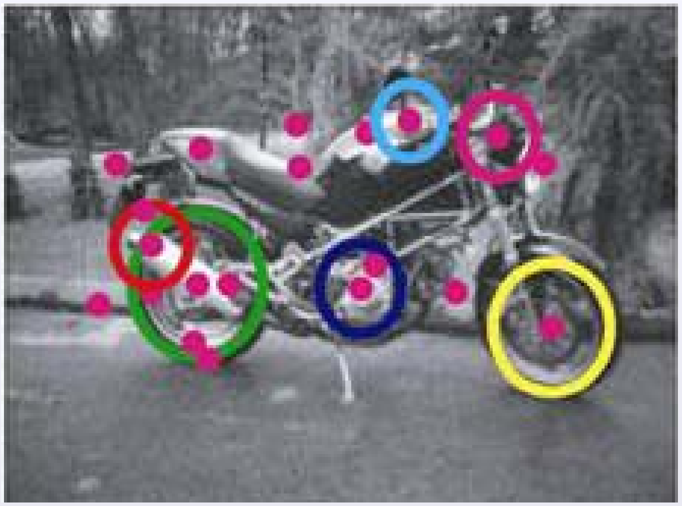
Describing Objects with Parts



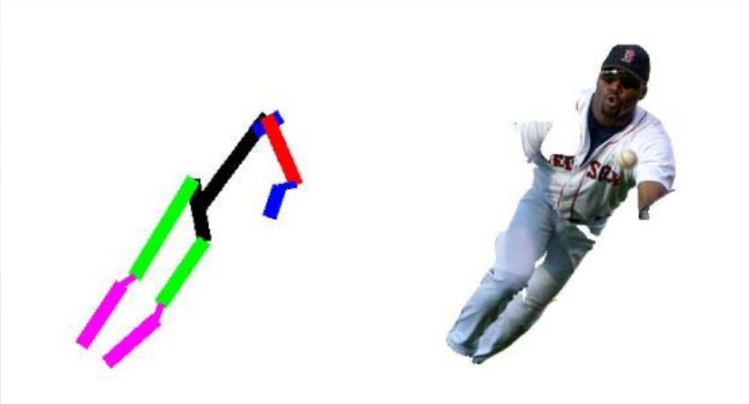
Pictorial Structures
Fischler & Elschlager, IEEE Trans. Comp. 1973



Cascaded SVM Detectors
Heisele, Poggio, et. al., NIPS 2001

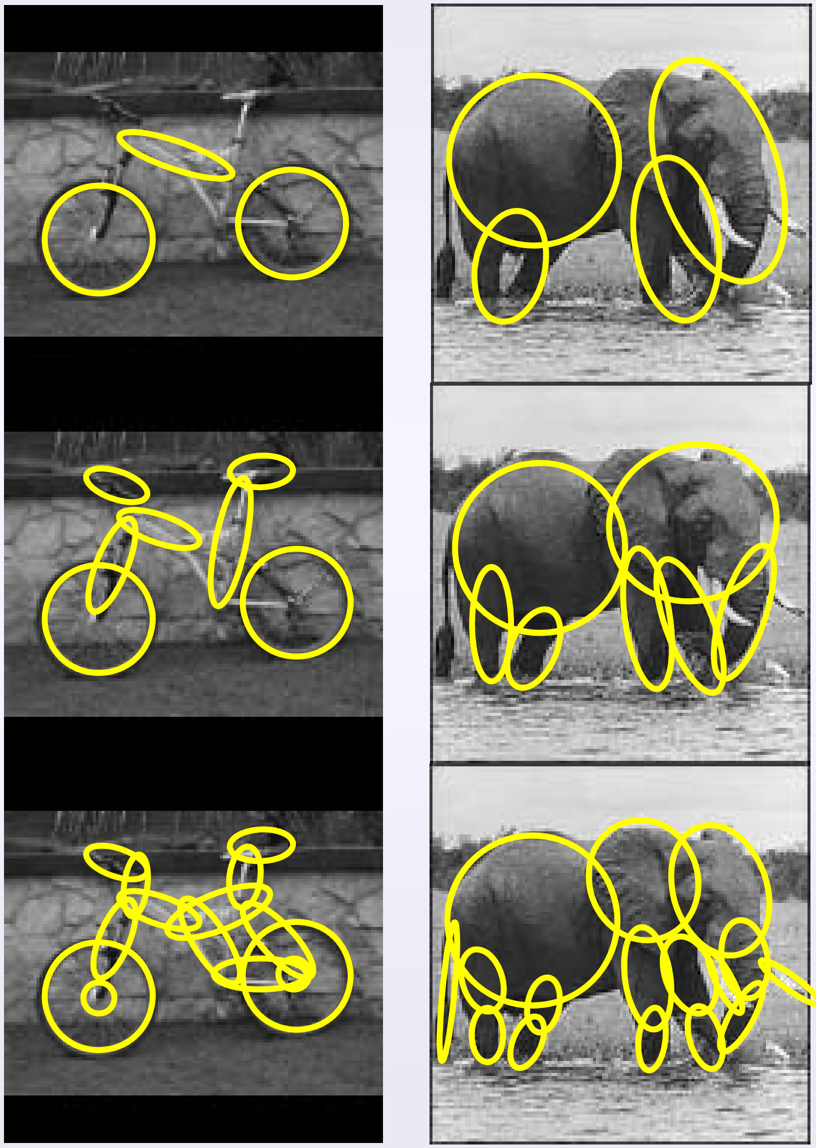


Constellation Model
Fergus, Perona, & Zisserman, CVPR 2003

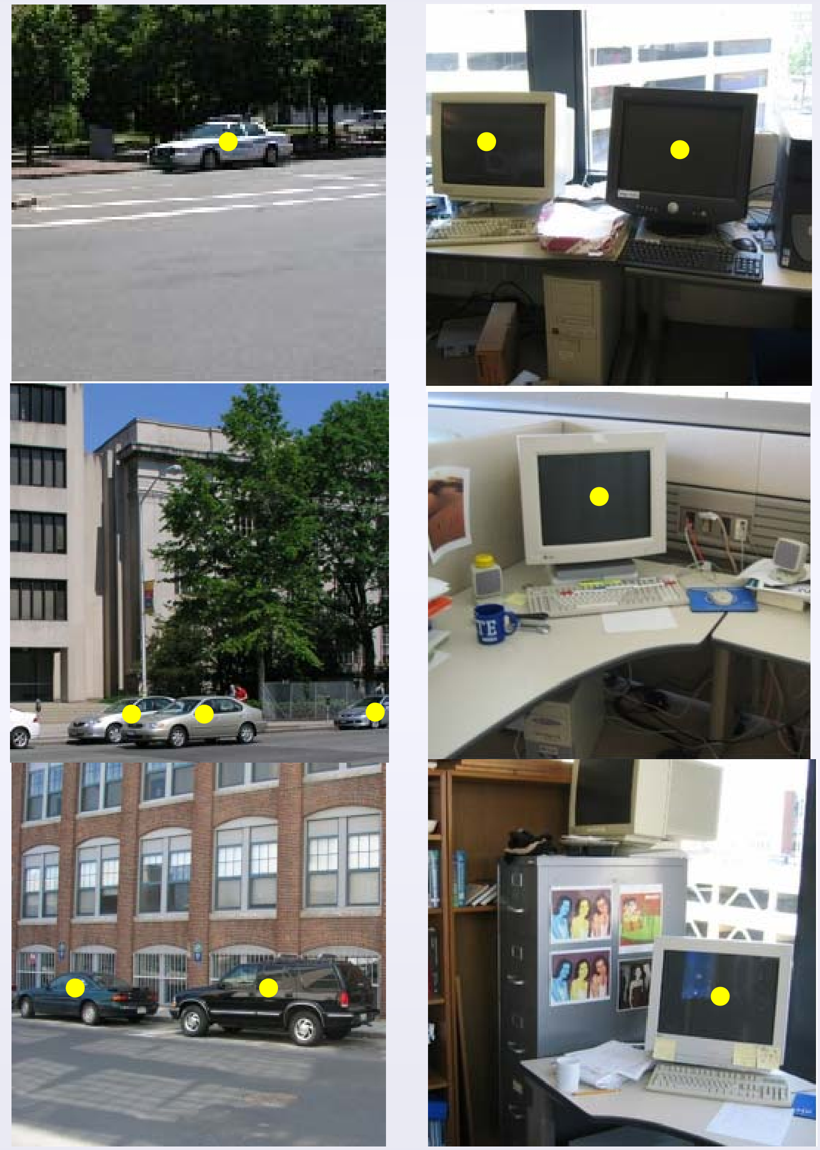


Model-Guided Segmentation
Mori, Ren, Efros, & Malik, CVPR 2004

Counting Objects & Parts

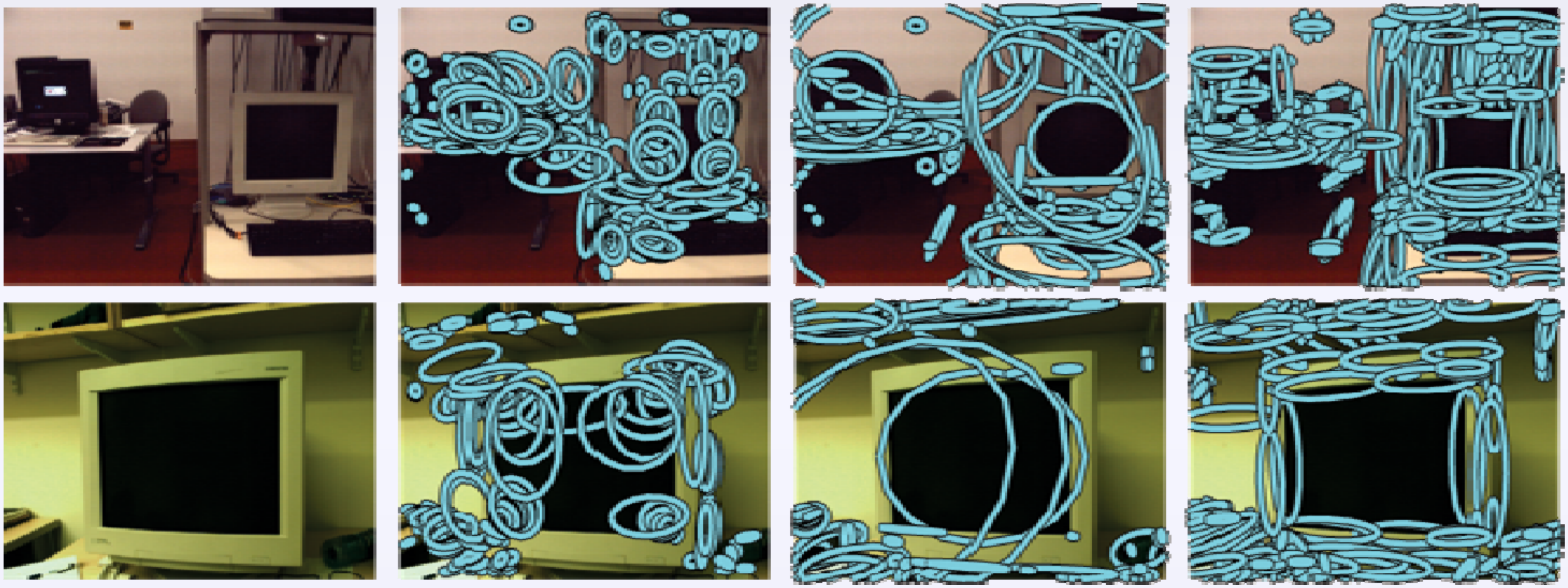


How many parts?



How many objects?

From Images to Features



Affinely Adapted Harris Corners

Maximally Stable Extremal Regions

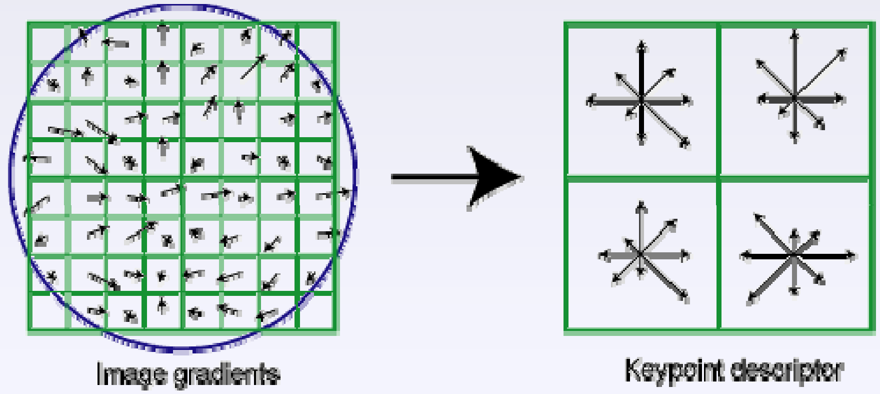
Linked Sequences of Canny Edges

- Some invariance to lighting & pose variations
- Dense, multiscale, over-segmentation of image

A Discrete Feature Vocabulary

SIFT Descriptors

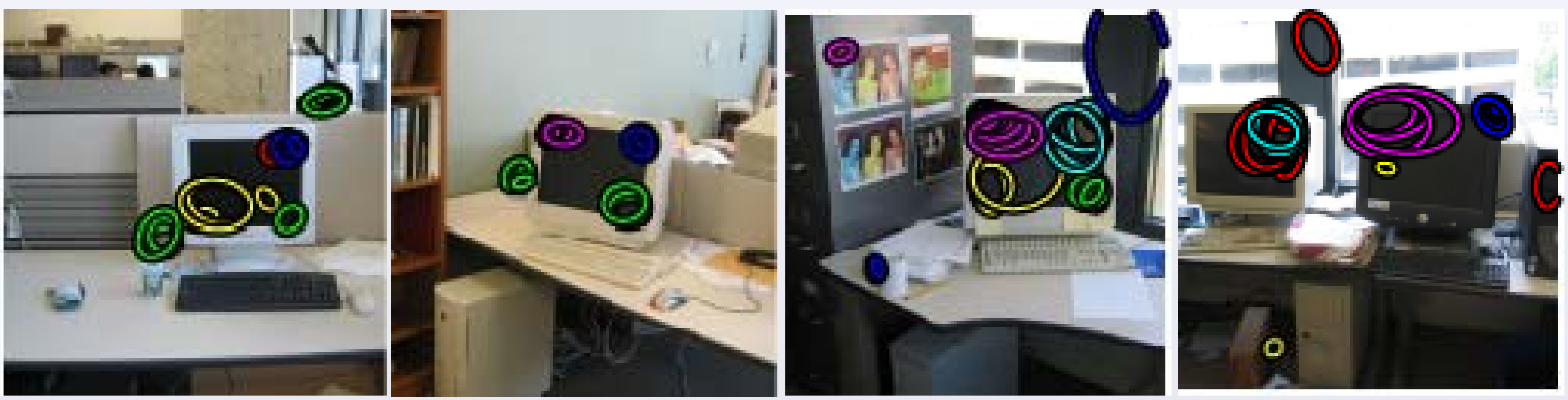
- Normalized histograms of orientation energy
- Compute ~1,000 word dictionary via K-means
- Map each feature to nearest *visual word*



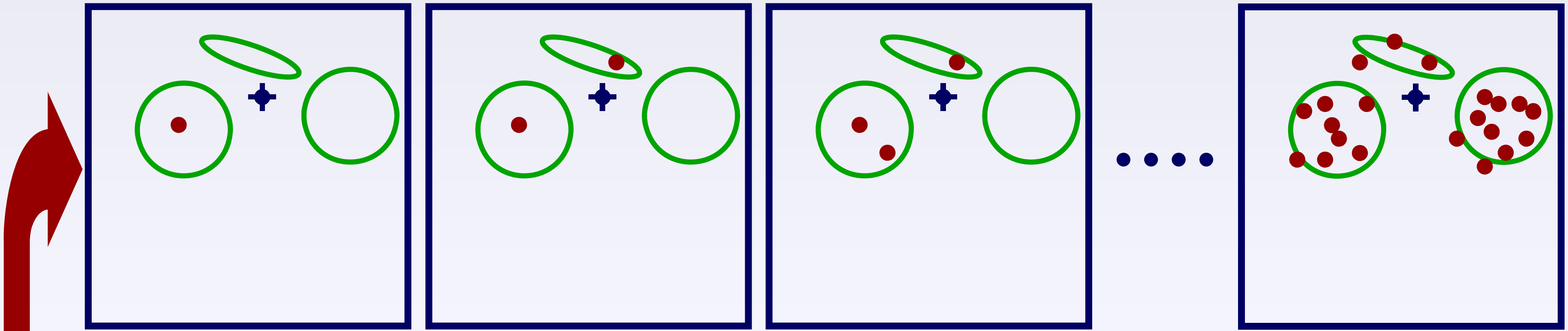
Lowe, IJCV 2004

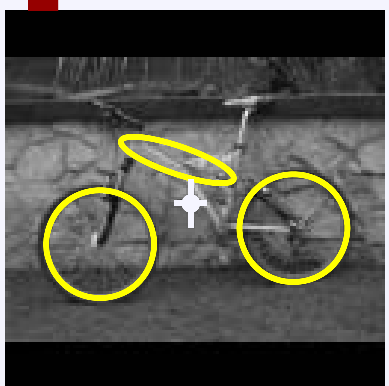
$w_{ji} \longrightarrow$ appearance of feature i in image j

$v_{ji} \longrightarrow$ 2D position of feature i in image j



Generative Model for Objects

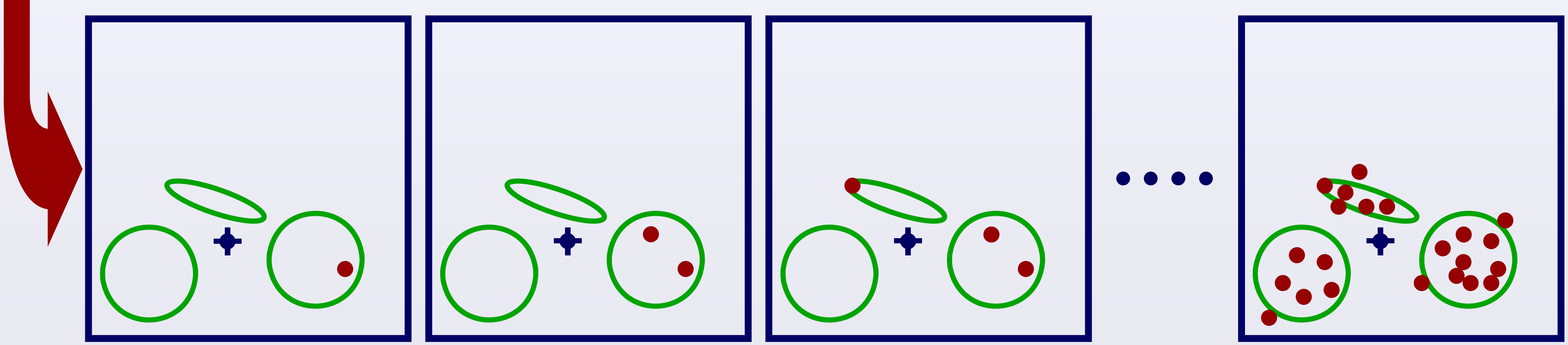




For each image: Sample a reference position

For each feature:

- Randomly choose one part
- Sample from that part's feature distribution



Objects as Mixture Models

- For a fixed reference position, our generative model is equivalent to a finite mixture model:

$$p(w_{ji}, v_{ji} \mid \rho_j) = \sum_{k=1}^{K} \pi_k \underbrace{\eta_k(w_{ji})}_{\text{Pr(appearance | part)}} \underbrace{\mathcal{N}(v_{ji}; \mu_k + \rho_j, \Lambda_k)}_{\text{Pr(position | part)}}$$

Feature appearance

Feature position

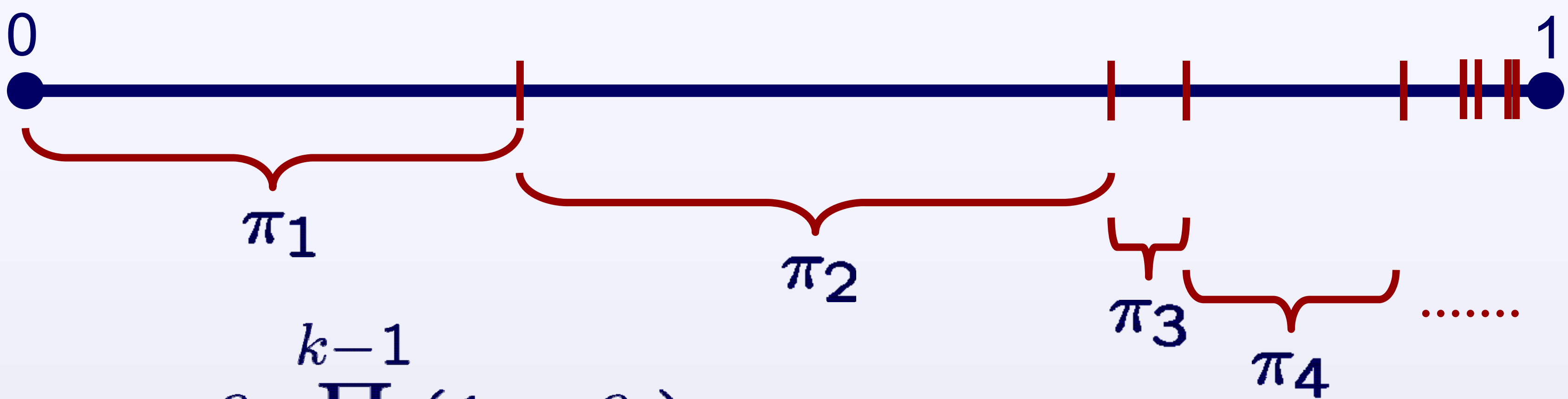
Pr(part)

- How many parts should we choose?
  - ➢ Too few reduces model accuracy
  - ➢ Too many causes overfitting & poor generalization

Dirichlet Process Mixtures

$$p(x) = \sum_{k=1}^{\infty} \pi_k f\left(x \mid \theta_k\right)$$

- *Dirichlet processes* define a prior distribution on weights assigned to mixture components:



$$\pi_k = \beta_k \prod_{\ell=1}^{k-1} (1 - \beta_\ell)$$

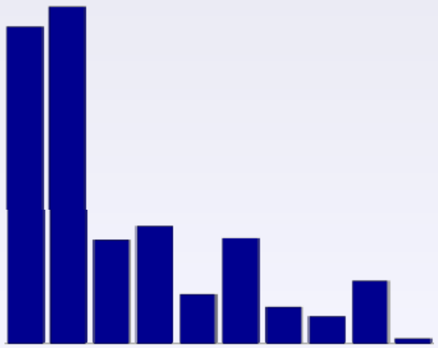
$$\beta_k \sim \text{Beta}(1, \alpha)$$

$\alpha \longrightarrow$ concentration parameter

Stick-Breaking Construction: *Sethuraman, 1994*

Why the Dirichlet Process?

$$p(x) = \sum_{k=1}^{\infty} \pi_k f\left(x \mid \theta_k\right)$$



Nonparametric $\neq$ No Parameters

- Model complexity grows as data observed:
  - Small training sets give *simple, robust* predictions
  - Reduced sensitivity to prior assumptions

Flexible but Tractable

- Literature showing attractive *asymptotic properties*
- Leads to simple, effective *computational methods*
  - Avoids challenging model selection issues

Objects as Distributions

$$p(w_{ji}, v_{ji} | \rho_j) = \sum_{k=1}^{\infty} \pi_k \underbrace{\eta_k(w_{ji})}_{\text{Pr(appearance | part)}} \underbrace{\mathcal{N}(v_{ji}; \mu_k + \rho_j, \Lambda_k)}_{\text{Pr(position | part)}}$$

Feature appearance

Feature position

- Parts are defined by *parameters*, which encode distributions on visual features:

$$\theta_k = \{\eta_k, \mu_k, \Lambda_k\}$$

- Objects are defined by *distributions* on the infinitely many potential part parameters:

$$G(\theta) = \sum_{k=1}^{\infty} \pi_k \delta(\theta, \theta_k) \qquad \pi \sim \text{Stick}(\alpha)$$

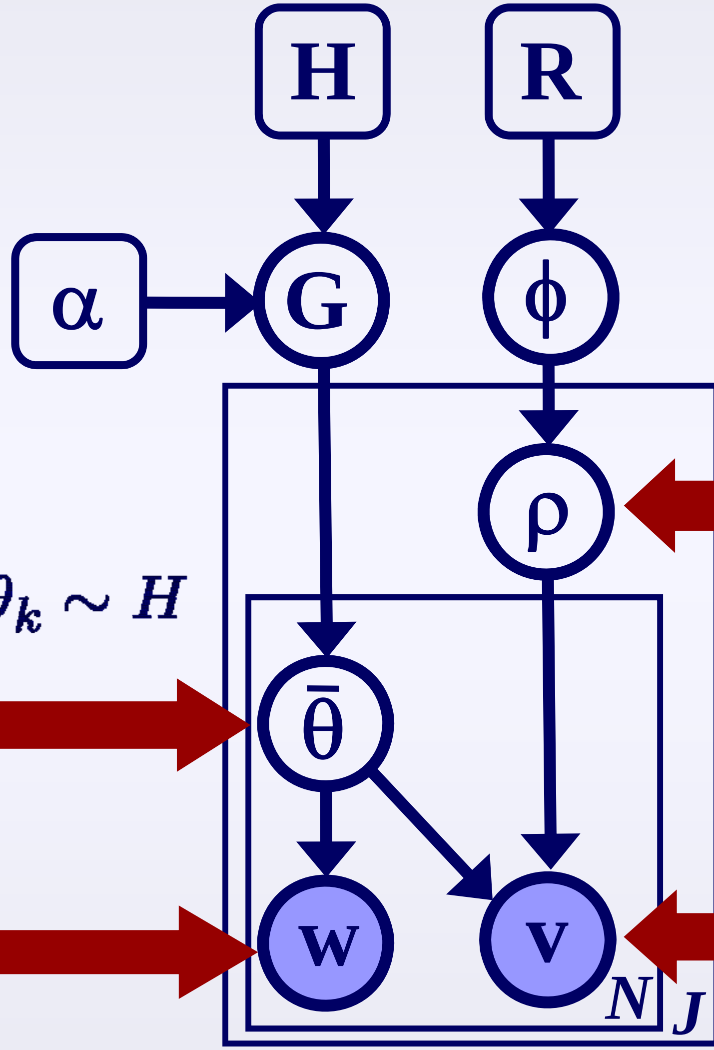
Dirichlet Process Object Model

Part-based object model sampled from DP prior:

$$G \sim \mathrm{DP}(\alpha, H)$$

$$G(\theta) = \sum_{k=1}^{\infty} \pi_k \delta(\theta, \theta_k)$$

$$\pi \sim \mathrm{Stick}(\alpha) \qquad \theta_k \sim H$$



For each of J images, sample a *reference position:*

$$\rho_j \sim \mathcal{N}(\rho; \phi)$$

For each of N features, sample *part parameters:*

$$\bar{\theta}_{ji} \sim G(\theta)$$

Sample multinomial *feature appearance:*

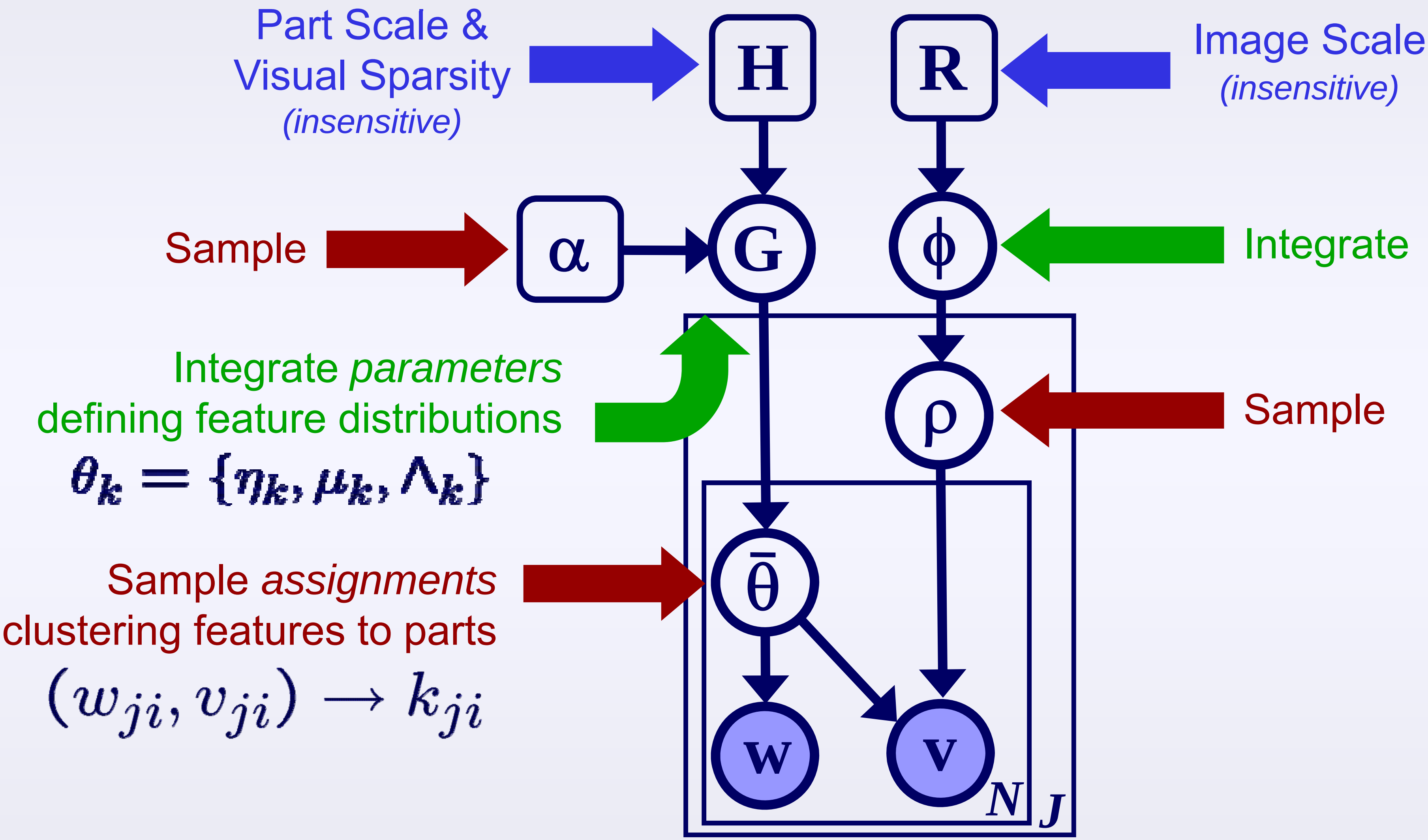
$$w_{ji} \sim \bar{\eta}_{ji}(w)$$

Sample Gaussian *feature position:*

$$v_{ji} \sim \mathcal{N}(v; \bar{\mu}_{ji} + \rho_j, \bar{\Lambda}_{ji})$$

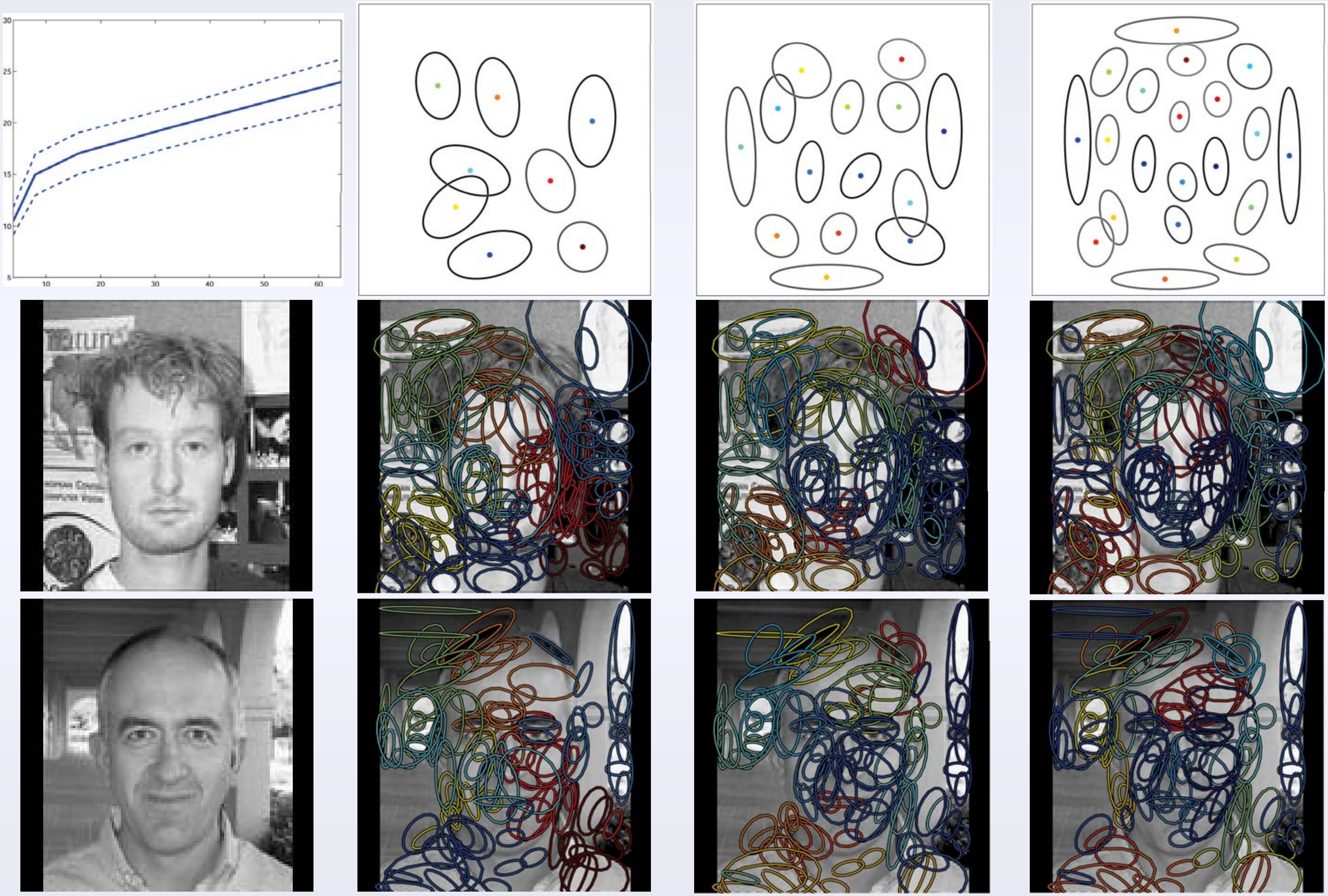
$$\bar{\theta}_{ji} = \{\bar{\eta}_{ji}, \bar{\mu}_{ji}, \bar{\Lambda}_{ji}\}$$

Learning DPs: Gibbs Sampling



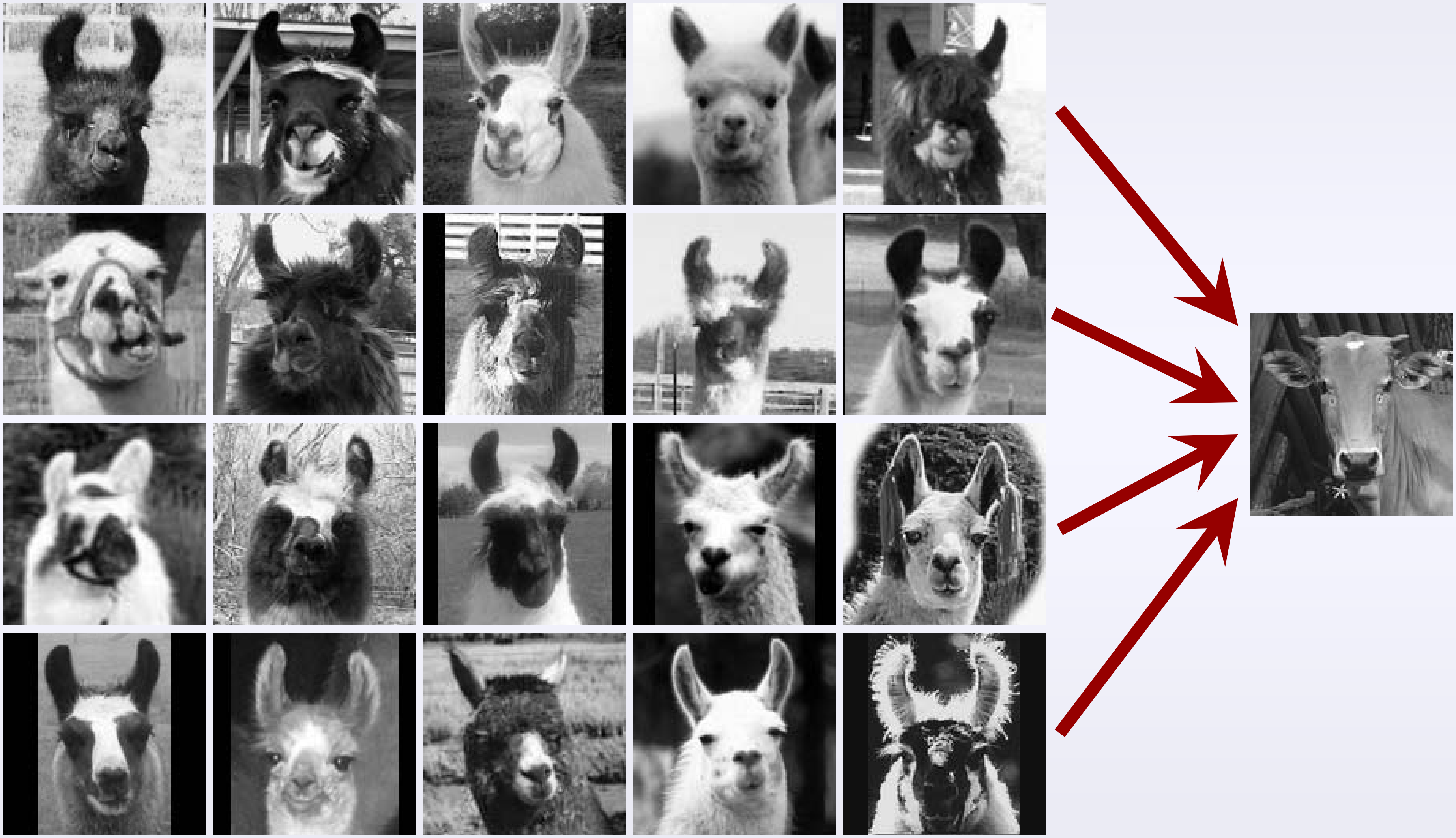
Dirichlet processes have many desirable analytic properties, which lead to efficient *Rao-Blackwellized* learning algorithms

Decomposing Faces into Parts



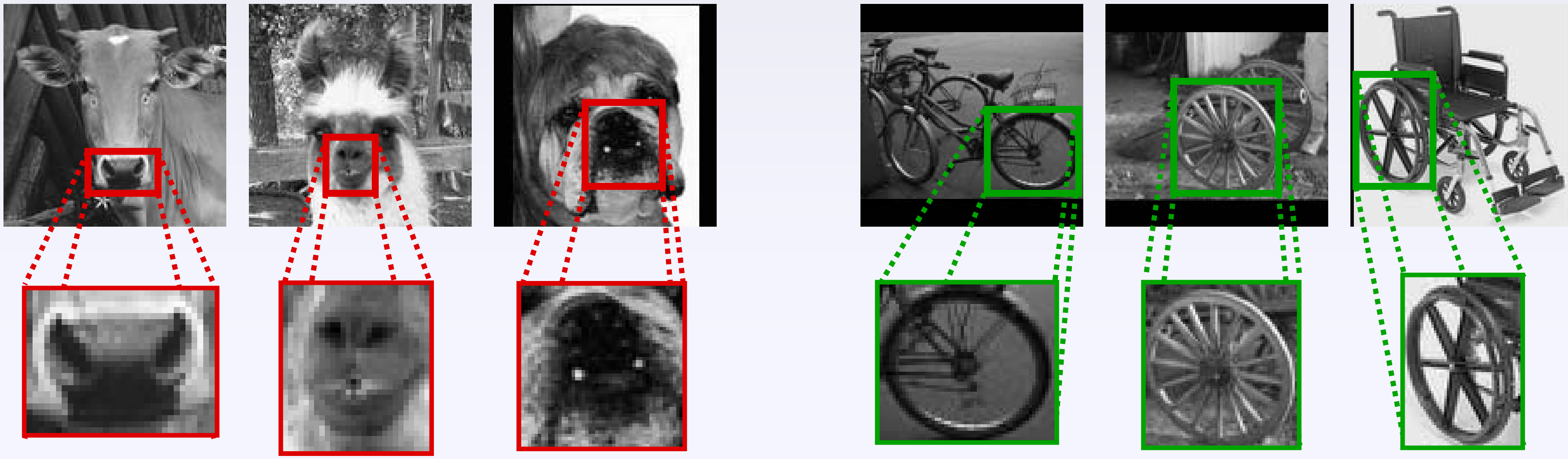
4 Images *16 Images* *64 Images*

Generalizing Across Categories



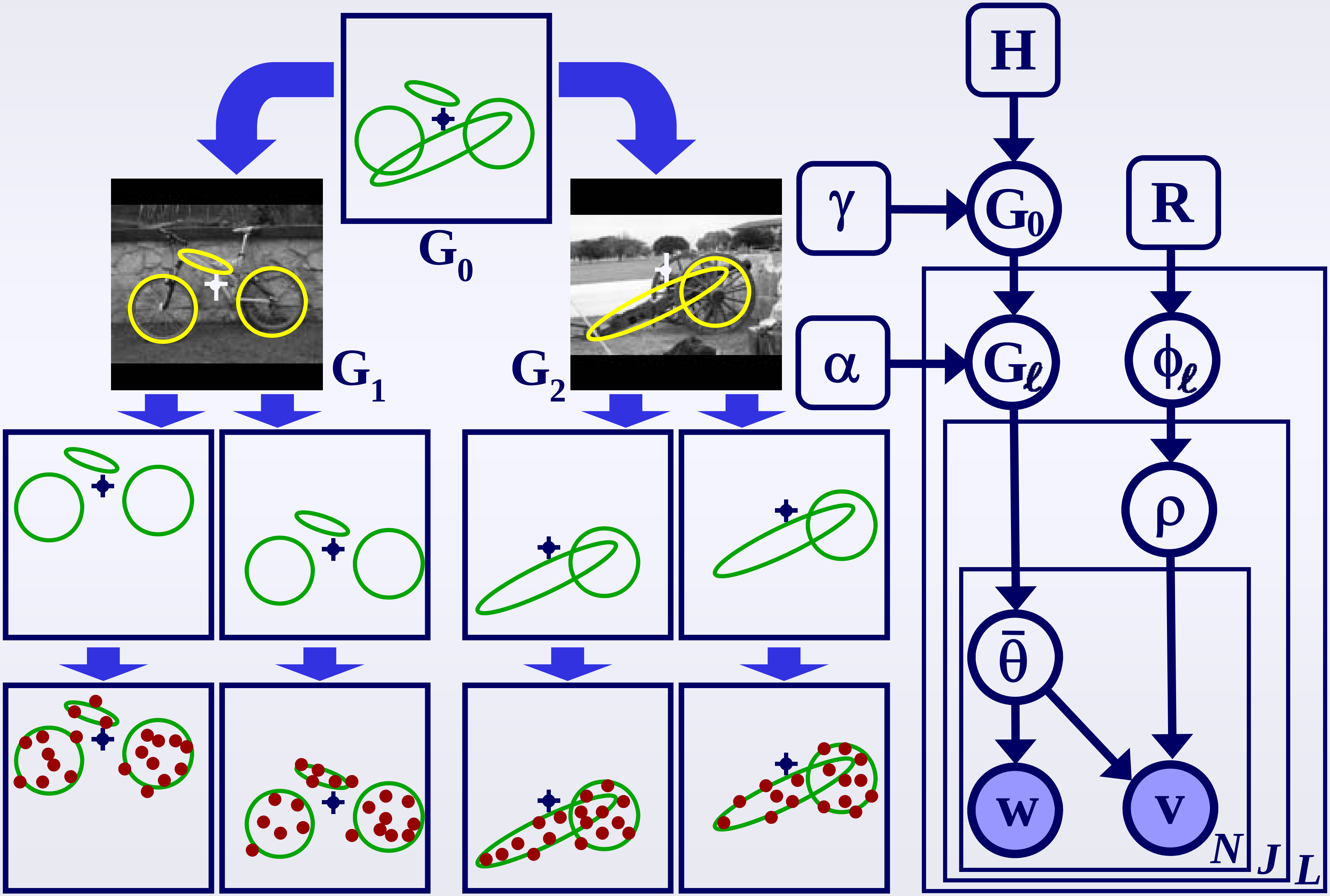
Can we transfer knowledge from one object category to another?

Learning Shared Parts

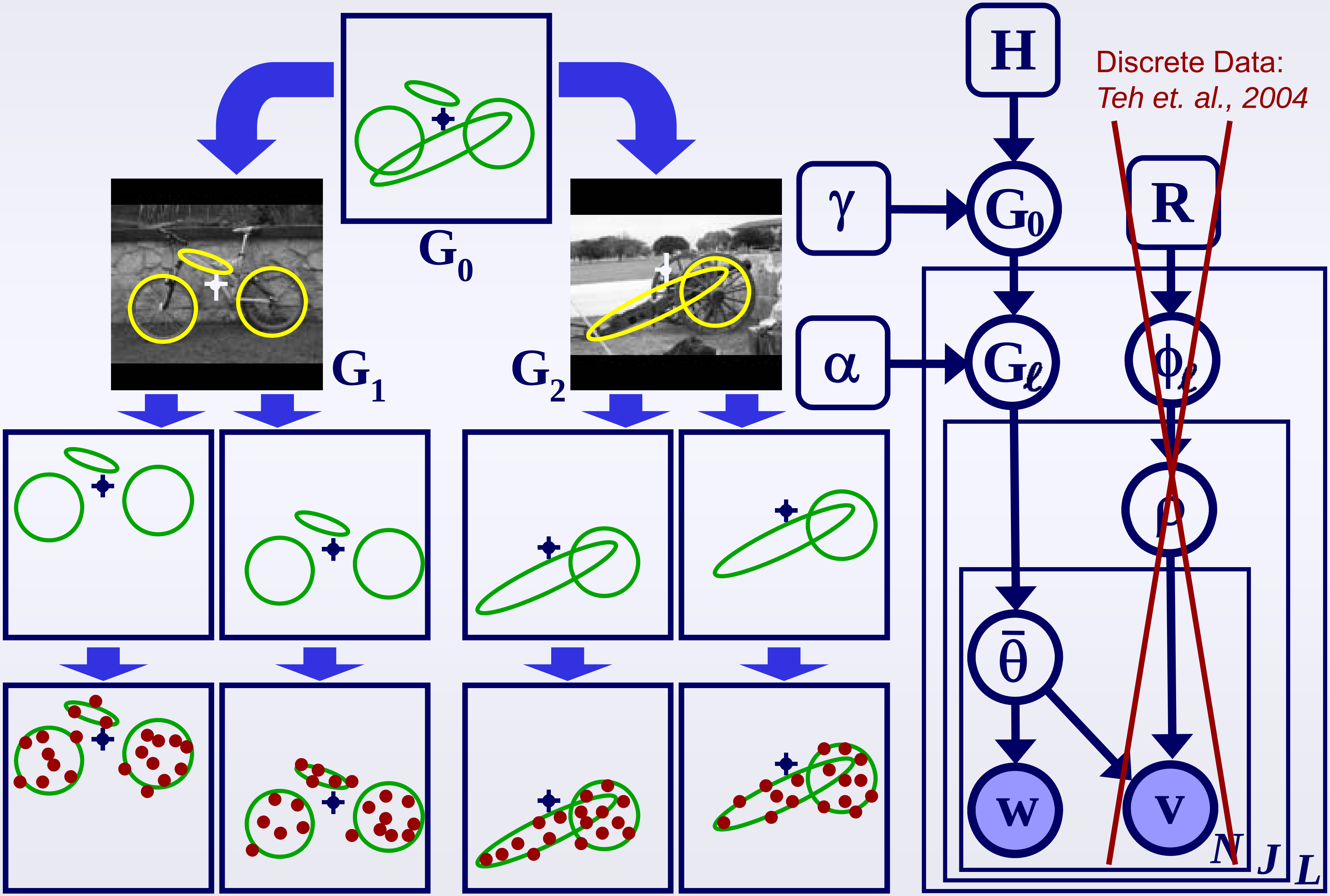


- Objects are often locally similar in appearance
- Discover *parts* shared across categories
  - ➢ How many total parts should we share?
  - ➢ How many parts should each category use?

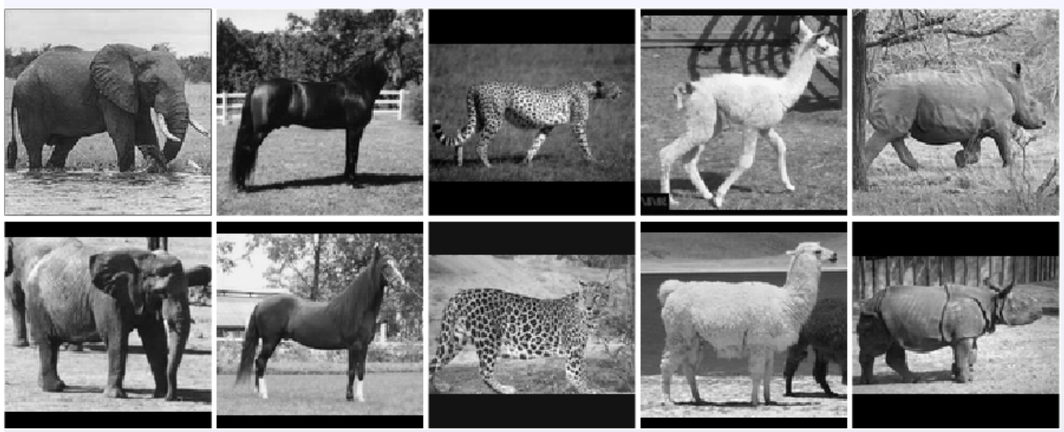
Hierarchical DP Object Model

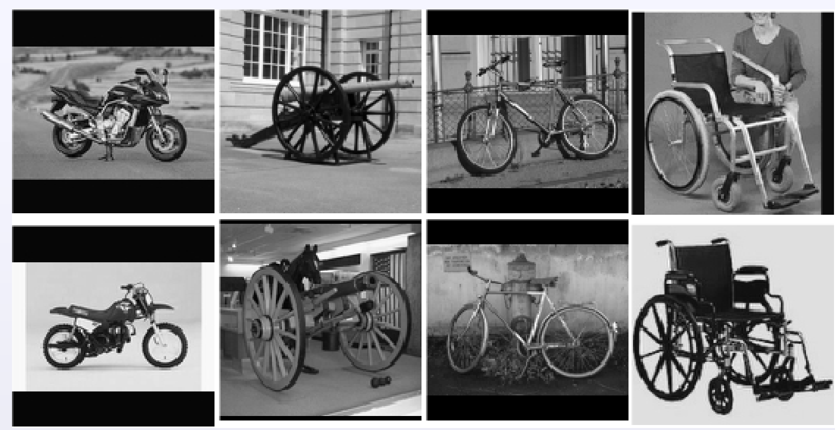


Hierarchical DP Object Model



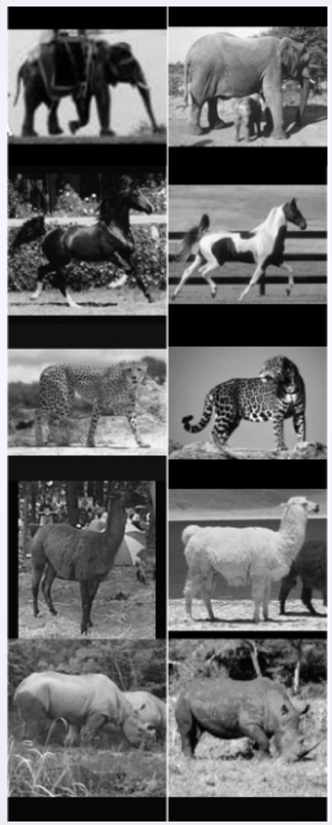
Sharing Parts: 16 Categories



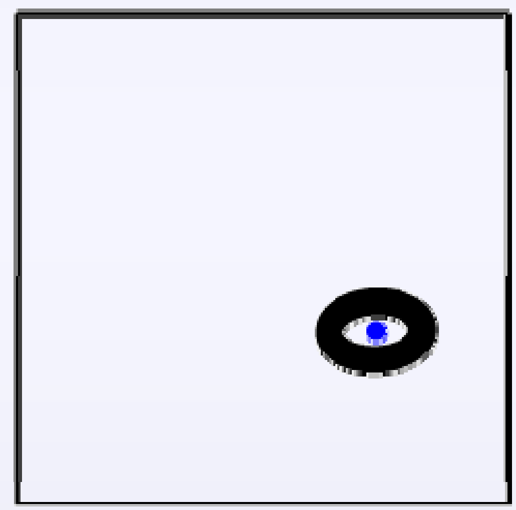


- Caltech 101 Dataset (Li & Perona)
- Horses (Borenstein & Ullman)
- Cat & dog faces (Vidal-Naquet & Ullman)
- Bikes from Graz-02 (Opelt & Pinz)
- Google…

Visualization of Shared Parts

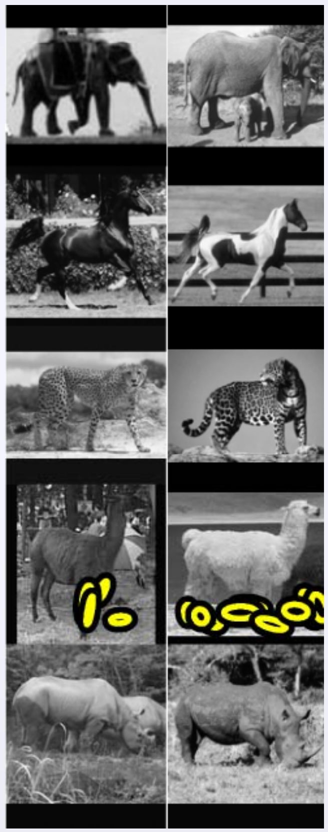


Pr(appearance | part)

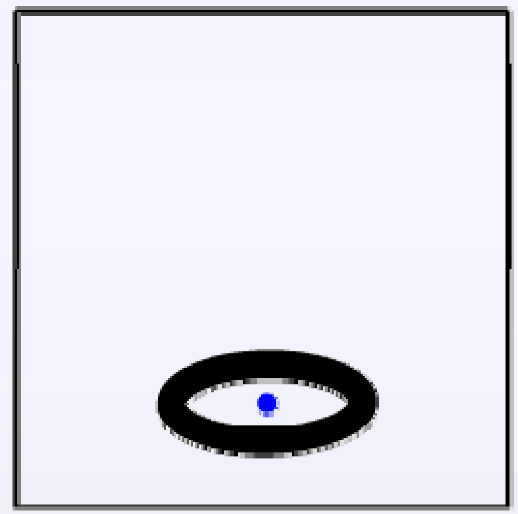


Pr(position | part)

Visualization of Shared Parts

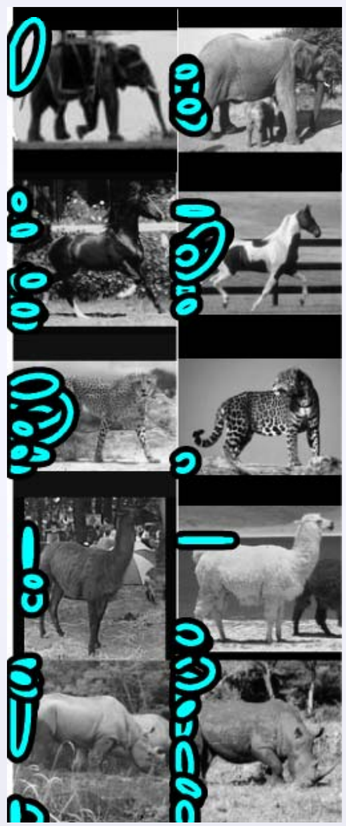


Pr(appearance | part)

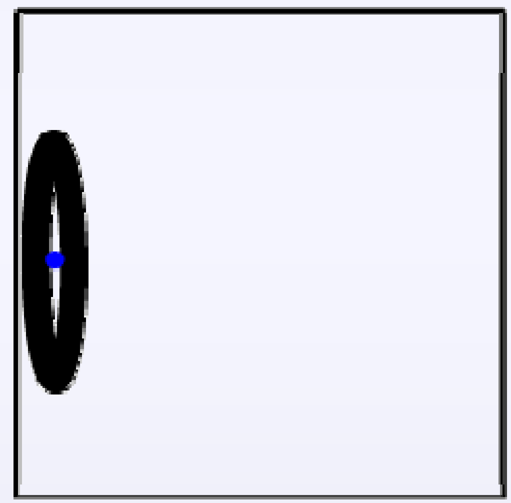


Pr(position | part)

Visualization of Shared Parts

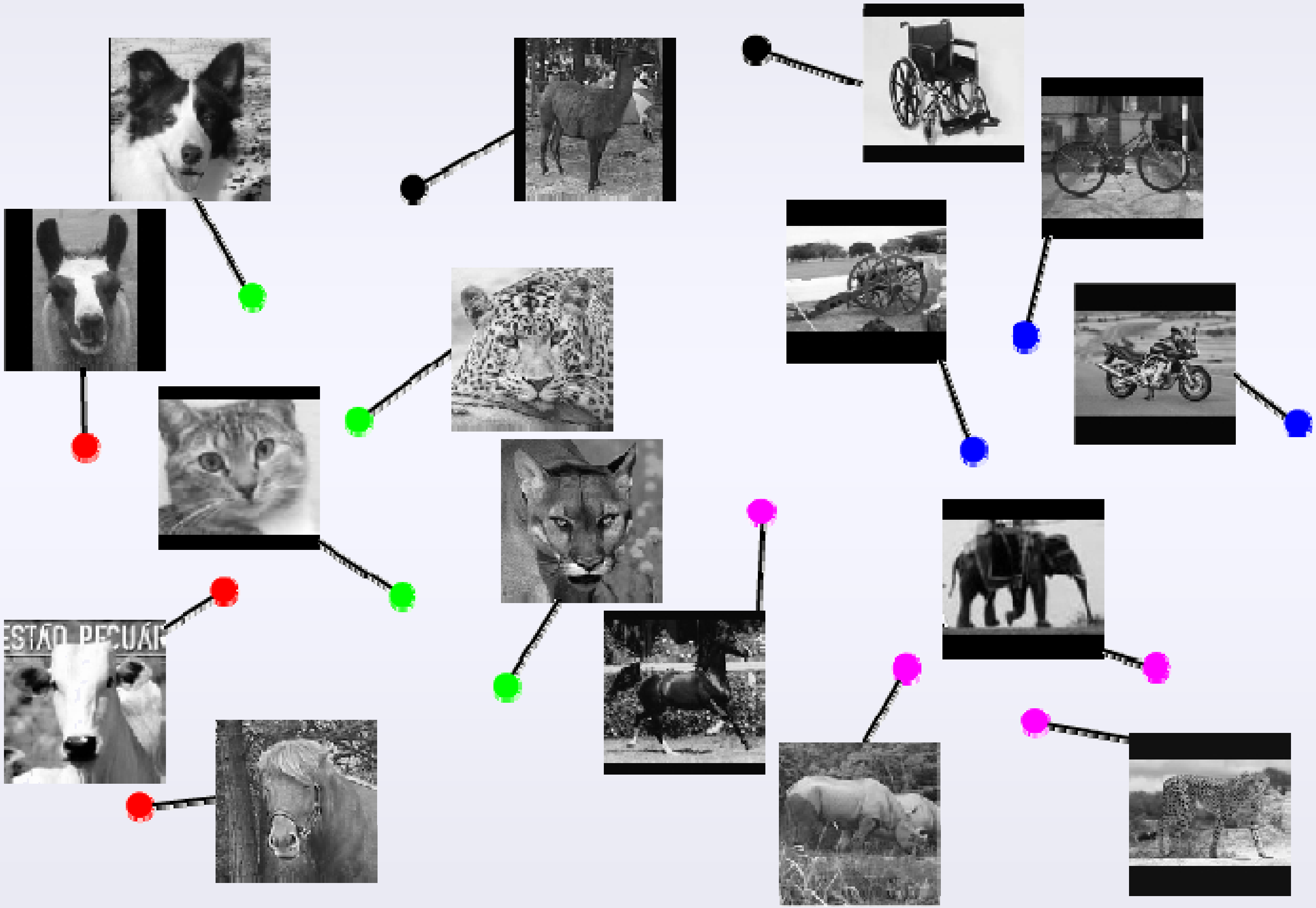


Pr(appearance | part)



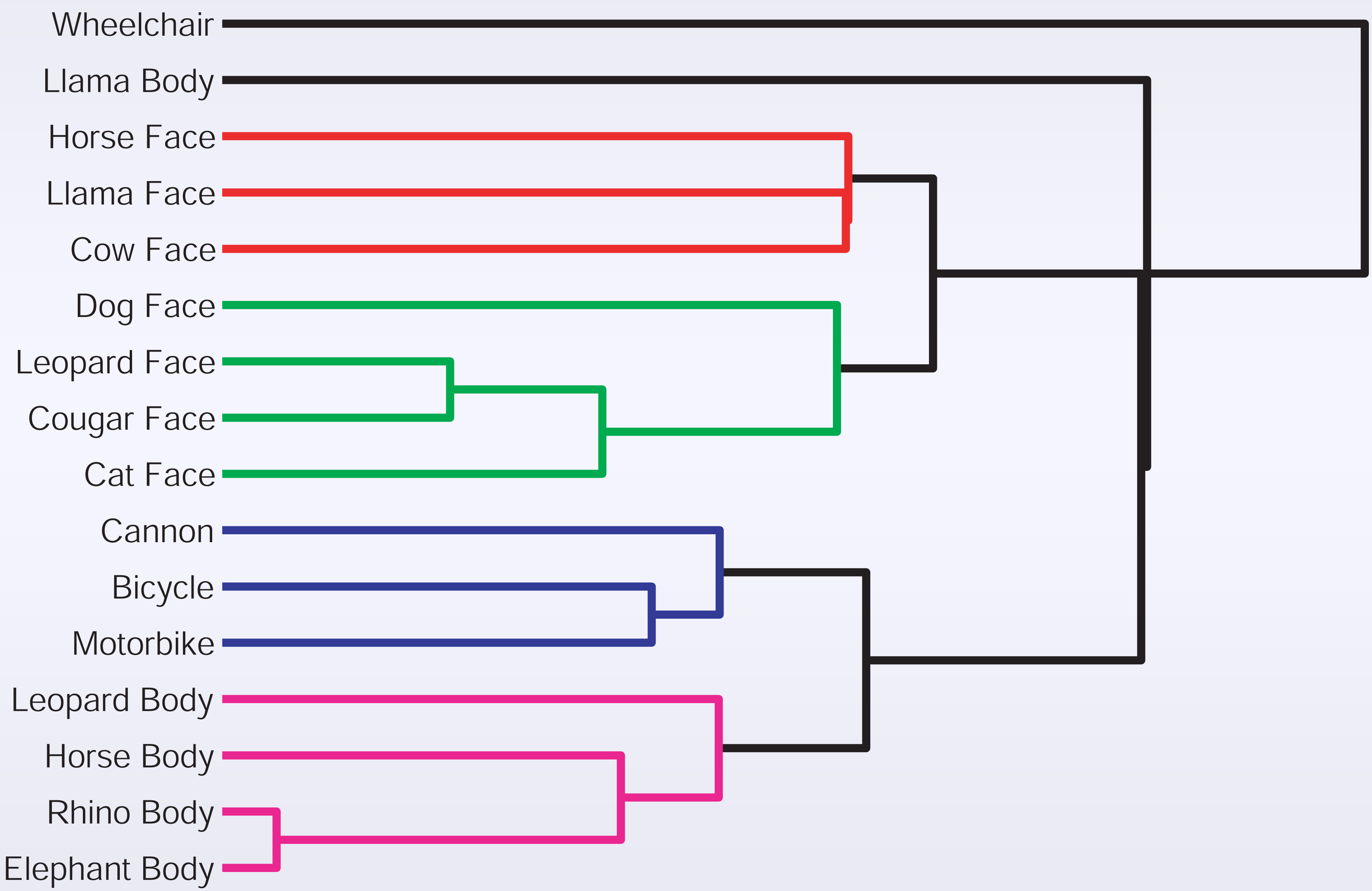
Pr(position | part)

Visualization of Part Densities



MDS Embedding of Pr(part | object)

Visualization of Part Densities

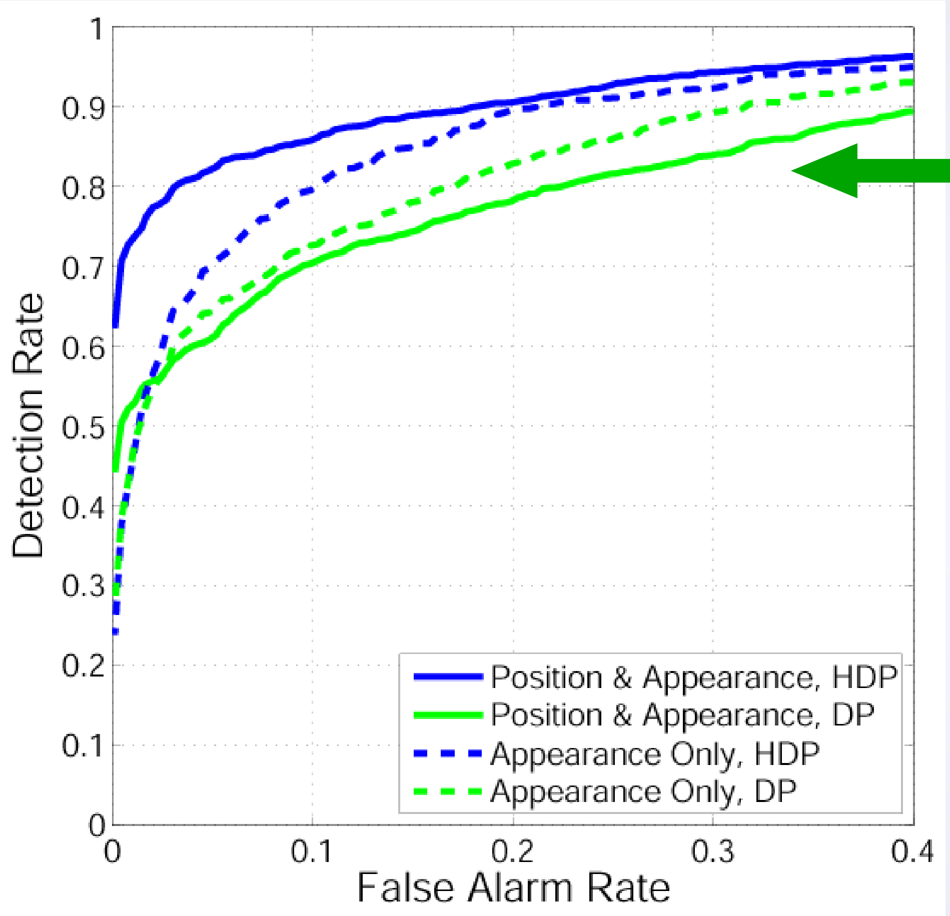


Hierarchical Clustering of Pr(part | object)

Detection Task

versus

Detection Results

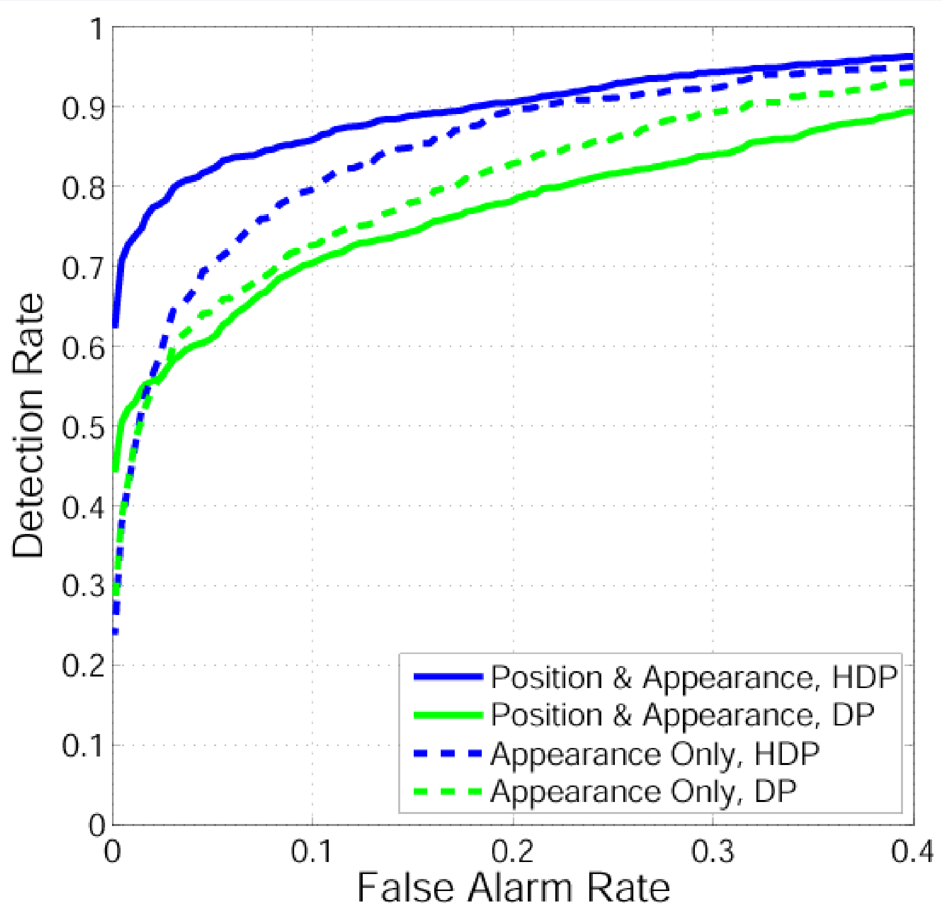


6 Training Images per Category
(ROC Curves)

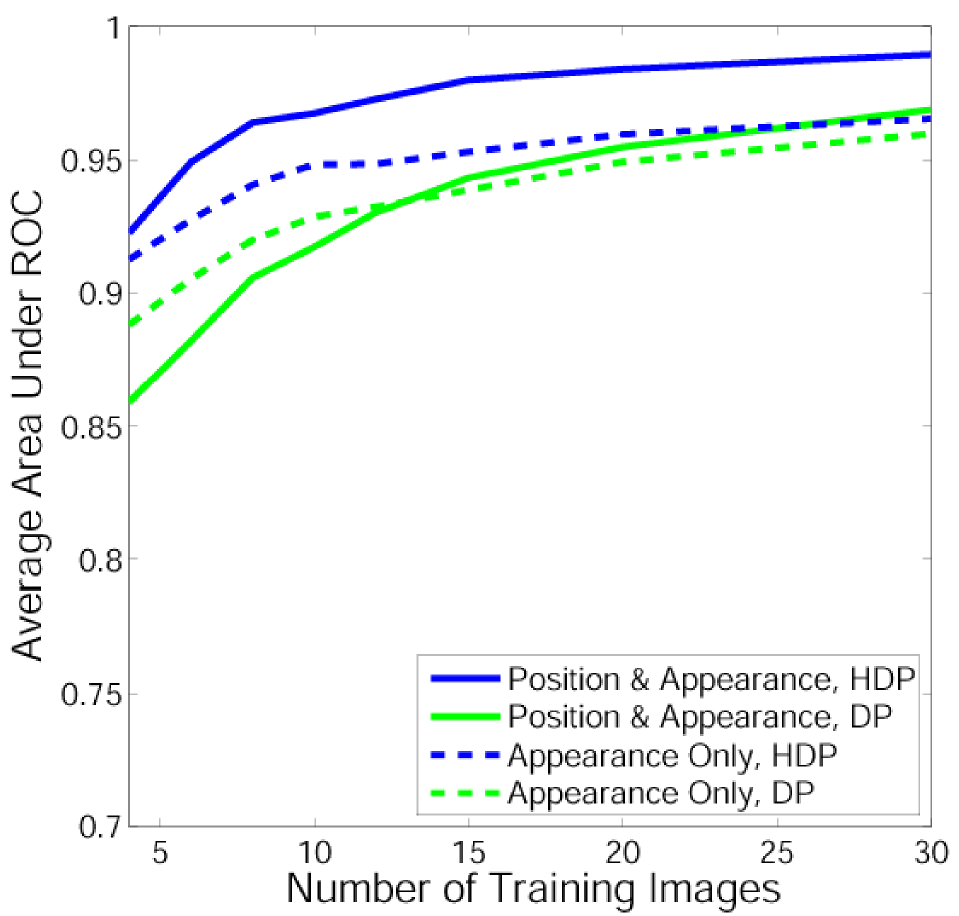
Shared Parts
more accurate than
Unshared Parts

Modeling feature positions ***improves shared*** detection, but ***hurts unshared*** detection

Detection Results

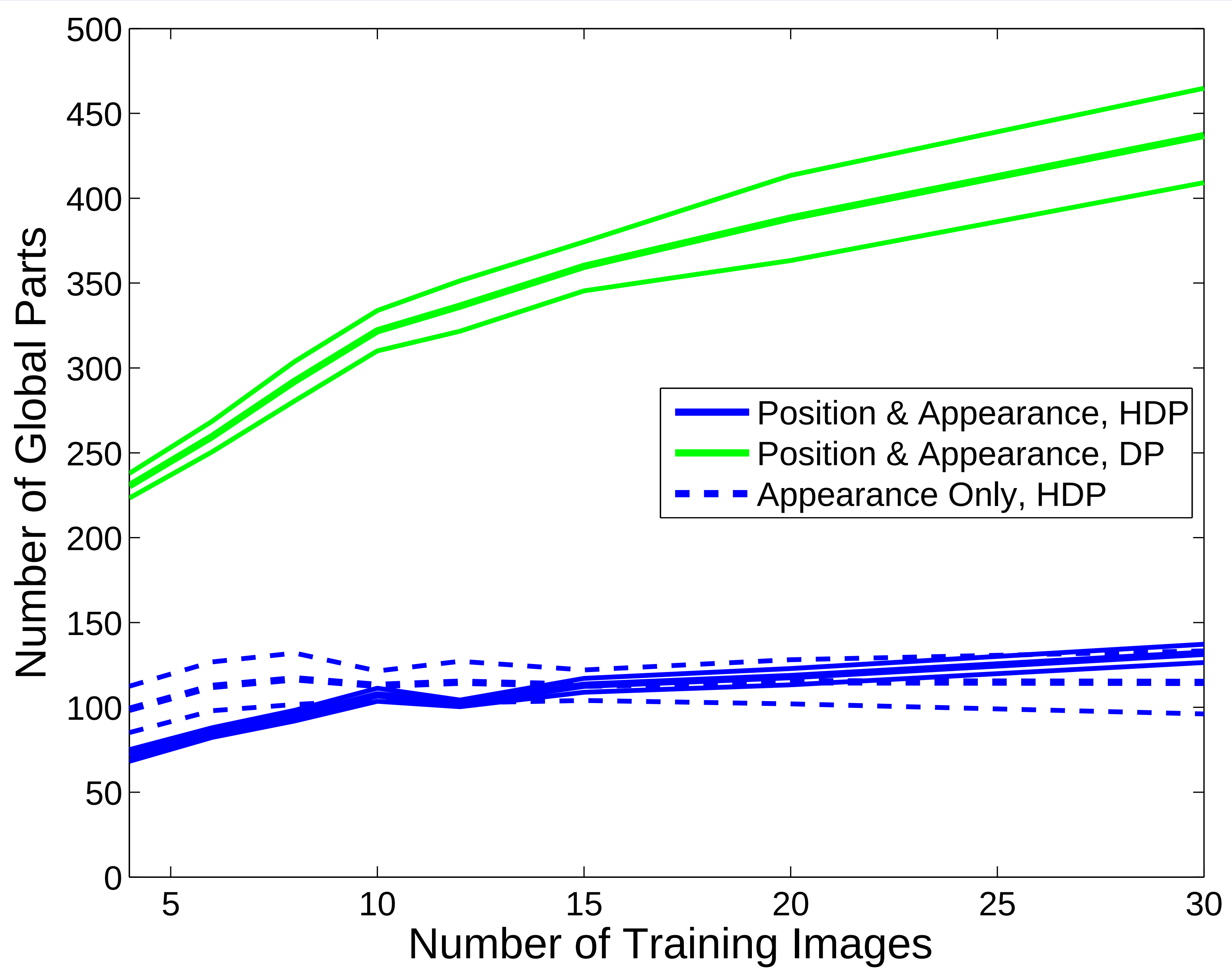


6 Training Images per Category
(ROC Curves)



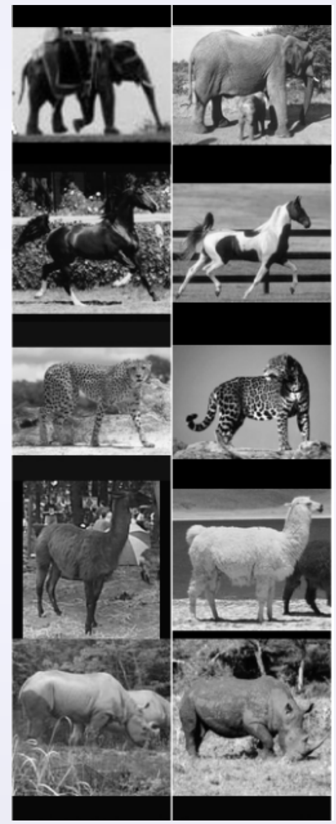
Detection vs. Training Set Size
(Area Under ROC)

Sharing Simplifies Models

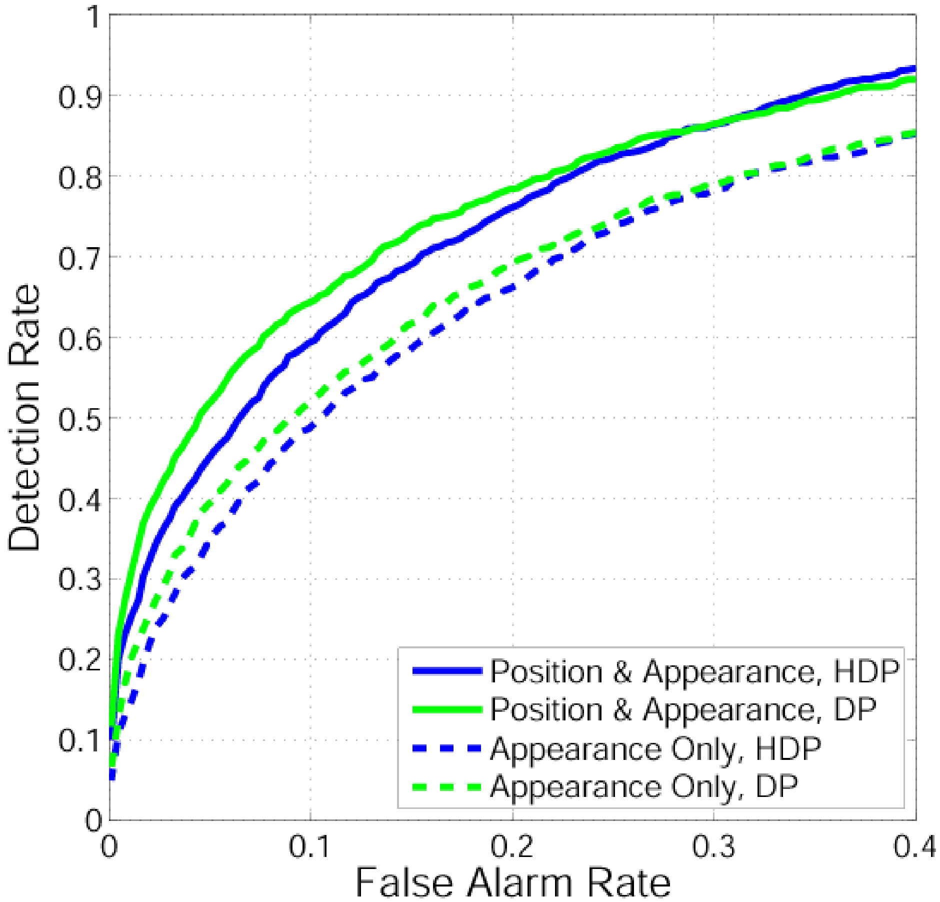


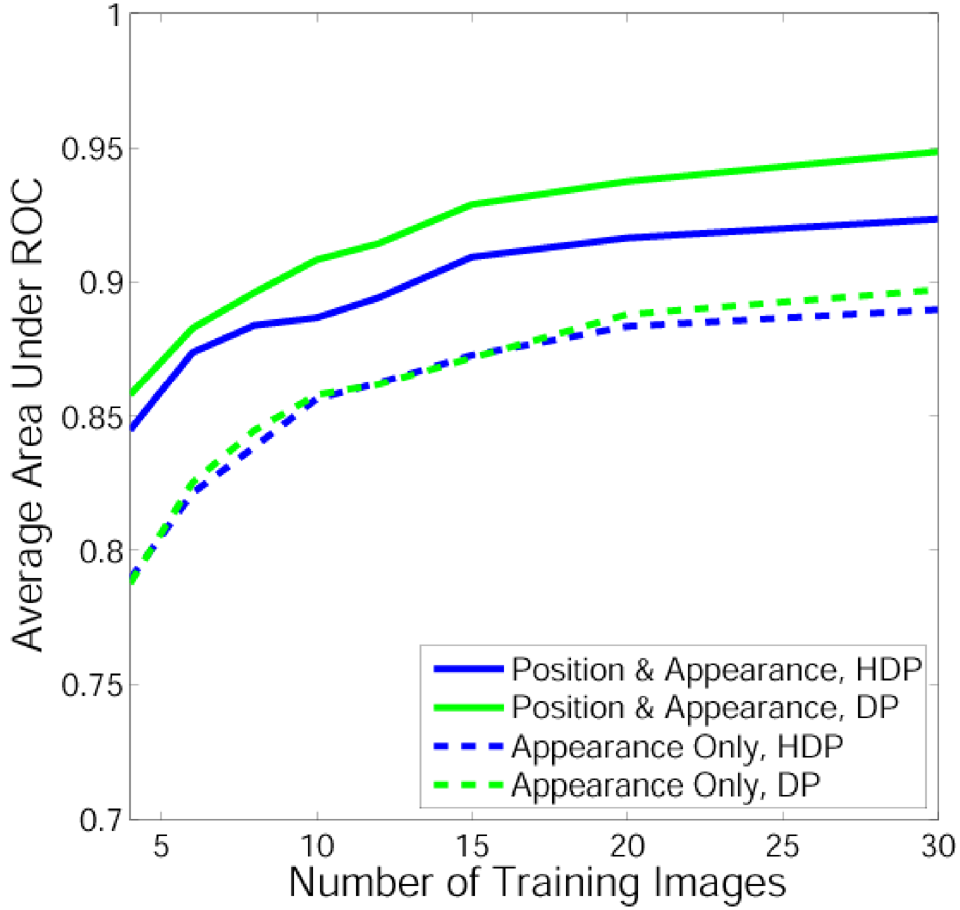
Recognition Task

versus



Recognition Results



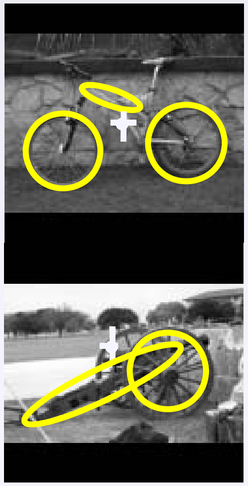


6 Training Images per Category
(ROC Curves)

Detection vs. Training Set Size
(Area Under ROC)

Outline

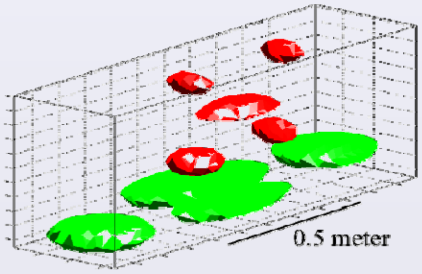
Object Recognition with Shared Parts



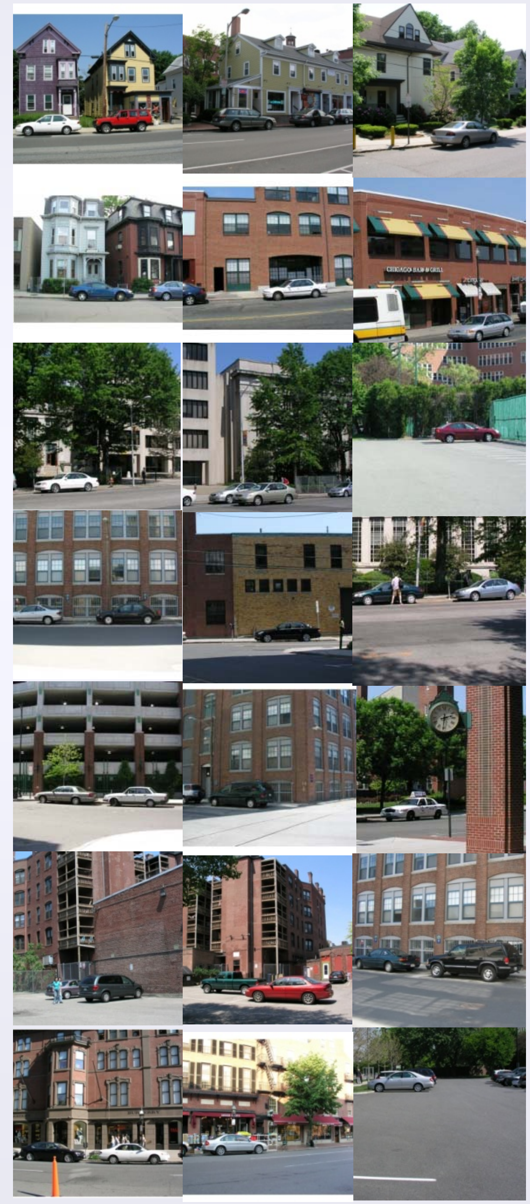
- Learning parts via Dirichlet processes
- Hierarchical DP model for 16 object categories

Multiple Object Scenes

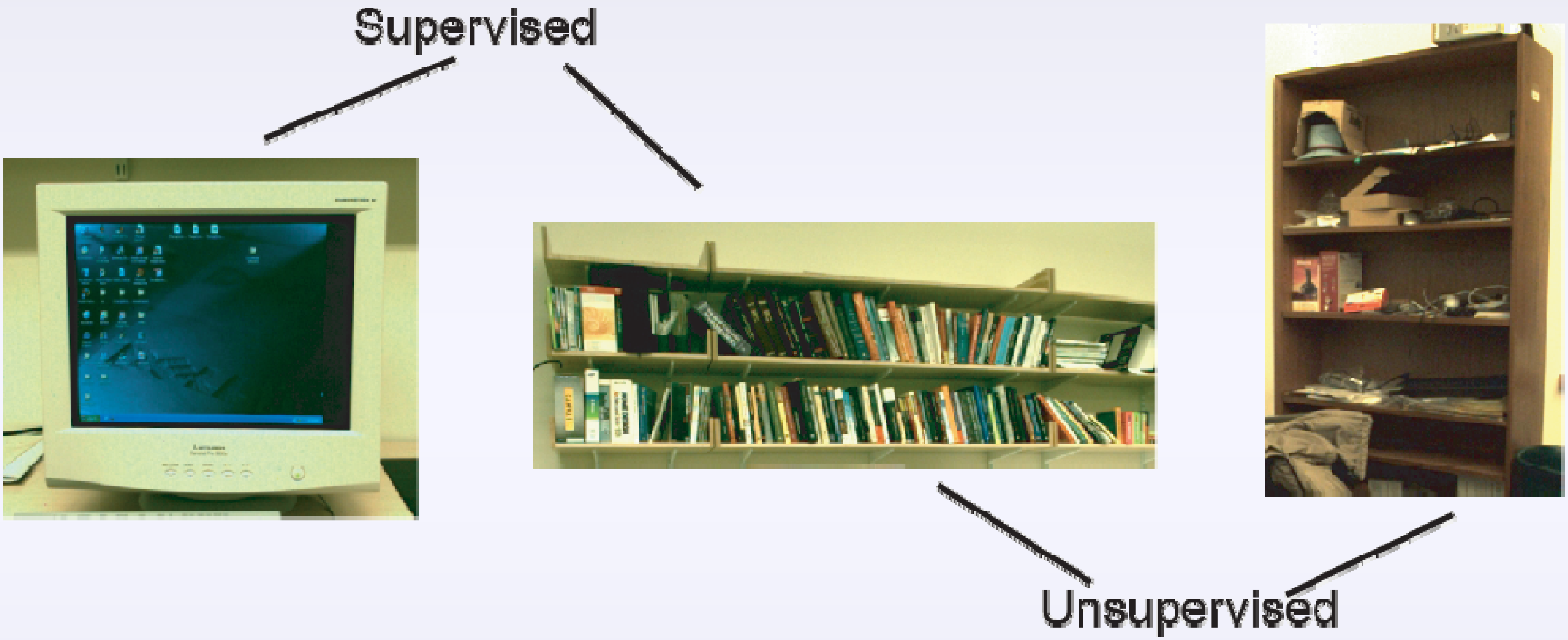
- Transformed Dirichlet processes
- Part-based models for 2D scenes
- Joint object detection & 3D reconstruction



Semi-supervised Learning

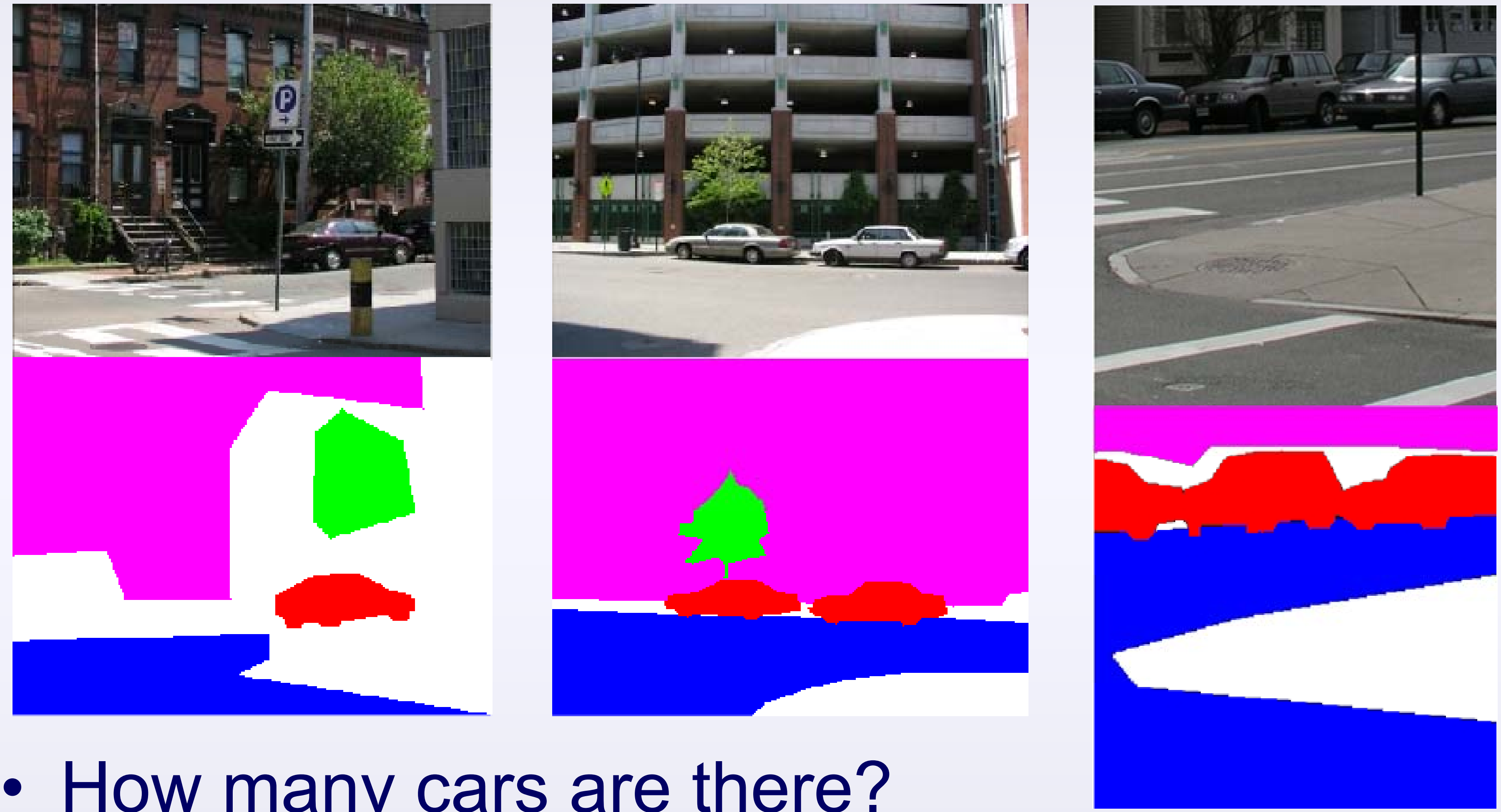


Object vs. Visual Categories



- Assume training data contains object category labels
- Discover underlying visual categories automatically

Multiple Object Scenes

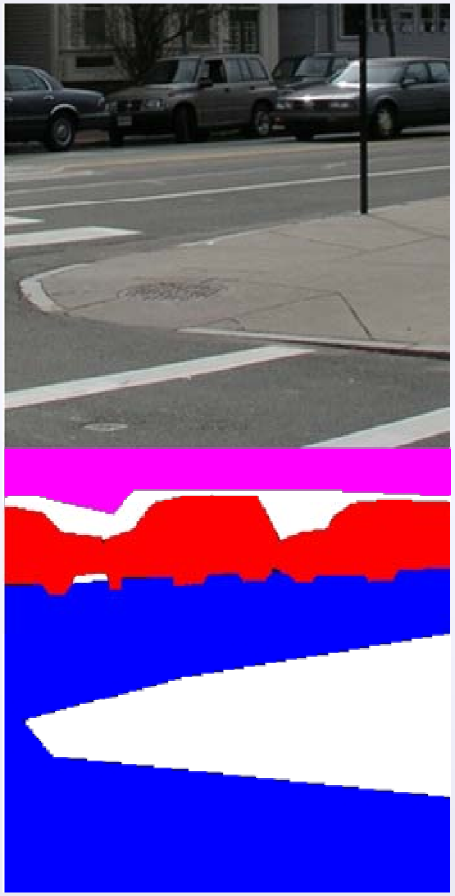


- How many cars are there?
- Where are those cars in the scene?

Standard dependent Dirichlet process models (Gelfand et. al., 2005) inappropriate

Spatial Transformations

- Let global DP clusters model objects in a *canonical* coordinate frame
- Generate images via a random *set of transformations:*



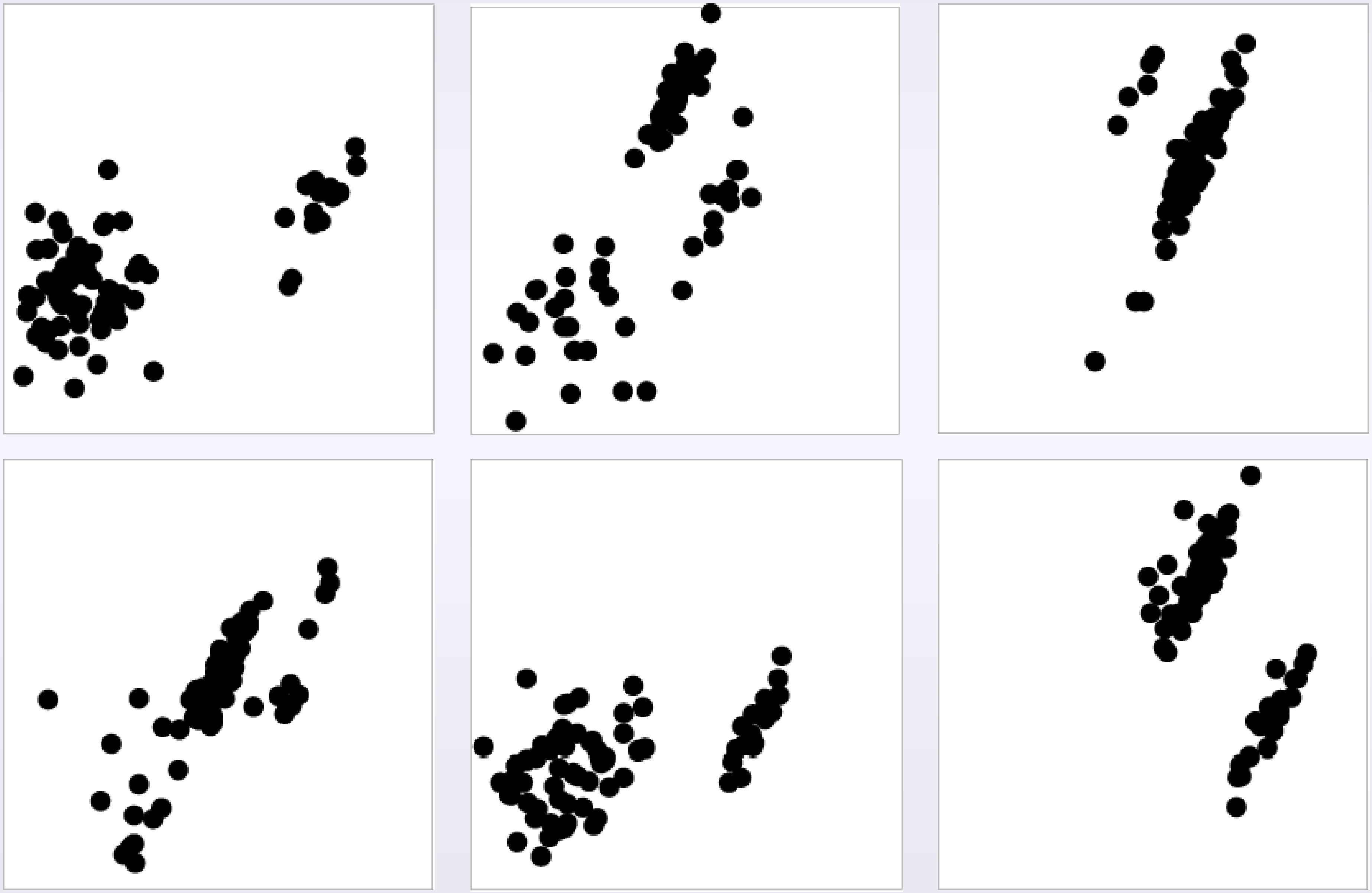
$$\tau((\mu, \Lambda); \rho) = (\mu + \rho, \Lambda)$$

Parameterized family of transformations

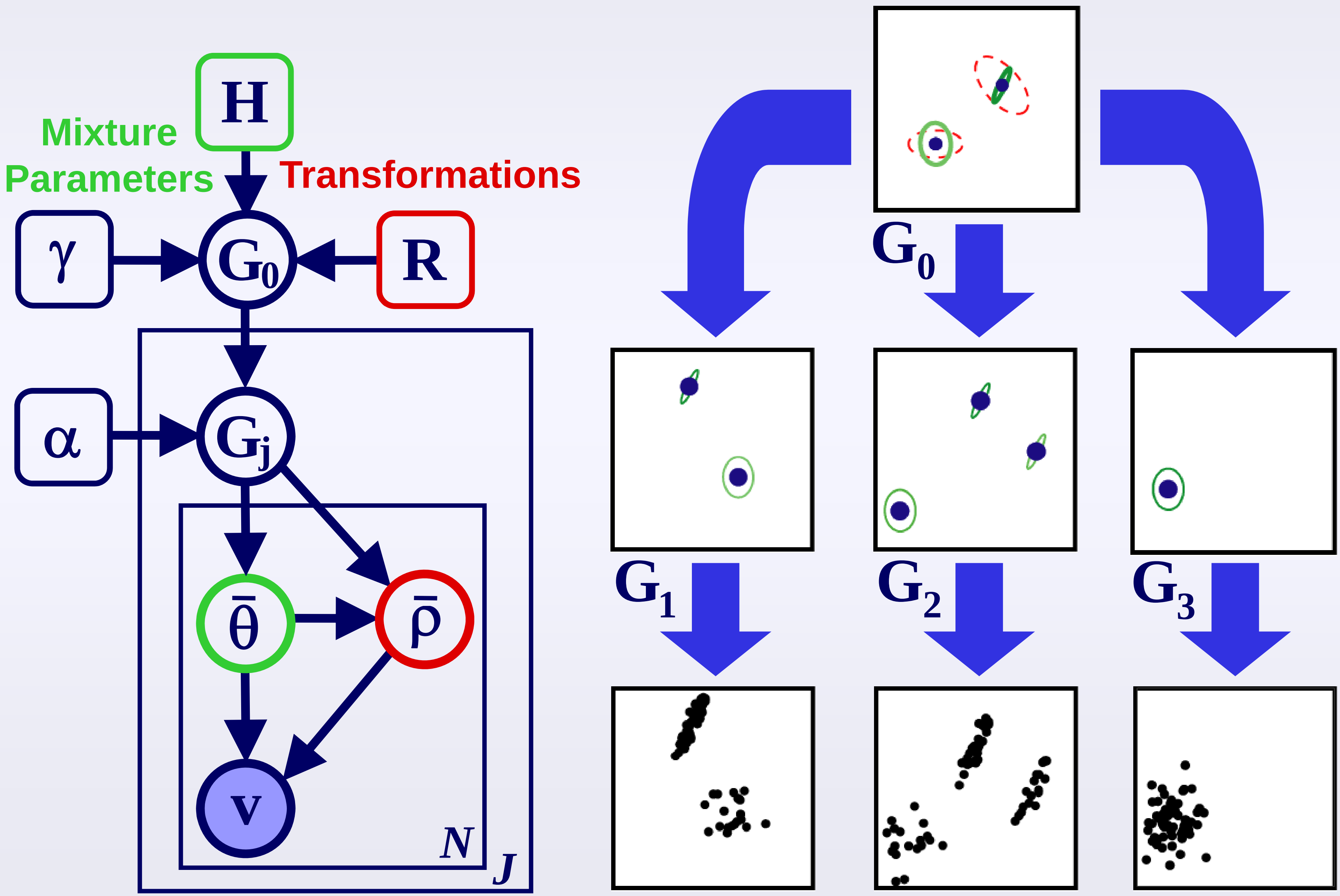
Shift cluster from canonical coordinate frame to object location in a given image

Layered Motion Models *(Wang & Adelson, Jojic & Frey)*
Nonparametric Transformation Densities *(Learned-Miller & Viola)*

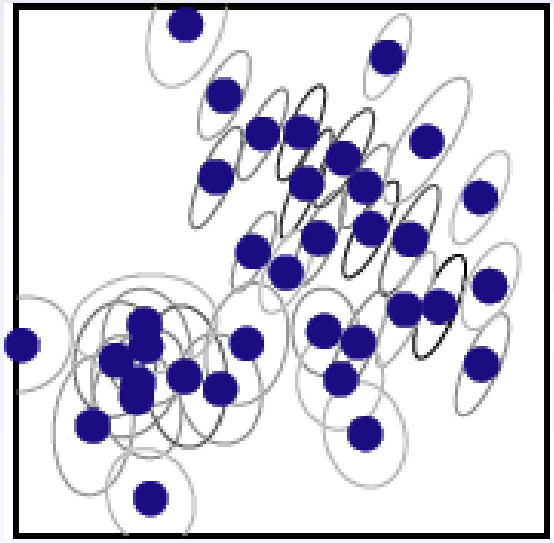
A Toy World: Bars & Blobs



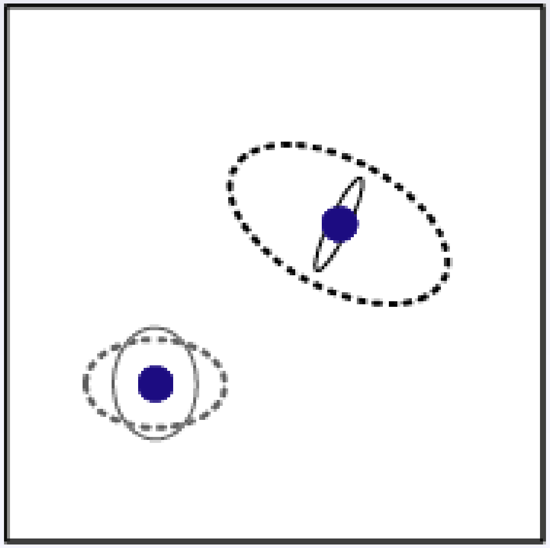
Transformed Dirichlet Process



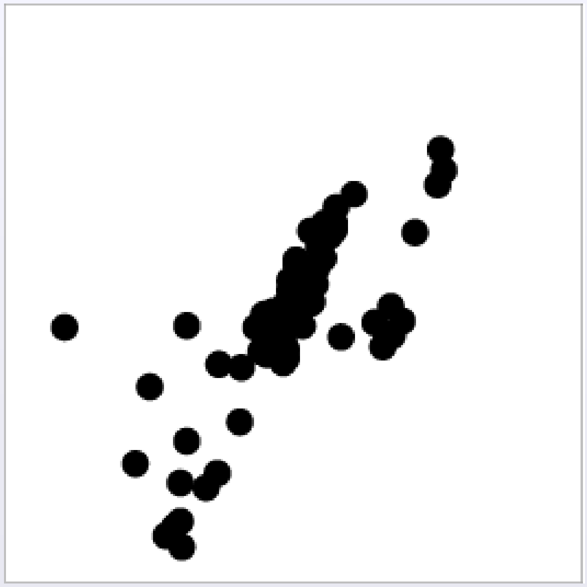
Importance of Transformations

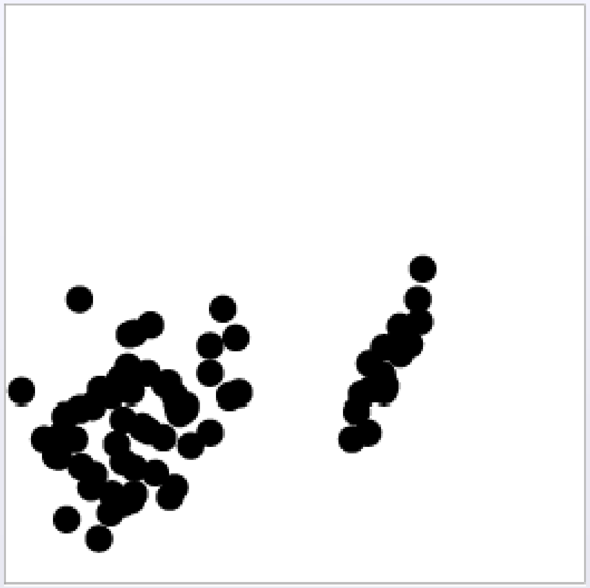


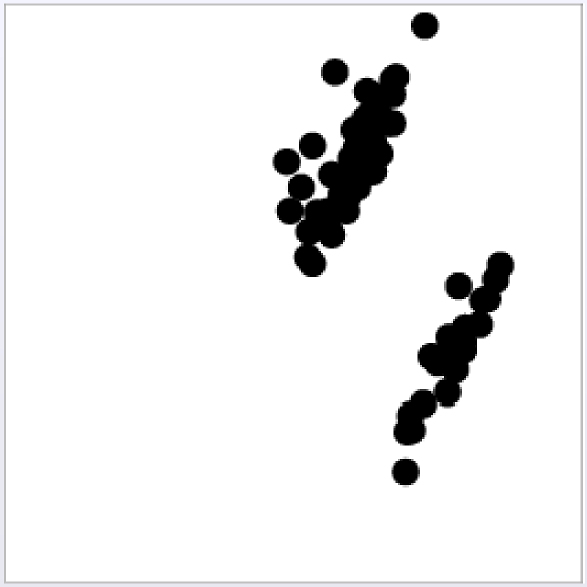
HDP



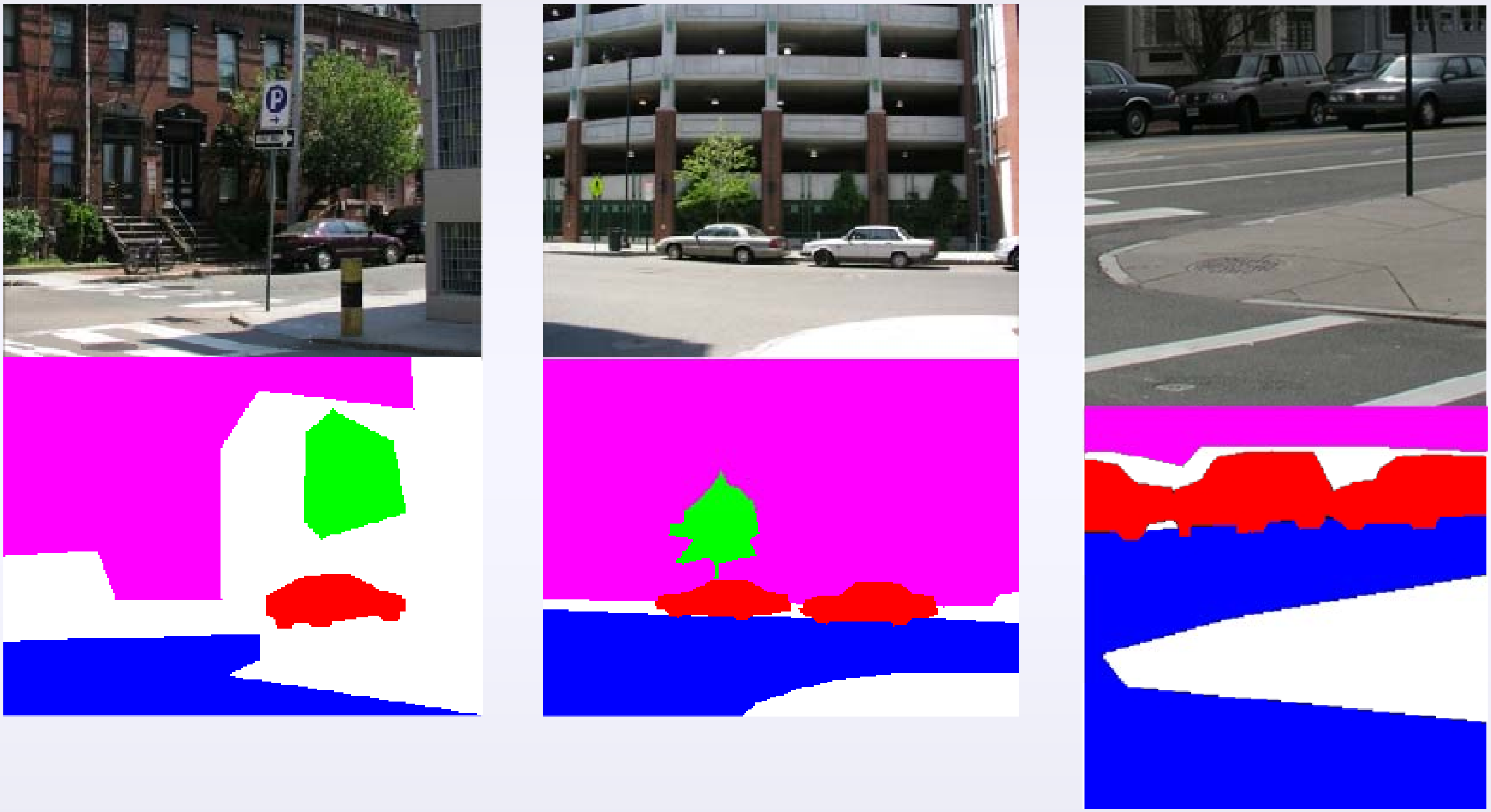
TDP







Counting & Locating Objects

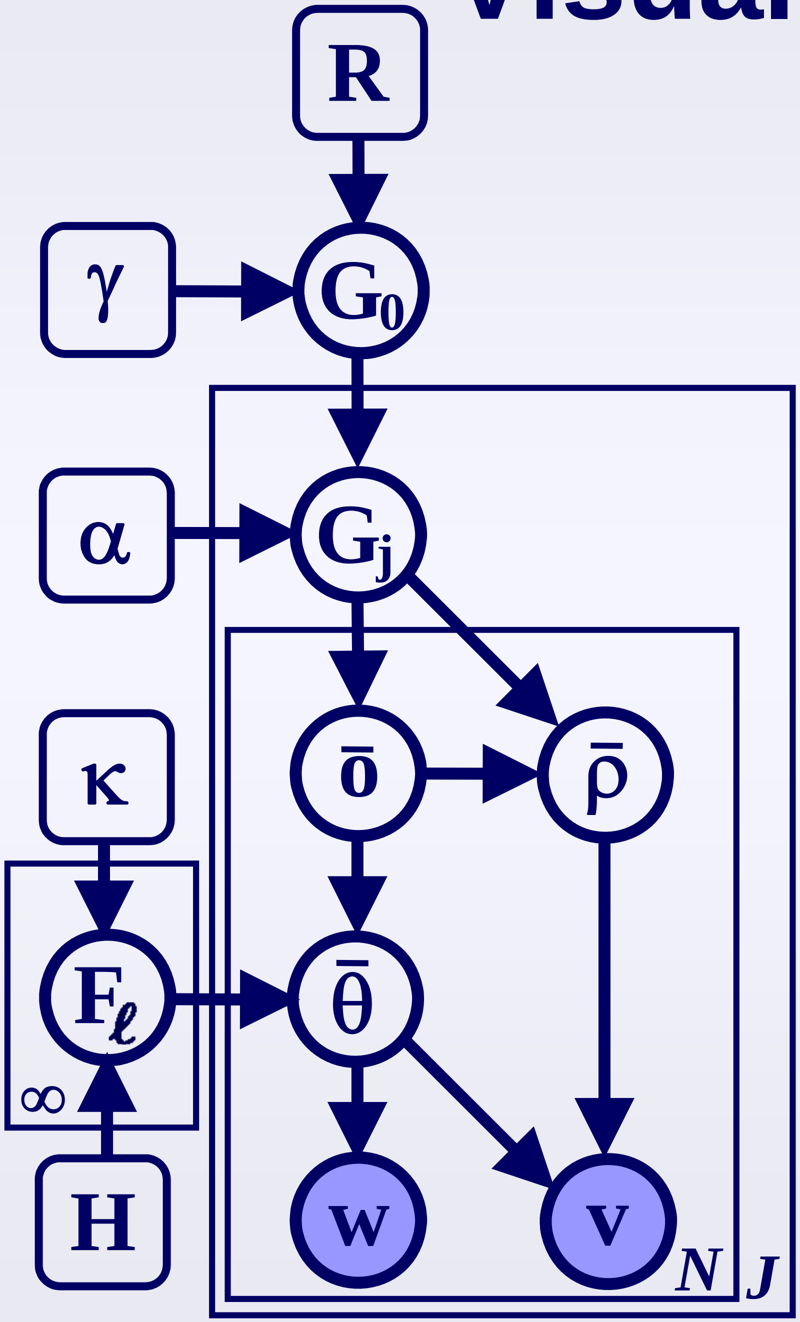


- How many cars are there?
- Where are those cars in the scene?

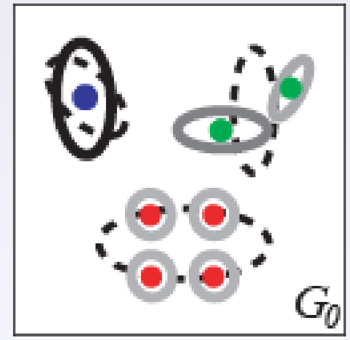
Dirichlet Processes

Transformations

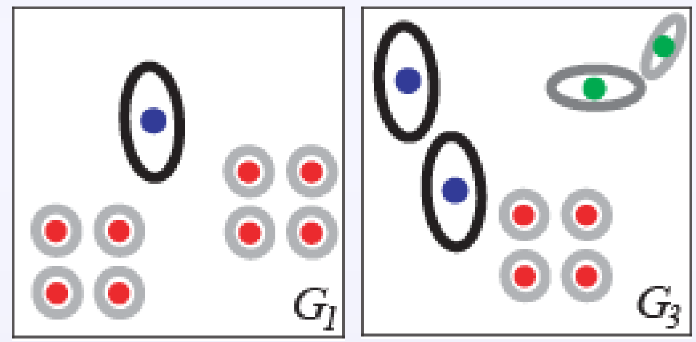
Visual Scene TDP



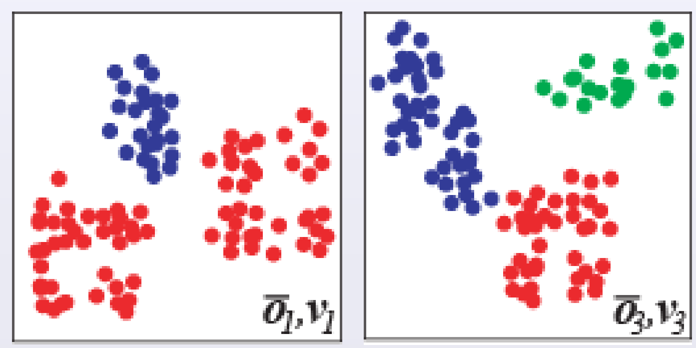
Global Density
Object category
Part size & shape
Transformation prior



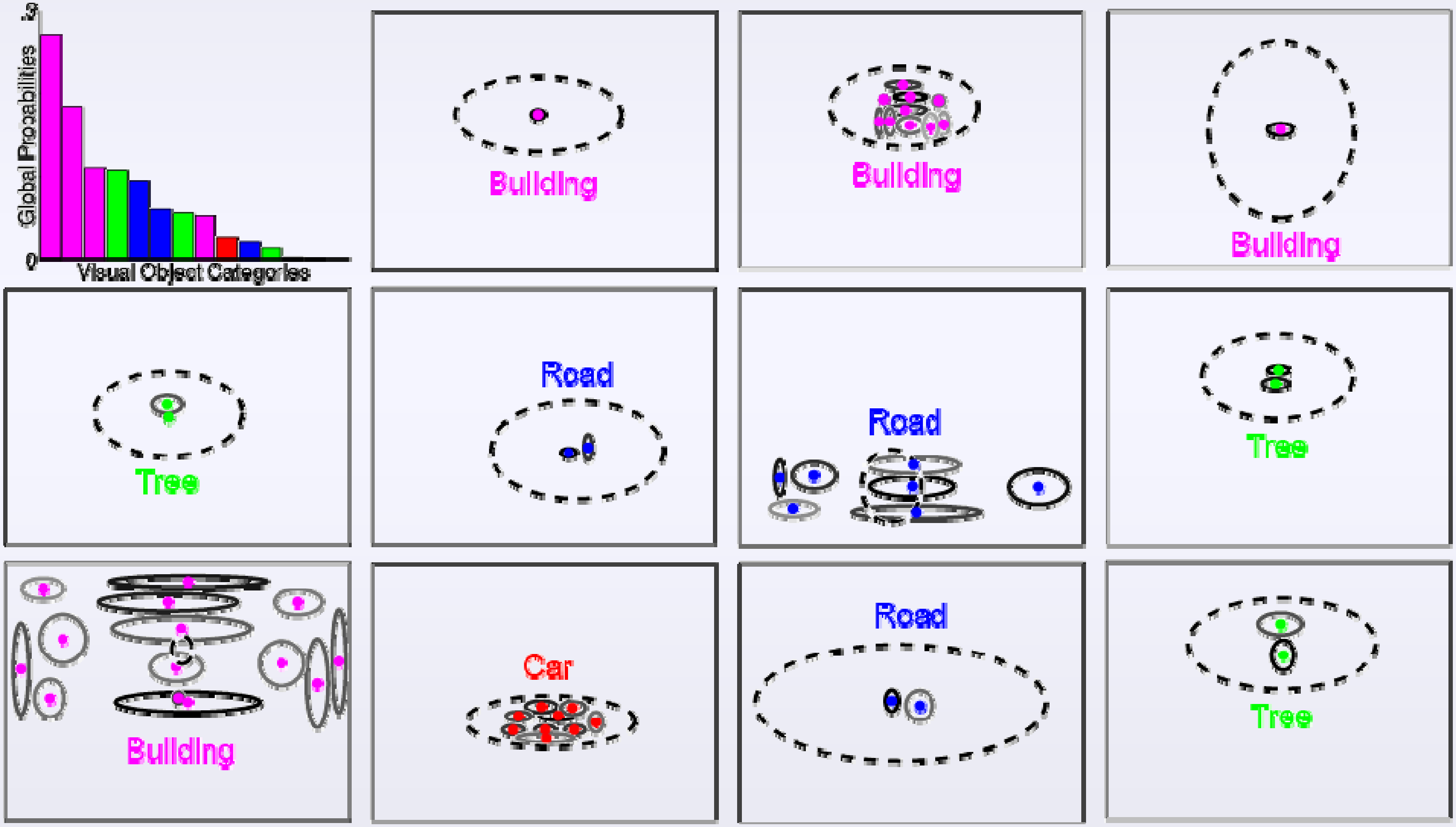
Transformed Densities
Object category
Part size & shape
Instance locations



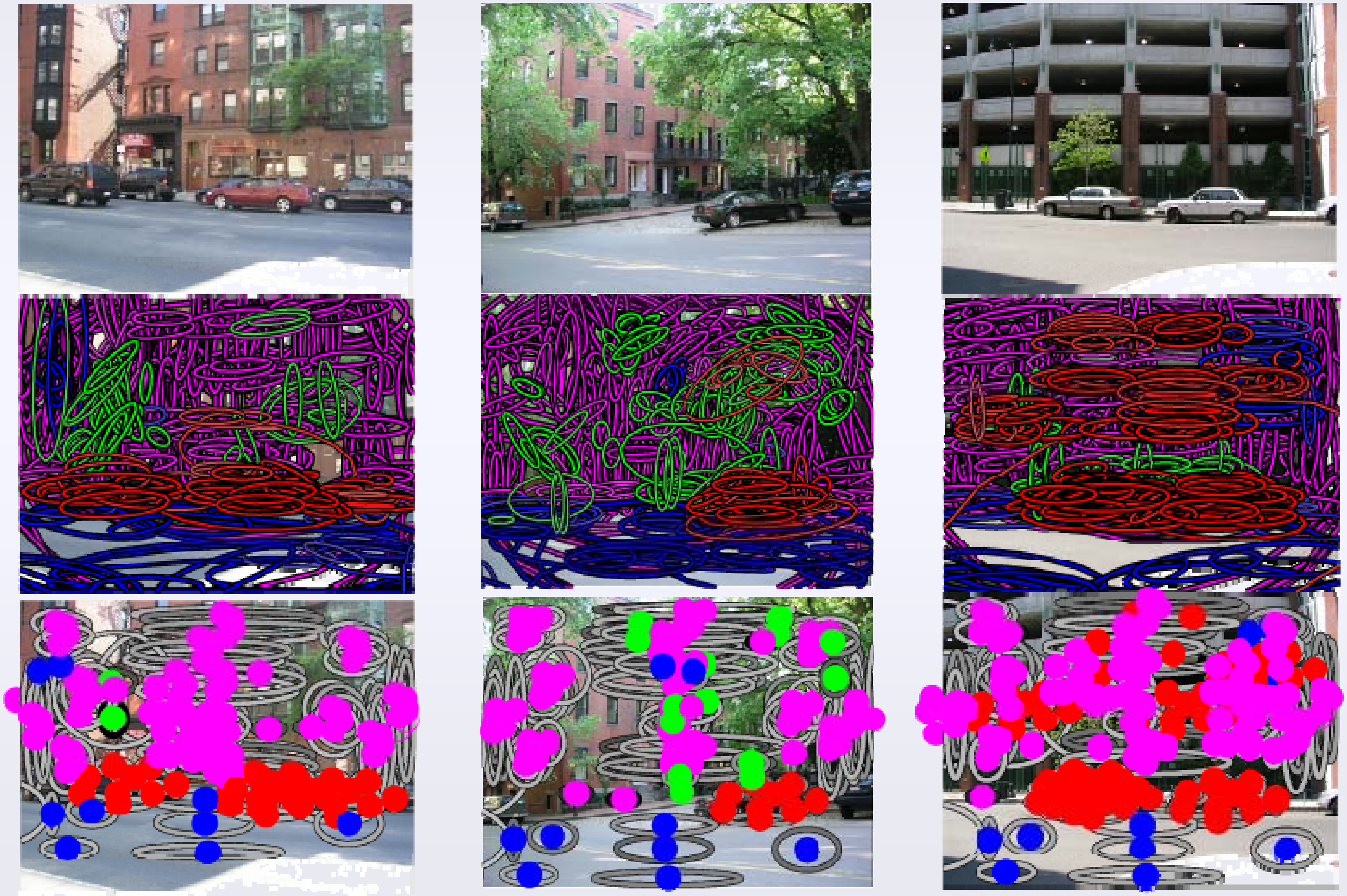
2D Image Features
Appearance
Location



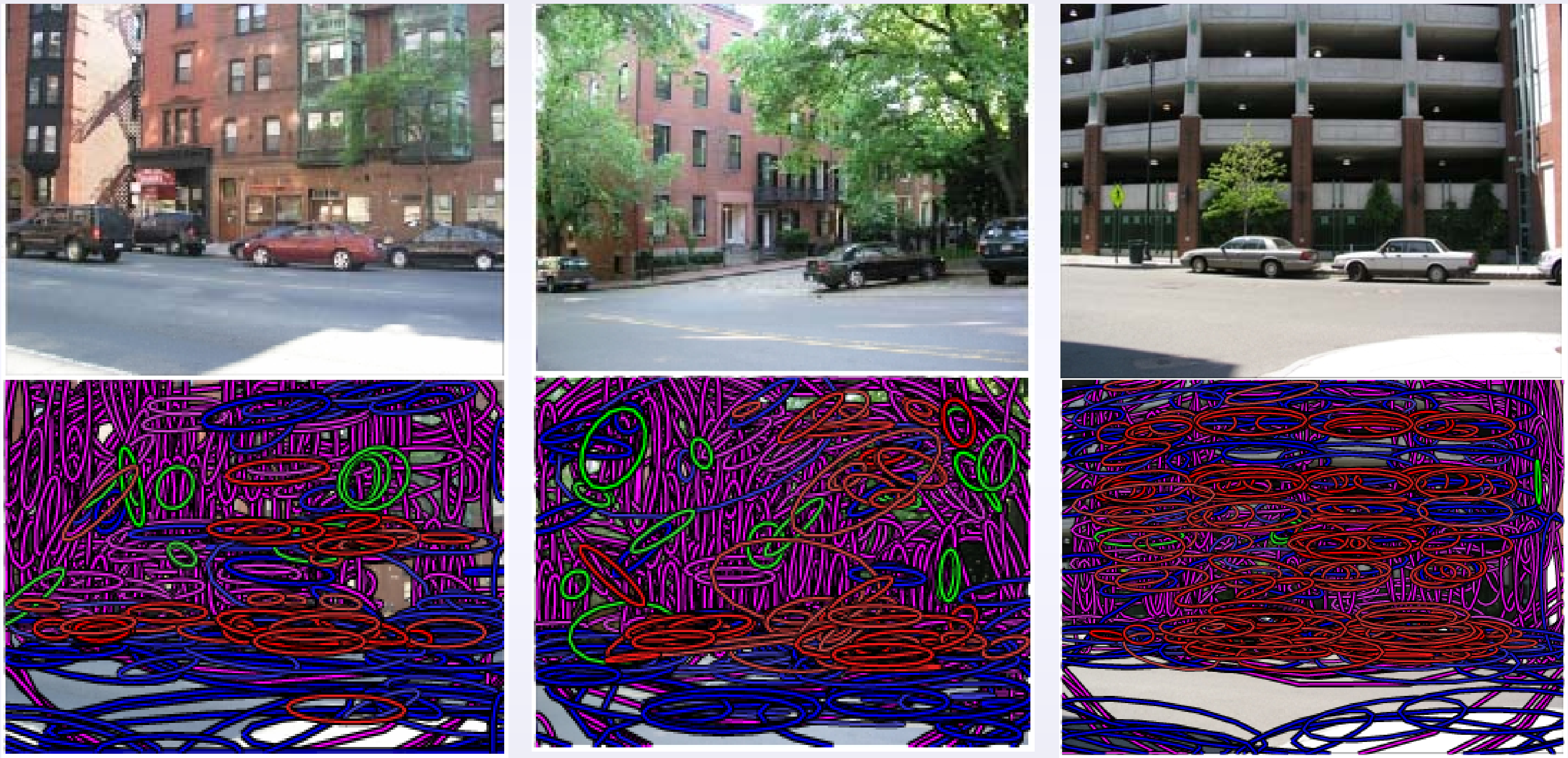
Street Scene Visual Categories



Street Scene Segmentations

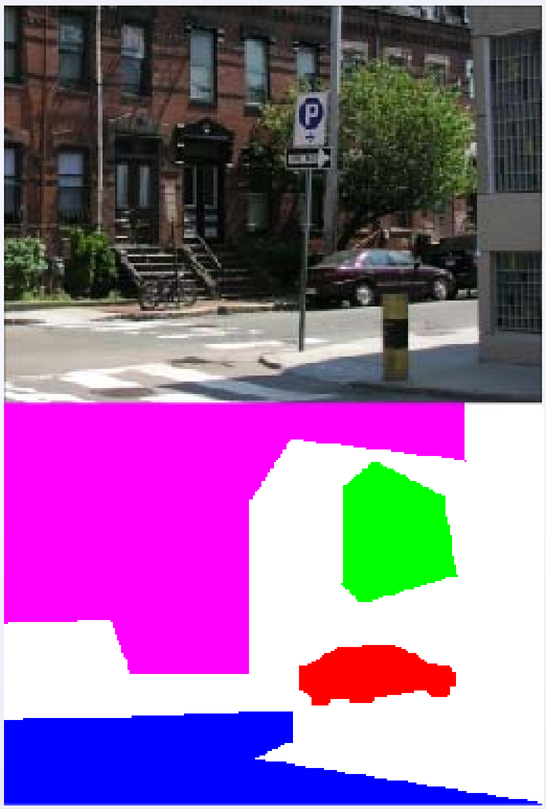


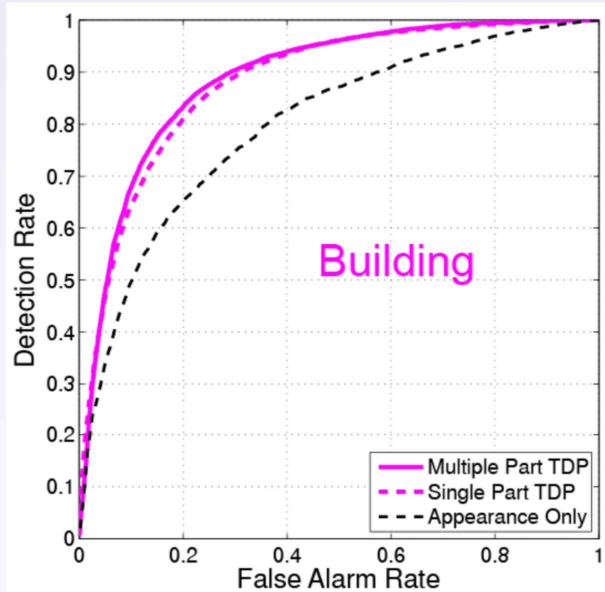
Appearance Only

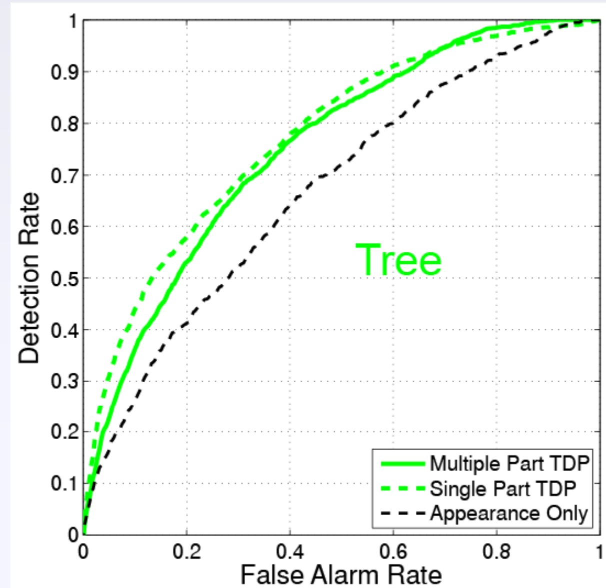


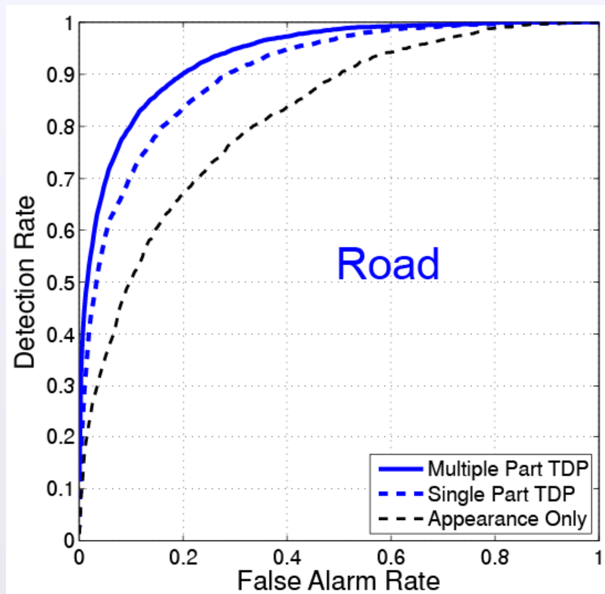
- “Bag of features” model, ignores feature positions
- Inferior segmentations, cannot count objects

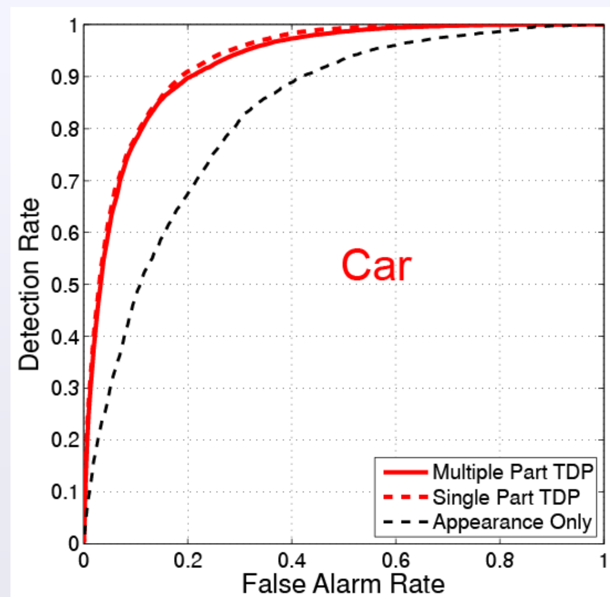
Segmentation Performance





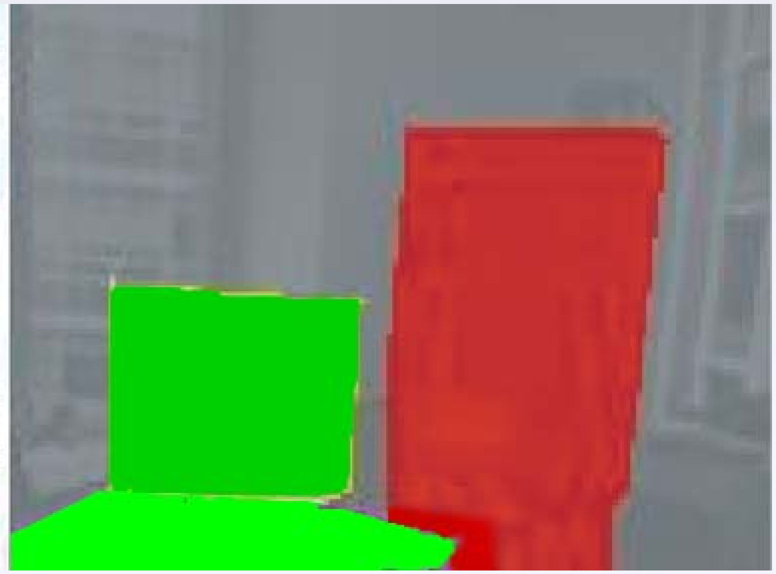






Objects & 3D Reconstruction

An Office Scene



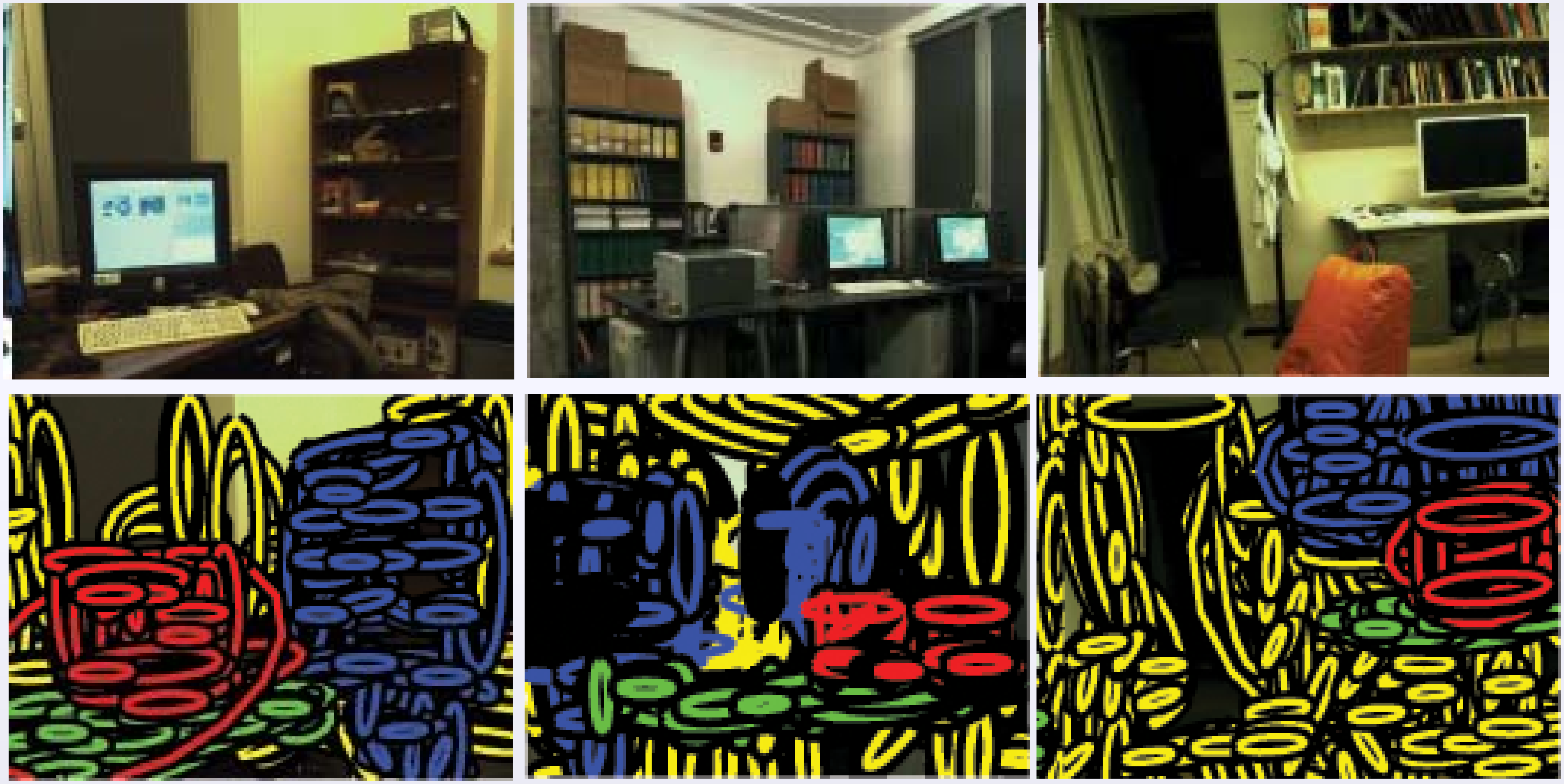
Green ⟷ Near

Red ⟷ Far

- Given 3D structure, segmentation is easier
- Identifying objects regularizes depth estimation

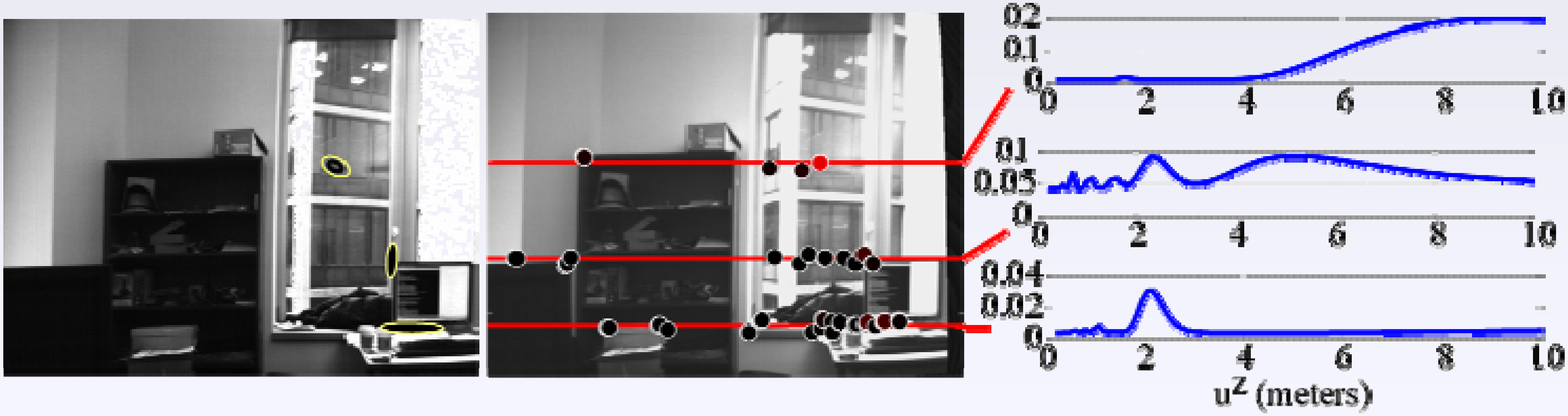
Office Scene Training Images

Objects at Multiple Scales

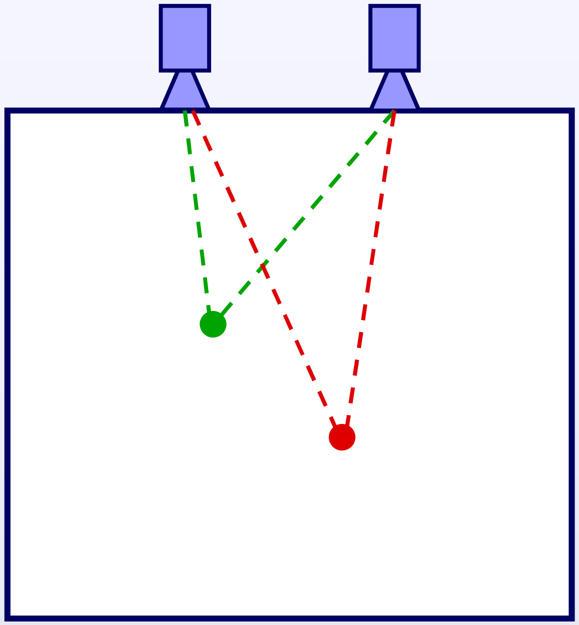


Computer Screens
Desks
Bookshelves

3D Structure from Stereo



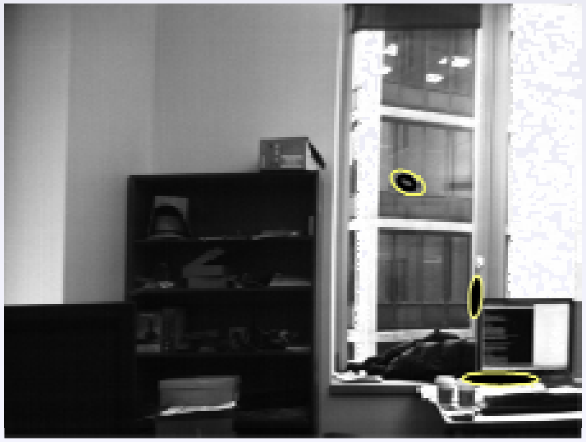
Reference (left) Image *Potential Matches* *Depth Densities*

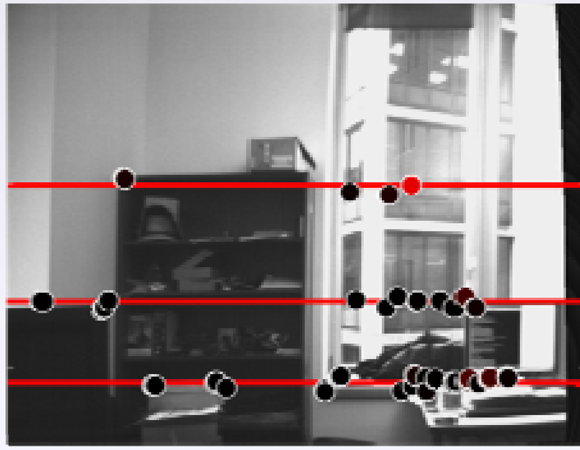


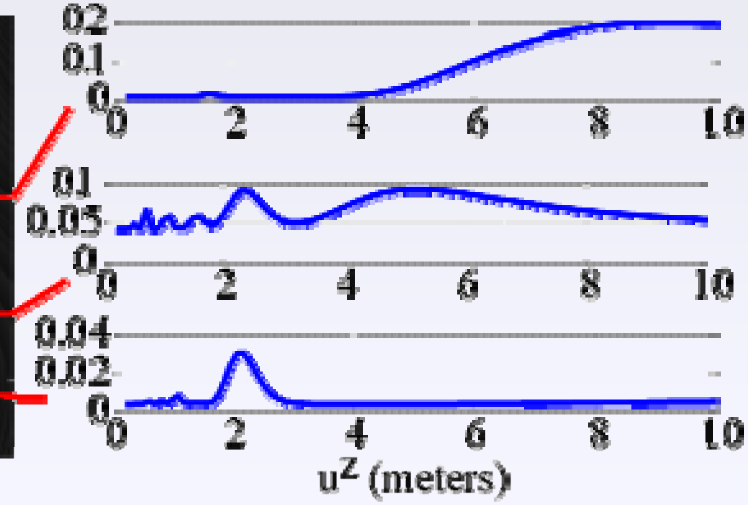
Overhead View

$$\textbf{\textit{Depth}} = \frac{\delta}{\textbf{\textit{Disparity}}}$$

Greedy Depth Estimates



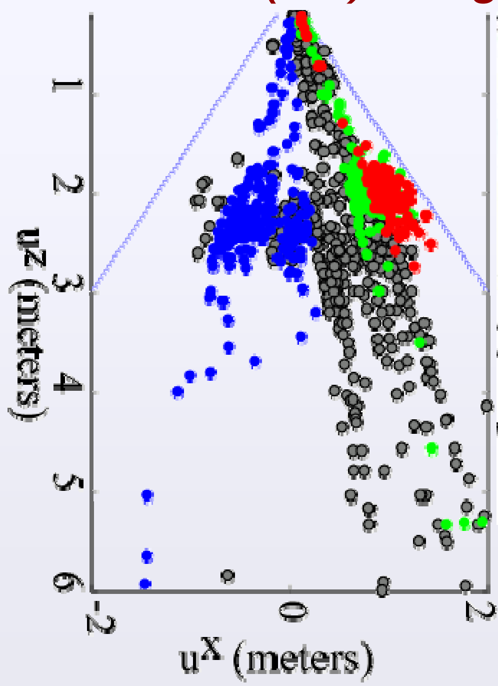


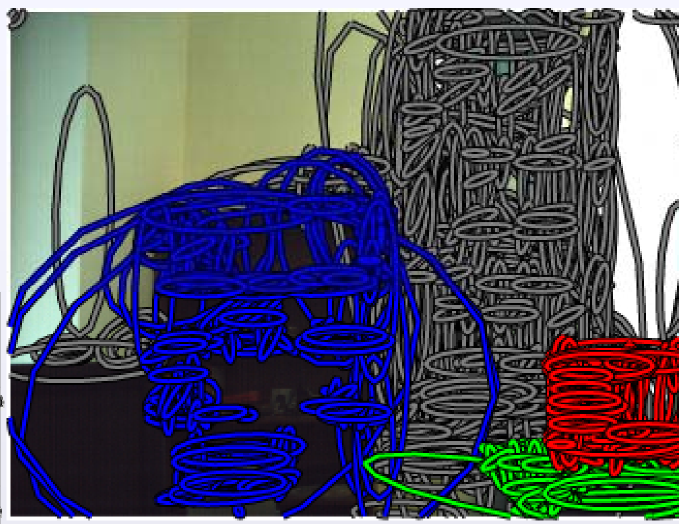


Reference (left) Image

Potential Matches

Depth Densities

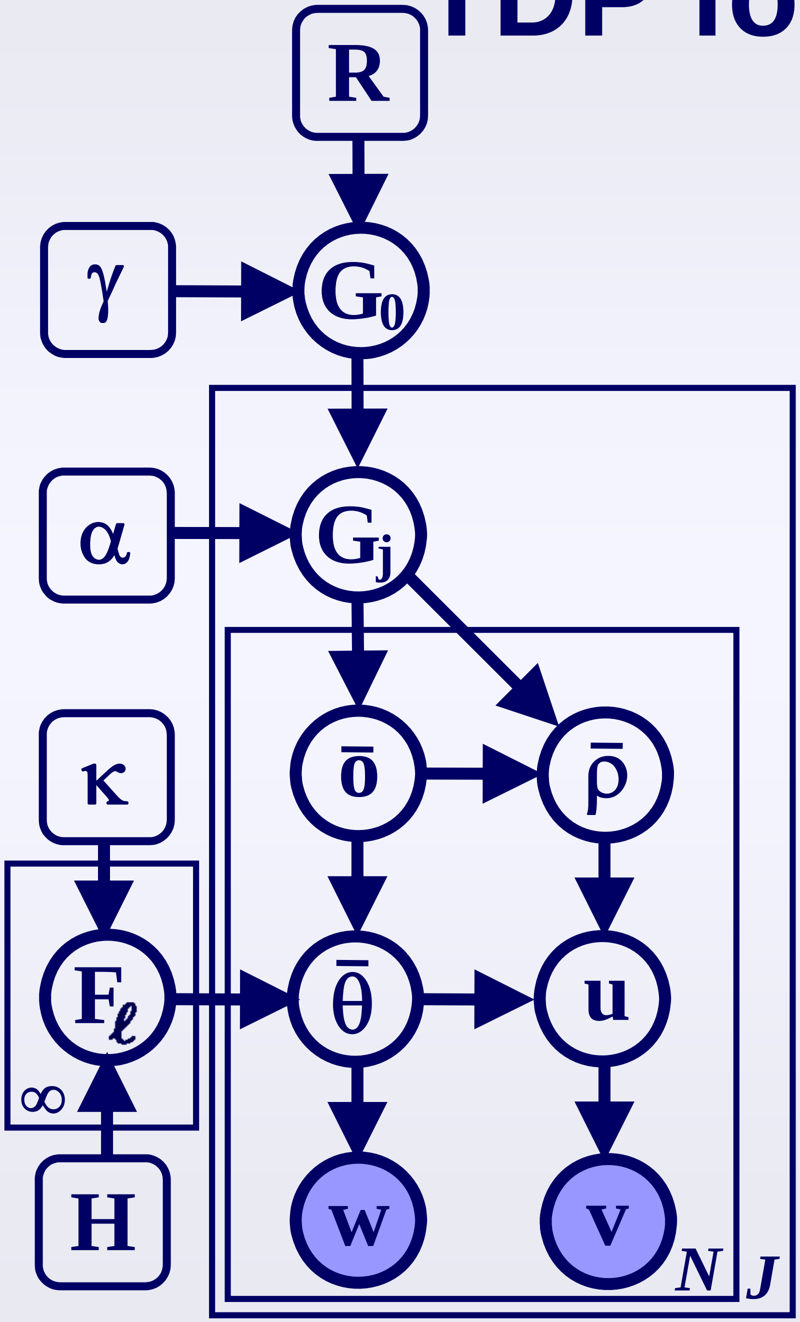




Green ⟷ Near

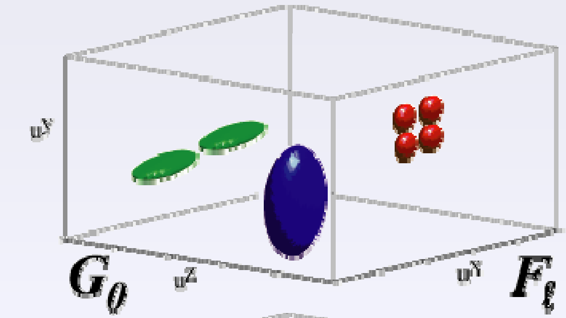
Red ⟷ Far

TDP for 3D Scenes



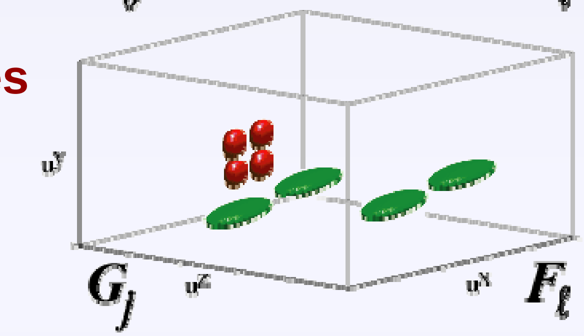
Global Density

Object category
Part size & shape
Transformation prior



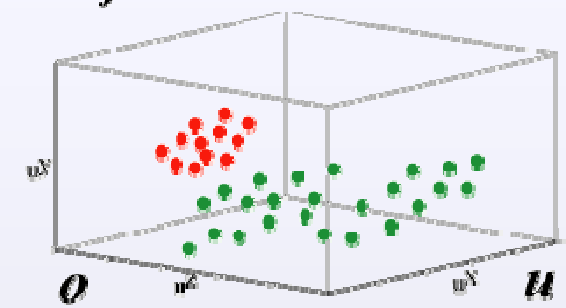
Transformed Densities

Object category
Part size & shape
Transformed locations



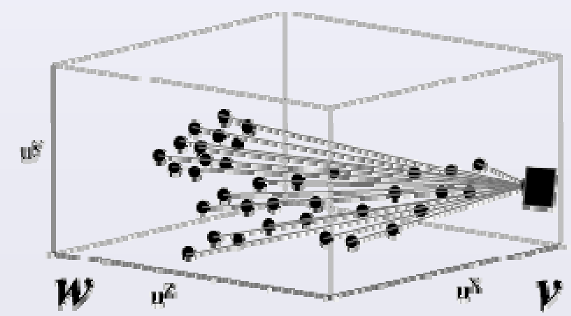
3D Scene Features

Object category
3D Location

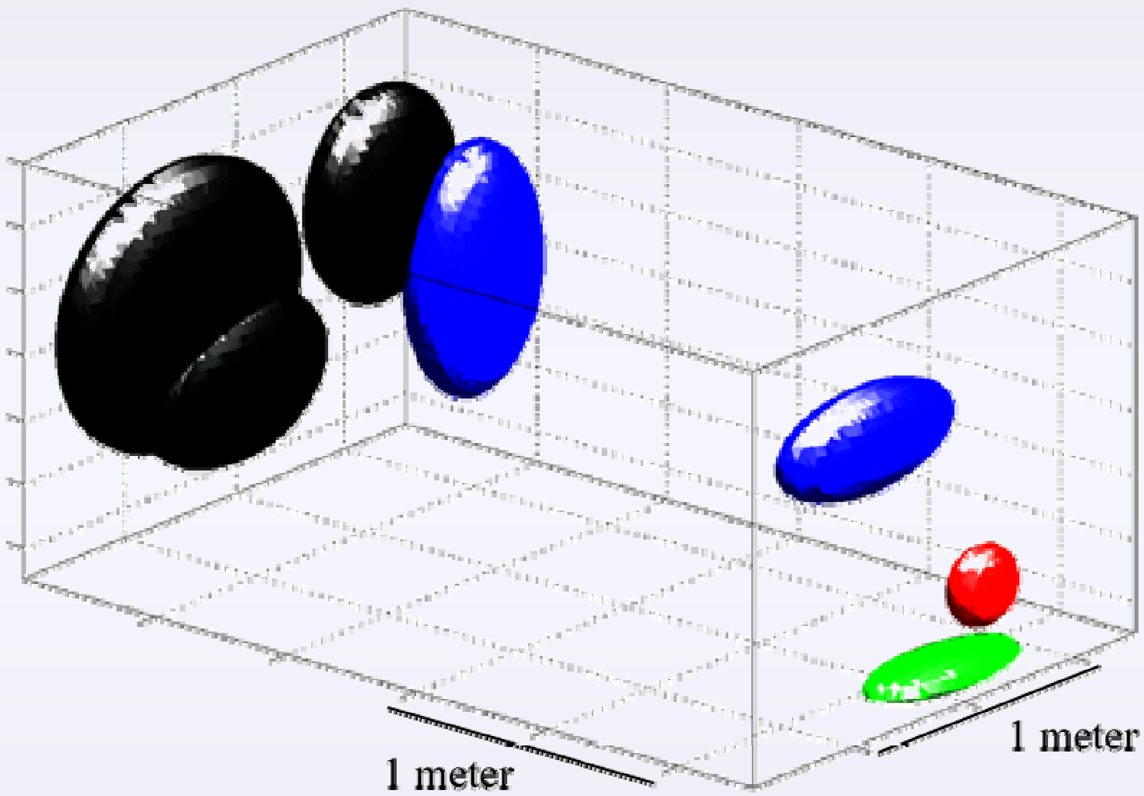


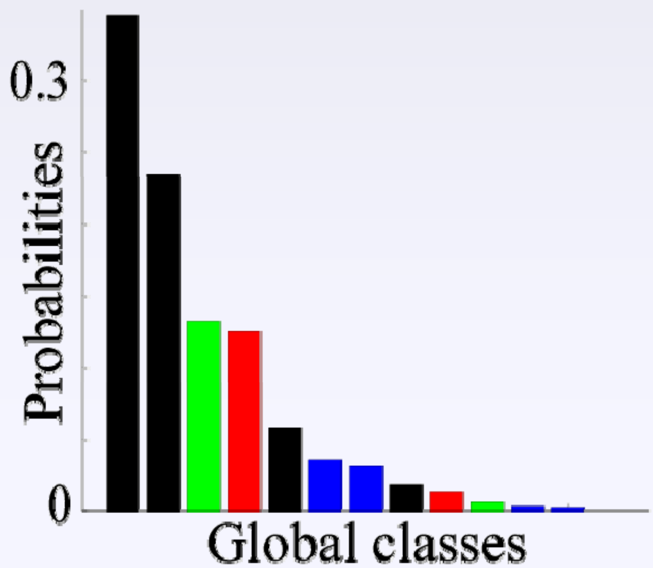
2D Image Features

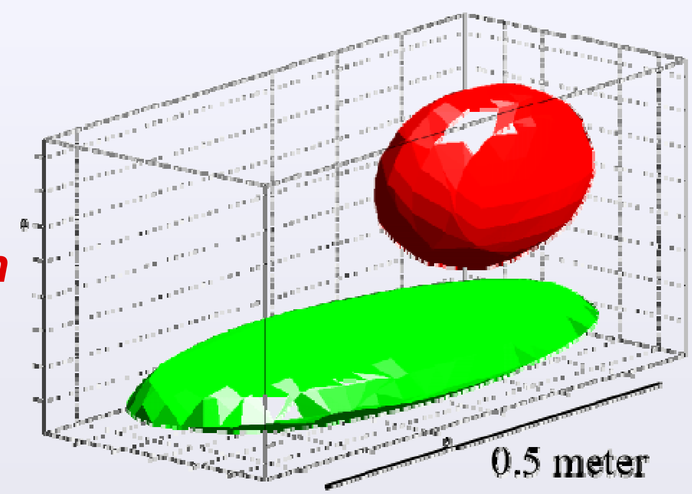
Appearance Descriptors
2D Pixel Coordinates



Single-Part Office Scene Model

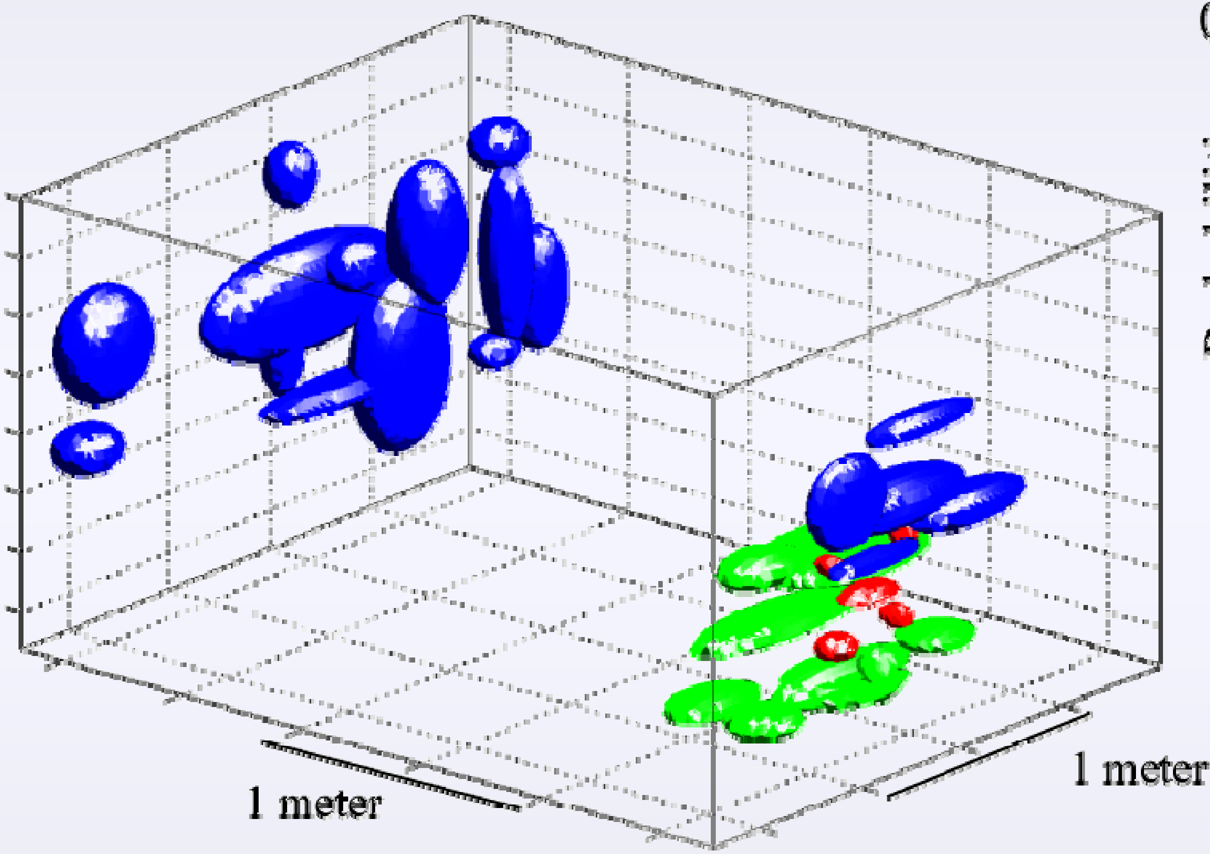


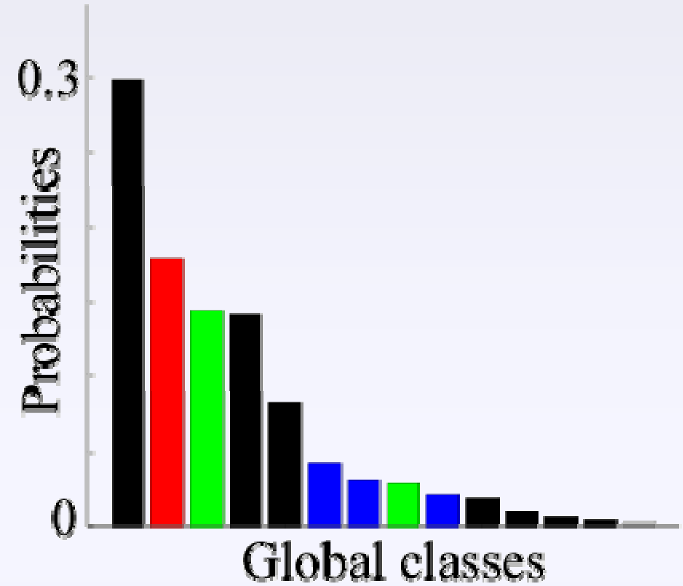


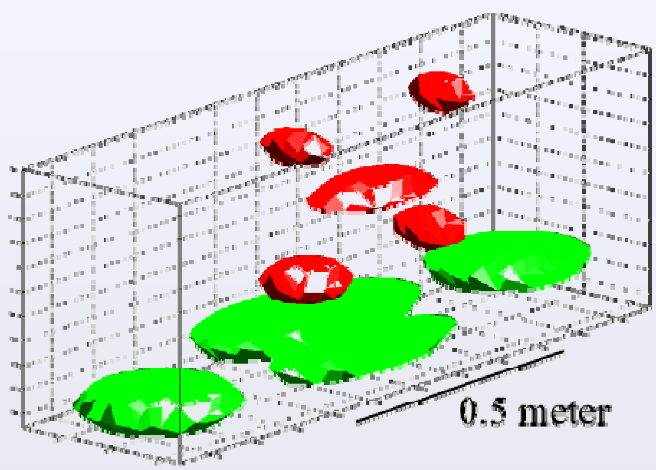


Background ***Bookshelves*** ***Computer Screen*** ***Desk***

Multi-Part Office Scene Model

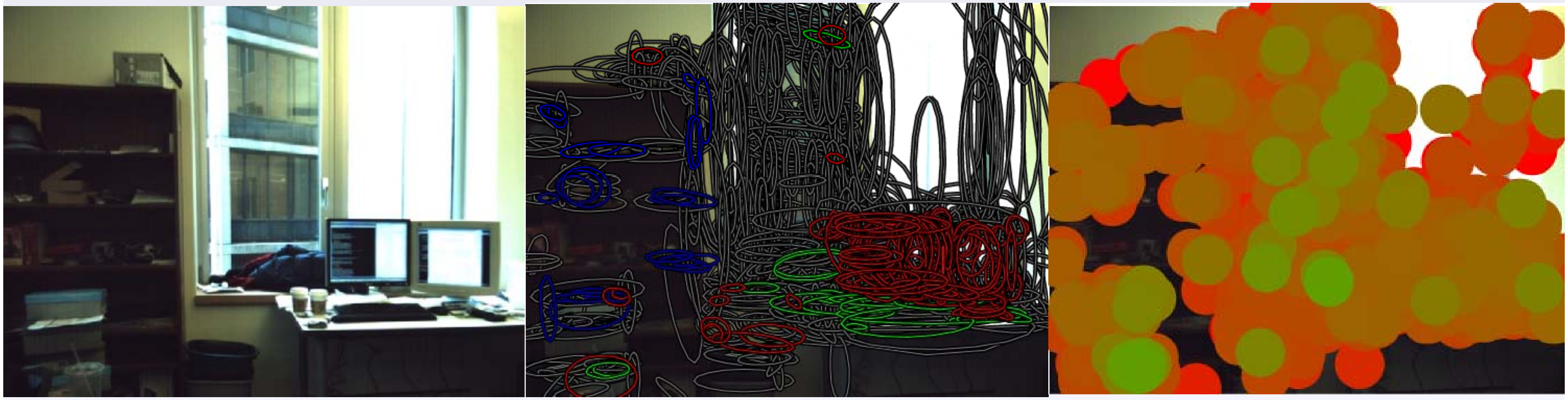


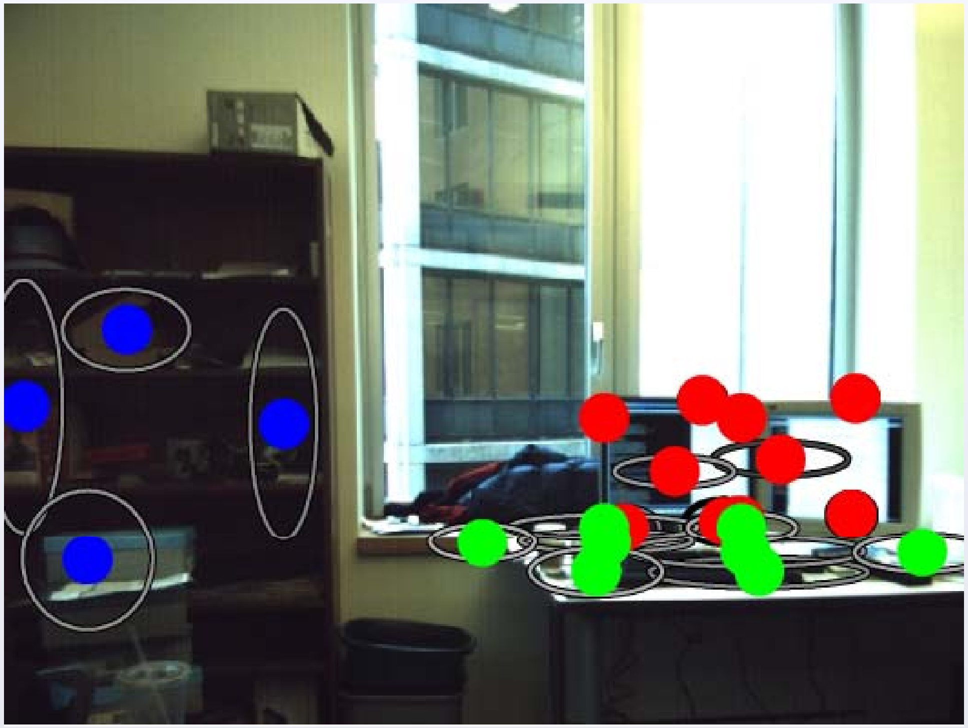


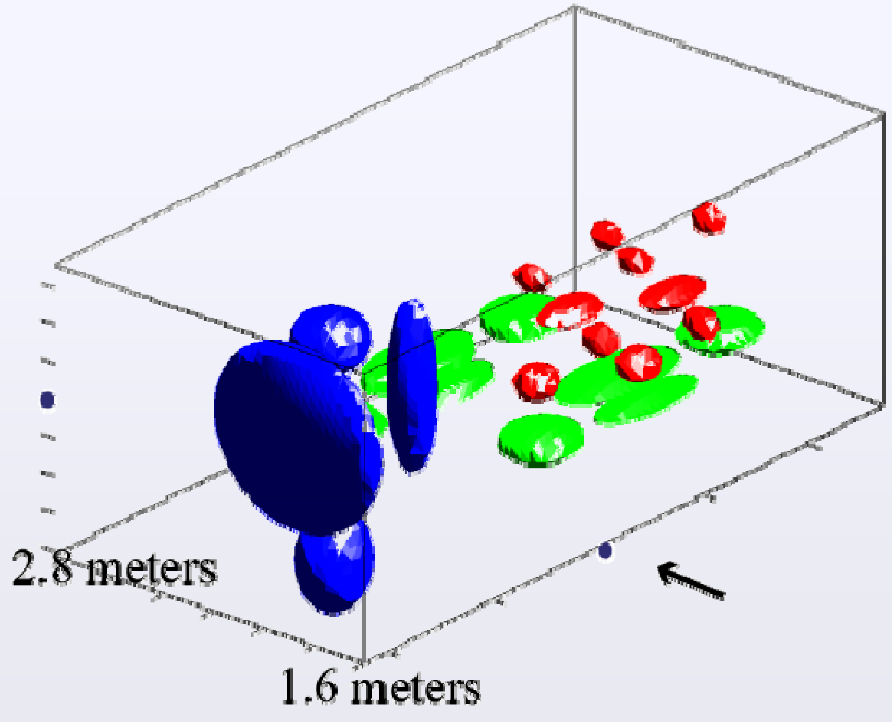


Background ***Bookshelves*** ***Computer Screen*** ***Desk***

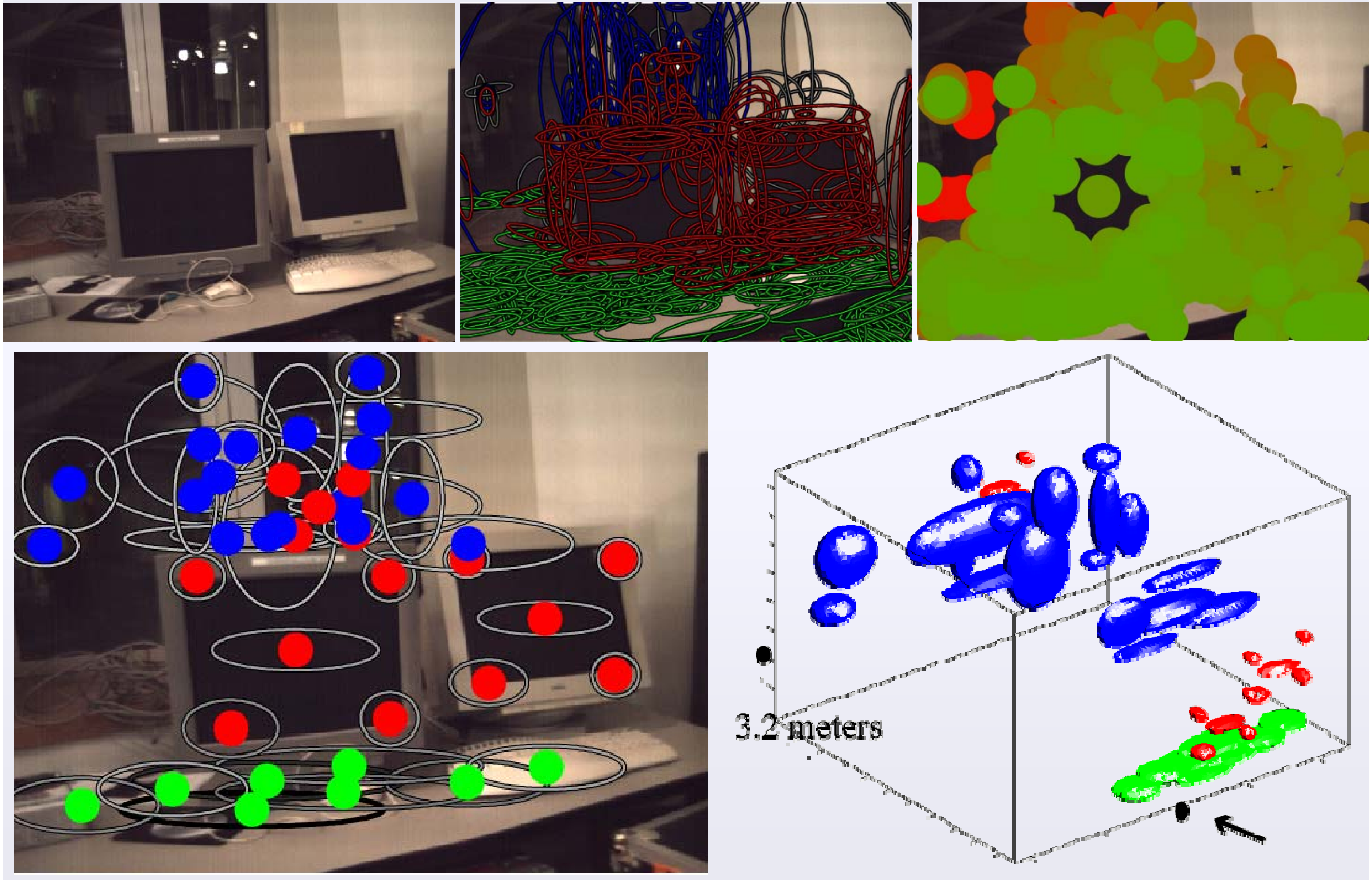
Stereo Test Image I



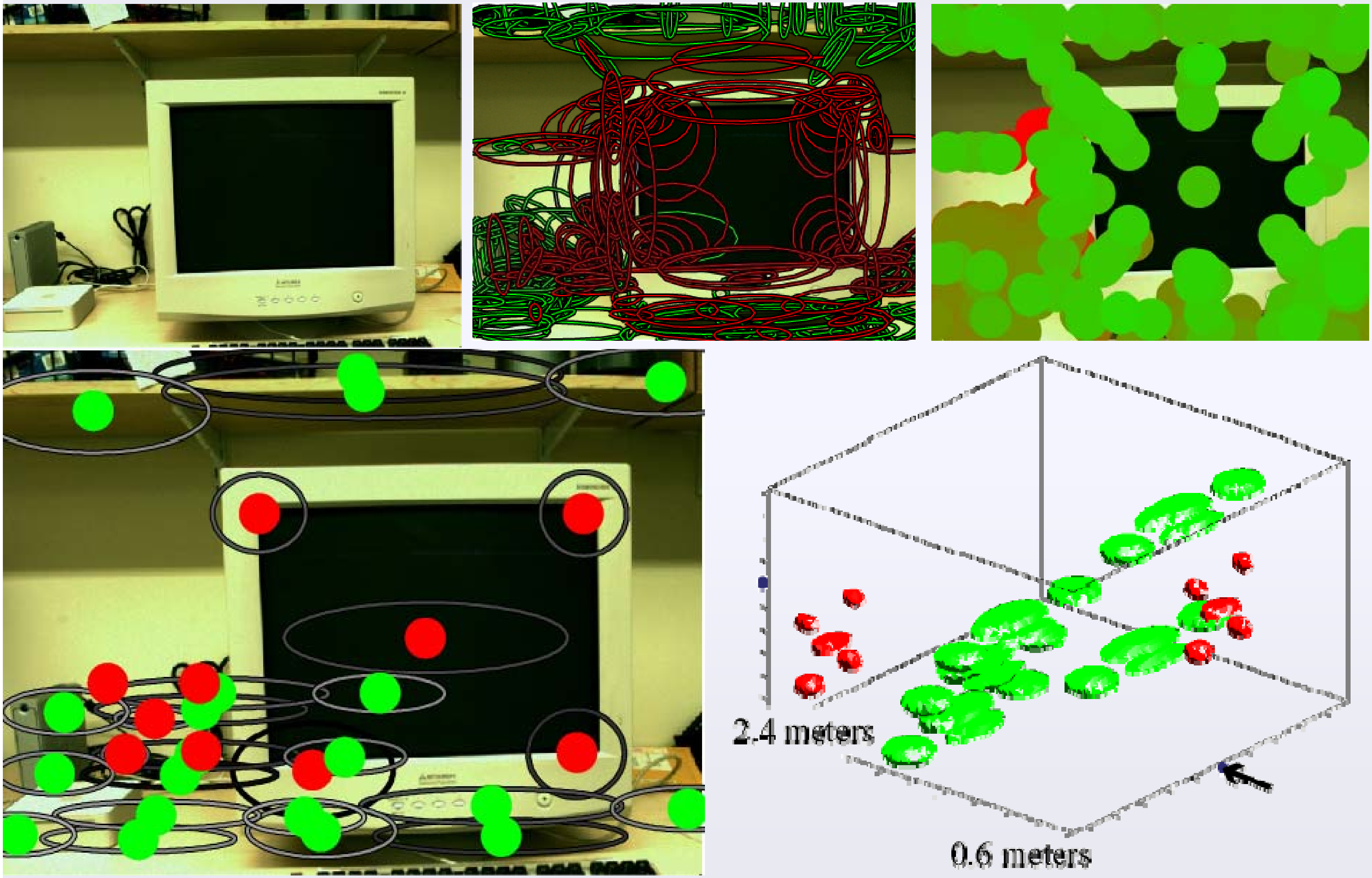




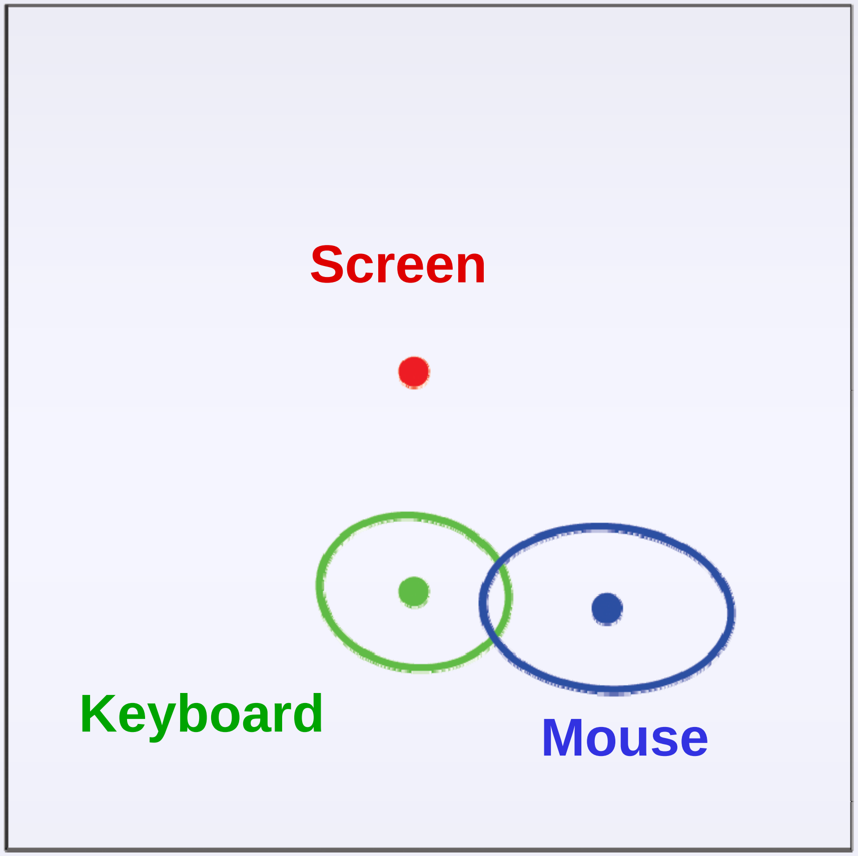
Stereo Test Image II



Ongoing Work: Monocular Test



Ongoing Work: Context

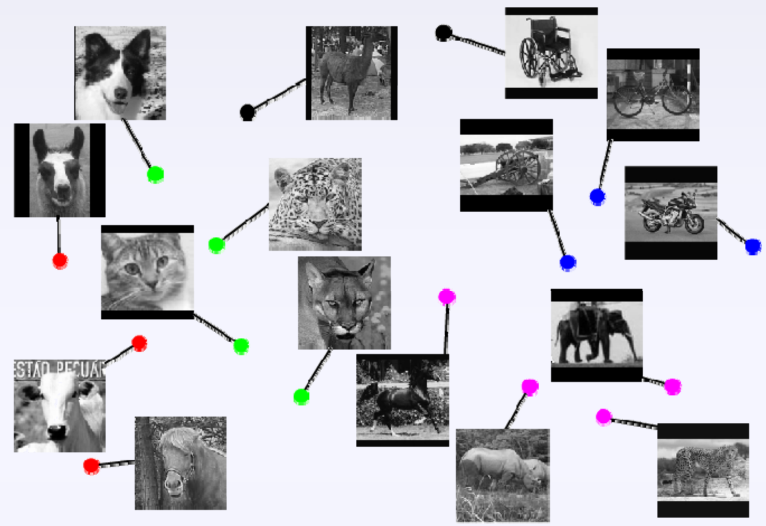


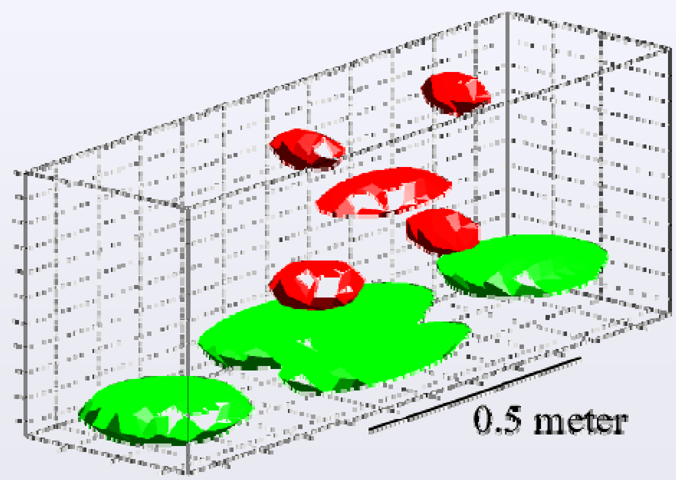
- Developed *fixed-order* contextual scene model
- Extension to Transformed DP model is an open problem
- Needed: Richer models for *background* scene structure

Conclusions

Transformed Dirichlet Processes allow…

- flexible *transfer* of knowledge among related object categories
- robust learning from small, *partially labeled* datasets
- *an integrated* view of object recognition & 3D reconstruction
- potential *scaling* of nonparametric methods to complex domains





Today

Sudderth guest lecture:

- Constellation Models (Fergus)
- Unsupervised Object Discovery with pLSA (Sivic)
- Scene Models (Li)
- Transformed Models (Sudderth)

Daphna Buschsbaum student presentation:

- **pLSA models of activity (Neibles)**

Moreels guest lecture:

- A probabilistic formulation of voting / SIFT (Moreels)

Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words

Juan Carlos Niebles, Hongcheng Wang, Li Fei-Fei

Daphna Buschsbaum student presentation

Video

Goal: Automatically Categorize or Localize Different Actions

- moving cameras
- non-stationary background
- moving target
- multiple activities

QuickTime™ and a
decompressor
are needed to see this picture.

Overview

QuickTime™ and a
decompressor
are needed to see this picture.

Approach

- Generative model
  - Bag of (spatio-temporal) video words
  - Actions are distributions over words
  - Videos are distributions over actions
  - Based on topic modeling of documents
    - pLSA
- Unsupervised learning of video “topics” (actions)
  - Use to categorize actions
  - Use to localize actions within video sequences

Interest Point Detector

QuickTime™ and a
decompressor
are needed to see this picture.

- From Piotr Dollàr, Vincent Rabaud, Garrison Cottrell, and Serge Belongie, *2005*

QuickTime™ and a
decompressor
are needed to see this picture.

Dollàr et. al.

- $\omega = 4/\tau$
- σ= spatial extent
- τ= temporal extent

QuickTime™ and a
decompressor
are needed to see this picture.

Interest points
Centered at local maxima
Of R

Dollàr et. al.

Interest Point Detector

QuickTime™ and a
decompressor
are needed to see this picture.

Dollàr et. al.

Cuboids

- A cuboid (or right prism) of data is extracted around each feature point (local maximum of the response function). Each cuboid has spatial and temporal extend

QuickTime™ and a
decompressor
are needed to see this picture.

Dollàr et. al.

Cuboids

- Size of the cuboid is set to contain most of the volume that contributed to the response function at that interest point; cuboids have a side length ≈ six times the scale at which they were detected.

QuickTime™ and a
decompressor
are needed to see this picture.

Dollàr et. al.

Feature/Word Representation

- Flatten cuboids into single vector. Approaches tried:
  - Brightness gradients
  - Optical flow
  - Gradient histograms
- PCA
- Cluster into “types”

QuickTime™ and a
decompressor
are needed to see this picture.

Dollàr et. al

Feature/Word Representation

- Flatten cuboids
  - **Brightness gradients**
  - Optical flow
  - Gradient histograms
- Cluster into “codewords”???

Niebles et. al.

QuickTime™ and a
decompressor
are needed to see this picture.

Generative Topic Model (Video pLSA)

QuickTime™ and a
decompressor
are needed to see this picture.

Learning Topics/Actions

Fitting Model:

- Distribution of words per action
  - Common across all videos
- Distribution of actions per video
  - Video specific
- Use Expectation Maximization algorithm to find values that maximize:

$$\prod_{i=1}^{M}\prod_{j=1}^{N} p(w_i \mid d_j)^{n(w_i, d_j)}$$

Where: $$p(w_i \mid d_j) = \sum_{k=1}^{K} p(z_k \mid d_j)\, p(w_i \mid z_k)$$

Experiments

- KTH human motions data
  - 6 classes performed by 25 actors
  - 3 actors used to learn video word vocabulary
  - Leave one out cross-validation (learn model on 24 actors, test on 25 for all actors)
- SFU figure skating data
  - 3 classes, 7 actors
  - Learn video word vocabulary from 6 actors
  - Leave one out cross-validation

Categorization

- Similar to learning, but with distribution of words per action $p(w_i|z_k)$ fixed:

$$p(w \mid d_{test}) = \sum_{k=1}^{K} p(z_k \mid d_{test}) \, p(w \mid z_k)$$

- Classified as:

$$\arg\max p(z_k \mid d_{test})$$

Categorization Results

QuickTime™ and a
decompressor
are needed to see this picture.

QuickTime™ and a
decompressor
are needed to see this picture.

Categorization Results

QuickTime™ and a
decompressor
are needed to see this picture.

Localization

QuickTime™ and a
decompressor
are needed to see this picture.

Localization Results

QuickTime™ and a
decompressor
are needed to see this picture.

Today

Sudderth guest lecture:

- Constellation Models (Fergus)
- Unsupervised Object Discovery with pLSA (Sivic)
- Scene Models (Li)
- Transformed Models (Sudderth)

Daphna B. student presentation:

- pLSA models of activity (Neibles)

Moreels guest lecture:

- **A probabilistic formulation of voting / SIFT (Moreels)**

Features-based Object Recognition

Pierre Moreels

UC Berkeley, Feb. 17, 2009

The recognition continuum

Individual objects

BMW logo

Categories

cars

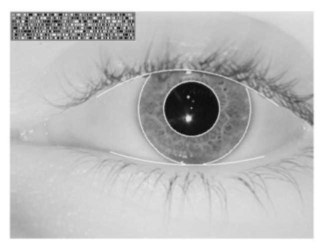
means of transportation

variability

Applications

Autonomous navigation

Help Daiki find his toys !



Identification, Security.

Outline

- Problem setup
- Features
- Coarse-to-fine algorithm
- Probabilistic model
- Experiments
- Conclusion

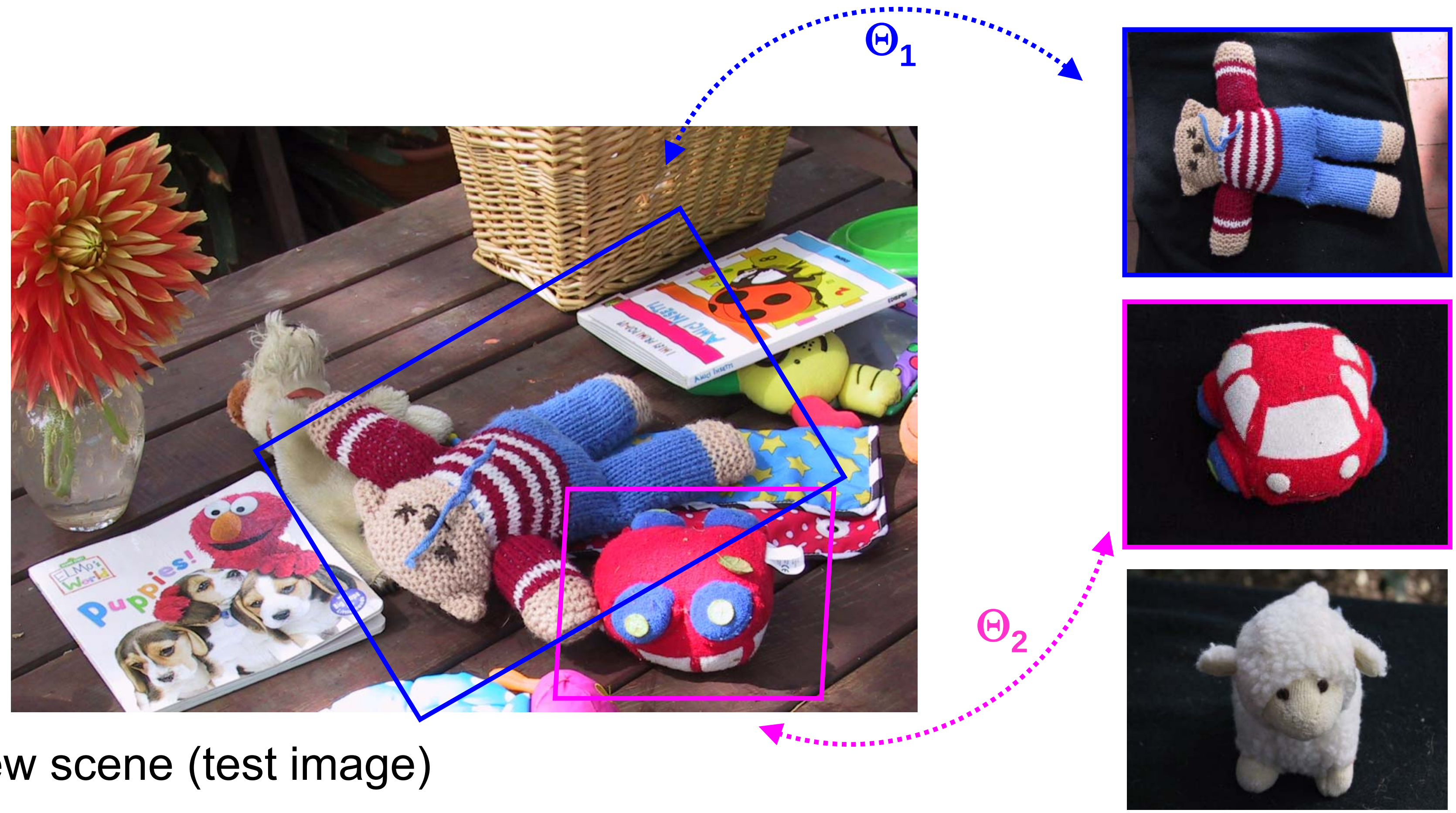
The detection problem

New scene (test image)

⋮ Models from database

Find models and their pose (location, orientation…)

Hypotheses – models + positions



New scene (test image)

Models from database

Θ = affine transformation

Matching features

New scene (test image)

Models from database

⇒ Set of correspondences = **assignment vector**

Features detection

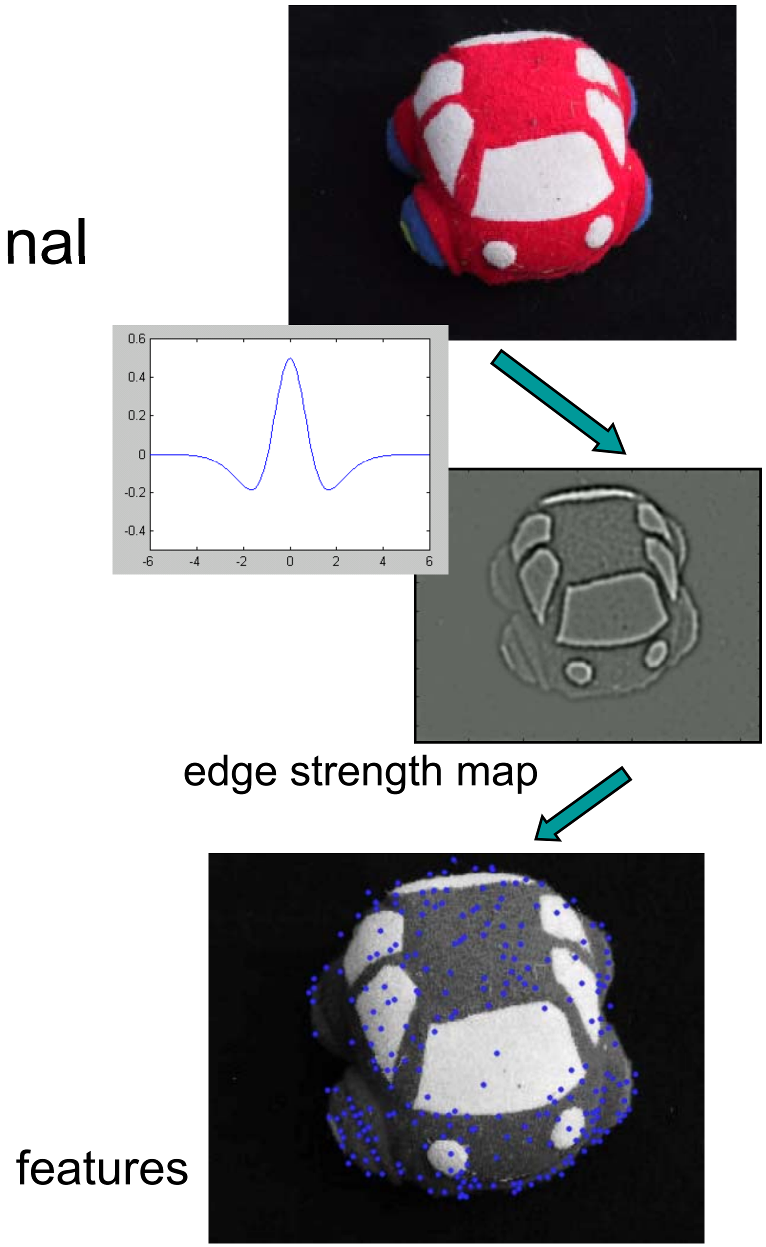
Image characterization by features

- Features = high information content

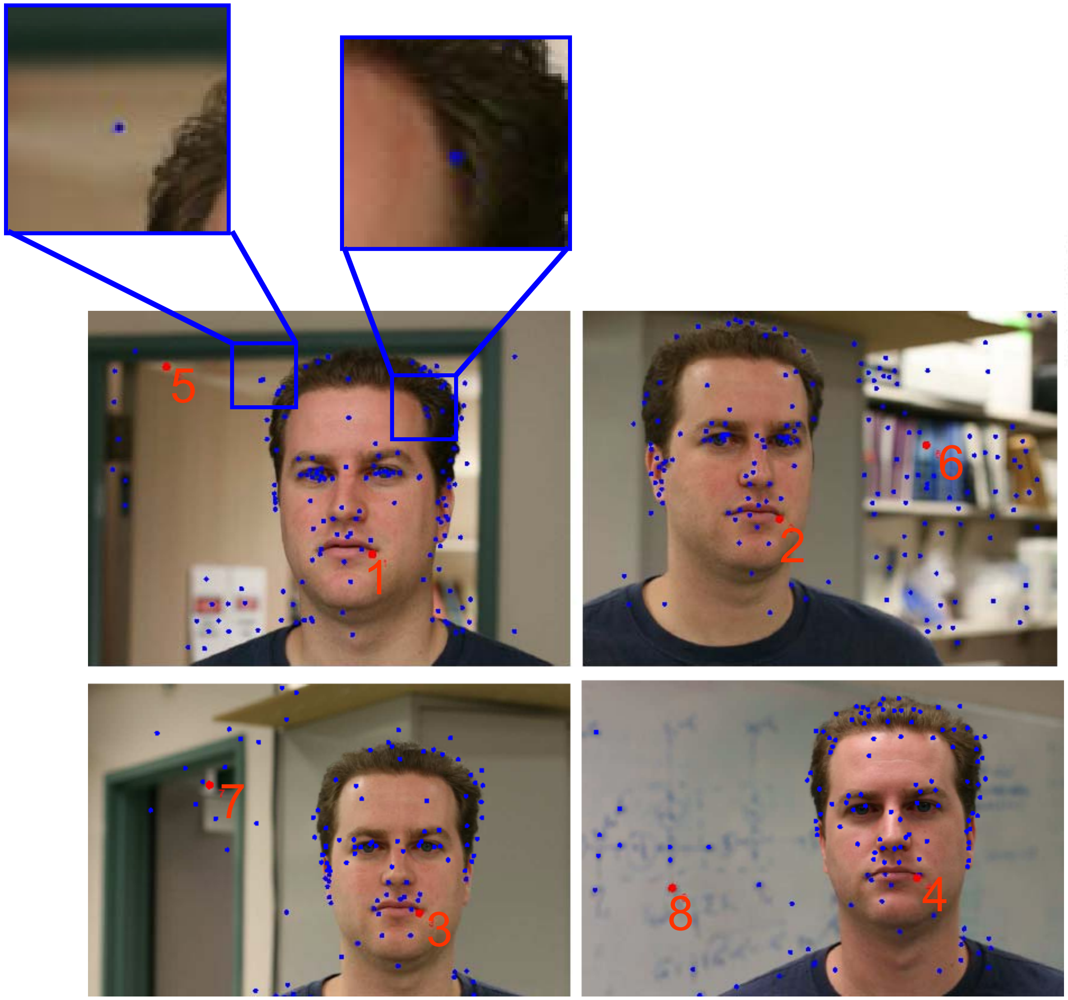
  ‘locations in the image where the signal changes two-dimensionally’ C.Schmid

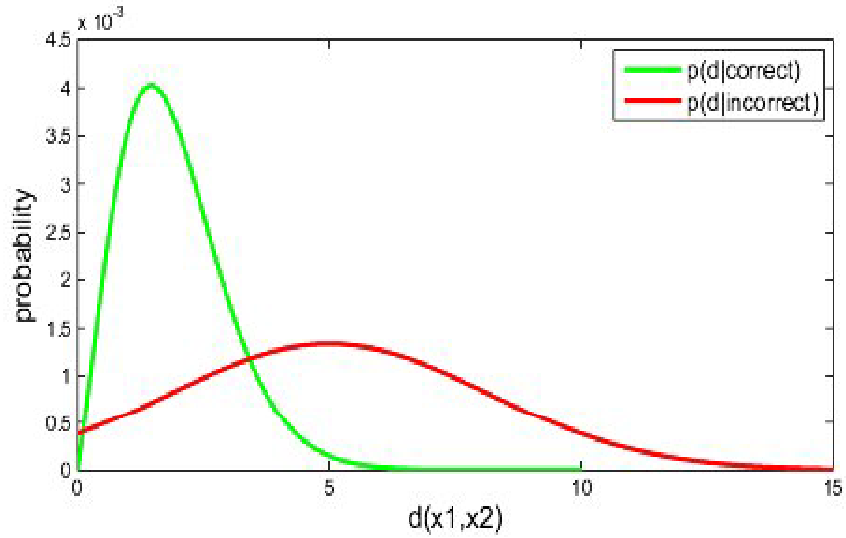
- Reduce the volume of information

  - [Sobel 68]
  - Diff of Gaussians [Crowley84]
  - [Harris 88]
  - [Foerstner94]
  - Entropy [Kadir&Brady01]



Correct vs incorrect descriptors matches



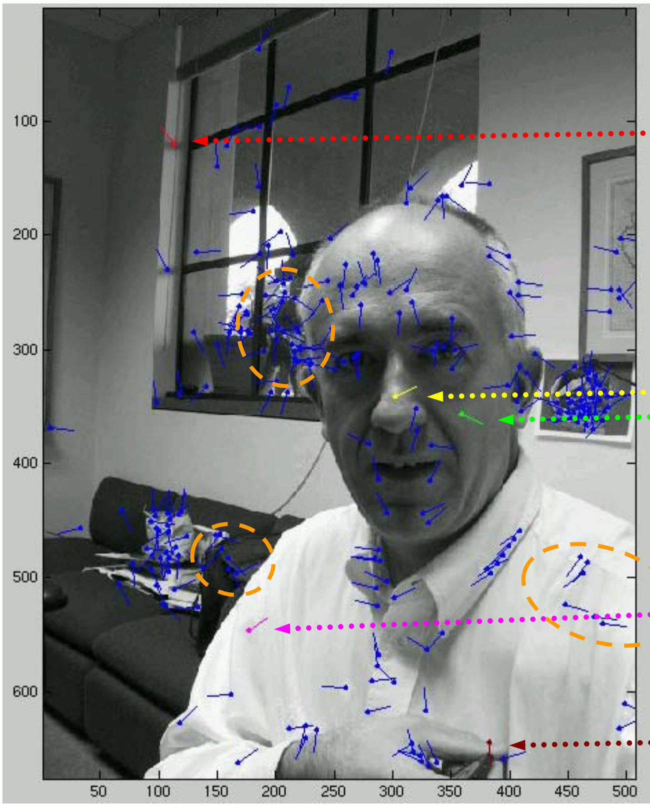


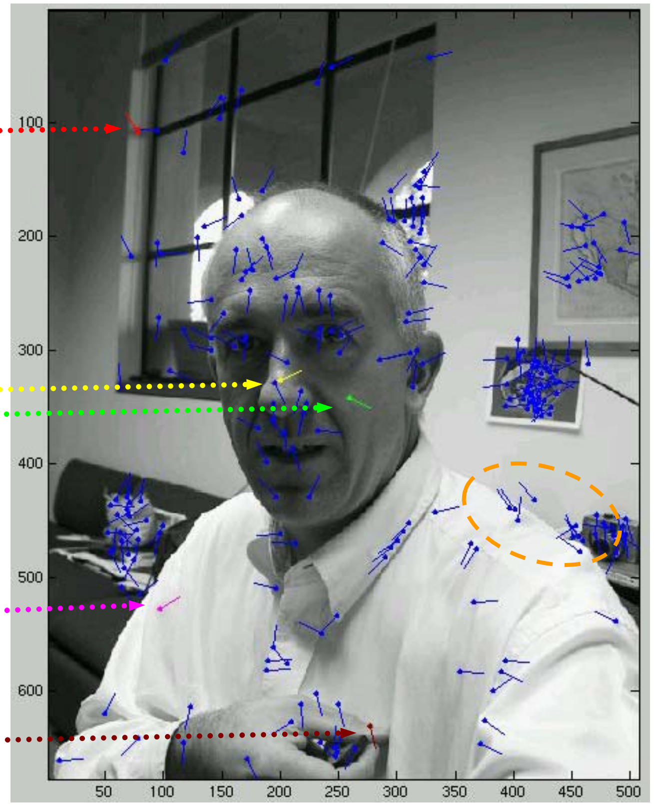
Mutual Euclidean distances in appearance space of **descriptors**

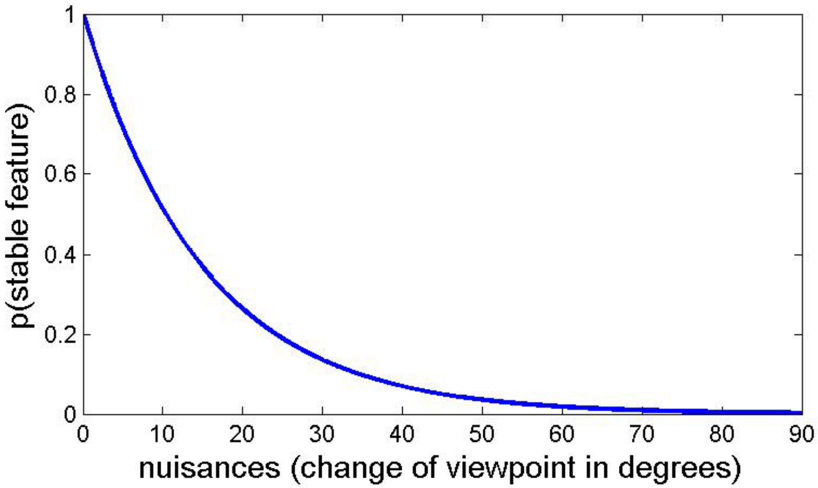
	1	2	3	4	5	6	7	8
1	0	0.1	0.1	0.1	0.7	0.5	0.5	0.8
2	0.1	0	0.1	0.2	0.7	0.6	0.5	0.9
3	0.1	0.1	0	0.0	0.5	0.4	0.5	0.8
4	0.1	0.2	0.0	0	0.5	0.4	0.5	0.9
5	0.7	0.7	0.5	0.5	0	0.5	0.8	1.0
6	0.5	0.6	0.4	0.4	0.5	0	0.4	0.9
7	0.5	0.5	0.5	0.5	0.8	0.4	0	1.0
8	0.8	0.9	0.8	0.9	1.0	0.9	1.0	0

- Pixels intensity within a patch
- Steerable filters [Freeman1991]
- SIFT [Lowe1999,2004]
- Shape context [Belongie2002]
- Spin [Johnson1999]
- HOG [Dalal2005]

Stability with respect to nuisances



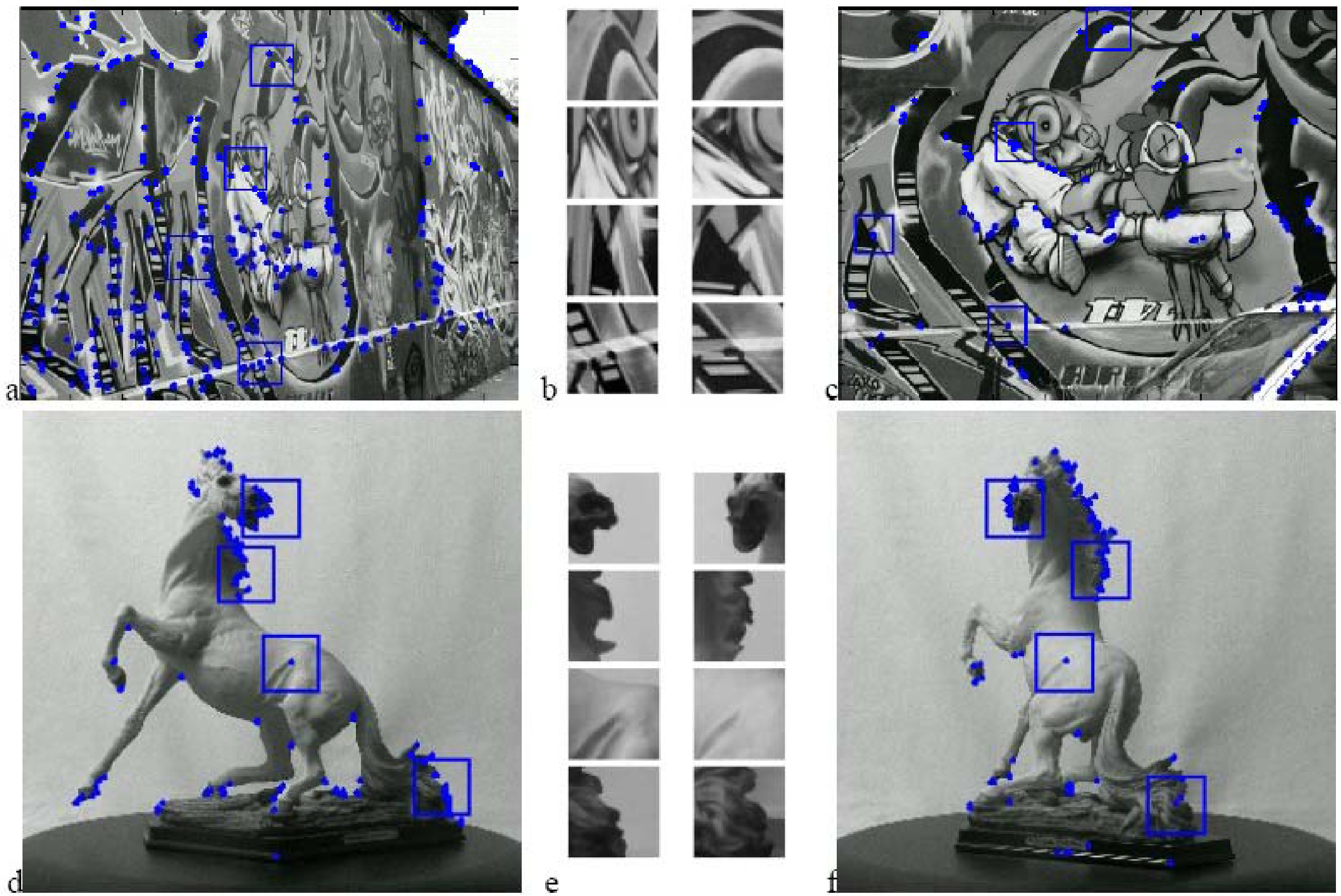




$\Rightarrow$ Which detector / descriptor combination is best for recognition ?

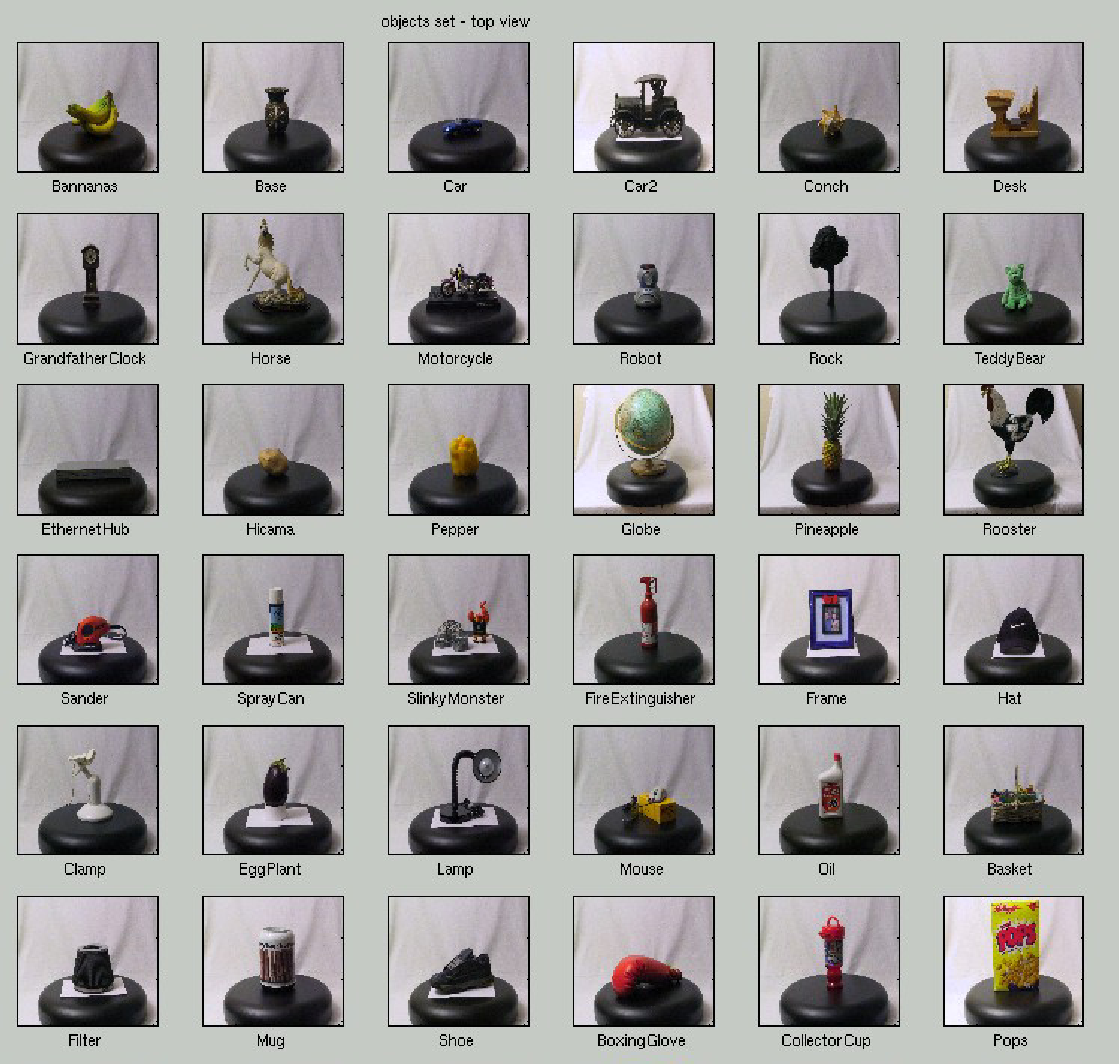
Past work on evaluation of features

- Use of flat surfaces, ground truth easily established
- In 3D images appearance changes more !

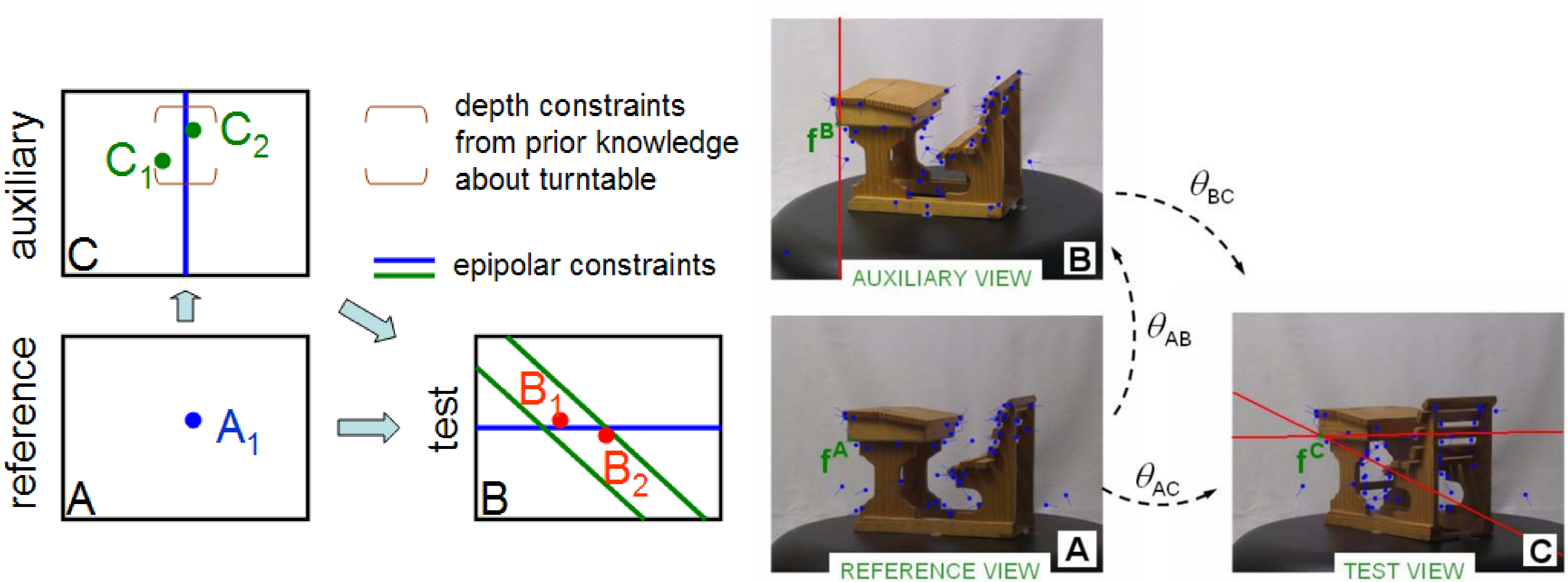


[Schmid&Mohr00] [Mikolajczyk&Schmid 03,05,05]

Database : 100 3D objects



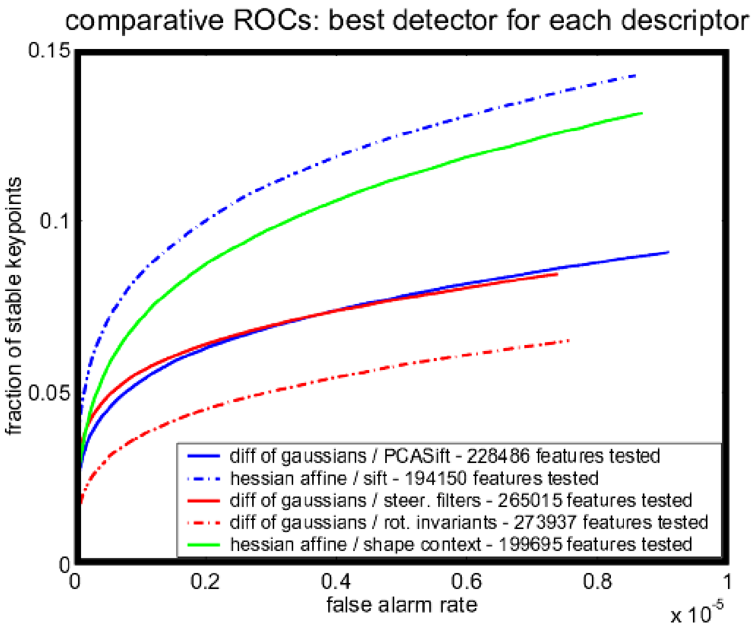
Testing setup

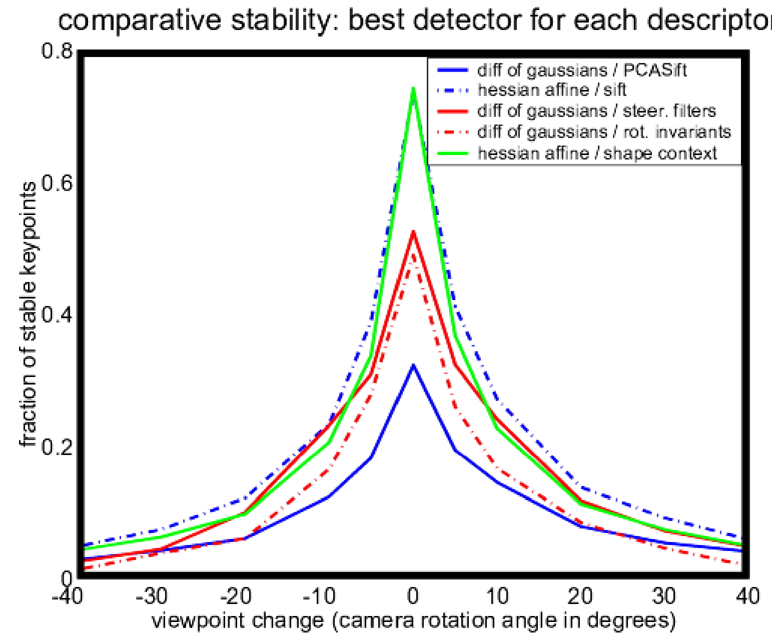


[Moreels&Perona ICCV05, IJCV07]

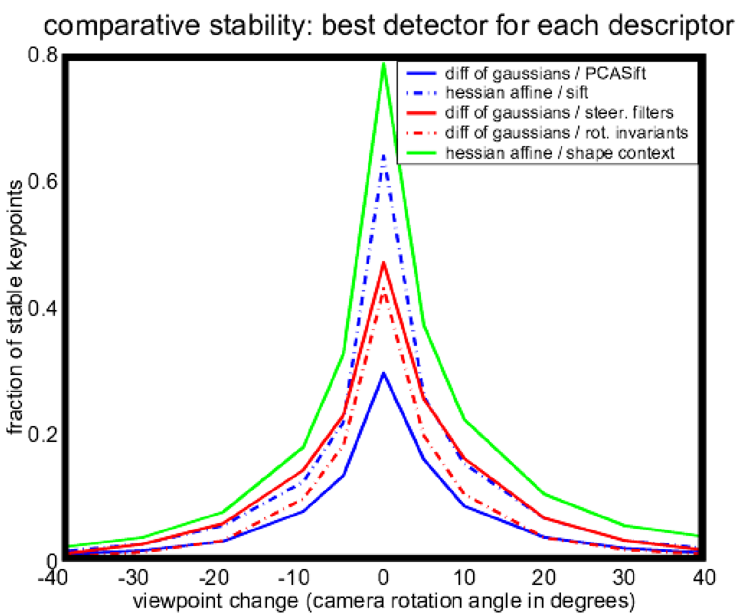
Used by [Winder, CVPR07]

Results – viewpoint change

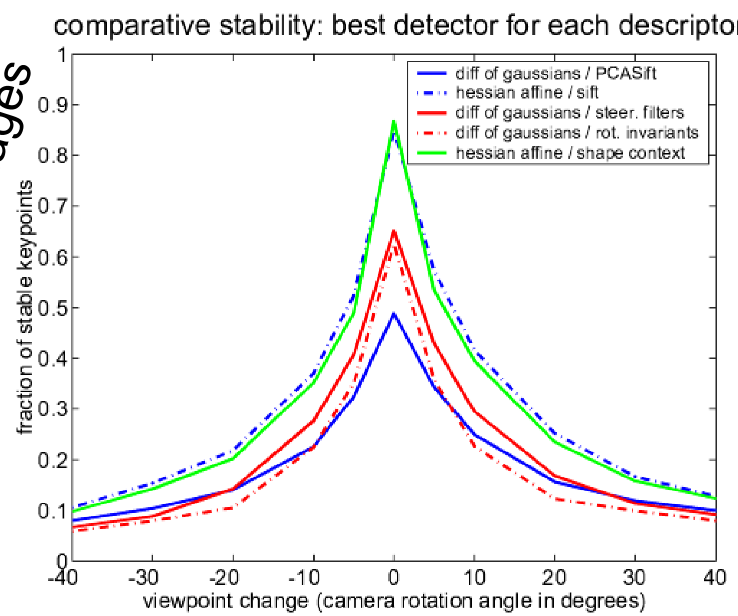




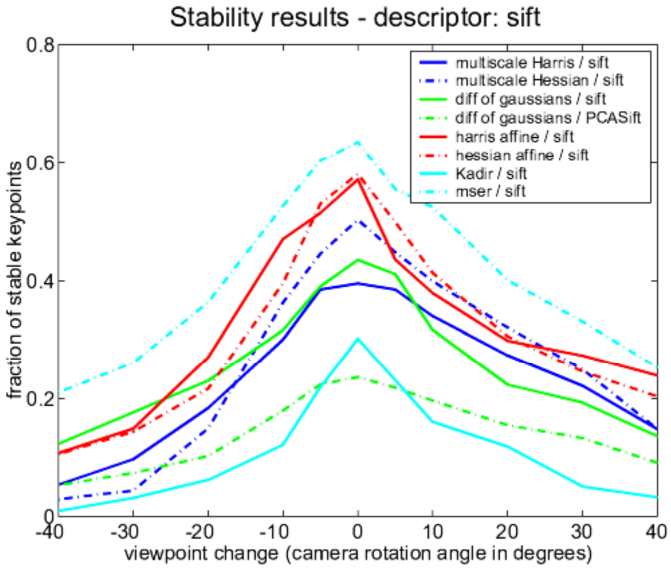
Mahalanobis distance

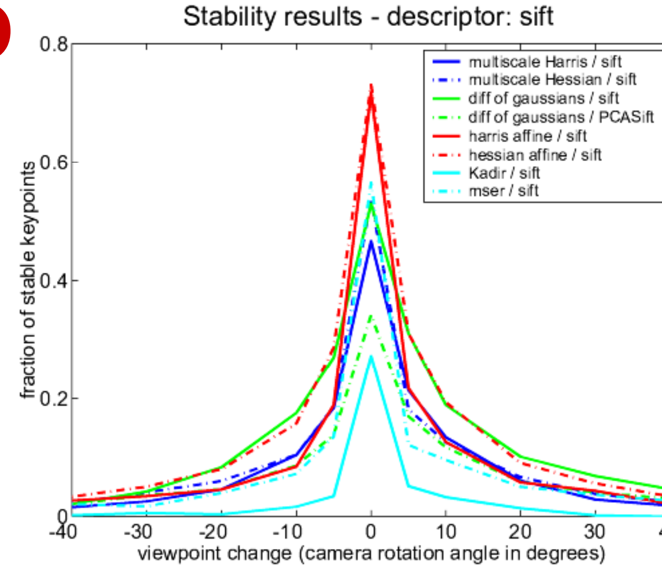


No 'background' images

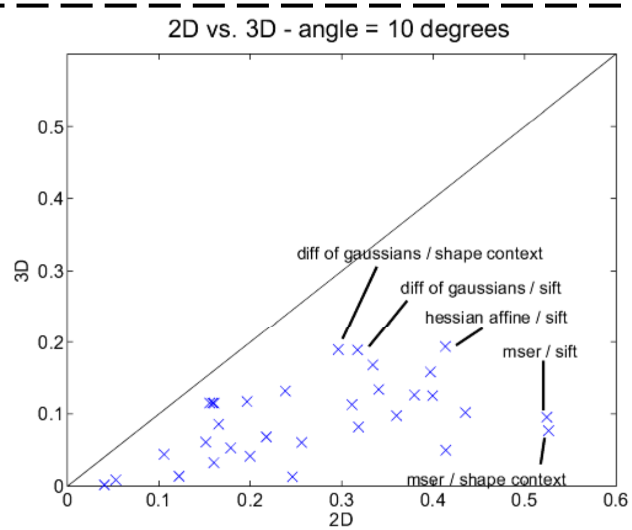


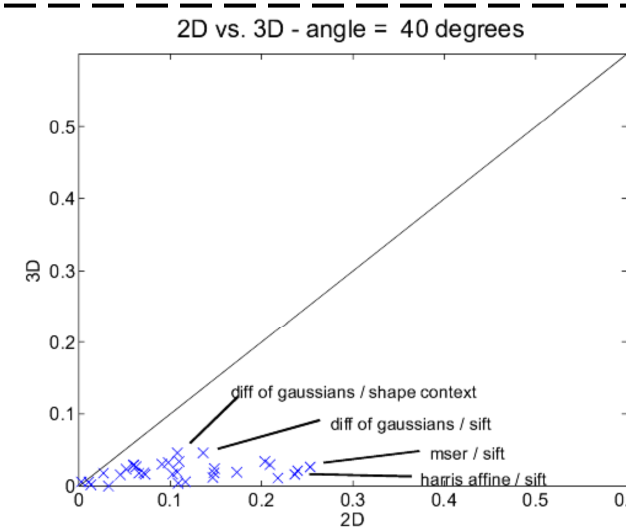
2D vs. 3D





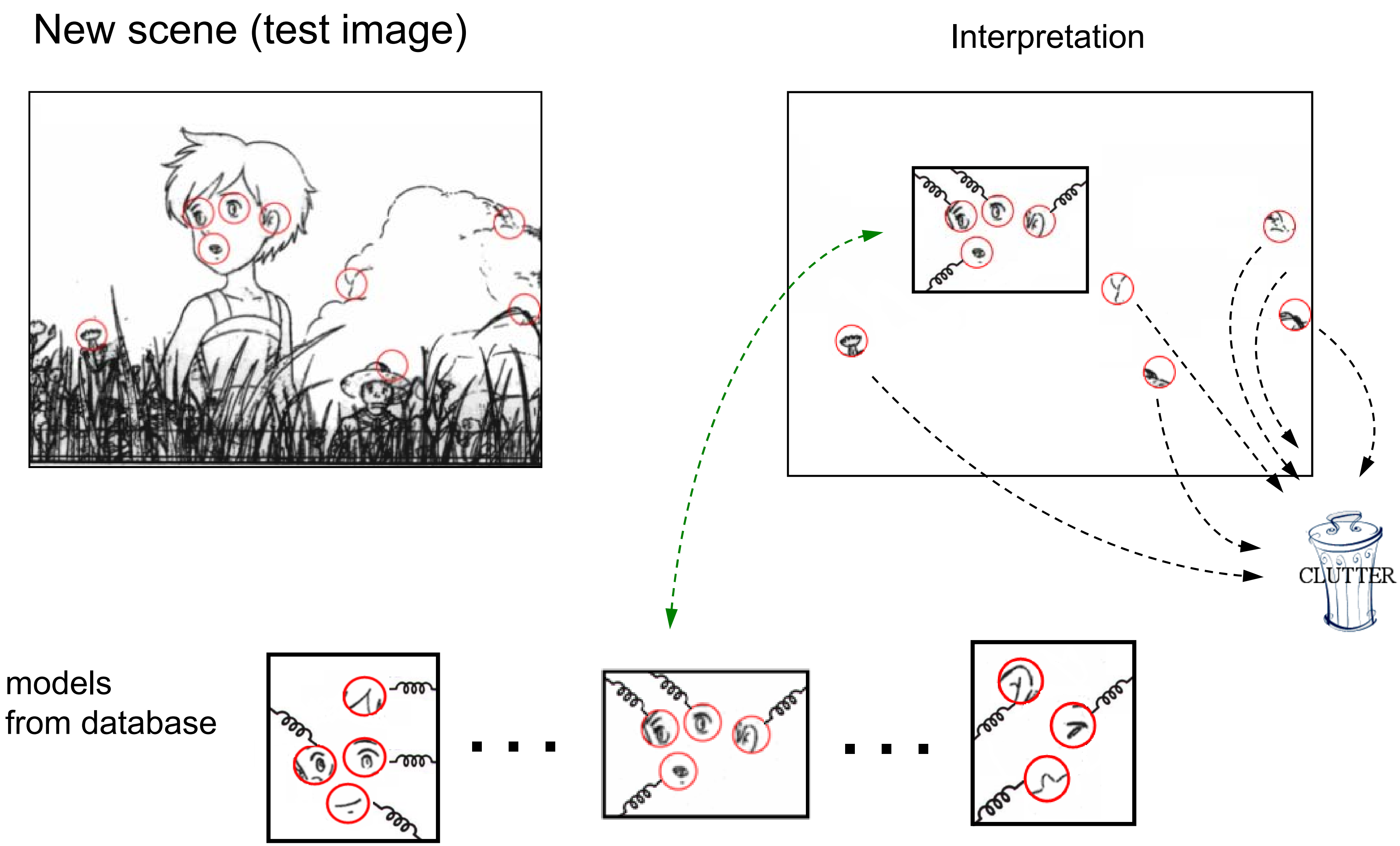
Ranking of detectors/descriptors combinations are modified when switching from 2D to 3D objects





Features matching algorithm

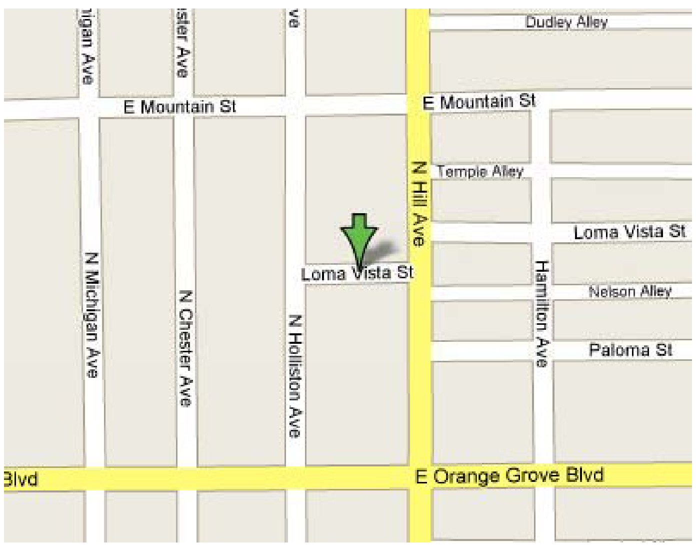
Features assignments

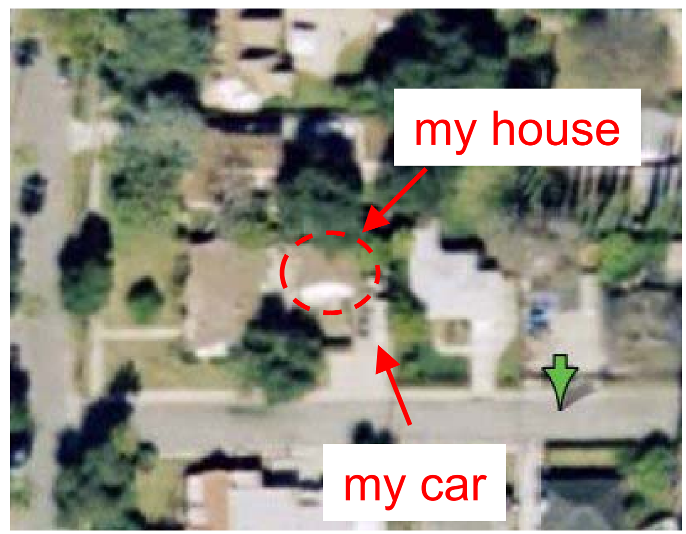


Coarse-to-fine strategy

- We do it every day !

Search for my place : Los Angeles area – Pasadena – Loma Vista - 1351





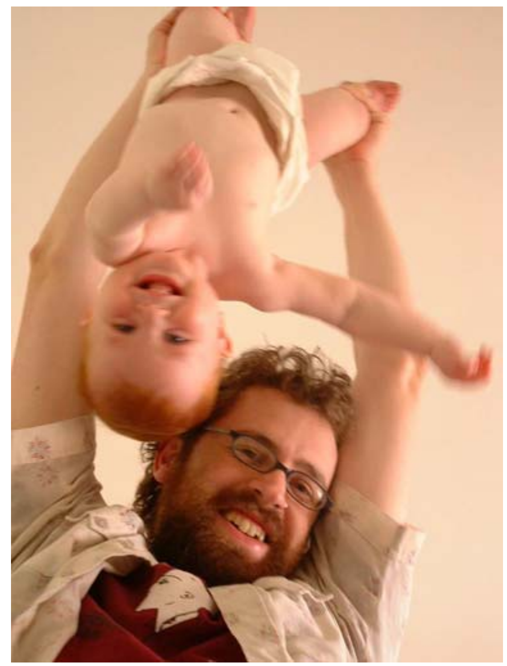
Coarse-to-Fine detection

- Progressively narrow down focus on correct region of hypothesis space
- Reject with little computation cost irrelevant regions of search space
- Use first information that is easy to obtain
- Simple building blocks organized in a cascade
- Probabilistic interpretation of each step

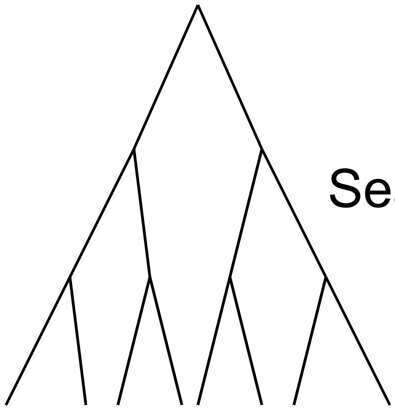
Coarse data : prior knowledge

- Which objects are likely to be there, which pose are they likely to have ?

unlikely
situations



Model voting



Search tree (**appearance** space – leaves = database features)

New scene (test image)

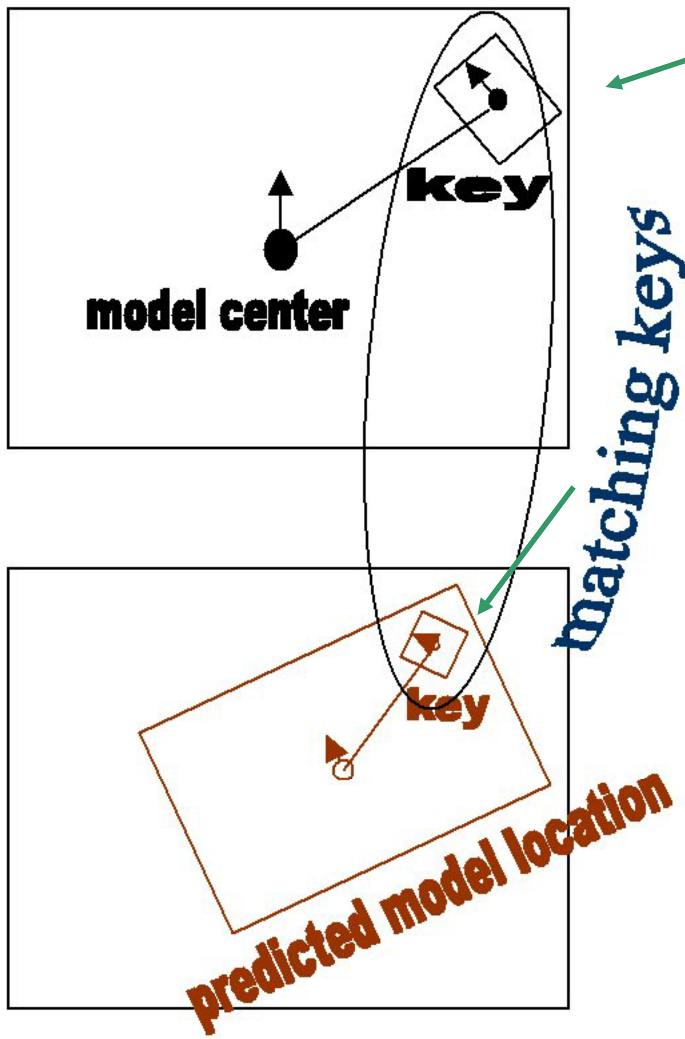
4 votes

2 votes

0 vote

⋮ Models from database

Use of rich geometric information



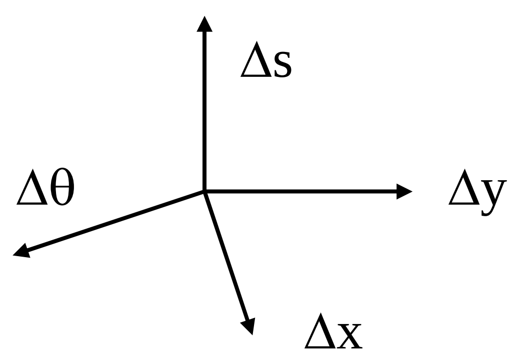
(x_1,y_1,s_1,θ_1)

(x_2,y_2,s_2,θ_2)

Transform predicted by this match:

- $\Delta x = x_2 - x_1$
- $\Delta y = y_2 - y_1$
- $\Delta s = s_2 / s_1$
- $\Delta\theta = \theta_2 - \theta_1$

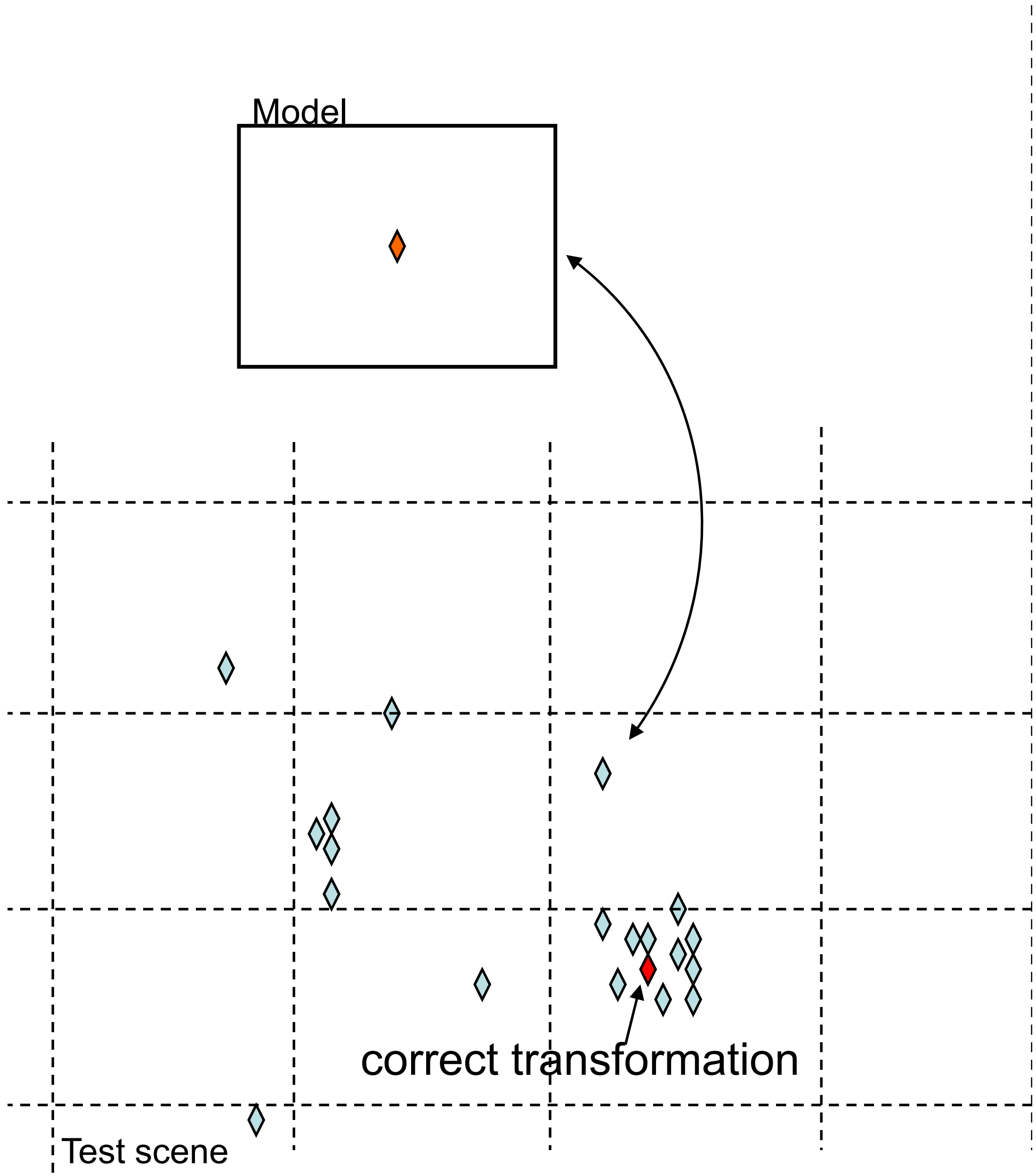
Each match is represented by a dot in the space of 2D similarities (Hough space)



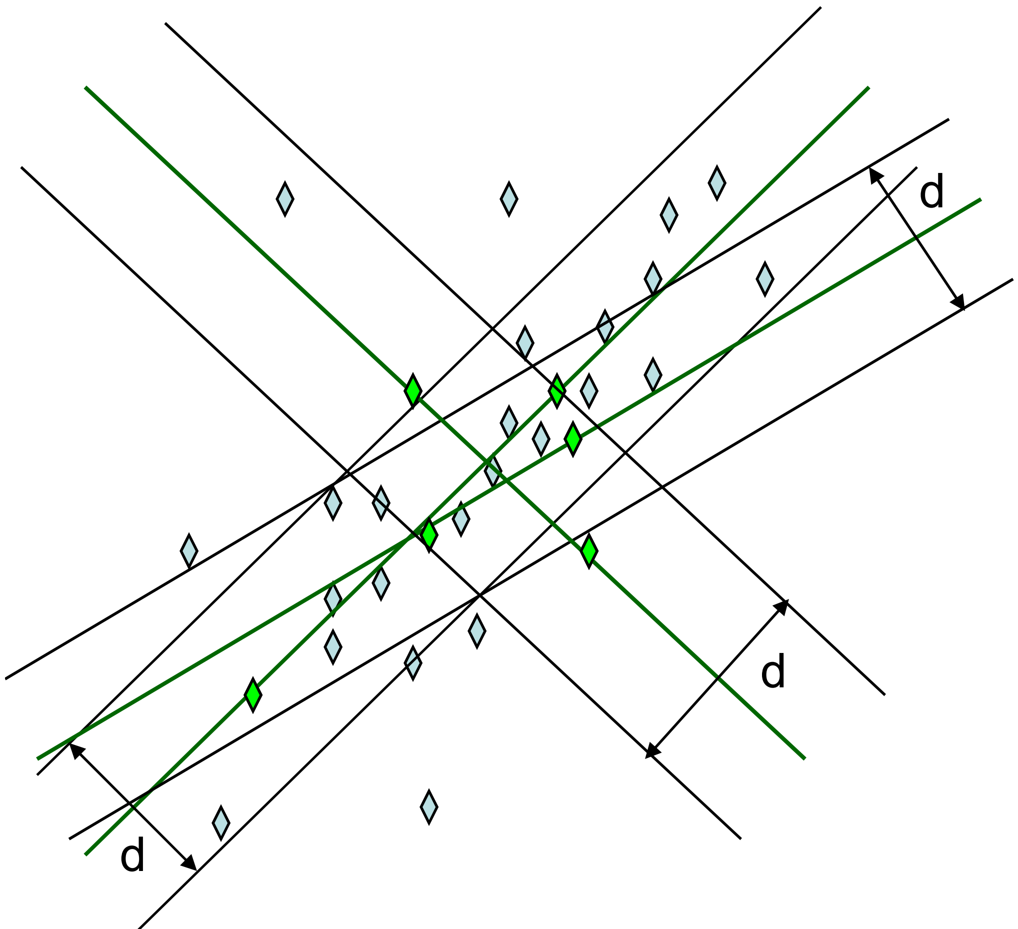
[Lowe1999,2004]

Coarse Hough transform

- Prediction of position of model center after transform
- The space of transform parameters is discretized into 'bins'
- Coarse bins to limit boundary issues and have a low false-alarm rate for this stage
- We count the number $\tilde{N}$ of votes collected by each bin.



Correspondence or clutter ? PROSAC



- Similar to RANSAC – robust statistic for parameter estimation
- Priority to candidates with good **quality** of appearance match
- 2D affine transform : 6 parameters
  $\Rightarrow$ each sample contains 3 candidate correspondences.

[Fischler 1973]
[Chum&Matas 2005]

Output of PROSAC : pose transformation
+ set of features correspondences

Probabilistic model

Score of an extended hypothesis

Features assignments

Hypothesis: model + position

observed features geometry + appearance

database of models

$$P(H, V | F, M) = \frac{P(F, H, V | M)}{P(F | M)}$$

constant

Votes per model

Votes per model pose bin (Hough transform)

$$P(F, H, V | M) = P(H|M) \cdot P(\bar{N}|H, M) \cdot P(\tilde{N}|\bar{N}, H, M) \cdot P(V|\tilde{N}, \bar{N}, H, M) \cdot P(F|V, \tilde{N}, \bar{N}, H, M)$$

Prior on model and poses

Prior on assignments (before actual observations)

Consistency (after PROSAC)

Consistency

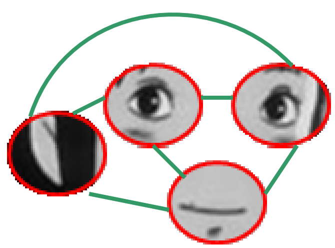
Consistency between observations and predictions from hypothesis

$$P(F|V,\tilde{N},\bar{N},H,M) = \prod_{V(i)\neq 0} p_{fg}(f_i|H,f_{V(i)}) \cdot \prod_{V(i)=0} p_{bg}(f_i)$$

Common-frame approximation : parts are conditionally independent once reference position of the object is fixed. [Lowe1999,Huttenlocher90,Moreels04]

Constellation model

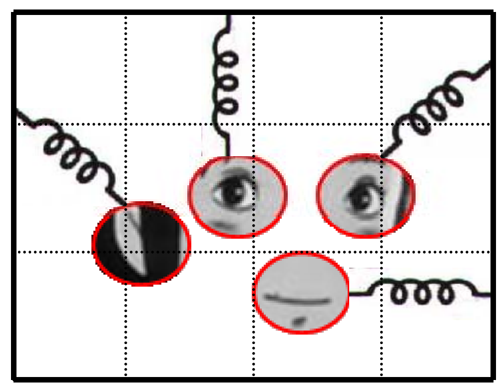
model m



$$p_m(x_1 \ldots x_N)$$

$$D.O.F. = O(Parts^2)$$

Common-frame



$$\prod_i p_m(x_i|\Theta_m)$$

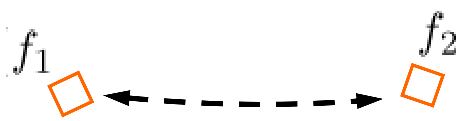
position of model m

$$D.O.F. = O(Parts)$$

Consistency

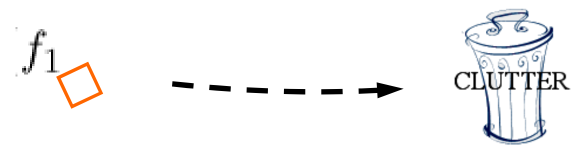
Consistency between observations and predictions from hypothesis

$$P(F|V,\tilde{N},\bar{N},H,M) = \prod_{V(i)\neq 0} p_{fg}(f_i|H,f_{V(i)}) \cdot \prod_{V(i)=0} p_{bg}(f_i)$$



foreground features

appearance geometry



‘null’ assignments

appearance geometry

$$p_{fg}(f_i|H,f_{V(i)}) = p_{fg,\mathcal{A}}(\mathcal{A}|H,\mathcal{A}_{V(i)}) \cdot p_{fg,\mathcal{X}}(\mathcal{X}|H,\mathcal{X}_{V(i)})$$

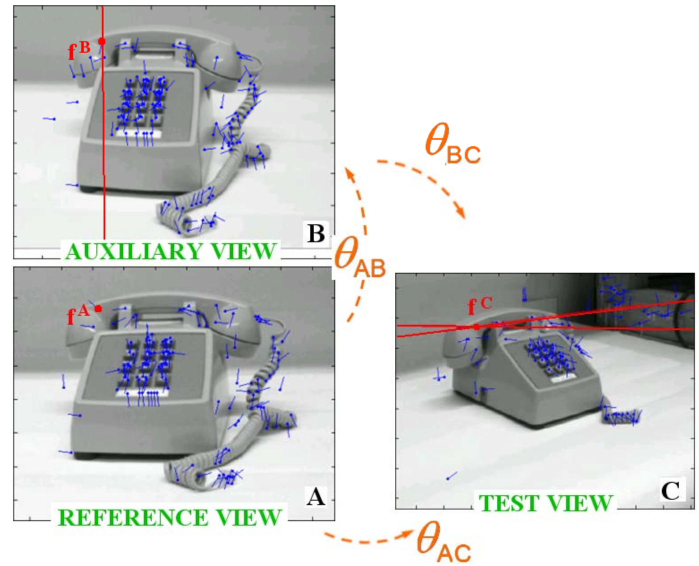
Consistency - appearance Consistency - geometry

Learning foreground & background densities

- Ground truth pairs of matches are collected

- Gaussian densities, centered on the nominal value that appearance / pose should have according to H

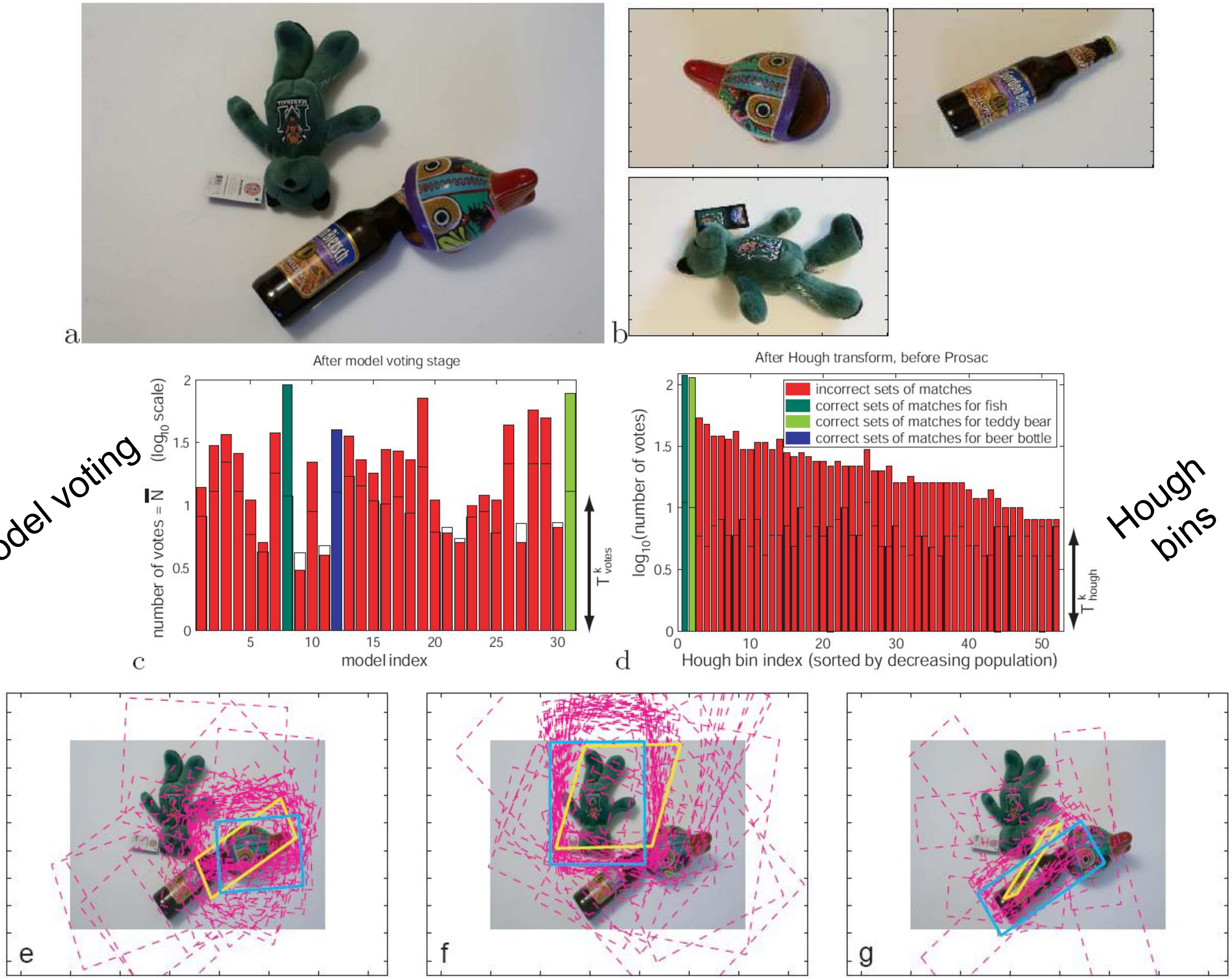
- Learning background densities is easy: match to random images.



[Moreels&Perona, IJCV, 2007]

Experiments

An example

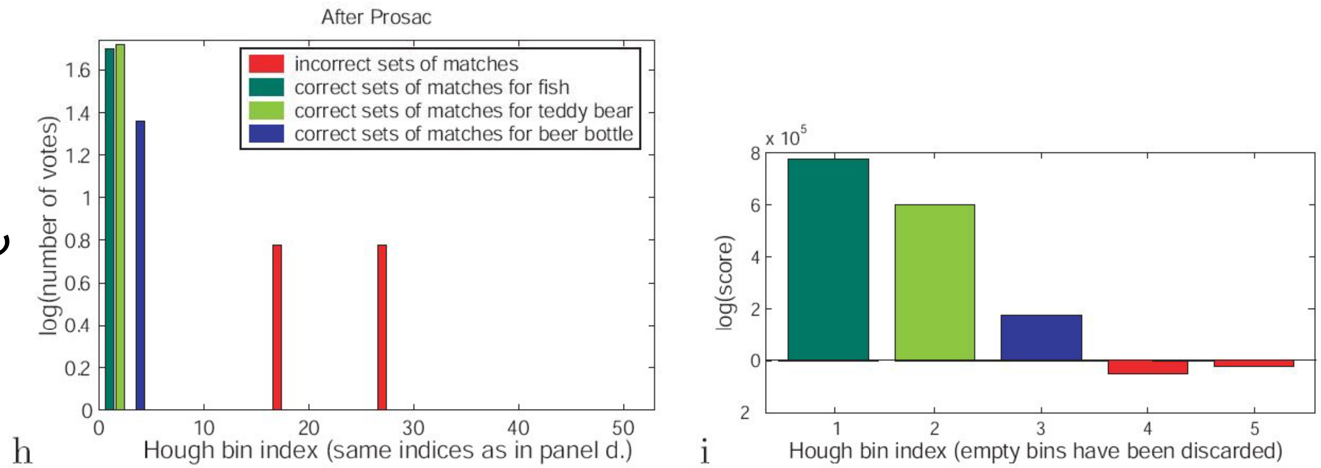


Model voting

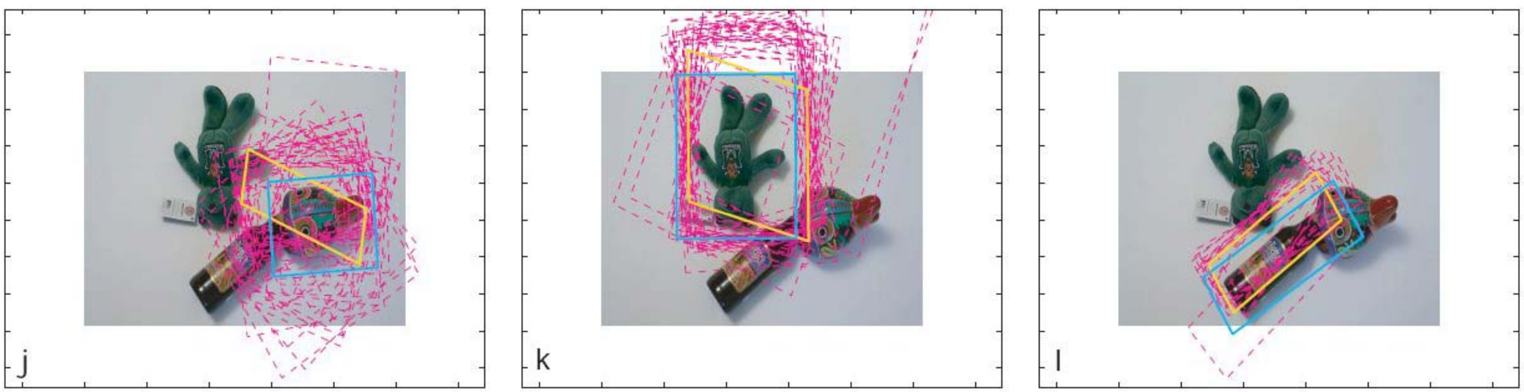
Hough bins

An example

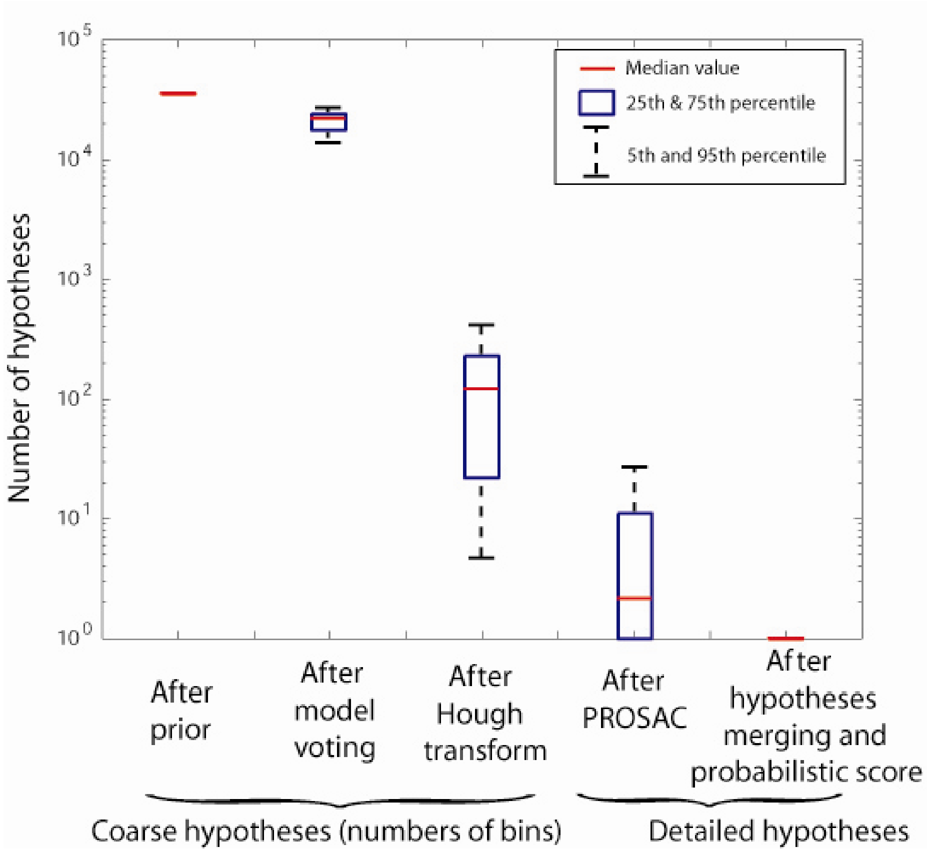
After PROSAC

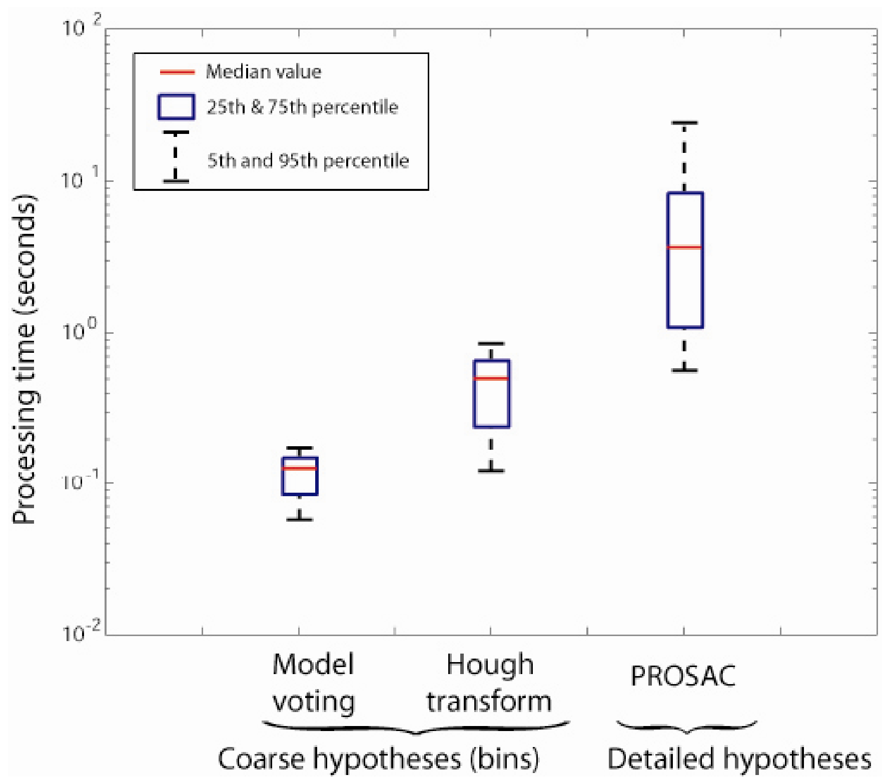


Probabilistic scores



Efficiency of coarse-to-fine processing





Giuseppe Toys database – Models

61 objects, 1-2 views/object

Giuseppe Toys database – Test scenes

141 test scenes

Home objects database – Models

49 objects, 1-2 views/object

Home objects database – Test scenes

141 test scenes

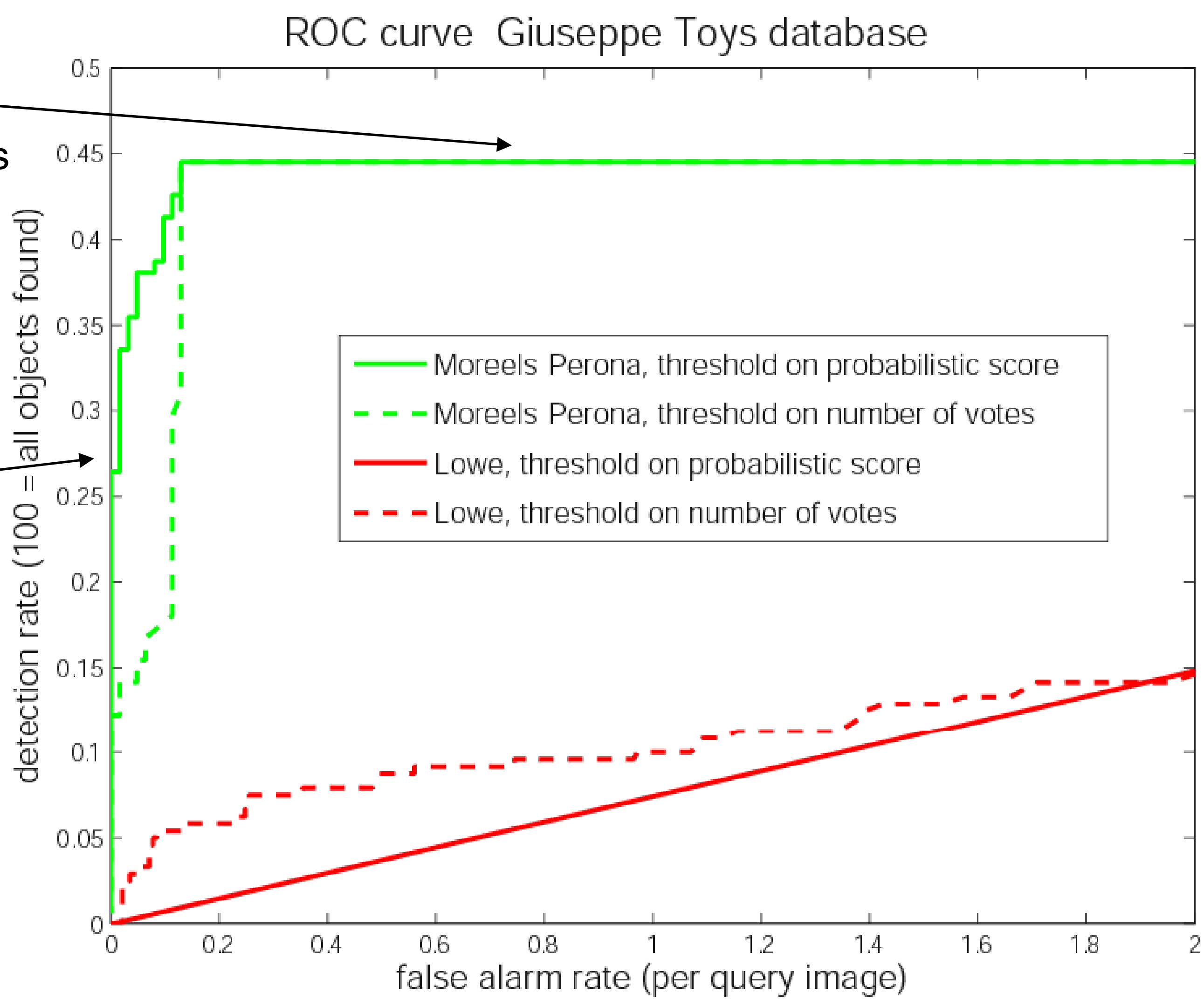
Results – Giuseppe Toys database

-

undetected objects:
features with poor appearance distinctiveness index to incorrect models

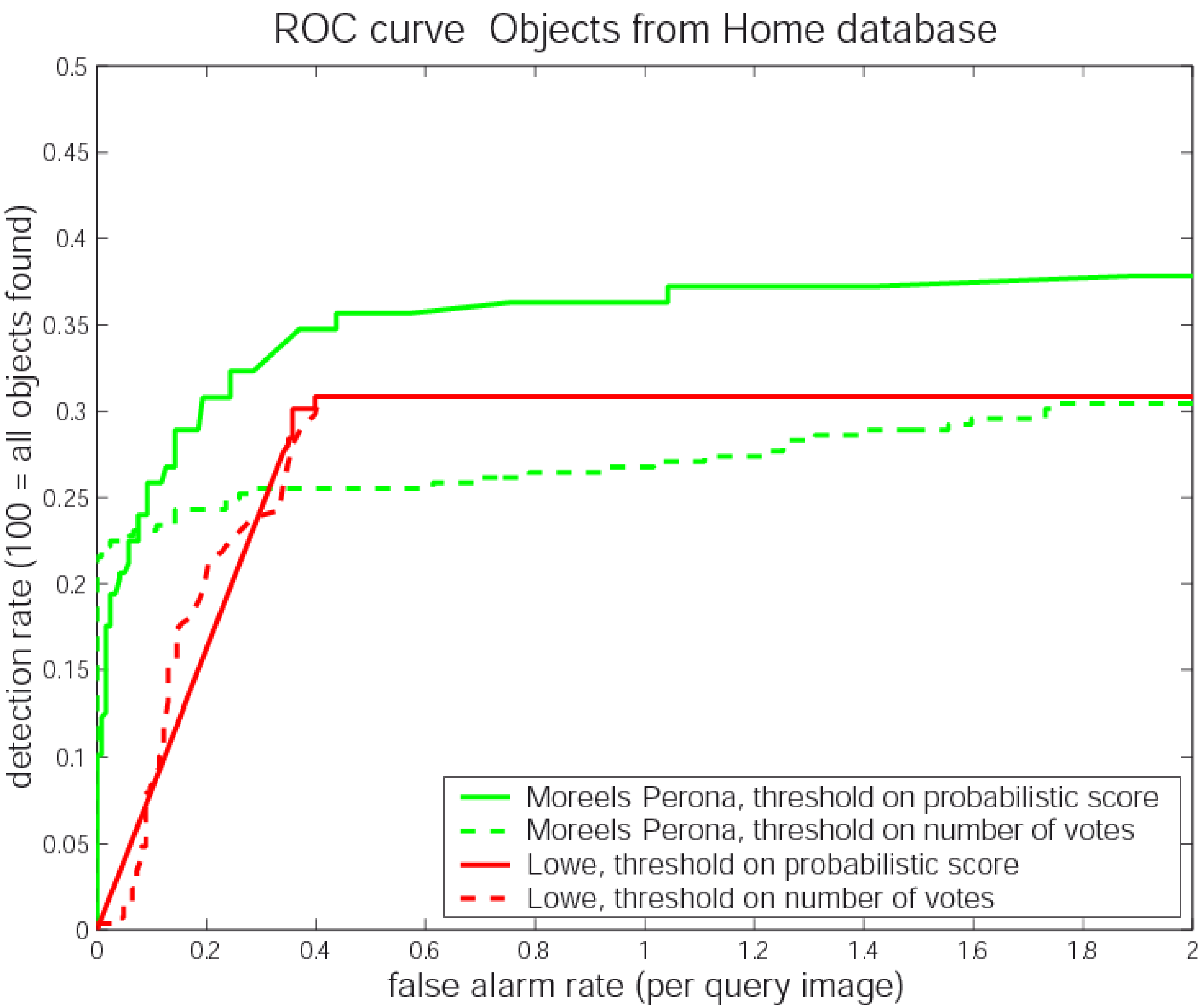
+

Lower false alarm rate
- more systematic verification of geometry consistency
- more consistent verification of geometric consistency



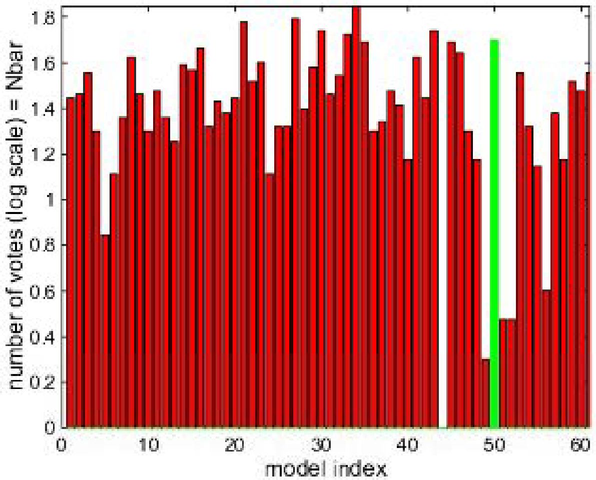
Lowe'99,'04

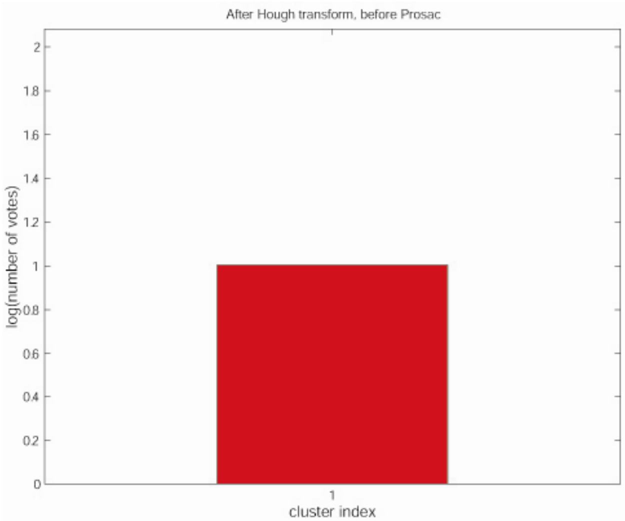
Results – Home objects database



Failure mode

Test image hand-labeled
before the experiments

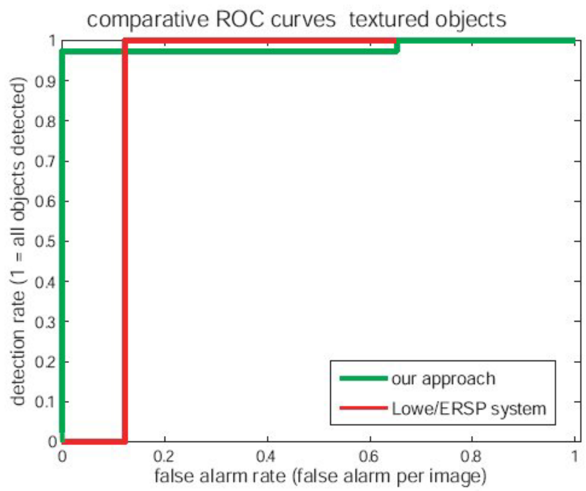




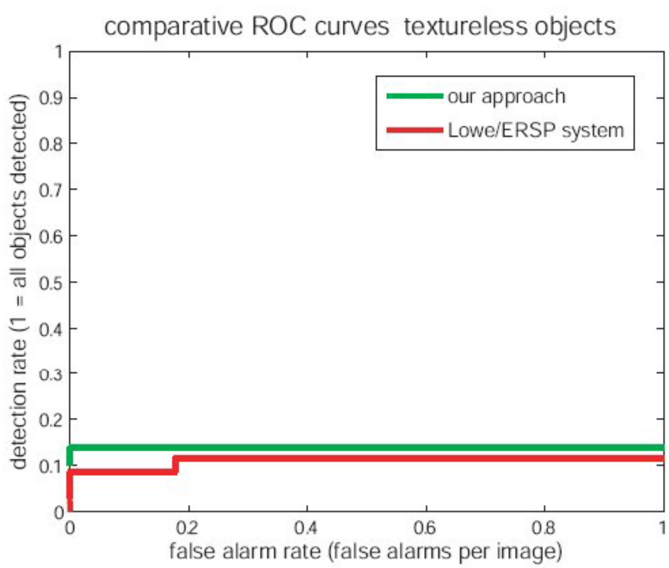
Test – Text and graphics

Training

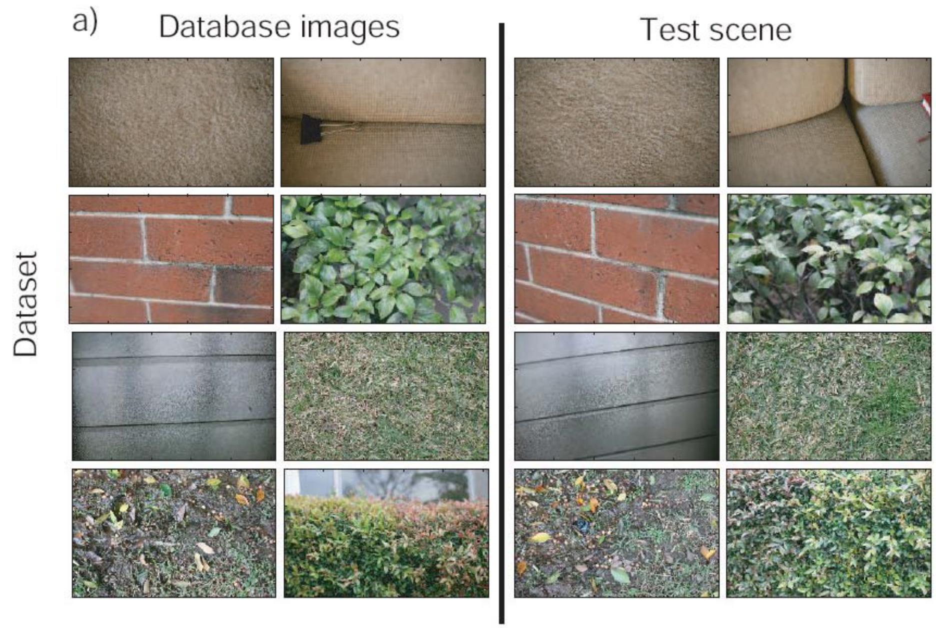
Test

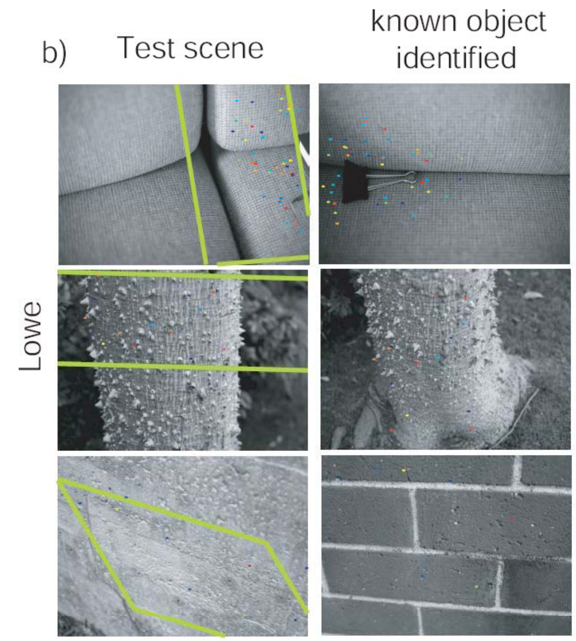


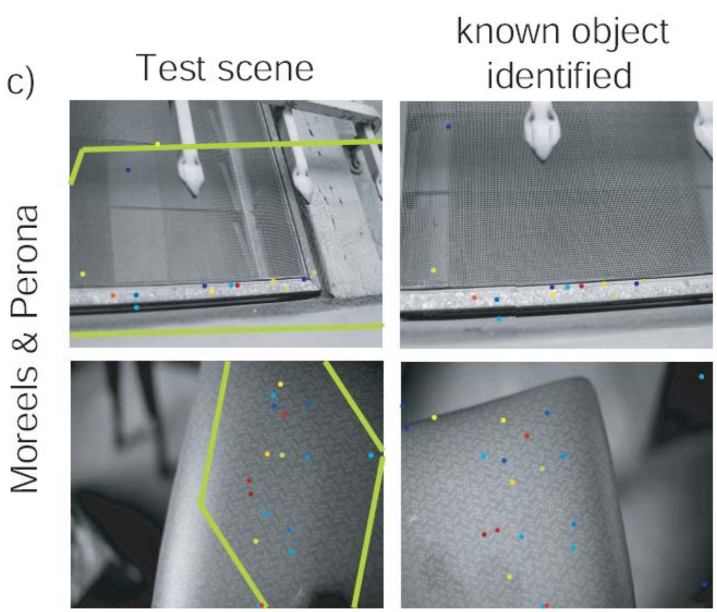
Test – no texture



Test – Clutter







d)

	same training texture		different training texture	
	Lowe	Moreels	Lowe	Moreels
false alarms	111	14	30	12
>30 matches	61	3	3	3
wrong texture	11	4	30	12

Conclusions

- Coarse-to-fine strategy prunes irrelevant search branches at early stages.
- Probabilistic interpretation of each step.
- Higher performance than Lowe, especially in cluttered environment.
- Front end (features) needs more work for smooth or shiny surfaces.

Today

Sudderth guest lecture:

- Constellation Models (Fergus)
- Unsupervised Object Discovery with pLSA (Sivic)
- Scene Models (Li)
- Transformed Models (Sudderth)

Daphna B. student presentation:

- pLSA models of activity (Neibles)

Moreels guest lecture:

- A probabilistic formulation of voting / SIFT (Moreels)