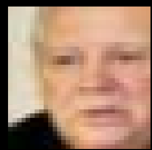
Objects and scenes:

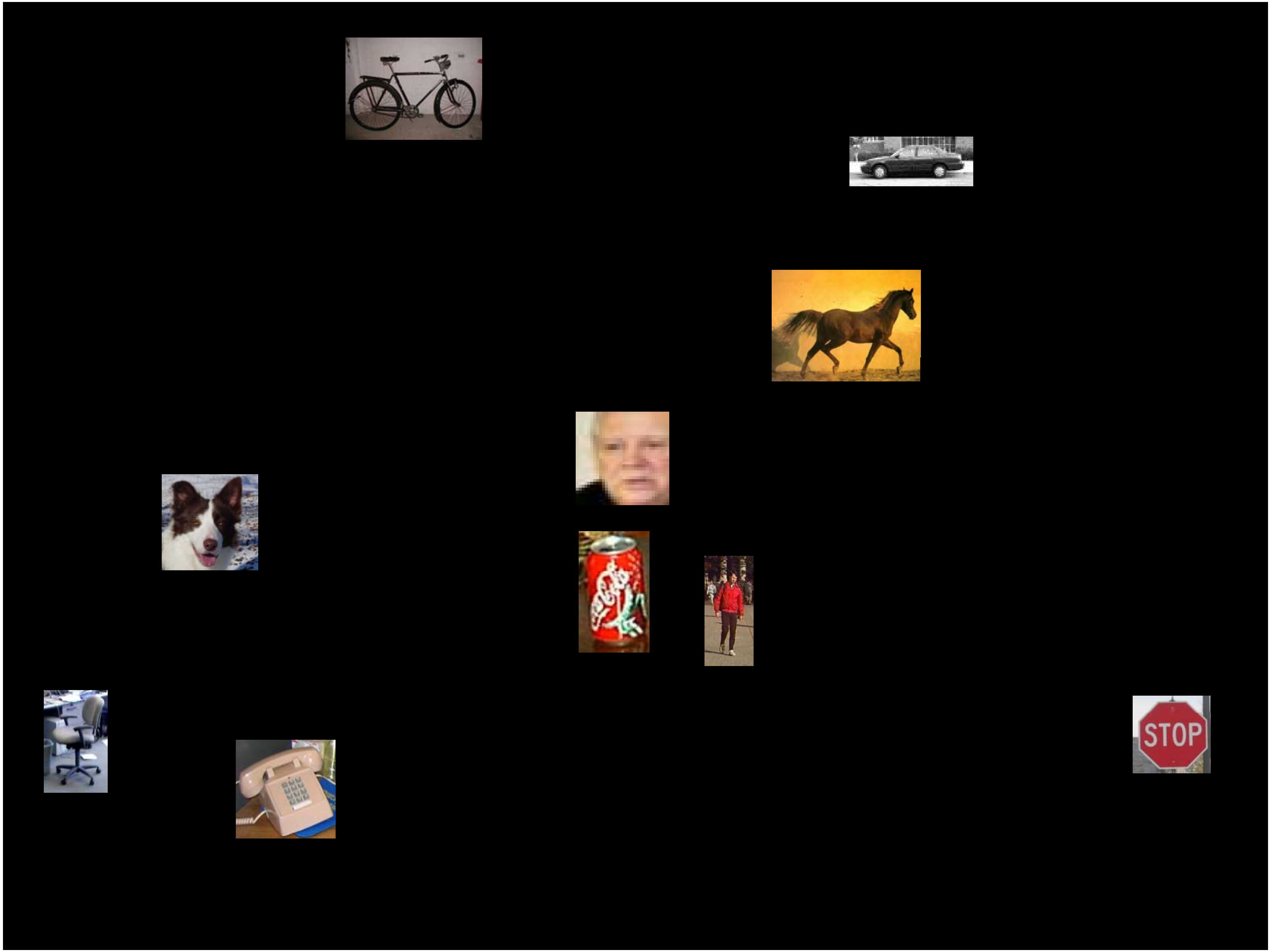
Recognizing Multiple Object Classes

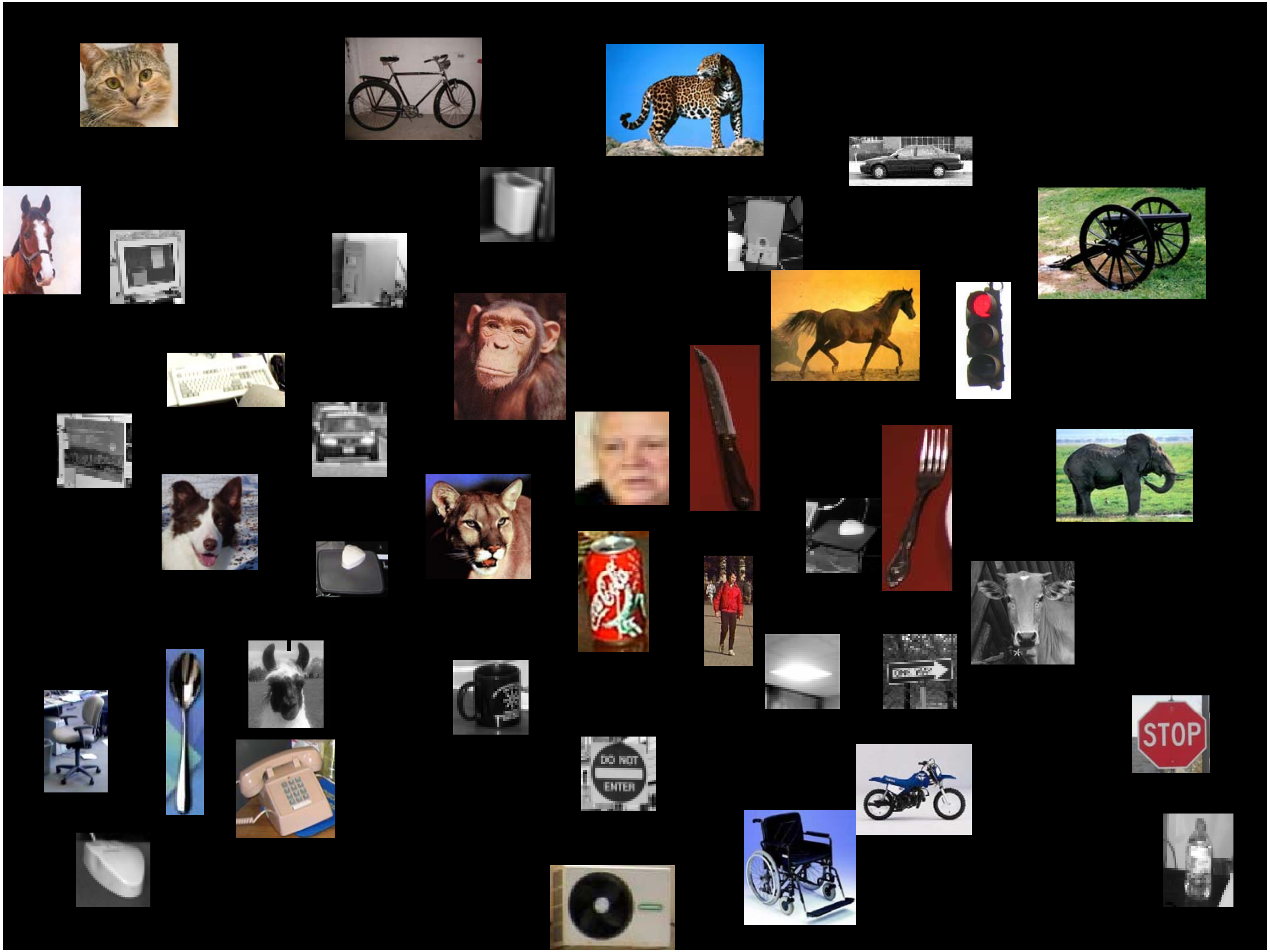
Josef Sivic and Ivan Laptev
http://www.di.ens.fr/~josef
INRIA, WILLOW, ENS/INRIA/CNRS UMR 8548
Laboratoire d’Informatique, Ecole Normale Supérieure, Paris

With slides from: A. Torralba, D. Hoiem, D. Ramanan and others.

Multiclass object detection




STOP


DO NOT
ENTER
STOP

Context: objects appear in configurations

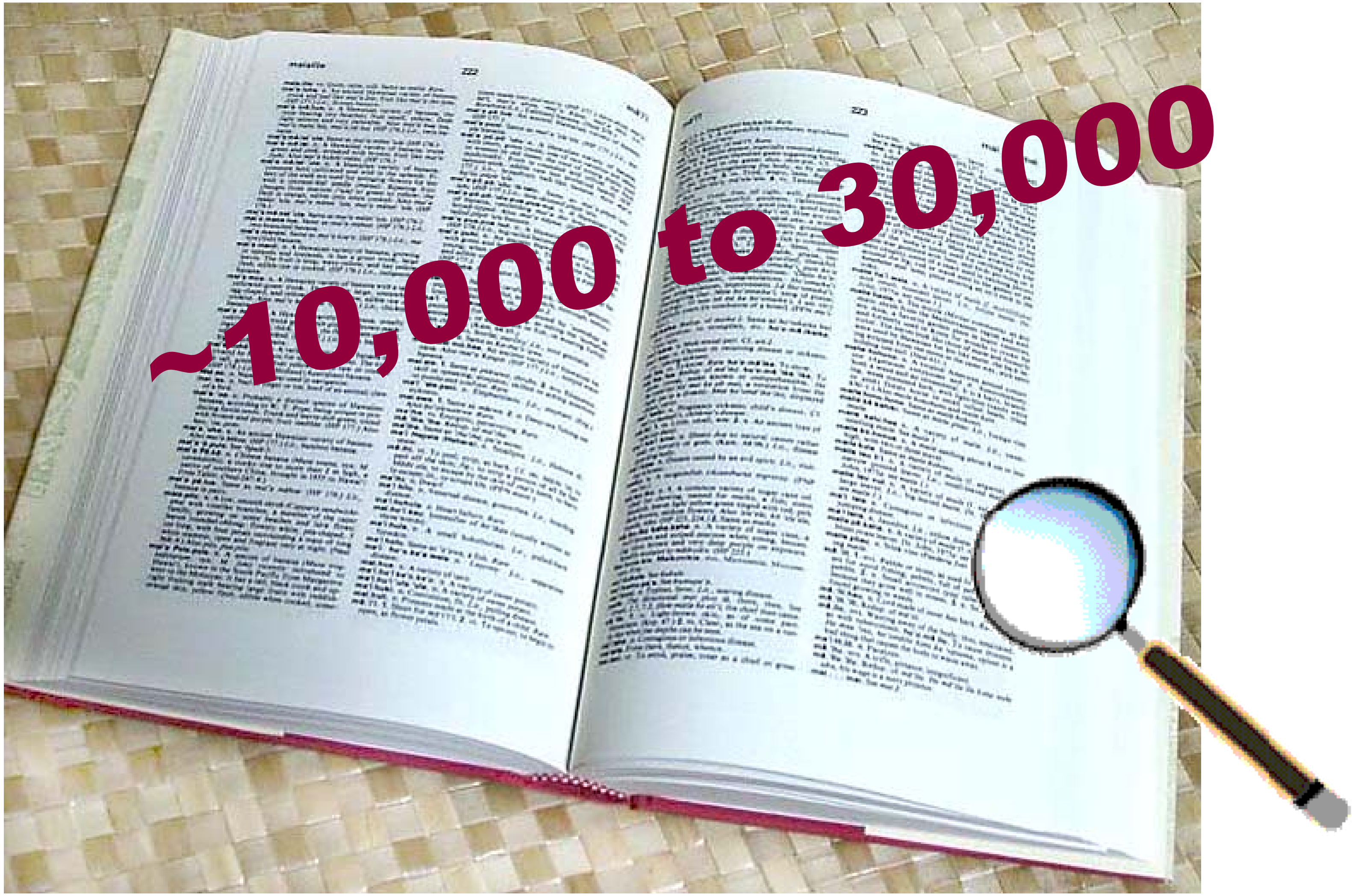
Generalization: objects share parts

How many categories?

How many categories?

Slide by Aude Oliva

How many object categories are there?



Biederman 1987

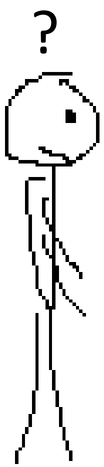
How many categories?

- Probably this question is not even specific enough to have an answer

Which level of categorization is the right one?

Car is an object composed of:

a few doors, four wheels (not all visible at all times), a roof, front lights, windshield



If you are thinking in buying a car, you might want to be a bit more specific about your categorization level.

Entry-level categories
(Jolicoeur, Gluck, Kosslyn 1984)

- Typical member of a basic-level category are categorized at the expected level
- Atypical members tend to be classified at a subordinate level.

A bird

An ostrich

We do not need to recognize the exact category

A new class can borrow information from similar categories

So, where is computer vision?

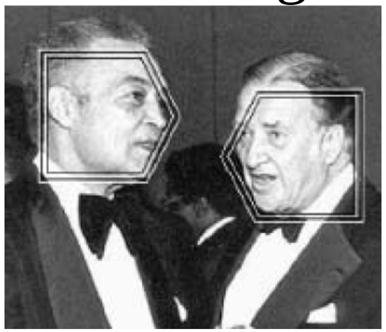
Well…

Multiclass object detection
the not so early days

Multiclass object detection
the not so early days

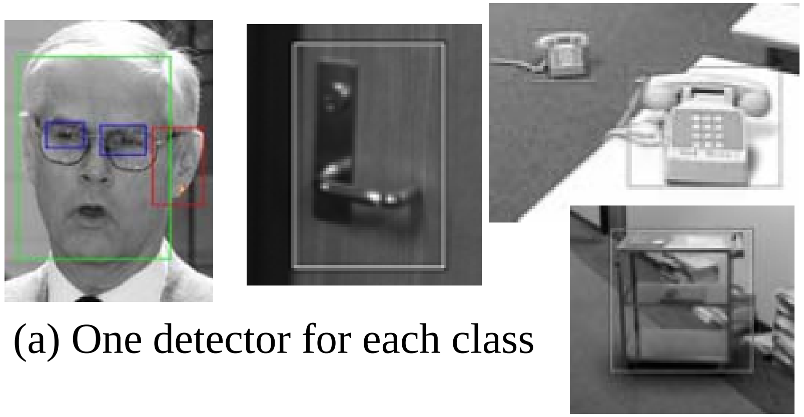
Using a set of independent binary classifiers was a common strategy:

- Viola-Jones extension for dealing with rotations



- two cascades for each view

- Schneiderman-Kanade multiclass object detection

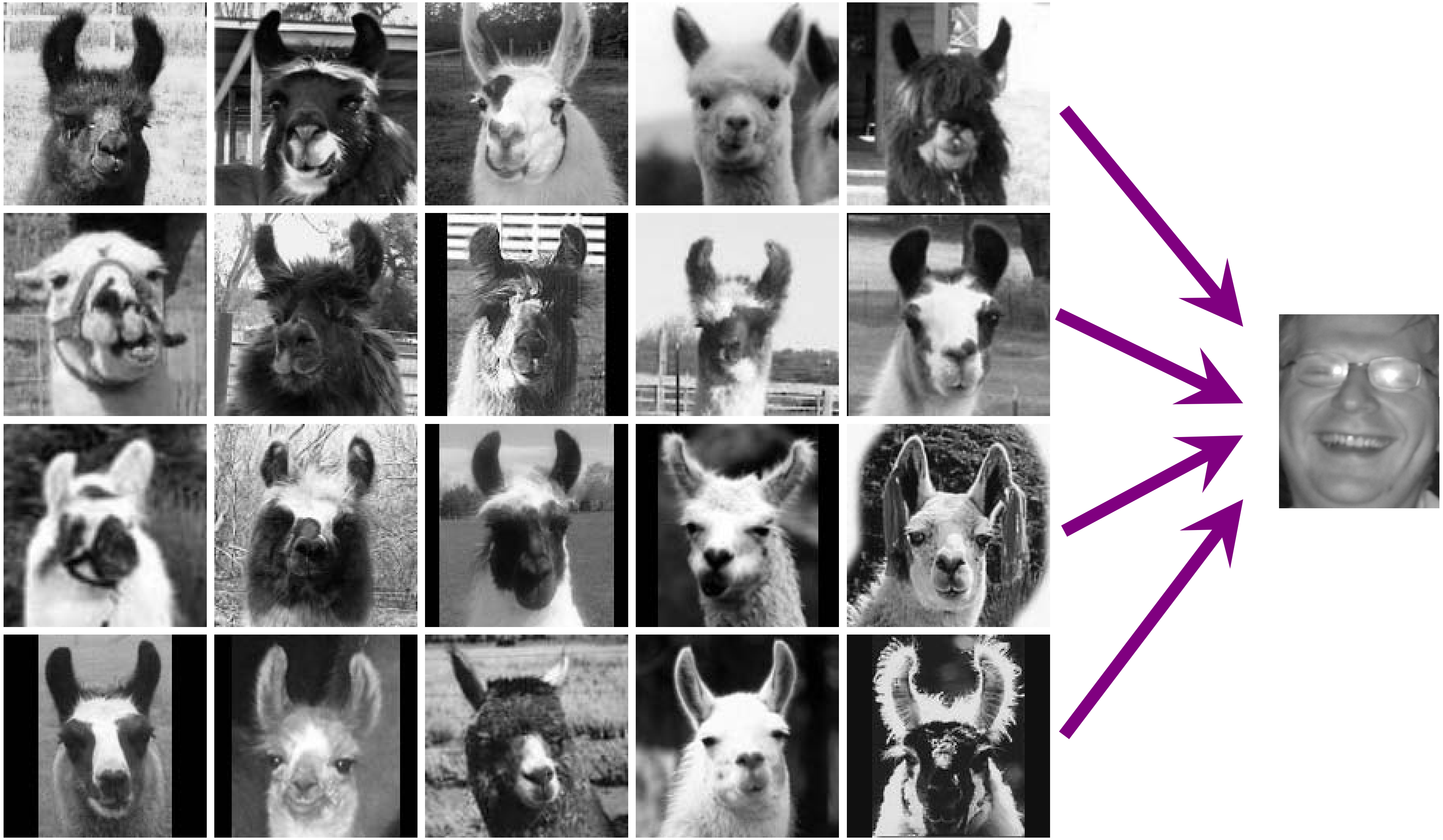


(a) One detector for each class

(b) For cars, classifiers are trained on 8 viewpoints

There is nothing wrong with this approach if you have access to lots of training data and you do not care about efficiency.

Generalizing Across Categories



Can we transfer knowledge from one object category to another?

Slide by Erik Sudderth

Shared features

- Is learning the object class 1000 easier than learning the first?

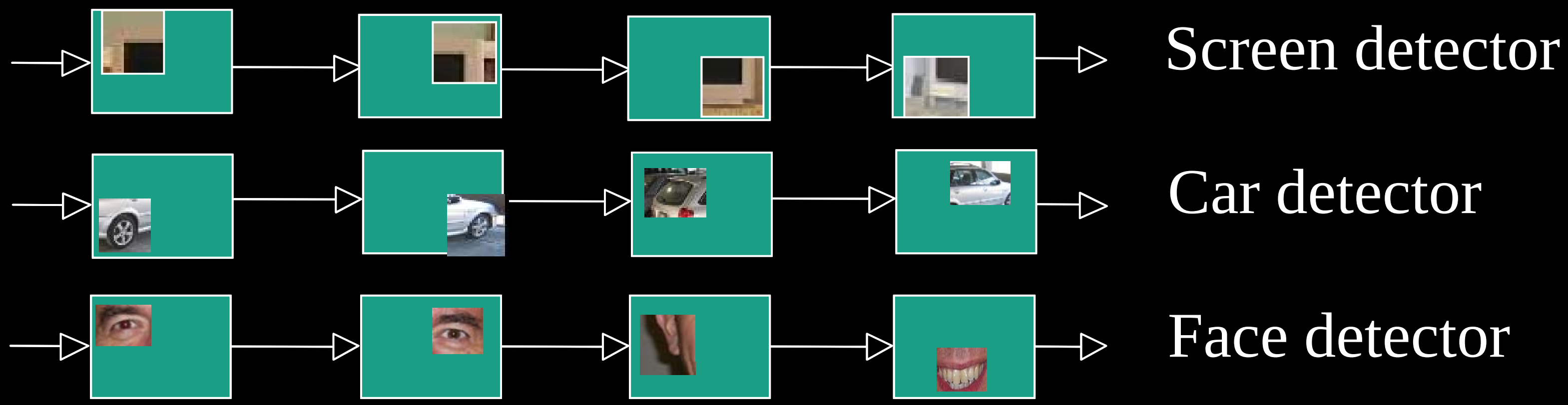


…

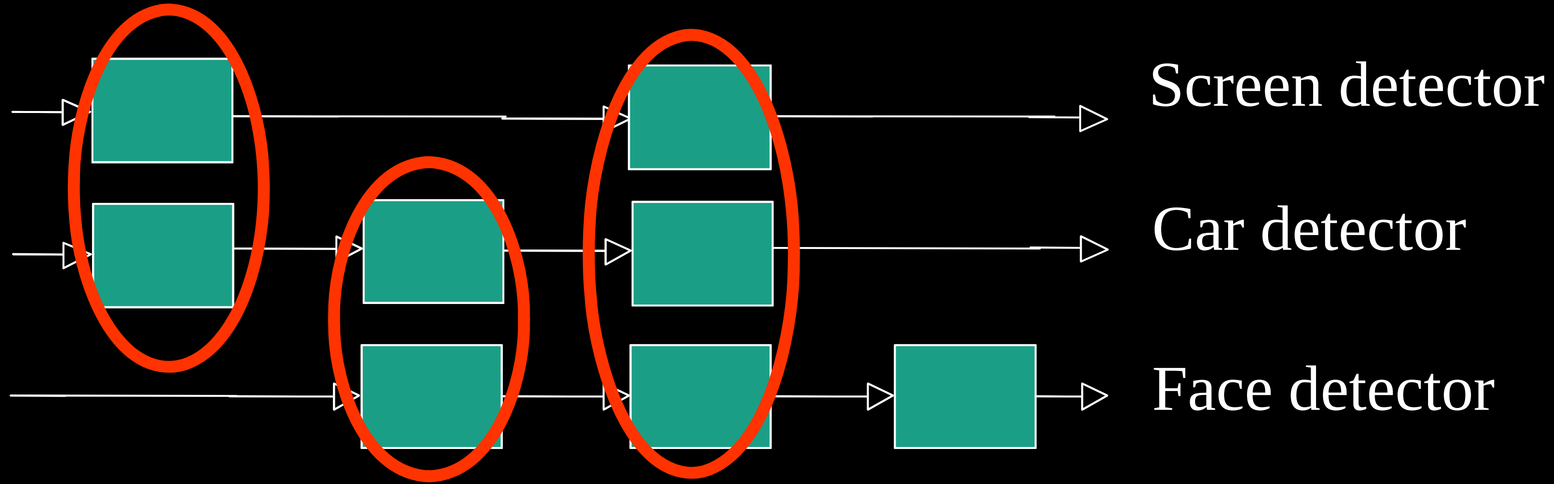

- Can we transfer knowledge from one object to another?
- Are the shared properties interesting by themselves?

Additive models and boosting

- Independent binary classifiers:

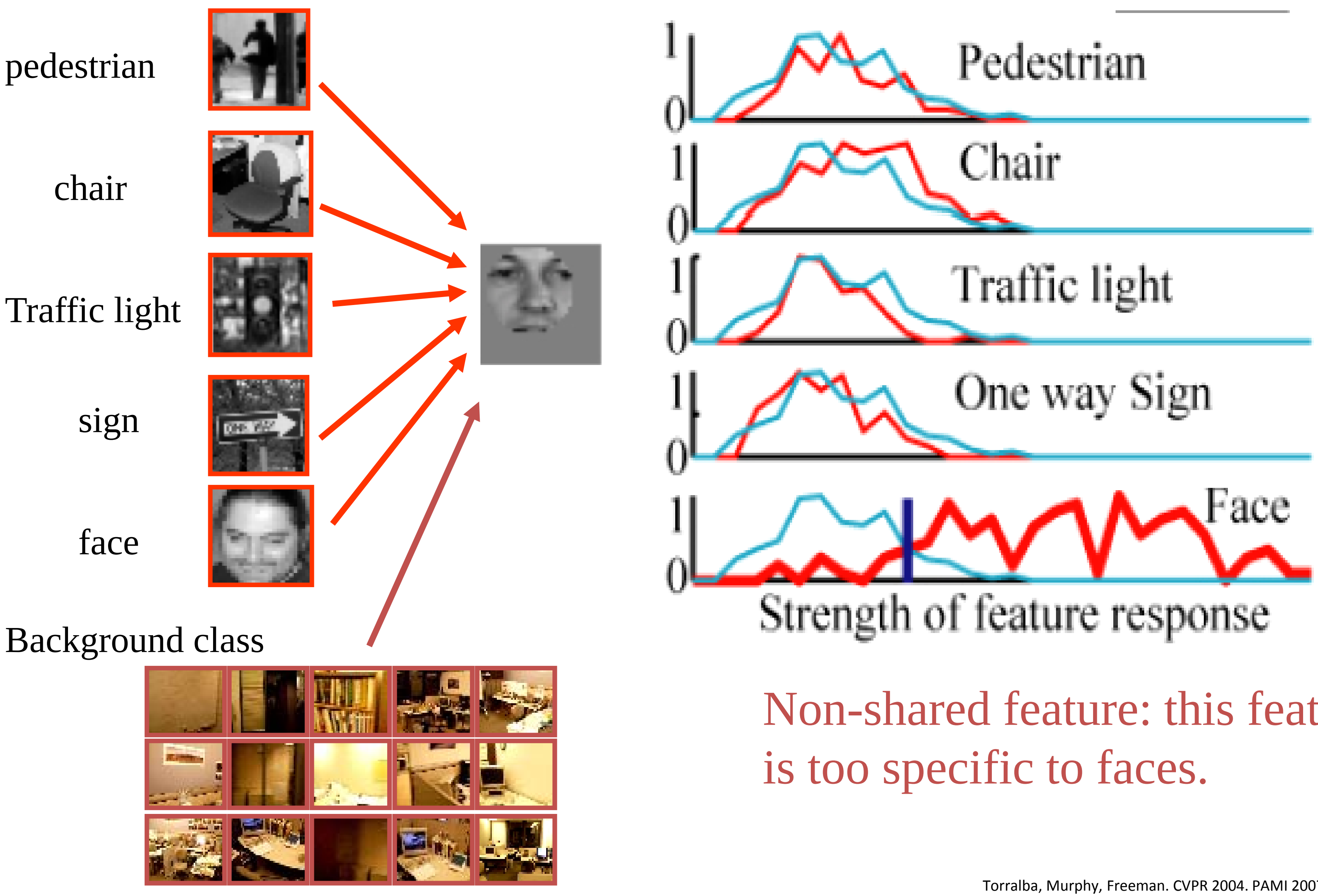


- Binary classifiers that share features:



Torralba, Murphy, Freeman. CVPR 2004. PAMI 2007

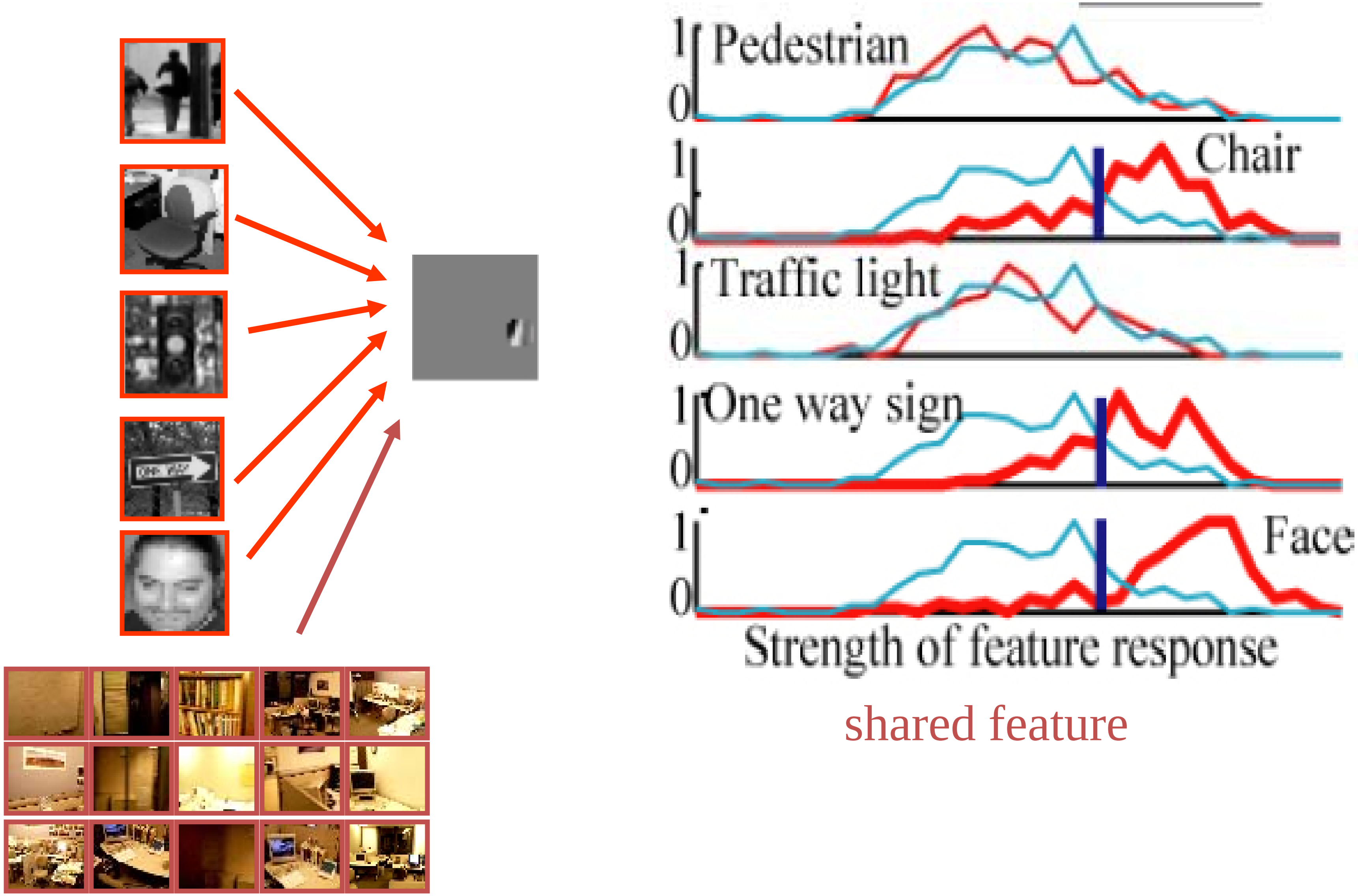
Specific feature



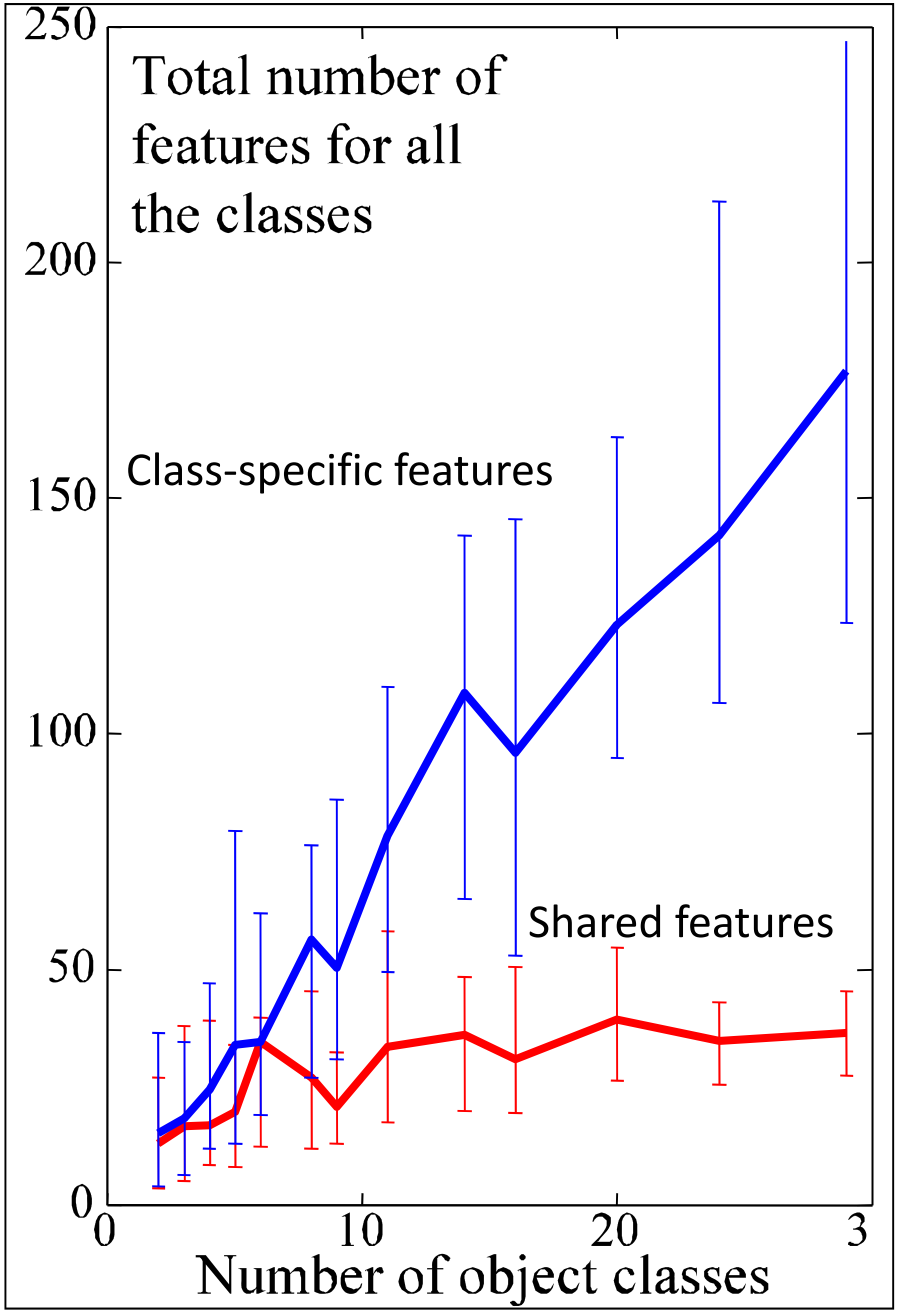
Non-shared feature: this feature is too specific to faces.

Torralba, Murphy, Freeman. CVPR 2004. PAMI 2007

Shared feature



Torralba, Murphy, Freeman. CVPR 2004. PAMI 2007



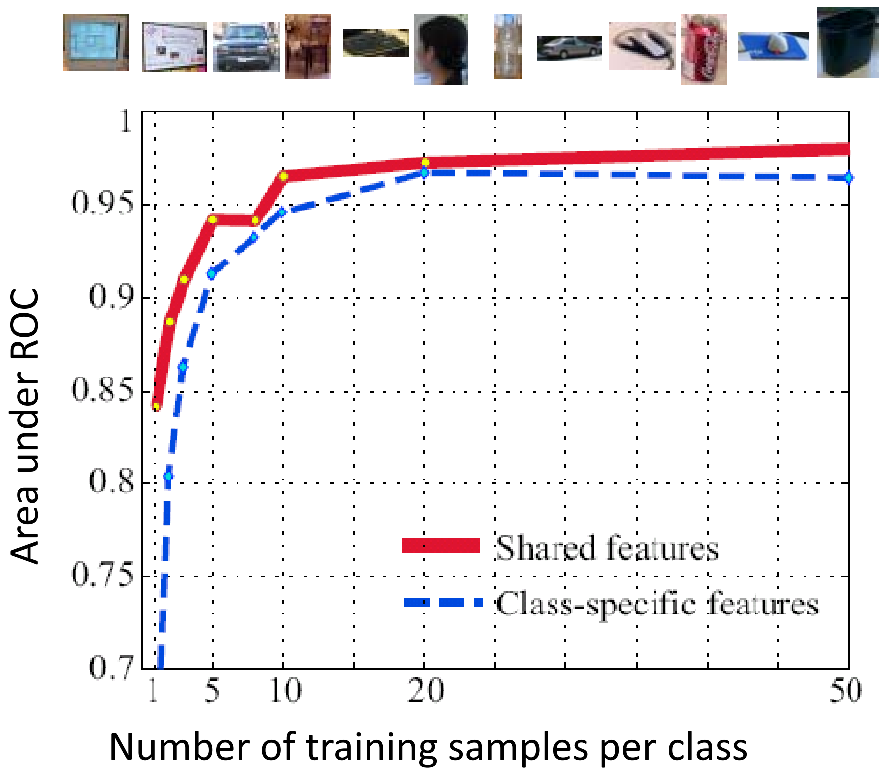
50 training samples/class
29 object classes
2000 entries in the dictionary

Results averaged on 20 runs

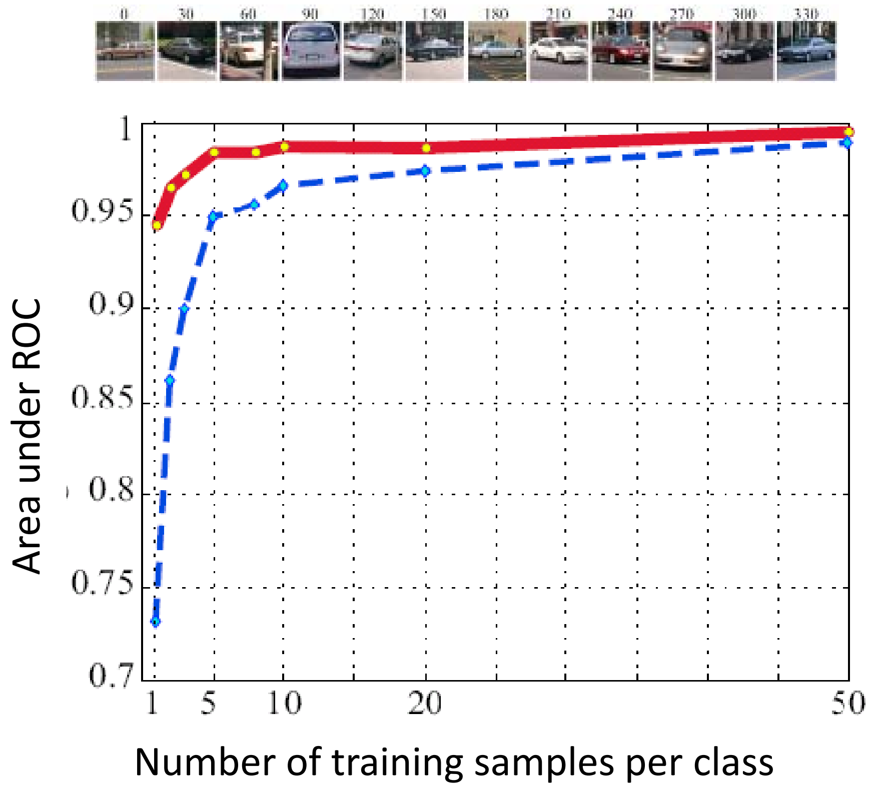
Torralba, Murphy, Freeman. CVPR 2004. PAMI 2007

Generalization as a function of object similarities

12 unrelated object classes

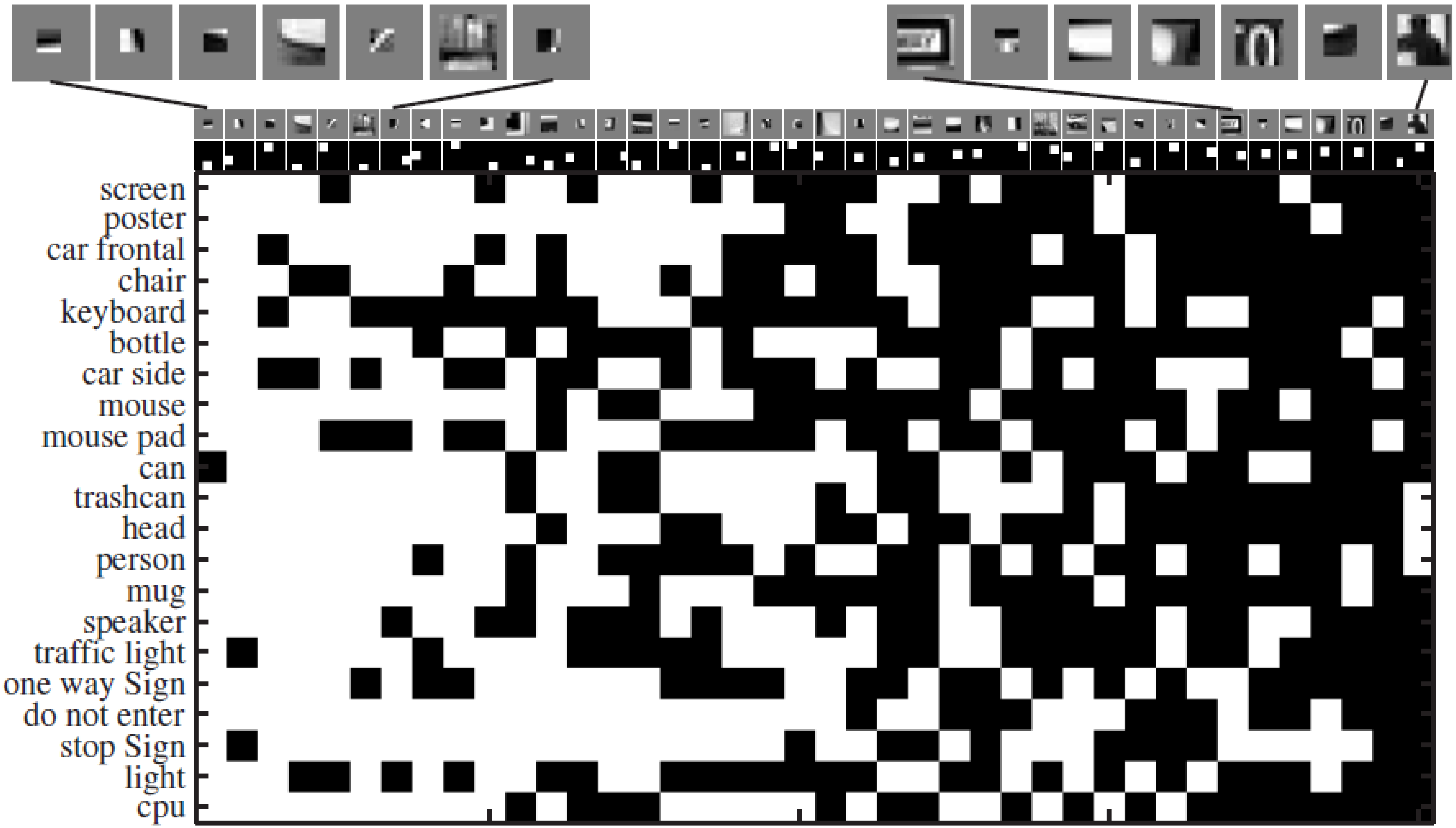


12 viewpoints



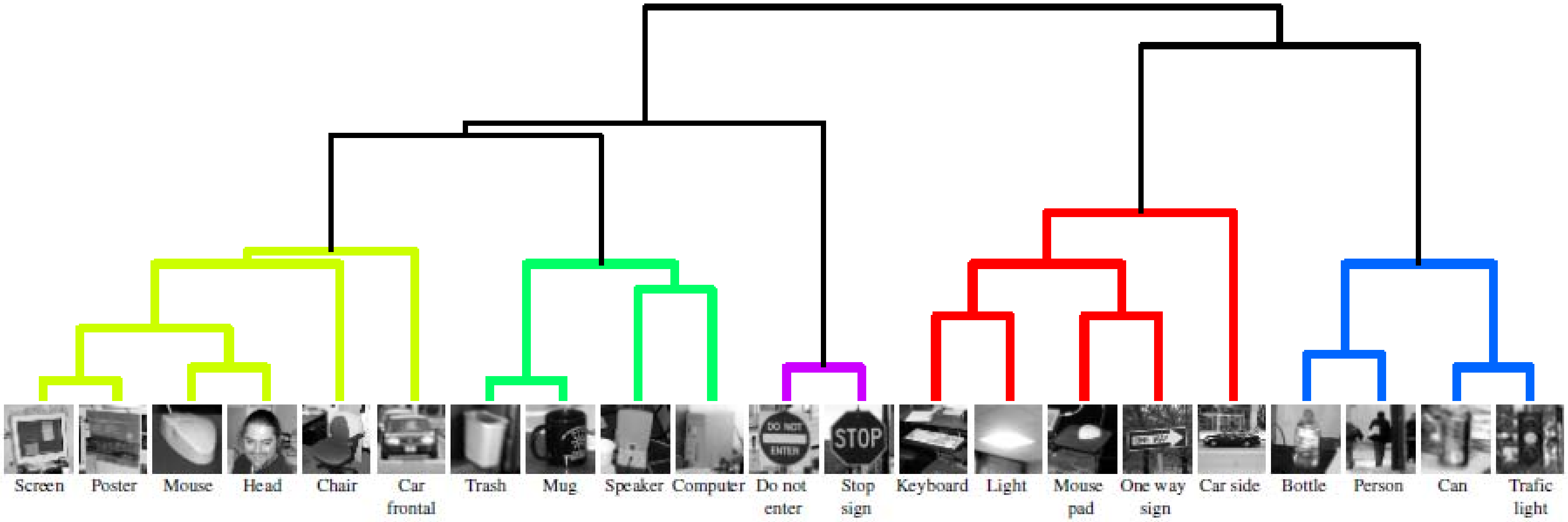
Torralba, Murphy, Freeman. CVPR 2004. PAMI 2007

Generic vs. specific features



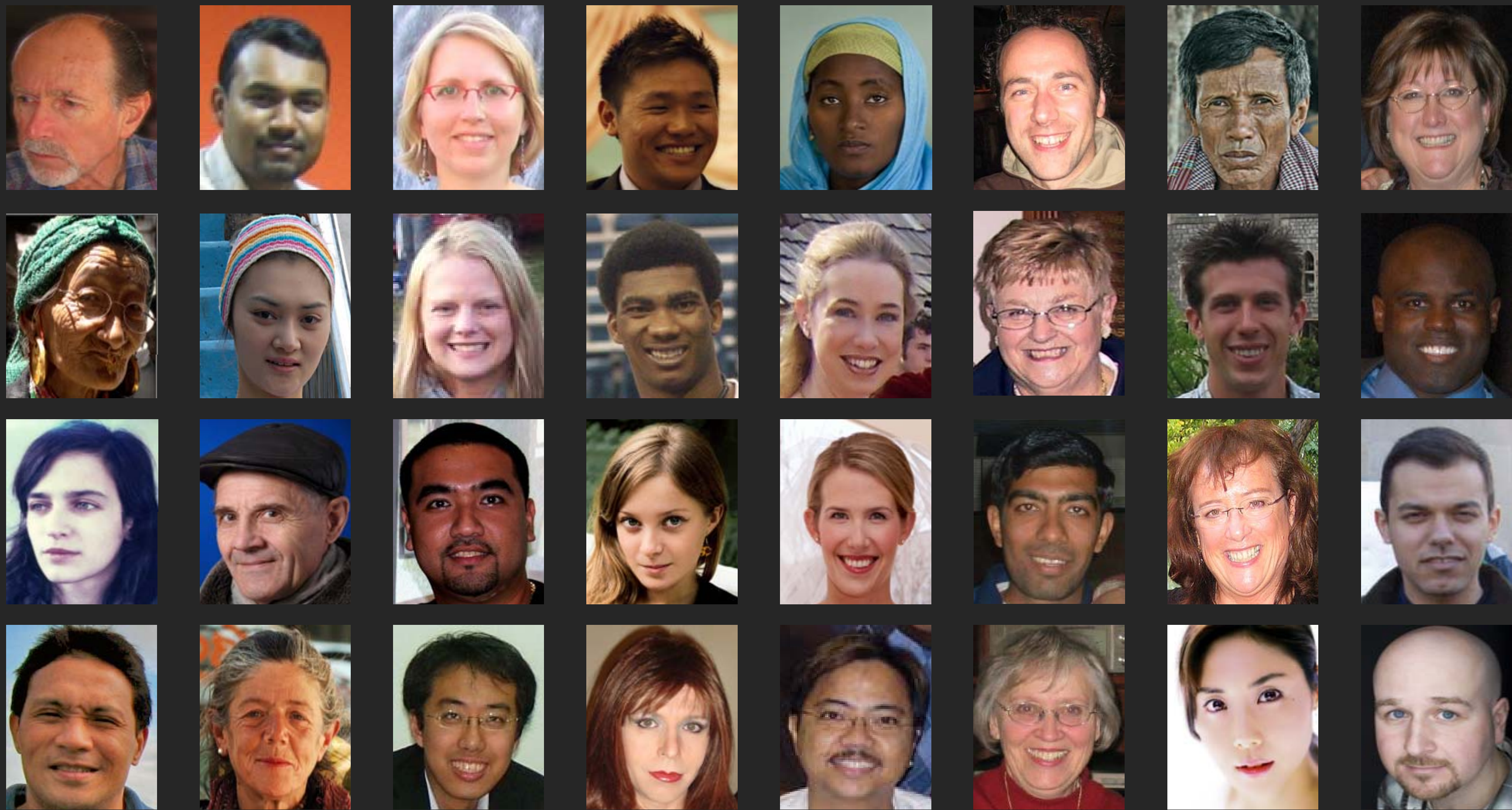
Torralba, Murphy, Freeman. CVPR 2004. PAMI 2007

Object clustering according to shared features



Torralba, Murphy, Freeman. CVPR 2004. PAMI 2007

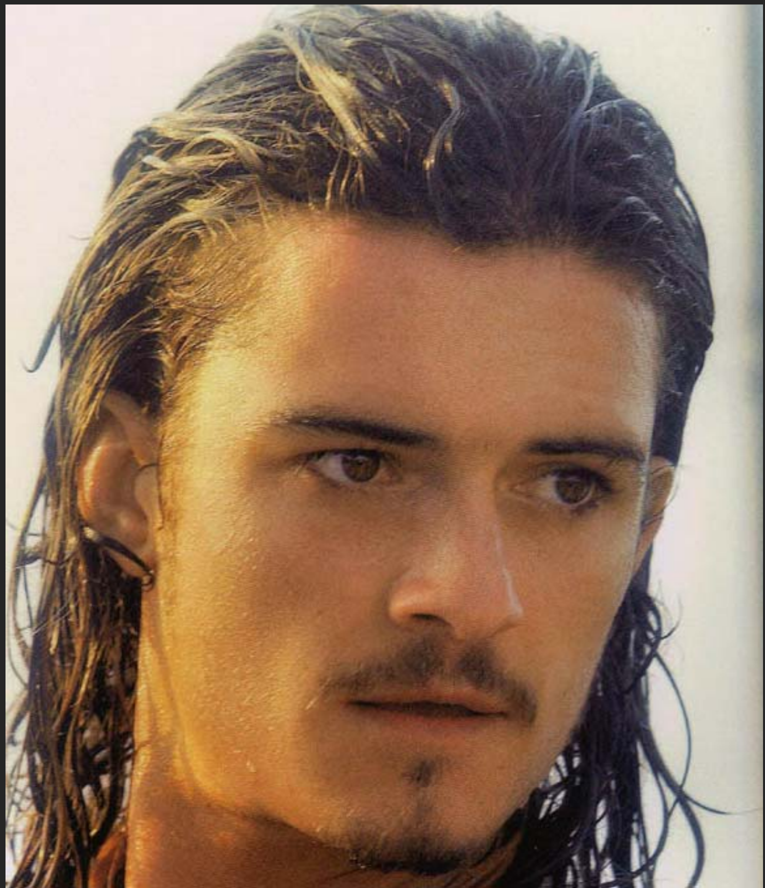
Another multi-class problem: Face recognition



We do not want to learn recognition of each person from scratch!

N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar ICCV 2009

Are these images of the same person?



N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar ICCV 2009

Prior approaches

Images → Low-level features → Verification

RGB
HOG
LBP
SIFT
...

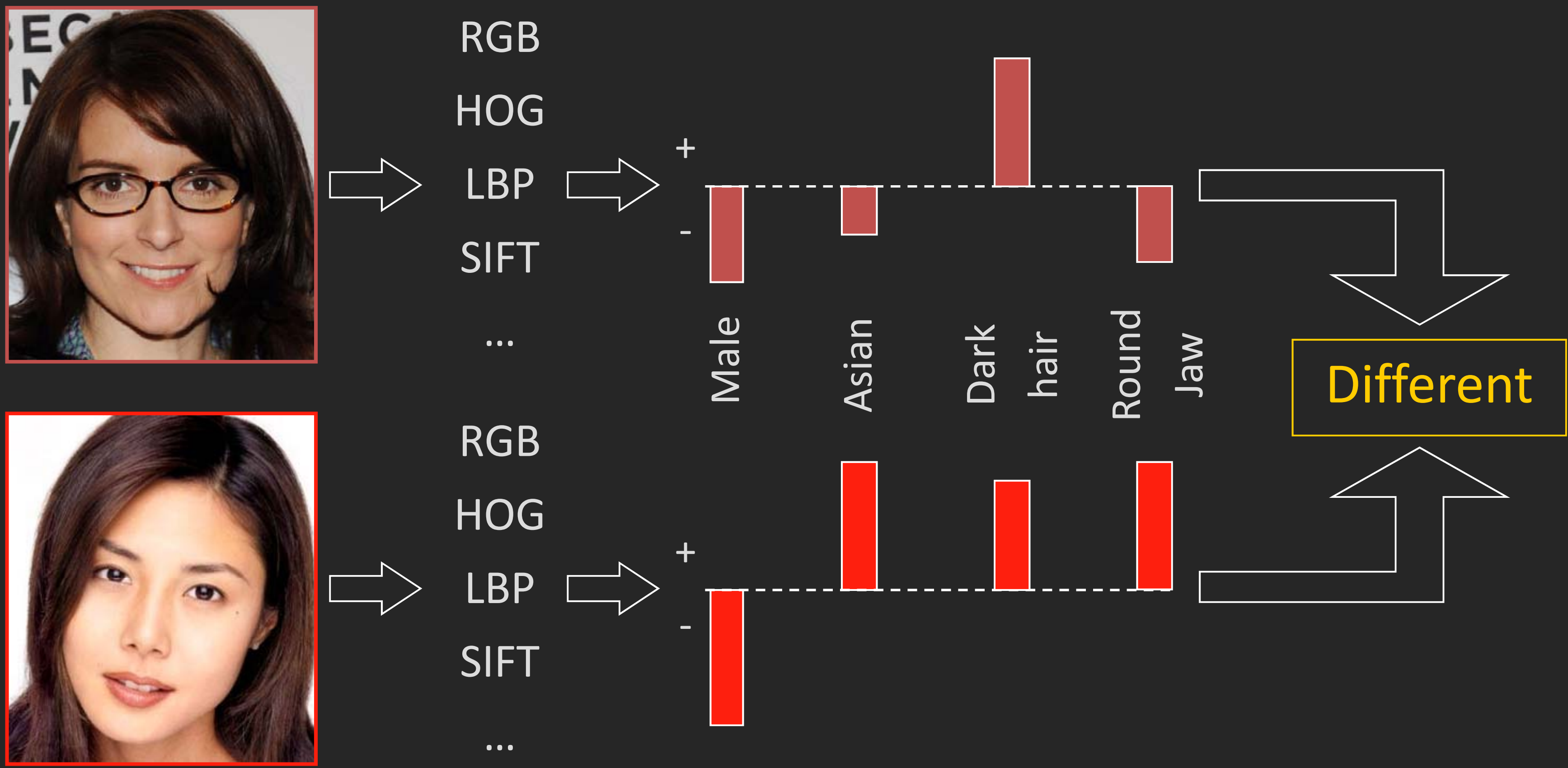
RGB
HOG
LBP
SIFT
...

Different

N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar ICCV 2009

Approach: attributes

Images → Low-level features → Attributes → Verification



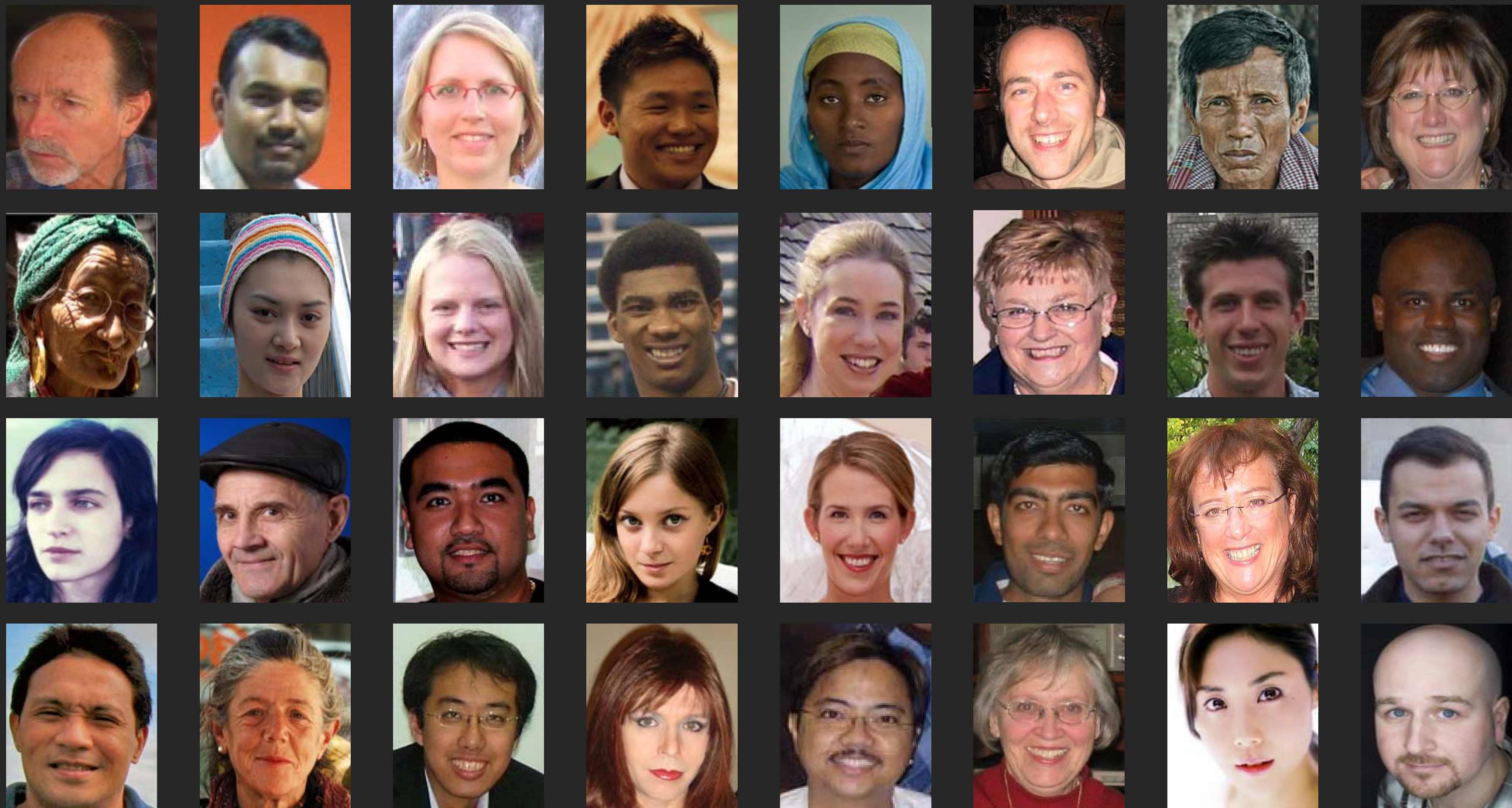
N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar ICCV 2009

Attributes can define categories

Caucasian Middle-aged

Female

Eyeglasses Dark hair



N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar ICCV 2009

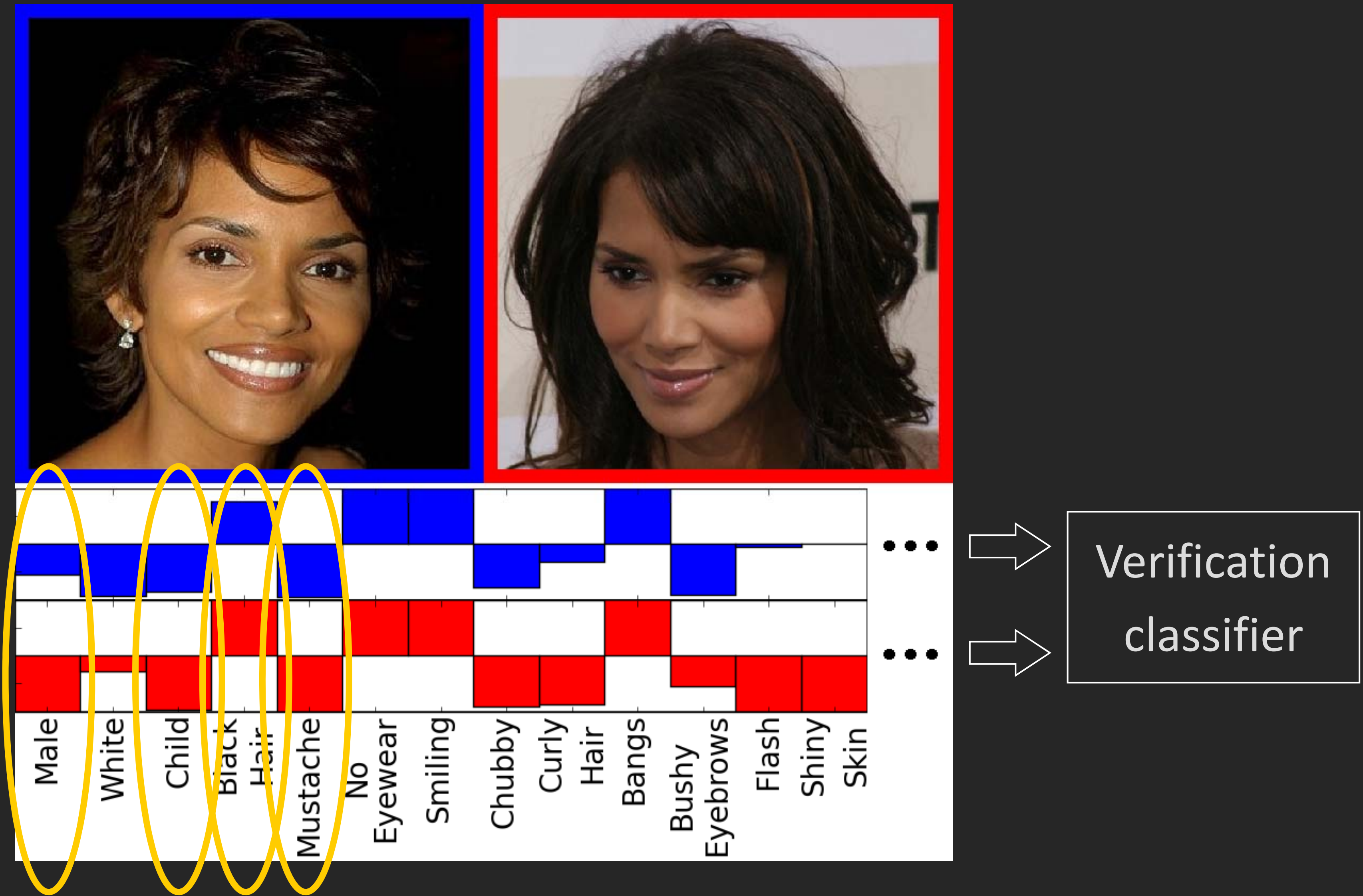
Some attributes may be irrelevant

Teeth showing   Tilted head

Outside

N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar ICCV 2009

Using attributes to perform verification



N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar ICCV 2009

Attributes are intuitive

Female

Young

Attractive

White

Black hair

Frontal pose

Mouth closed

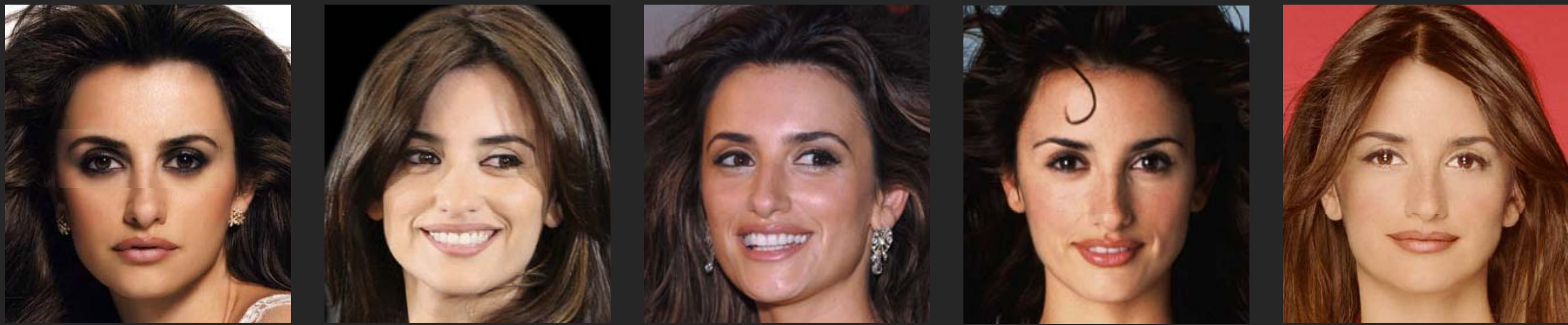
Eyes open

N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar ICCV 2009

Describe faces using similes

N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar ICCV 2009

Training simile classifiers

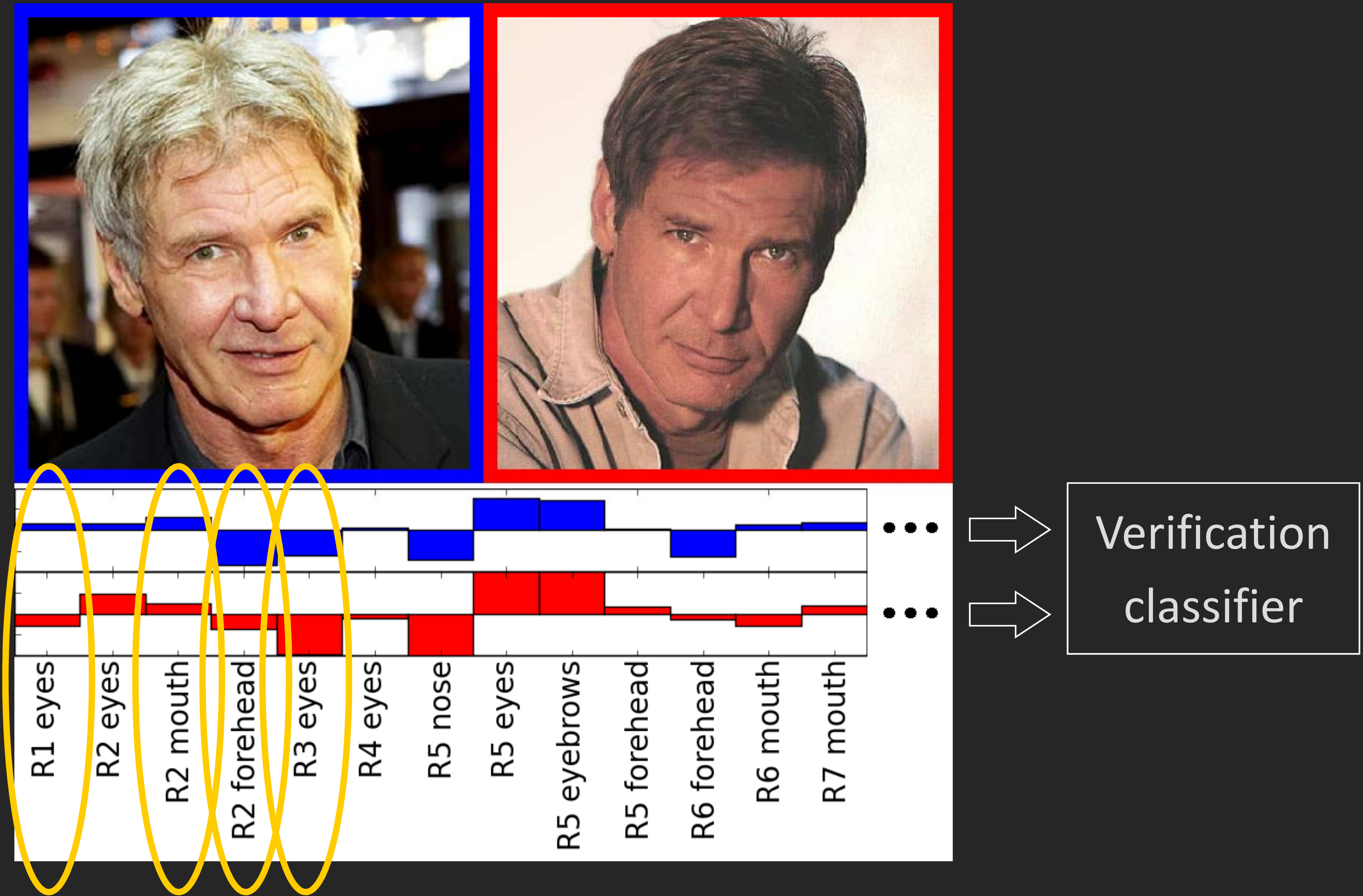


Images of Penelope Cruz ’s eyes

Images of other people ’s eyes

N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar ICCV 2009

Using simile classifiers for verification

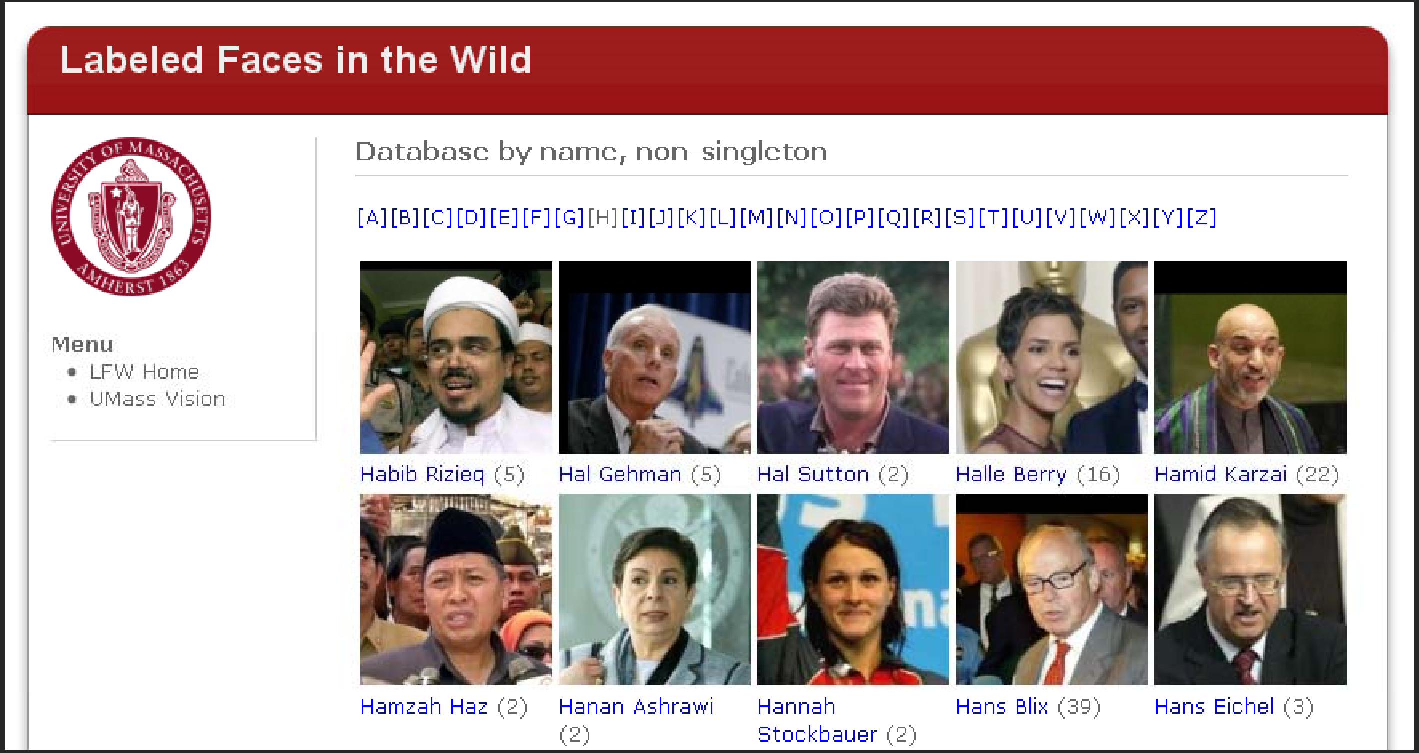


N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar ICCV 2009

Experimental evaluation

LFW Image-Restricted Benchmark:

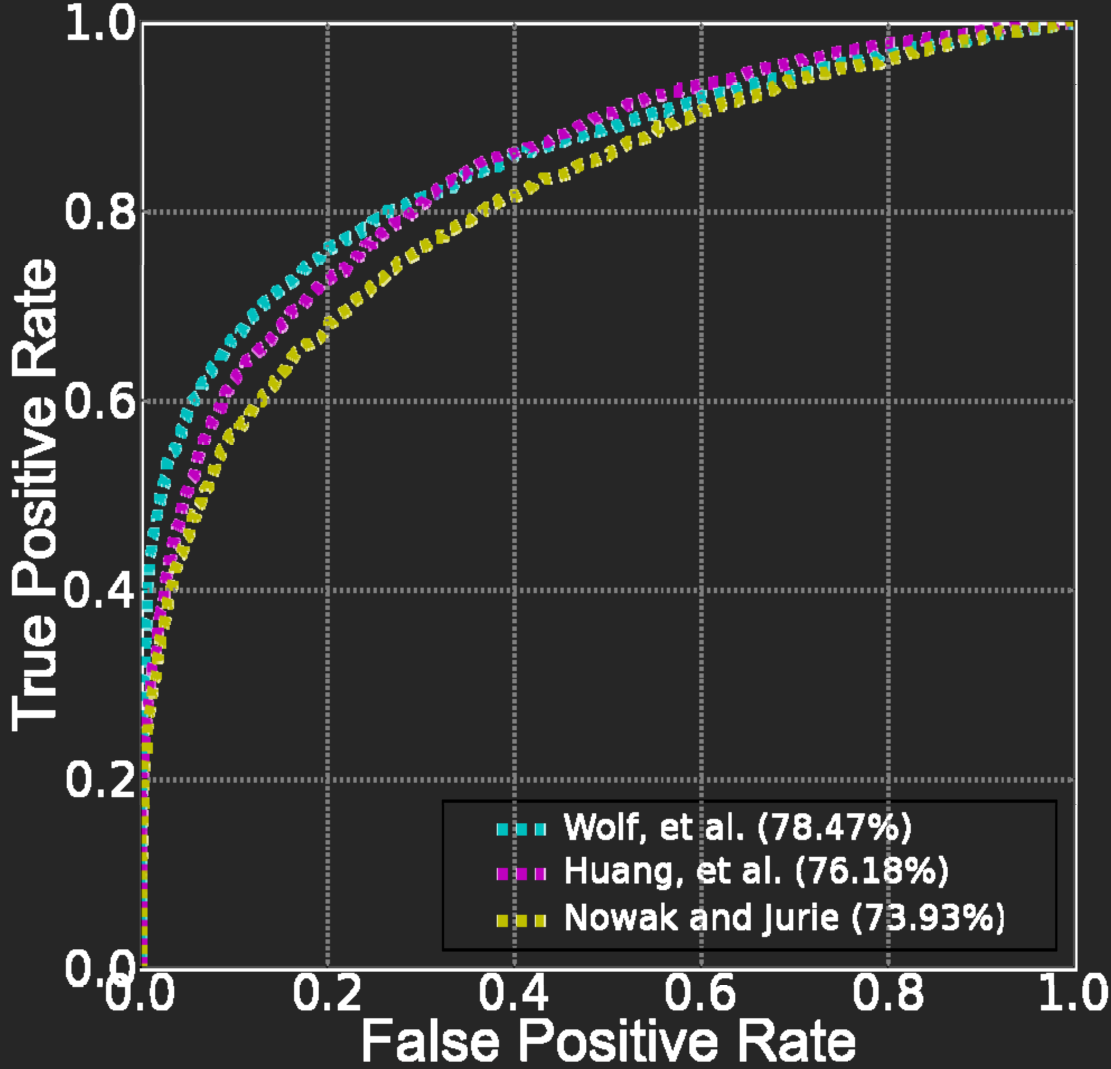
- 6,000 face pairs (3,000 same, 3,000 different)
- 10-fold cross-validation



http://vis-www.cs.umass.edu/lfw

N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar ICCV 2009

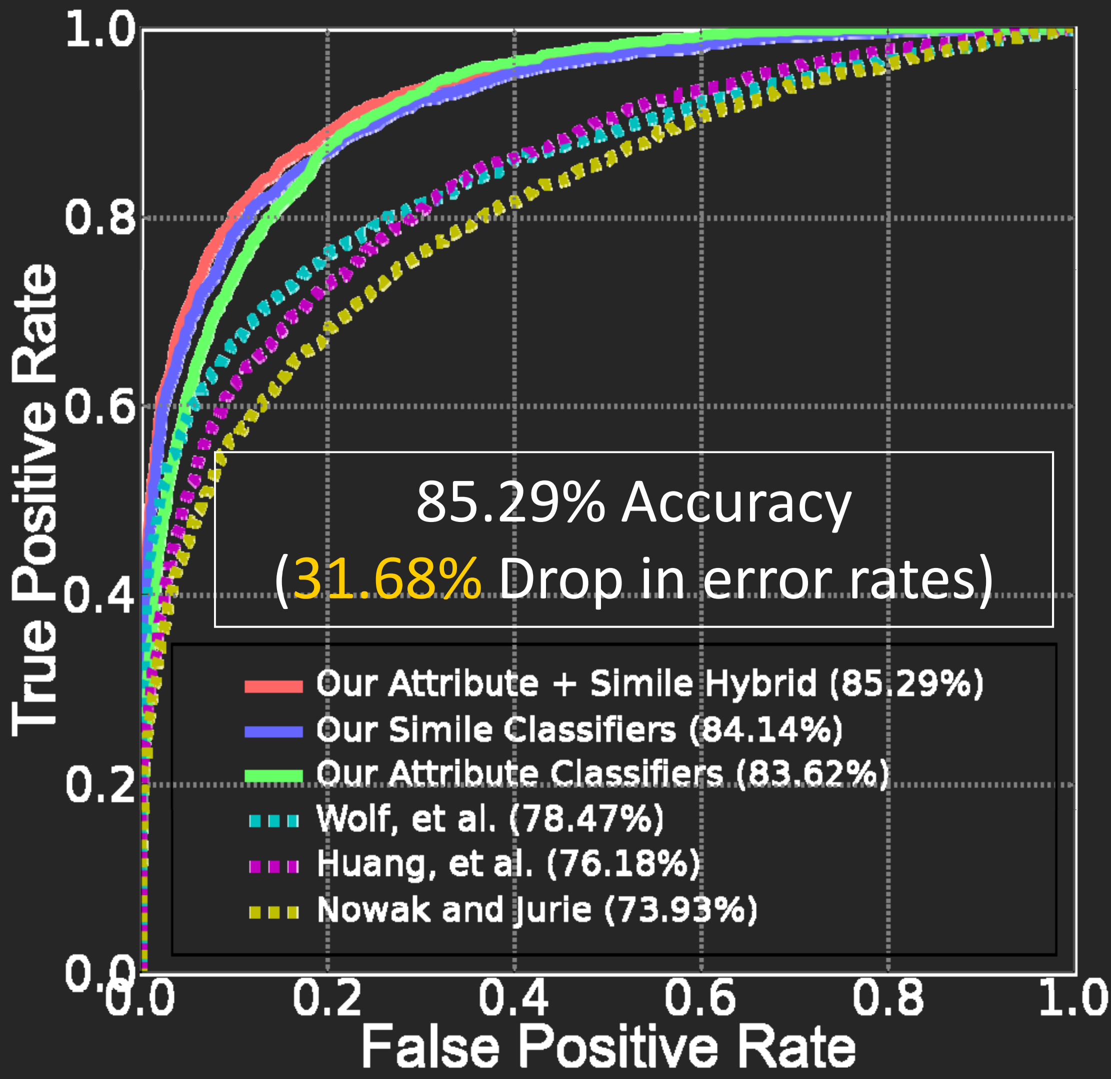
Previous state-of-the-art on LFW



as of May 2009

N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar ICCV 2009

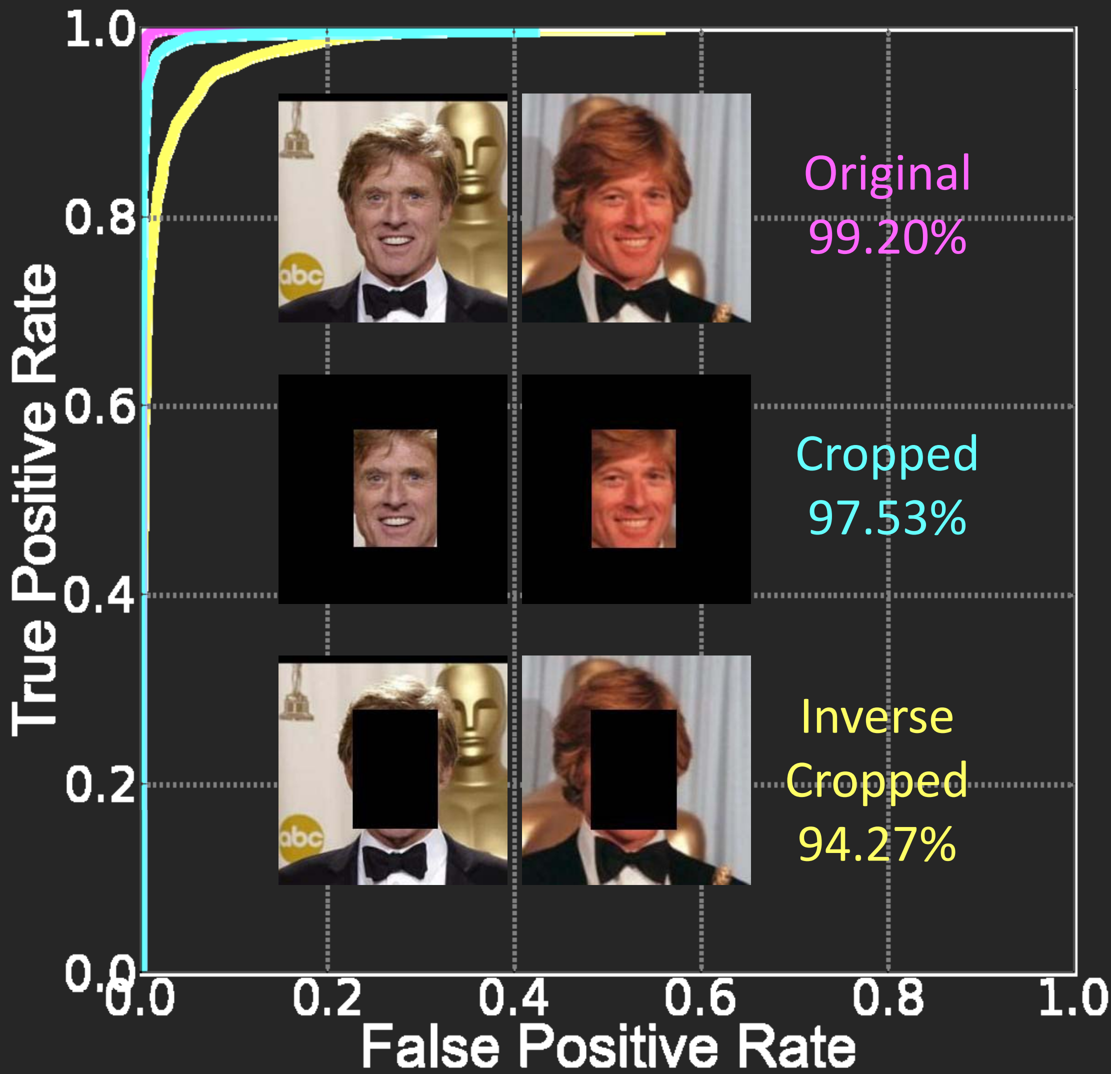
Kumar et al. 2009 on LFW



as of May 2009

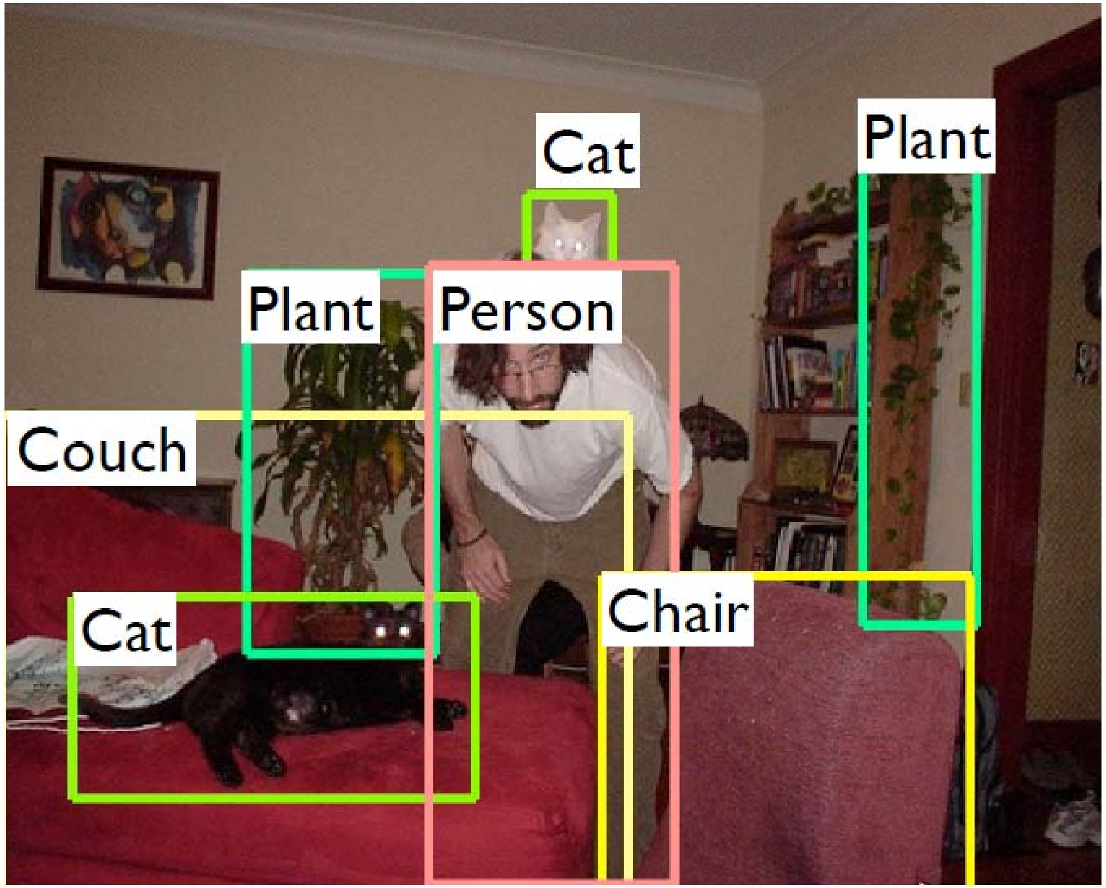
N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar ICCV 2009

Human face verification performance



N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar ICCV 2009

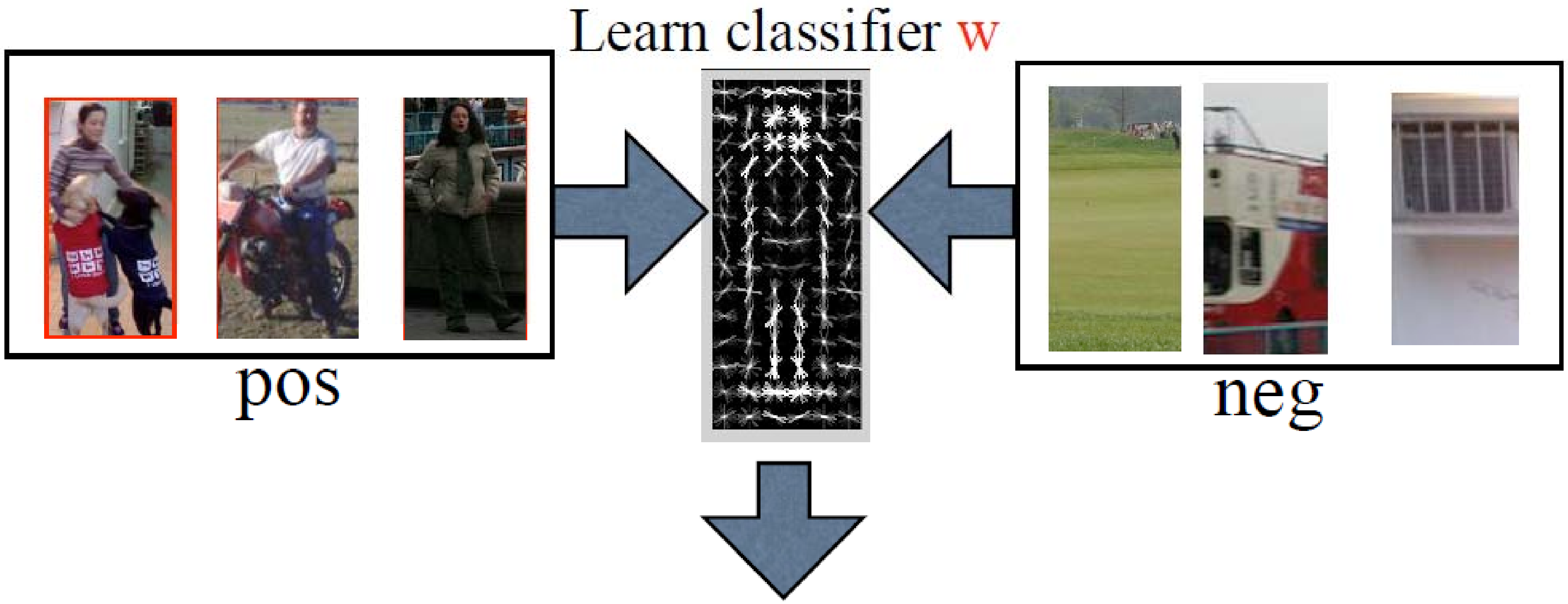
What about multiple objects in the same image?



Multiclass object detection

C. Desai, D. Ramanan, C. Fowlkes ICCV 2009

Scanning-window pattern classification



$w^T x > 0$?

Face detection

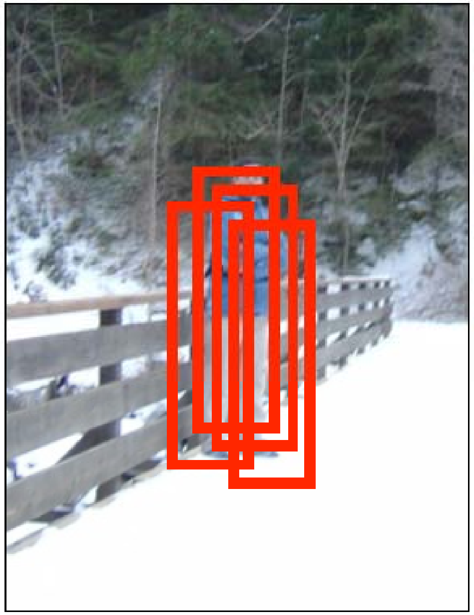
Rowley, Baluja, & Kanade. CVPR 96
Viola & Jones IJCV 01

Pedestrian detection
(and other objects)

Oren et al. CVPR 97
Dalal & Triggs CVPR 05
Felzenswalb et al. PAMI 09

C. Desai, D. Ramanan, C. Fowlkes ICCV 2009

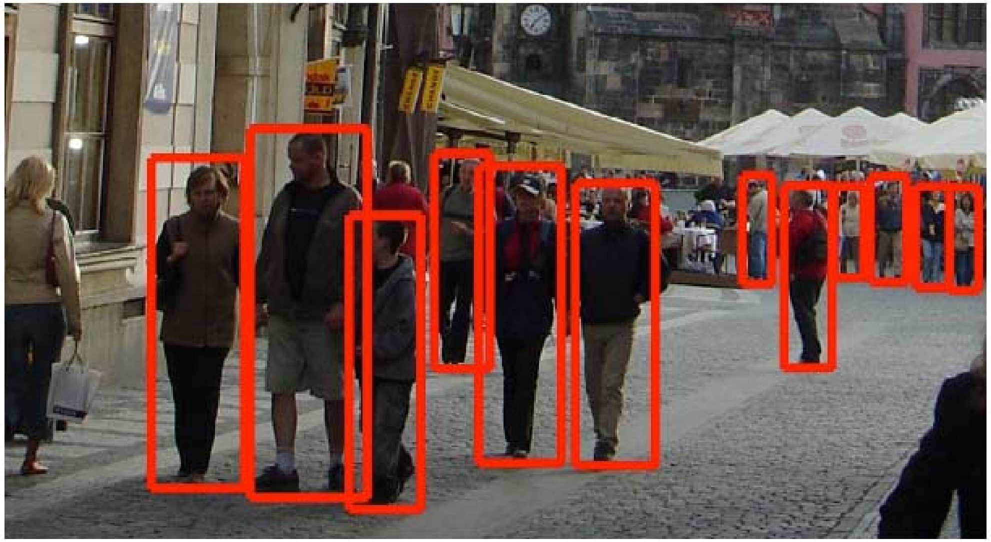
Non-maxima suppression (NMS)



We need to suppress overlapping detections
Many heuristics (mode finding, greedy selection)

C. Desai, D. Ramanan, C. Fowlkes ICCV 2009

NMS in cluttered scenes



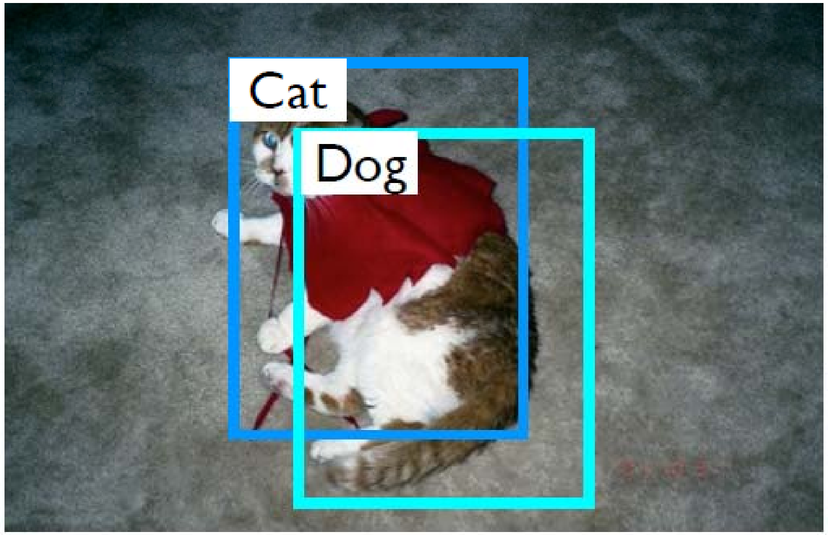
NMS should exploit spatial statistics of objects in real scenes

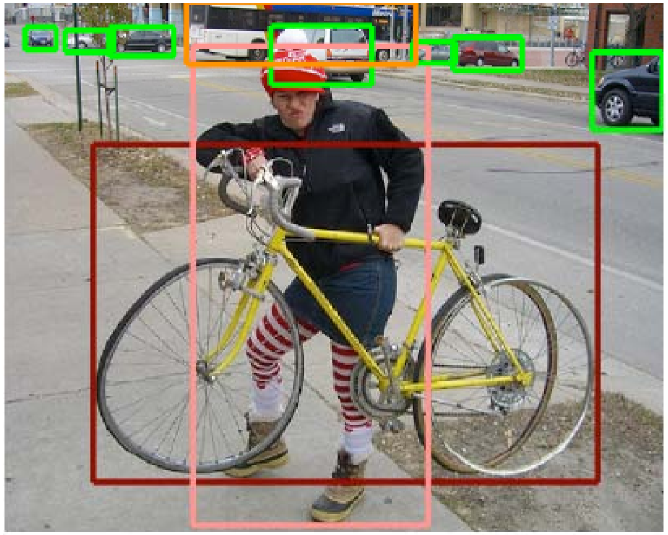
Is there a principled way to learn how to perform NMS?

C. Desai, D. Ramanan, C. Fowlkes ICCV 2009

Inter-class NMS

Mutual exclusion: two objects cannot occupy the same 3D volume





May not be a strict constraint due to porous or transparent objects

C. Desai, D. Ramanan, C. Fowlkes ICCV 2009

Taxonomy of spatial interactions

	within-class	between-class
negative	NMS	mutual exclusion
positive	textures of objects	spatial cueing

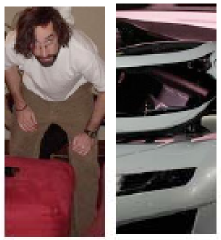
Most past work focuses on positive interactions,
heuristically performing NMS & mutual exclusion.

Our contribution: a model for all of the above

Our inspiration: Torralba, Murphy, & Freeman NIPS 04
Kumar & Hebert ICCV 05
He, Zemel, & Carreira-Perpinan CVPR 05
Galleguillos, Rabinovich & Belongie CVPR 08
Hoeim, Efros, & Hebert IJCV 08

Object detection as ....

Classification



x = image window

$y \in \{0,1\}$

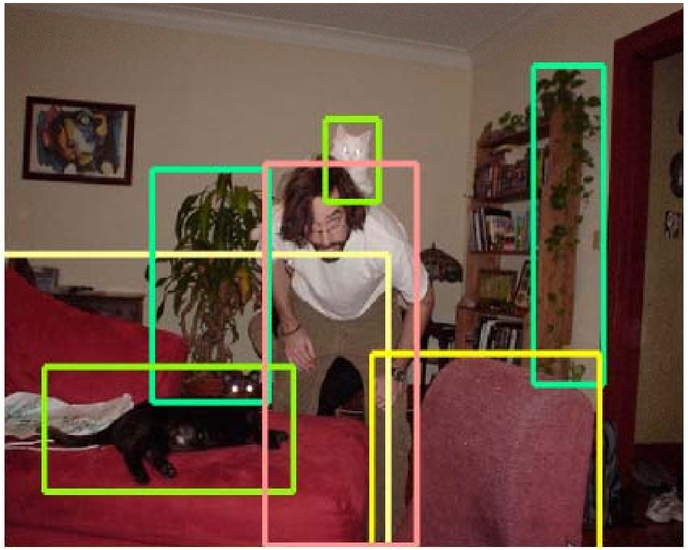
C. Desai, D. Ramanan, C. Fowlkes ICCV 2009

Object detection as a structured labeling task

Classification

x = image window
$y \in \{0,1\}$

Structured, sparse label



X = entire image
Y = [...4...3...2.7..1..]

C. Desai, D. Ramanan, C. Fowlkes ICCV 2009

Global scoring function

$$S_w(X, Y)$$

$$X = \{x_i\} \quad Y = \{y_i\}$$

x_i = feature vector extracted from i^{th} window (e.g. HOG)

y_i = class label (0..K) for i^{th} window

Global scoring function

$$S_w(X,Y) = \underbrace{\boxed{\sum_i w_{y_i}^T x_i}}_{\text{sum of per-window classifier scores}}$$

w_{yi} = template for class y_i

w_{car}

w_{bus}

$\vdots$

$X = \{x_i\}$ $Y = \{y_i\}$

x_i = feature vector extracted from i^{th} window (e.g. HOG)

y_i = class label (0..K)

sum of per-window classifier scores

C. Desai, D. Ramanan, C. Fowlkes ICCV 2009

Global scoring function

$$S_w(X,Y) = \sum_i w_{y_i}^T x_i + \sum_{i,j} w_{y_i,y_j}^T d_{ij}$$

$X = \{x_i\}$ $Y = \{y_i\}$

x_i = feature vector extracted from i^{th} window (e.g. HOG)

y_i = class label (0..K)

w_{yi} = template for class y_i

w_{car}

w_{bus}

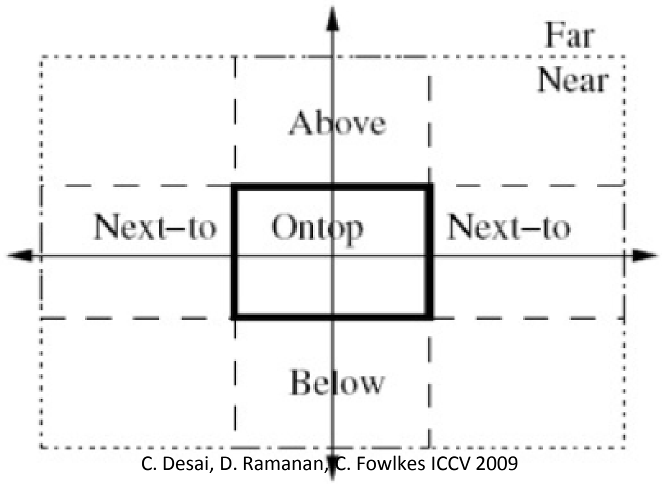
$\vdots$

d_{ij} = spatial context descriptor for window i and j

w_{yiyj} = spatial interaction model for class y_i & y_j

sum of per-window classifier scores

sum of pairwise window-label interactions



C. Desai, D. Ramanan, C. Fowlkes ICCV 2009

Inference

$$S_w(X,Y) = \sum_i w_{y_i}^T x_i + \sum_{i,j} w_{y_i,y_j}^T d_{ij}$$

$$L(X) = \underset{Y}{\operatorname{argmax}}\, S_w(X,Y)$$

Looks like an MRF - can we use standard inference techniques?

Our model is not sub-modular

Sub-modular interactions: neighboring labels should be similar

NMS interactions: neighboring labels should be different

C. Desai, D. Ramanan, C. Fowlkes ICCV 2009

Greedy inference

$$L(X) = \underset{Y}{\operatorname{argmax}}\, S(X,Y) \qquad S(X,Y) = \sum_i w_{y_i}^T x_i + \sum_{i,j} w_{y_i,y_j}^T d_{ij}$$

Analogous to common NMS schemes

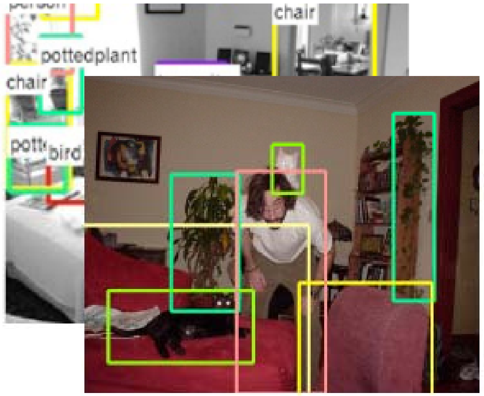
(1) Initialize all labels to bg
Initialize per-window scores with local template
(2) Select highest scoring un-instanced window
(3) Instance it and add pairwise contribution to remaining windows
(4) Stop when remaining windows score < 0

Effectiveness: Greedy solution close to optimal in practice
(See Numhauser et al. 78 for theoretical arguements)

C. Desai, D. Ramanan, C. Fowlkes ICCV 2009

Learning

Training data consists of pairs of $\{X_n, Y_n\}$



$$S_w(X,Y) = \sum_i w_{y_i}^T x_i + \sum_{i,j} w_{y_i,y_j}^T d_{ij}$$

$$S_w(X,Y) = w^T \Psi(X,Y)$$

C. Desai, D. Ramanan, C. Fowlkes ICCV 2009

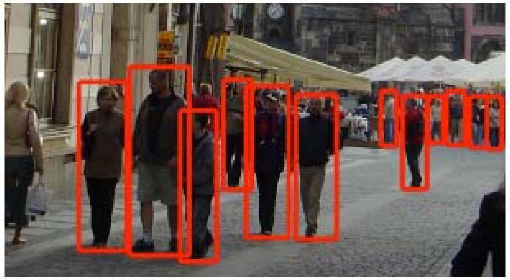
Learning with SVMs

$$\underset{w}{\operatorname{argmin}} \quad \frac{1}{2} w^T w$$

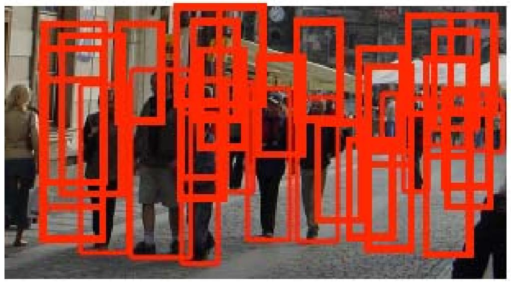
$$\text{s.t.} \quad \forall n, H_n \neq Y_n \qquad w^T \Psi(X_n, Y_n) - w^T \Psi(X_n, H_n) \geq 1$$

"Find a small w such that for each image, score of true label Y_n dominates all other hypothesized labels H_n by at least 1 unit"

Y_n



H_n



Only a tiny fraction of exponential number of constraints are necessary (i.e., support vectors)

Structured Prediction
Tsochantaridis et al. ICML 04

C. Desai, D. Ramanan, C. Fowlkes ICCV 2009

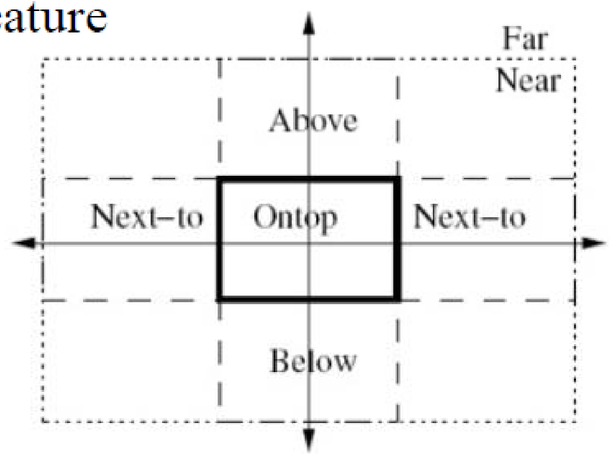
Experiments

1) We use PASCAL 2007 training and test data
   20 classes, 5000 training images, 5000 test images

2) Baseline: Felzenswalb et al. PAMI 09 (with default NMS)

3) Local feature = [score of baseline detector 1]
   (We learn bias and offset for each local detector)

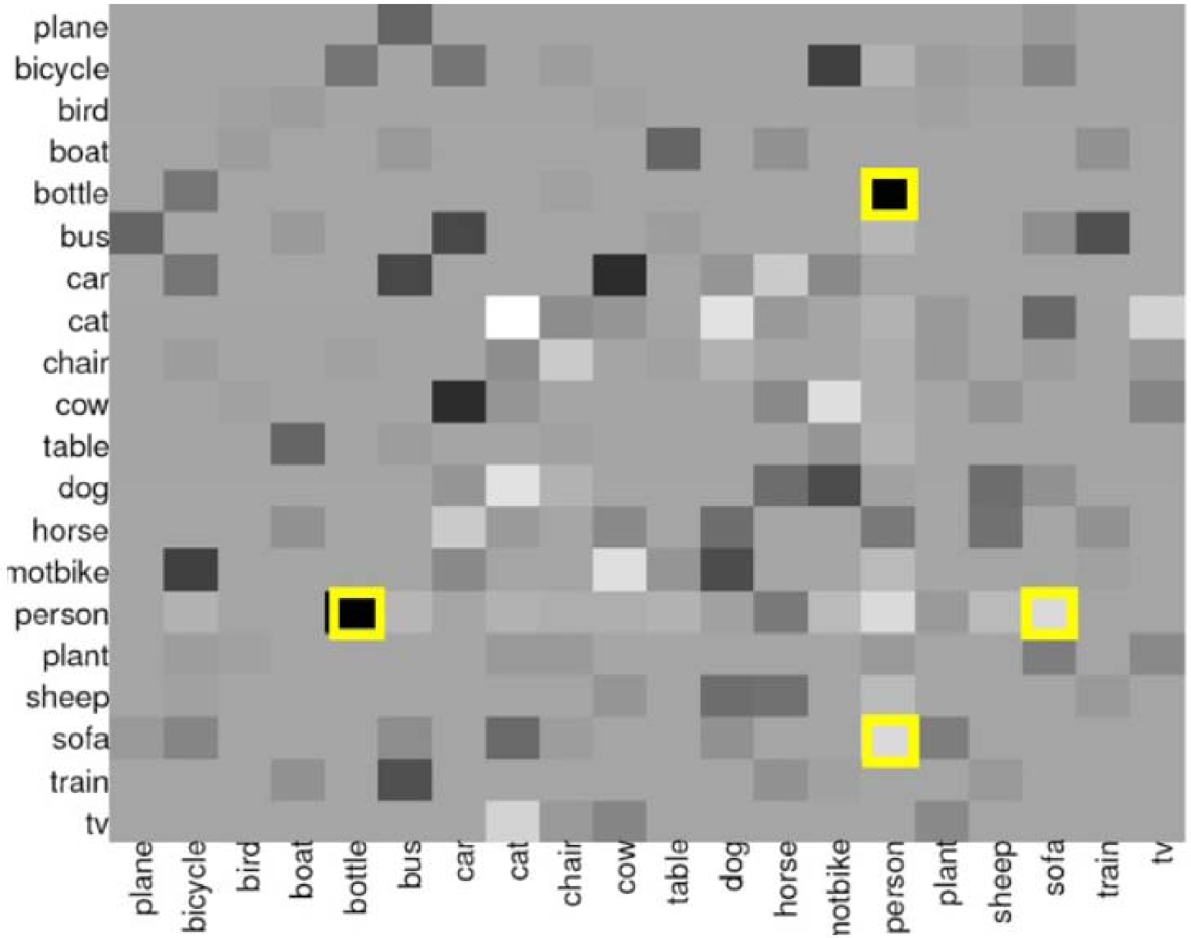
4) Pairwise feature



+ 50% overlap feature

C. Desai, D. Ramanan, C. Fowlkes ICCV 2009

Overlap feature in pairwise potential

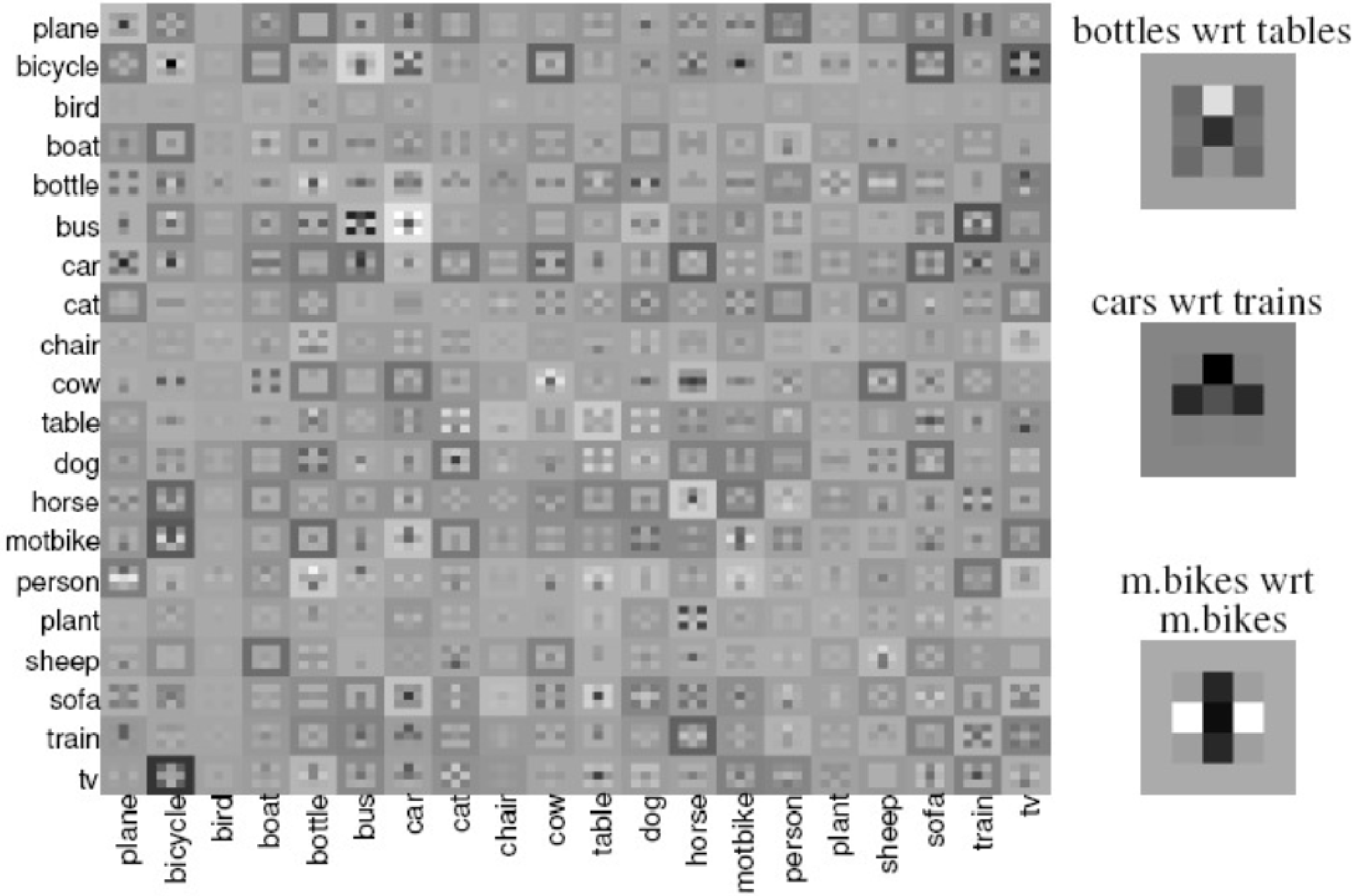


Mutual exclusion can be subtle

Parameters are trained with knowledge of local detectors

C. Desai, D. Ramanan, C. Fowlkes ICCV 2009

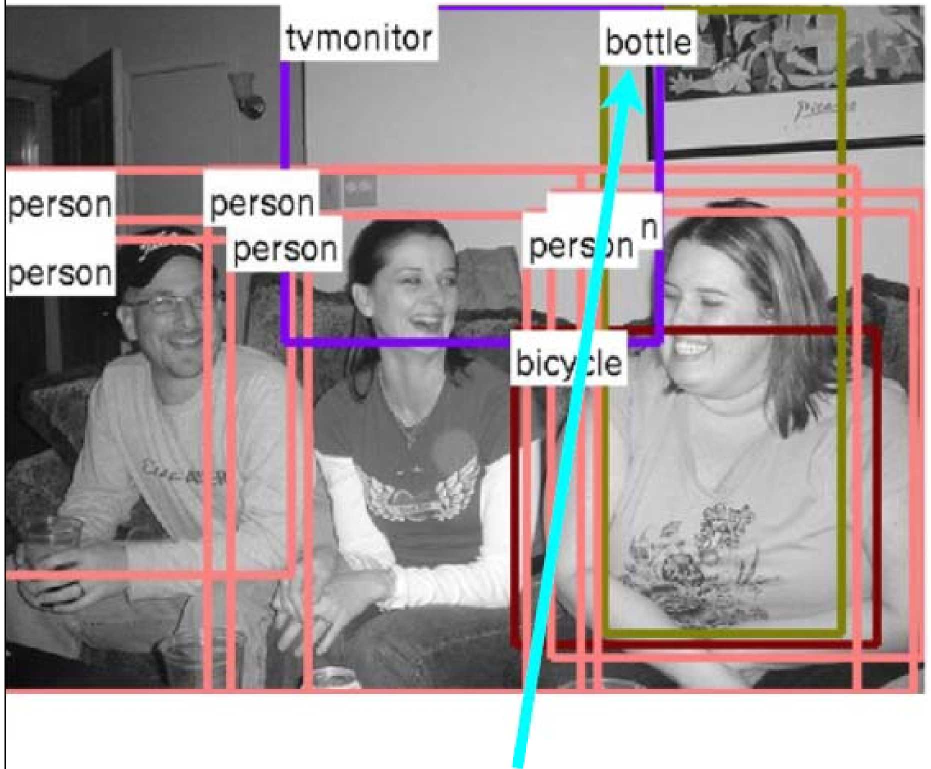
Remaining pairwise potentials



C. Desai, D. Ramanan, C. Fowlkes ICCV 2009

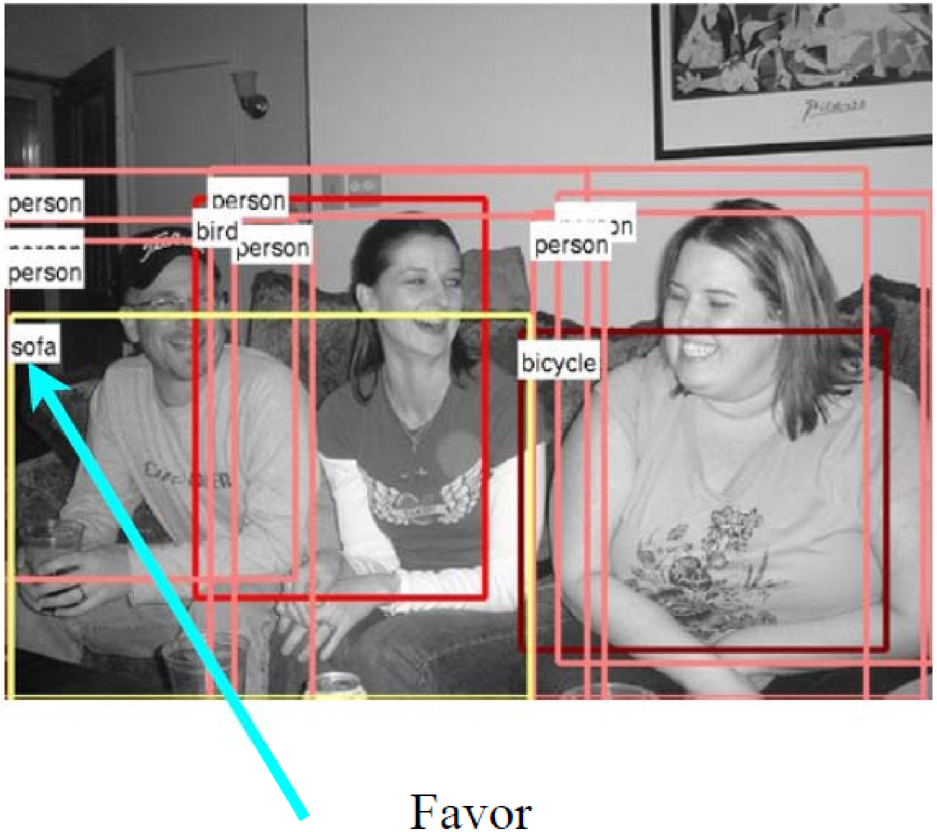
Results

Top 10 detections for baseline



Inhibit
overlapping people & bottles
because local detectors confuse them

Our top 10 detections

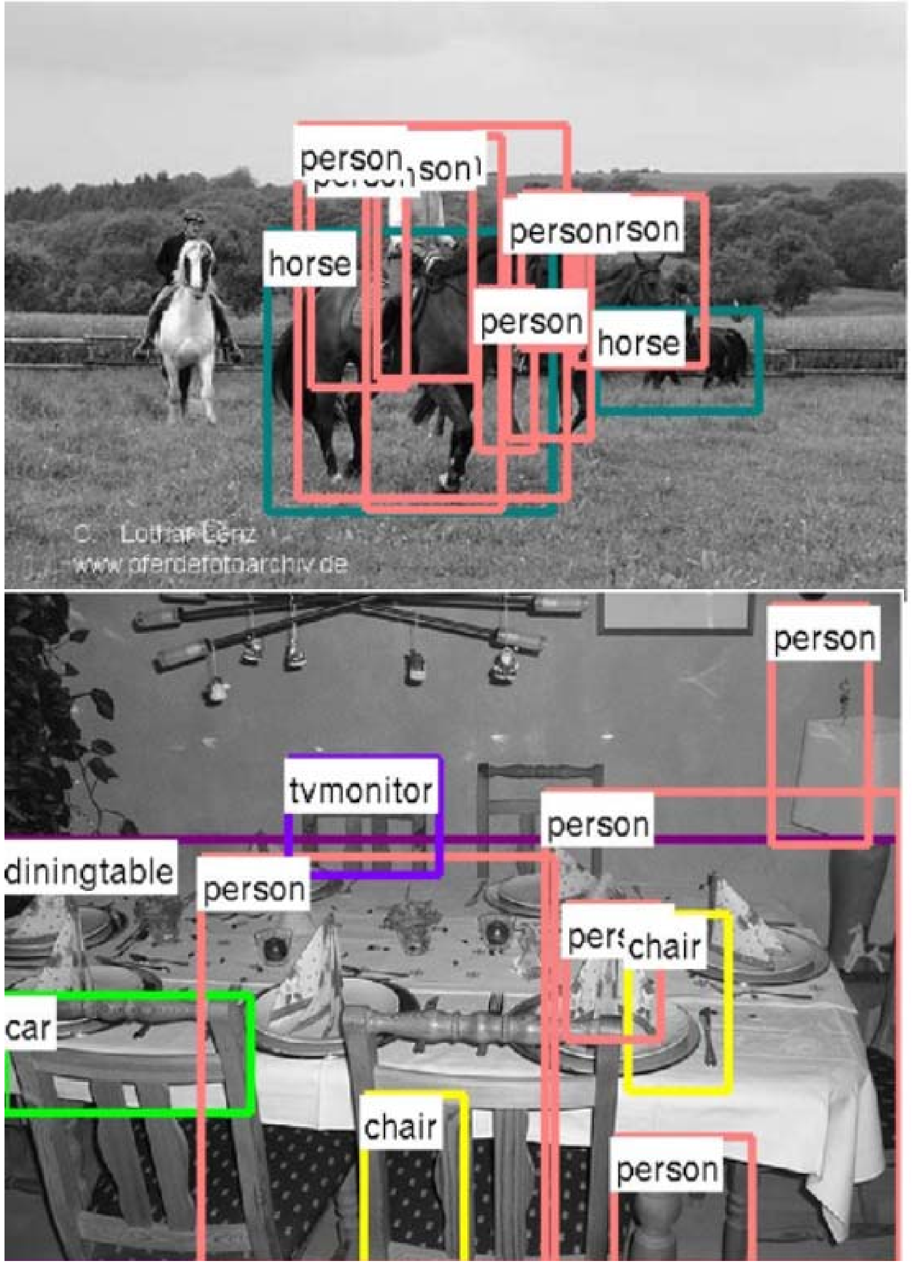


Favor
overlapping people & sofas
because people sit on sofas

C. Desai, D. Ramanan, C. Fowlkes ICCV 2009

Results

Baseline



Our model

Default NMS heuristics

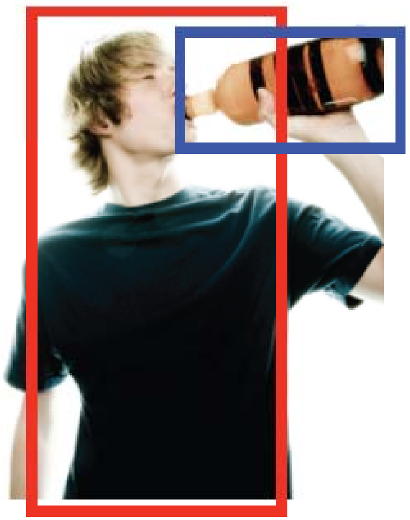
Default heuristics don't work for Mutual Exclusion

	Winning PASCAL07 score	Felzenszwalb et al. PAMI 09 code	Mutual Exclusion	Our model
plane	.262	0.278	0.270	**0.288**
bike	.409	0.559	0.444	**0.562**
bird	**.098**	0.014	0.015	0.032
boat	.094	**0.146**	0.125	0.142
bottle	.214	0.257	0.185	**0.294**
bus	**.393**	0.381	0.299	0.387
car	.432	0.470	0.466	**0.487**
cat	**.240**	0.151	0.133	0.124
chair	.128	**0.163**	0.145	0.160
cow	.140	0.167	0.109	**0.177**
table	.098	0.228	0.191	**0.240**
dog	**.162**	0.111	0.091	0.117
horse	.335	0.438	0.371	**0.450**
motbike	.375	0.373	0.325	**0.394**
person	.221	0.352	0.342	**0.355**
plant	.120	0.140	0.091	**0.152**
sheep	**.175**	0.169	0.091	0.161
sofa	.147	0.193	0.188	**0.201**
train	.334	0.319	0.318	**0.342**
TV	.289	**0.373**	0.359	0.354

Our model outperforms Felzenszwalb et al.'s baseline for most classes

C. Desai, D. Ramanan, C. Fowlkes ICCV 2009

Alternate scores for multiclass detection



Building a 'drinking detector' requires finding people and bottles simultaneously

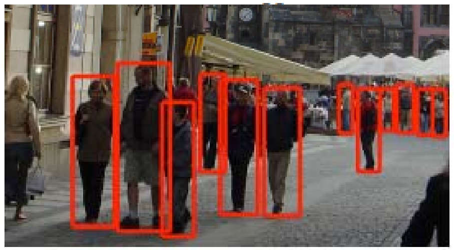
Per-class AP's don't score this

Under more appropriate scoring criteria, our model does significantly better (see paper)

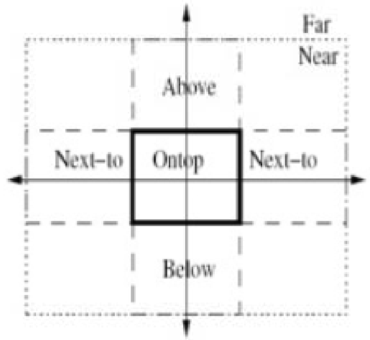
C. Desai, D. Ramanan, C. Fowlkes ICCV 2009

A look back

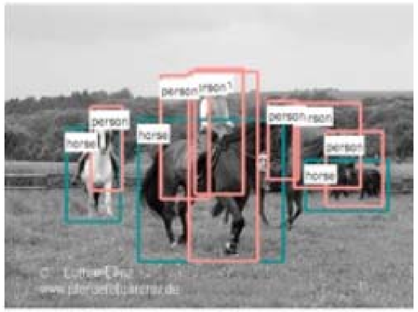
1) Object detection = sparse, structured labeling task



2) Modeling spatial layouts of objects helps



3) Structured prediction provides machinery for such models



$$\underset{Y}{\operatorname{argmax}}\, w^T \Psi(X, Y)$$

C. Desai, D. Ramanan, C. Fowlkes ICCV 2009

What to do about The Object That Cannot Be Named?

Slides by Derek Hoiem
Computer Science Department
University of Illinois at Urbana- Champaign

A. Farhadi, I. Endres, and D. Hoiem 2010

A failure/success story

A. Farhadi, I. Endres, and D. Hoiem 2010

Photo by Ivan Makarov

Dealing with inevitable failure

Failure in categorization should not mean failure in recognition

A. Farhadi, I. Endres, and D. Hoiem 2010

What to do about the

Object That Cannot Be Named?

A. Farhadi, I. Endres, and D. Hoiem 2010

Example

Assisted Driving

A. Farhadi, I. Endres, and D. Hoiem 2010

Example

Security

A. Farhadi, I. Endres, and D. Hoiem 2010

Key steps

1. We need richer, more interconnected object representations

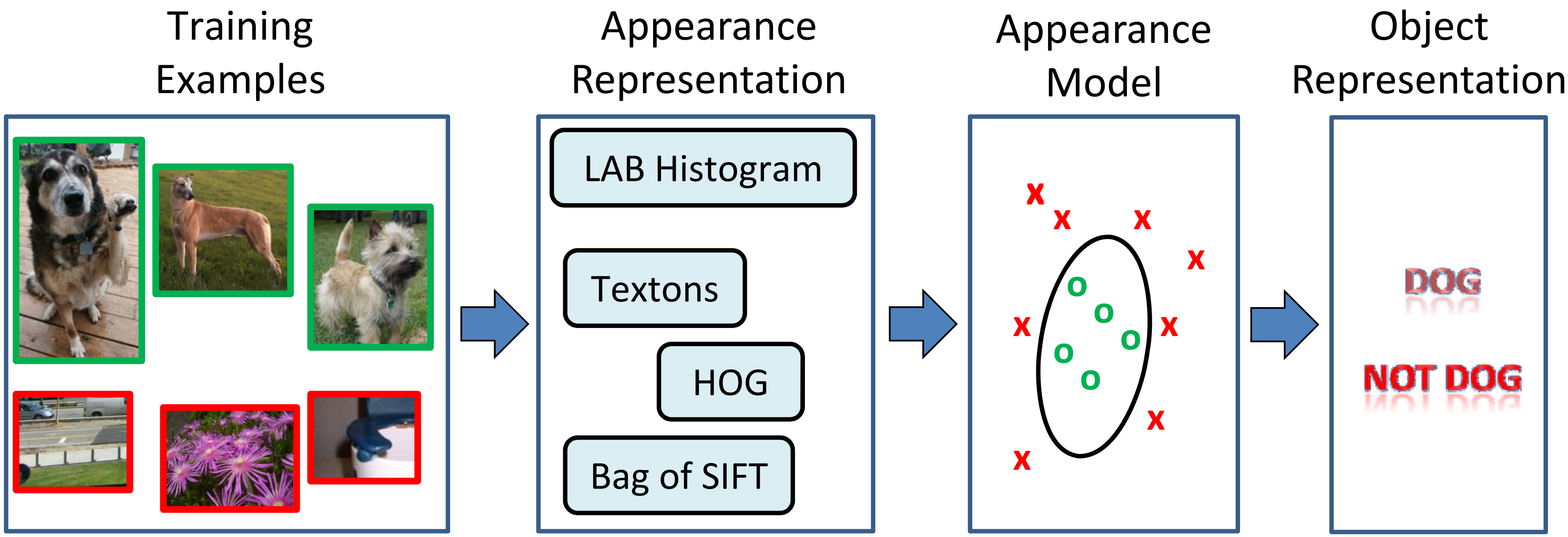
A. Farhadi, I. Endres, and D. Hoiem 2010

What makes a good object representation?

- **Prediction**
  - Where will it go, what will it do, how could I use it?
- **Description**
  - What is it, what is it doing, what does it look like?
- **Generalization**
  - Applicable beyond the immediate task
- **Composition**
  - New, related objects and tasks are easier to learn

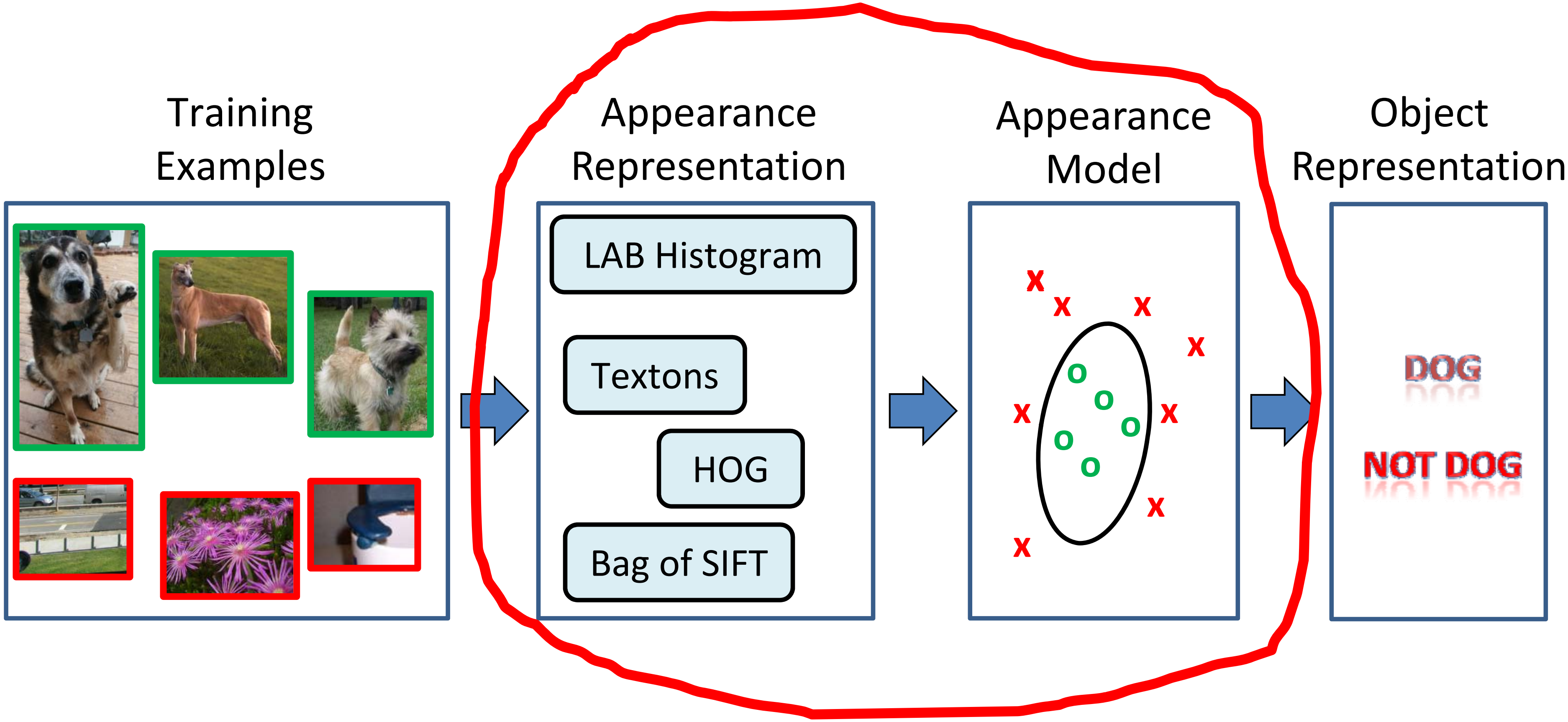
A. Farhadi, I. Endres, and D. Hoiem 2010

Current View of Recognition



A. Farhadi, I. Endres, and D. Hoiem 2010

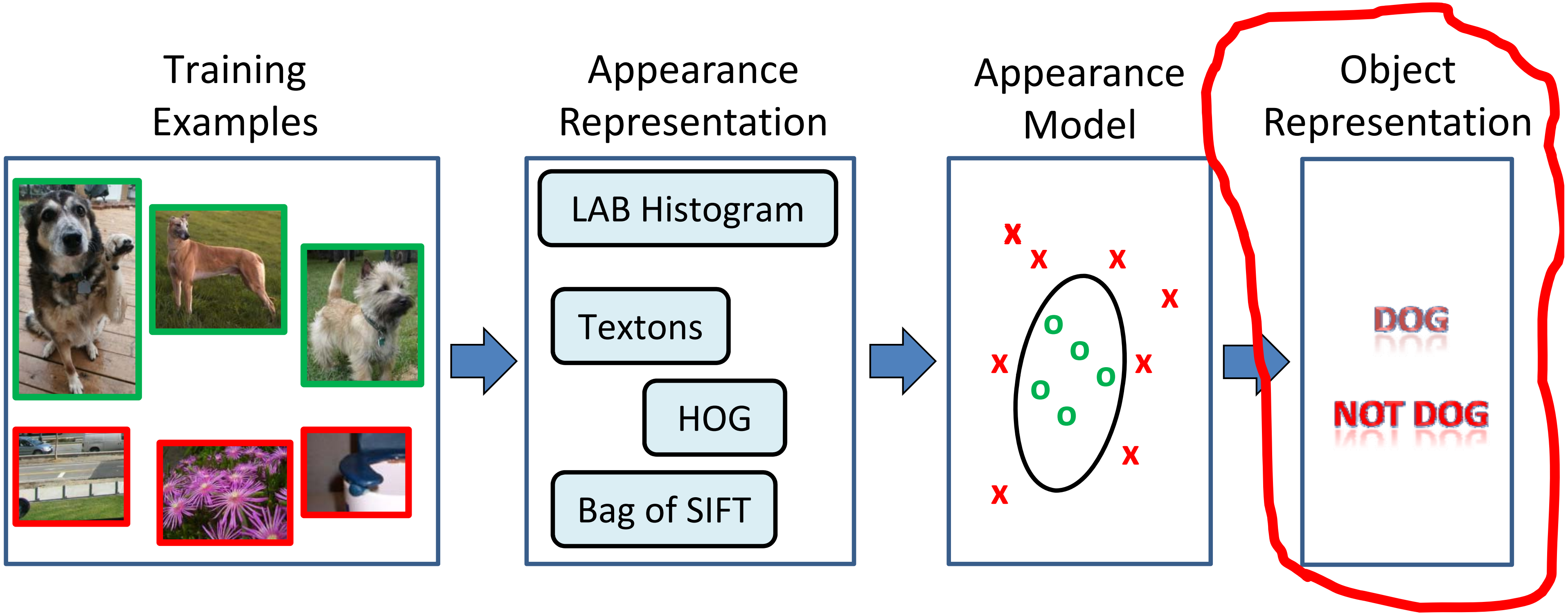
Current View of Recognition



Lots of effort – fancy stuff

A. Farhadi, I. Endres, and D. Hoiem 2010

Current View of Recognition



Not much changed

A. Farhadi, I. Endres, and D. Hoiem 2010

Value of basic categories

DOG ➡

Has head
Is animal
Is furry
Is small
Can be pet
Eats meat

A. Farhadi, I. Endres, and D. Hoiem 2010

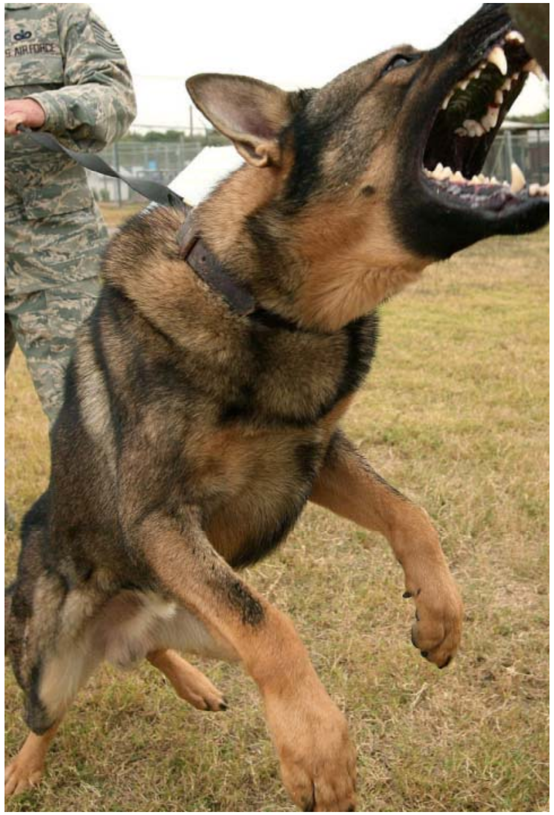
Limitations of basic categories

They provide limited prediction and description

DOG

?=

DOG



A. Farhadi, I. Endres, and D. Hoiem 2010

Limitations of basic categories

They do not apply to objects from novel categories

Familiar Objects


Cat


Horse


Dog

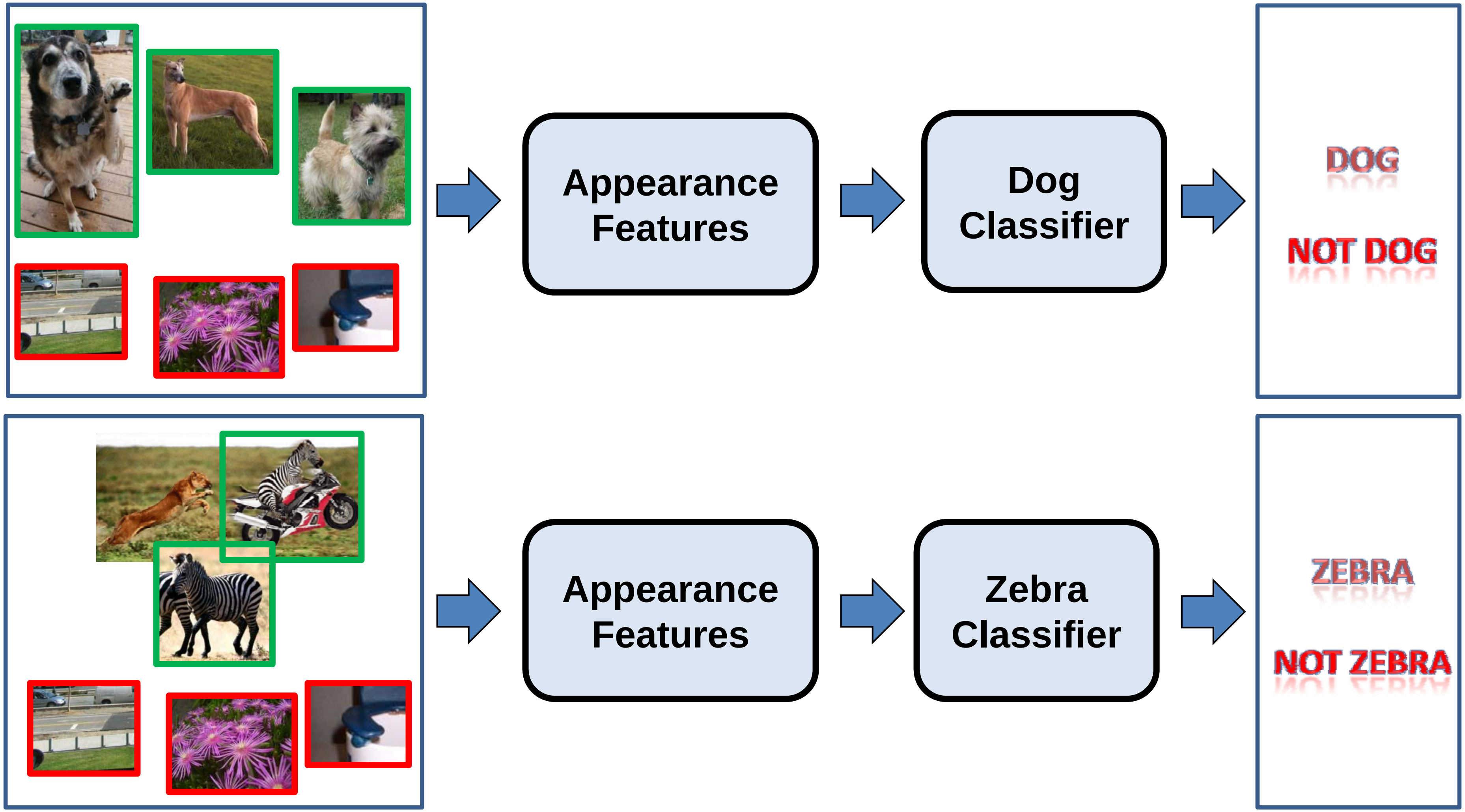
New Object


???

A. Farhadi, I. Endres, and D. Hoiem 2010

Limitations of basic categories

They do not make it easier to learn new categories



Category-based representation

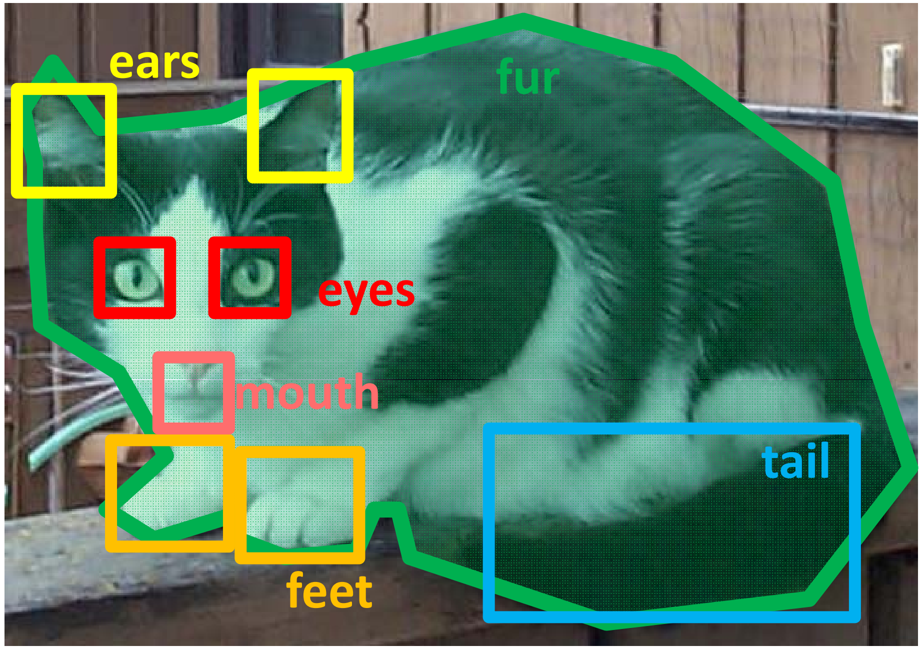
- Limited description and prediction
- No generalization to objects outside of learned categories
- Provides little guidance for learning

So what would make a better representation?

A. Farhadi, I. Endres, and D. Hoiem 2010

Attribute-based Representation

Learn intermediate structure with object categories



Multiple Categories
animal, land animal, …, cat

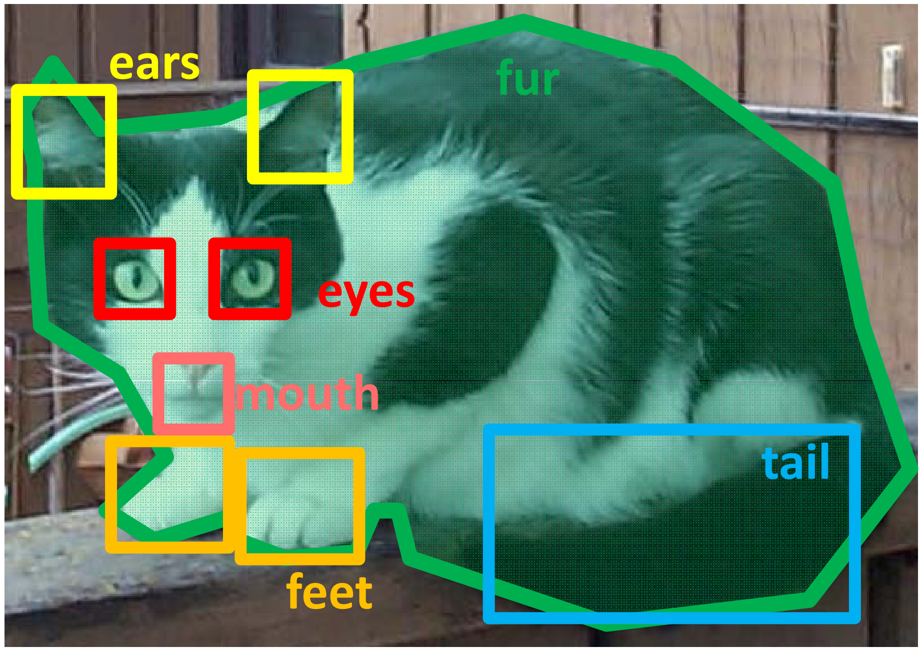
Viewpoint/pose
lying down, left side, facing camera

Function
fast runner, climb trees, eat small animals, jump high, household pet, scratch

A. Farhadi, I. Endres, and D. Hoiem 2010

What we mean by attributes

- Properties that we want to describe or predict
- Shared across basic categories
- Made explicit through supervision



Multiple Categories
animal, land animal, …, cat

Viewpoint/pose
lying down, left side, facing camera

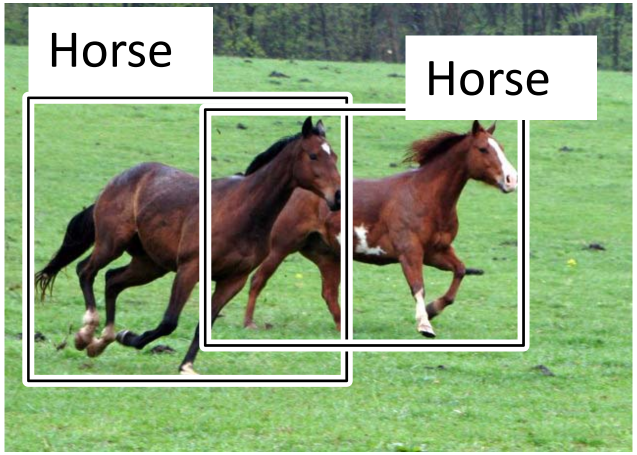
Function
fast runner, climb trees, eat small animals, jump high, household pet, scratch

A. Farhadi, I. Endres, and D. Hoiem 2010

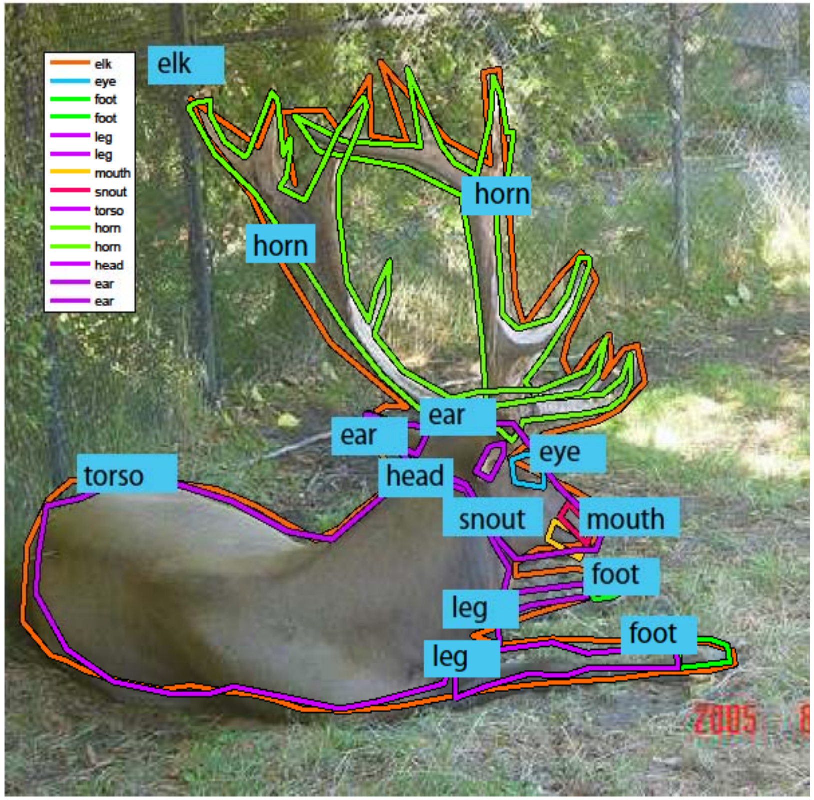
What do these attributes get us?

Image Level

Object Level



Detailed Attributes Level



Categories
Animal
Land animal
Mammal
Four legged animal
Elk

Pose
Lying down = 1
Back = 1
…

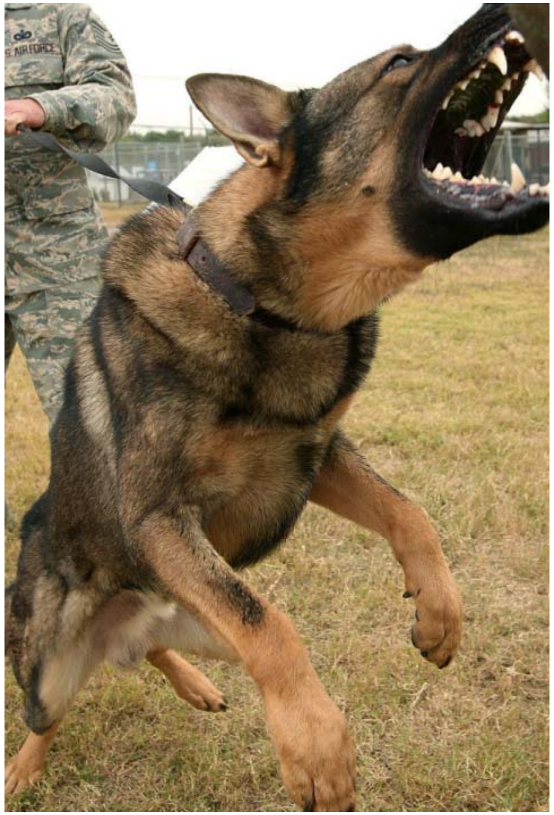
Functional
Can see
Can walk
Herbivorous
…

Material
Pixel segmentations

A. Farhadi, I. Endres, and D. Hoiem 2010

Advantages of supervised attributes

- Enables verbal description of objects and images

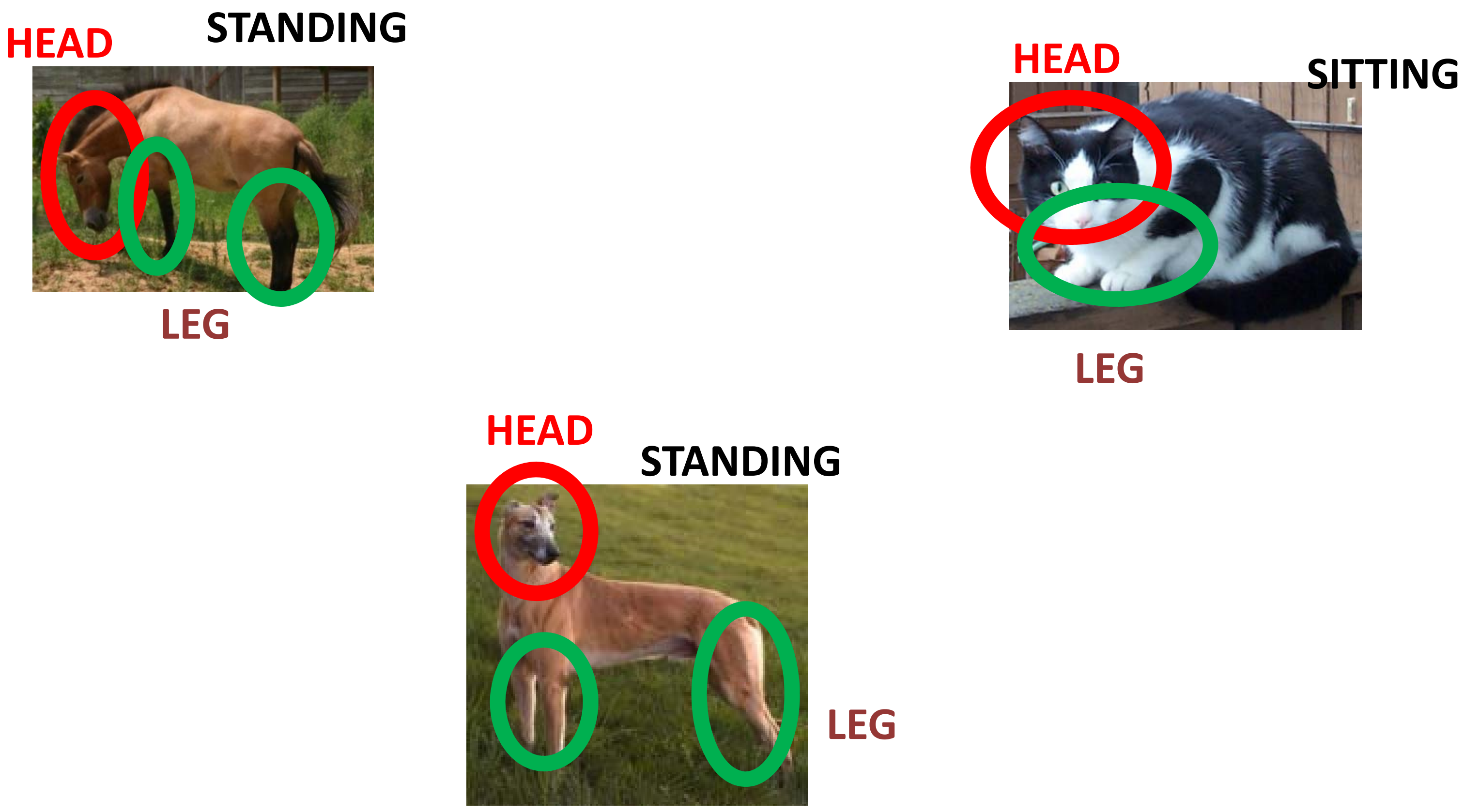


Large angry dog with pointy teeth

A. Farhadi, I. Endres, and D. Hoiem 2010

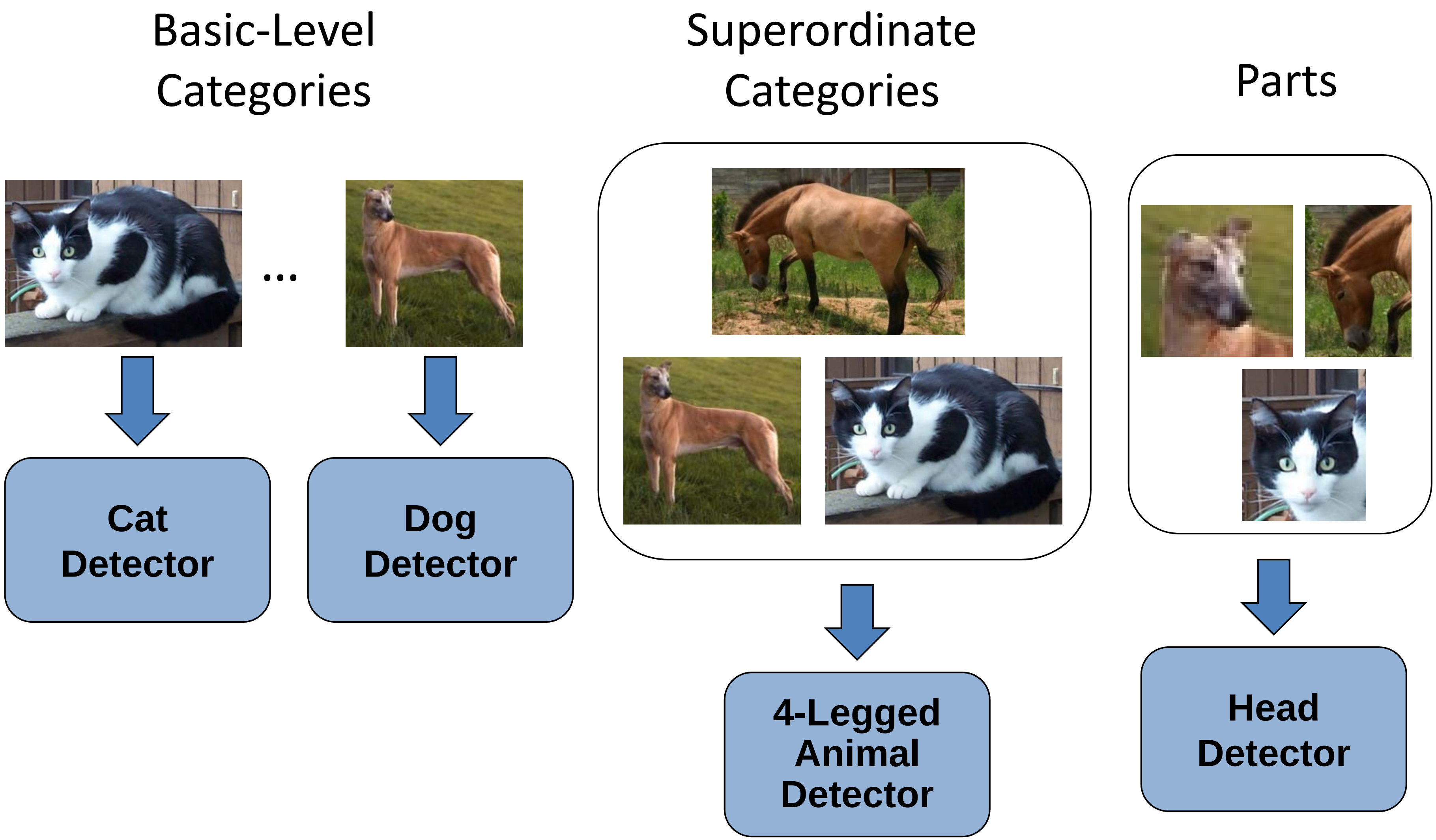
Advantages of supervised attributes

- Provides correspondence for objects from different categories



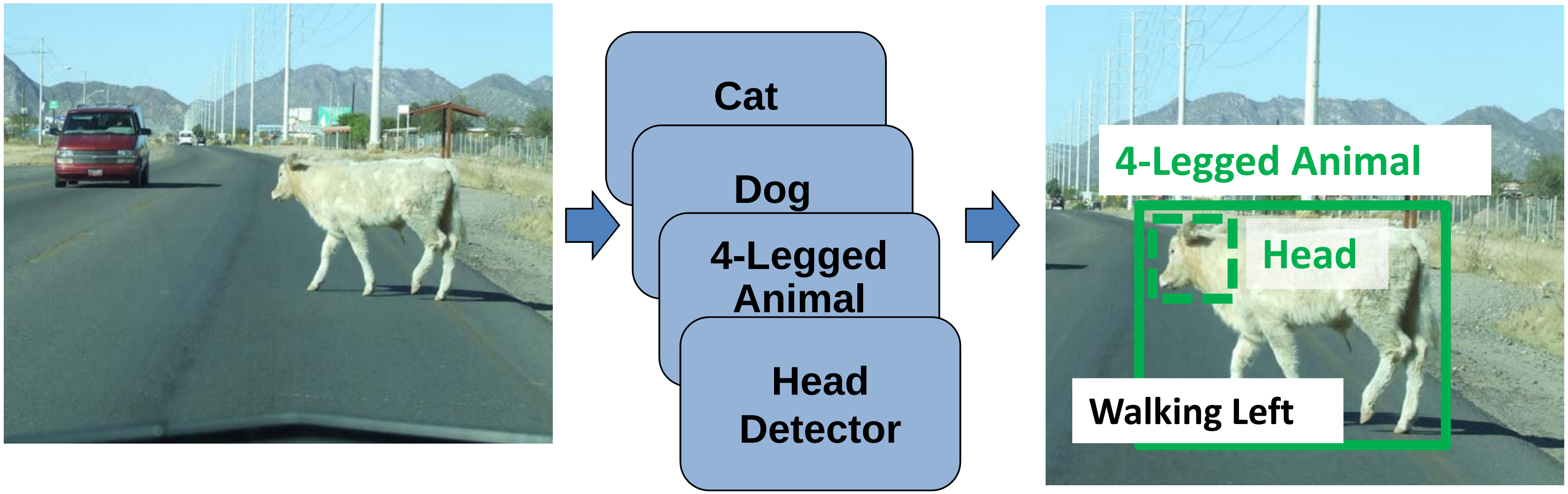
A. Farhadi, I. Endres, and D. Hoiem 2010

Domain-based Recognition



A. Farhadi, I. Endres, and D. Hoiem 2010

Domain-based Recognition



A. Farhadi, I. Endres, and D. Hoiem 2010

Domain-based recognition: overview

Trained Detectors

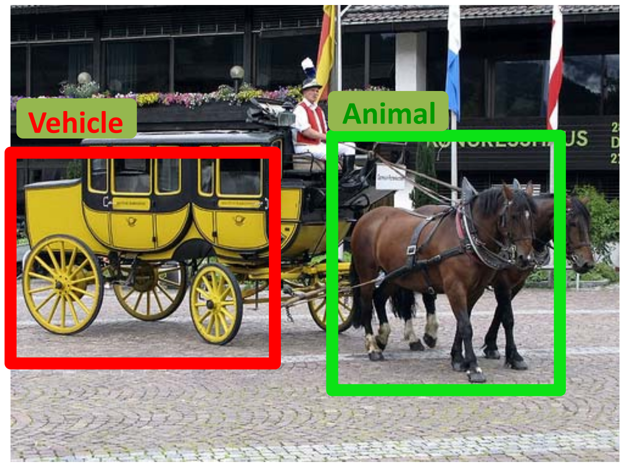
Basic Level Categories
Elephant, Dog, Eagle,
Camel, Lizard, Bat,
Dog, Penguin, Monkey, …

Broad Categories
Four-legged Animal,
Mammal, Water Animal,
Animal

Parts
Leg, Horn, Wing, Head, Eye,
Ear, Foot, Mouth, Nose, Tail

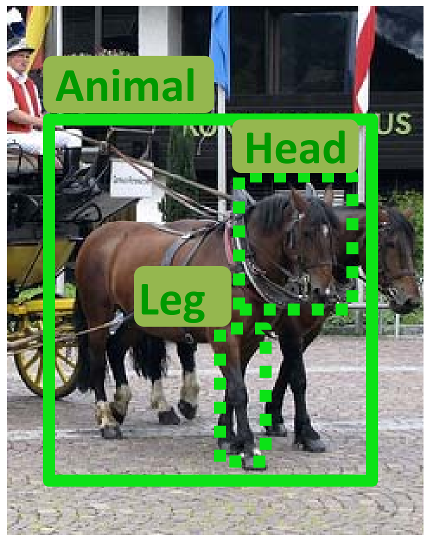
Voting using Shared Spatial Models

Object Localization



Attribute Predictors

Object Description



Four-legged
Mammal

Can run
Can Jump
Is Herbivorous
Facing right

A. Farhadi, I. Endres, and D. Hoiem 2010

CORE Dataset

Cross-category **O**bject **RE**cognition

- 2780 Images – from ImageNet
- 3192 Objects – 28 Categories
- 26695 Parts – 71 types
- 30046 Attributes – 34 types
- 1052 Material Images – 10 types

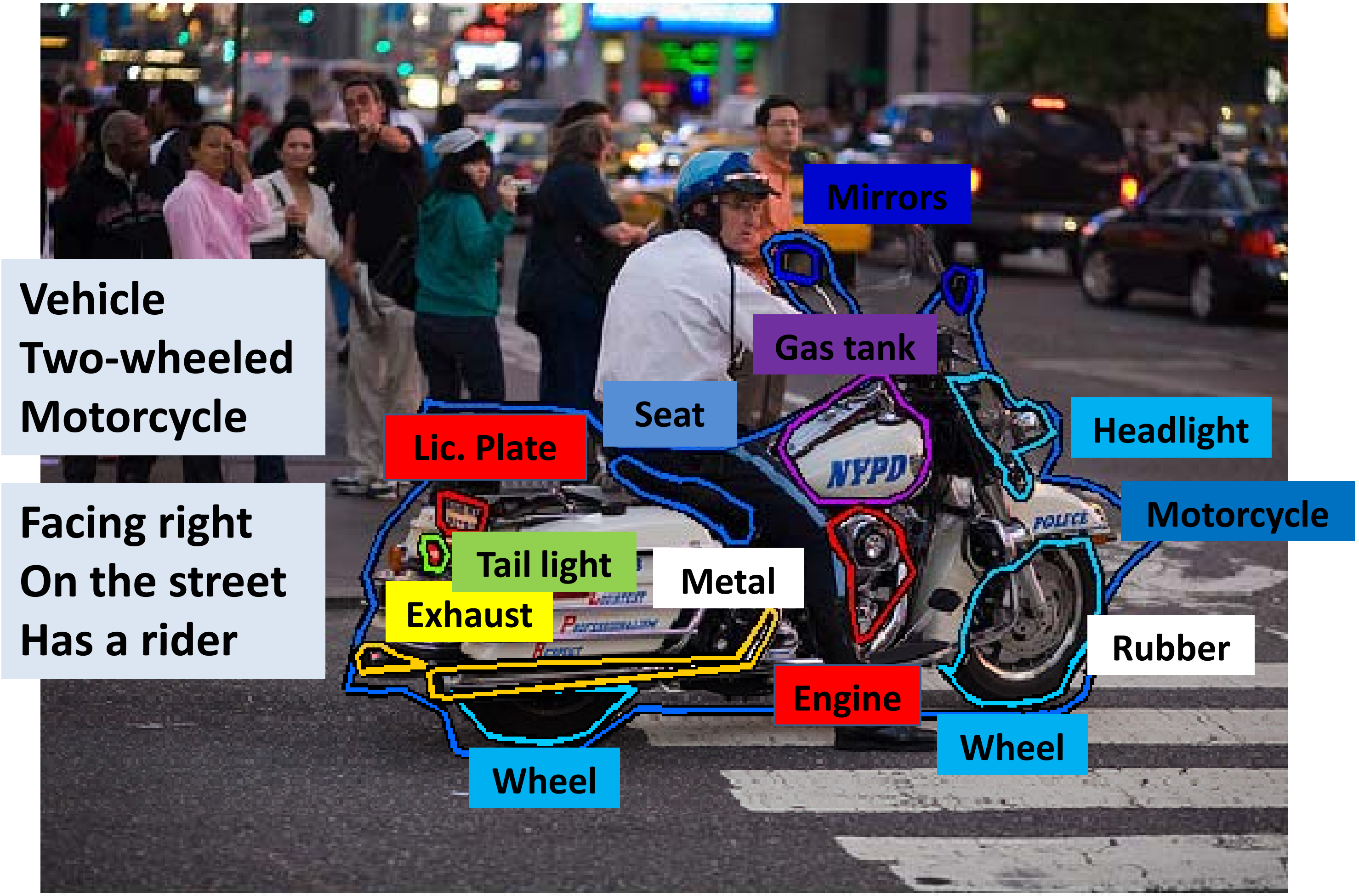
Download or browse online:

http://vision.cs.uiuc.edu/CORE

A. Farhadi, I. Endres, and D. Hoiem 2010

CORE Dataset

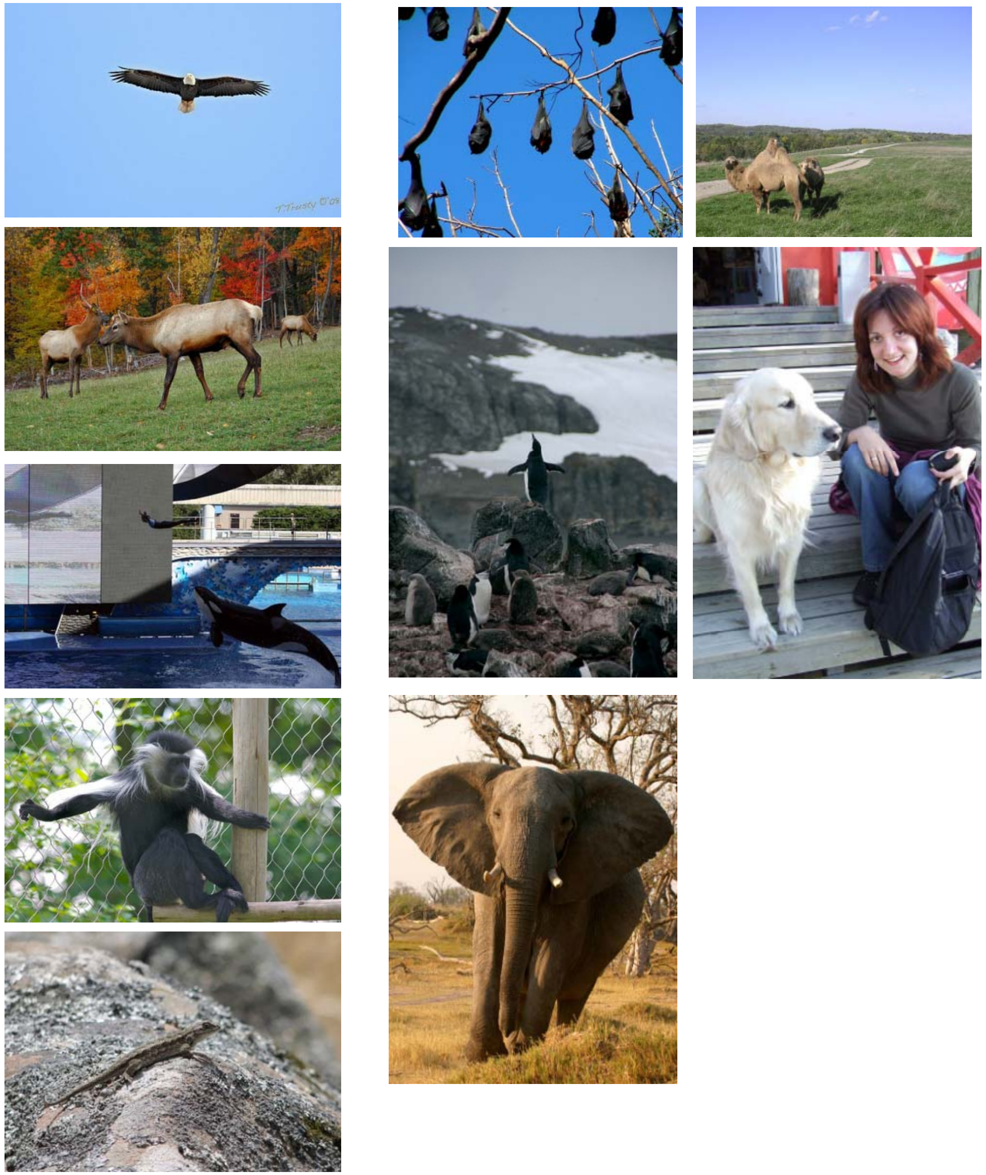
Annotation Example



A. Farhadi, I. Endres, and D. Hoiem 2010

Dataset examples: animals

Categories Seen During Training and Testing



Categories Seen Only During Testing

A. Farhadi, I. Endres, and D. Hoiem 2010

Dataset examples: vehicles

Categories Seen During Training and Testing







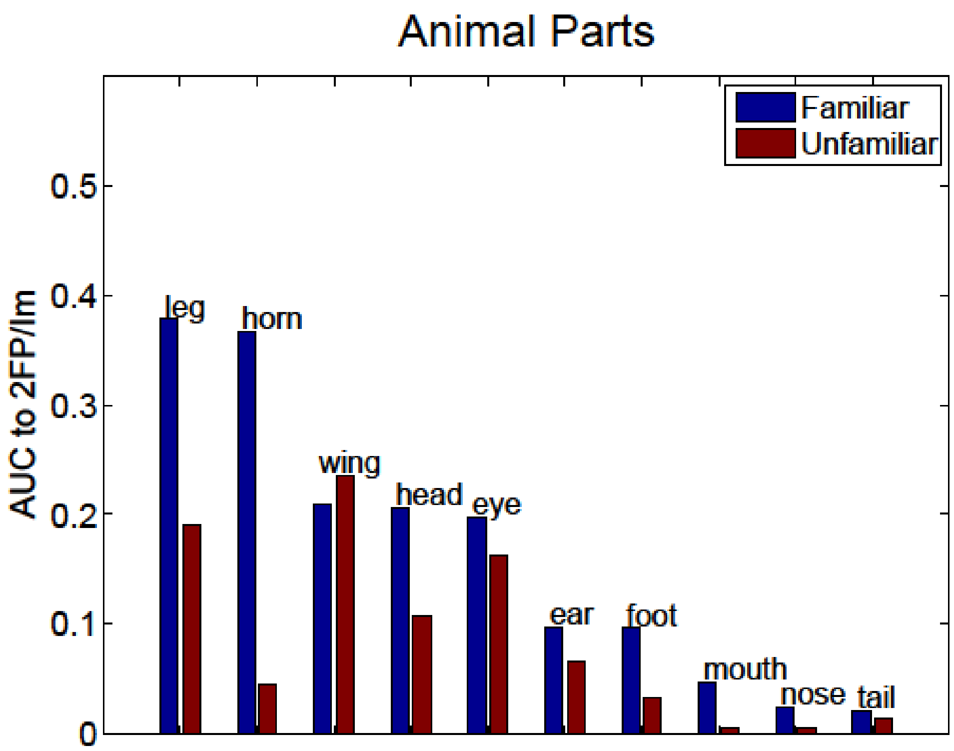


Categories Seen Only During Testing

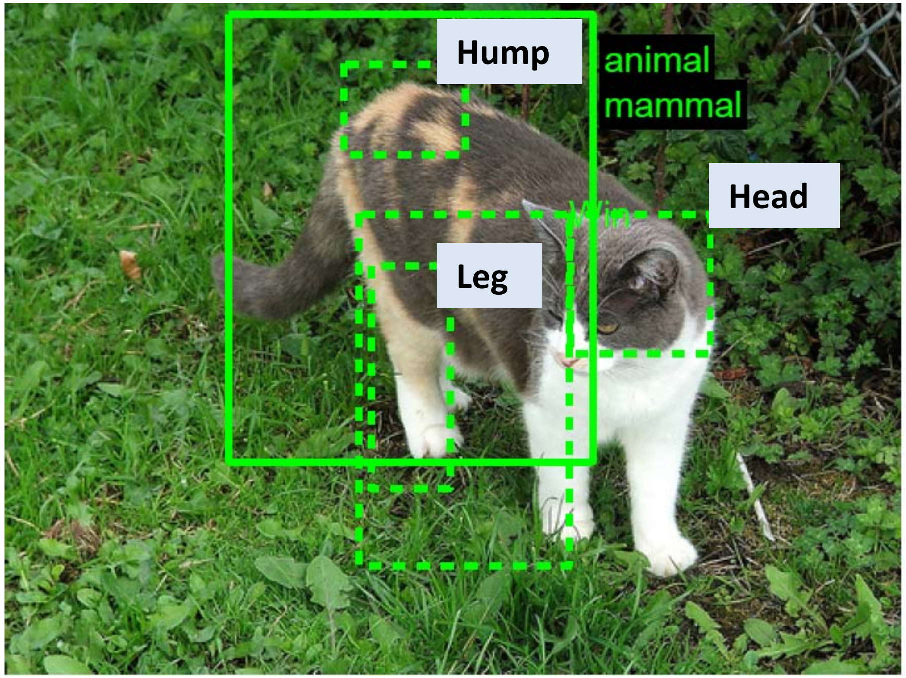



A. Farhadi, I. Endres, and D. Hoiem 2010

Result: Part detectors can generalize across categories

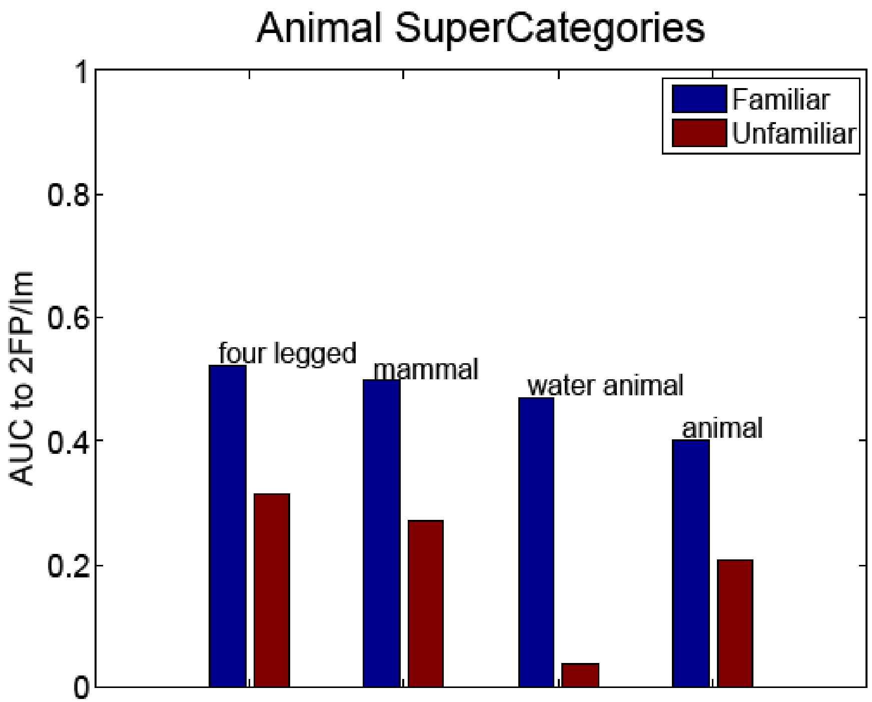


Part Detections for Novel Object

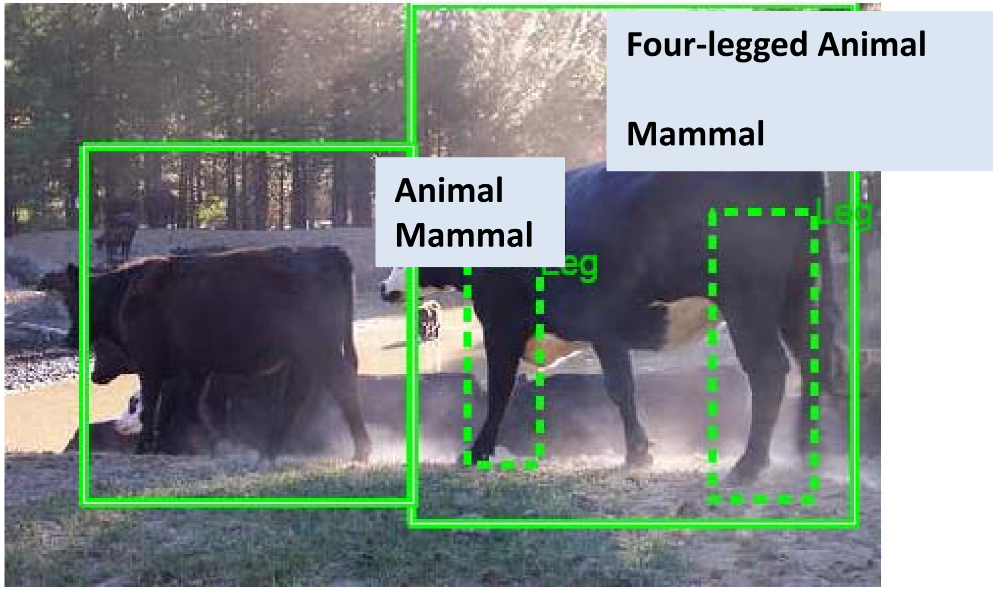


Detectors trained using (Felzenszwalb Girshik McAllester Ramanan 2009) method

Result: Broad category detectors can generalize across basic categories

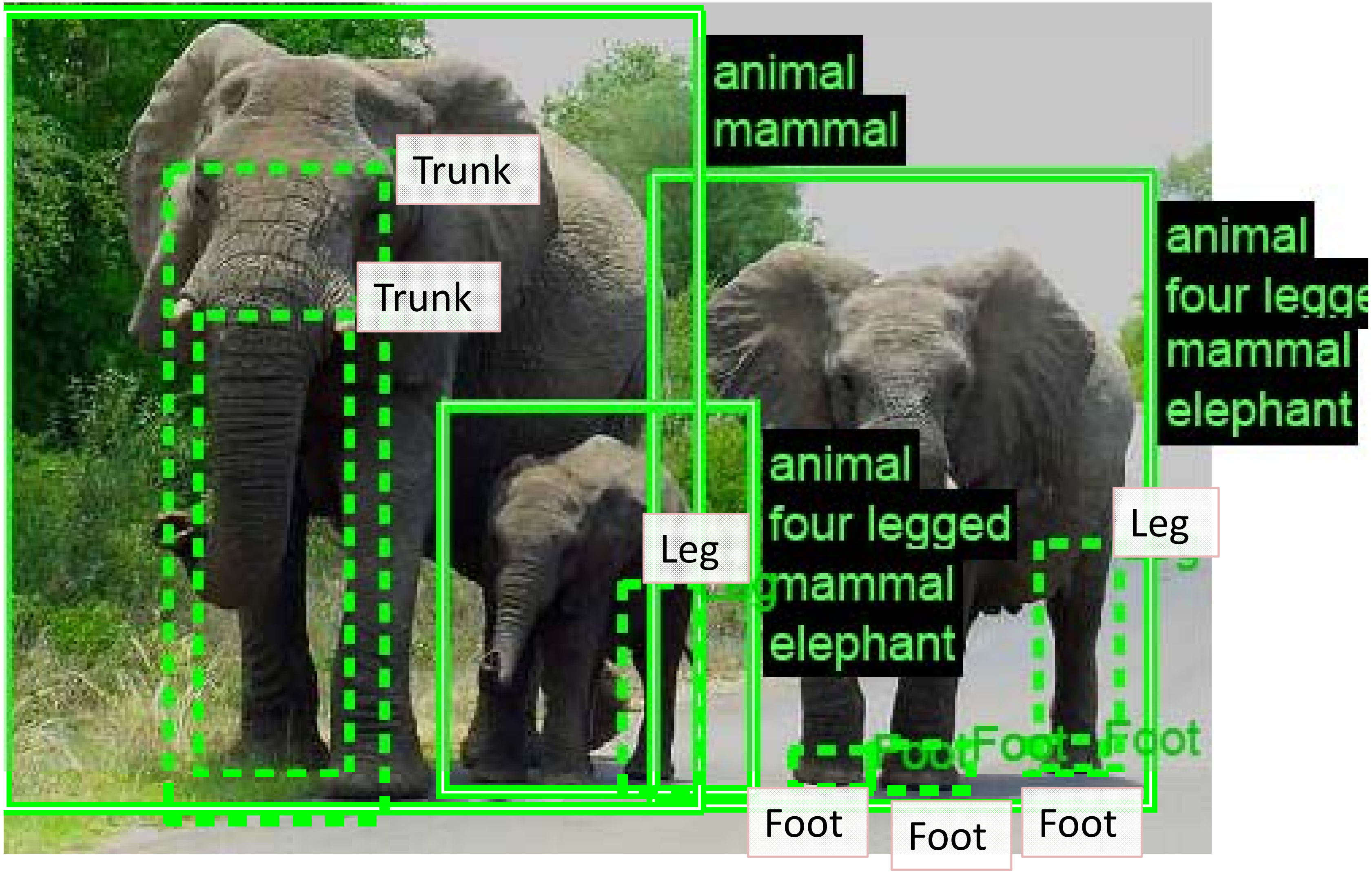


Category Detections for Novel Object



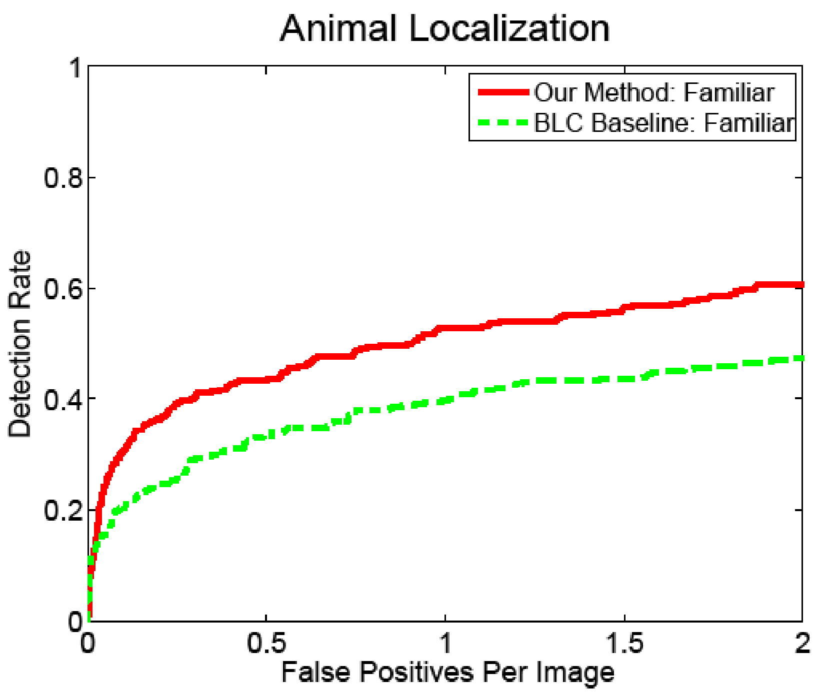
Detectors trained using (Felzenszwalb Girshik McAllester Ramanan 2009) method

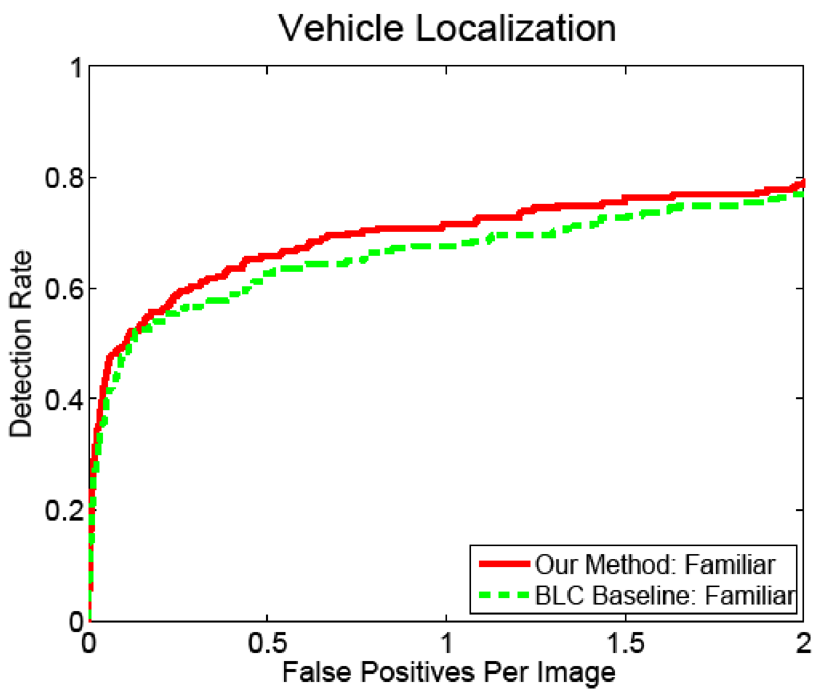
describe objects from familiar categories



describe objects from familiar categories

ROC for Localization of Familiar Objects

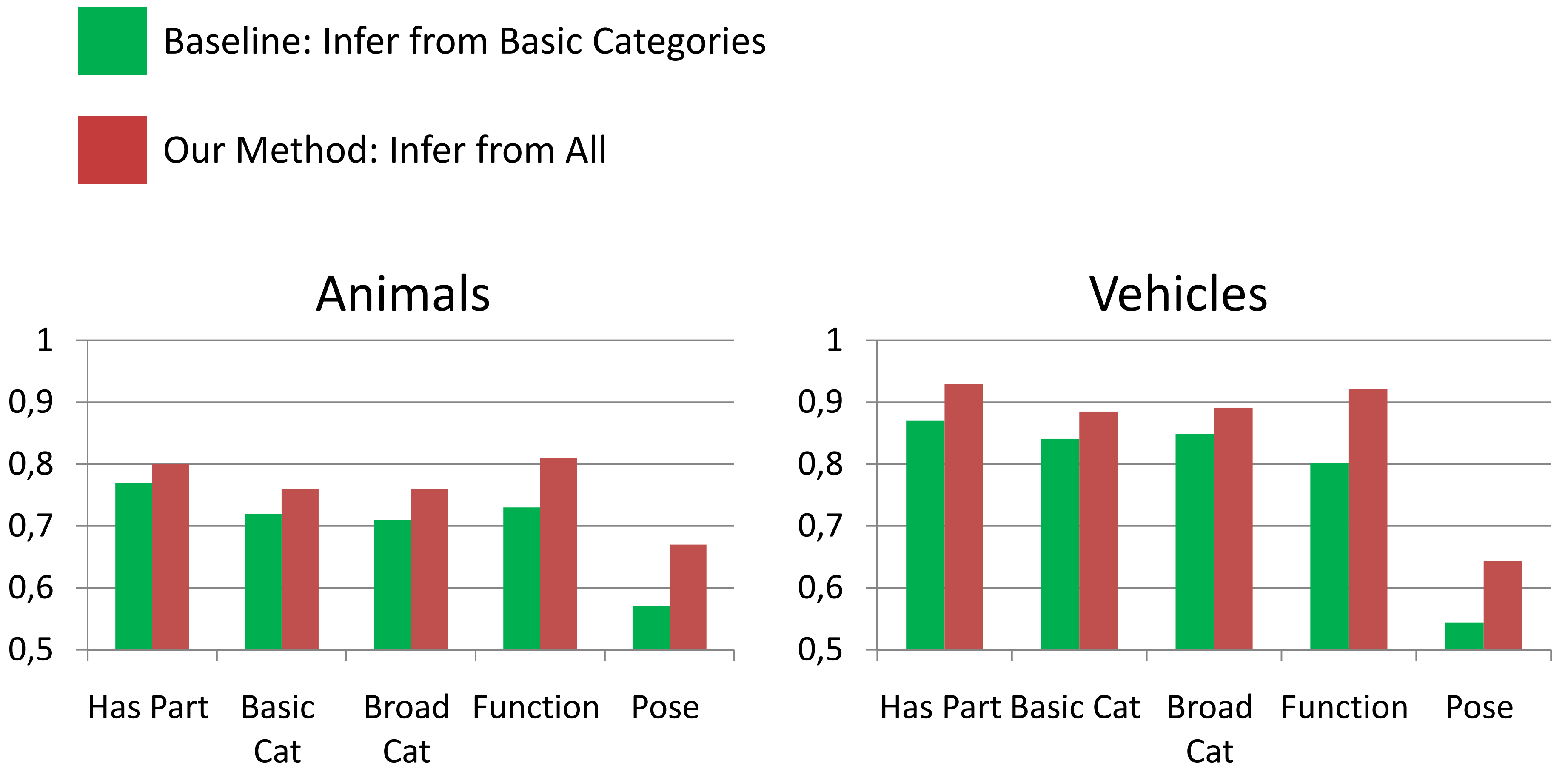




A. Farhadi, I. Endres, and D. Hoiem 2010

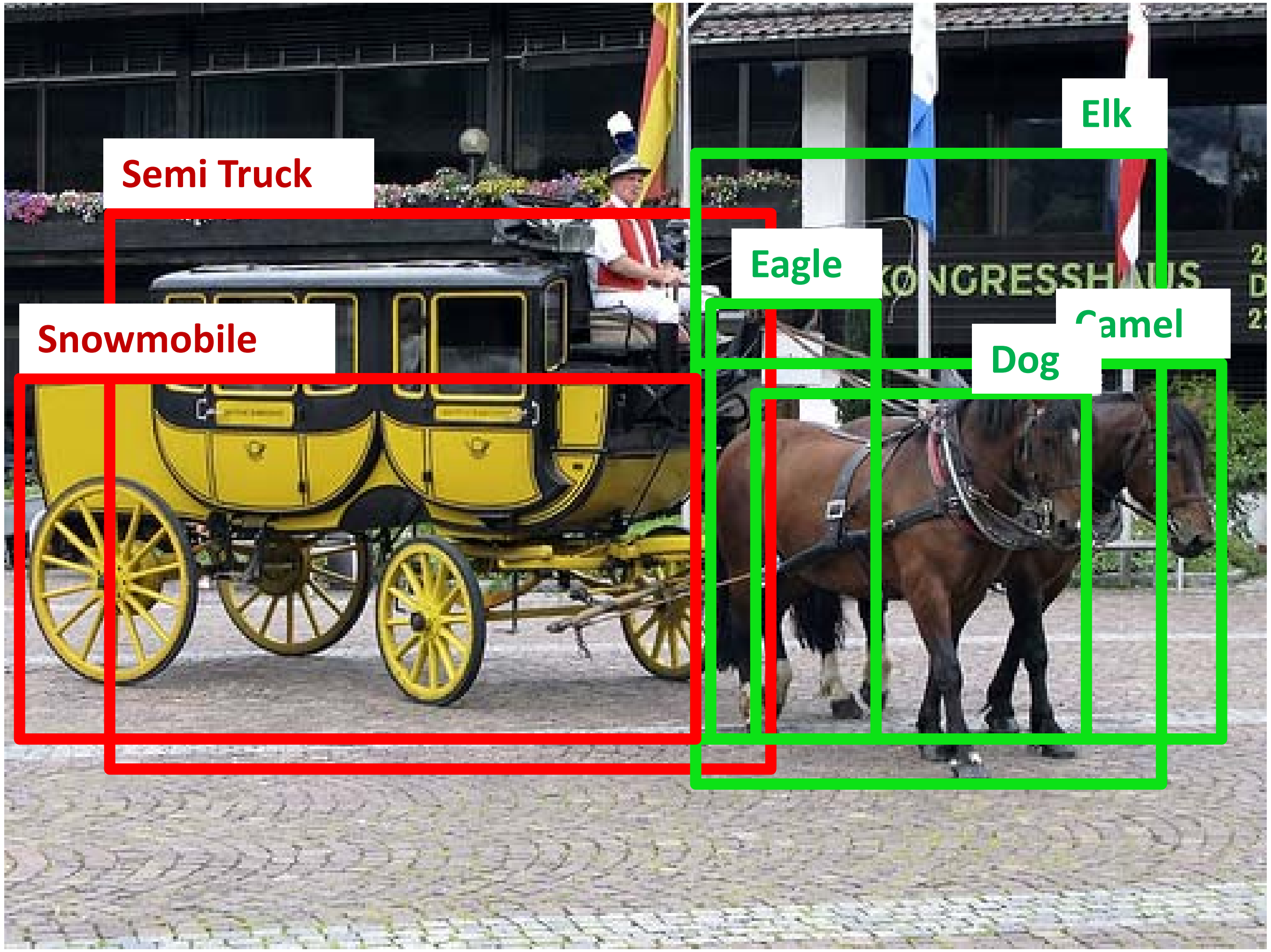
describe objects from familiar categories

AUC for Attribute Prediction for Familiar Objects



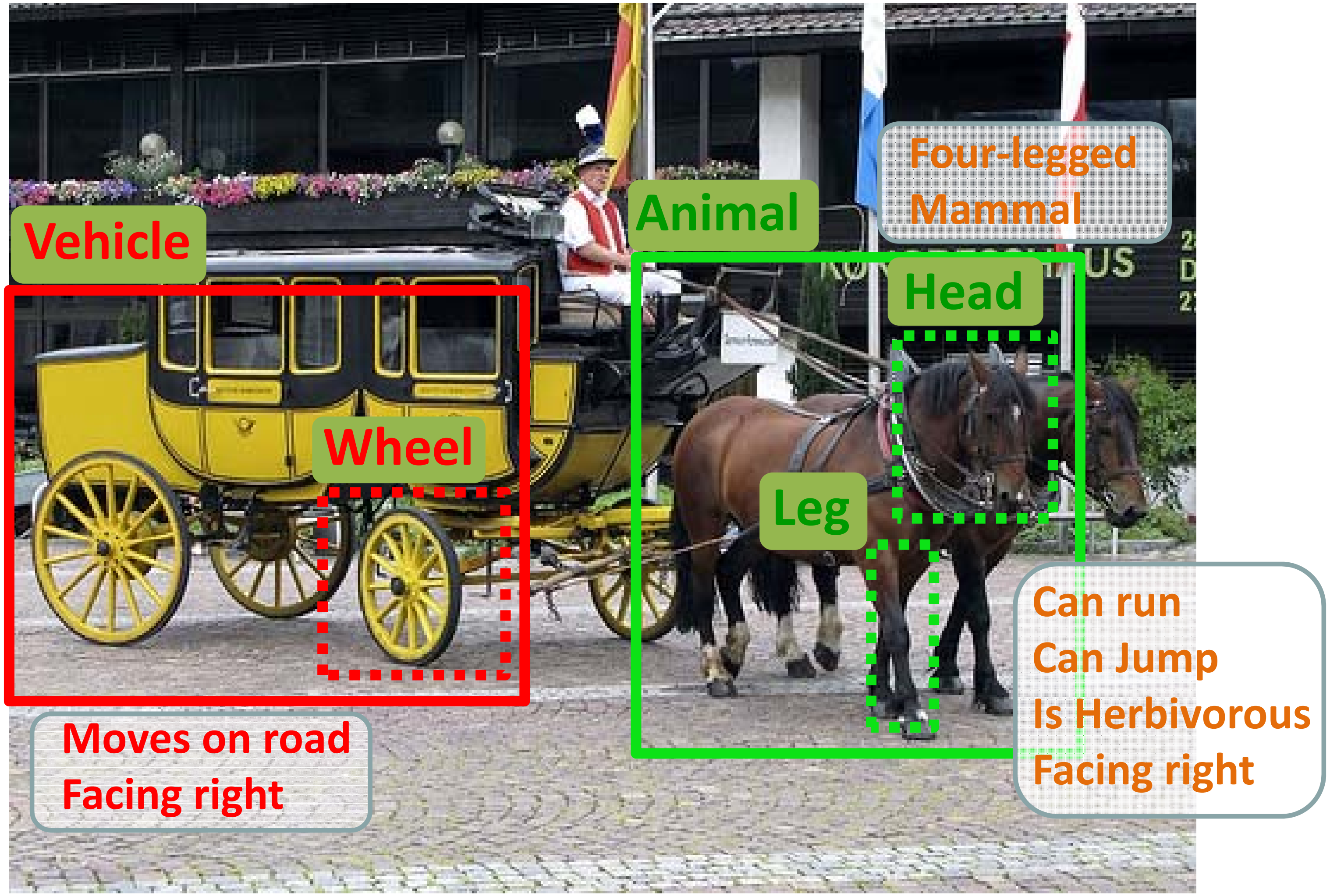
A. Farhadi, I. Endres, and D. Hoiem 2010

Result using only basic categories



A. Farhadi, I. Endres, and D. Hoiem 2010

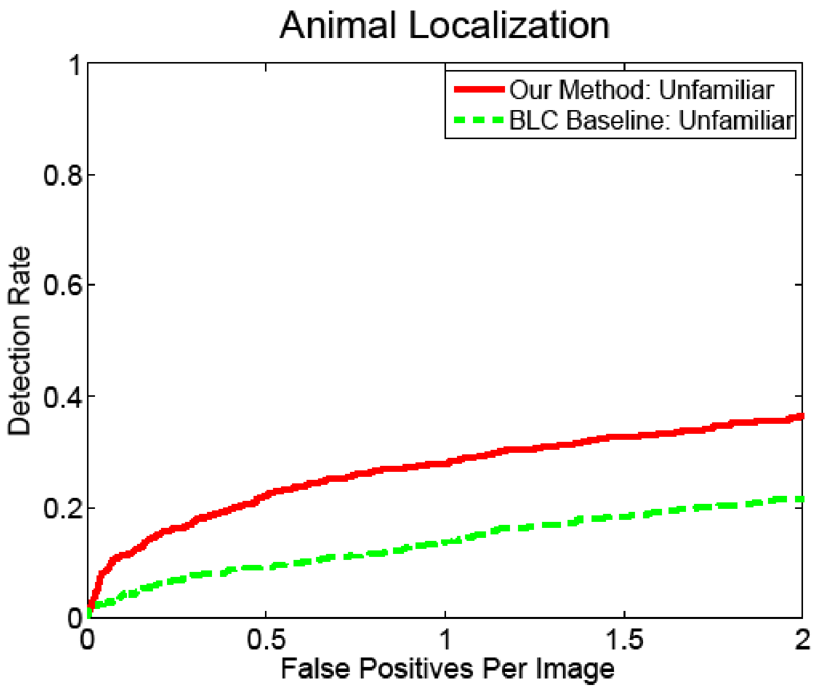
Result 3: We can find and describe objects from novel categories

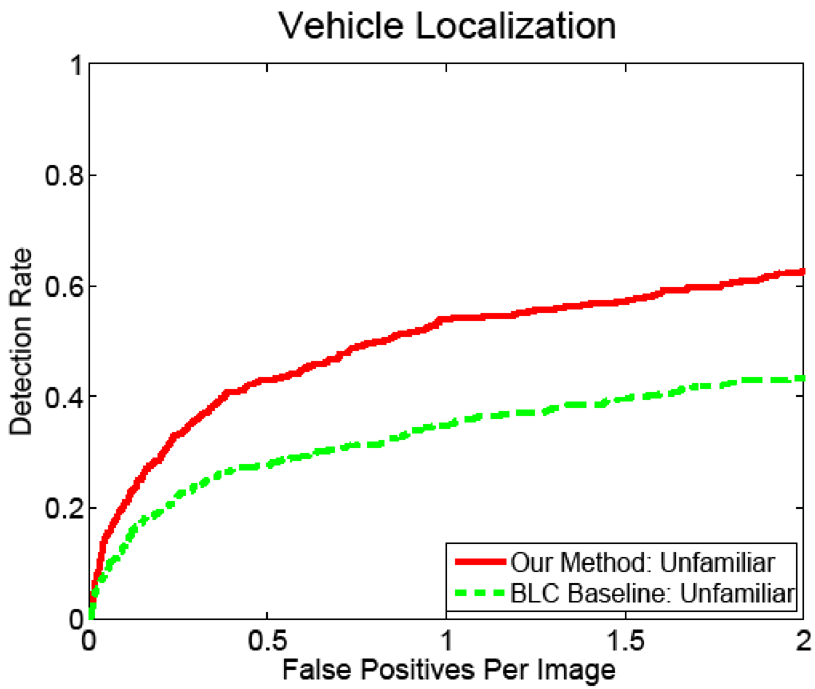


A. Farhadi, I. Endres, and D. Hoiem 2010

Result 3: We can find and describe objects from *novel categories*

ROC for Localization of Unfamiliar Objects

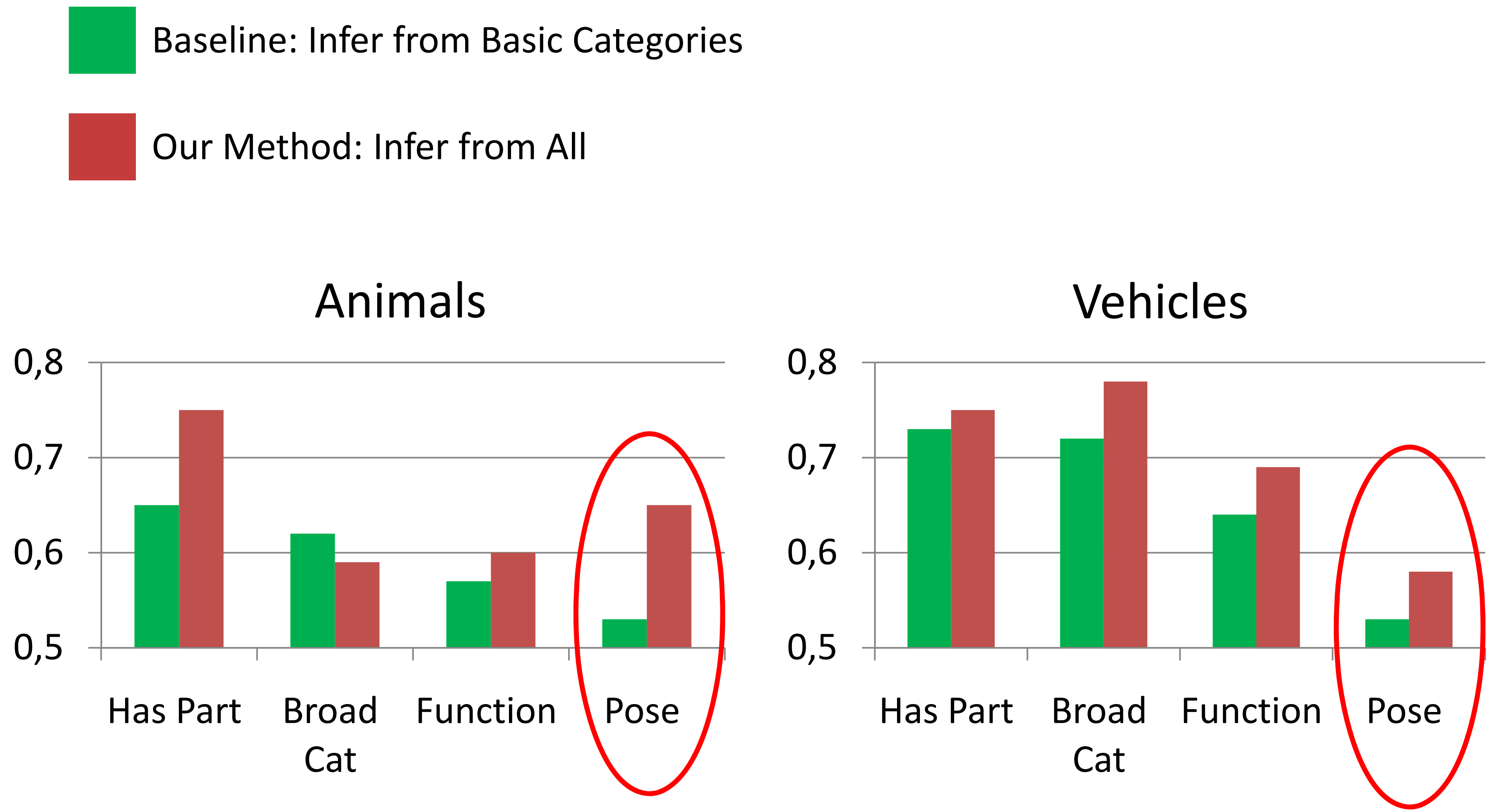




A. Farhadi, I. Endres, and D. Hoiem 2010

Result 3: We can find and describe objects from *novel categories*

AUC for Attribute Prediction for Unfamiliar Objects



Summary of Findings

- Current detectors are good enough to recognize general parts and broad categories
- Learning to recognize parts and broad categories improves both detection and description
- By going beyond categories, we can partially recognize novel objects

A. Farhadi, I. Endres, and D. Hoiem 2010