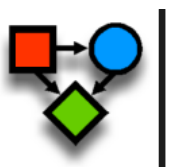
From Concepts to Events: A Progressive Process for Multimedia Content Analysis

Nicu Sebe
University of Trento
sebe@disi.unitn.it

with Zhigang Ma, Jasper Uijlings, Yi Yang, Alex Hauptmann

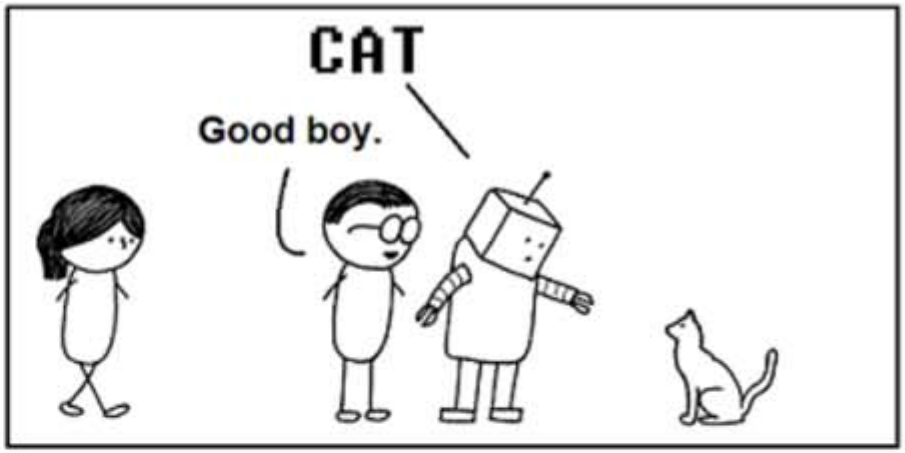
Research Group

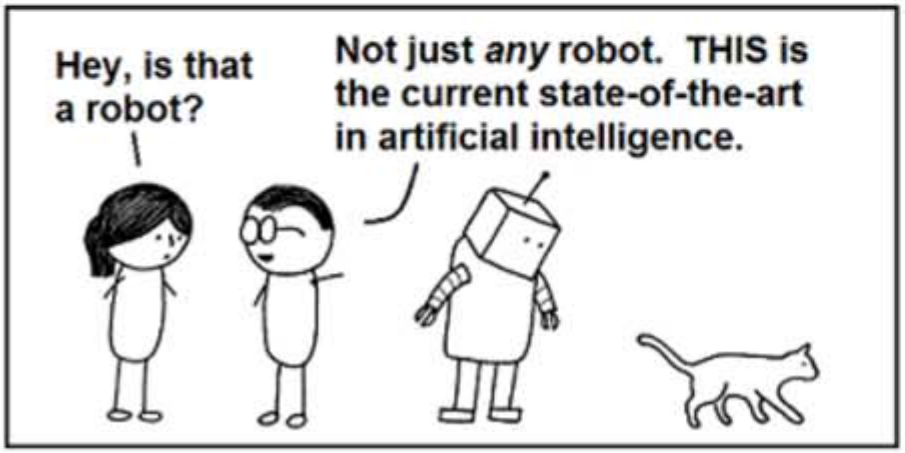
- M-HUG: Multimedia & Human Understanding Group (http://mhug.disi.unitn.it/)
  - 12 PhD students: China, Romania, Belarus, Iran, Brazil, US, Italy
  - 8 PostDocs: China, Romania, Serbia, Spain, Italy, Iran

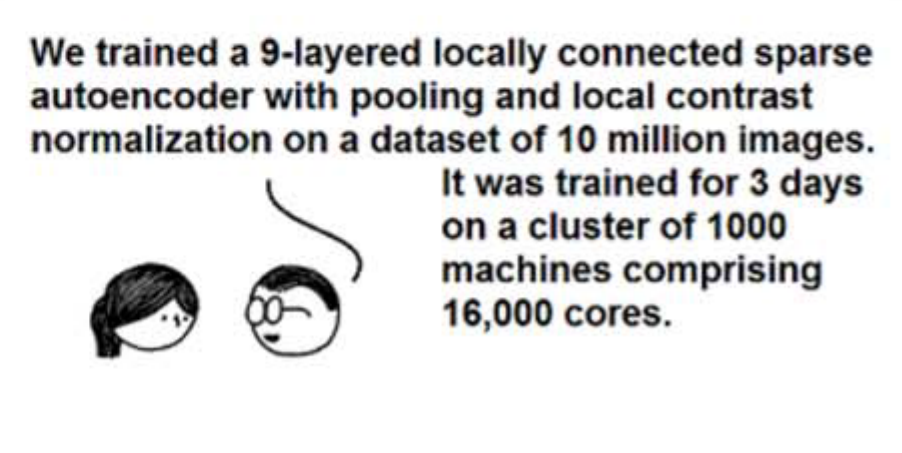


Joining the protest!!!!

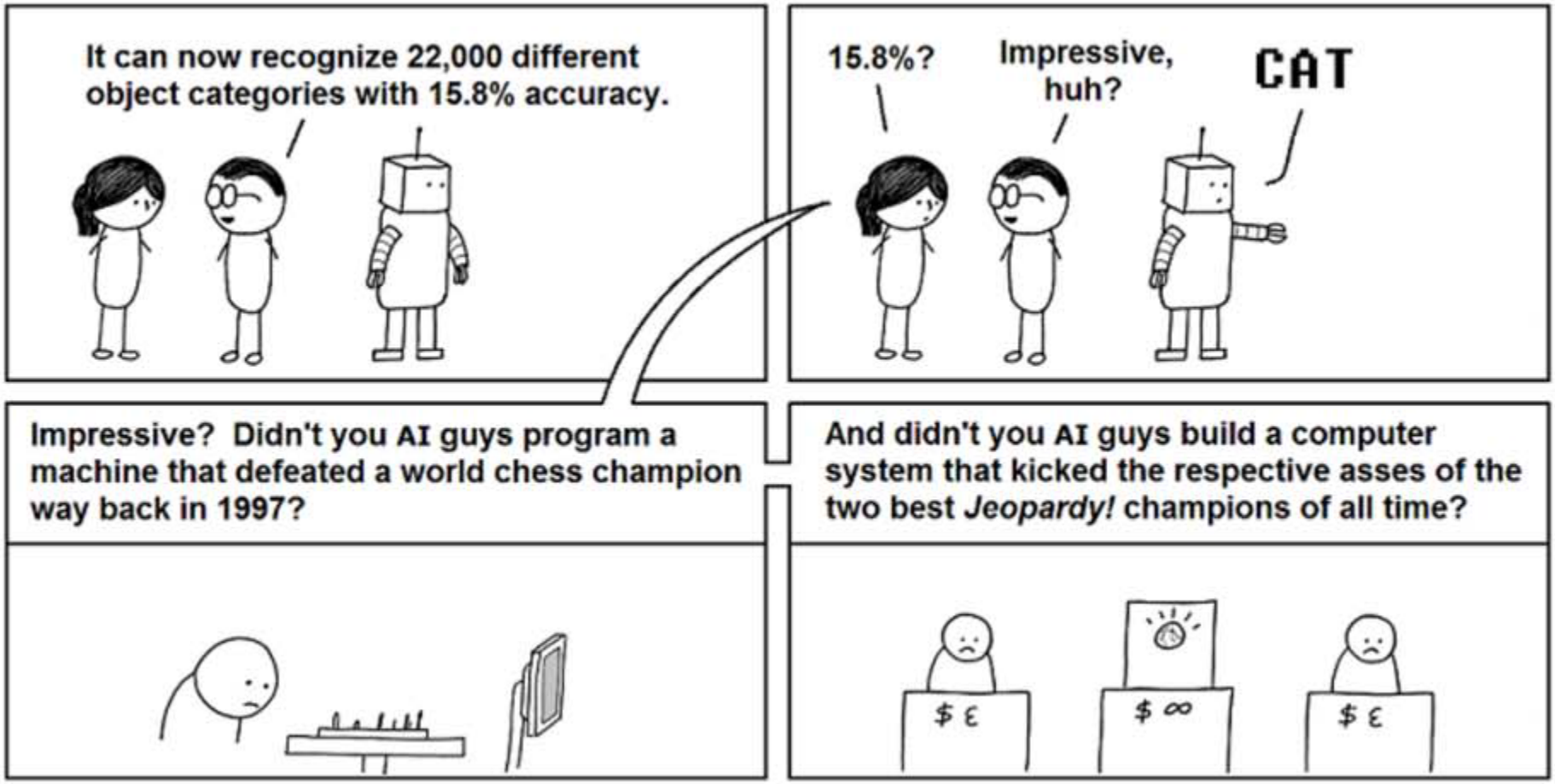
State of the art?



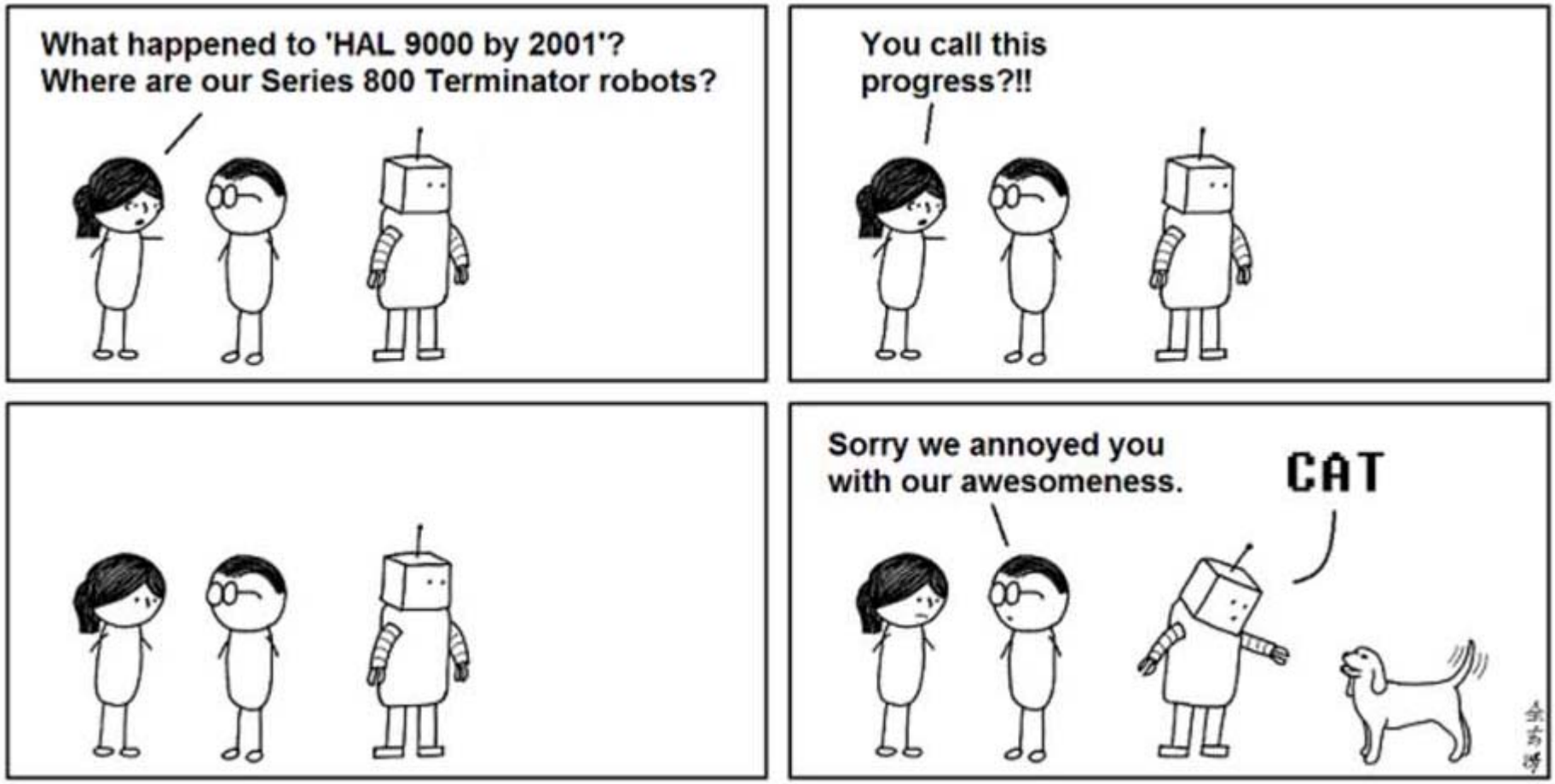




State of the art?



State of the art?



How to attain better multimedia content understanding?

Feature Selection

Noise?

Redundancy?

QUESTION 1:

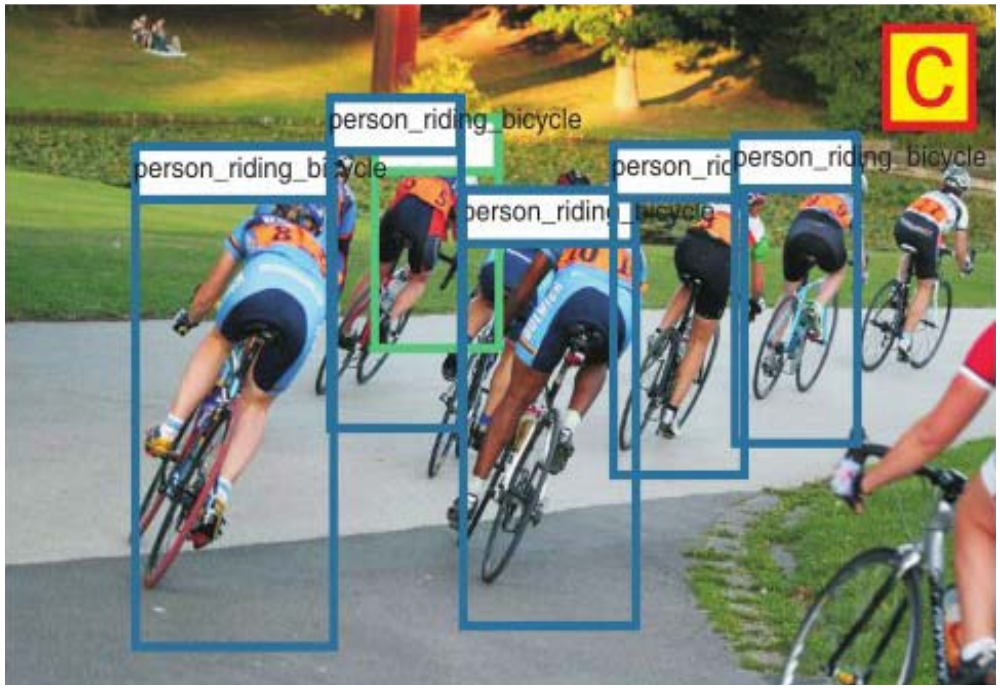
Is it possible to get a compact feature representation? Would the accuracy be improved as a result?

Semi-supervised Learning

QUESTION 2:

Is there any way to attain a reasonable performance when only few labeled images and videos are available?

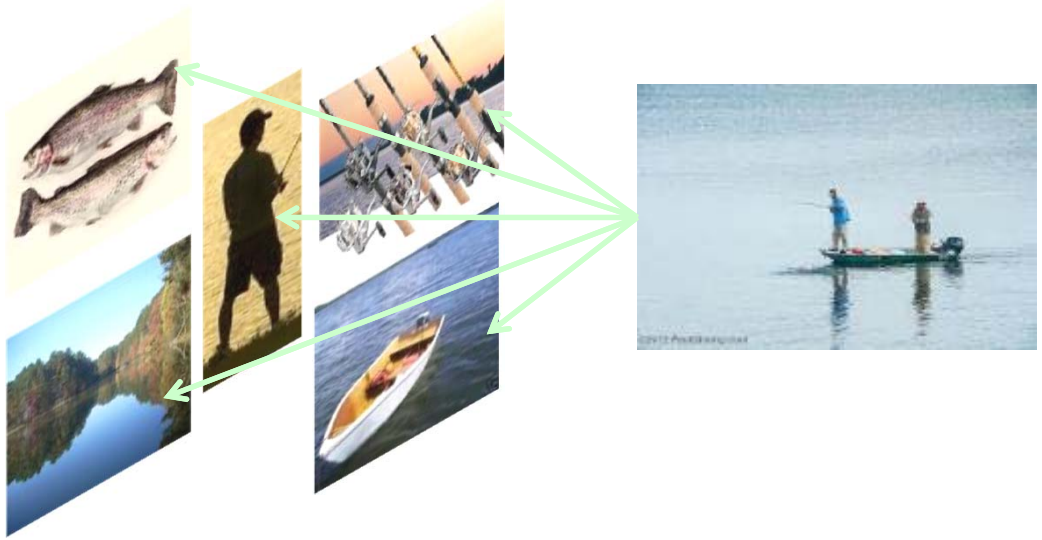
Multimodal Analysis



QUESTION 3:

Can we use other modalities (e.g., text) to improve the analysis? Can visual information help text retrieval?

Classifier-specific Representation



QUESTION 4:

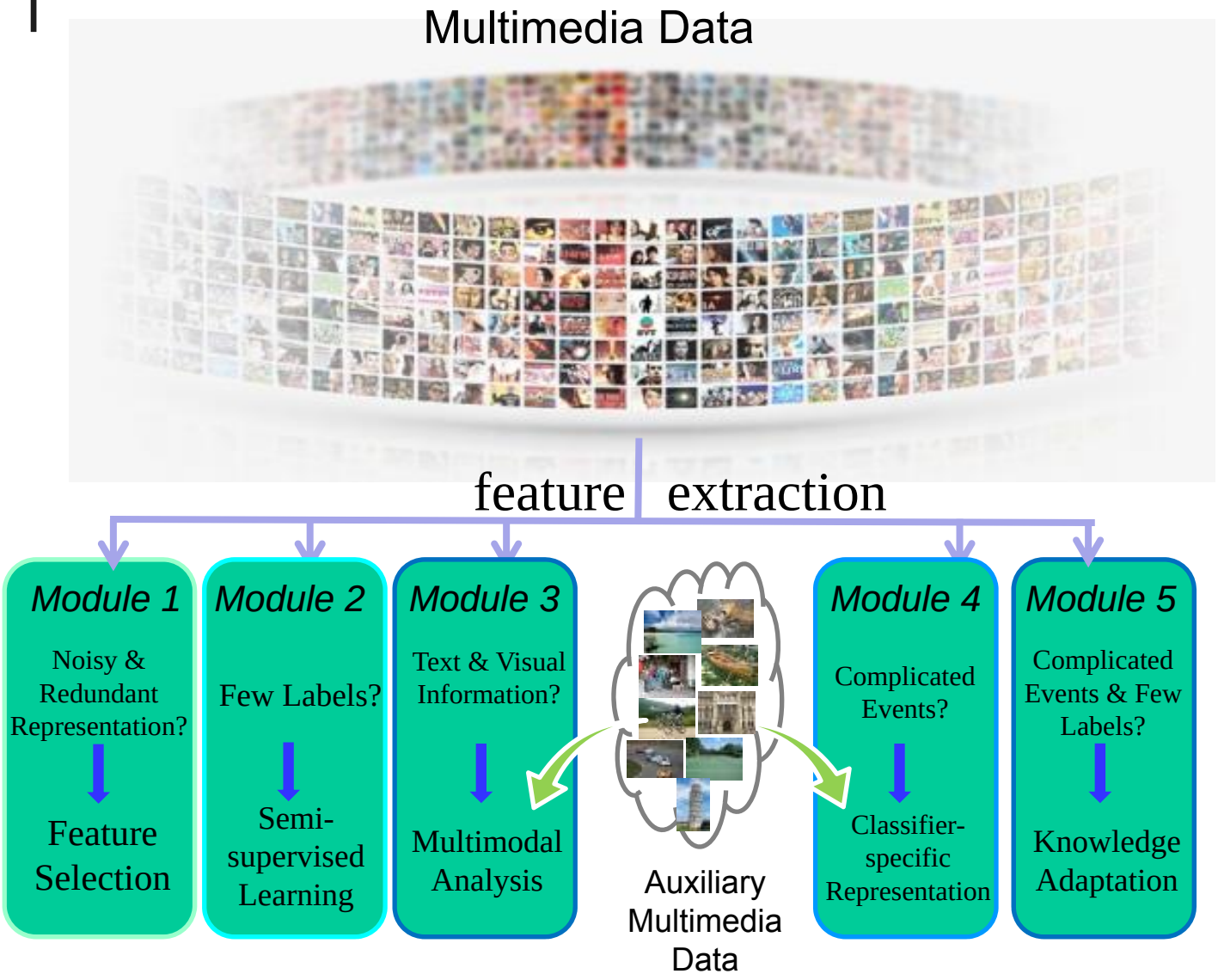
Can we skip the explicit concept detection process but learn an intermediate representation using the available multimedia archives related to various concepts for complicated events?

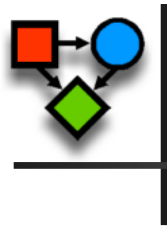
Knowledge Adaptation

QUESTION 5:

How can we guarantee reasonable multimedia event detection accuracy when only few positive exemplars are provided?

A Generic Framework





Things and Stuff:
A Computer Vision Perspective

Some slides courtesy of Heitz & Koller, Uijlings et. al

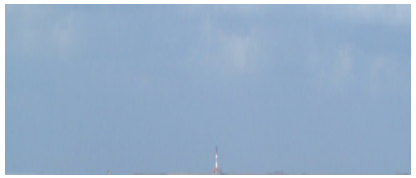
Things vs. Stuff

Thing: An object with a specific size and shape






Stuff: Material defined by a homogeneous or repetitive pattern of fine-scale properties, but has no specific or distinctive spatial extent or shape

Finding Things

Context is key!

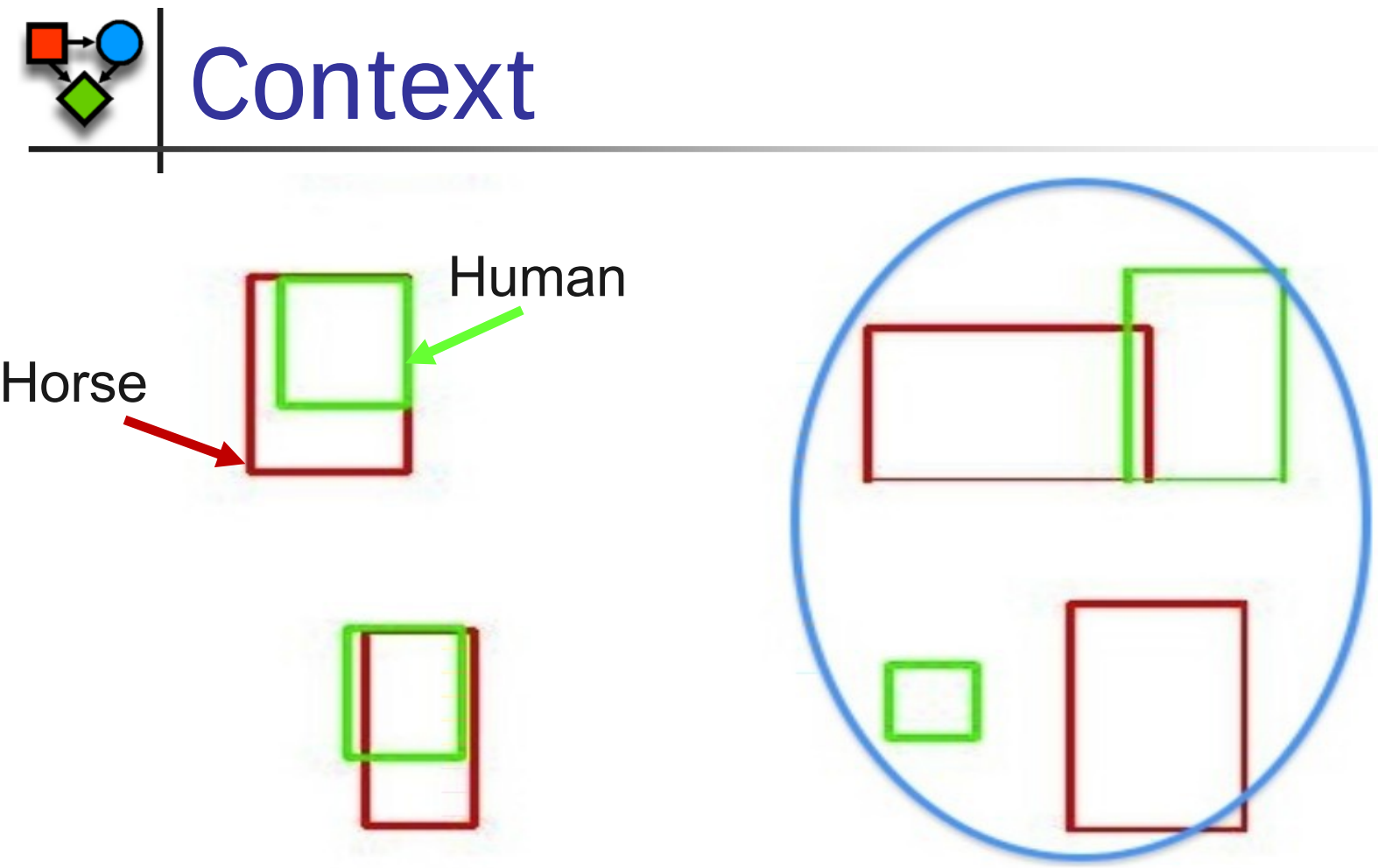
Context

Zürich: a city and its trams

D. Gatica-Perez

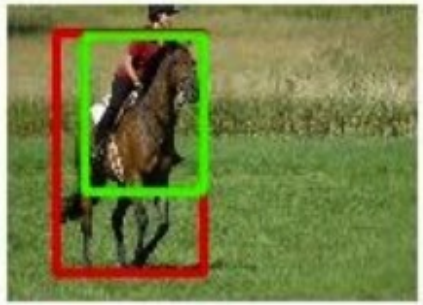
Context


Das Rauchen ist in sämtlichen Fahrzeugen des ZVV verboten.
Reisen Sie immer mit einem gültigen Fahrausweis.
Darbietungen, Werbeaktionen, Betteln und Unterschriftensammlungen sind verboten.
Das Beschädigen der Einrichtung hat strafrechtliche Konsequenzen.
Parkieren Sie Ihre Schuhe nicht auf den Sitzflächen.
en Fahrausweis bezahlen Sie mindestens einen Zuschlag von CHF 80.–, beim zweiten Mal von CHF 120.– beim dritten Mal von CHF 150.–. Ihre Personalien werden in jedem Fall aufgenommen.
ravelling without a valid ticket may be required to pay a penalty of CHF 80.– or more (CHF 120.– and 150.– for second and third offense). Their identity will be on record.



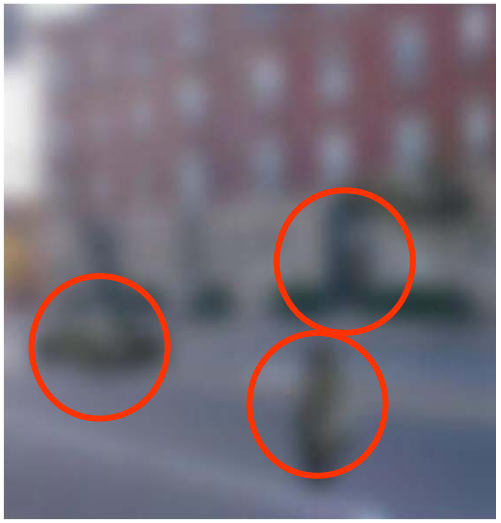
Disambiguation using relative locations of detected boxes

Context



Riding Horse or Feeding Horse?

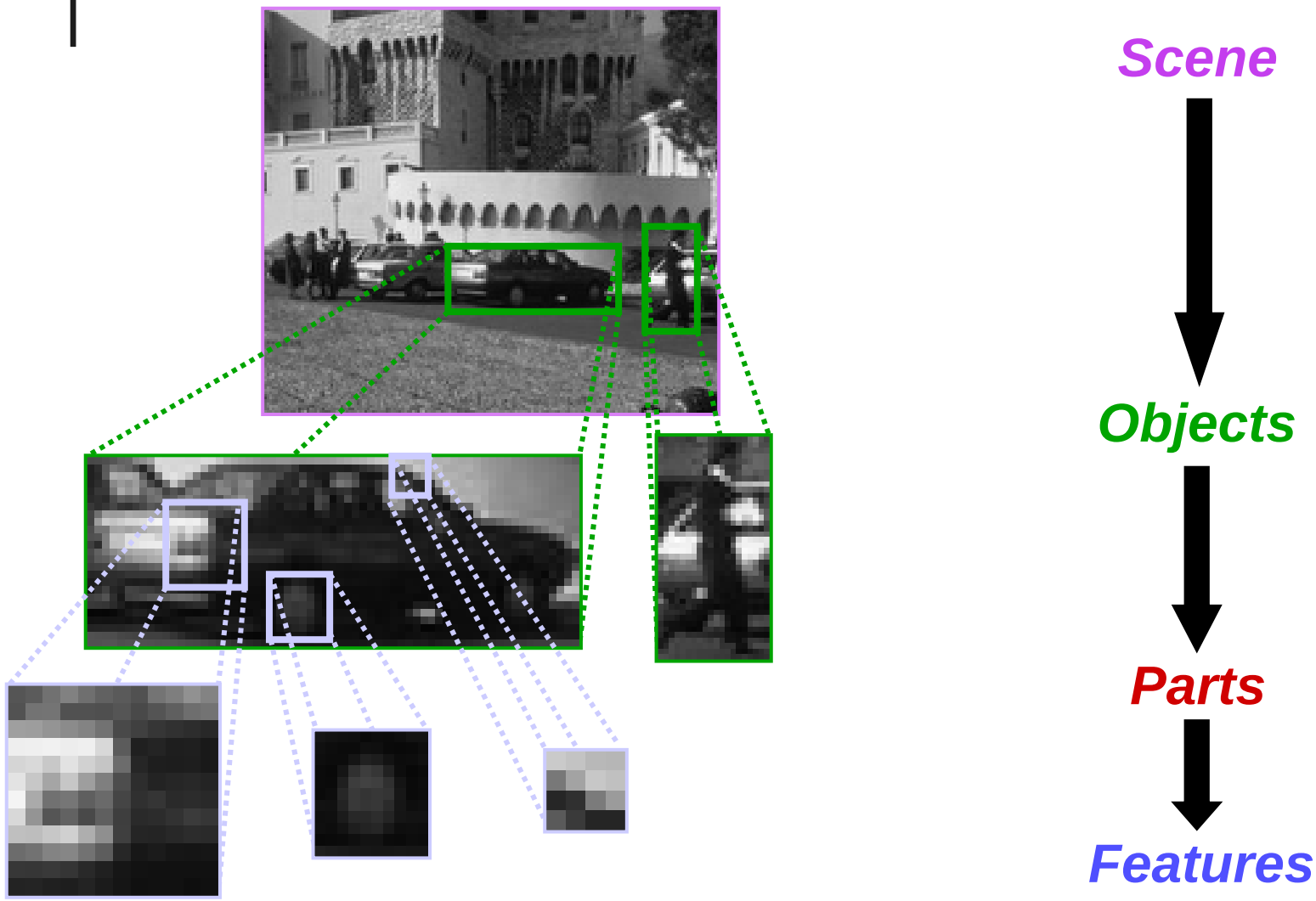
Context



Outline

- Sliding Window
- Selective Search vs. Sliding Window
- What is Context?
- The Things and Stuff (TAS) model
- Results

Scenes, Objects, and Parts



E. Sudderth, A. Torralba, W. Freeman, A. Willsky. ICCV 2005.

The statistical viewpoint

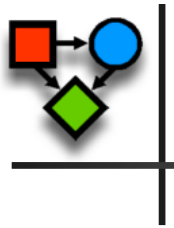
$$p(zebra \mid image)$$

vs.

$$p(no\ zebra \mid image)$$

- Bayes rule:

$$\underbrace{\frac{p(zebra \mid image)}{p(no\ zebra \mid image)}}_{\text{posterior ratio}} = \underbrace{\frac{p(image \mid zebra)}{p(image \mid no\ zebra)}}_{\text{likelihood ratio}} \cdot \underbrace{\frac{p(zebra)}{p(no\ zebra)}}_{\text{prior ratio}}$$



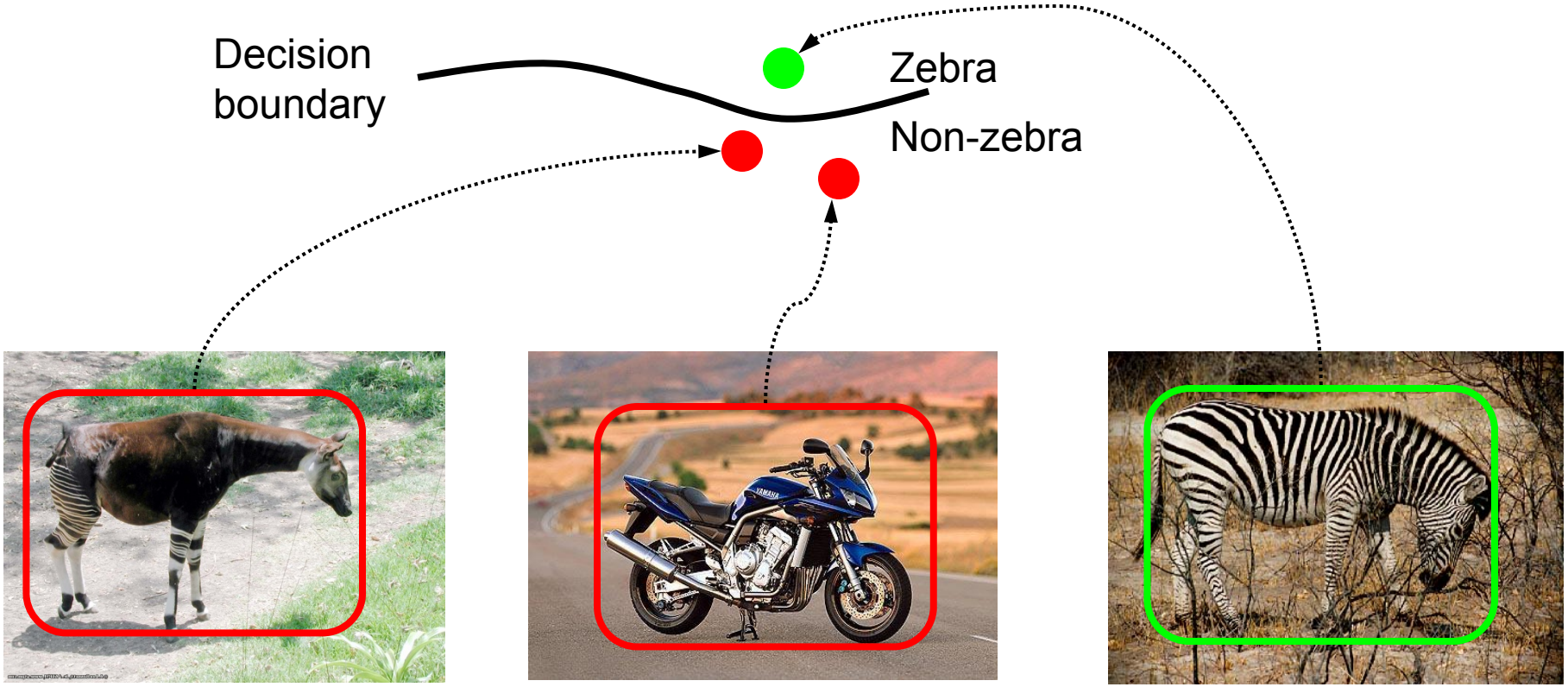
The statistical viewpoint

$$\underbrace{\frac{p(zebra \mid image)}{p(no\ zebra \mid image)}}_{\text{posterior ratio}} = \underbrace{\frac{p(image \mid zebra)}{p(image \mid no\ zebra)}}_{\text{likelihood ratio}} \cdot \underbrace{\frac{p(zebra)}{p(no\ zebra)}}_{\text{prior ratio}}$$

- **Discriminative methods model posterior**
- **Generative methods model likelihood and prior**

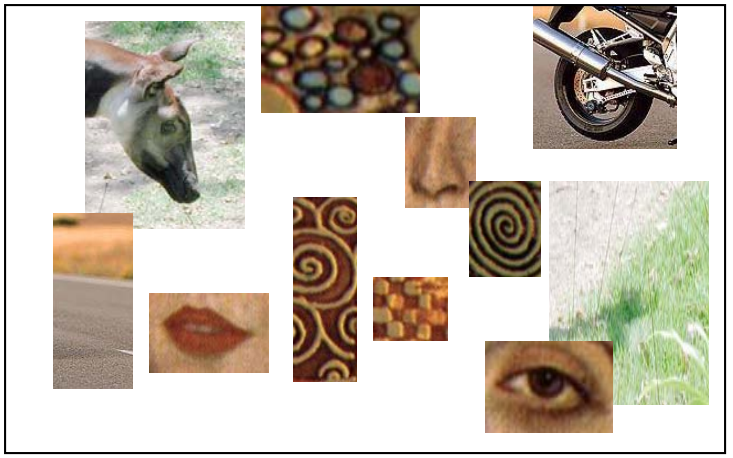
Discriminative

- Direct modeling of $\frac{p(zebra \mid image)}{p(no\ zebra \mid image)}$



Generative

- Model $p(image \mid zebra)$ and $p(image \mid no\ zebra)$



$p(image \mid zebra)$	$p(image \mid no\ zebra)$
Low	Middle
High	Middle→Low

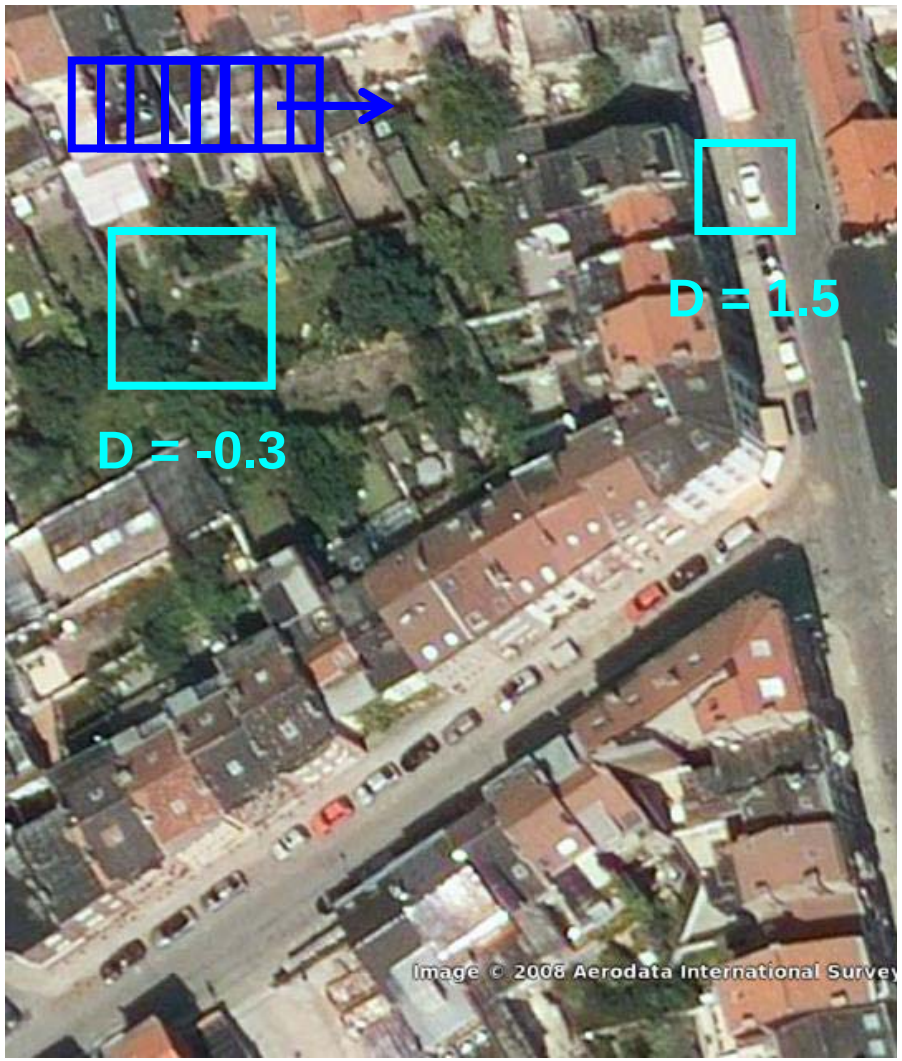
Three main issues

- Representation
  - How to represent an object category
- Learning
  - How to form the classifier, given training data
- Recognition
  - How the classifier is to be used on novel data

Object Detection

- Task: Find the things
- Example: Find all the cars in this image
  - Return a “bounding box” for each
- Evaluation:
  - Maximize true positives
  - Minimize false positives

Sliding Window Detection



- Consider every bounding box
  - All shifts
  - All scales
  - Possibly all rotations
- Each such window gets a score:
  - D(W)
- Detections: Local peaks in D(W)
- Pros:
  - Covers the entire image
  - Flexible to allow variety of D(W)'s
- Cons:
  - Brute force – can be slow
  - Only considers features in box

Object-based Classification

Given an image/video collection:

Find the objects containing a specific object

Object-based Classification

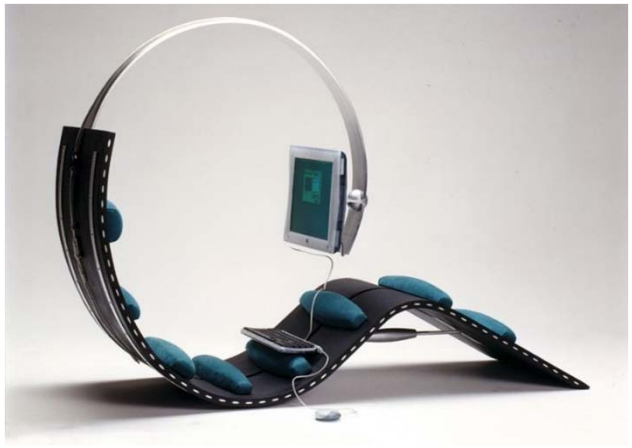
Problems:

- Viewpoint changes
- Location
- Illumination conditions

Object-based Classification

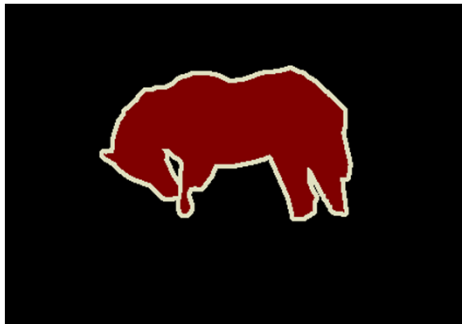
Problems:

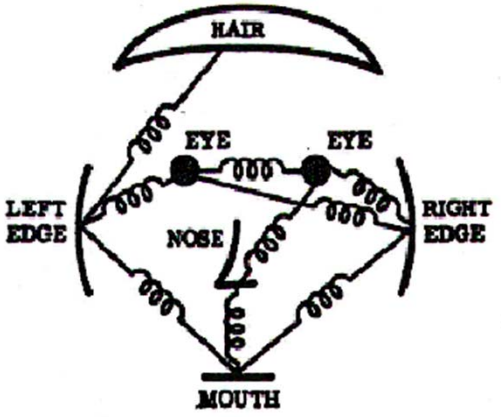
- Same functionality, different manifestations



Object-based Classification

- First intuition: First find the object, then recognize



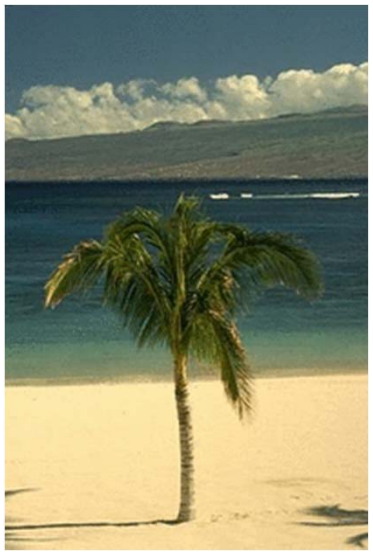


segmentation

Part-based model
(Fischler and Elschlager 1973)

Segmentation

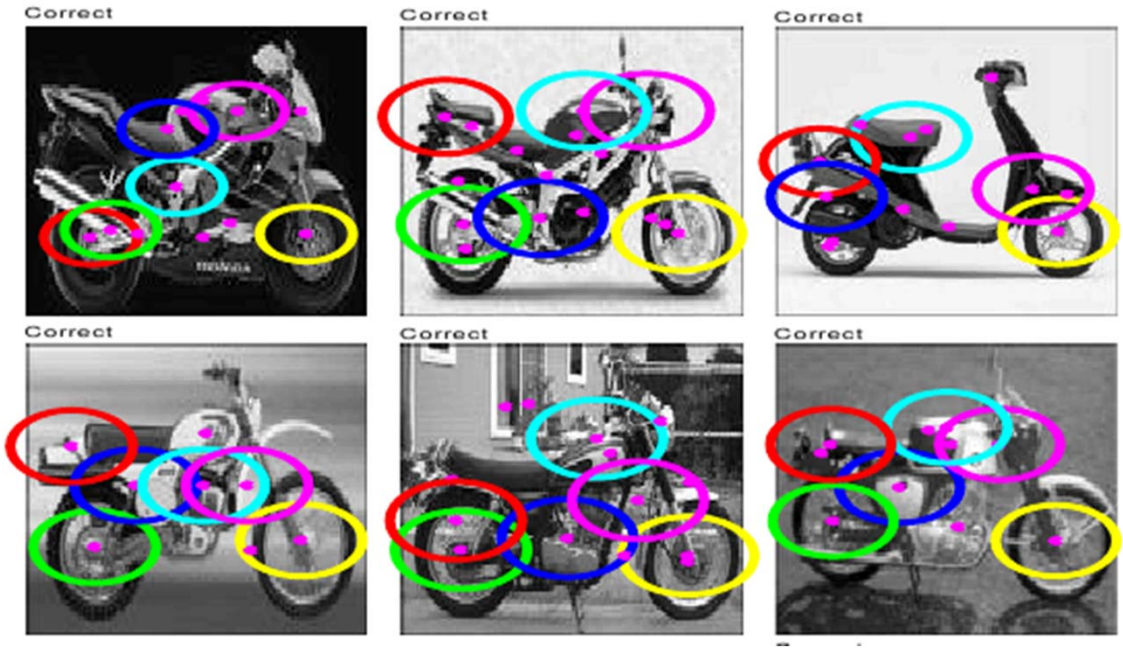
- Segmentation traditionally aimed for unique partitioning



But this never resulted in the necessary accuracy for subsequent recognition

Part-based Classification

No segmentation. Collection of spatially related local image details



Fergus, Perona, Zisserman, 2003

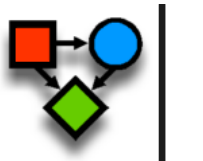
Bag-of-Words

- No segmentation
- No location

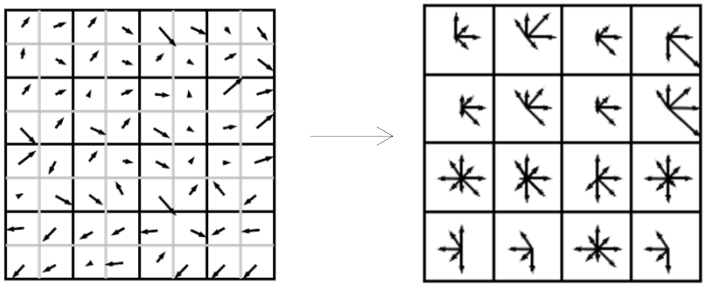
Sivic et al. 2003, Csurka et al. 2004

Bag-of-Words

Bag-of-Words

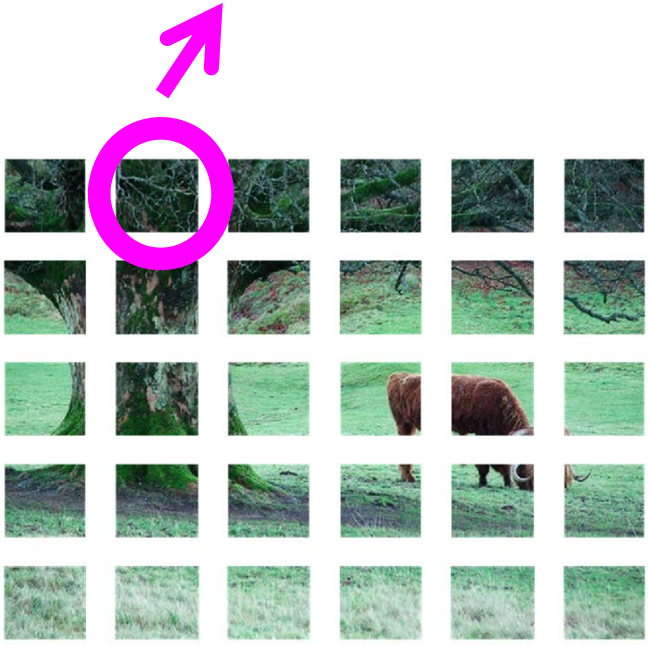


Bag-of-Words



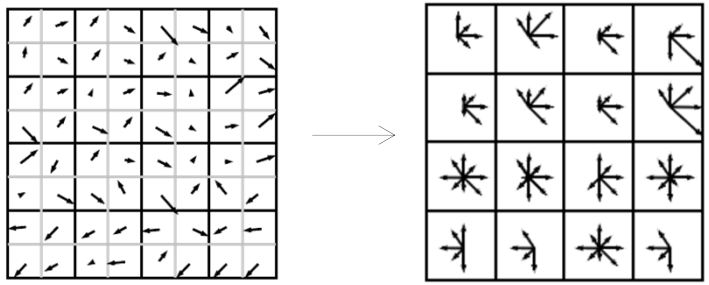
SIFT 4x4

Pixel-wise gradient responses



Bag-of-Words

Descriptor Space

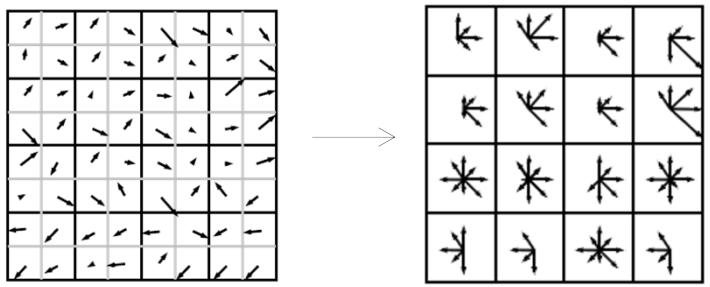


SIFT 4x4

Pixel-wise gradient responses

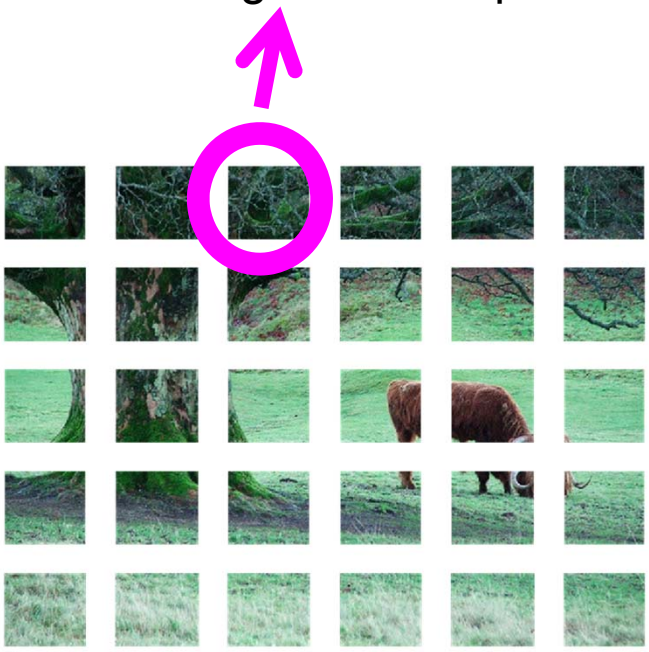
Bag-of-Words

Descriptor Space



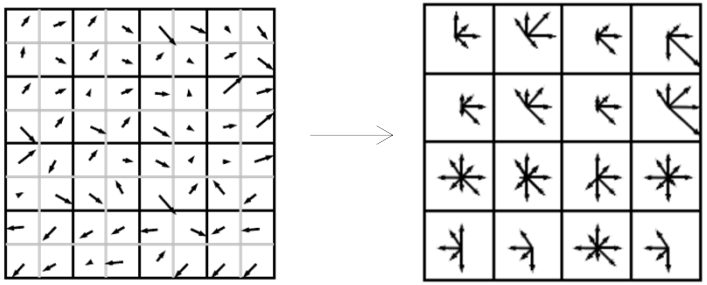
SIFT 4x4

Pixel-wise gradient responses



Bag-of-Words

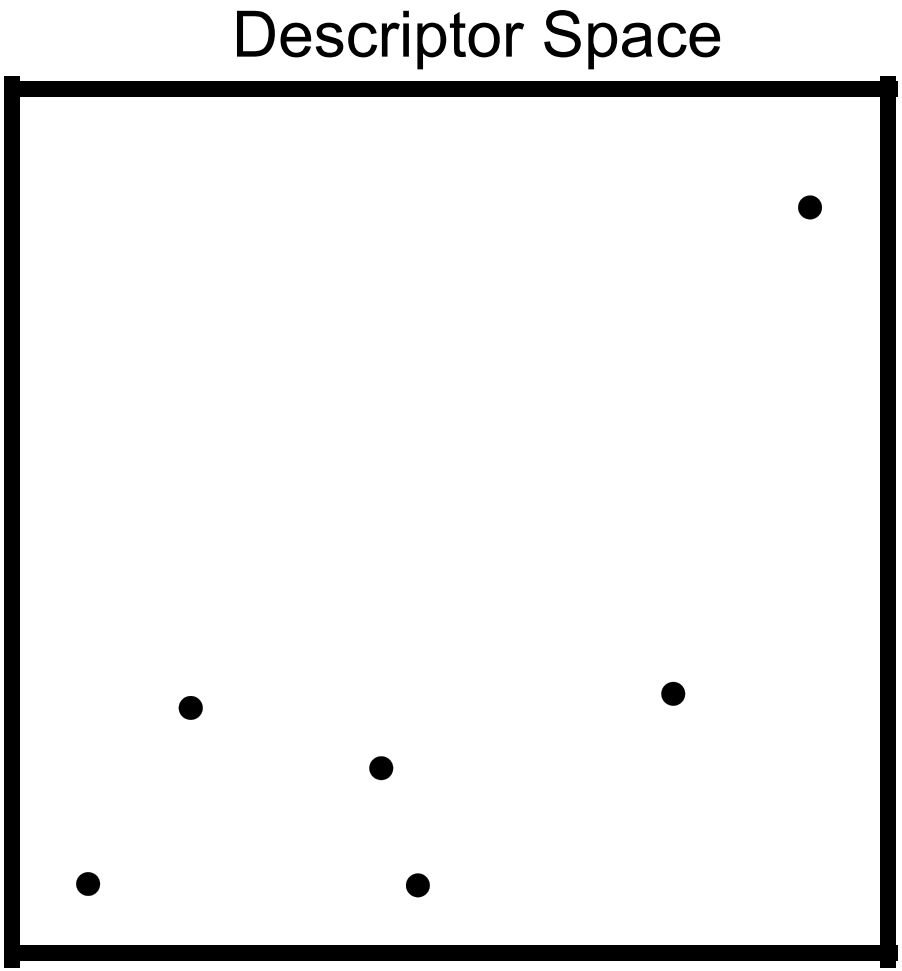
Descriptor Space

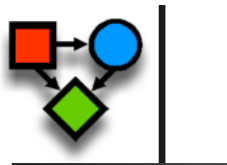


SIFT 4x4

Pixel-wise gradient responses

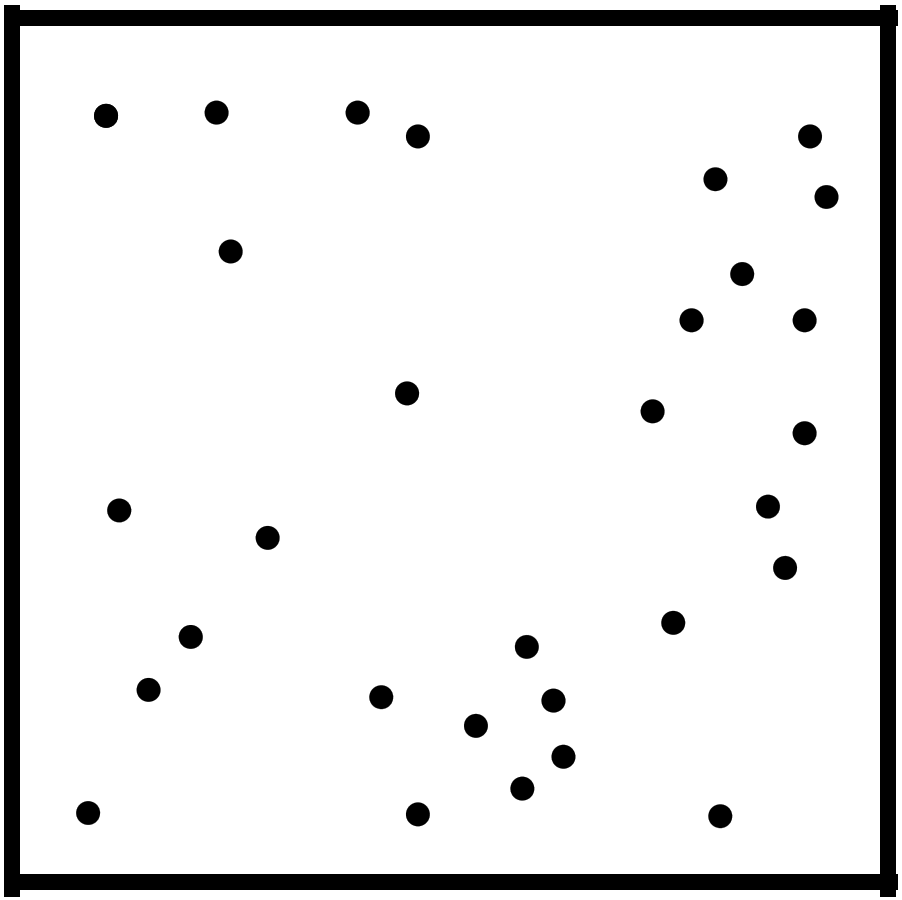
Bag-of-Words



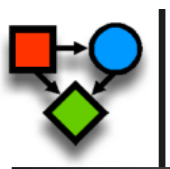


Bag-of-Words

Descriptor Space

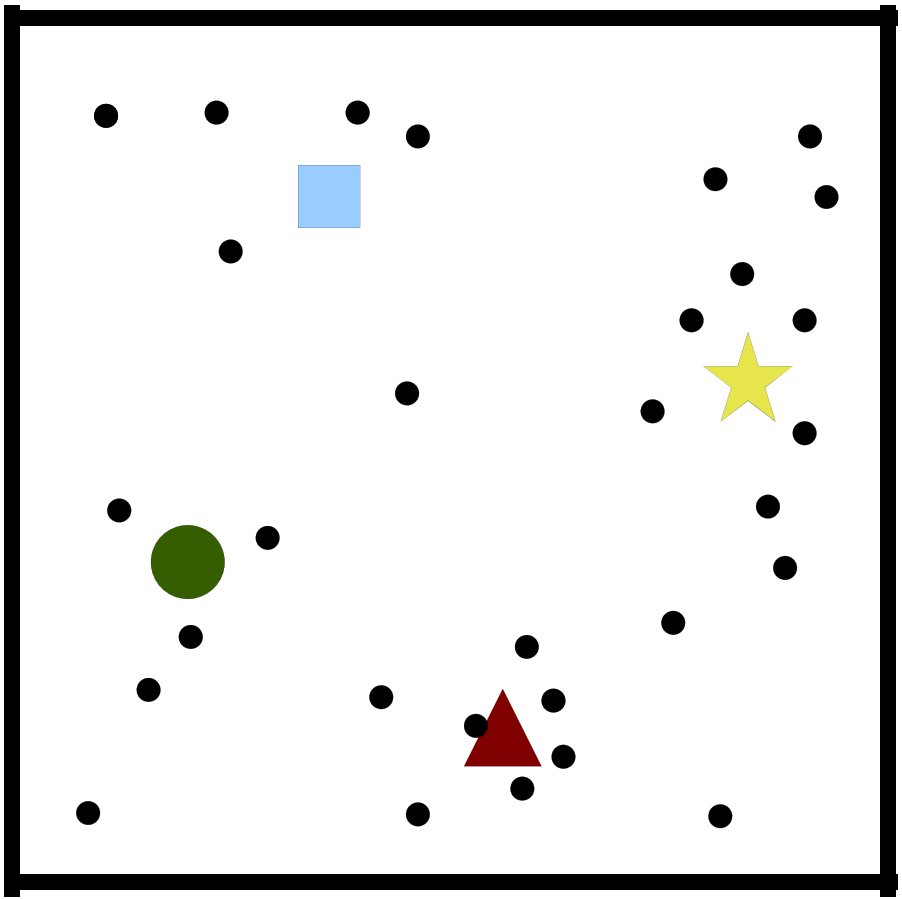


Add SIFT descriptors from many training images



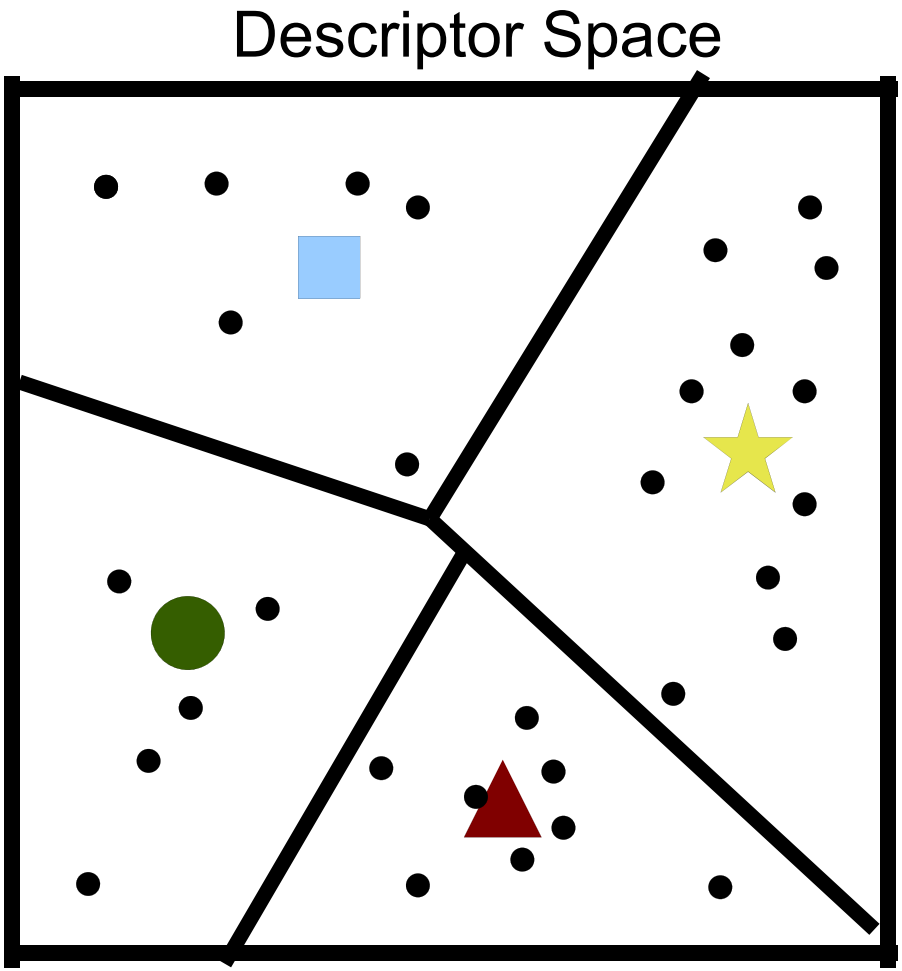
Bag-of-Words

Descriptor Space

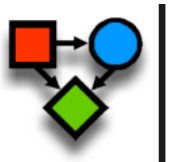


Clustering

Bag-of-Words

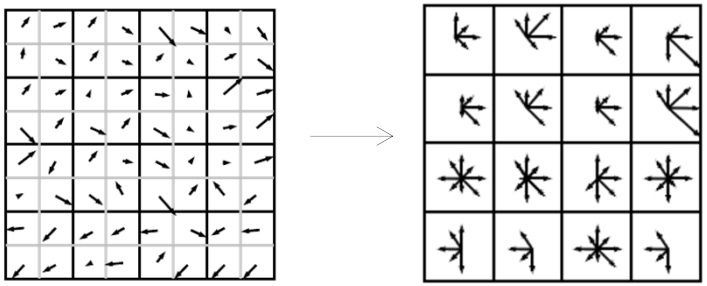


The clusters partition the descriptor space. Each cluster is called a “Visual Word”

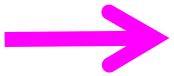


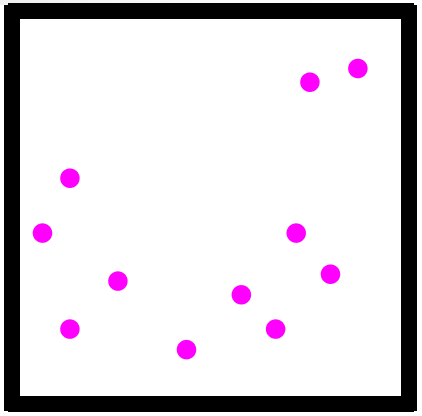
Bag-of-Words

Descriptor Space

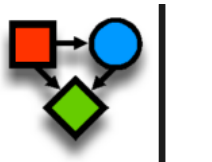


SIFT 4x4

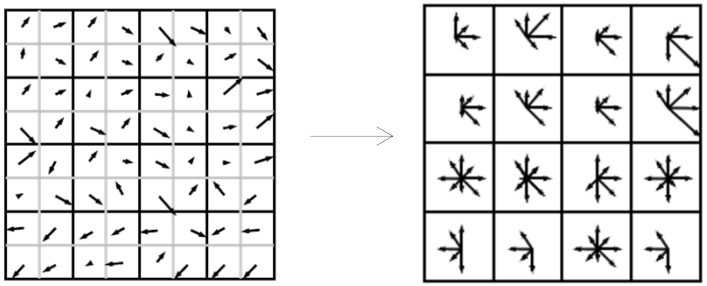




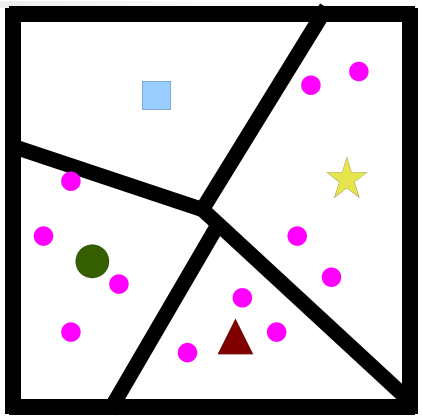
Pixel-wise gradient responses



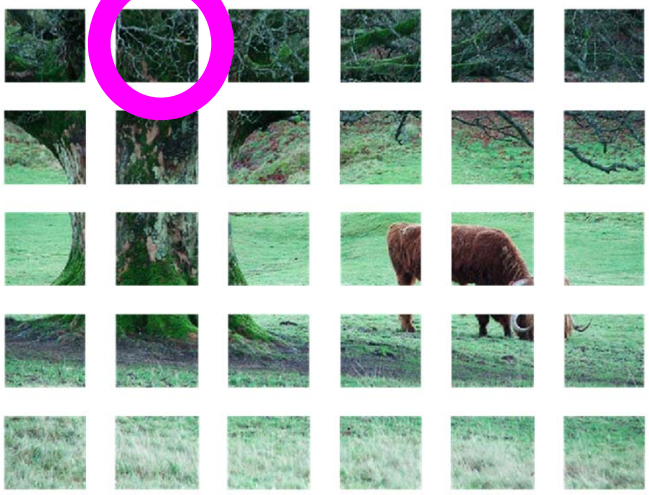
Bag-of-Words



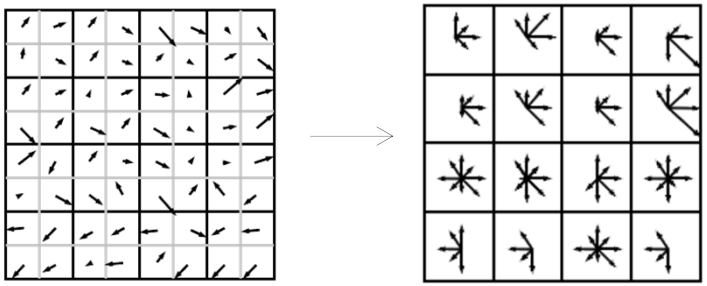
Descriptor Space



Pixel-wise gradient responses

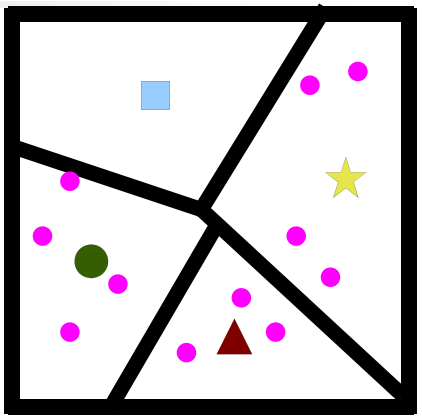


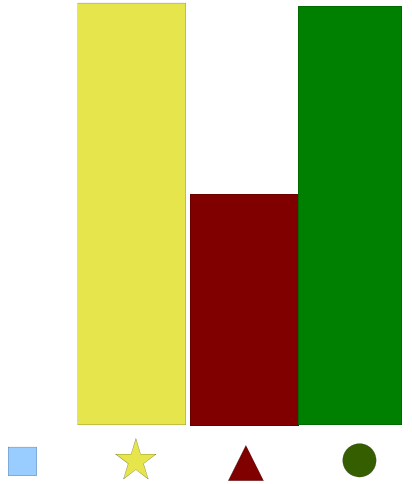
Bag-of-Words



Pixel-wise gradient responses

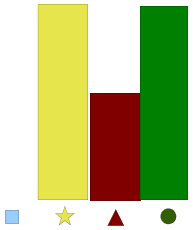
Descriptor Space

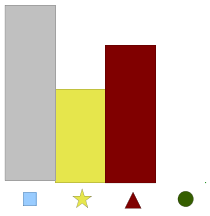


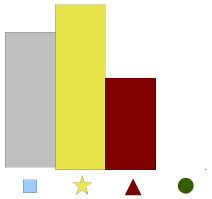


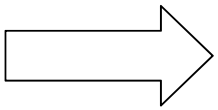
Global Representation

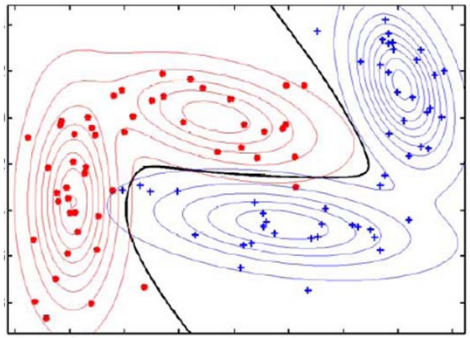
Bag-of-Words











Classification (SVM)

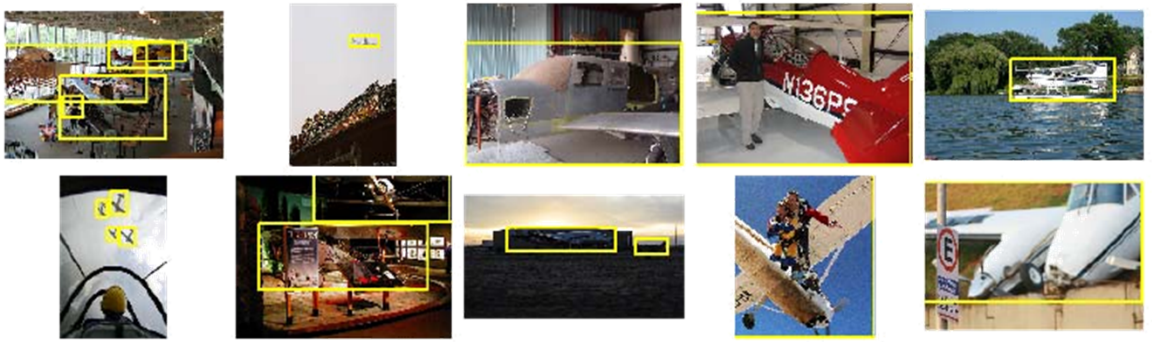
Bag-of-Words details

- Extreme dense sampling at every pixel
- Local patches of 16 by 16 pixels
- SIFT, Opponent SIFT, and RGB-SIFT
- Visual Vocabulary using Random Forests, 4 binary trees of depth 10 = 4096 visual words
- SVM with Histogram Intersection kernel

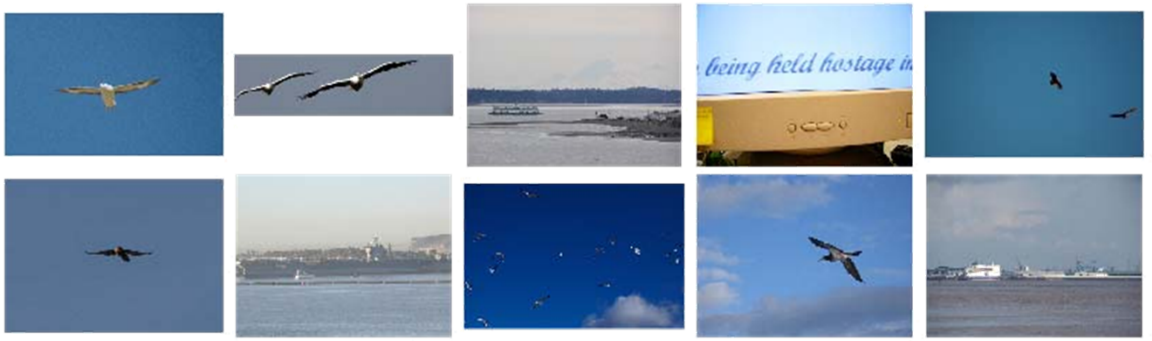
Ranked Images: Aeroplane

- Class images: Highest ranked

- Class images: Lowest ranked



- Non-class images: Highest ranked



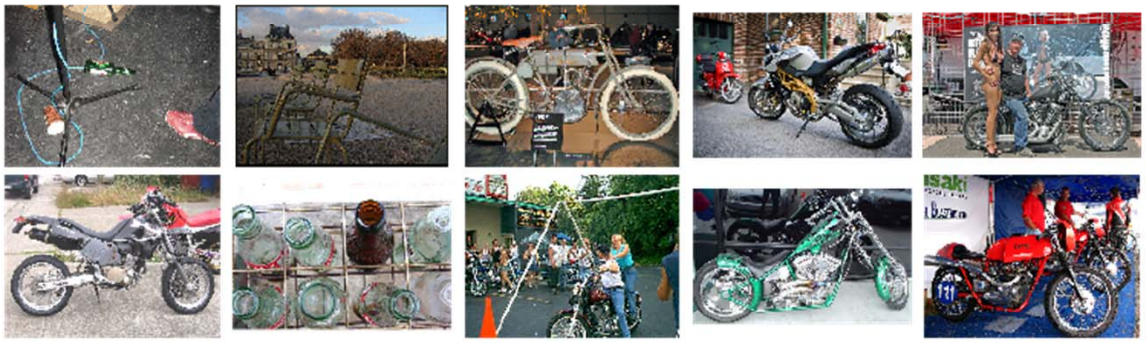
- Context?

Ranked Images: Bicycle

- Class images:
  Highest ranked

- Class images:
  Lowest ranked

- Non-class images:
  Highest ranked



- "Texture"?

Ranked Images: Cat

- Class images:
  Highest ranked

- Class images:
  Lowest ranked

- Non-class images:
  Highest ranked

- "Composition"?

Conclusions Bag-of-Words

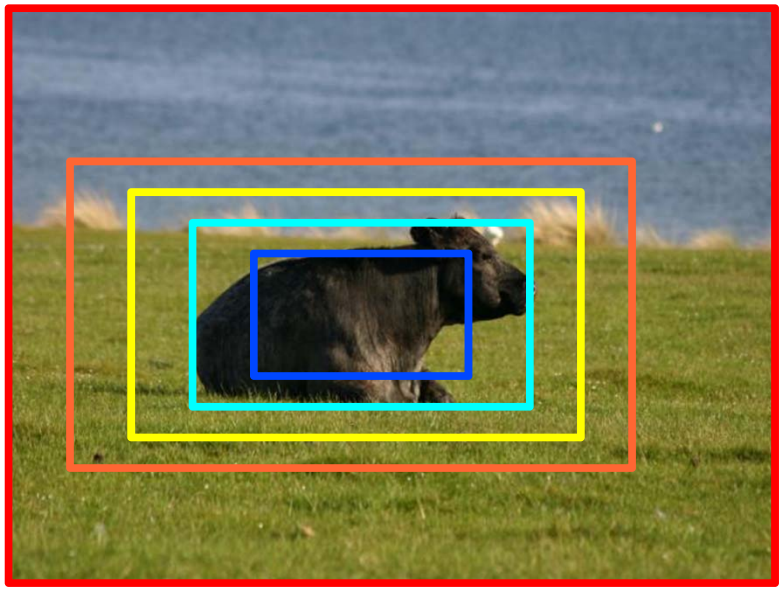
- Works well enough for retrieval purposes
- No segmentation
- No object location

Conclusions Bag-of-Words

- Works well enough for retrieval purposes
- No segmentation
- No object location

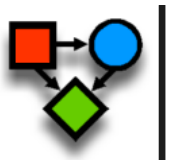
- What do we lose by ignoring object location?
- Which parts of the image are important for recognition?

The Visual Extent of an Object

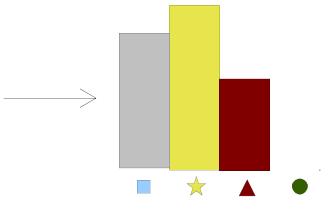


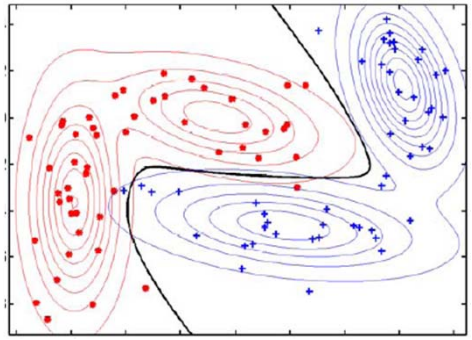
- What do we lose by ignoring object location?
- Which parts of the image are important for recognition?

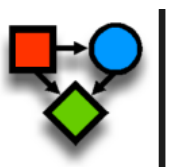
Uijlings, Smeulders, Scha, IJCV 2012



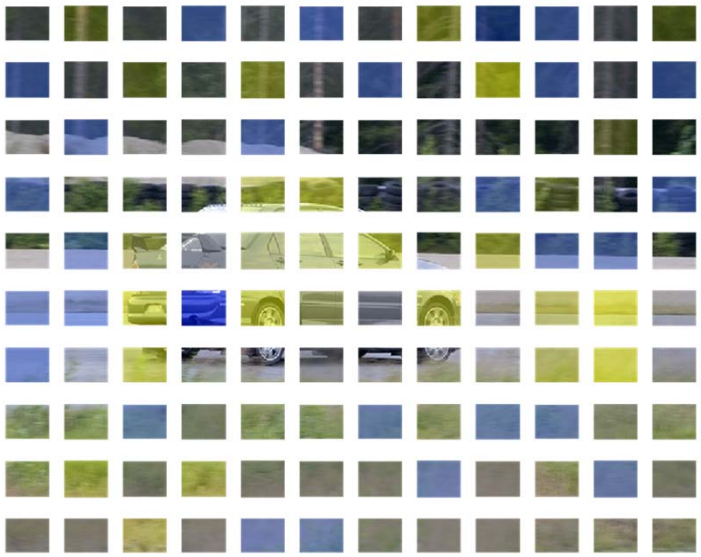
How BoW classifies images

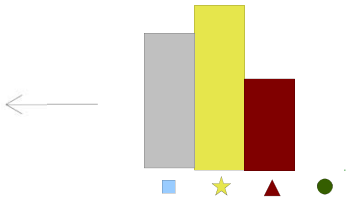


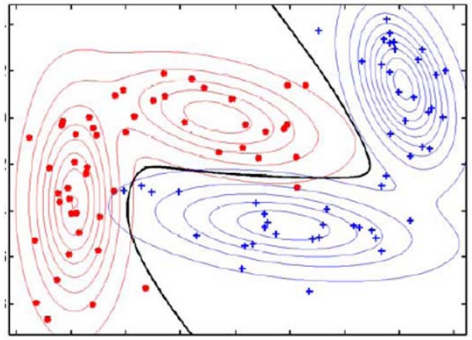




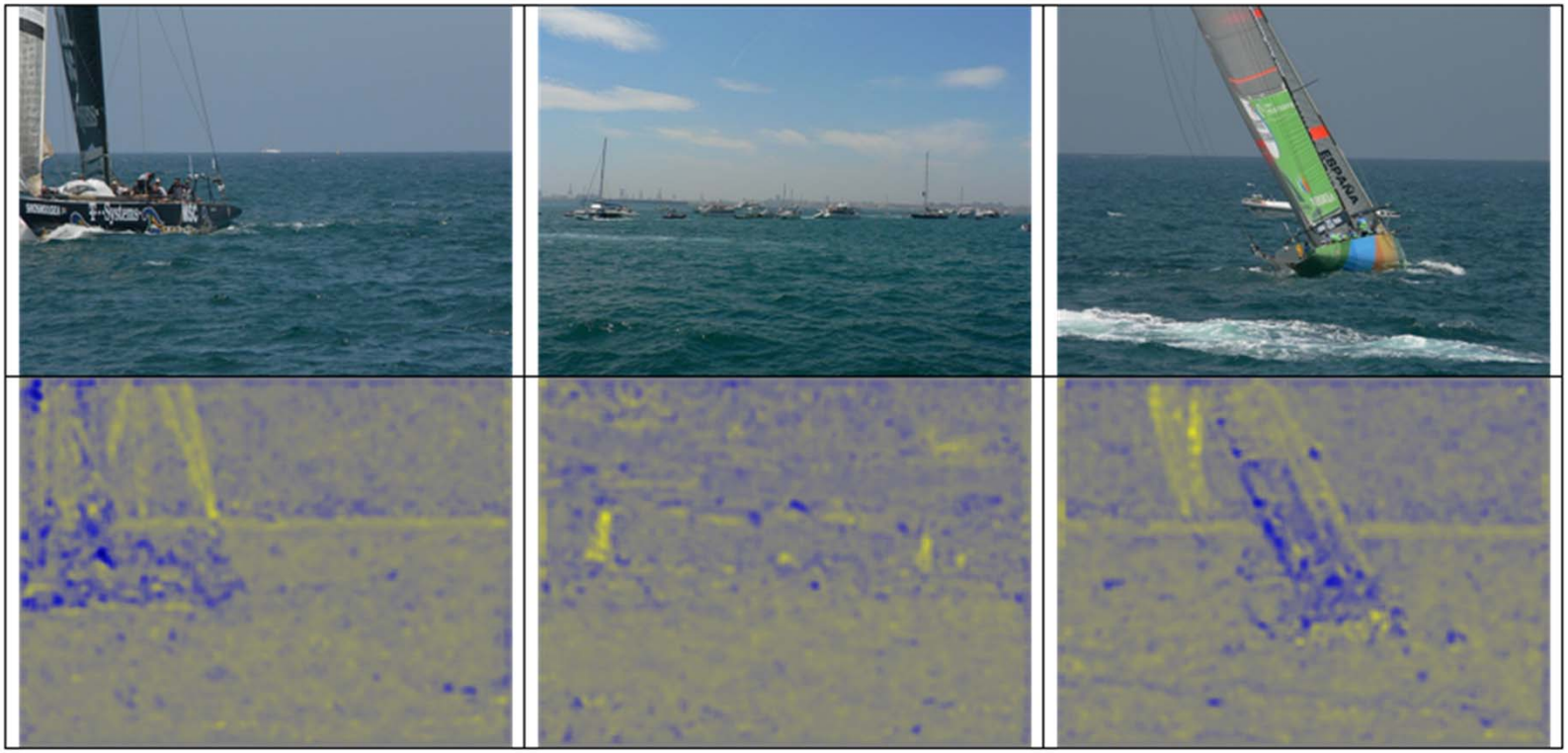
How BoW classifies images

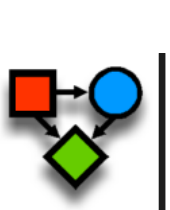




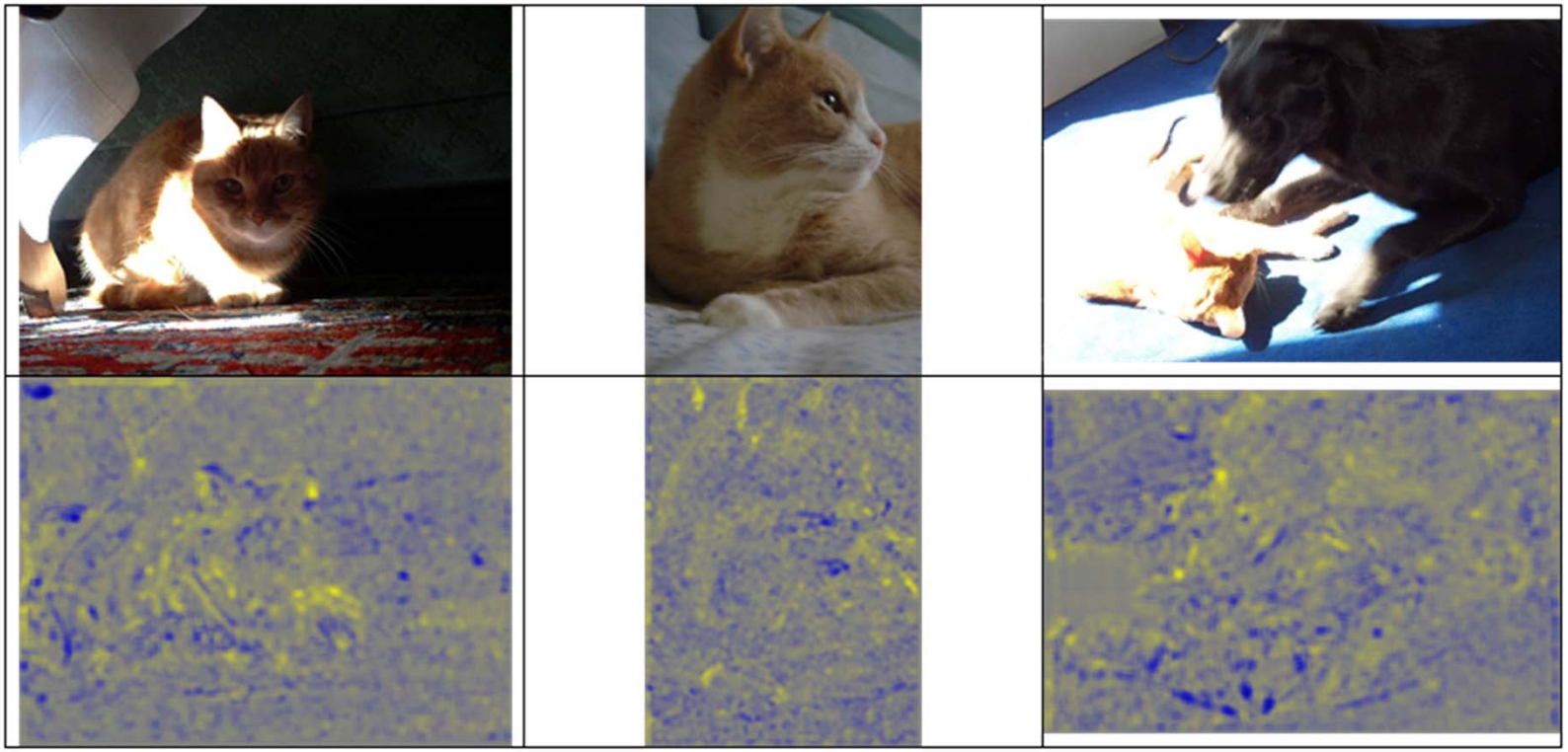


How BoW classifies images

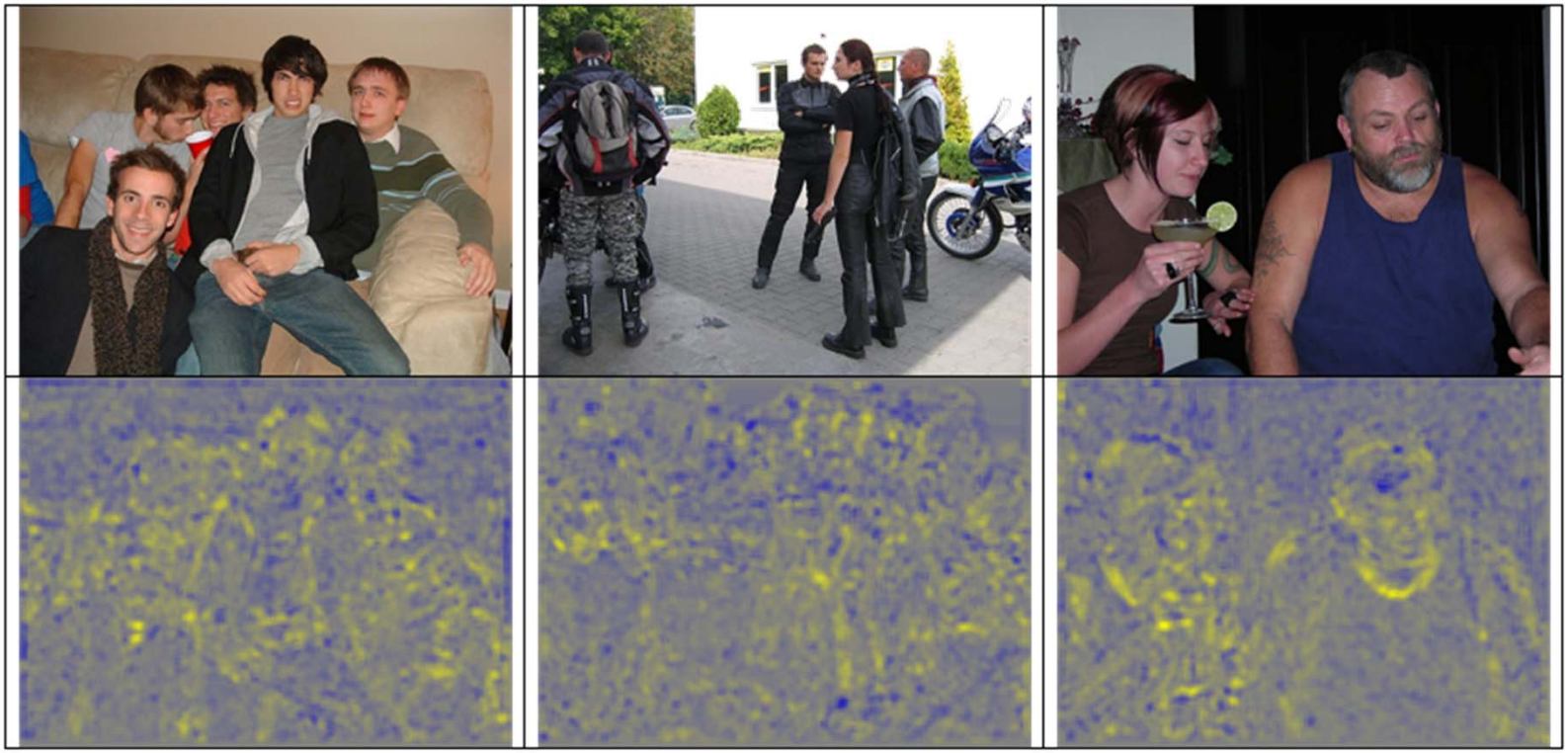




How BoW classifies images



How BoW classifies images



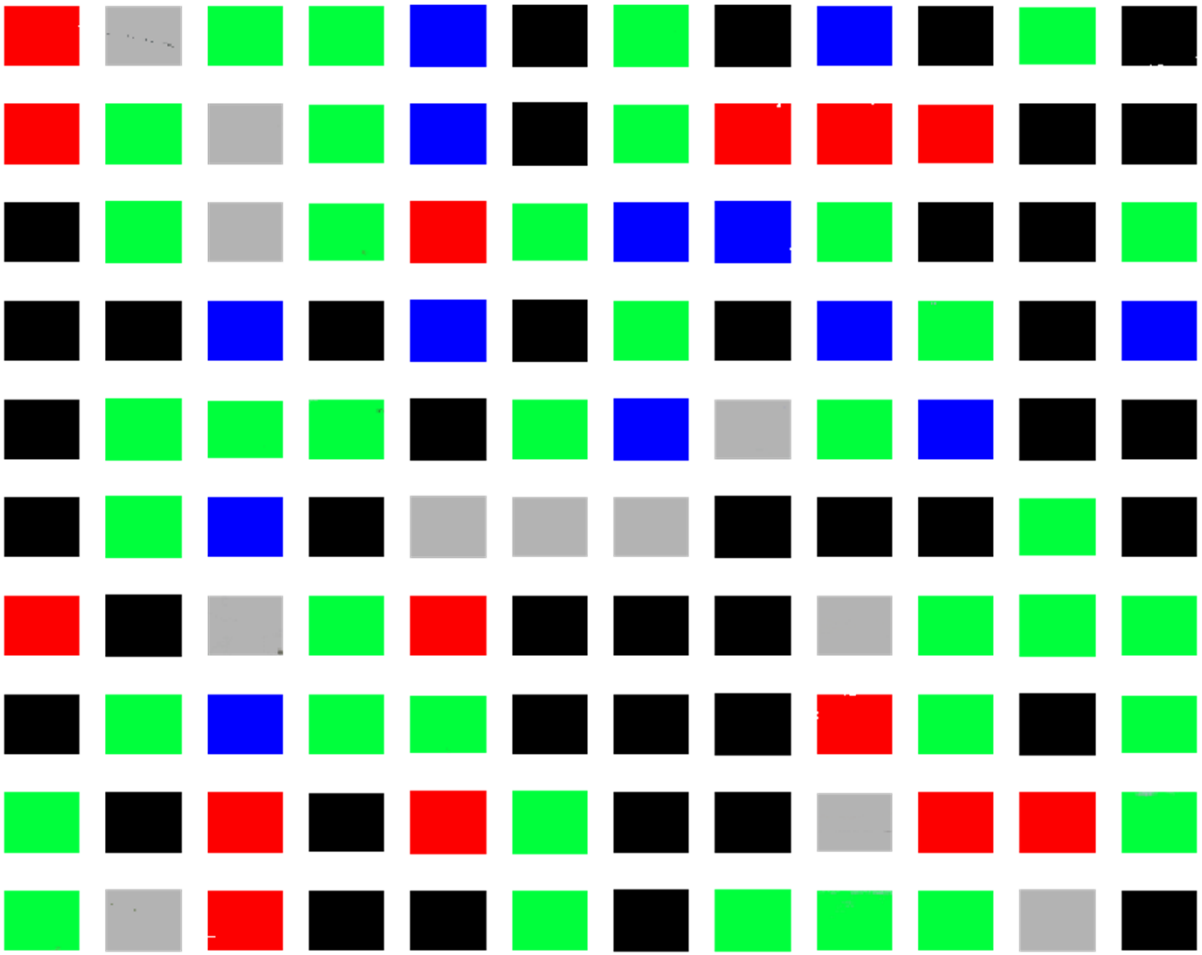
How BoW classifies images

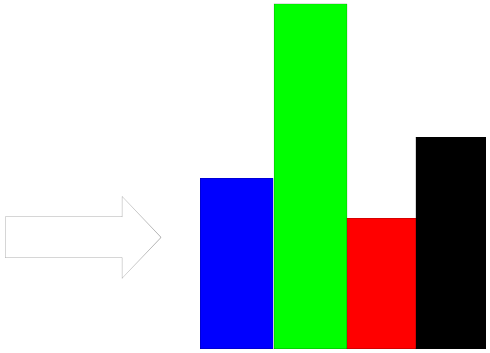
- Bag-of-Words works really on local details, although details are slightly larger than patches
- Bag-of-Words uses details from both the object and its surroundings
- Individual details are not very object or surrounding specific

Uijlings, Smeulders, Scha, IJCV 2012

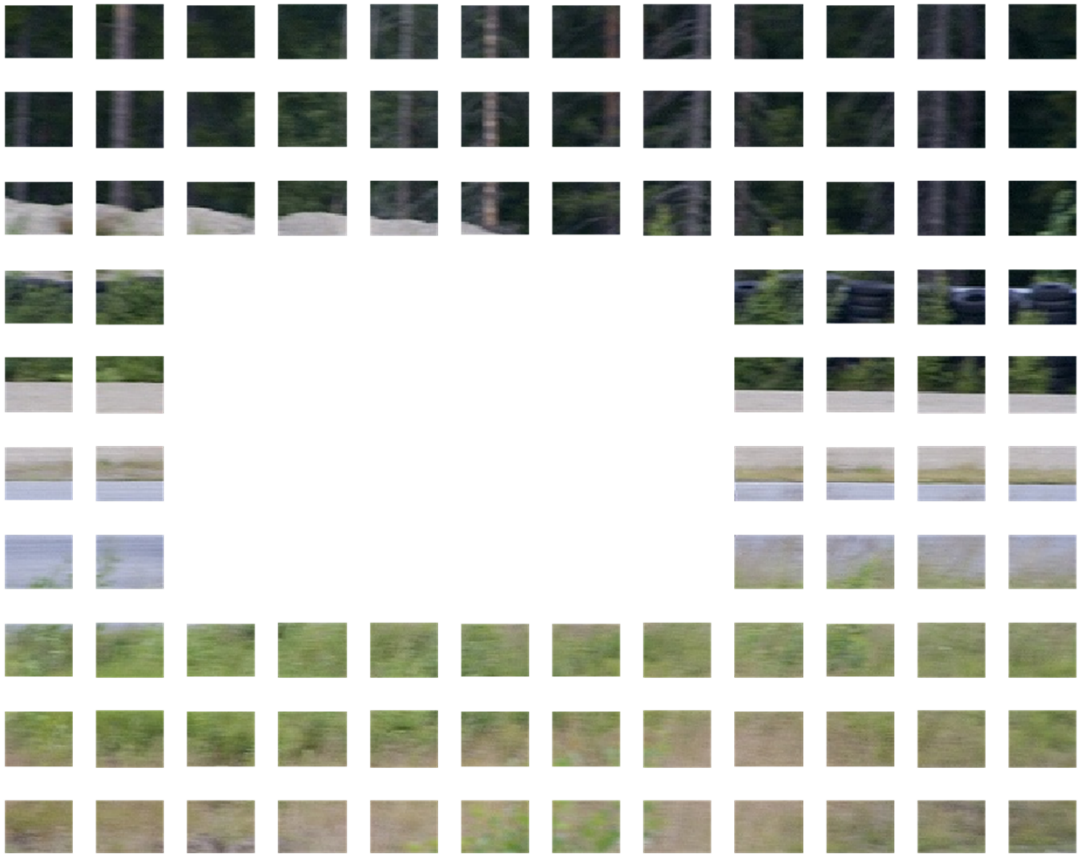
Importance of Object Localization

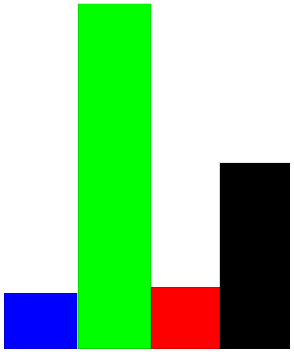
Importance of Object Localization

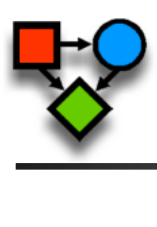




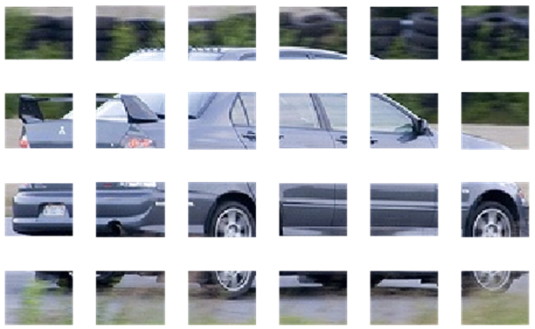
Importance of Object Localization

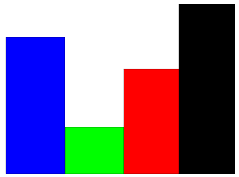






Importance of Object Localization

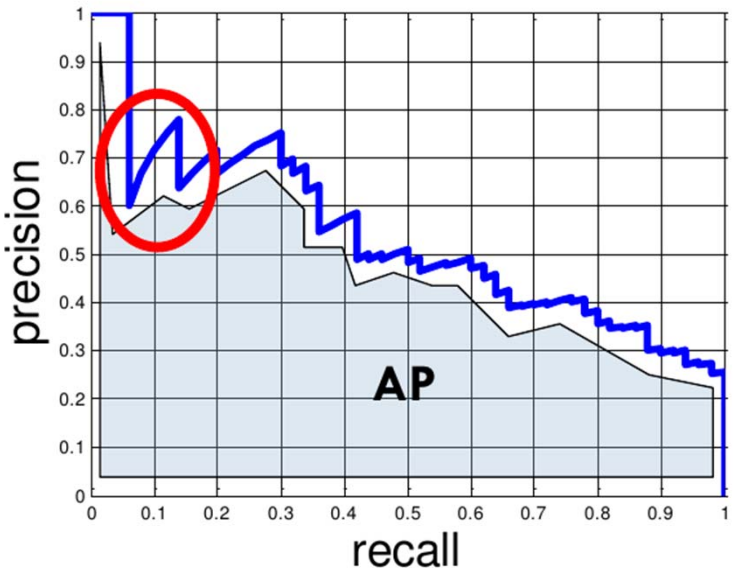




Importance of Object Localization

- Global Bag-of-Words: 0.54 MAP
- Object Only: 0.68 MAP

Average Precision:



Importance of Object Localization

- Knowing the object location increases performance by 26%, from 0.54 to 0.68 MAP
- When object location is known, the surround adds very little information

Uijlings, Smeulders, Scha, IJCV 2012

Importance of Object Localization

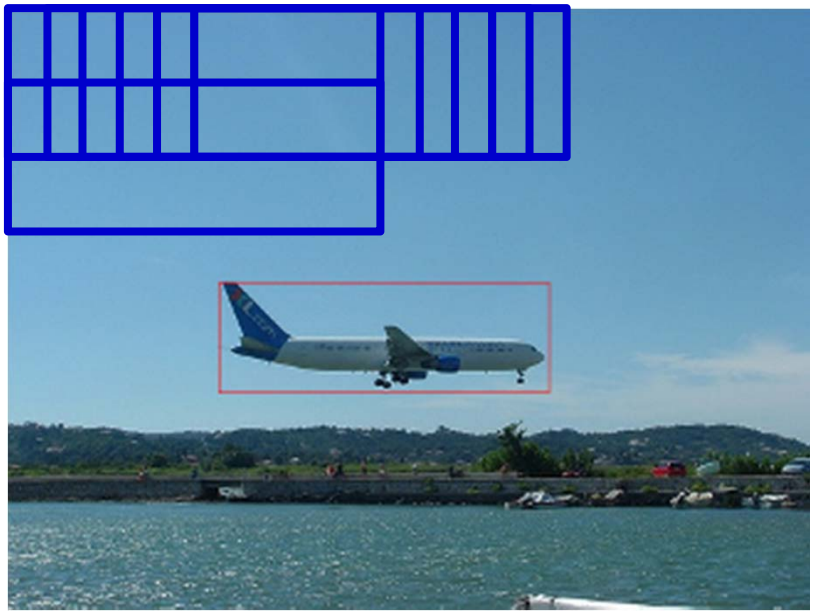
- Knowing the object location increases performance by 26%, from 0.54 to 0.68 MAP
- When object location is known, the surround adds very little information

- Need to incorporate the notion of object location

Outline

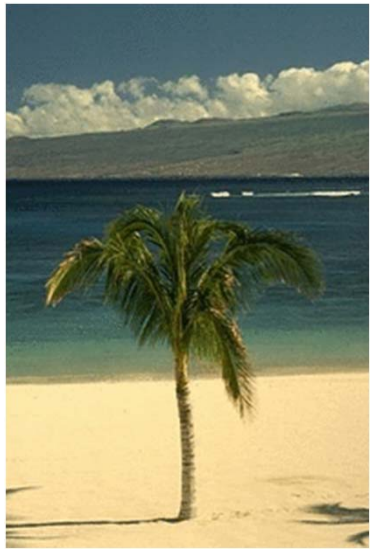
- Sliding Window
- Selective Search vs. Sliding Window
- What is Context?
- The Things and Stuff (TAS) model
- Results

Localisation: Exhaustive Search



- 100,000-1,000,000 locations. Imposes huge computational constraints on subsequent methods

Localisation: Segmentation



- 10-100 locations but captures few objects

Localisation: Segmentation

- Sliding window: 100,000-1,000,000 locations. Imposes huge computational constraints on subsequent methods
- Segmentation: 10-100 locations, but captures few objects

Segmentation as Selective Search

- Rethink segmentation:
  - **High Recall**
  - Coarse locations are sufficient (boxes)
  - Fast to compute

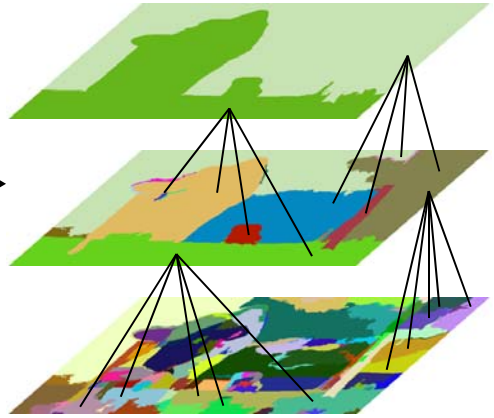
Segmentation as Selective Search

- An image is intrinsically hierarchical. A segmentation at a single scale cannot find all objects

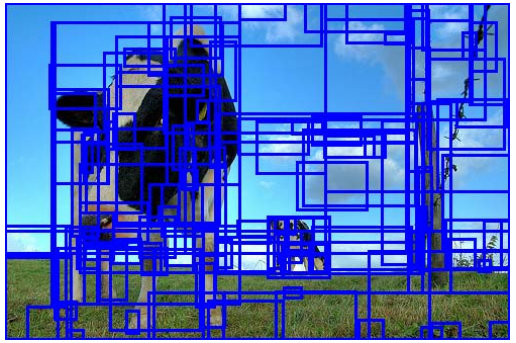
Segmentation as Selective Search

- Use all locations from a hierarchical grouping

Oversegmentation
(Felzenszwalb 2004)



Hierarchical grouping
of segments



Object hypotheses from
all hierarchy levels

Segmentation as Selective Search

Color cues work best

Texture cues work, color fails

- No single segmentation strategy works everywhere
- We need a set of complementary segmentation strategies

Segmentation as Selective Search

- Hierarchical Grouping
- Use of a variety of color spaces with complementary invariance properties
- Different grouping criteria: Colour, Texture, Size, Insideness
- 2 methods:
  - Fast: uses 8 different hierarchical groupings
  - Quality: uses 80 different hierarchical groupings

Van de Sande, et al, ICCV 2011

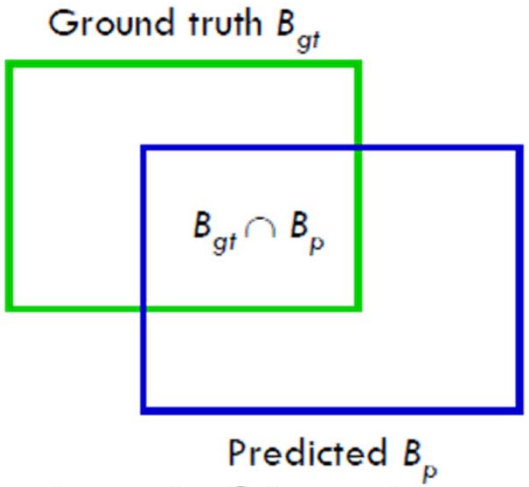
Segmentation as Selective Search

Version	Diversification Strategies	MABO	# windows	time (s)
Single Grouping	HSV C+R+S+F $k = 100$	0.693	362	0.71
Structured Sampling Fast	HSV, Lab C+T+S+F, T+S+F $k = 50, 100$	0.799	2147	3.79
Structured Sampling Quality	HSV, Lab, rgI, H, I C+T+S+F, T+S+F, F, S $k = 50, 100, 150, 300$	0.878	10,108	17.15

MAVO: Mean Average Best Overlap
rgI is normalized R and G and intensity. H is the Hue from HSV.
C = color, T = texture, S = Size, F = Fill/Insideness
k is the parameter for the initial oversegmentation. Higher k means fewer, larger initial regions

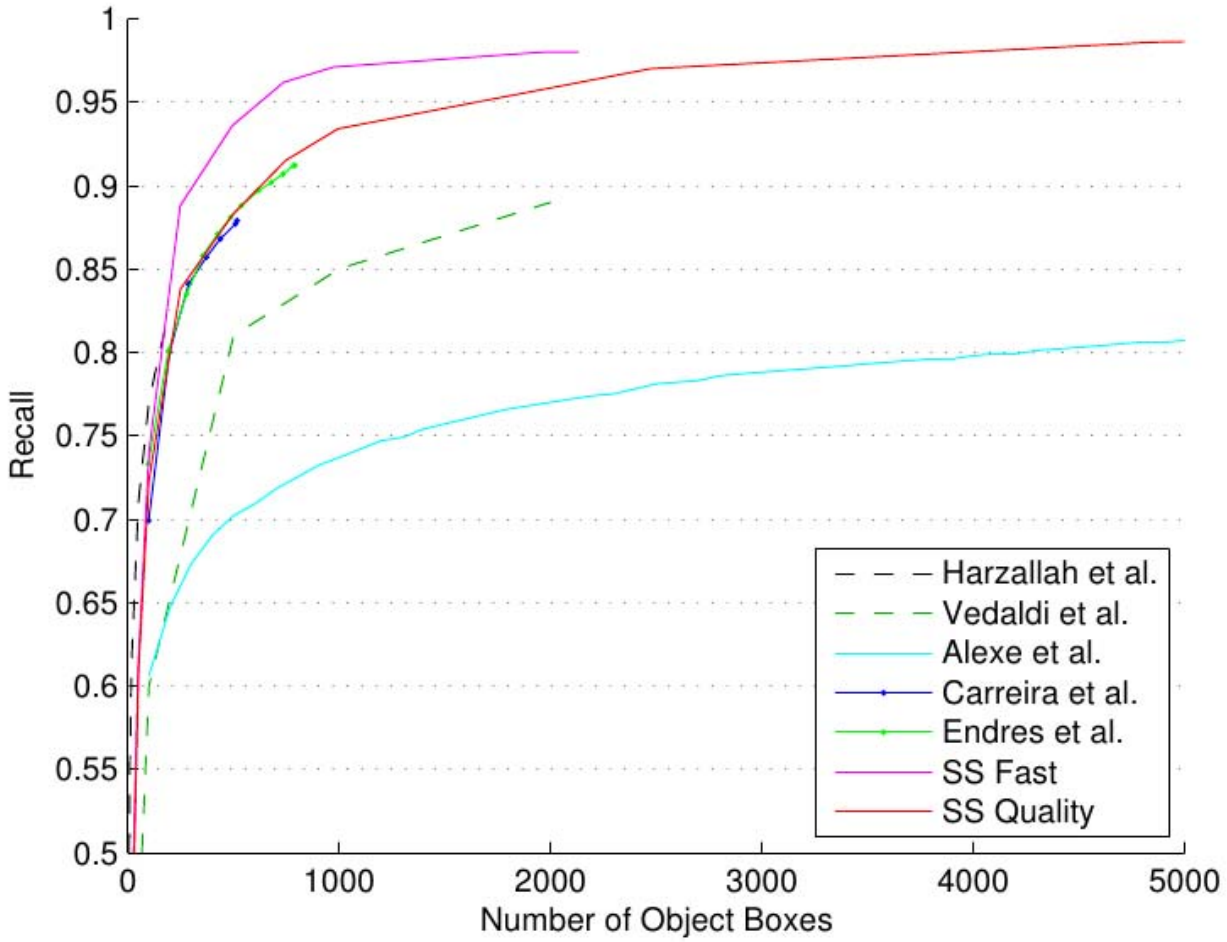
Evaluation of Locations

- Pascal Overlap Criterion

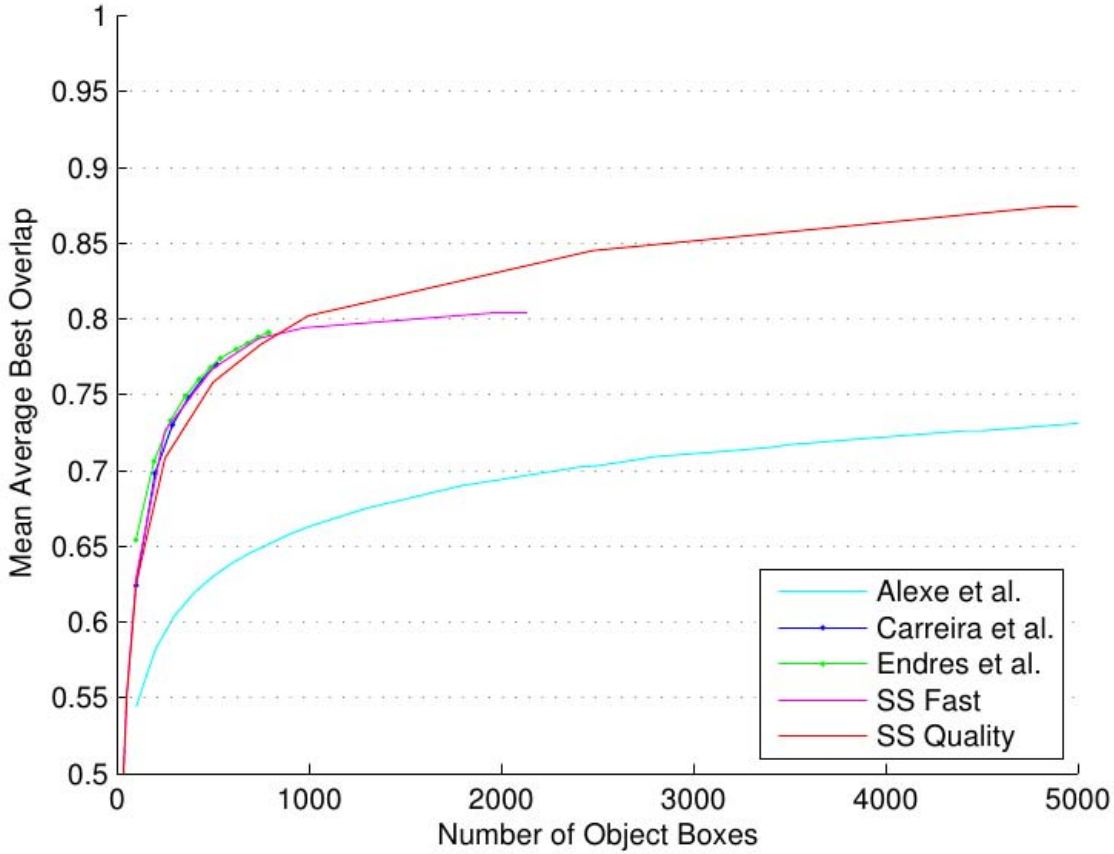


- Correctly localised if best overlap > 50%
- Recall is the % of objects for which there is a location with > 50% overlap

Evaluation of Locations



Evaluation of Locations

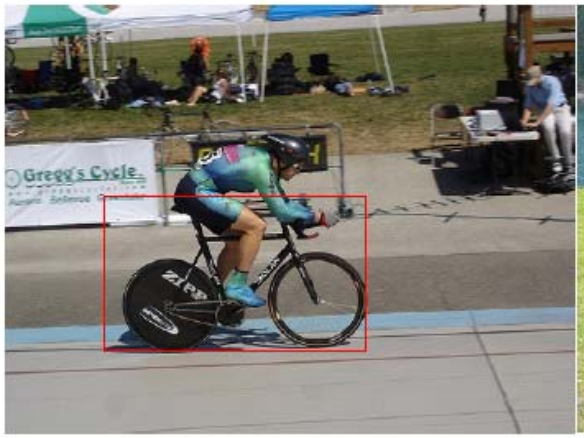


Evaluation of Locations

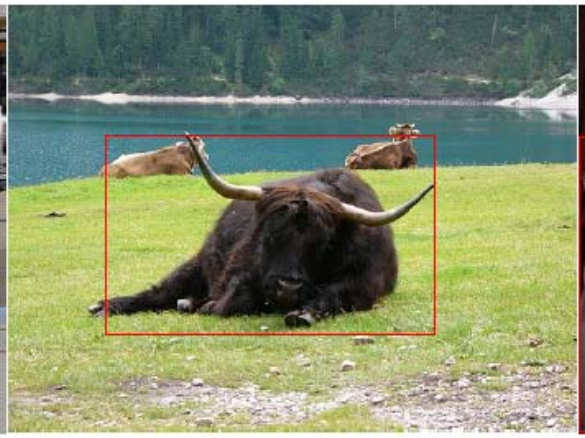
method	recall	MABO	# windows
Arbelaez *et al.* [2]	0.752	0.649	418
Alexe *et al.* [1]	0.824	0.747	10,000
Harzallah *et al.* [14]	0.830	-	200 per class
Carreira *et al.* [3]	0.879	0.770 ± 0.084	517
Endres *et al.* [8]	0.912	0.791 ± 0.082	790
Felzenszwalb *et al.* [10]	0.933	0.829 ± 0.052	100,352 per class
Vedaldi *et al.* [29]	0.940	-	10,000 per class
Single Grouping	0.840	0.690	289
SS "Fast"	0.980	0.804 ± 0.046	2,134
SS "Quality"	0.991	0.879 ± 0.039	10,097

Evaluation of Locations

- What does a .88 Best Overlap score mean?



Overlap 88.4%

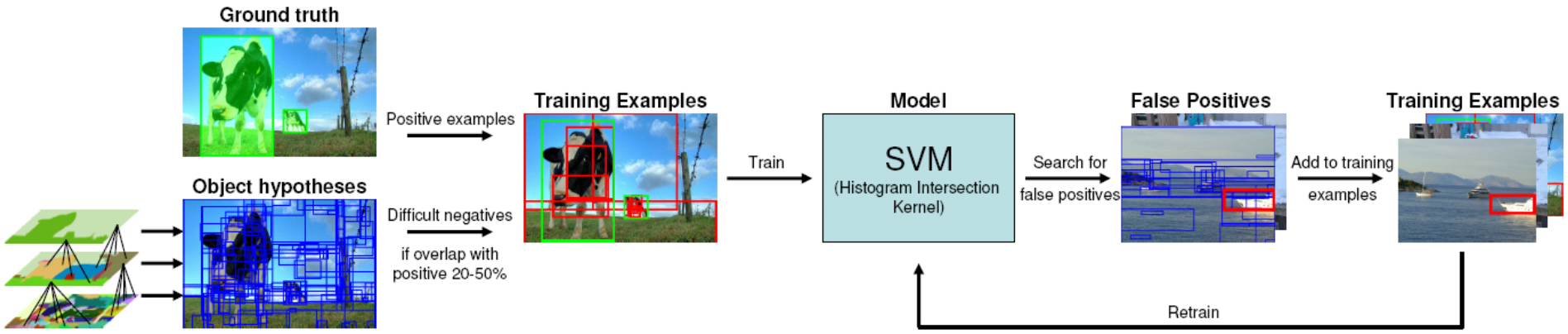


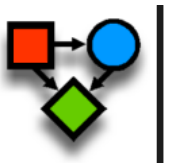
Overlap 87.9%

Overlap 87.4%

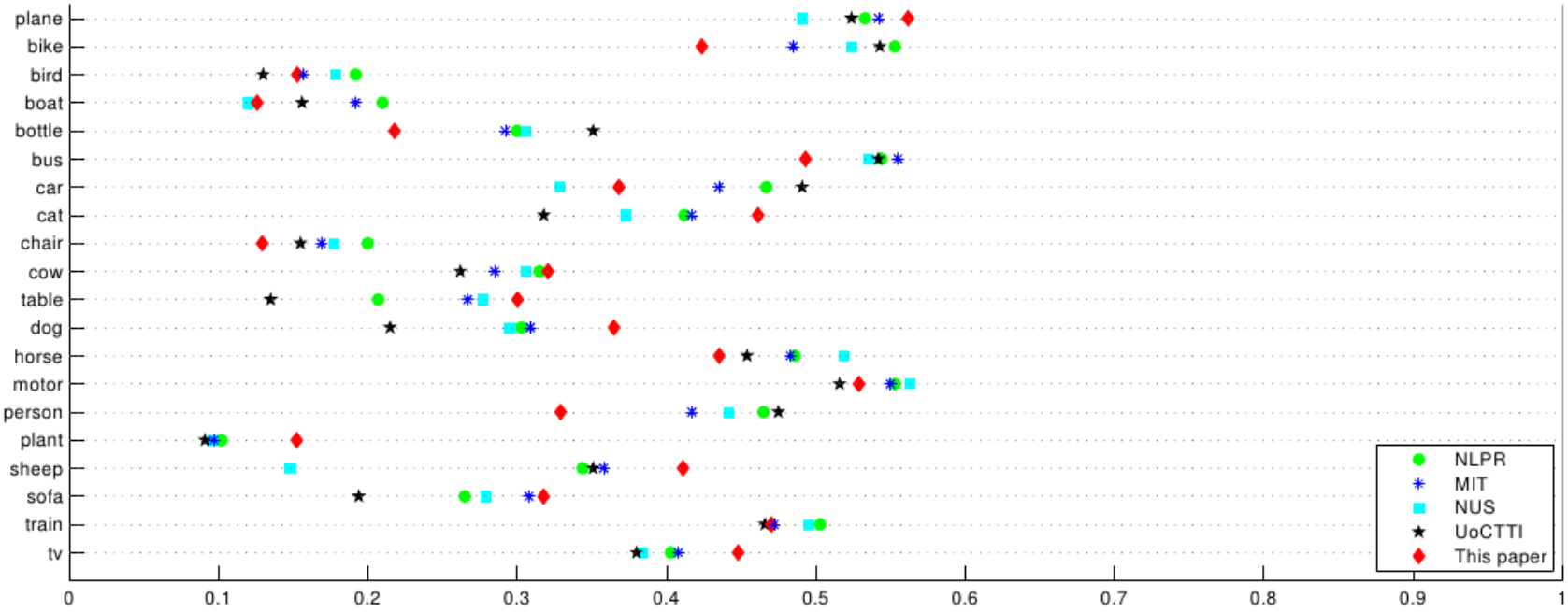
Selective Search in Object Localisation

- Goal: Identify and find the location of the objects. An object is found if the Pascal Overlap (MABO) score > 50%





Selective Search in Object Localisation

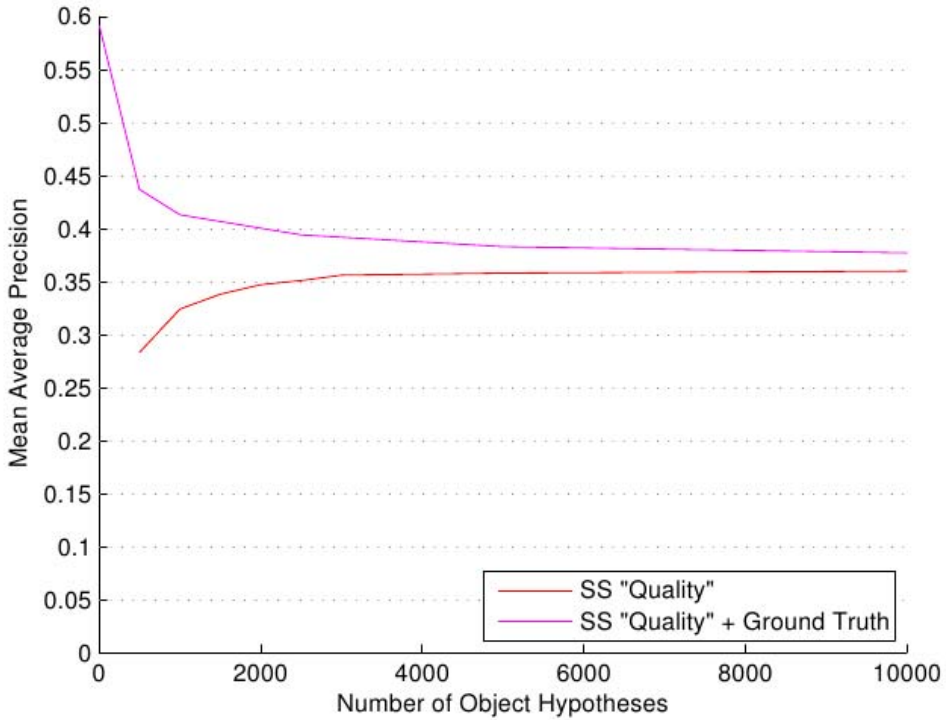


Pascal VOC 2010

- Best results for 9 out of 20 object classes
- Works especially well on non-rigid object classes
- All competing methods are based on exhaustive search with HOG-features

Sande, et al. ICCV 2011

Object Localisation

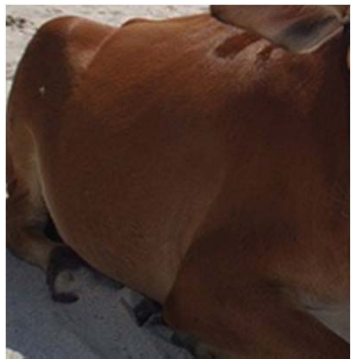


- Quality of locations is close to optimal for this Bag-of-Words system

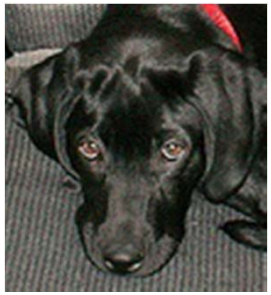
Conclusions Selective Search

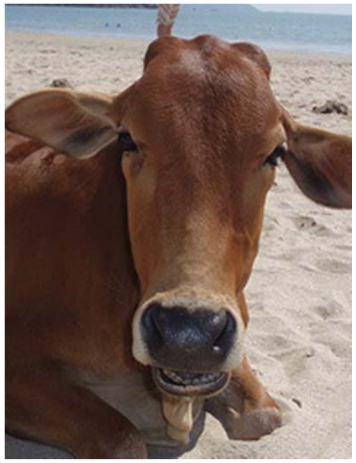
- Results in a small yet high quality set of potential object locations
- Works by rethinking segmentation:
  - Focus more on Recall than Precision
  - Hierarchical grouping to deal with objects at multiple scales
  - Multiple complementary strategies to deal with high variety in image conditions
- Enables use of more expensive features

Is Exact Localisation Optimal?



Is Exact Localisation Optimal?





Is Exact Localisation Optimal?

Parts were earlier used in “visual identification”
Learning to Locate Informative Features for Visual Identification, IJCV 2008,
A. Ferencz, E. Learned-Miller, J. Malik

Is Exact Localisation Optimal?

Parts may be more discriminative because of pose change, often caused by interaction between the objects

Is Exact Localisation Optimal?

For occluded objects only the non-occluded part is informative.

Is Exact Localisation Optimal?

In crowded scenes, compared to an individual object:
a collection is both more easy to find and may be more discriminative

Is Exact Localisation Optimal: NO

- Parts may be more discriminative for some classes
- Interacting objects may change pose, retaining typical appearance only for object parts
- Occluded objects are hard to find when searching for complete objects
- In crowded scenes groups are more easy to recognize

The Windows that Tell the Story of an Image, J.R.R. Uijlings and A.W.M. Smeulders

The Most Telling Window

- May focus on:
  - Object Parts
  - Complete Objects
  - Object Collections

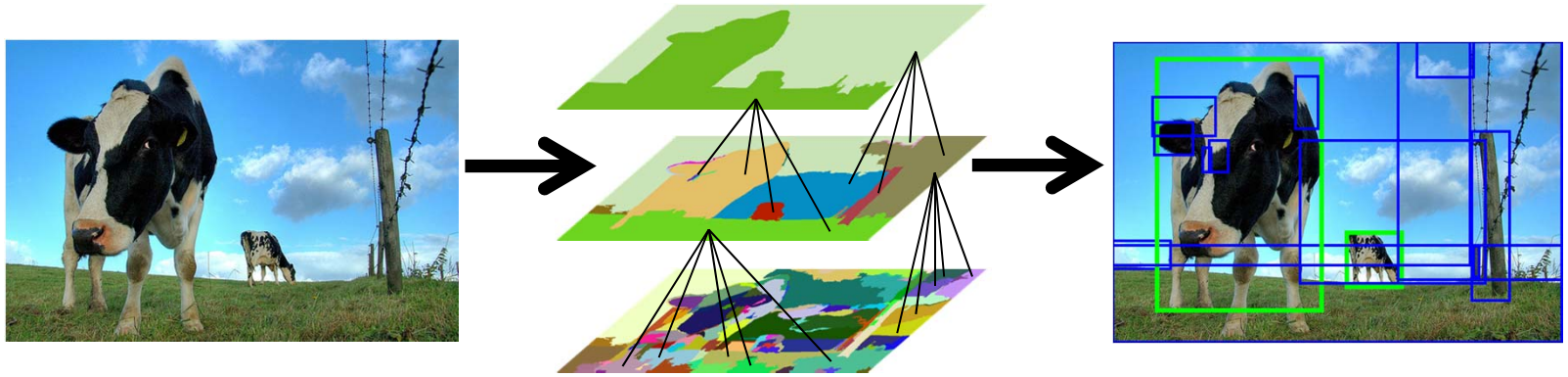
The Windows that Tell the Story of an Image, J.R.R. Uijlings and A.W.M. Smeulders.

Methodology: Object Location

- Most Dominant: Sliding Windows

- But yields 100.000 – 1.000.000 windows: infeasible for powerful Bag-of-Words implementation
- Solution: Selective Search

Methodology: Object Location

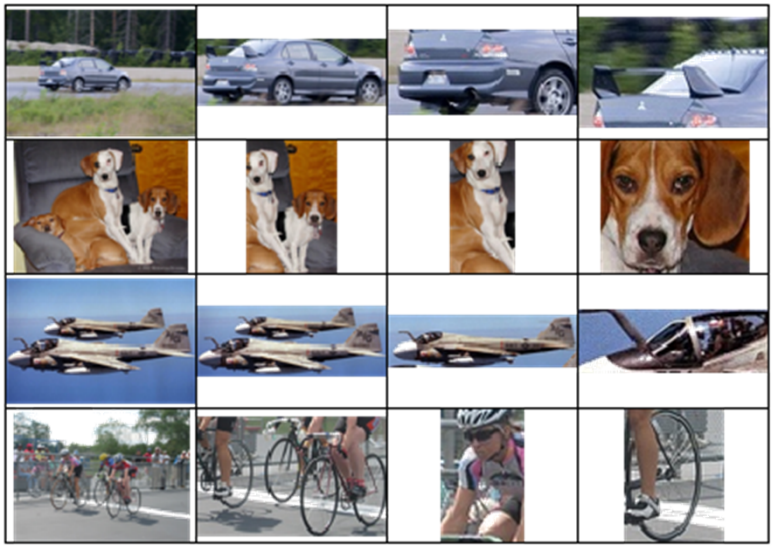


- Selective Search which uses multiple, complementary, hierarchical segmentations

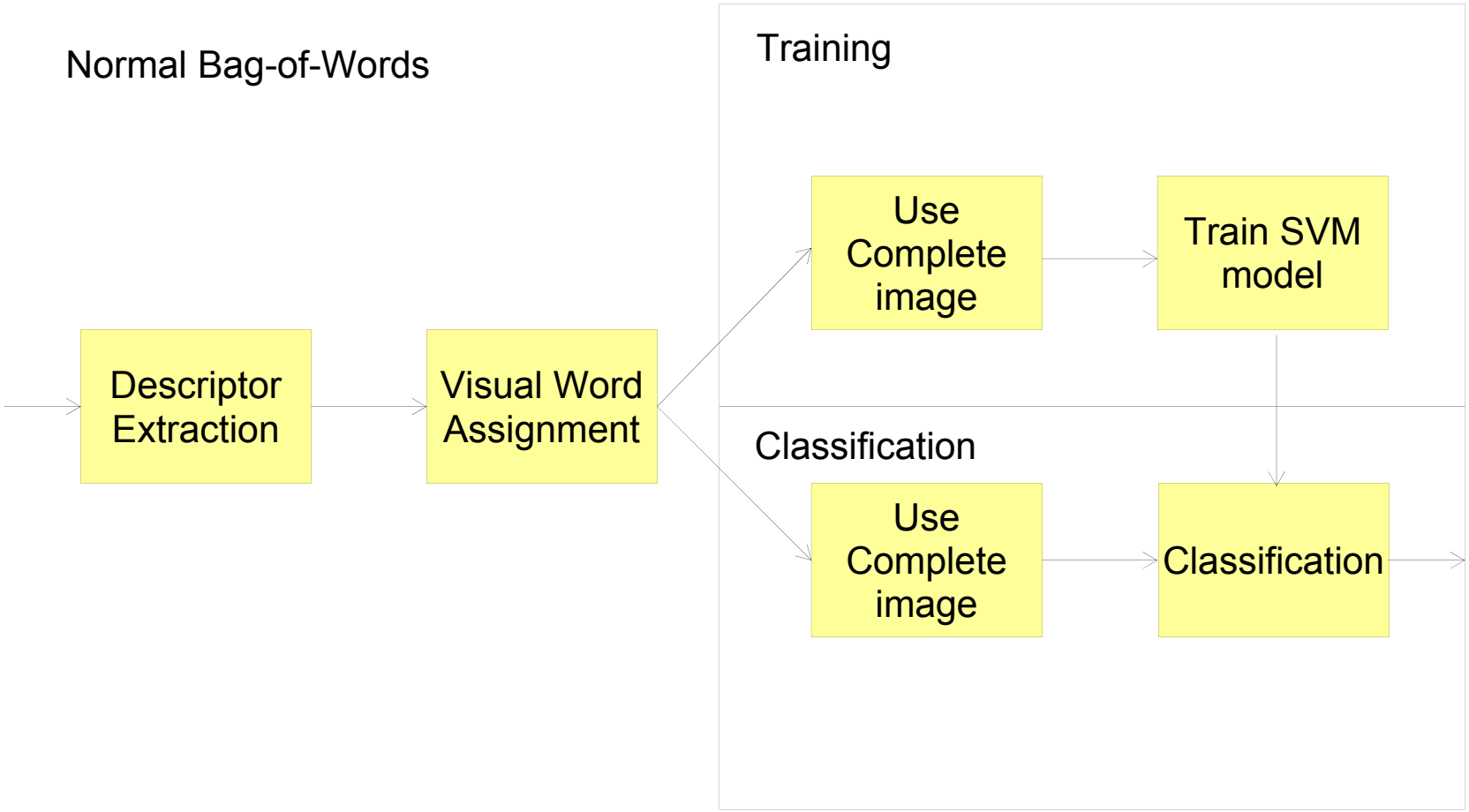
Methodology: Object Location

- Small set of class-independent locations
- Captures parts, objects, and collections

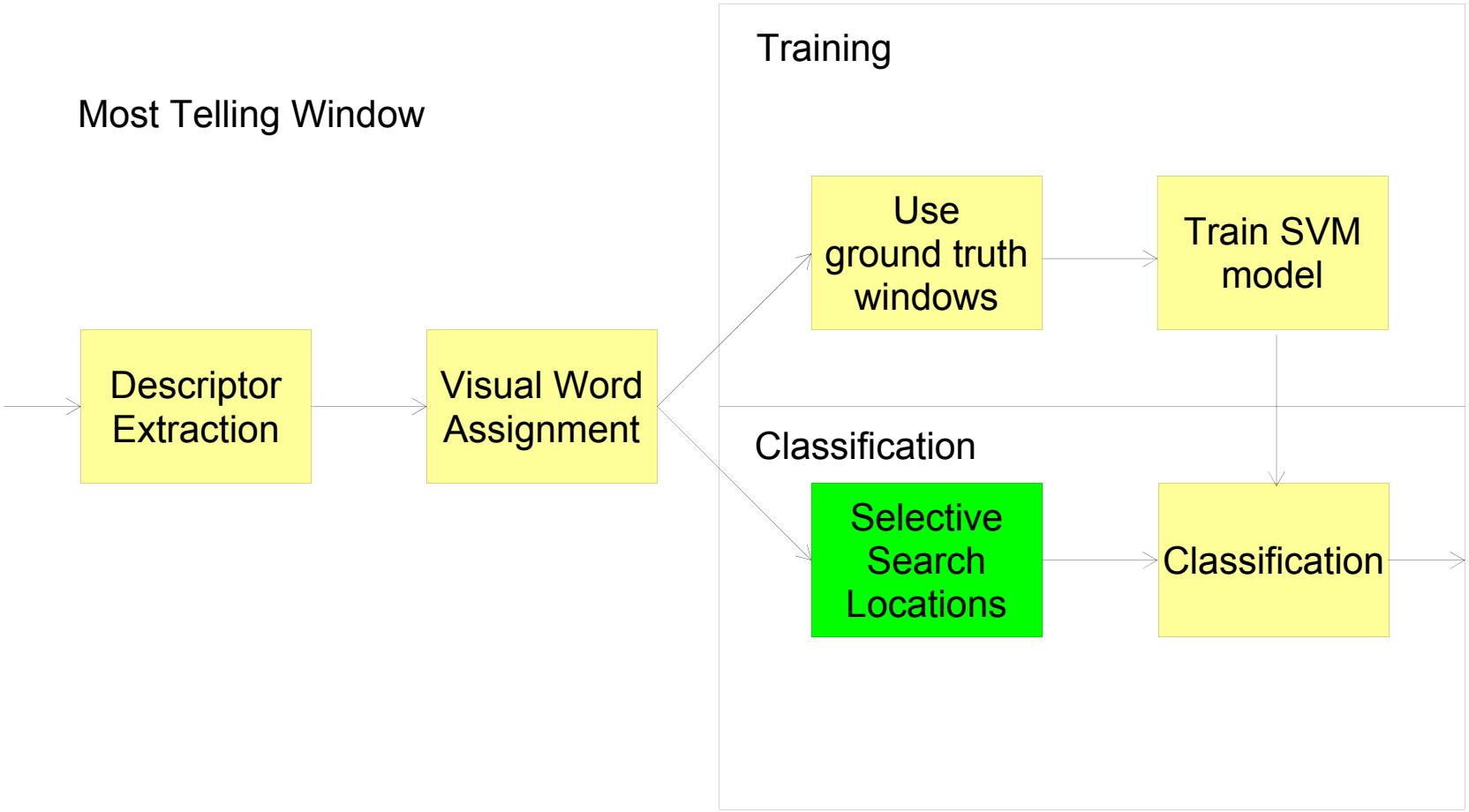
Example windows generated by the method:



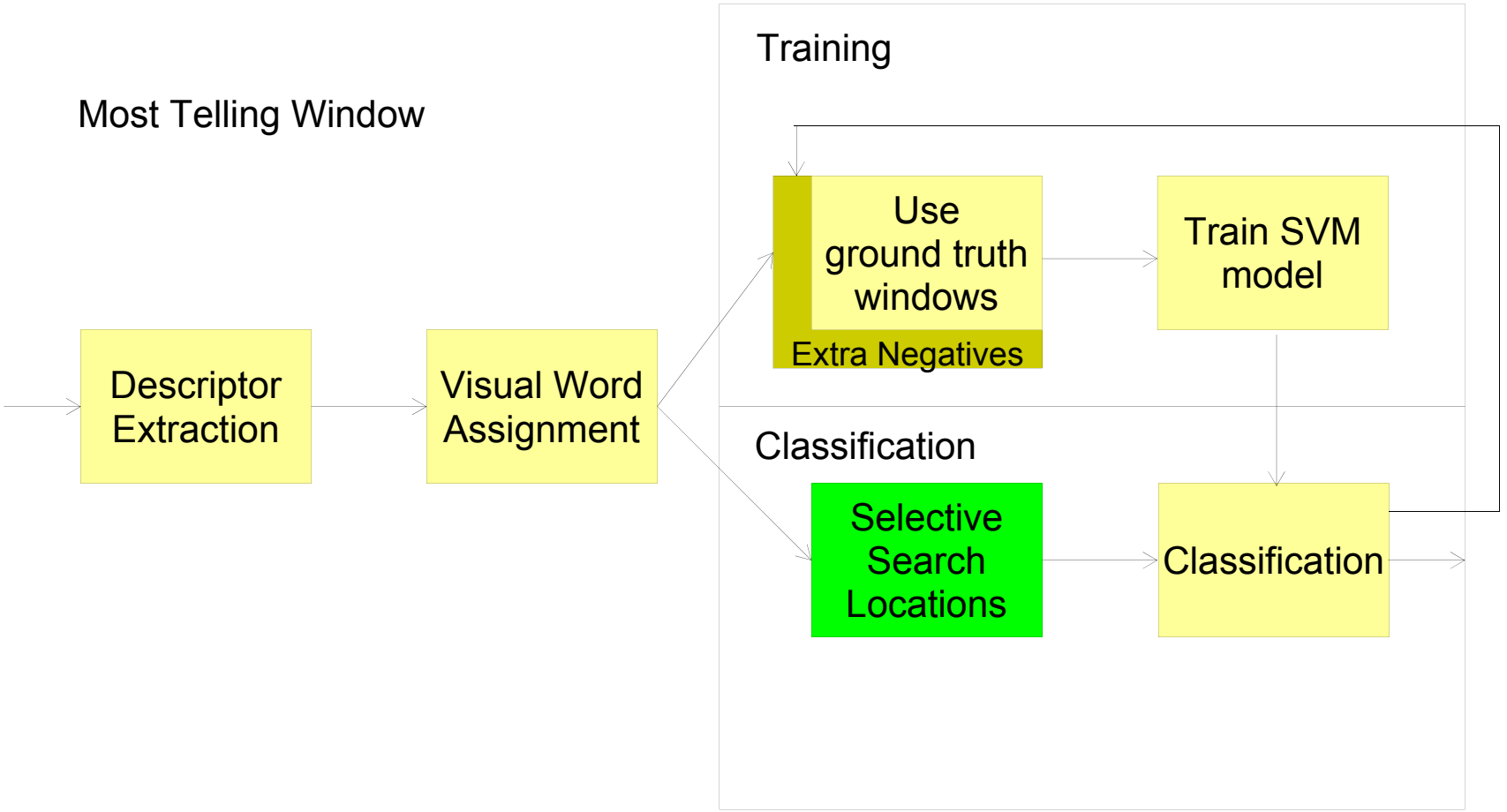
Methodology: Framework



Methodology: Framework



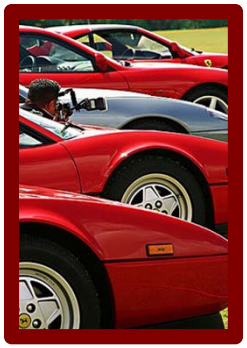
Methodology: Framework



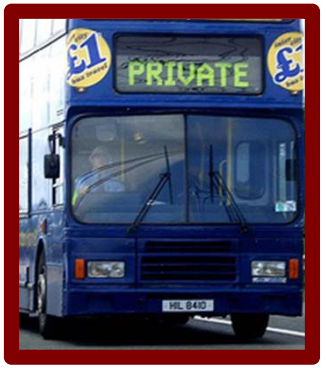
Retraining: e.g. Laptev 2009, Felzenszwalb et al. 2010

Localisation vs Most Telling Window

Localisation



Most Telling Window



No negative examples from positive images!

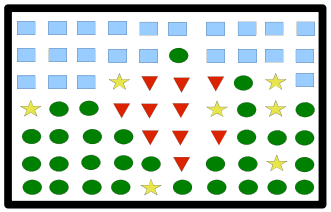
Localisation vs Most Telling Window

- Large difference in motivation:
  - Parts
  - Complete objects
  - Collections of objects

- Subtle difference in training windows
- Significant difference in final results
- (Of course, it would be better to also obtain new positive examples in retraining loop)

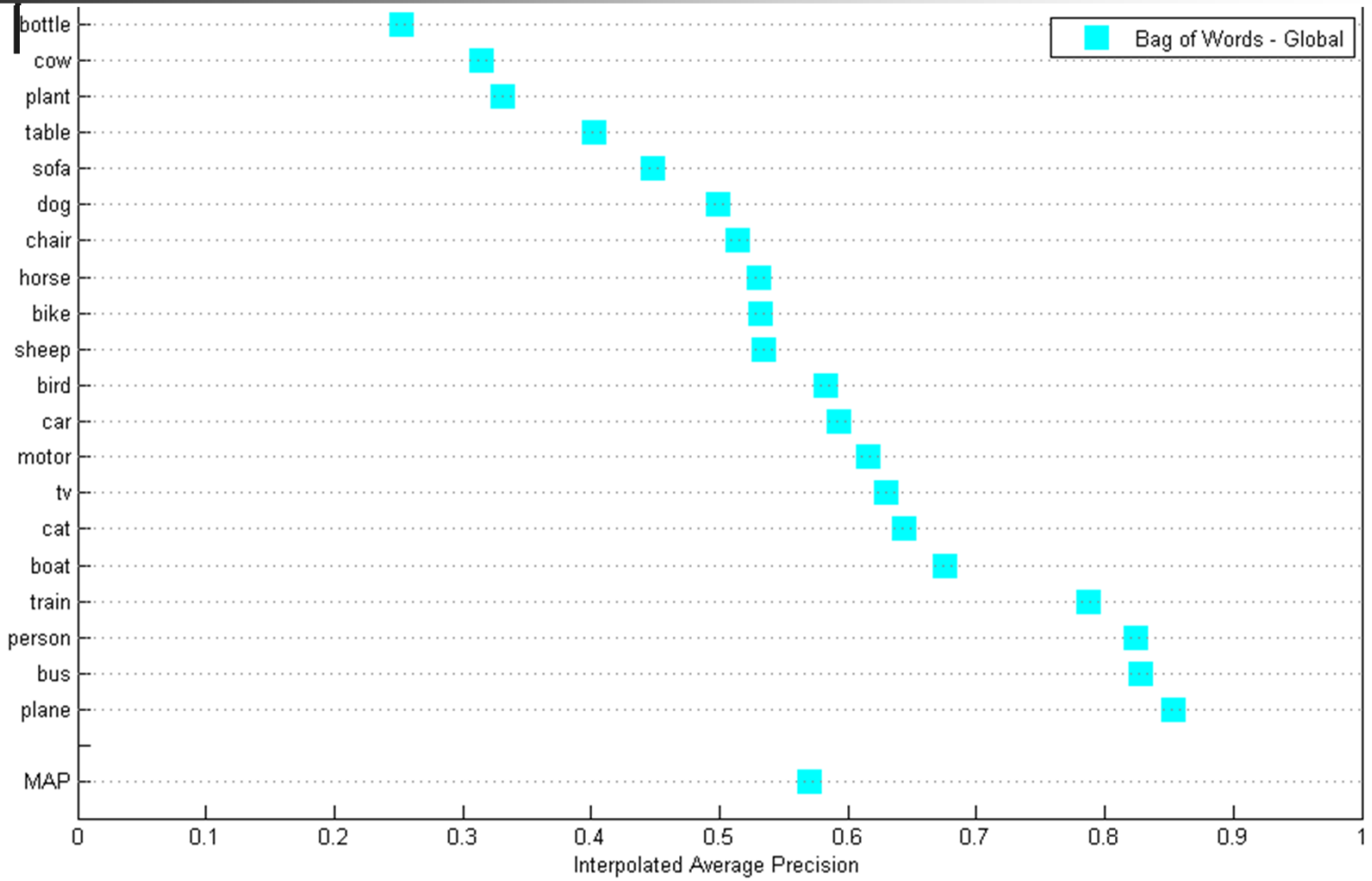
Implementation details

- Pixel-wise sampling
- (Colour) SIFT descriptors (Lowe04, Sande2010)
- K-means visual vocabulary
- Hard assignment.
- Store “Visual Word Images”



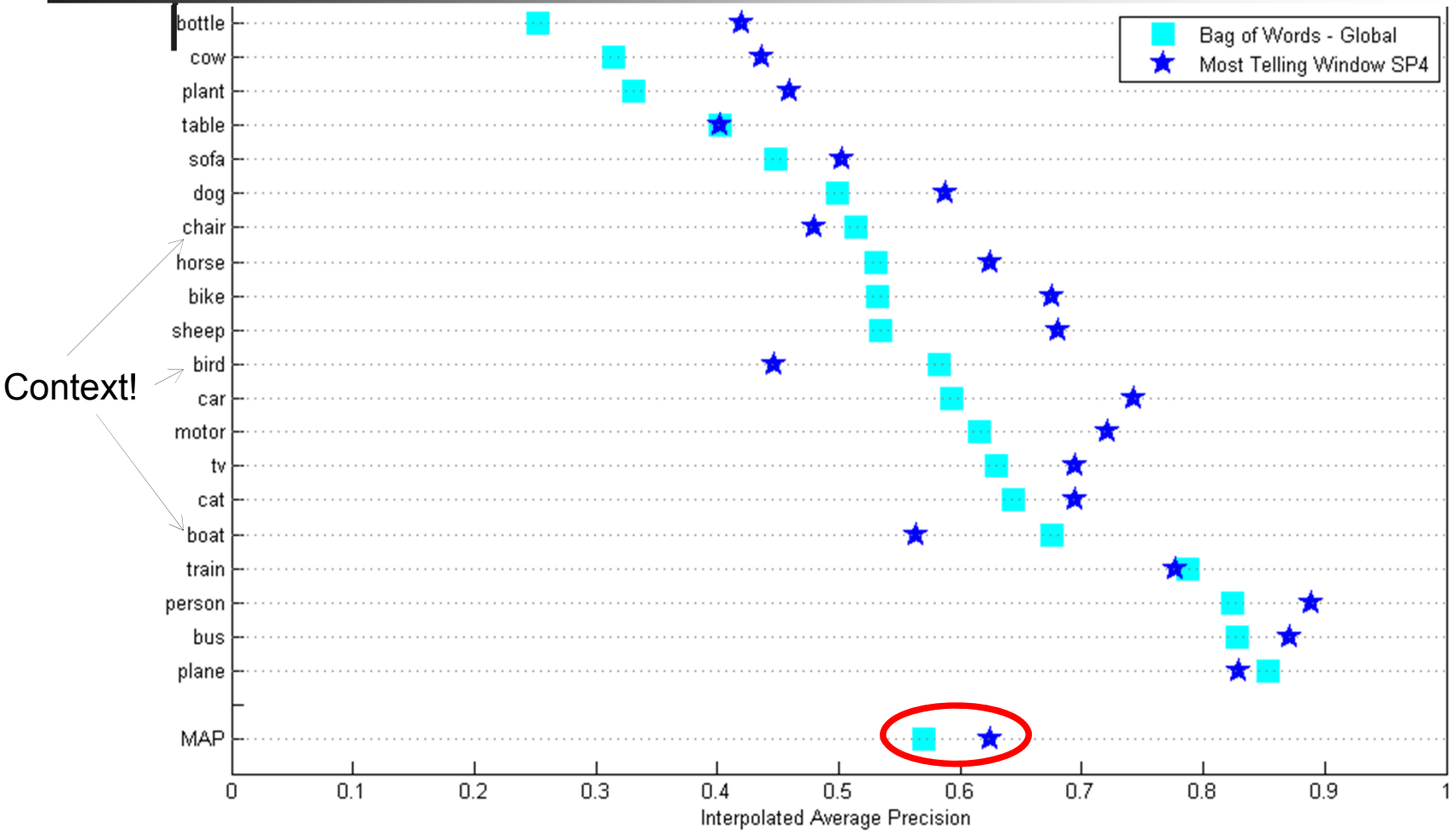
- Spatial Pyramid (Lazebnik06). BoW:1x1,2x2,1x3. MTW:2x2/4x4
- Bag-of-Words GPU acceleration (Sande2011)
- Selective Search
- Support Vector Machine with Histogram Intersection kernel. Fast additive classification (Maji 2009)

Results



Comparable with top scores reported in e.g. Chatfield et al. BMVC 2011
- We: Pixel-wise sampling, 5 Colour SIFT (Sande 2010), kmeans vocabulary 4096
- Chatfield et al.: dense sampling, grey-SIFT only, Fisher/Sparse coding

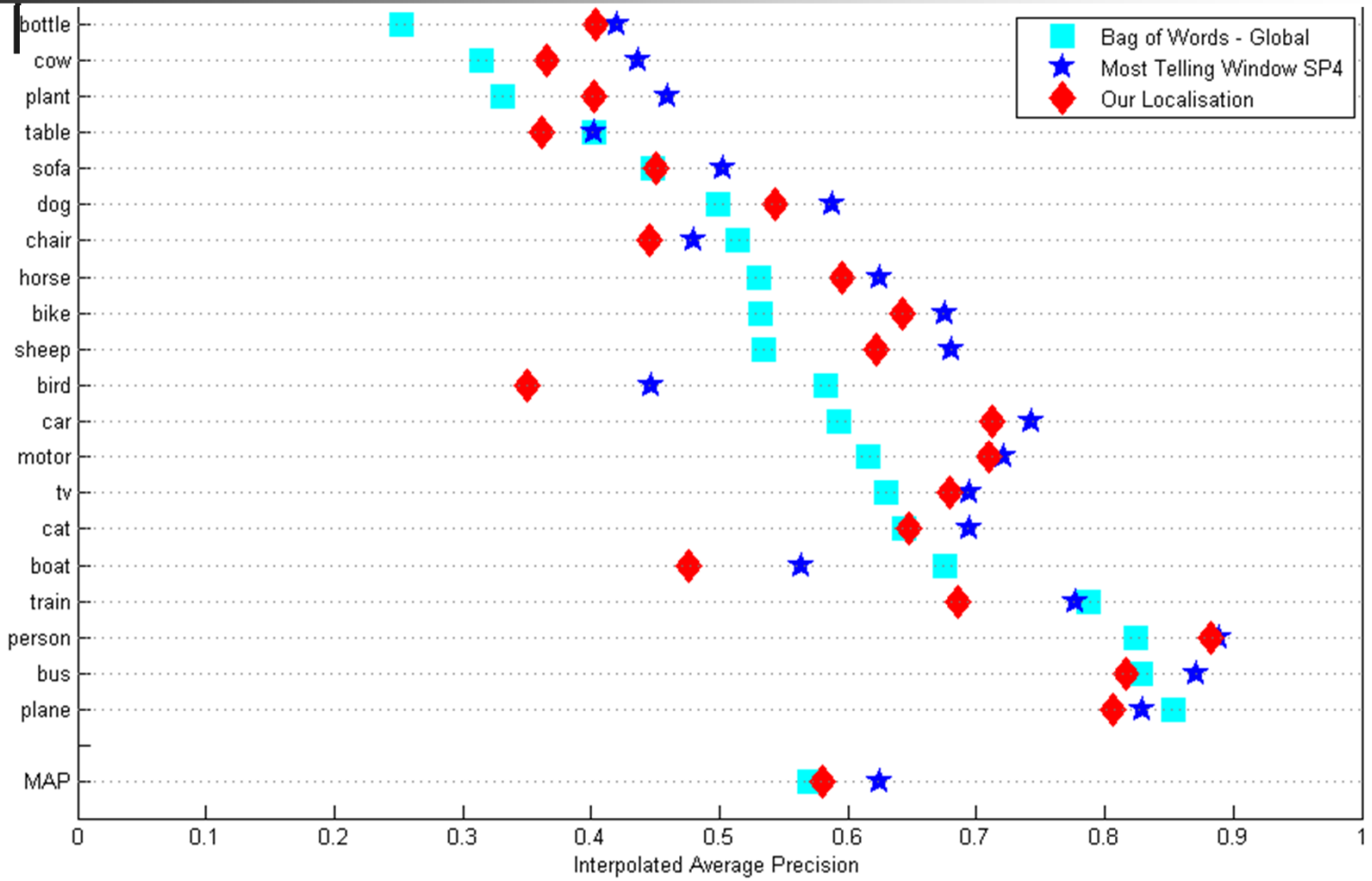
Results



Context!

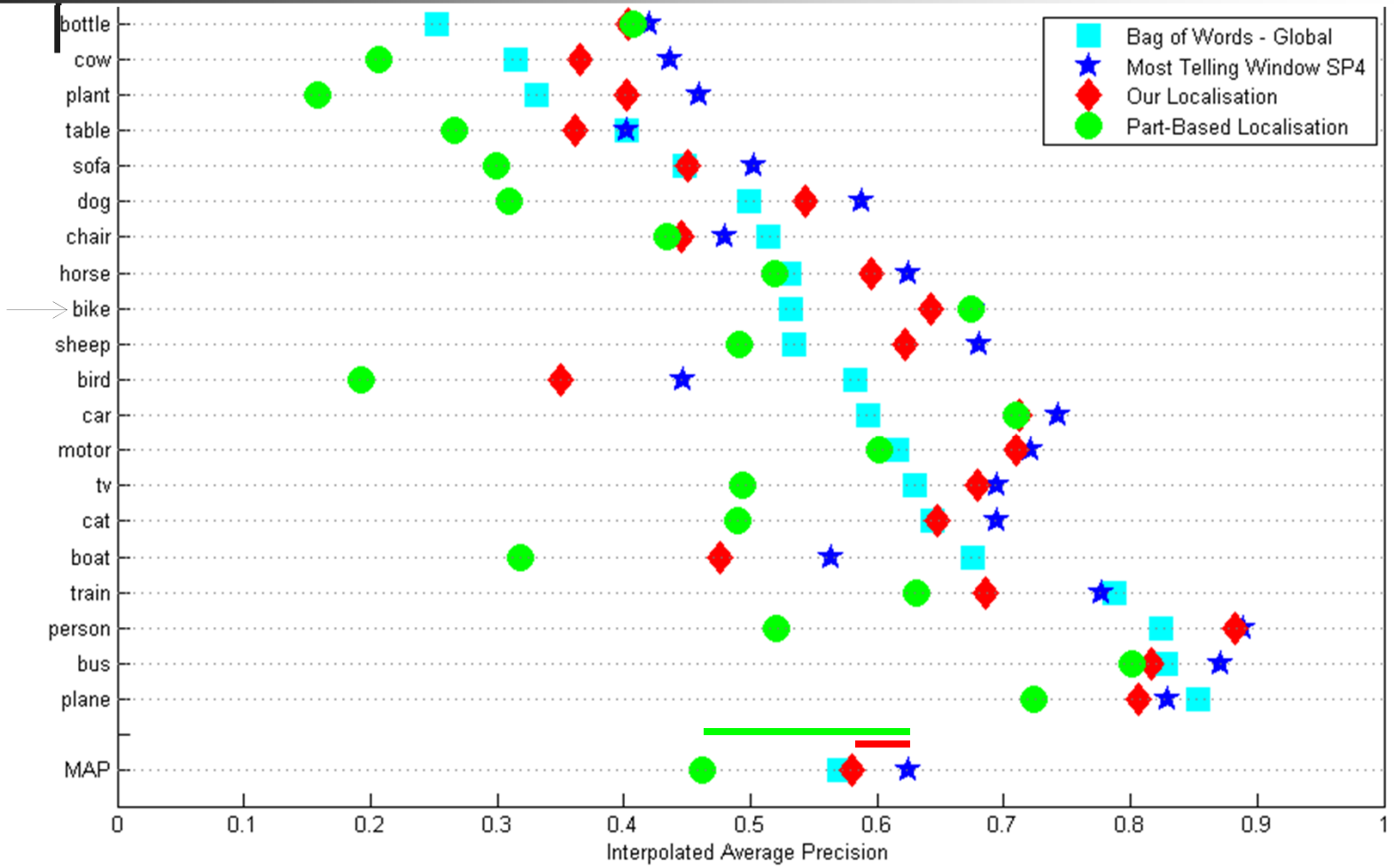
Significant improvement by using not the whole image but its Most Telling Window

Results



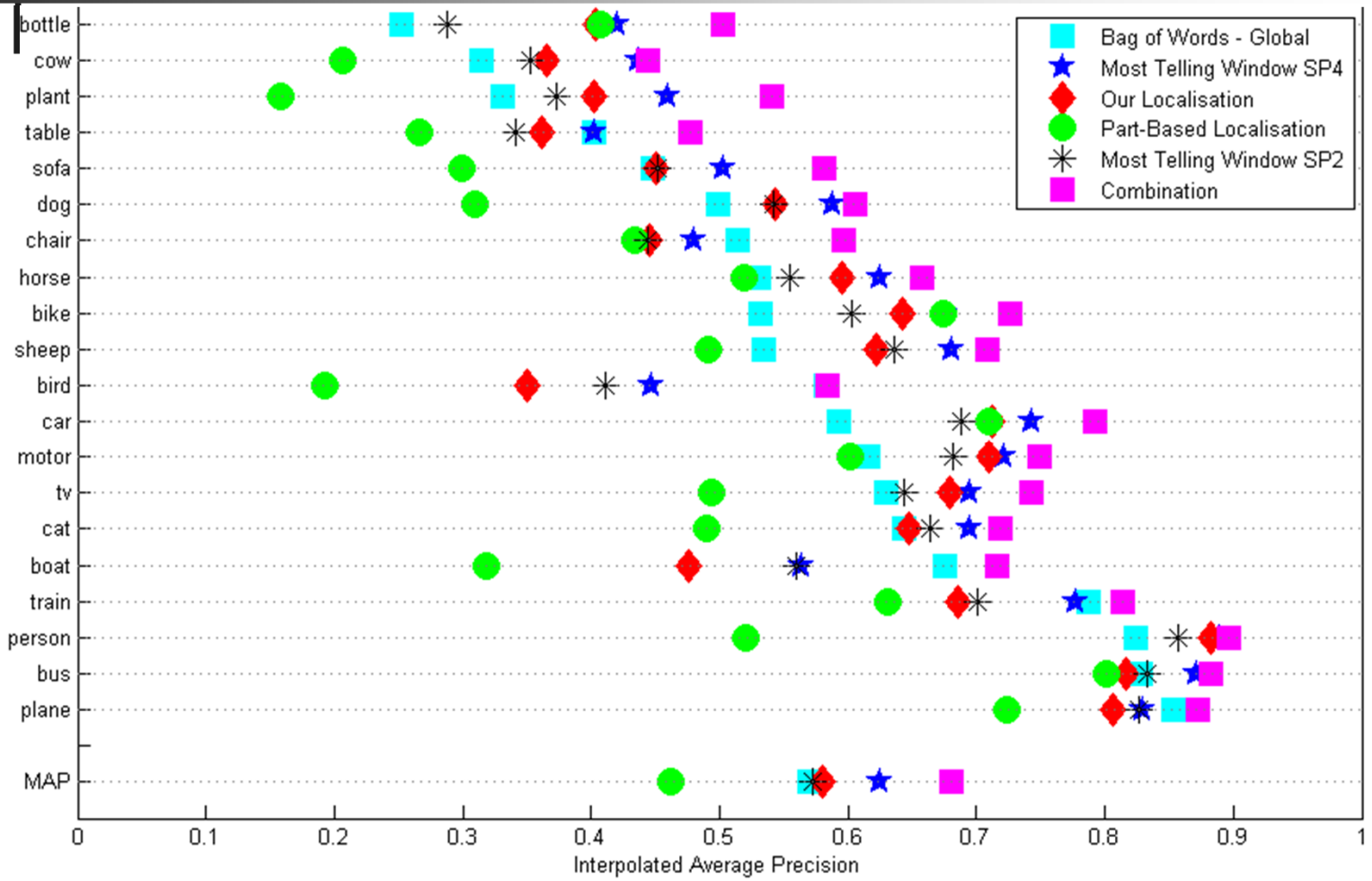
Most Telling Window consistently outperforms Exact Localisation (using same basic framework)

Results



Scores Detection Task: Felzenszwalb: 0.253 MTW: 0.317, Our localisation: 0.336, Discrepancy in results on detection and classification suggests that exact localisation tends to hallucinate objects that are not there while Most Telling Window finds object approximately.

Results



Final combination by cross-validation using weighted addition of classifier output:

- 2 parts Most Telling Window SP 4x4
- 1 part Most Telling Window SP 2x2
- 2 parts Localisation (Felzenszwalb 2010)
- 1 part global Bag-of-Words

3 variations of global Bag-of-Words and our exact localisation were discarded. Location is crucial!

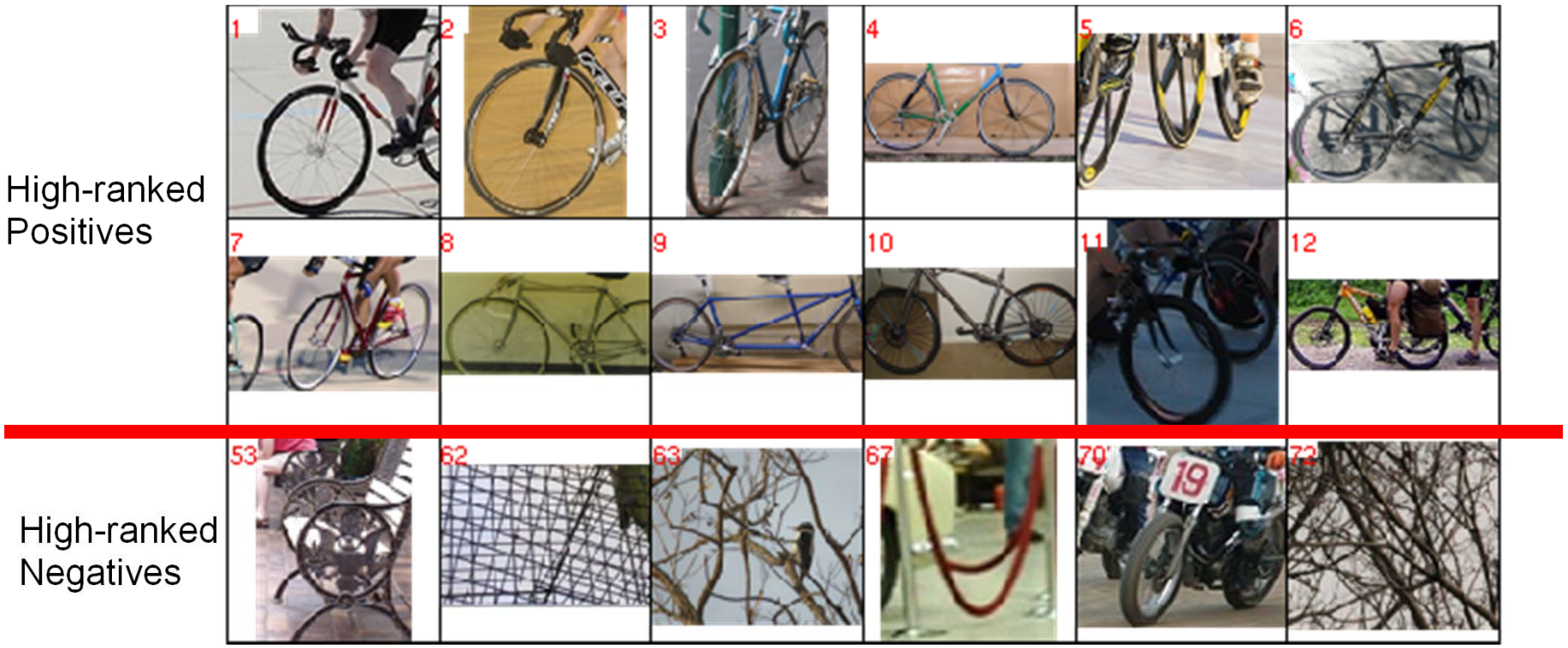
Visualising the Most Telling Window of top-ranked images

High-ranked Positives

High-ranked Negatives

Aeroplane

Visualising the Most Telling Window of top-ranked images



Visualising the Most Telling Window of top-ranked images

High-ranked Positives

High-ranked Negatives

Cat

Visualising the Most Telling Window of top-ranked images

High-ranked Positives

High-ranked Negatives

Cow

Visualising the Most Telling Window of top-ranked images

High-ranked Positives

High-ranked Negatives

Motorcycle

Visualising the Most Telling Window of top-ranked images

High-ranked Positives

High-ranked Negatives

Person

Conclusions Most Telling Window

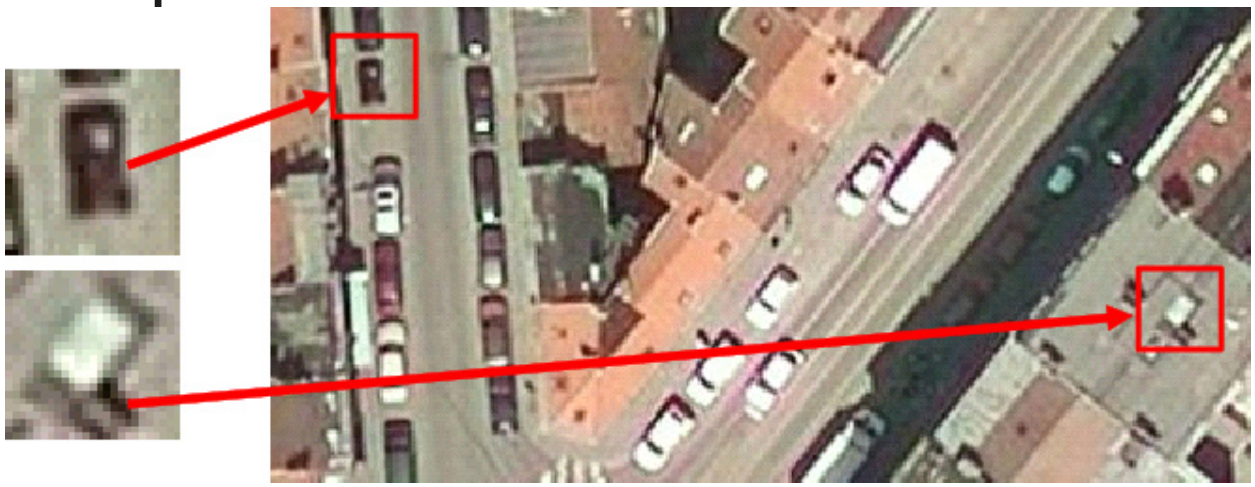
- The Most Telling Window is the window that is the most discriminative for classifying the presence of an object: can be (1) object part; (2) whole object; (3) object collection
- First time that window within the image yields better results by itself than whole image?
- The Most Telling Window works better than exact localisation
- Suboptimal positive windows suggest room for improvement
- Selective Search enables powerful, local Bag-of-Words
  - Class independent parts, wholes, and collections

Outline

- Sliding Window
- Selective Search vs. Sliding Window
- What is Context?
- The Things and Stuff (TAS) model
- Results

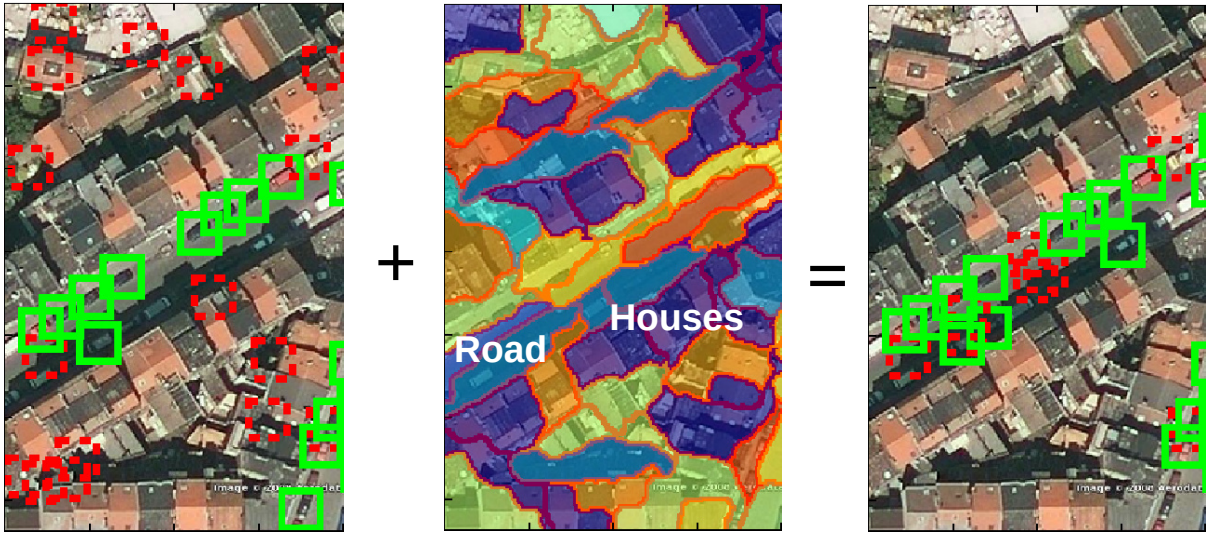
Satellite Detection Example

Detections in Context



Task: Identify all cars in the satellite image

Idea: The surrounding context adds info to the local window detector



Prior:
Detector Only

Region Labels

Posterior:
TAS Model

Error Analysis

Typically…

True Positives are
IN CONTEXT

False Positives are
OUT OF CONTEXT

We need to look outside the bounding box!

Types of Context

- Scene-Thing:

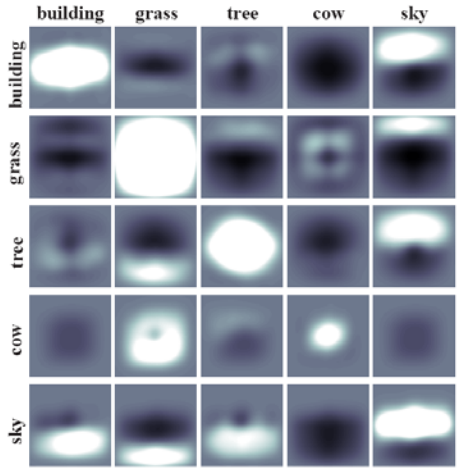
[Torralba et al., LNCS 2005]

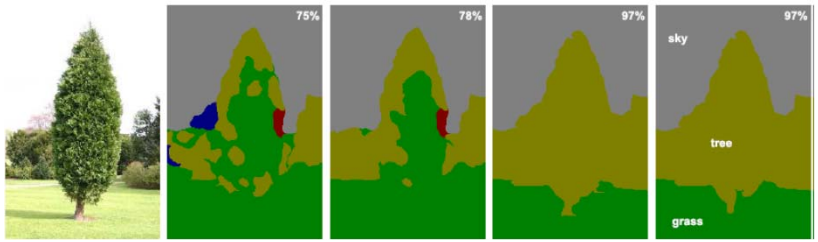
car “likely”

keyboard “unlikely”

- Stuff-Stuff:

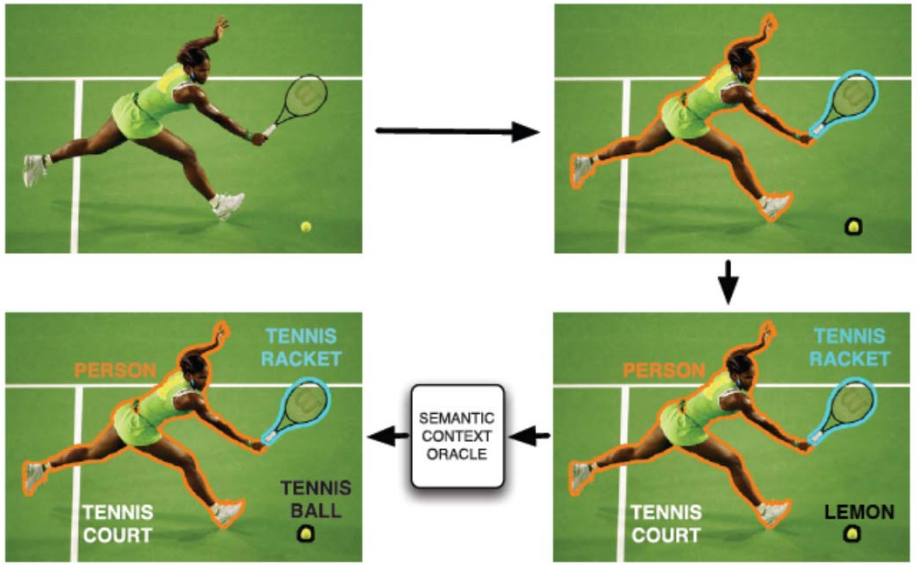
[Gould et al., IJCV 2008]





- Thing-Thing:

[Rabinovich et al., ICCV 2007]



Types of Context

- Stuff-Thing:
  - Based on spatial relationships

- Intuition:

“Cars drive on roads”

“Cows graze on grass”

“Boats sail on water”

Outline

- Selective Search vs. Sliding Window
- What is Context?
- The Things and Stuff (TAS) model
- Results

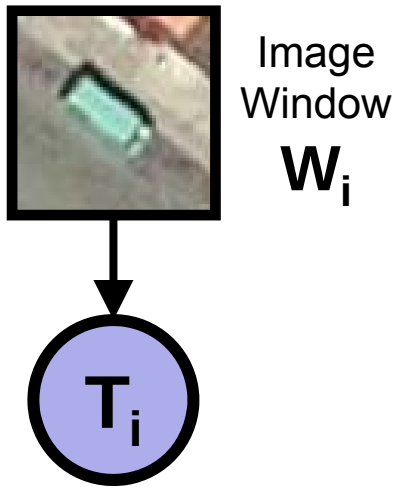
Things

- Detection “candidates”
  - Low detector threshold -> “over-detect”
  - Each candidate has a detector score

Things

- Candidate detections
  - Image Window $\mathbf{W_i}$ + Score
- Boolean random variable $\mathbf{T_i}$
  - Presence/absence of the target class in the window i
- Thing model: conditional prob. from window features to prob. that window contains the object

$$P(T_i|W) = \frac{1}{1+\exp(\alpha + \beta \cdot D(W))}$$

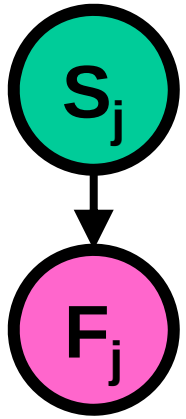


Stuff

- Coherent image regions
  - Segment image into regions: coarse “superpixels”
  - For each region extract color & texture feature vector $\mathbf{F}_j$ in R^n
  - Generative model: each region has a hidden class label $\mathbf{S}_j$ in {1...C}
- Stuff model
  - Naïve Bayes

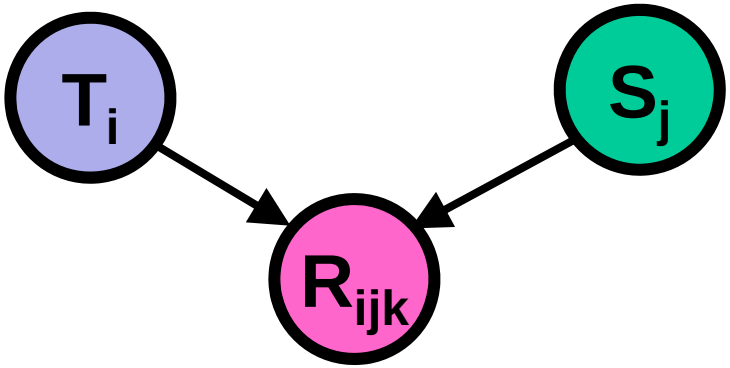
$$P(S_j, F_j) = P(S_j)P\left(F_j \middle| S_j\right)$$

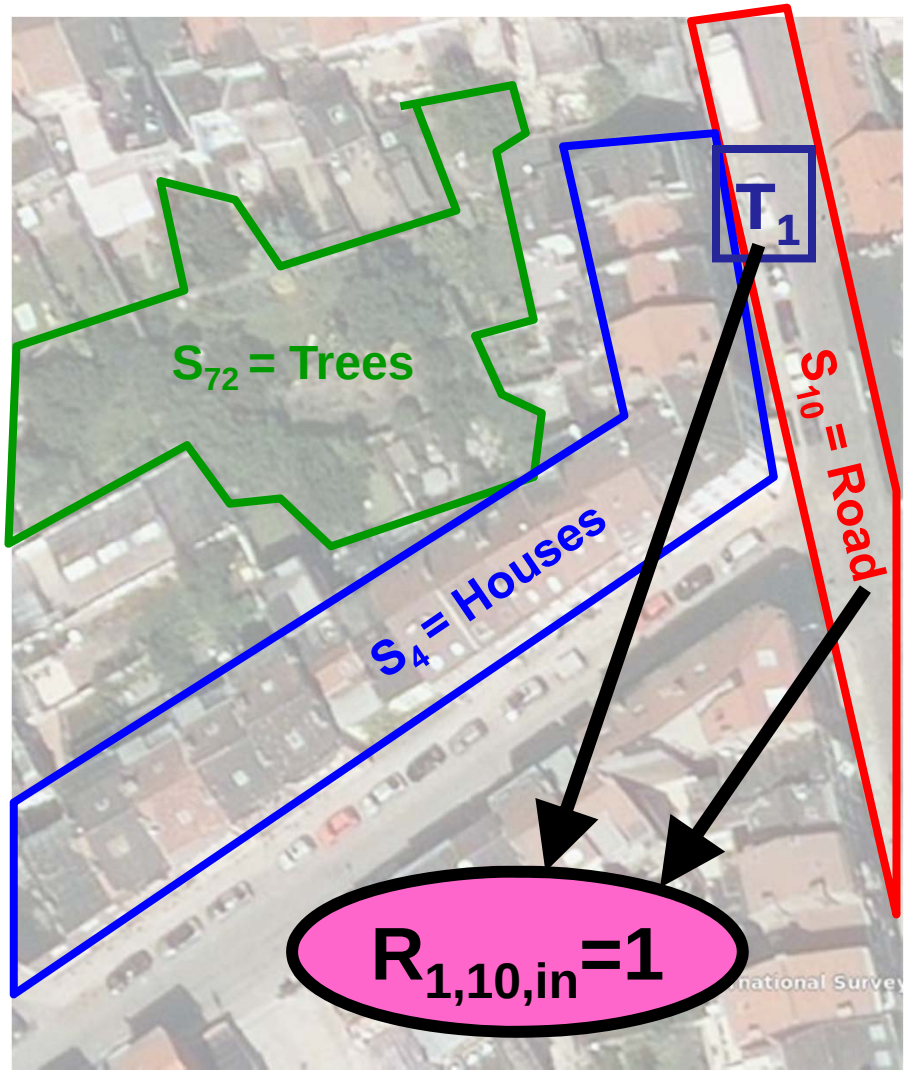
$$F_j \middle| \left(S_j = s\right) \sim \mathrm{N}\left(\mu_s, \Sigma_s\right)$$



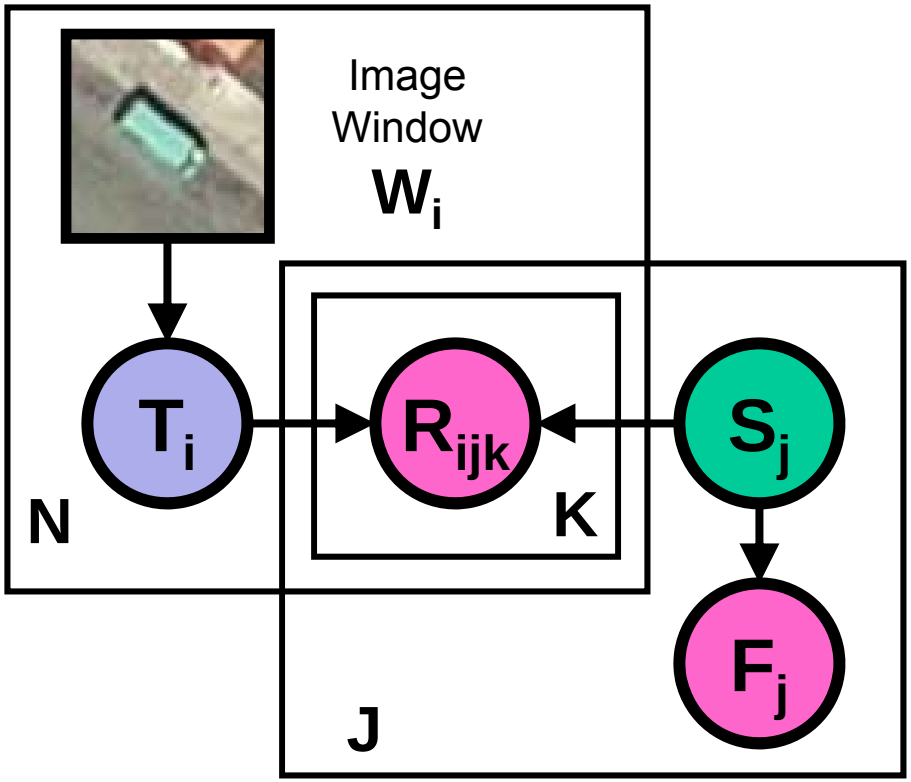
Relationships

- Descriptive Relations
  - "Near", "Above", "In front of", etc.
- Choose set $\mathcal{R} = \{r_1 \ldots r_K\}$
- R_{ijk}=1: Detection i and region j have relation k
- Relationship model





The TAS Model



W_i: Window

T_i: Object Presence

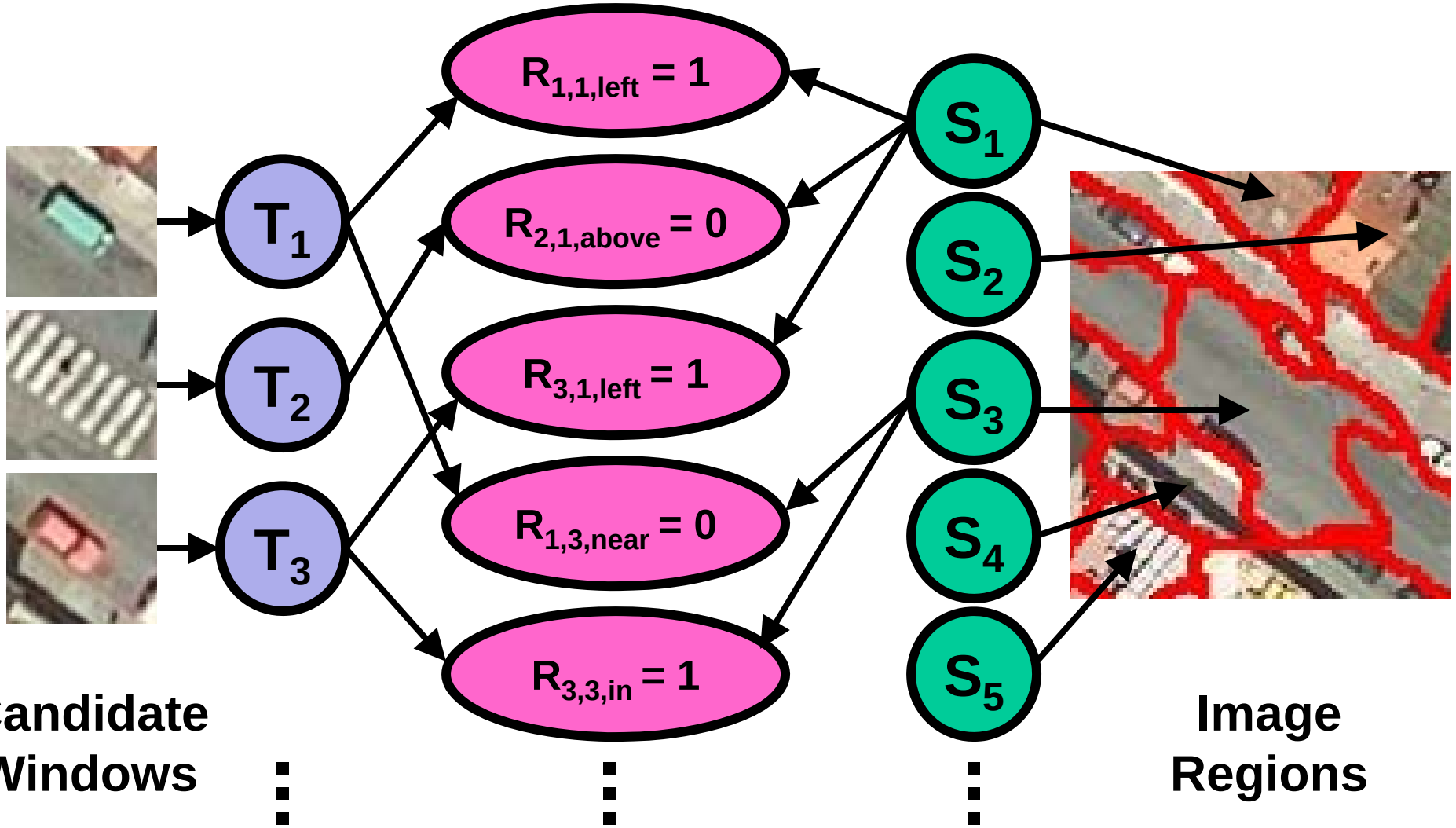
S_j: Region Label

F_j: Region Features

R_{ijk}: Relationship

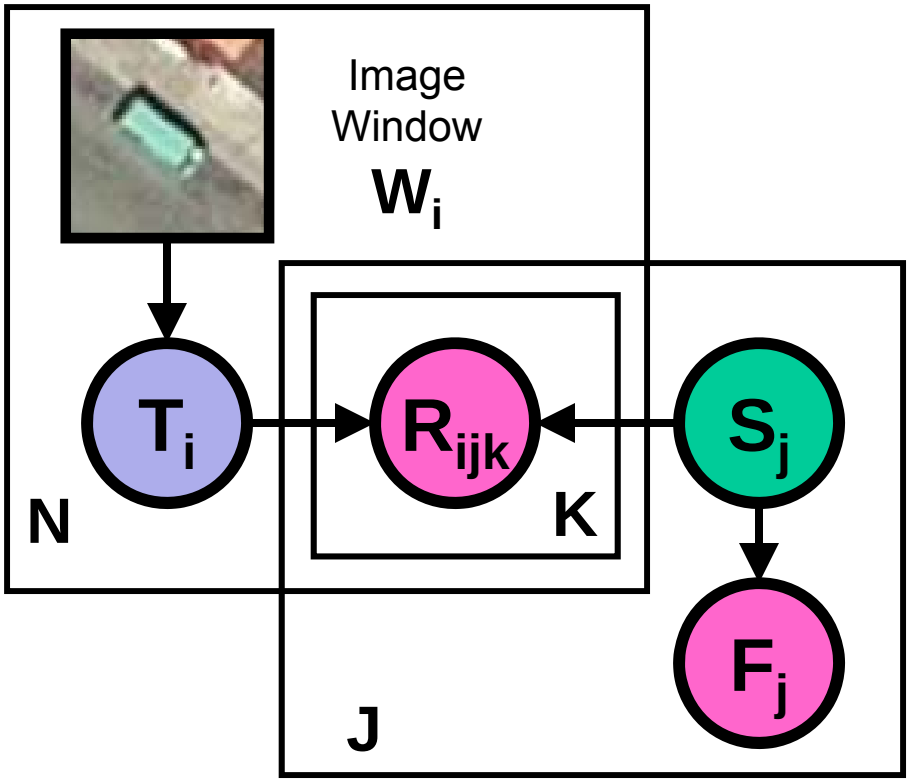
Supervised in Training Set	Always Observed	Always Hidden

Unrolled Model



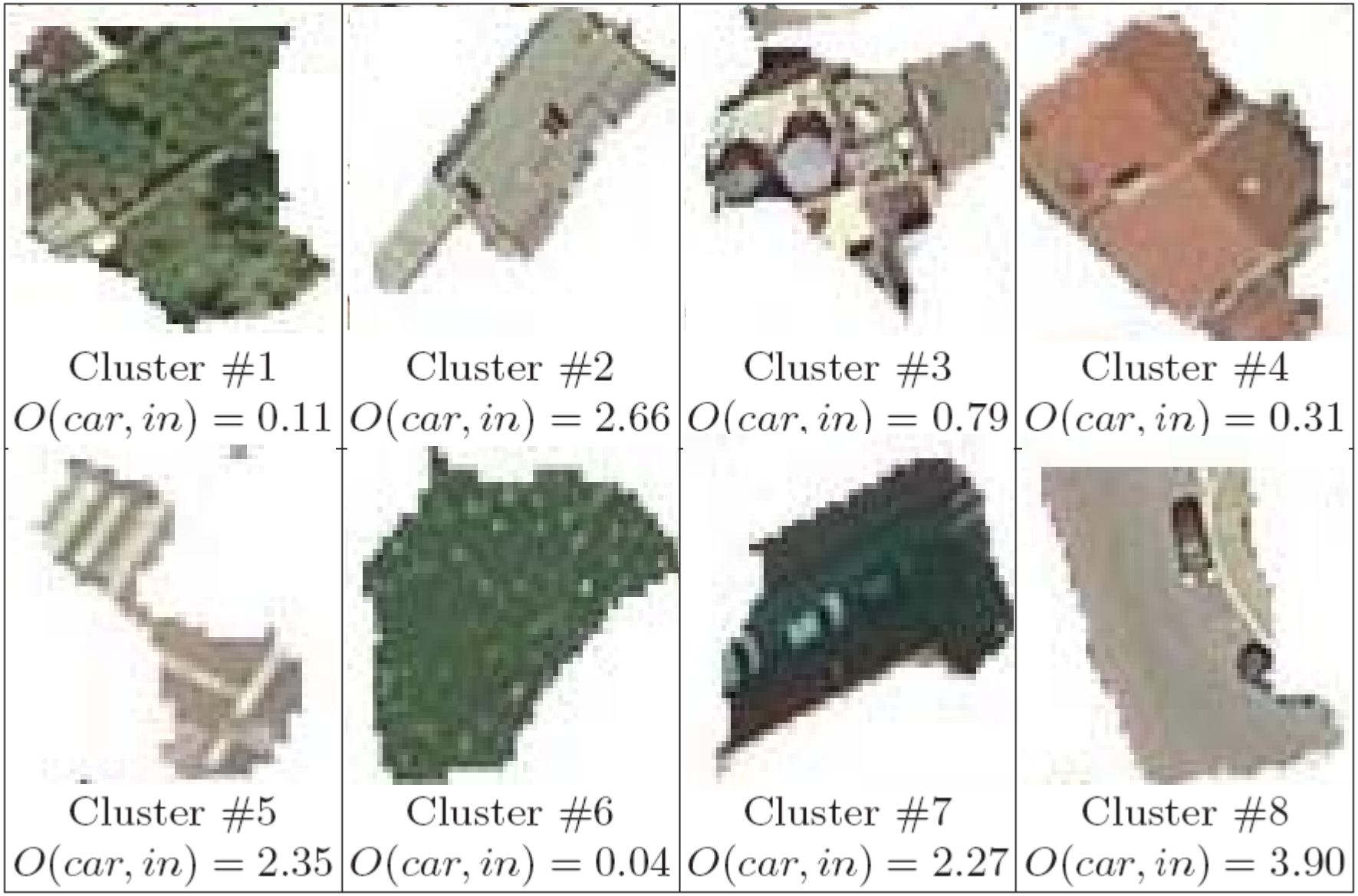
Learning the Parameters

- Assume we know $\mathcal{R}$
- S_j is hidden
  - Everything else observed
- Expectation-Maximization
  - "Contextual clustering"
- Parameters are readily interpretable



Supervised in Training Set	Always Observed	Always Hidden

Learned Satellite Clusters



Which Relationships to Use?

- Rijk = spatial relationship between candidate i and region j

Rij1 = candidate in region

Rij2 = candidate closer than 2 bounding boxes (BBs) to region

How do we avoid overfitting?

...

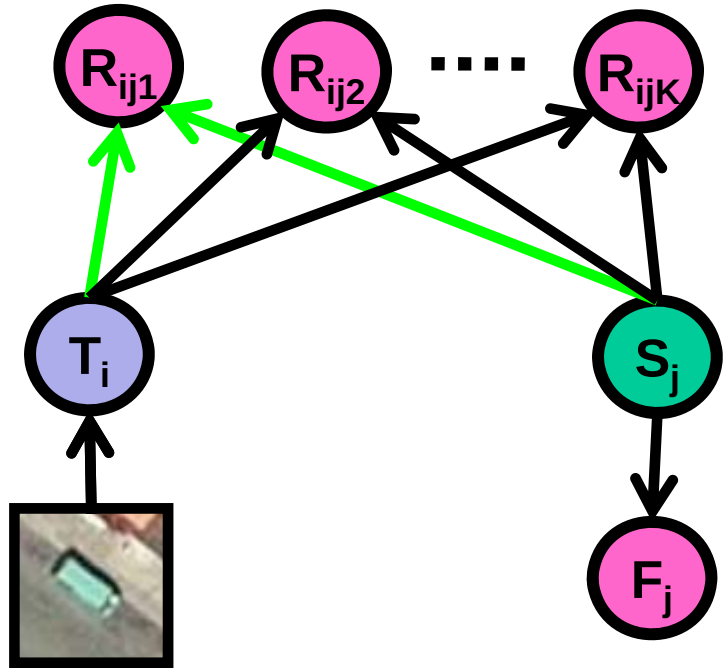
RijK = candidate near region boundary

Inference

- So far, we assumed a known set of relationships
- But, different data may require different types of contextual relationships => learn which one to use
  - Define a large set C of “candidate relationships” (i.e., all possible relationships to be included)
  - Search through C for the subset of “active” relationships R that best facilitates the use of context
  - If a relationship is “inactive” => remove the edges from all T_i and S_j to the R_{ijk} variables for this particular k.
  - With this view of “activating” relationships by including the edges in the Bayesian Network, we can formulate our search for R as a structure learning problem

Learning the Relationships

- Intuition
  - "Detached" R_{ijk} = inactive relationship
- Structural EM iterates:
  - Learn parameters
  - Decide which edge to toggle
- Evaluate with $\ell(T|F,W,R)$
  - Requires inference
  - Better results than using standard $E[\ell(T,S,F,W,R)]$



Learning the Relationships

Algorithm LearnTAS

Input: Candidate relationships $\mathcal{C}$, Dataset $\mathcal{D} = \{(\boldsymbol{W}[m], \boldsymbol{T}[m], \boldsymbol{F}[m], \boldsymbol{R}[m])\}$

$\mathcal{R} \leftarrow \emptyset$ (all relationships "inactive")

Repeat until convergence

Repeat until convergence (EM over Parameters)

E-step: $Q[m] \leftarrow P(\boldsymbol{S} \mid \boldsymbol{T}, \boldsymbol{F}, \boldsymbol{R}; \theta_{\mathcal{R}}) \quad \forall m$

M-step: $\theta_{\mathcal{R}} \leftarrow \text{argmax} E_Q \left[\sum_m \ell(\boldsymbol{S}, \boldsymbol{T}, \boldsymbol{F}, \boldsymbol{R} \mid \boldsymbol{W}; \theta_{\mathcal{R}})\right]$

Repeat until convergence (Greedy Structure Search)

Forall k, $score_k = \sum_m \ell(\boldsymbol{T} \mid \boldsymbol{F}, \boldsymbol{R}; \theta_{\mathcal{R} \oplus k})$ (score with k "activated")

$\mathcal{R} \leftarrow \mathcal{R} \oplus k^*$ where $k^* = \text{argmax } score_k$

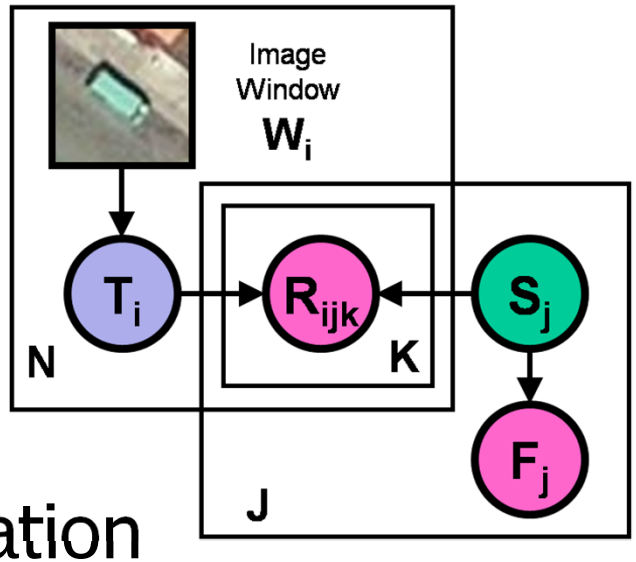
Return Set $\mathcal{R}$ of "active" relationships, TAS parameters $\theta_{\mathcal{R}}$

Learning a TAS model. Here ℓ represents the log-likelihood of the data, and $\oplus$ represents the set exclusive-or operation.

Inference

- Goal: find the probability that each window contains the object

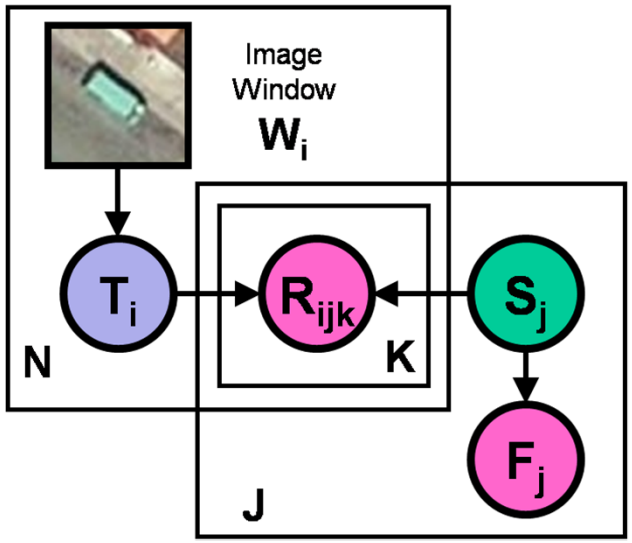
$$P(\boldsymbol{T} \mid \boldsymbol{F}, \boldsymbol{R}, \boldsymbol{W}) = \sum_{\boldsymbol{S}} P(\boldsymbol{T}, \boldsymbol{S} \mid \boldsymbol{F}, \boldsymbol{R}, \boldsymbol{W})$$



- This expression involves a summation over an exponential set of values for the S vector of variables
  - solve the inference problem approximately using a Gibbs sampling MCMC method (Geman&Geman, 1987)

Inference

- Block Gibbs Sampling
  - Initial assignment to the variables
  - in each Gibbs iteration resample all of the S's and then resample all the T's according to the following two probabilities:



$$P(S_j \mid \boldsymbol{T}, \boldsymbol{F}, \boldsymbol{R}, \boldsymbol{W}) \propto P(S_j) P(F_j \mid S_j) \prod_i P(R_{ij} \mid T_i, S_j)$$

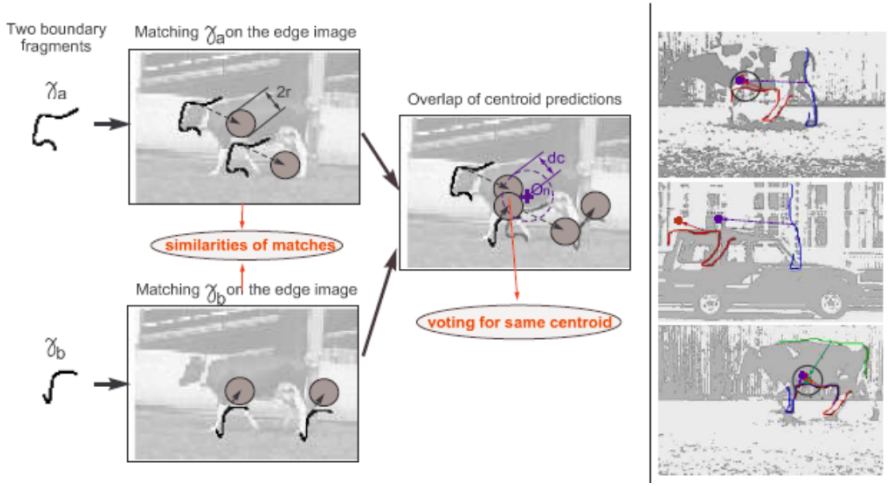
$$P(T_i \mid \boldsymbol{S}, \boldsymbol{F}, \boldsymbol{R}, \boldsymbol{W}) \propto P(T_i \mid W_i) \prod_j P(R_{ij} \mid T_i, S_j).$$

Outline

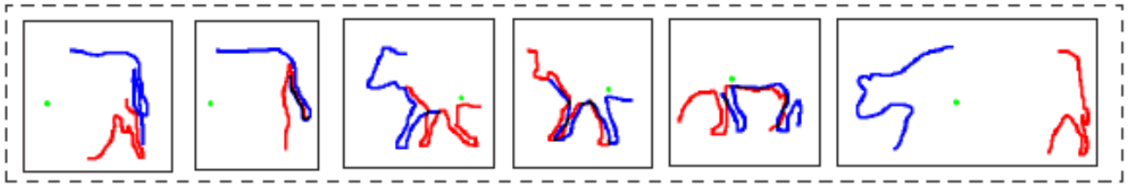
- Selective Search vs. Sliding Window
- What is Context?
- The Things and Stuff (TAS) model
- Results

Features: Edge fragments

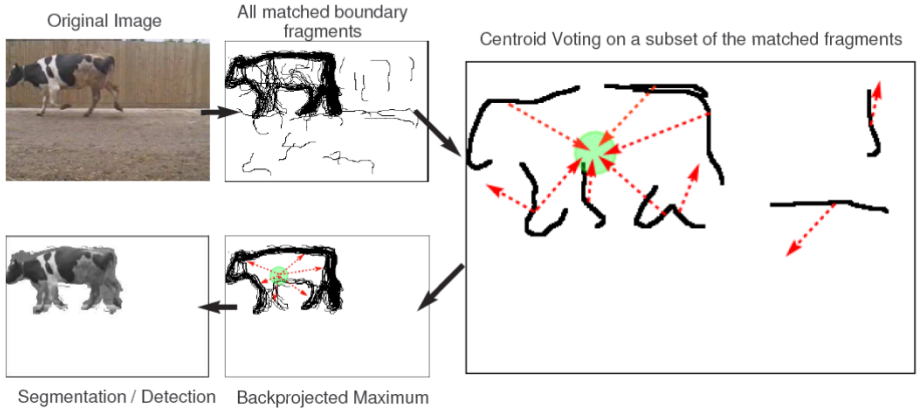
Opelt, Pinz, Zisserman, ECCV 2006



Weak detector = Match of edge chain(s) from training image to edgemap of test image



BOOSTING!



Base Detector - HOG

- HOG Detector: [Dalal & Triggs, CVPR, 2006]

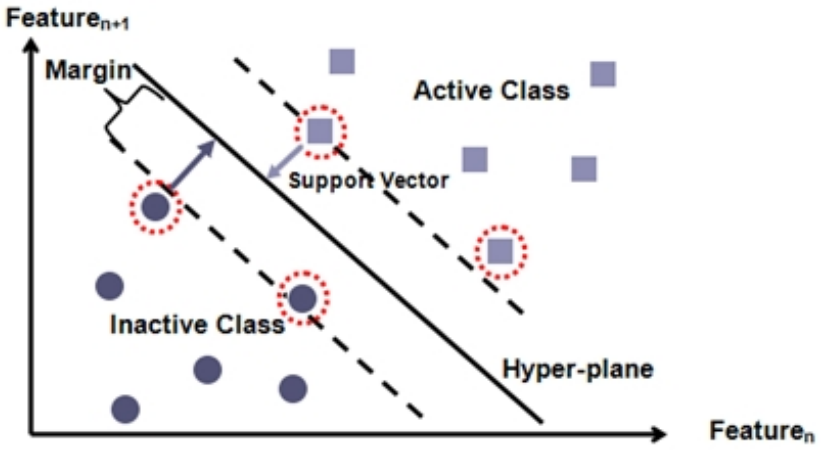
Feature Vector X

input image

weighted pos wts

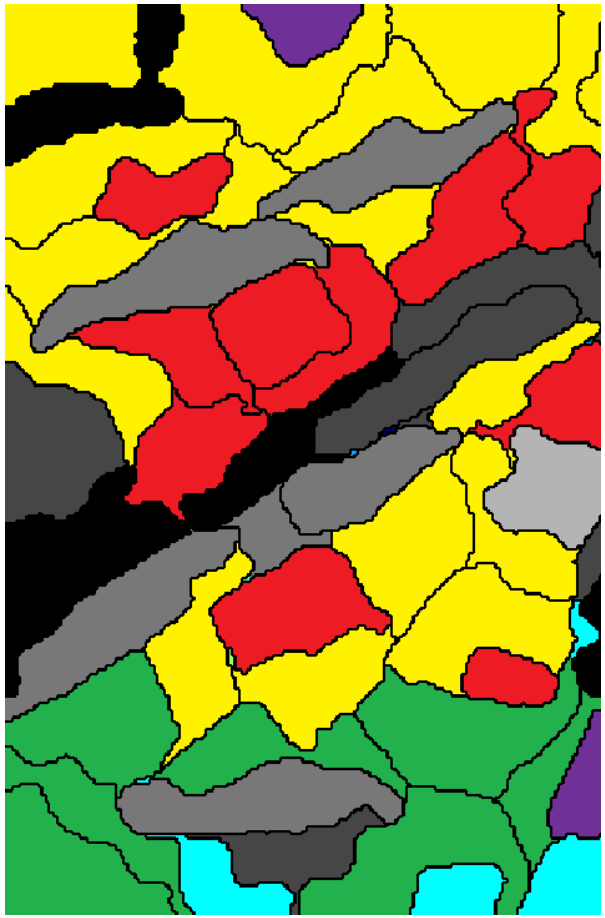
weighted neg wts

SVM Classifier



Results - Satellite

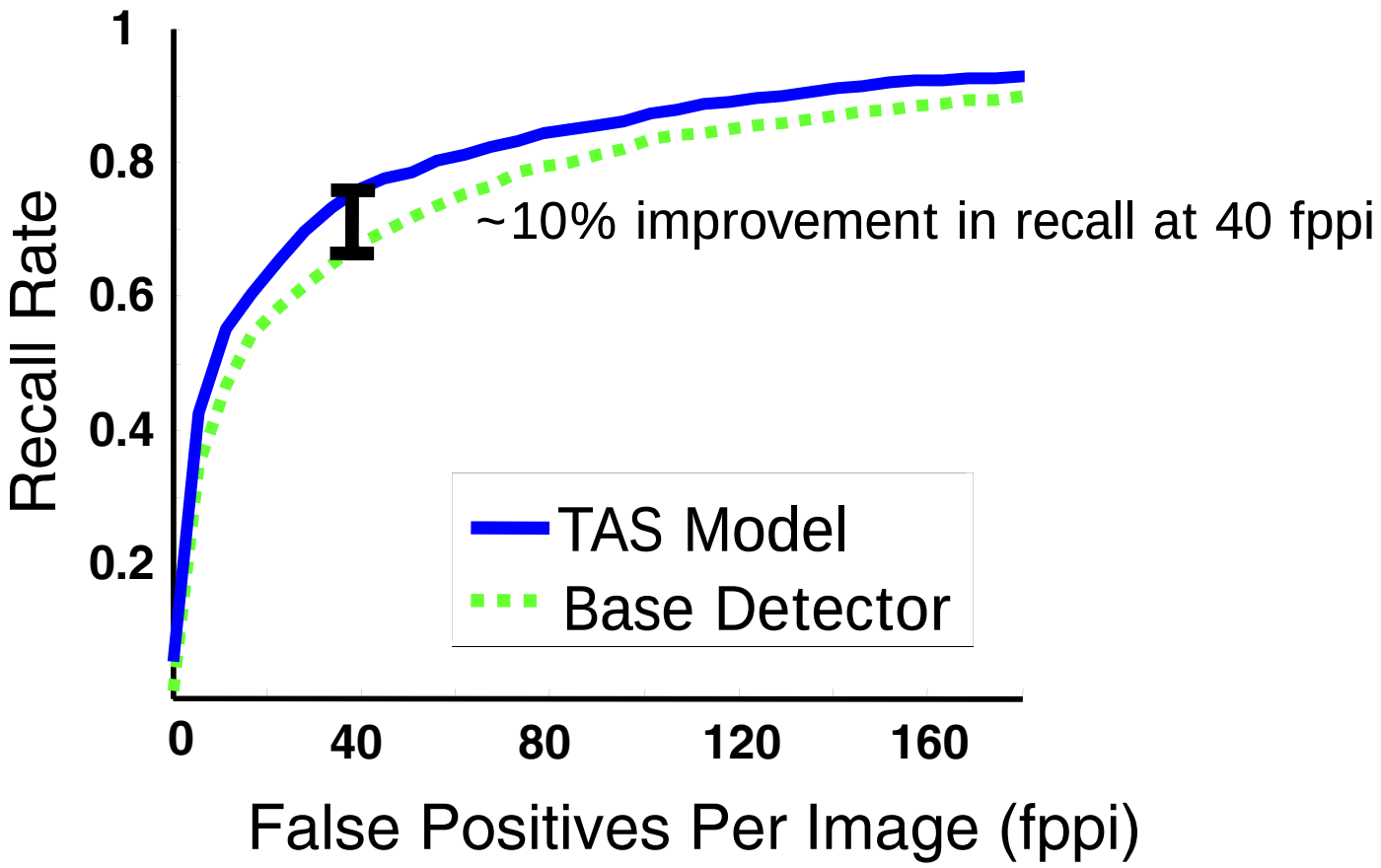
Prior:
Detector Only



Posterior:
Region Labels

Posterior:
Detections

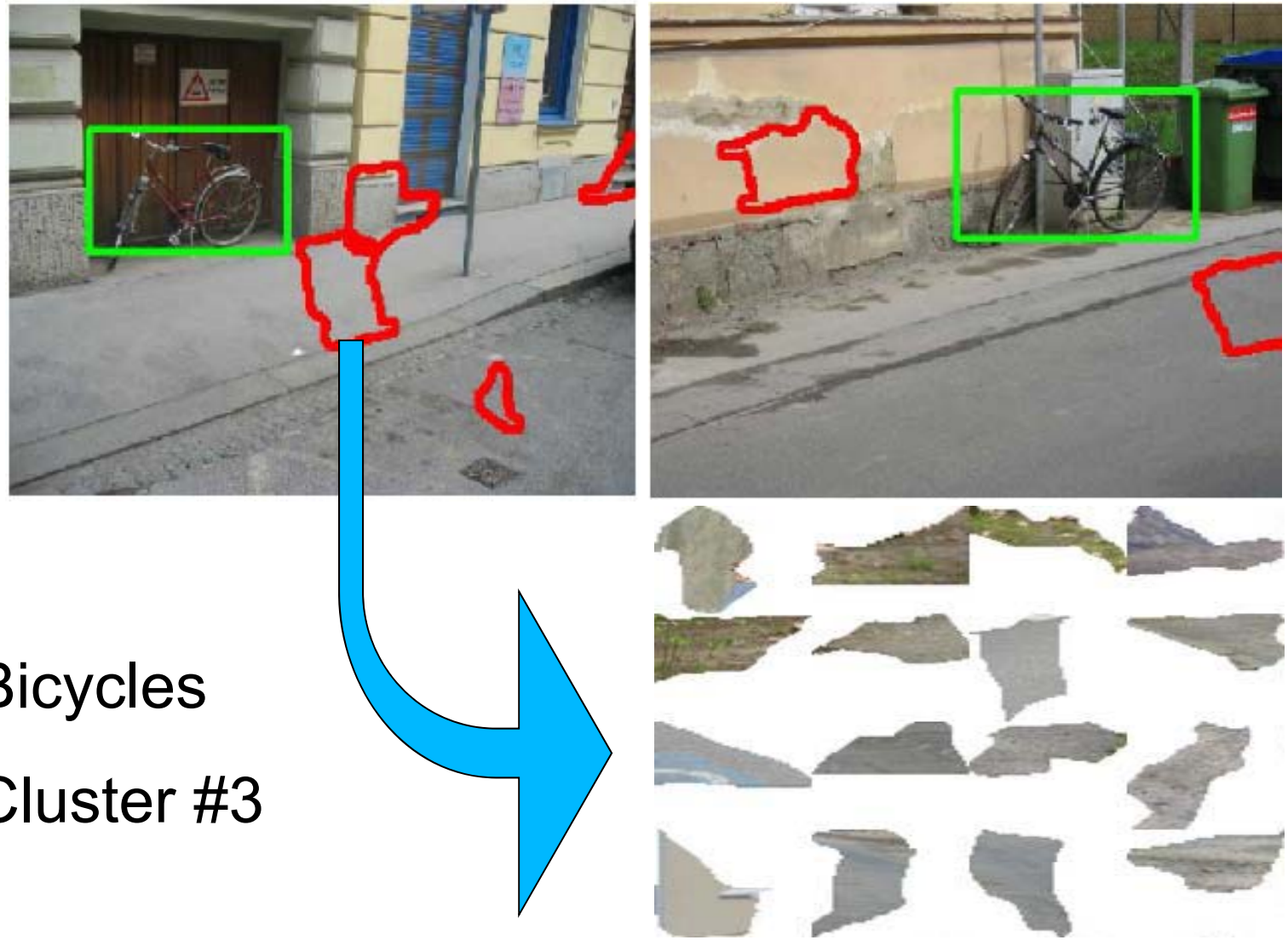
Results - Satellite



Base Detector Error Analysis

Cows

Discovered Context - Bicycles



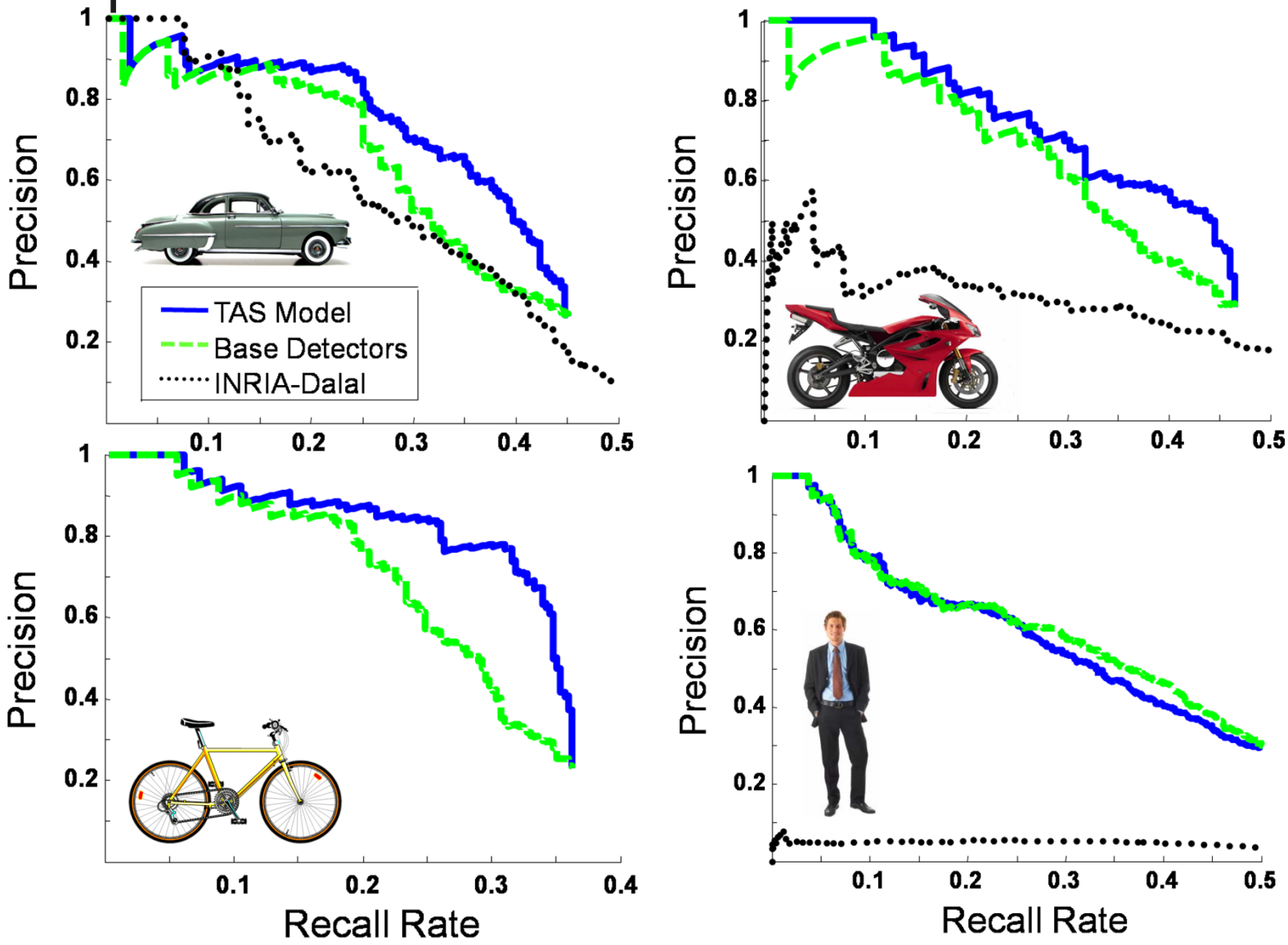
Bicycles

Cluster #3

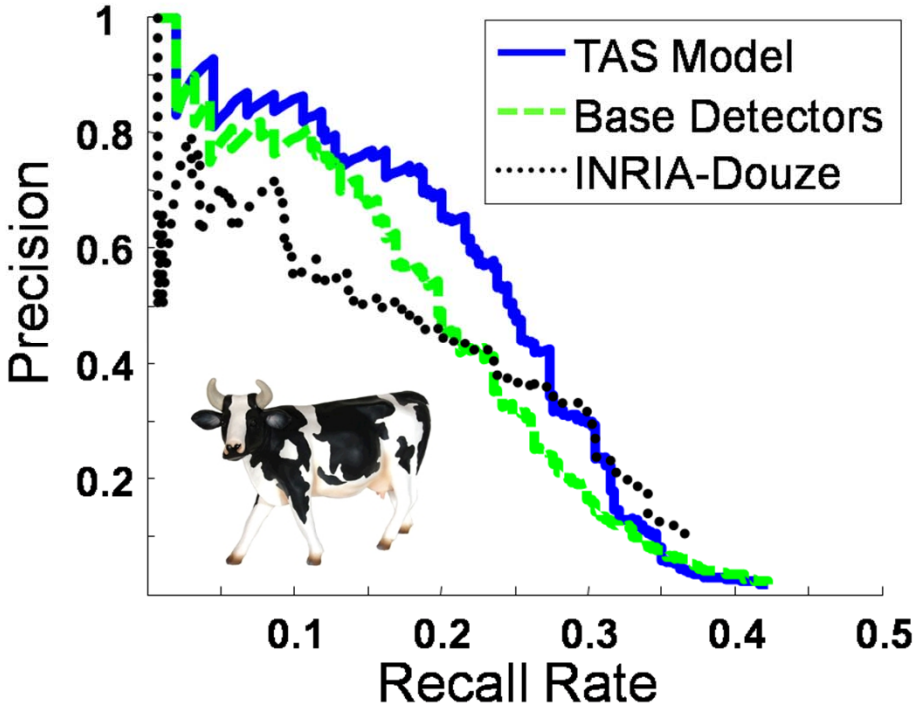
TAS Results – Bicycles

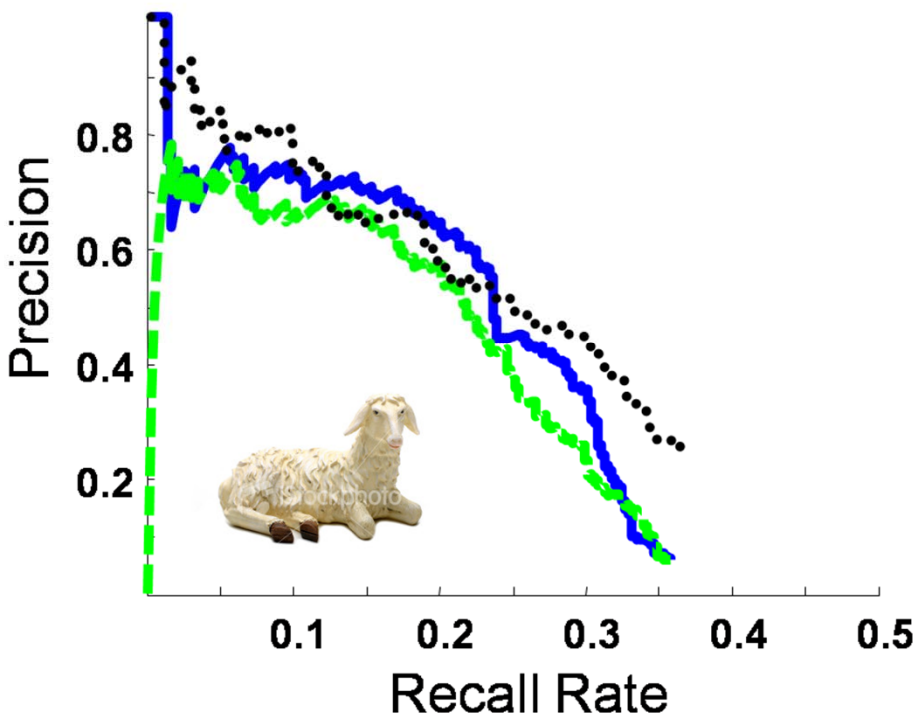
- Examples
  - Discover “true positives”
  - Remove “false positives”

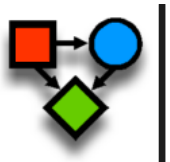
Results – VOC 2005



Results – VOC 2006







Object Class	Base AP	TAS AP (Fixed R)	TAS AP (Learned R)	Improvement (TAS - Base)
Cars	0.325	0.360	**0.363**	0.038
Motorbikes	0.341	**0.390**	0.373	0.032
People	**0.346**	0.346	0.337	-0.009
Bicycles	0.281	0.310	**0.325**	0.044
Cows	0.224	0.241	**0.258**	0.034
Sheep	0.206	0.233	**0.248**	0.042

Conclusions

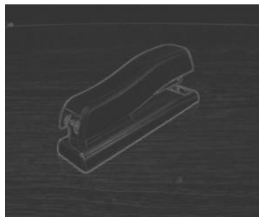
- Detectors can benefit from context
- The TAS model captures an important type of context
- Can improve *any* sliding window/selective search detector using TAS
- The TAS model can be interpreted and matches our intuitions
- We can learn which relationships to use

Traditional Recognition Approach

Features are not learned

Input data (pixels) → feature representation (hand-crafted) → Learning Algorithm (e.g., SVM)

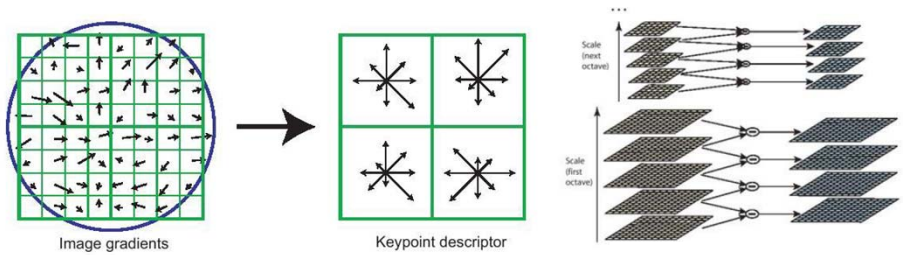
Image



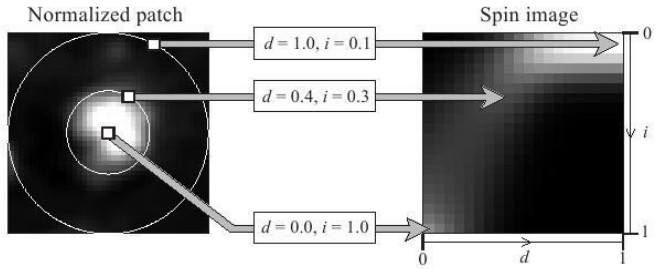
Low-level vision features (edges, SIFT, HOG, etc.)

Object detection / classification

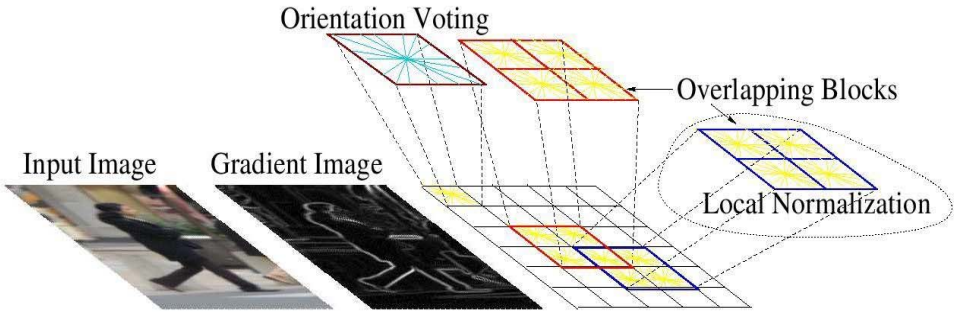
Computer vision features



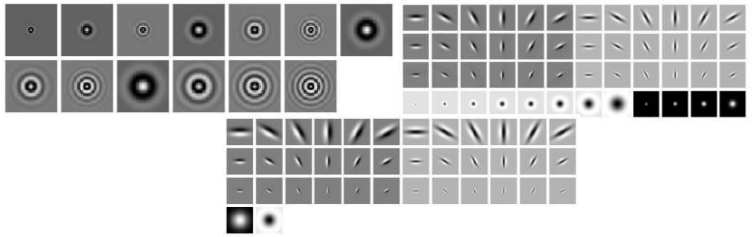
SIFT



Spin image



HoG



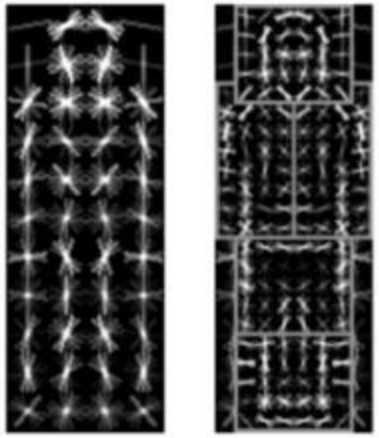
Textons

and many others:

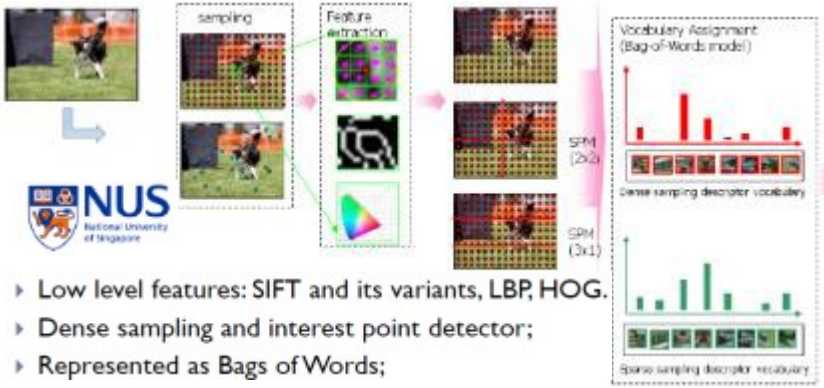
SURF, MSER, LBP, Color-SIFT, Color histogram, GLOH, …..

Motivation

- Features are key to recent progress in recognition
- Multitude of hand-designed features currently in use
- Where next? Better classifiers? Building better features?



Felzenszwalb, Girshick, McAllester and Ramanan, PAMI 2007

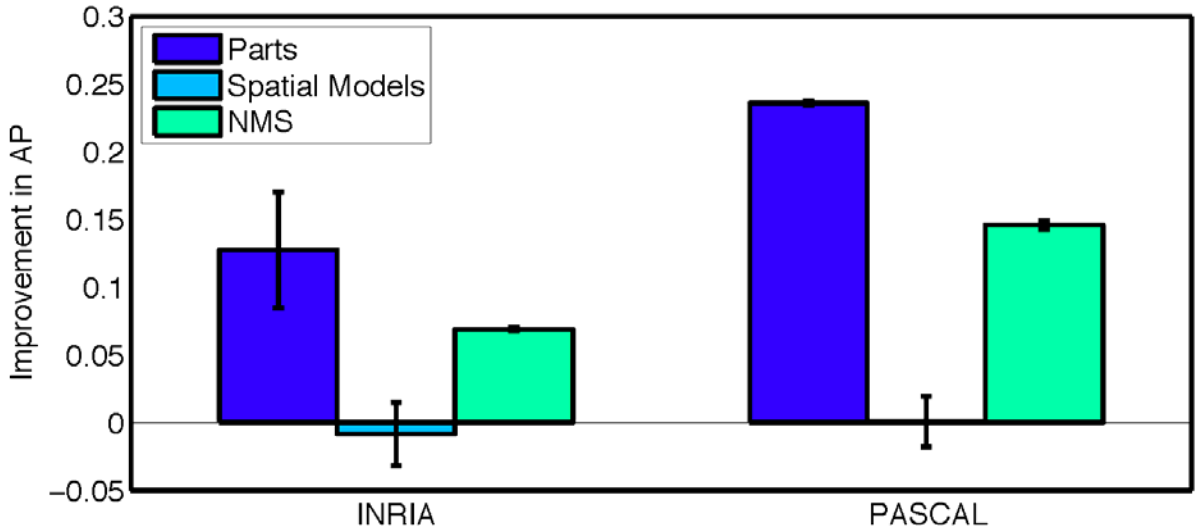


Yan & Huang
(Winner of PASCAL'10 classification competition)

Slide: R. Fergus

What Limits Current Performance?

- Replace each part with humans (Amazon Turk):



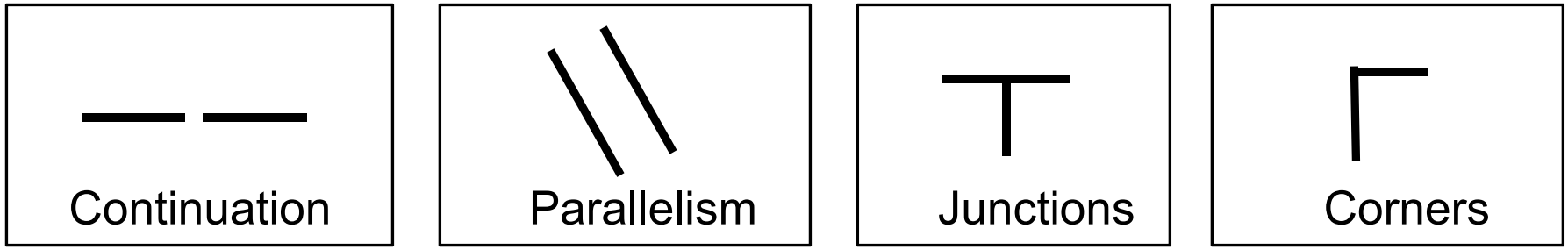
Parikh & Zitnick, CVPR'10

- Also removal of part deformations has small (<2%) effect.
  - Are "Deformable Parts" necessary in the Deformable Parts Model? Divvala, Hebert, Efros, ECCV 2012

Slide: R. Fergus

Mid-Level Representations

- Mid-level cues

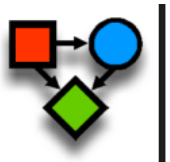


Tokens" from Vision by D.Marr:

- Object parts:

- Difficult to hand-engineer → What about learning them?

Slide: R. Fergus



Learning Feature Hierarchy

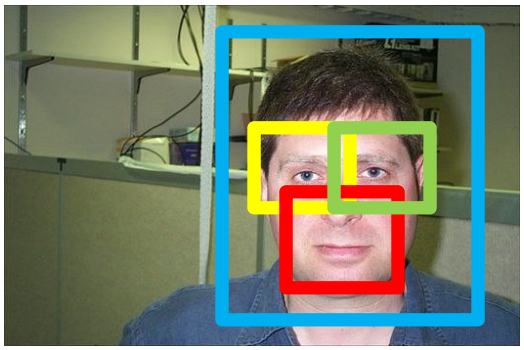
- Learn hierarchy
- All the way from pixels → classifier
- One layer extracts features from output of previous layer

- Train all layers jointly

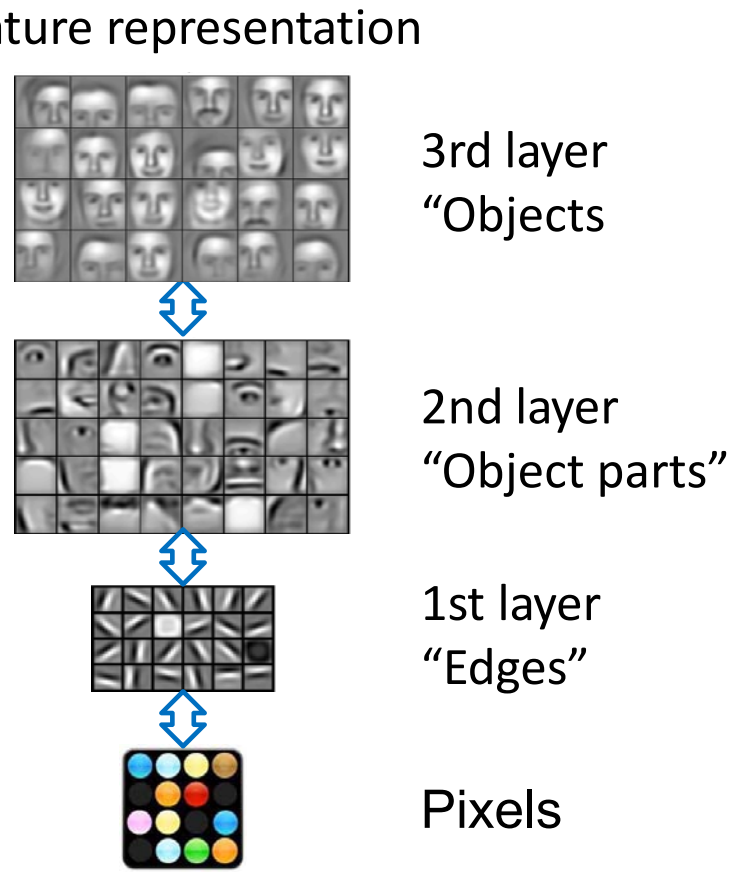
Slide: R. Fergus

Learning Feature Hierarchy

- Learn **useful higher-level features** from images



Lee et al., ICML 2009;
CACM 2011

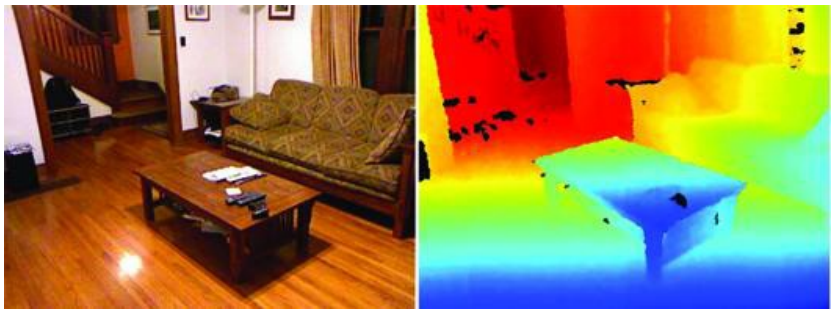


- **Fill in representation gap** in recognition

Learning Feature Hierarchy

- Better performance

- Other domains (unclear how to hand engineer):
  - Kinect
  - Video
  - Multi spectral



- Feature computation time
  - Dozens of features now regularly used [e.g., MKL]
  - Getting prohibitive for large datasets (10's sec /image)

Slide: R. Fergus

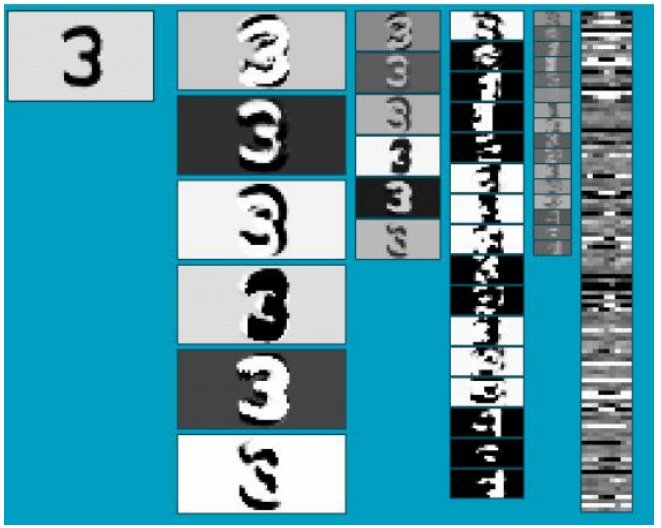
Approaches to Learning Features

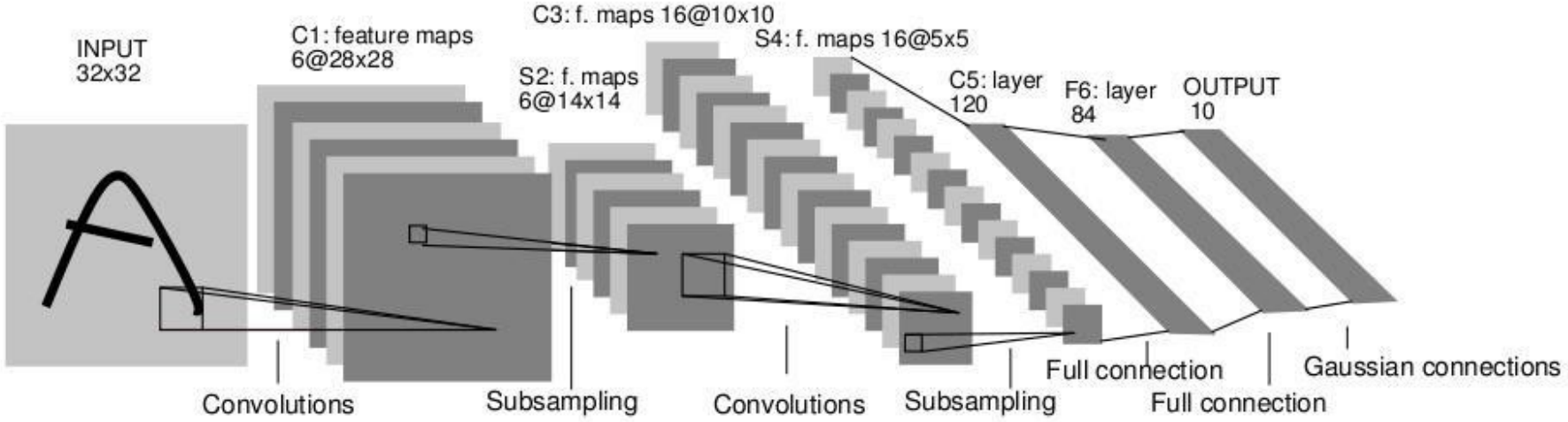
- Supervised Learning
  - End-to-end learning of deep architectures (e.g., deep neural networks) with back-propagation
  - Works well when the amounts of labels is large
  - Structure of the model is important (e.g. convolutional structure)

- Unsupervised Learning
  - Learn statistical structure or dependencies of the data from unlabeled data
  - Layer-wise training
  - Useful when the amount of labels is not large

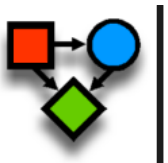
Convolutional Neural Networks

- LeCun et al. 1989
- Neural network with specialized connectivity structure



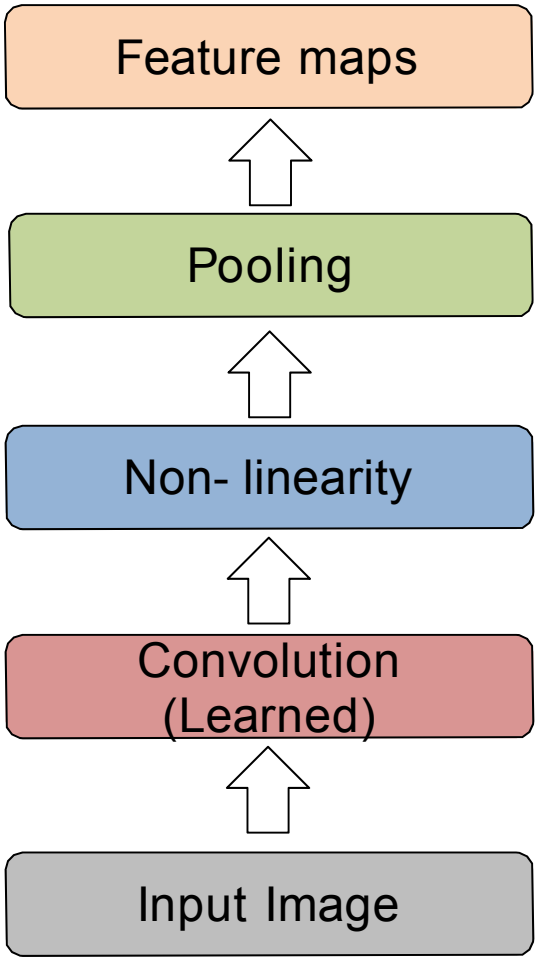


Slide: R. Fergus

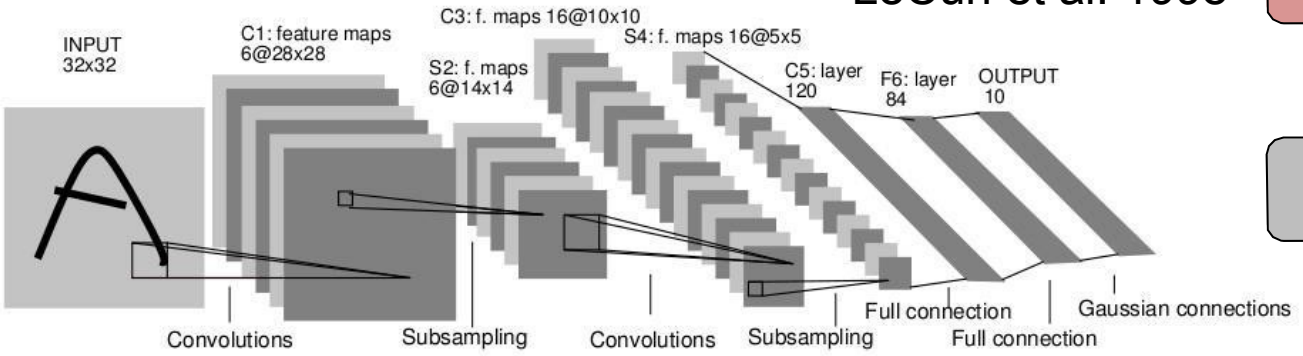


Convolutional Neural Networks

- Feed-forward:
  - Convolve input
  - Non-linearity (rectified linear)
  - Pooling (local max)
- Supervised
- Train convolutional filters by back-propagating classification error



LeCun et al. 1998

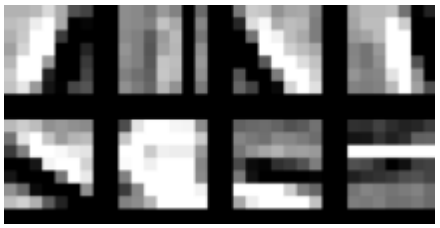


Slide: R. Fergus

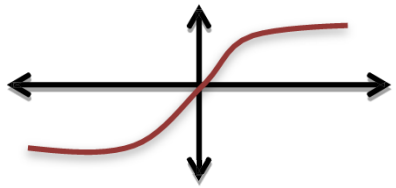
Components of Each Layer

Pixels / Features

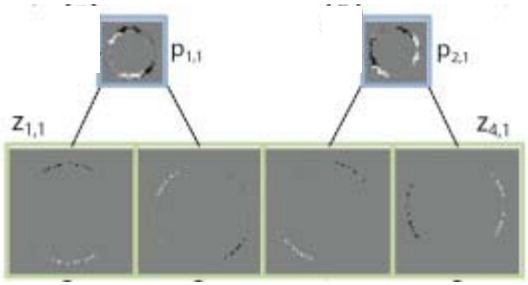
Filter with Dictionary (convolutional or tiled)



+ Non-linearity



Spatial/Feature (Sum or Max)



[Optional]

Normalization between feature responses

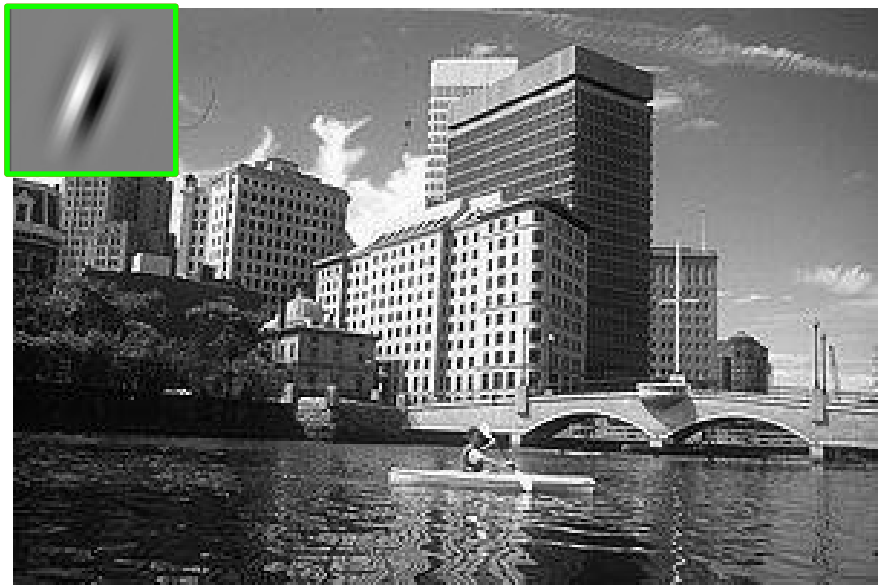
Output Features

Slide: R. Fergus

Filtering

Convolutional

- Dependencies are local
- Translation equivariance
- Tied filter weights (few params)
- Stride 1,2,… (faster, less mem.)



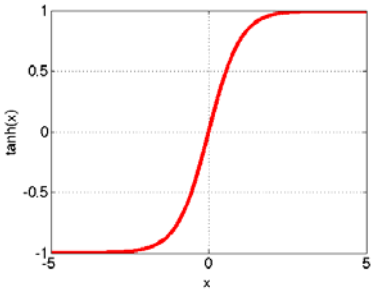
Input

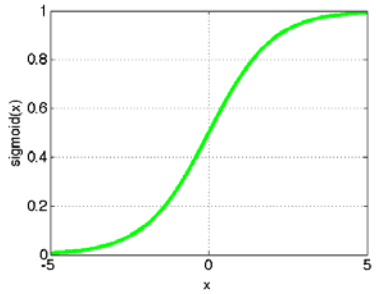
Feature Map

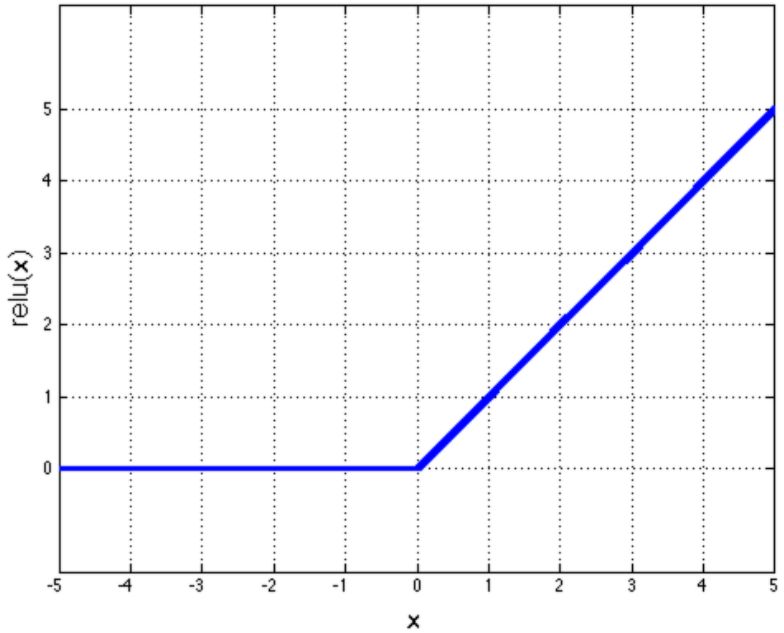
Slide: R. Fergus

Non-Linearity

- Non-linearity
  - Per-element (independent)
  - Tanh
  - Sigmoid: 1/(1+exp(-x))
  - Rectified linear
    - Simplifies backprop
    - Makes learning faster
    - Avoids saturation issues
    - Preferred option



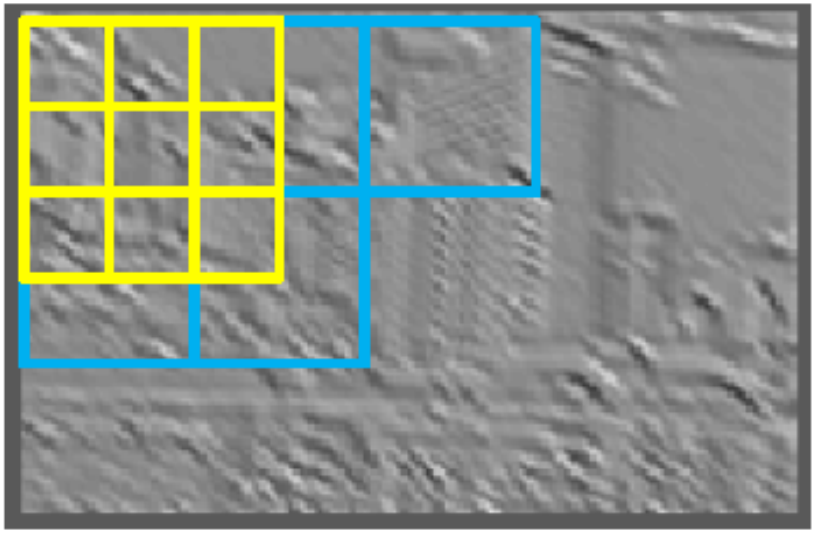


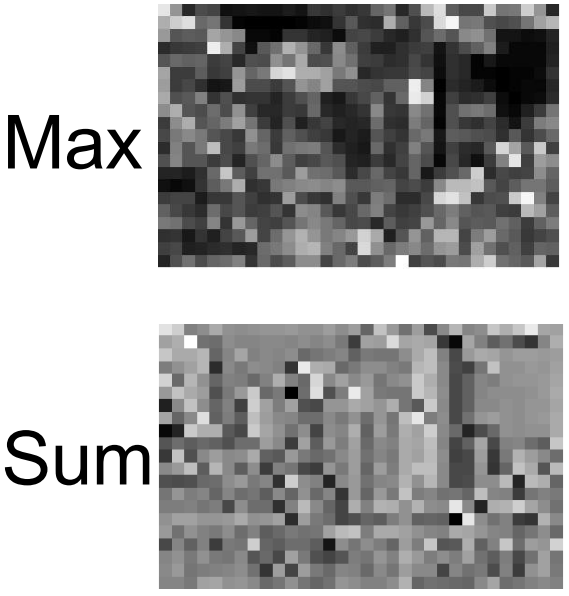


Slide: R. Fergus

Pooling

- Spatial Pooling
  - Non-overlapping / overlapping regions
  - Sum or max
  - Boureau et al. ICML'10 for theoretical analysis



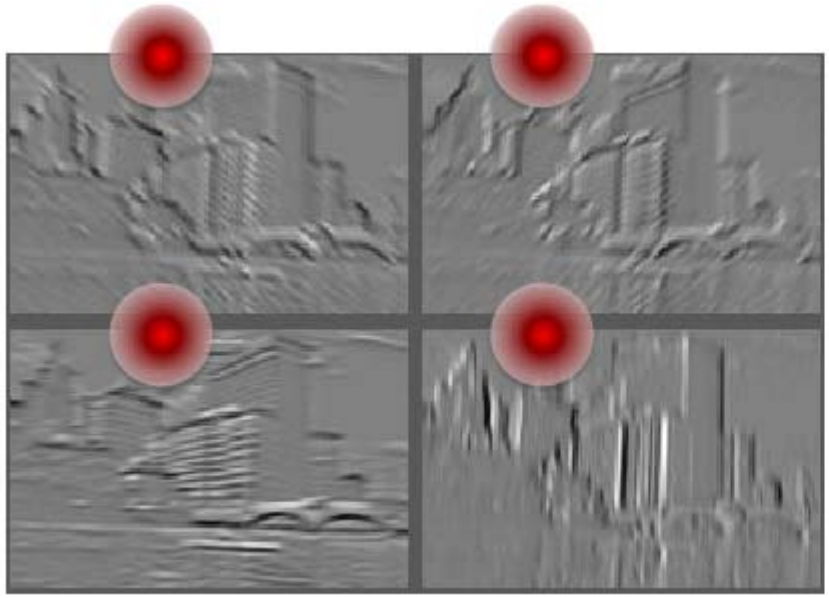


Slide: R. Fergus

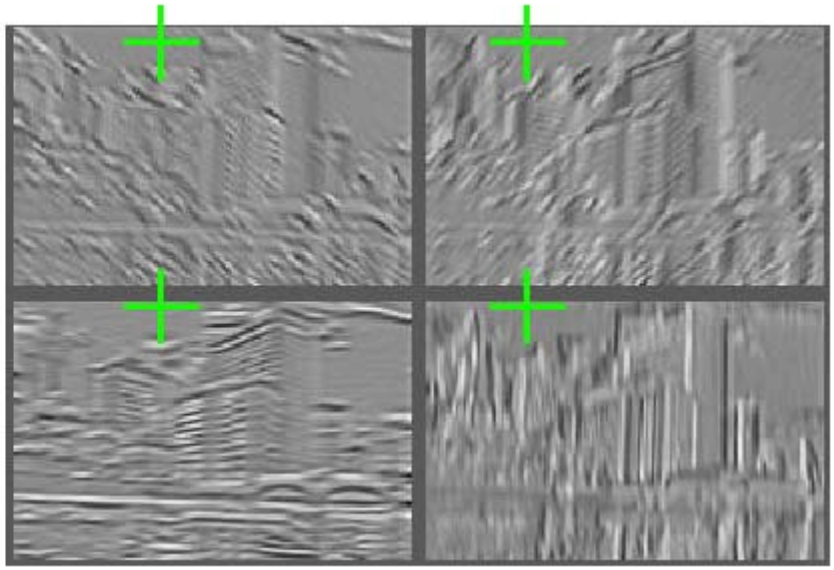
Normalization

- Contrast normalization (across feature maps)
  - Local mean = 0, local std. = 1, "Local" → 7x7 Gaussian
  - Equalizes the features maps

Feature Maps



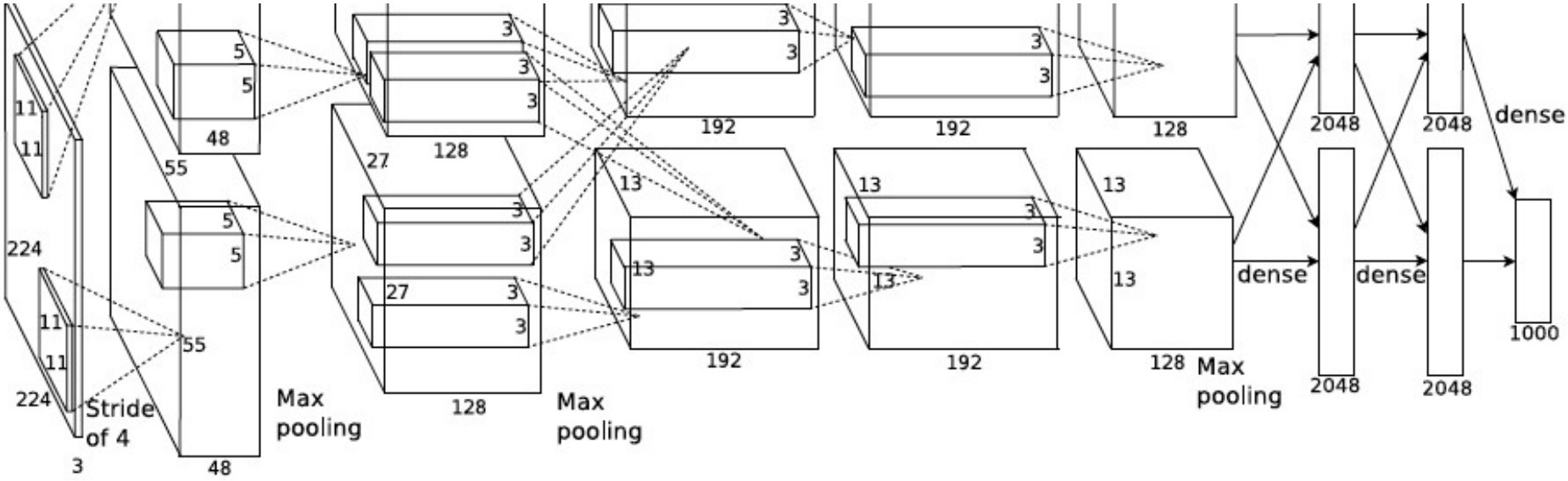
Feature Maps
After Contrast Normalization



Slide: R. Fergus

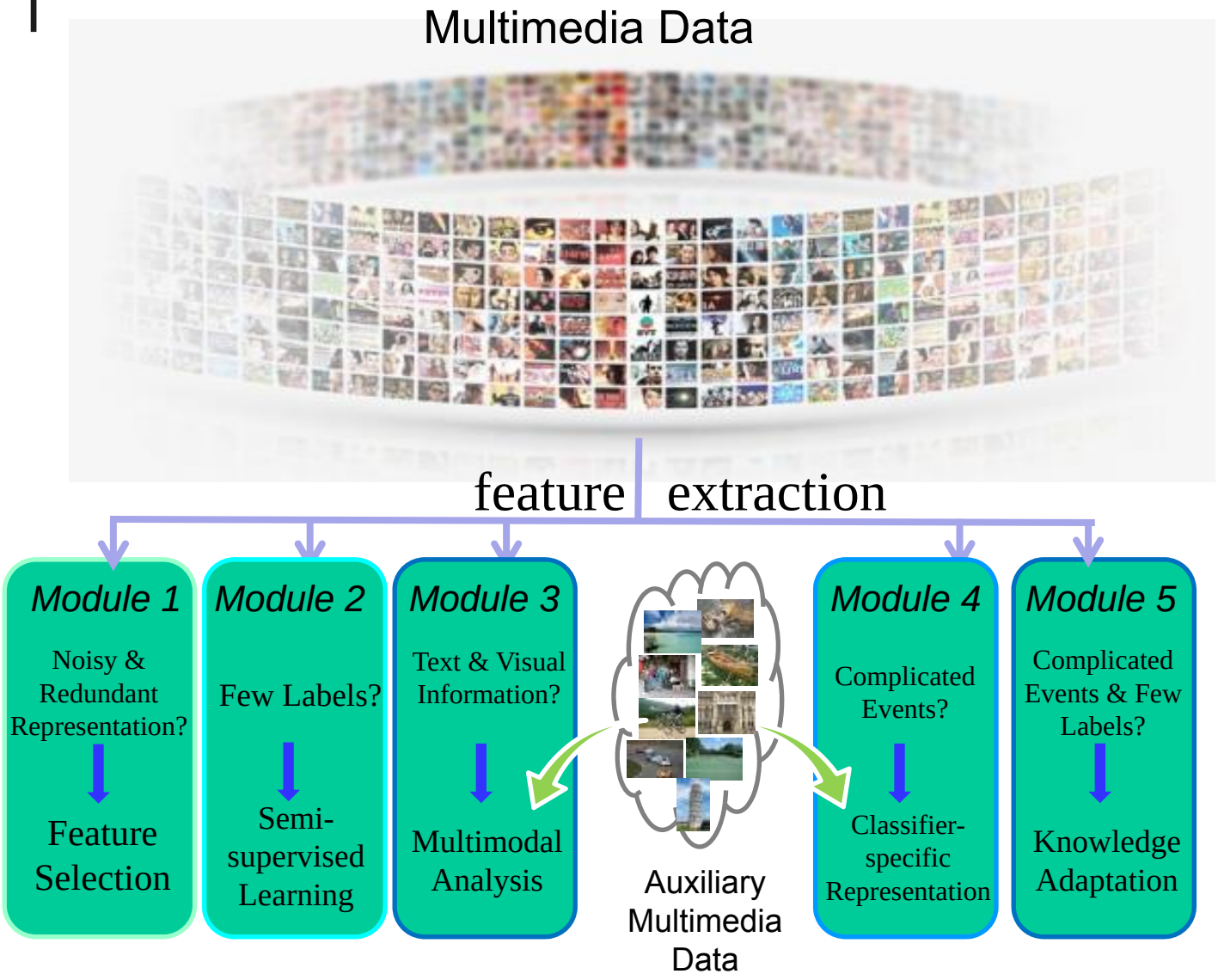
Krizhevsky et al. [NIPS 2012]

- Same model as LeCun'98 but:
  - Bigger model (8 layers)
  - More data (10^6 vs 10^3 images)
  - GPU implementation (50x speedup over CPU)
  - Better regularization (DropOut)



- 7 hidden layers, 650,000 neurons, 60,000,000 parameters
- Trained on 2 GPUs for a week

A Generic Framework



Outline

- Feature Selection (Q1)
- Semi-supervised Feature Analysis (Q2)
- Visual Info Helps Text Retrieval (Q3)
- Classifier-specific Representation (Q4)
- Knowledge Adaptation for MED (Q5)

Feature Selection

Noise?

Redundancy?

QUESTION 1:

Is it possible to get a compact feature representation? Would the accuracy be improved as a result?

Motivation

- Images are represented by various features
- Feature selection eliminates noise and redundancy
- Feature selection can improve both classification accuracy and computational efficiency
- Web images are usually multi-labeled

Z. Ma, F. Nie, Y. Yang, J. Uijlings and N. Sebe: “Web Image Annotation via Subspace-Sparsity Collaborated Feature Selection". IEEE Transactions on Multimedia, 14(4): 1021-1030, 2012.

Related Work

- Feature Selection
  - traditional approach: individual evaluation of features, e.g., Duda et al. [1]
    - Problem: low efficiency, does not consider feature correlation
  - sparse feature selection: joint evaluation, e.g., Yang et al. [2]
    - Problem: does not consider concept correlation

[1] R. O. Duda, P. E. Hart, and D. G. Stork. Pattern Classification (2nd ed.). Wiley-Interscience, New York, USA, 2001.
[2] Y. Yang, H. T. Shen, Z. Ma, Z. Huang, and X. Zhou. L21-norm regularized discriminative feature selection for unsupervised learning. In IJCAI, 2011.

Methodology

- Sparse model
- Shared subspace learning

Advantages:

- Batch-mode: evaluates features jointly across all data points
- considers the correlation between different concept labels

Formulation

- Feature Selection
  - Joint feature selection with sparsity
  - $l_{2,1}$- norm regularized model [3]

$$\min_W \text{loss}(W) + \gamma \|W\|_{2,1}$$

$$\|W\|_{2,1} = \sum_{i=1}^{d} \sqrt{\sum_{j=1}^{c} W_{ij}^2}$$

loss function

guarantees the optimized W to be sparse, i.e., some (many) of its rows shrink to zero. W can be viewed as the combination coefficients for the most discriminative features

[3] A. Argyriou, T. Evgeniou, and M. Pontil. Multi-task feature learning. In NIPS, 2007

Formulation (Cont'd)

- Shared subspace learning [4]

training data ground truth labels

$$\min_{W,V,P,Q} loss(W^T X, Y) + \mu\Omega(V,P)$$

$$s.t. \quad Q^T Q = I, \quad W = V + QP$$

shared subspace weights

assumes multi-label images share common attributes, e.g.,. an image labeled "parade", "people" and "street" share "people" with another one labeled "party" and "people"

- Objective Function:

controls the complexity

$$\arg\min_{W,P,Q} \left\| X^T W - Y \right\|_{2,1} + \alpha \|W\|_{2,1} + \beta \|W - QP\|_F^2$$

$$s.t. \quad Q^T Q = I$$

regulates the information to each specific label

[4] R. Ando, and T. Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. JMLR, 2005.

Experiments

- Datasets:
  - MSRA-MM 2.0: web images, diverse, multi-labeled
  - NUS-WIDE: Flickr images, diverse, multi-labeled, large-scale

	MSRA-MM 2.0	NUS-WIDE
Class Number	100	81
Training Set Size	10,000	10,000
Testing Set Size	32,266	199,347

- Features: Color Correlogram
  Edge Direction Histogram
  Wavelet Texture

Experiments (Cont'd)

- Comparison algorithms:
  - Sub-Feature Uncovering with Sparsity (SFUS)
  - All Features
  - Fisher Score
  - Sparse Multinomial Logistic Regression via Bayesian L1 Regularisation (SBMLR), NIPS
  - Spectral feature selection (SPEC), ICML
  - Group Lasso with Logistic Regression (GLRR), ACM MM
  - Feature Selection via Joint l2,1 –Norms Minimization (FSNM), NIPS

Experiments (Cont'd)

- Results comparison – 10c (class number) training data

MAP/MicroAUC/MacroAUC ± Standard Deviation on MSRA dataset

All Features	Fisher Score	SBLMR	SPEC	FSNM	GLRR	SFUS
0.062±0.001	0.060±0.002	0.056±0.002	0.058±0.001	0.061±0.002	0.060±0.001	**0.063±0.001**
0.840±0.001	0.861±0.005	0.869±0.003	0.852±0.002	0.875±0.002	0.846±0.001	**0.878±0.002**
0.655±0.006	0.655±0.003	0.643±0.006	0.650±0.004	0.658±0.006	0.653±0.005	**0.662±0.005**

MAP/MicroAUC/MacroAUC ± Standard Deviation on NUS dataset

All Features	Fisher Score	SBLMR	SPEC	FSNM	GLRR	SFUS
0.081±0.002	0.080±0.002	0.072±0.008	0.078±0.002	0.092±0.001	0.082±0.002	**0.094±0.003**
0.842±0.003	0.851±0.003	0.871±0.005	0.847±0.003	0.869±0.002	0.853±0.002	**0.877±0.002**
0.726±0.003	0.728±0.004	0.718±0.028	0.722±0.003	0.753±0.002	0.732±0.003	**0.756±0.003**

Experiments (Cont'd)

- Results comparison – 20c (class number) training data

MAP/MicroAUC/MacroAUC ± Standard Deviation on MSRA dataset

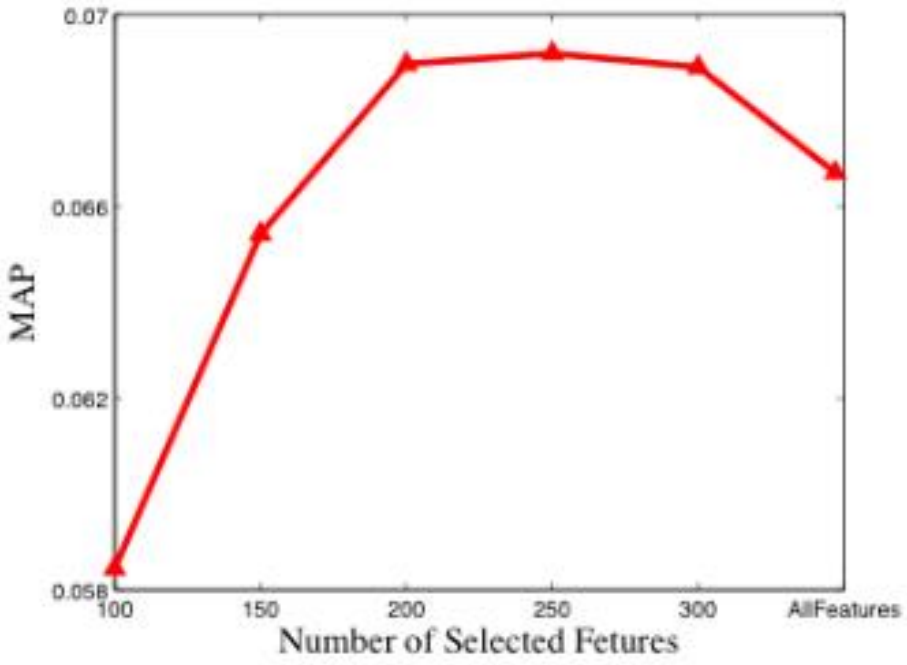
All Features	Fisher Score	SBLMR	SPEC	FSNM	GLRR	SFUS
0.067±0.004	0.066±0.002	0.059±0.001	0.066±0.001	0.068±0.001	0.067±0.001	**0.070±0.001**
0.859±0.011	0.876±0.004	0.883±0.004	0.868±0.001	0.888±0.002	0.866±0.002	**0.888±0.002**
0.676±0.013	0.680±0.004	0.666±0.004	0.679±0.002	0.687±0.002	0.680±0.002	**0.690±0.002**

MAP/MicroAUC/MacroAUC ± Standard Deviation on NUS dataset

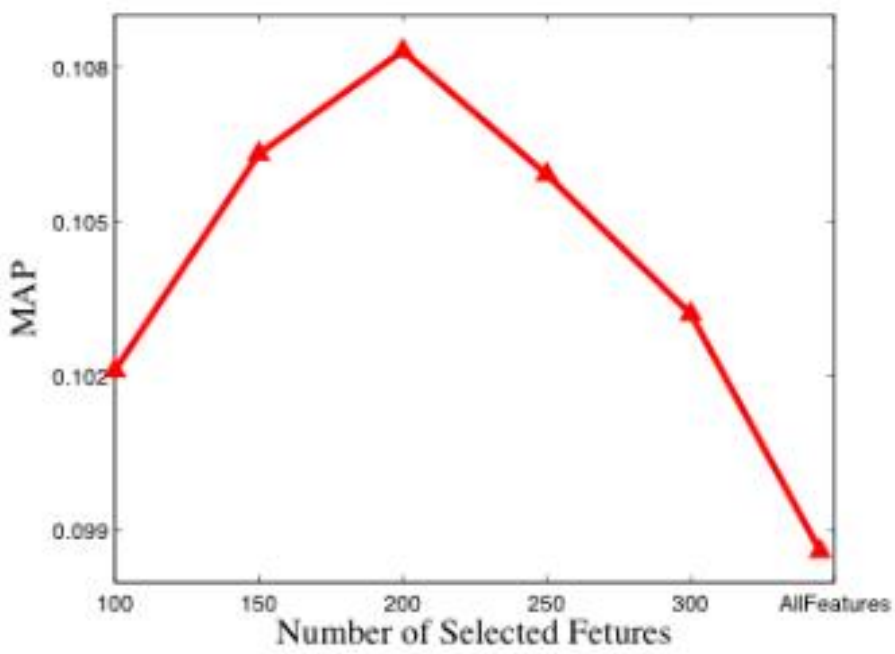
All Features	Fisher Score	SBLMR	SPEC	FSNM	GLRR	SFUS
0.099±0.001	0.098±0.004	0.073±0.007	0.094±0.001	0.105±0.003	0.105±0.002	**0.108±0.002**
0.874±0.001	0.880±0.005	0.887±0.006	0.875±0.001	0.888±0.003	0.885±0.003	**0.891±0.003**
0.767±0.001	0.770±0.005	0.733±0.024	0.763±0.001	0.785±0.004	0.780±0.001	**0.789±0.003**

Experiments (Cont'd)

- Influence of selected features



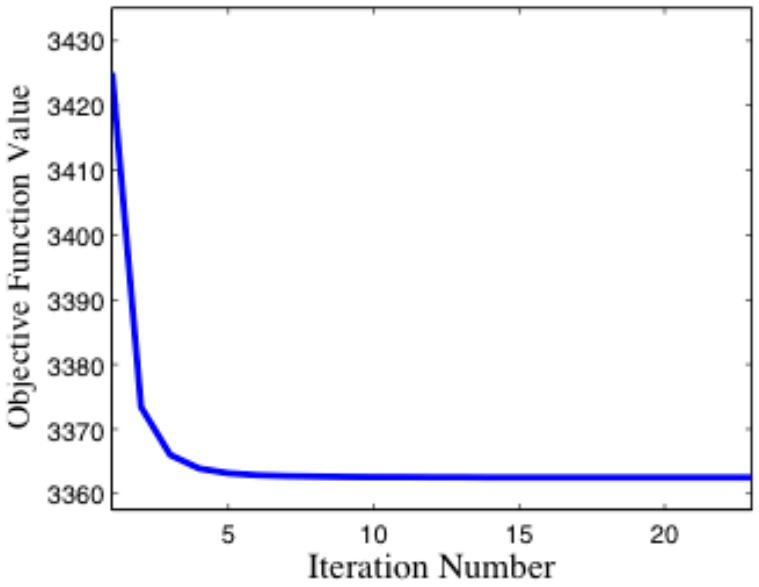
(a) MSRA-MM



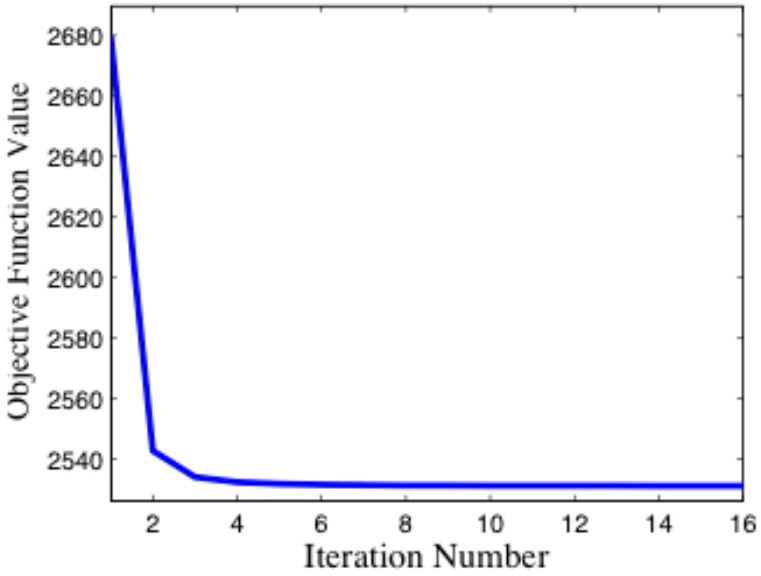
(b) NUS-WIDE

Experiments (Cont'd)

- Convergence



(a) MSRA-MM



(b) NUS-WIDE

Summary

- Integration of shared subspace learning and joint feature selection with sparsity
- Evaluating feature importance jointly
- Consideration of the correlation between labels
- Promising results on large-scale web image sets

Z. Ma, F. Nie, Y. Yang, J. Uijlings and N. Sebe: “Web Image Annotation via Subspace-Sparsity Collaborated Feature Selection". IEEE Transactions on Multimedia, 14(4): 1021-1030, 2012.

Outline

- Feature Selection (Q1)
- Semi-supervised Feature Analysis (Q2)
- Visual Info Helps Text Retrieval (Q3)
- Classifier-specific Representation (Q4)
- Knowledge Adaptation for MED (Q5)

Semi-supervised Learning

QUESTION 2:

Is there any way to attain a reasonable performance when only few labeled images and videos are available?

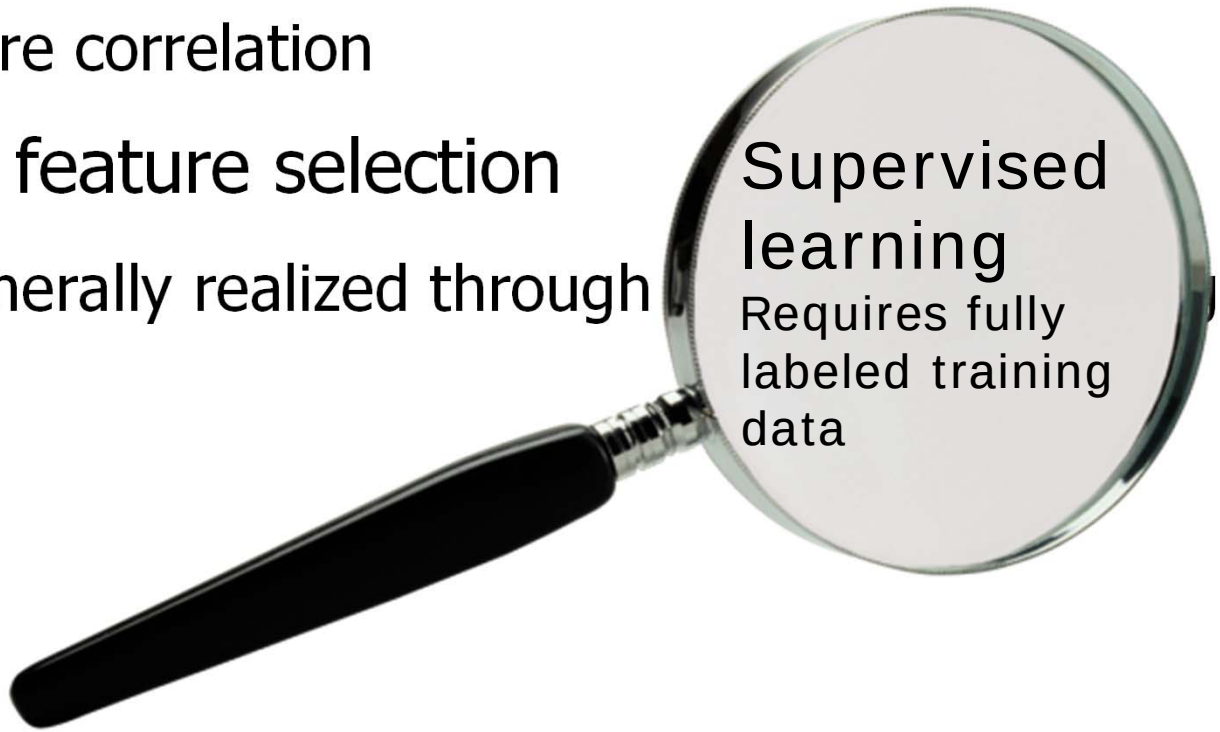
Motivation

- Multimedia data are represented by various features
- For classification purpose, some noisy and irrelevant features may be not useful
- Semi-supervised learning uses the limited available labels in an effective way
- It is natural to integrate semi-supervised learning with feature selection

Z. Ma, F. Nie, Y. Yang, J. Uijlings, N. Sebe and A. G. Hauptmann: “Discriminating Joint Feature Analysis for Multimedia Content Understanding". IEEE Transactions on Multimedia, 14(6): 1662-1672, 2012.

Problems

- Semi-supervised feature selection
  - traditional method: low efficiency, does not consider feature correlation
- Sparse feature selection
  - is generally realized through



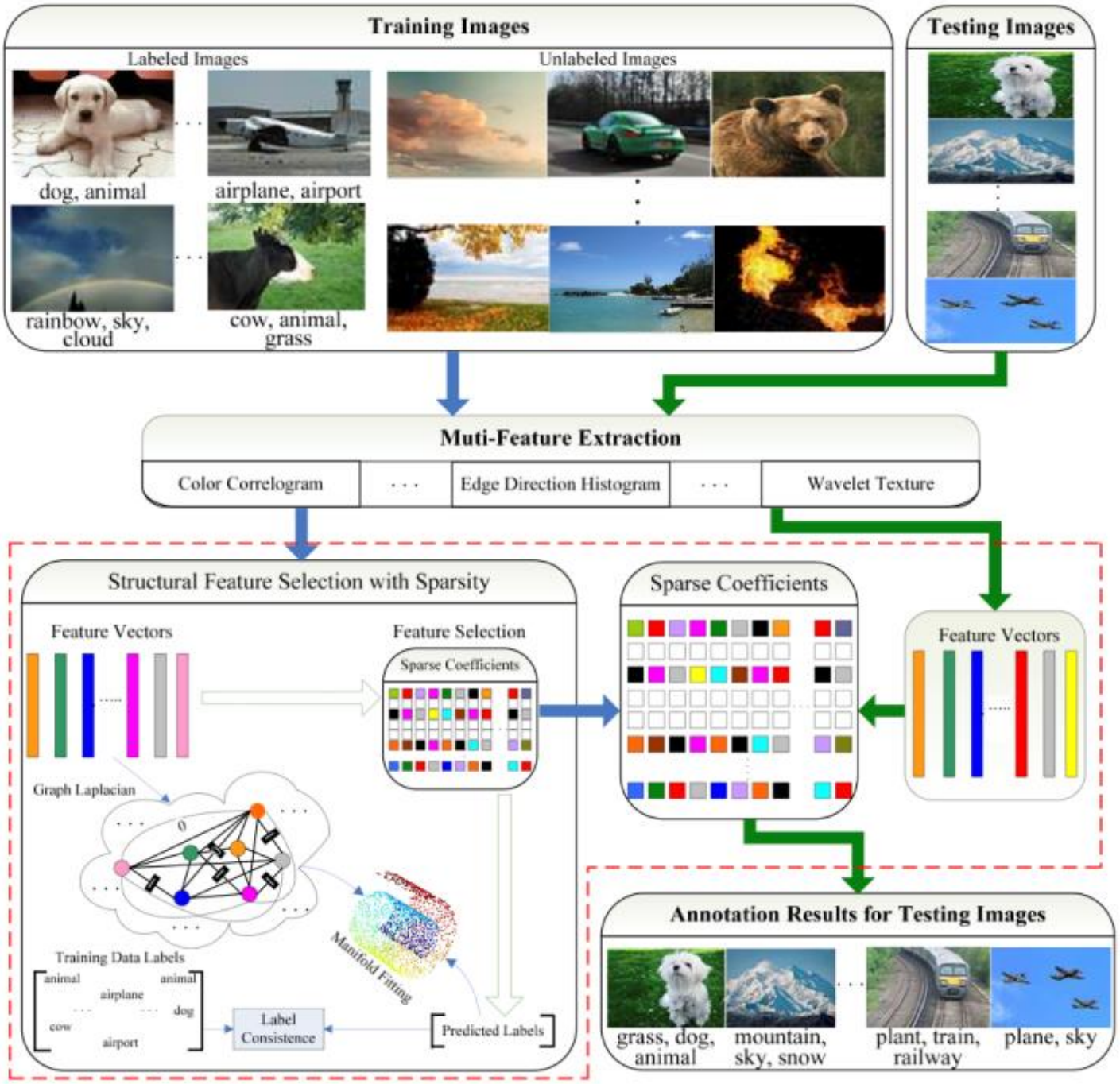
Methodology

- Efficient feature analysis
  - Sparse feature selection
  - Semi-supervised via graph Laplacian

Advantages:

- Batch-mode: evaluating features jointly across all data points
- Semi-supervised: not so expensive as supervised learning

FRAMEWORK


Training Images
Labeled Images
Unlabeled Images
dog, animal
airplane, airport
rainbow, sky, cloud
cow, animal, grass
Testing Images
Muti-Feature Extraction
Color Correlogram
Edge Direction Histogram
Wavelet Texture
Structural Feature Selection with Sparsity
Feature Vectors
Feature Selection
Sparse Coefficients
Graph Laplacian
Training Data Labels
animal
airplane
animal
dog
cow
airport
Label Consistence
Manifold Fitting
Predicted Labels
Sparse Coefficients
Feature Vectors
Annotation Results for Testing Images
grass, dog, animal
mountain, sky, snow
plant, train, railway
plane, sky

Formulation

- Feature Selection
  - Joint feature selection with sparsity
  - $l_{2,1}$- norm regularized model [3]

$$\min_{W} \text{loss}(W) + \gamma \|W\|_{2,1}$$

- Incorporate semi-supervised learning
  - Use graph Laplacian

Formulation

- Semi-supervised learning
  - Graph Laplacian construction: L = D – A

$$D_{ii} = \sum_{j=1}^{n} A_{ij} \qquad A_{ij} = \begin{cases} 1 & x_i \text{ and } x_j \text{ are } k \text{ nearest neighbors;} \\ 0 & \text{otherwise.} \end{cases}$$

- Objective function: use manifold regularization

$$\arg\min_{W,b} \mathrm{Tr}\left(W^{\mathrm{T}} X L X^{\mathrm{T}} W\right) + \mu \left\| X_l^{\mathrm{T}} W + 1_n b^{\mathrm{T}} - Y_l \right\|_{\mathrm{F}}^2 + \gamma \|W\|_{2,1}$$

labeled training data bias term ground-truth labels

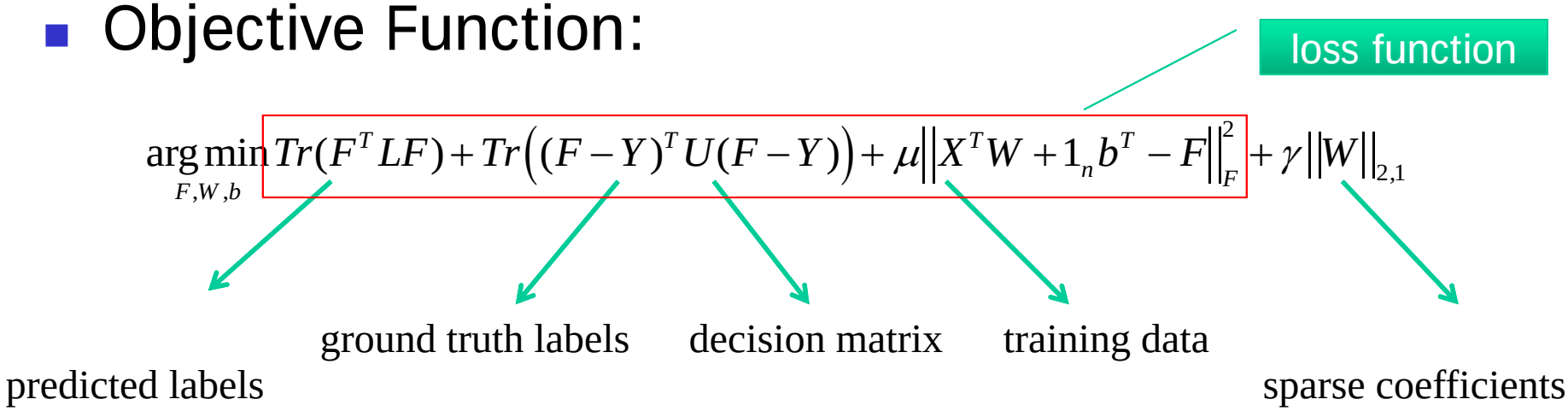
- Define a predicted label matrix F for all training data
  - smooth on Y_l and the manifold structure

$$\arg\min_{F} \mathrm{Tr}\left(F^{\mathrm{T}} L F\right) + \mathrm{Tr}\left((F - Y)^{\mathrm{T}} U (F - Y)\right)$$

$U_{ii} = \infty$ if x_i is labeled and $U_{ii} = 1$ otherwise

Formulation

- Objective Function:



  - We are able to get F, W, and b simultaneously
  - The optimal W obtained can be utilized directly for classification as W does feature selection

Experiments

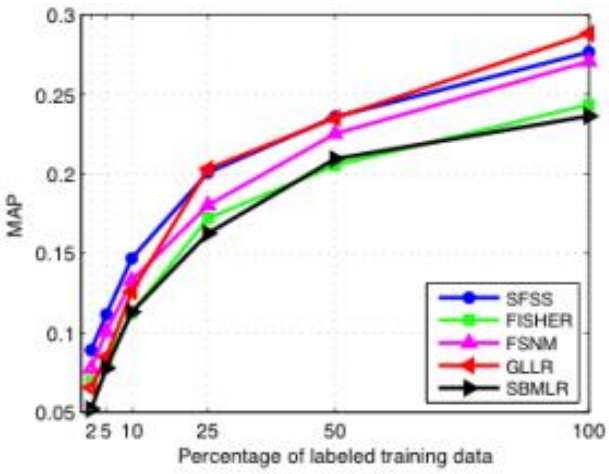
- Datasets
  - Image annotation:
    - Corel-5K: 50 classes, 5000 images
    - MSRA-MM 2.0: 81 classes, 42266 images
    - NUS-WIDE: 100 classes, 209347 images
  - Video concept recognition:
    - Kodak: 22 concepts, 3590 video frames
    - CareMedia: 5 concepts, 3913 video sequences
    - 3D motion data analysis
    - HumanEva: 10 classes, 10000 frames

Experiments (Cont'd)

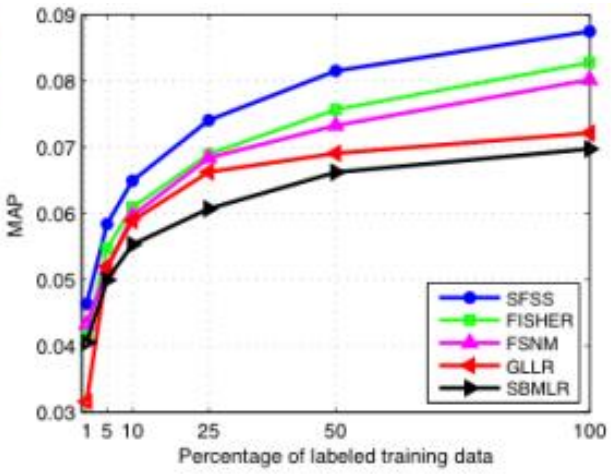
- Comparison algorithms:
  - Structural Feature Selection with Sparsity (SFSS)
  - Fisher Score (FISHER)
  - Sparse Multinomial Logistic Regression via Bayesian L1 Regularisation (SBMLR), NIPS
  - Group Lasso with Logistic Regression (GLRR), ACM MM
  - Feature Selection via Joint l2,1 –Norms Minimization (FSNM), NIPS
  - Semi-supervised Feature Selection via Spectral Analysis (sSelect), ICDM
  - Locality sensitive semi-supervised feature selection (LSDF), Nerocomputing

Experiments (Cont'd)

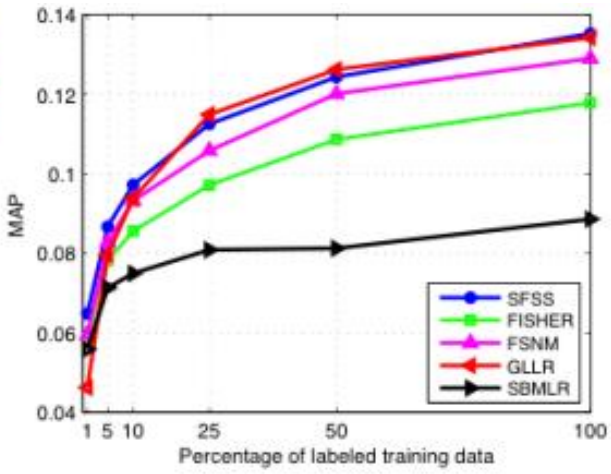
- Comparison on image annotation



(a) Corel-5K



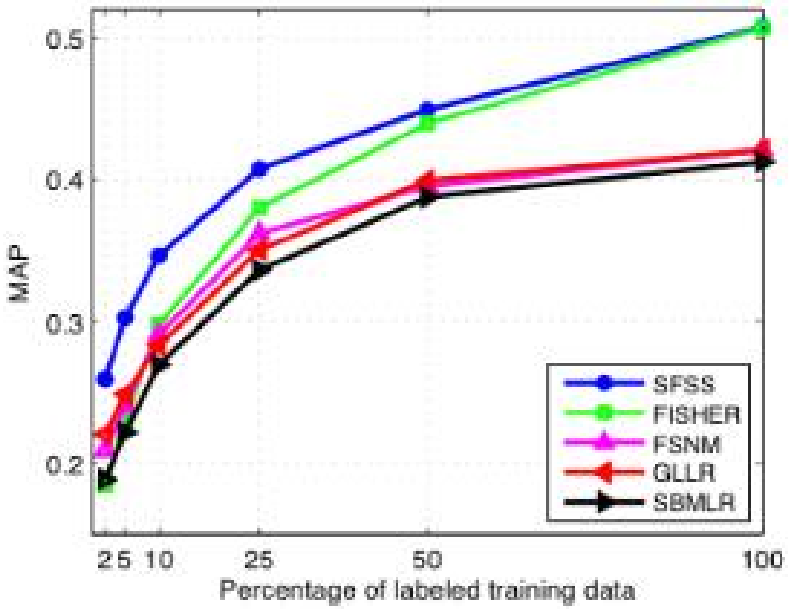
(b) MSRA-MM



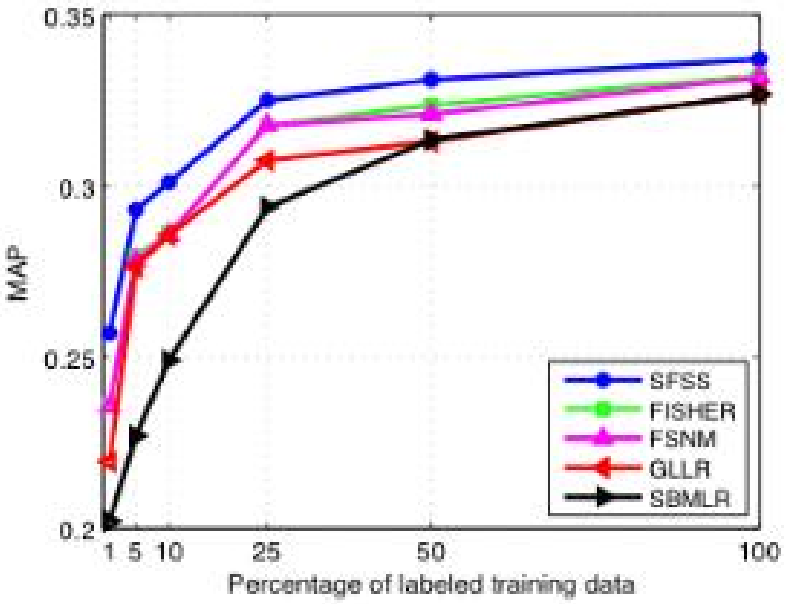
(c) NUS-WIDE

Experiments (Cont'd)

- Comparison on video concept recognition



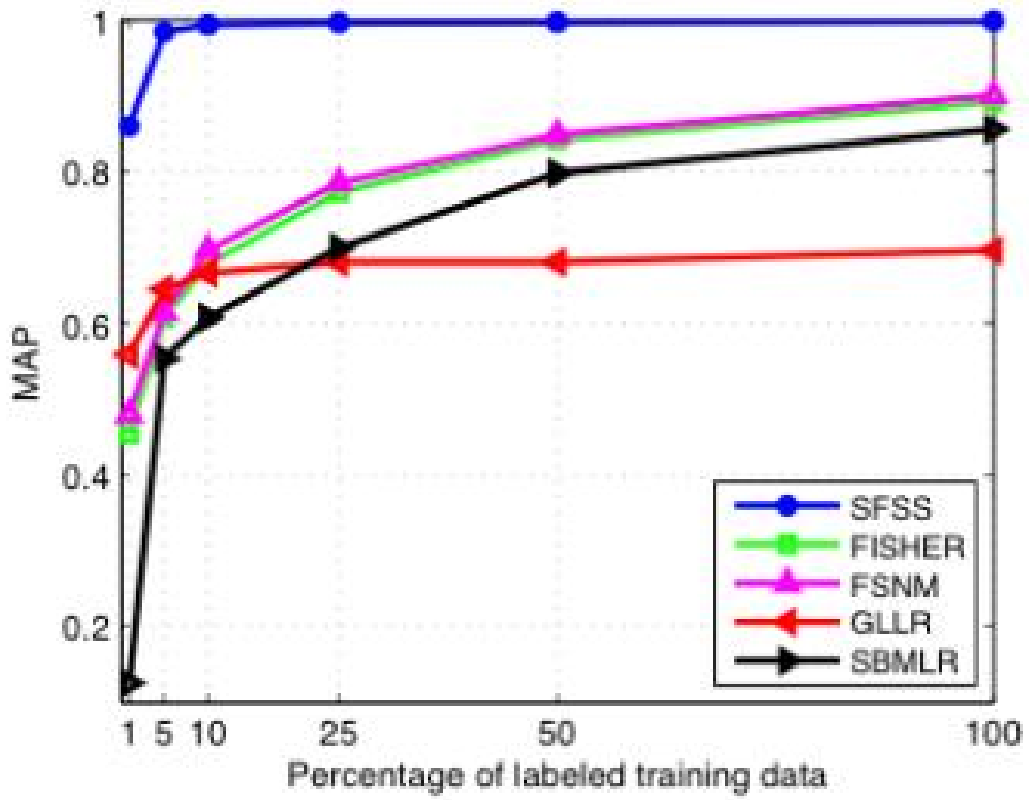
(a) Kodak



(b) CareMedia

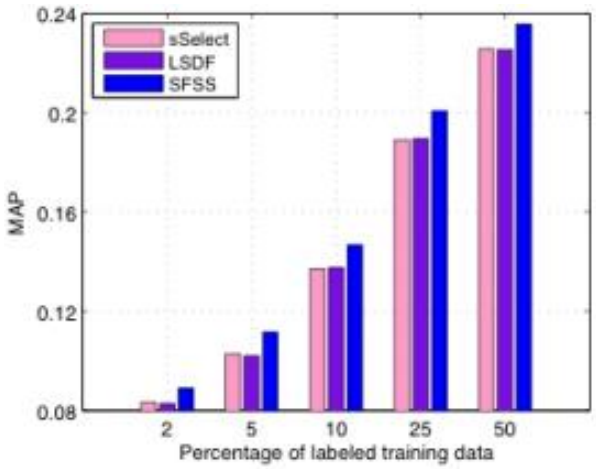
Experiments (Cont'd)

- Comparison on 3D motion data analysis

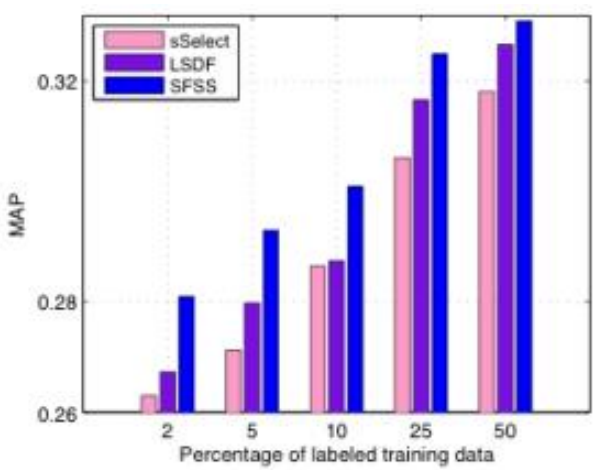


Experiments (Cont'd)

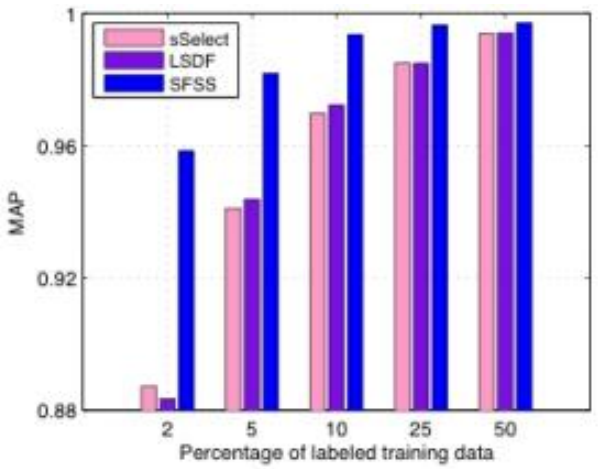
- Comparison with semi-supervised algorithms



(a) Corel-5K



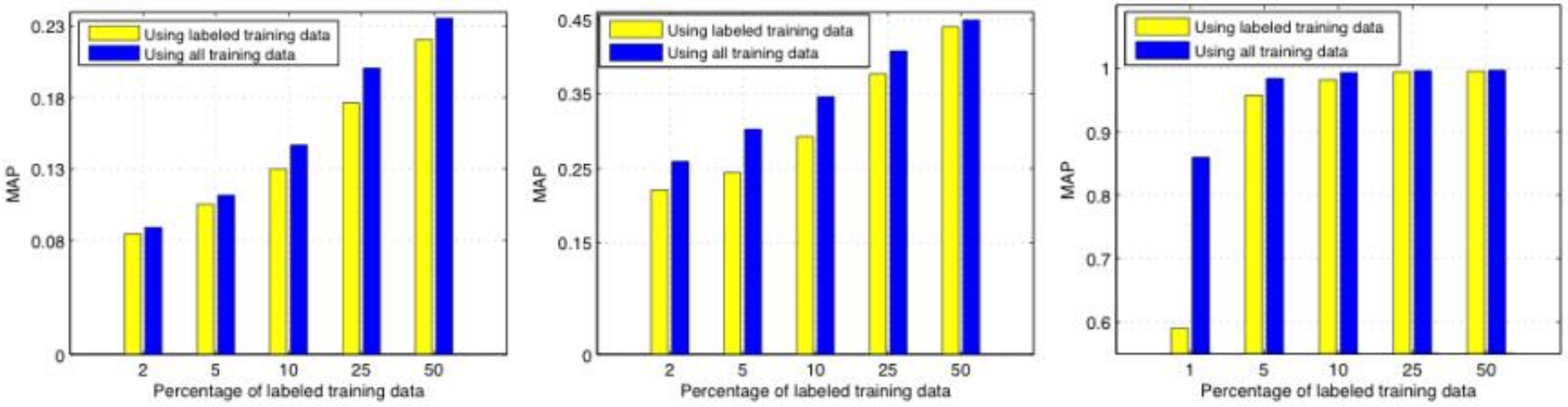
(b) CareMedia



(c) HumanEva

Experiments (Cont'd)

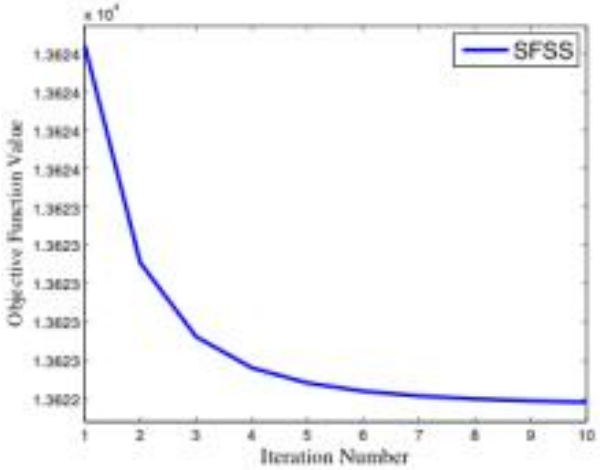
- Influence of unlabeled data



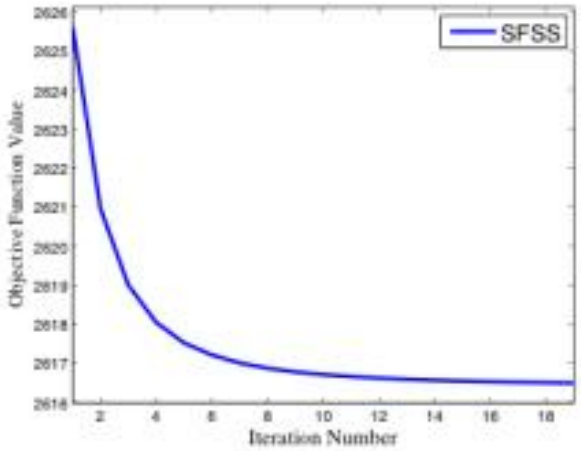
(a) Corel-5K (b) Kodak (c) HumanEva

Experiments (Cont'd)

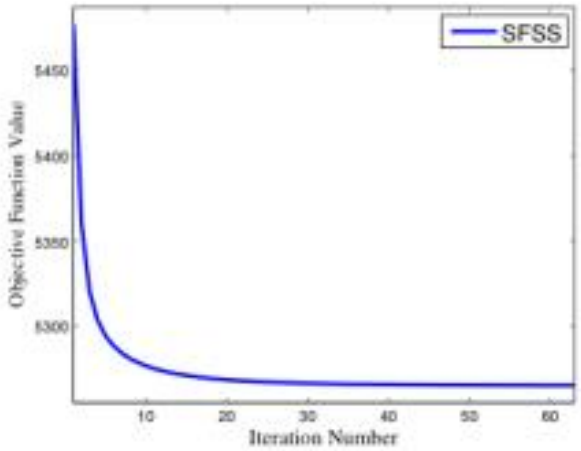
- Convergence



(a) Corel-5K



(b) Kodak



(c) HumanEva

Summary

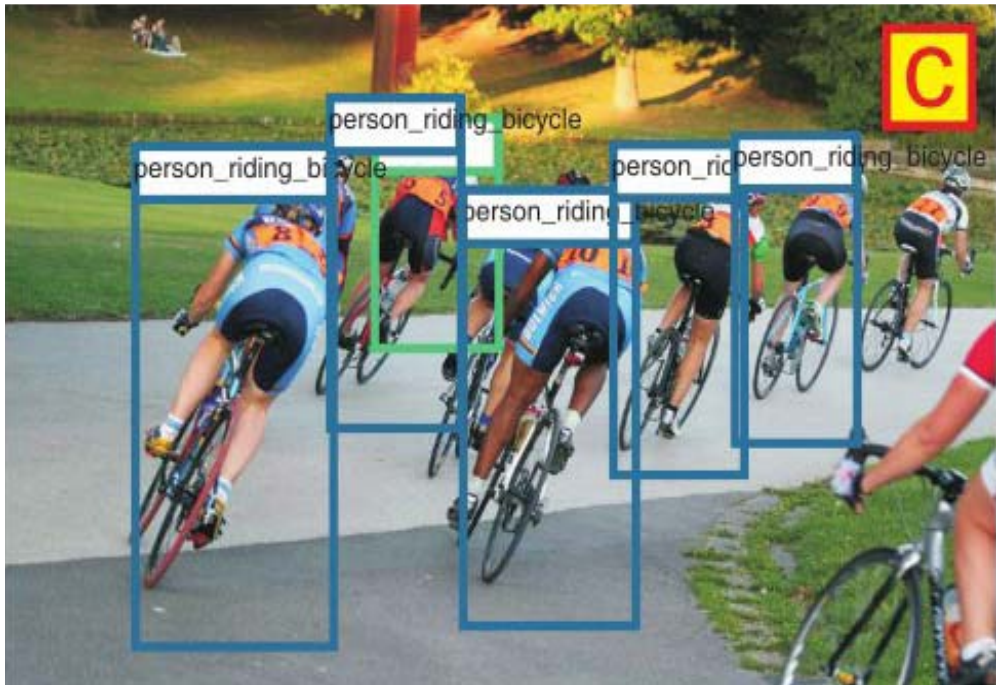
- Harnessing discriminating features closely related to the concept labels
- Cost saving
- Boosting performance with the usage of unlabeled data
- Analysis of multimedia data structure helps multimedia content understanding
- Clear advantages when few training data are labeled
- Applicable to a variety of applications

Z. Ma, F. Nie, Y. Yang, J. Uijlings, N. Sebe and A. G. Hauptmann: “Discriminating Joint Feature Analysis for Multimedia Content Understanding". IEEE Transactions on Multimedia, 14(6): 1662-1672, 2012.

Outline

- Feature Selection (Q1)
- Semi-supervised Feature Analysis (Q2)
- Visual Info Helps Text Retrieval (Q3)
- Classifier-specific Representation (Q4)
- Knowledge Adaptation for MED (Q5)

Multimodal Analysis



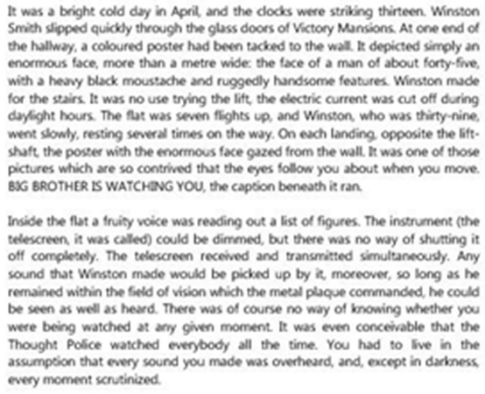
QUESTION 3:

Can we use other modalities (e.g., text) to improve the analysis? Can visual information help text retrieval?

Motivation

- Text-analysis used for vision

[Barnard ICCV 2001]
[Berg ECCV 2010]



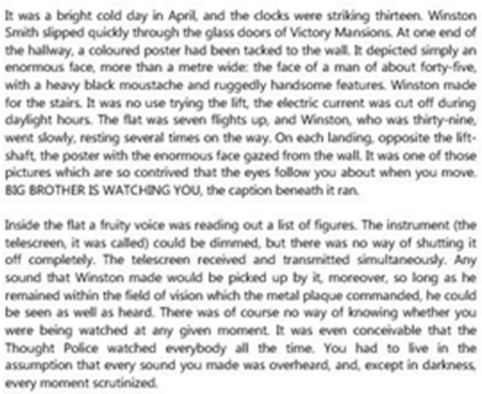
It was a bright cold day in April, and the clocks were striking thirteen. Winston Smith slipped quickly through the glass doors of Victory Mansions. At one end of the hallway, a coloured poster had been tacked to the wall. It depicted simply an enormous face, more than a metre wide: the face of a man of about forty-five, with a heavy black moustache and ruggedly handsome features. Winston made for the stairs. It was no use trying the lift, the electric current was cut off during daylight hours. The flat was seven flights up, and Winston, who was thirty-nine, went slowly, resting several times on the way. On each landing, opposite the lift-shaft, the poster with the enormous face gazed from the wall. It was one of those pictures which are so contrived that the eyes follow you about when you move. BIG BROTHER IS WATCHING YOU, the caption beneath it ran.

Inside the flat a fruity voice was reading out a list of figures. The instrument (the telescreen, it was called) could be dimmed, but there was no way of shutting it off completely. The telescreen received and transmitted simultaneously. Any sound that Winston made would be picked up by it, moreover, so long as he remained within the field of vision which the metal plaque commanded, he could be seen as well as heard. There was of course no way of knowing whether you were being watched at any given moment. It was even conceivable that the Thought Police watched everybody all the time. You had to live in the assumption that every sound you made was overheard, and, except in darkness, every moment scrutinized.

E. Bruni, J. Uijlings, M. Baroni, N. Sebe, Distributional semantics with eyes: Using image analysis to improve computational representations of word meaning. ACM Multimedia 2012

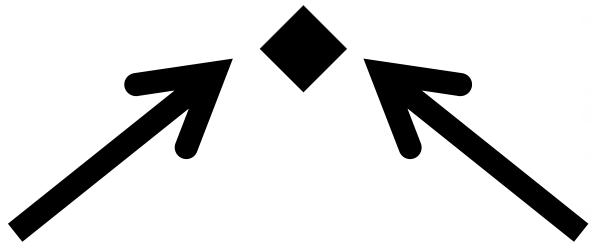
Motivation

- Text-analysis used for vision
- Multimodal analysis



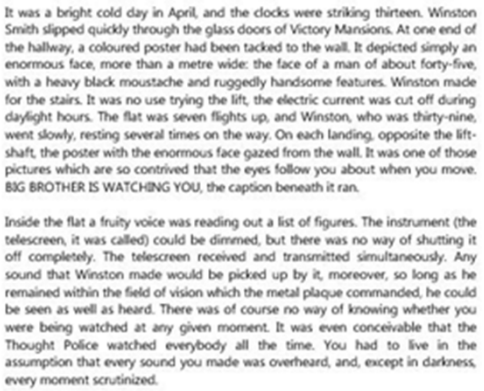
It was a bright cold day in April, and the clocks were striking thirteen. Winston Smith slipped quickly through the glass doors of Victory Mansions. At one end of the hallway, a coloured poster had been tacked to the wall. It depicted simply an enormous face, more than a metre wide: the face of a man of about forty-five, with a heavy black moustache and ruggedly handsome features. Winston made for the stairs. It was no use trying the lift, the electric current was cut off during daylight hours. The flat was seven flights up, and Winston, who was thirty-nine, went slowly, resting several times on the way. On each landing, opposite the lift-shaft, the poster with the enormous face gazed from the wall. It was one of those pictures which are so contrived that the eyes follow you about when you move. BIG BROTHER IS WATCHING YOU, the caption beneath it ran.

Inside the flat a fruity voice was reading out a list of figures. The instrument (the telescreen, it was called) could be dimmed, but there was no way of shutting it off completely. The telescreen received and transmitted simultaneously. Any sound that Winston made would be picked up by it, moreover, so long as he remained within the field of vision which the metal plaque commanded, he could be seen as well as heard. There was of course no way of knowing whether you were being watched at any given moment. It was even conceivable that the Thought Police watched everybody all the time. You had to live in the assumption that every sound you made was overheard, and, except in darkness, every moment scrutinized.



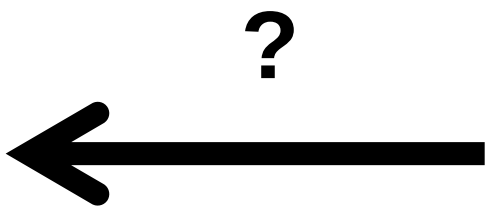
Motivation

- Text-analysis used for vision
- Multimodal analysis
- Vision-analysis used for text



It was a bright cold day in April, and the clocks were striking thirteen. Winston Smith slipped quickly through the glass doors of Victory Mansions. At one end of the hallway, a coloured poster had been tacked to the wall. It depicted simply an enormous face, more than a metre wide: the face of a man of about forty-five, with a heavy black moustache and ruggedly handsome features. Winston made for the stairs. It was no use trying the lift, the electric current was cut off during daylight hours. The flat was seven flights up, and Winston, who was thirty-nine, went slowly, resting several times on the way. On each landing, opposite the lift-shaft, the poster with the enormous face gazed from the wall. It was one of those pictures which are so contrived that the eyes follow you about when you move. BIG BROTHER IS WATCHING YOU, the caption beneath it ran.

Inside the flat a fruity voice was reading out a list of figures. The instrument (the telescreen, it was called) could be dimmed, but there was no way of shutting it off completely. The telescreen received and transmitted simultaneously. Any sound that Winston made would be picked up by it, moreover, so long as he remained within the field of vision which the metal plaque commanded, he could be seen as well as heard. There was of course no way of knowing whether you were being watched at any given moment. It was even conceivable that the Thought Police watched everybody all the time. You had to live in the assumption that every sound you made was overheard, and, except in darkness, every moment scrutinized.



this research

Distributional Semantics

- What is the semantic relatedness between two words?
- Applications:
  - Query expansion
  - Textual advertising
  - Information extraction
  - Word sense disambiguation

Distributional Semantics

- Distributional Hypothesis:
  Word-meaning can be derived from context
  [Harris, Charles and Miller, Firth, Wittgenstein, ...]

He filled the wampimuk, passed it around and we all drunk some

We found a little, hairy wampimuk sleeping behind the tree

this research

Distributional Semantics

- Distributional Hypothesis:
  Word-meaning can be derived from context
  [Harris, Charles and Miller, Firth, Wittgenstein, ...]

Few people write that bananas are yellow

Research Questions

- Text vs images: Which semantics are captured?
- Do images improve upon text-only semantics?
- How does the distributional hypothesis work for images?

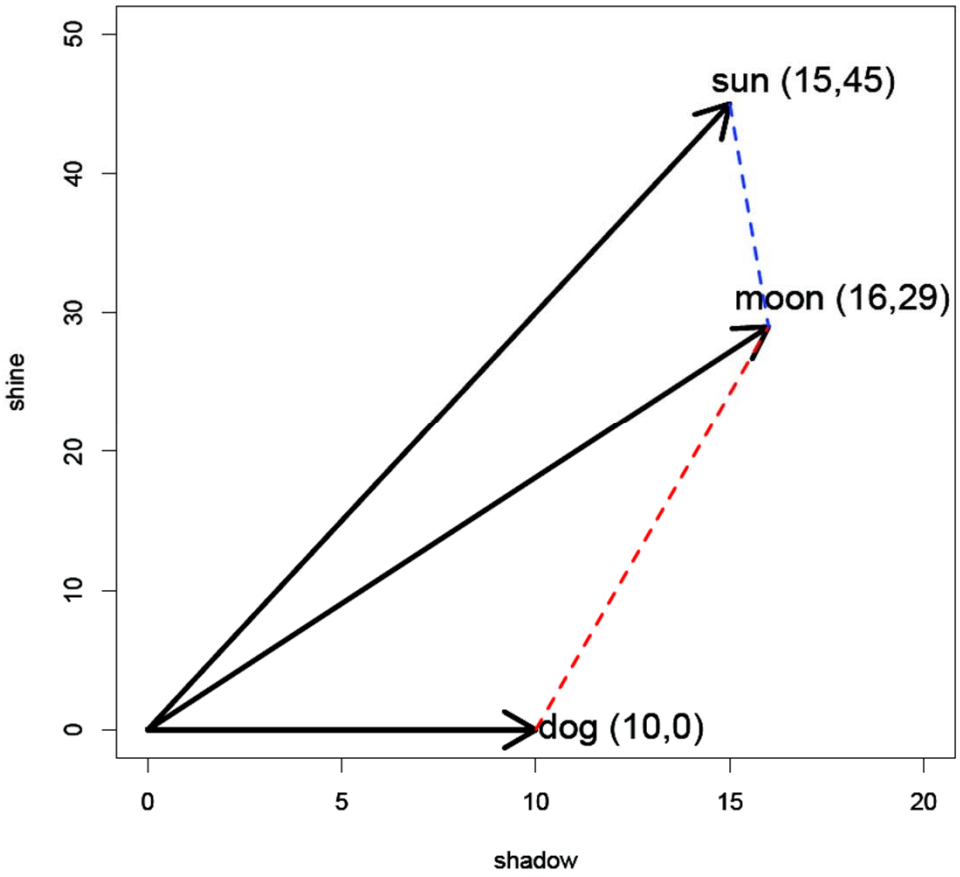
Distributional semantics from text

he curtains open and the moon shining in on the barely
ars and the cold , close moon " . And neither of the w
rough the night with the moon shining so brightly , it
made in the light of the moon . It all boils down , wr
surely under a crescent moon , thrilled by ice-white
sun , the seasons of the moon ? Home , alone , Jay pla
m is dazzling snow , the moon has risen full and cold
un and the temple of the moon , driving out of the hug
in the dark and now the moon rises , full and amber a
bird on the shape of the moon over the trees in front
But I could n't see the moon or the stars , only the
rning , with a sliver of moon hanging among the stars
they love the sun , the moon and the stars . None of
the light of an enormous moon . The plash of flowing w
man 's first step on the moon ; various exhibits , aer
the inevitable piece of moon rock . Housing The Airsh
oud obscured part of the moon . The Allied guns behind

Distributional semantics from text

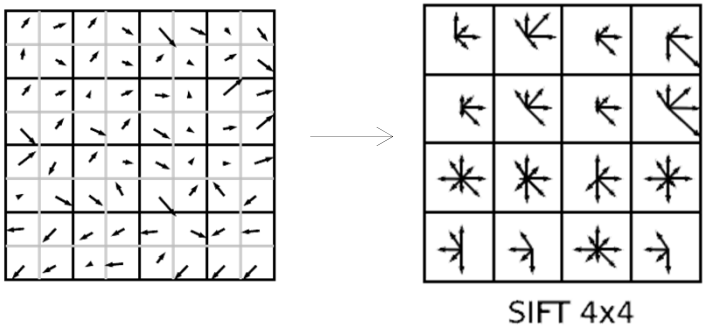
$$\cos(\mathbf{a}, \mathbf{b}) = \frac{\sum_{i=1}^{i=n} a_i \times b_i}{\sqrt{\sum_{i=1}^{i=n} a_i^2} \times \sqrt{\sum_{i=1}^{i=n} b_i^2}}$$

	shadow	shine
moon	16	29
sun	15	45
dog	10	0

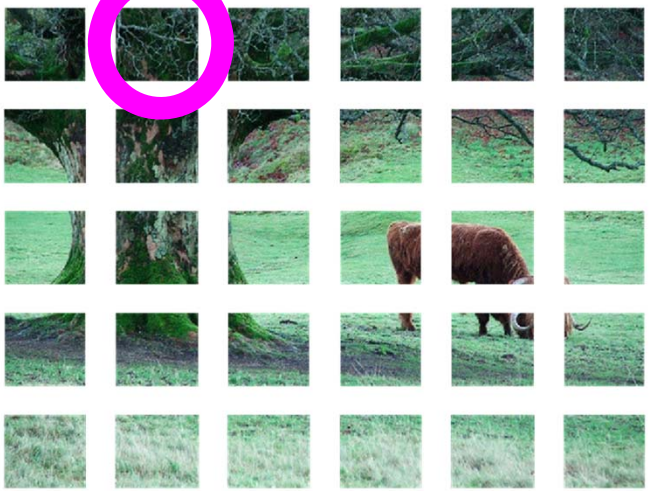


Distributional semantics from images

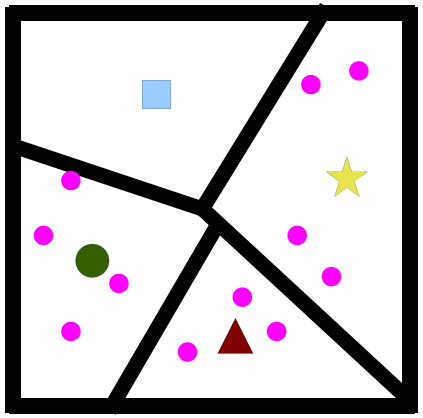
Bag-of-Words



Pixel-wise gradient responses



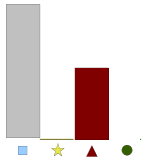
Descriptor Space

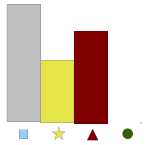


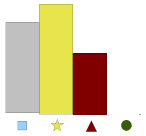
Global Representation

Distributional semantics from images

moon







Multimodal distributional semantics

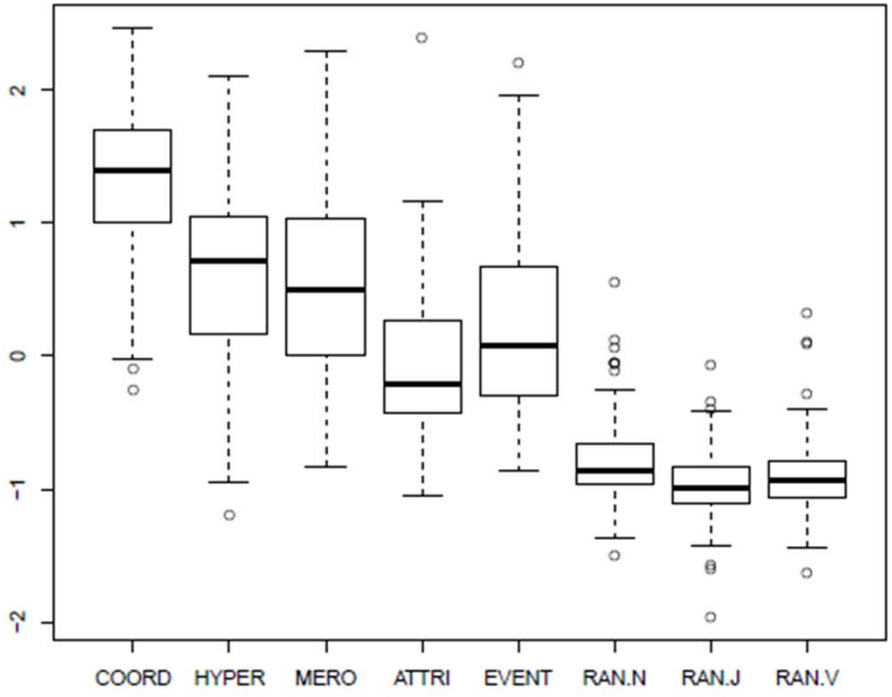
- Concatenation

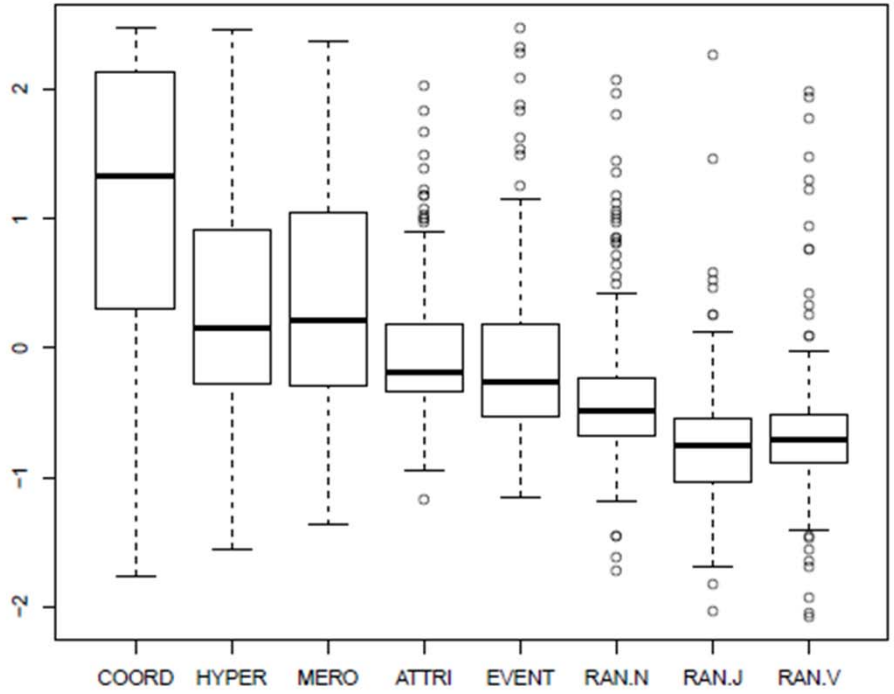
	shadow	shine		
moon	16	29	0	4
sun	15	45	2	9
dog	10	0	9	1

Semantics: text vs images

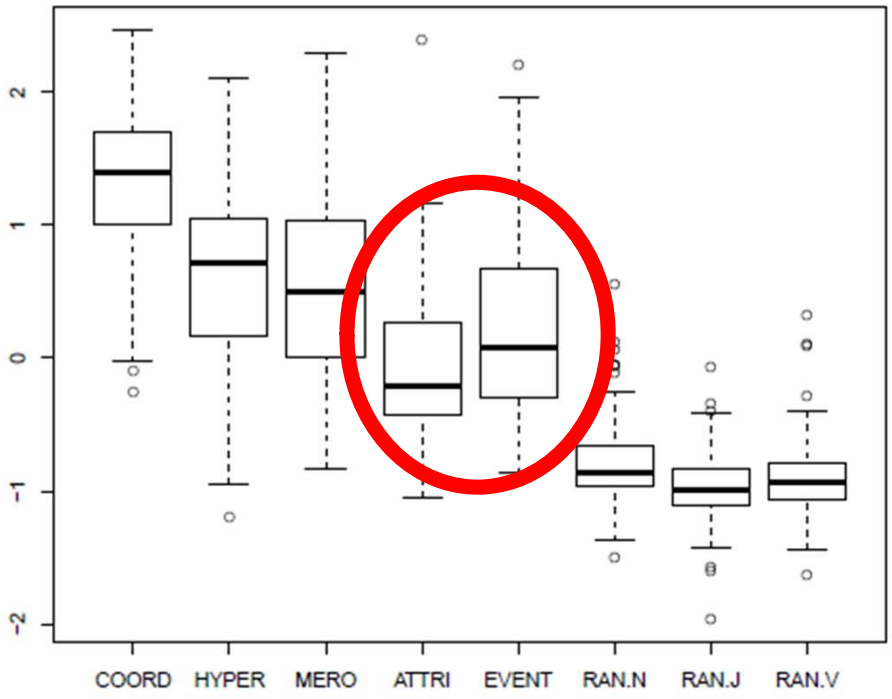
- BLESS dataset [Baroni 2011]
  - 200 pivot words
  - Human collected relata words in 8 categories
    - Coordinate: alligator - lizard
    - Hypernym (is-a): alligator - reptile
    - Meronym (part): alligator - teeth
    - Attribute: alligator - aquatic
    - Event (verb): alligator - swim
    - Random noun: alligator - trombone
    - Random adjectives: alligator - electronic
    - Random verbs: alligator – conclude
  - average 7-33 relata per category per pivot

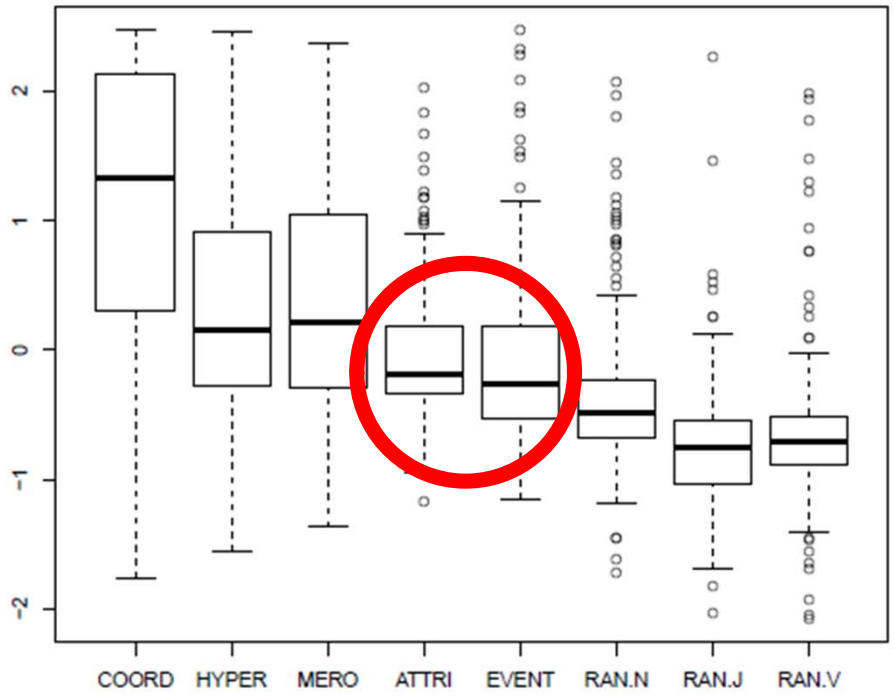
Semantics: text vs images





Semantics: text vs images





Semantics: text vs images

concept	*text*	*image*
cabbage	leafy	white
carrot	fresh	orange
cherry	ripe	red
deer	wild	brown
dishwasher	electric	white
hat	white	old
hatchet	sharp	short
onion	fresh	white
oven	electric	new
plum	juicy	red
sparrow	wild	little
tanker	heavy	grey

Improving Distributional Semantics with Eyes

- Datasets with human semantic judgements
  - WordSim (WS) [1]: 353 word pairs
  - MEN [2] : 3000 word pairs

[1] http://www.cs.technion.ac.il/~gabr/resources/data/wordsim353/
[2] http://clic.cimec.unitn.it/~elia.bruni/MEN

Improving Distributional Semantics with Eyes

Spearman correlations with human semantics:

Model	*MEN*	*WS*
Text	0.68	0.70
Image	0.43	0.36
Linear	0.73	0.67
Smoothed	**0.76**	**0.75**

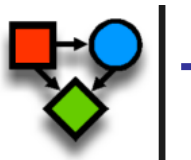
The illustrated distributional hypothesis

The meaning of a word can be derived from context

he curtains open and the moon shining in on the barely
ars and the cold , close moon " . And neither of the w
rough the night with the moon shining so brightly , it
made in the light of the moon . It all boils down , wr
surely under a crescent moon , thrilled by ice-white
sun , the seasons of the moon ? Home , alone , Jay pla
m is dazzling snow , the moon has risen full and cold
un and the temple of the moon , driving out of the hug
in the dark and now the moon rises , full and amber a
bird on the shape of the moon over the trees in front
But I could n't see the moon or the stars , only the
rning , with a sliver of moon hanging among the stars
they love the sun , the moon and the stars . None of
the light of an enormous moon . The plash of flowing w
man 's first step on the moon ; various exhibits , aer
the inevitable piece of moon rock . Housing The Airsh
oud obscured part of the moon . The Allied guns behind

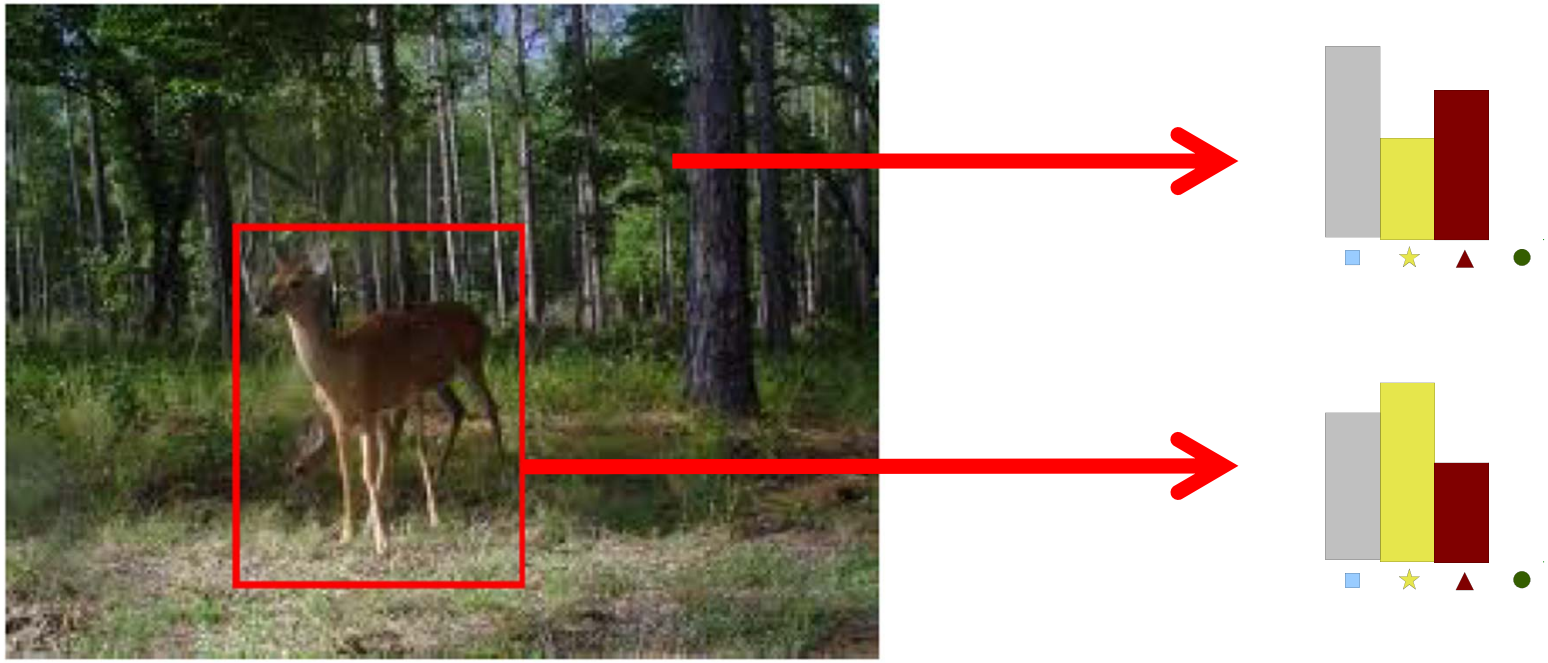
The illustrated distributional hypothesis

The meaning of a word can be derived from context



The illustrated distributional hypothesis

The meaning of a word can be derived from context



Pascal VOC 2007 (20 object classes, 5000 test images)

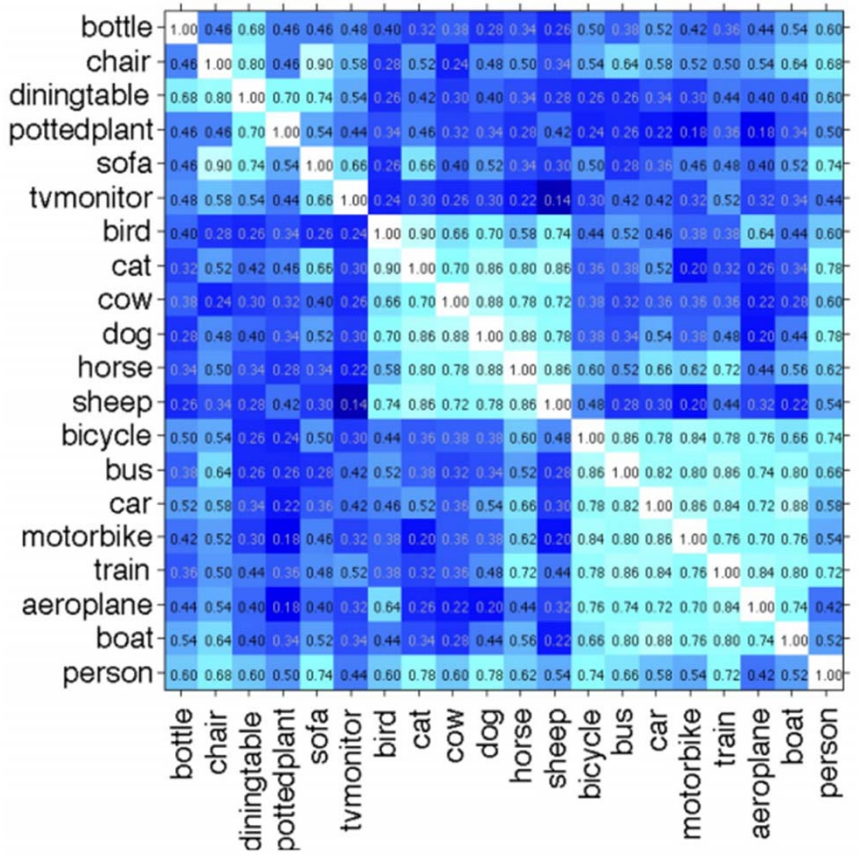
1) Ground truth object locations
2) Selective Search Localisation

The illustrated distributional hypothesis

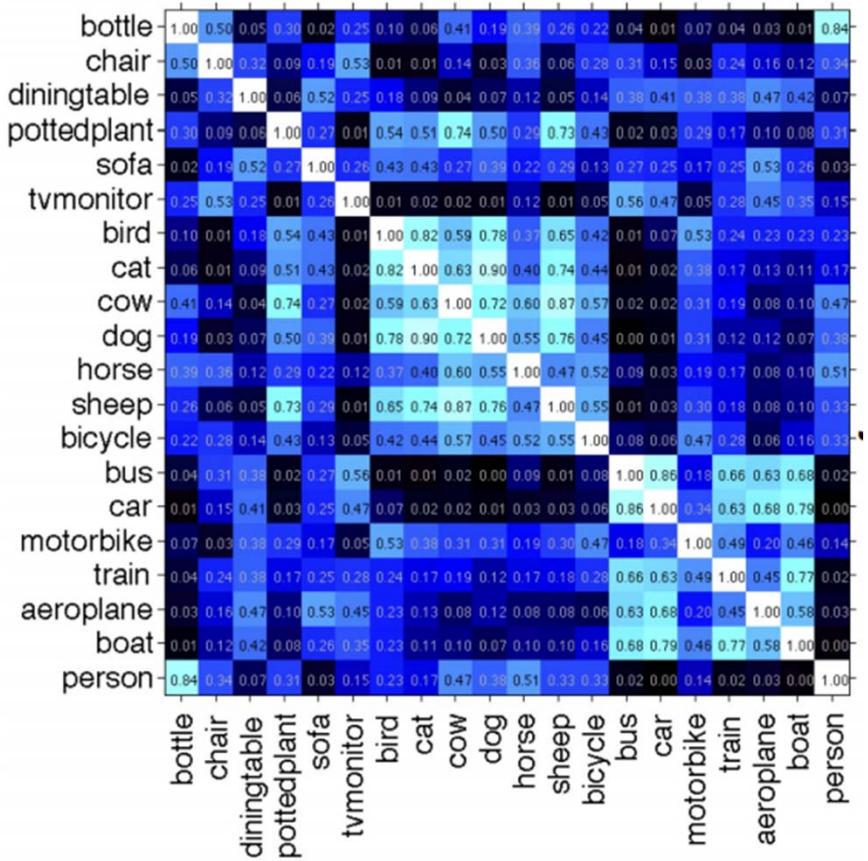
	Global (baseline)	Ground Truth	Selective Search
Object	-	0.39	0.36
Surround	-	0.50	0.51
Object+Surround	**0.47**	**0.54**	**0.54**

The illustrated distributional hypothesis

Human correlations

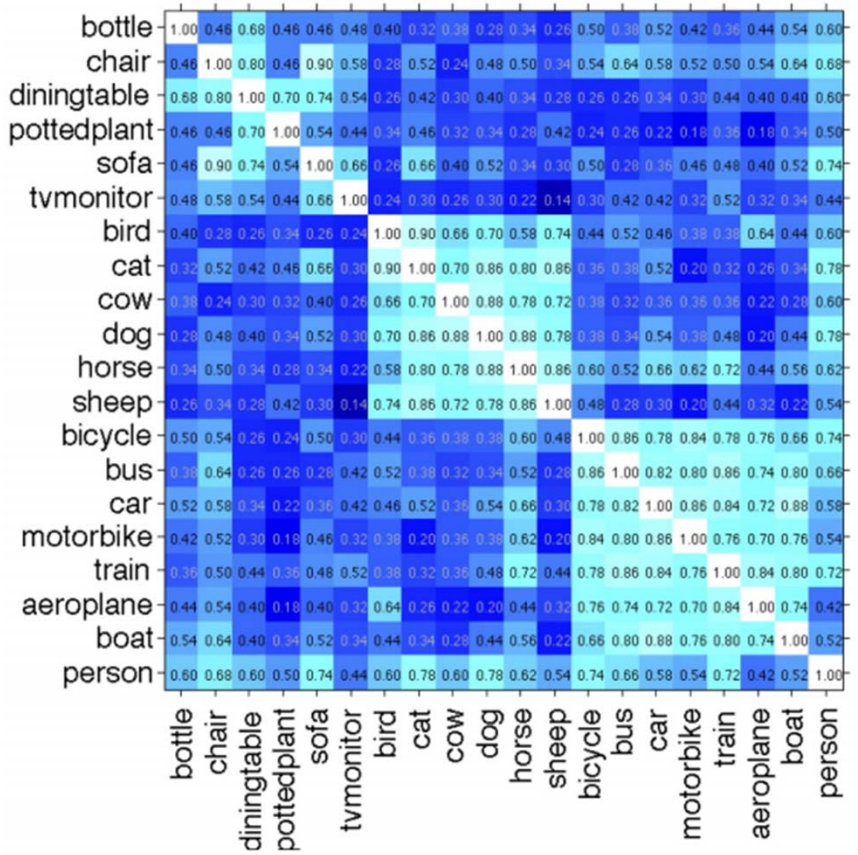


Object appearance correlations
(automatic localization)

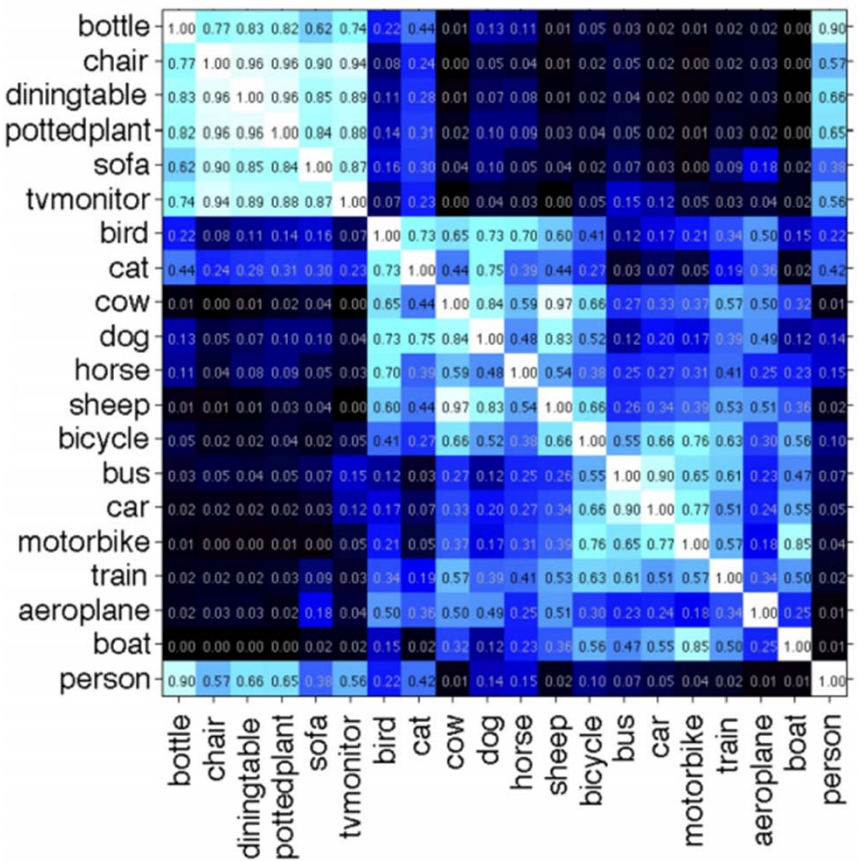


The illustrated distributional hypothesis

Human correlations



Surround appearance correlations (automatic localization)



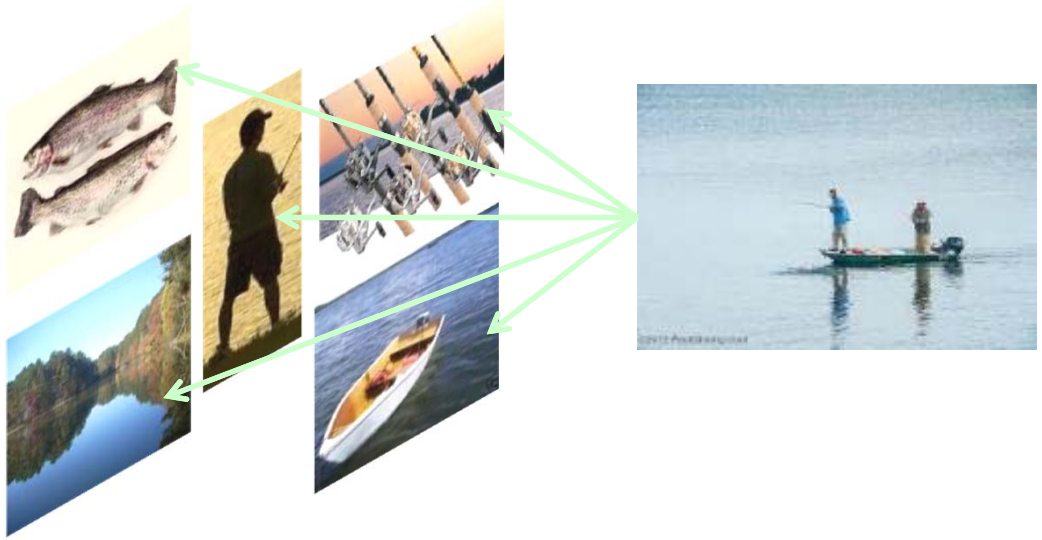
Summary

- Image and text have complementary semantic information
- Image features improve text-based task
- Distributional hypothesis for images works mostly on context

Outline

- Feature Selection (Q1)
- Semi-supervised Feature Analysis (Q2)
- Visual Info Helps Text Retrieval (Q3)
- Classifier-specific Representation (Q4)
- Knowledge Adaptation for MED (Q5)

Classifier-specific Representation



QUESTION 4:

Can we skip the explicit concept detection process but learn an intermediate representation using the available multimedia archives related to various concepts for complicated events?

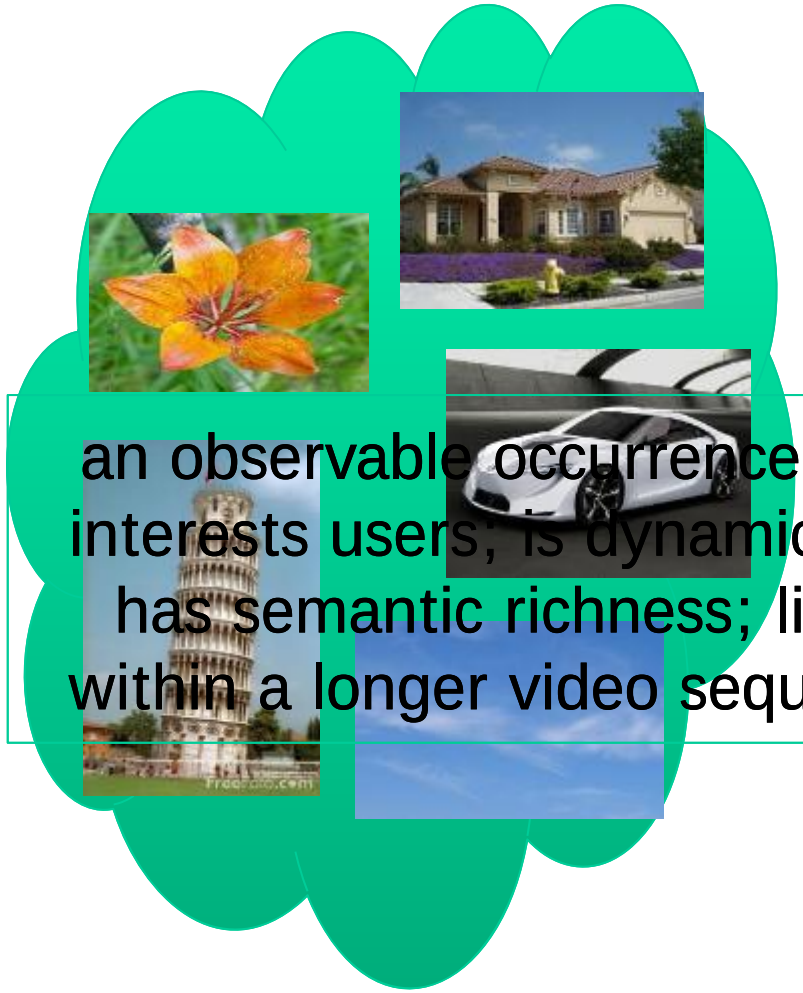
Multimedia Event Detection

Detect the occurrence of an event within a video clip based on an Event Kit, which contains some text description and some example videos

- National Institute of Standards and Technology

Concept

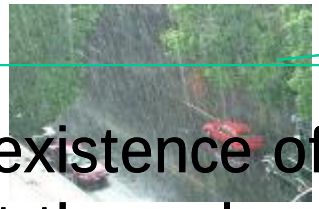
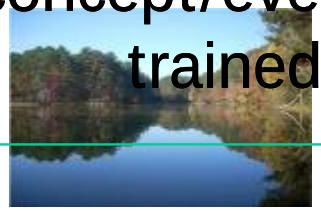
the abstract or general idea inferred from specific instances; usually describable by a single shot



an observable occurrence that interests users; is dynamic and has semantic richness; lives within a longer video sequence

Event

Annotation vs Detection

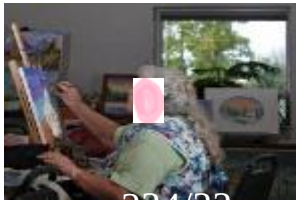





Detect the existence of concept/event through pre-trained detectors

Associate multime lia one se na gs)







Progress on Video Content Detection

- TREC Video Retrieval Evaluation (TRECVID)

Sports News Repetitive pattern Surveillance

- "Event detection in Internet multimedia (MED)" 2010: more complicated events, e.g., *Assembling a shelter*

Difficulty of MED

Inspiration

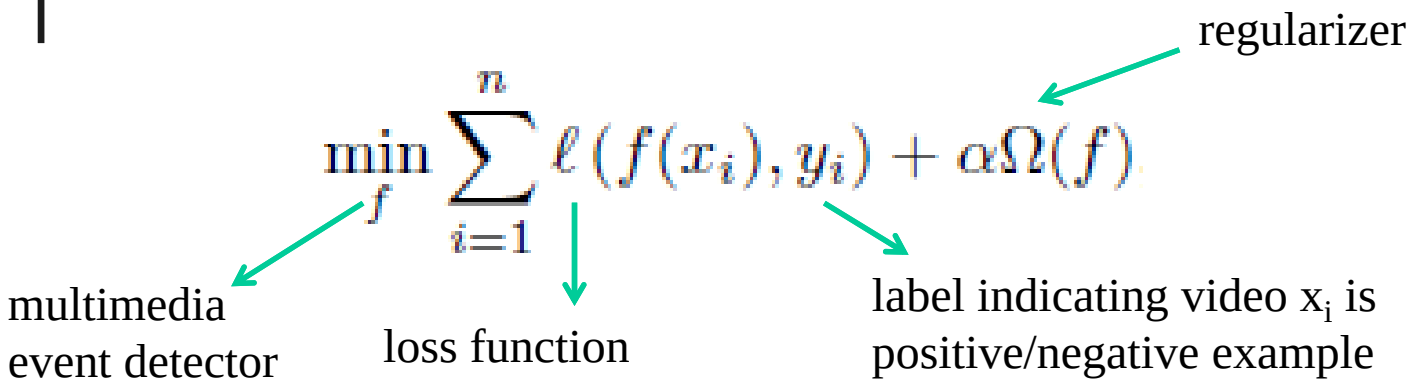
- Learning to refine multimedia representation
  - Limit: the refinement and the classifier training are independent
- Concepts-based representation
  - Limit: heavily dependent on concept detectors
- Available multimedia archives (concepts & events)

Methodology

- Learn an intermediate representation of videos by exploiting the target videos and external video archives together
- Integrate representation inference and classifier training into a joint framework
- Merits:
  - The optimization of classifier is event based
  - No need for pre-trained concept detectors

Z. Ma, Y. Yang, N. Sebe, K. Zheng, A. G. Hauptmann: “Multimedia Event Detection Using a Classifier-Specific Intermediate Representation“, IEEE Transactions on Multimedia, 15(7):1628-1637, 2013

Traditional approach



- Given a standard concept-based representation
- Use *m* annotated external videos $\{x_{n+1}, \ldots, x_{n+m}\}$ from *c* classes to pre-train *c* classifiers $g_k|_{k=1}^{c}$ (one for each intermediate concept, e.g., fish or boat for “landing a fish”)
- For each training or testing video $x_i (1 \leq i \leq n)$ the classifiers $g_k|_{k=1}^{c}$ are applied to detect the intermediate concepts
- Problem: $g_k|_{k=1}^{c}$ and *f* are trained independently

Formulation

- Joint learning of classifier and representation with external videos
  - exploit the shared components
  - assume that external-based videos and concept-based videos have a common intermediate representation

$$\min_{W,\Theta} \|X\Theta W - Y\|_{2,p} + \alpha \|W\|_F^2$$

$$s.t. \Theta^T \Theta = I$$

- avoid arbitrary scaling
- preserve information

$$\|A\|_{2,p} = \left(\sum_{i=1}^{d} (\sum_{j=1}^{c} |A_{ij}|)^{\frac{p}{2}} \right)^{\frac{1}{p}}$$

$X \in \mathbb{R}^{(n+m)\times d}$: target & external videos

$Y \in \mathbb{R}^{(n+m)\times(c+2)}$: labels

Θ : intermediate representation

Experiments

- Datasets
  - Target videos: TRECVID MED 2011 (15 events)
  - External videos: TRECVID MED 2011 development set (3 events)
- External videos: TRECVID 2011 semantic indexing task development set
  - concepts with few positive examples are removed
  - 65 concepts related to human, environment and objects
- Features: SIFT & CSIFT & MoSIFT

Events Visualization

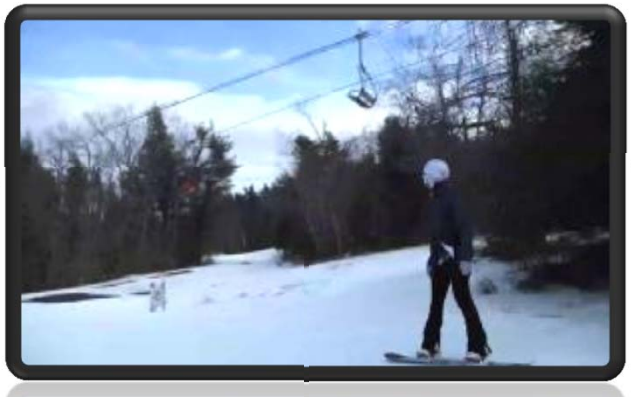
Making a cake

Batting a run

Assembling a shelter

Events Visualization

Feeding an animal



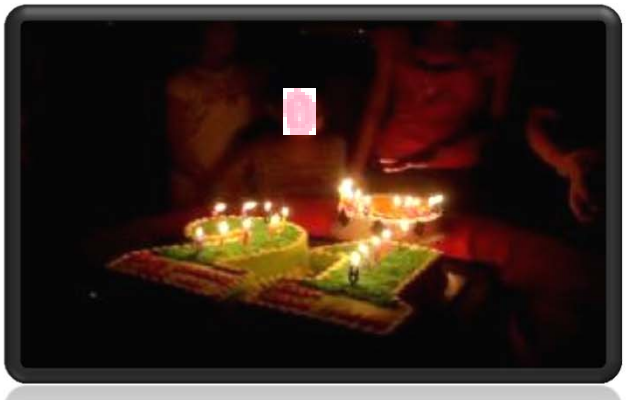
Attempting a board trick

Working on a woodworking project

Landing a fish

Wedding ceremony

Events Visualization

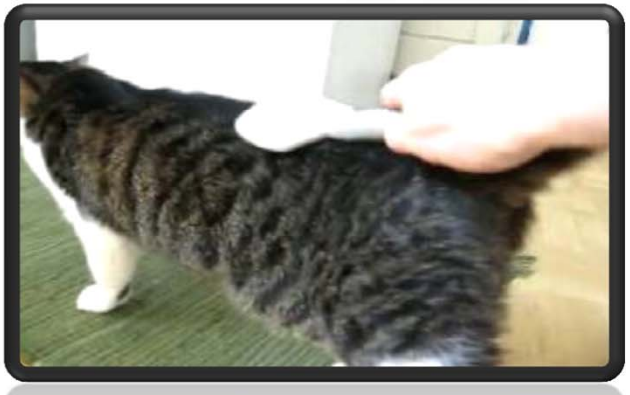


Birthday party

Changing a vehicle tire

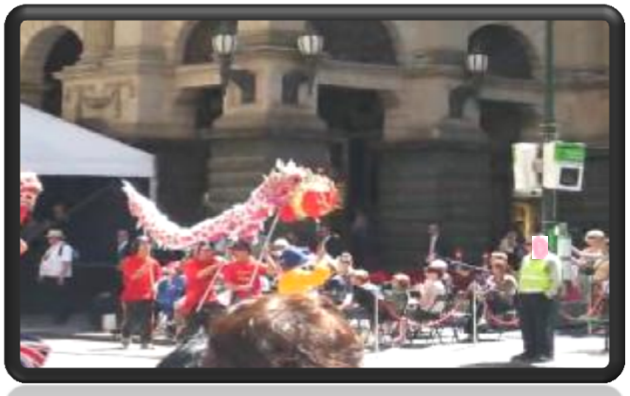
Flash mob gathering

Getting a vehicle unstuck



Grooming an animal

Events Visualization

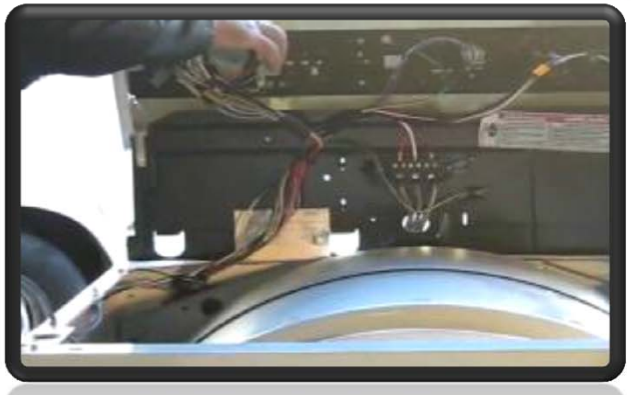


Parade

Making a sandwich

Working on a sewing project

Parkour



Repairing an appliance

Experiments

- Comparison algorithms:
  - Semantic Analysis via Intermediate Representation (SAIR)
  - AdaBoost
  - TaylorBoost, CVPR
  - SVM
  - LDA
  - Semantic Concept Representation (SCR), ECCV

Experiments

- MED performance comparison (MinNDC/AP)

Event	AdaBoost	TaylorBoost	SVM	LDA	SCR	SAIR
Attempting a board trick	1.218	0.995	0.826	0.998	**0.742**	0.775
	0.086	0.094	0.225	0.131	**0.274**	0.248
Feeding an animal	1.343	1.001	**0.963**	1.001	0.981	0.964
	0.037	0.043	0.087	0.045	0.079	**0.089**
Landing a fish	1.119	0.932	0.665	0.938	0.704	**0.626**
	0.065	0.097	0.260	0.103	0.234	**0.281**
Wedding ceremony	1.015	1.001	0.466	1.001	0.582	**0.441**
	0.084	0.067	0.483	0.073	0.322	**0.493**
Working on a woodworking project	1.203	1.001	0.726	1.001	0.940	**0.711**
	0.055	0.046	**0.294**	0.096	0.091	0.283

Experiments

- MED performance comparison (MinNDC/AP)

Event	AdaBoost	TaylorBoost	SVM	LDA	SCR	SAIR
Birthday party	1.211	1.001	0.885	1.001	0.939	**0.882**
	0.030	0.019	**0.079**	0.021	0.051	0.076
Changing a vehicle tire	1.187	1.001	0.670	1.001	0.862	**0.636**
	0.006	0.006	0.023	0.006	0.013	**0.030**
Flash mob gathering	1.139	1.001	0.629	1.001	**0.509**	0.568
	0.050	0.042	0.198	0.059	**0.291**	0.228
Getting a vehicle unstuck	1.031	0.902	0.802	0.970	**0.586**	0.711
	0.019	0.027	0.051	0.018	**0.107**	0.083
Grooming an animal	1.317	1.001	0.856	0.925	**0.814**	0.856
	0.006	0.013	0.046	0.025	**0.056**	0.047

Experiments

- MED performance comparison (MinNDC/AP)

Event	AdaBoost	TaylorBoost	SVM	LDA	SCR	SAIR
Making a sandwich	1.355	1.001	**0.821**	1.001	0.843	0.858
	0.008	0.009	**0.034**	0.010	0.029	0.030
Parade	1.091	0.991	0.654	1.001	0.712	**0.632**
	0.035	0.028	0.093	0.019	0.083	**0.108**
Parkour	1.156	0.955	0.570	1.001	0.566	**0.449**
	0.014	0.005	0.047	0.009	0.050	**0.055**
Repairing an appliance	0.971	1.001	0.550	0.822	0.664	**0.508**
	0.027	0.018	0.102	0.029	0.056	**0.109**
Working on a sewing project	1.188	1.001	0.706	0.974	0.833	**0.612**
	0.012	0.008	0.037	0.016	0.027	**0.054**
Average	1.163	0.986	0.719	0.976	0.752	**0.682**
	0.035	0.035	0.137	0.044	0.118	**0.148**

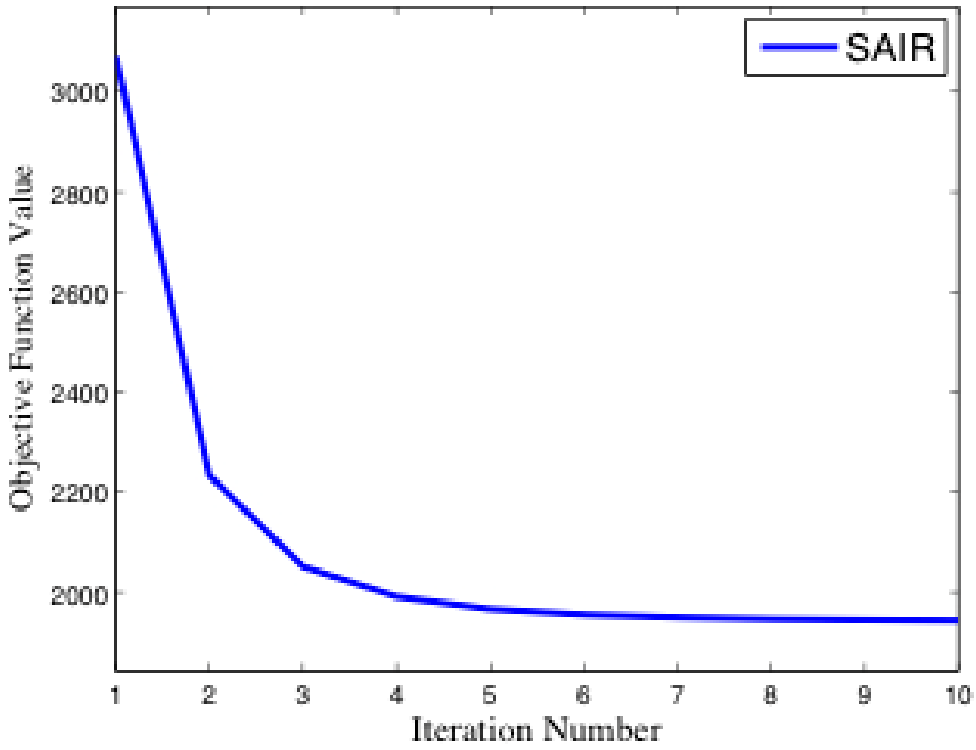
Experiments

- Performance comparison between using 30 and 65 external concepts (MinNDC/AP)

Event	SCR (30C)	SCR (65C)	SAIR (30C)	SAIR (65C)
Attempting a board trick	0.811	0.742	0.764	0.775
	0.215	0.274	0.246	0.248
Feeding an animal	0.976	0.981	0.961	0.964
	0.071	0.079	0.091	0.089
Landing a fish	0.722	0.704	0.625	0.626
	0.214	0.234	0.286	0.281

Experiments

- Convergence



Summary

- The intermediate representation is tightly coupled with the classifier
- Mutual benefit is attained
- External videos provide extra cues
- Promising results on TRECVID MED videos

Z. Ma, Y. Yang, N. Sebe, K. Zheng, A. G. Hauptmann: “Multimedia Event Detection Using a Classifier-Specific Intermediate Representation“, IEEE Transactions on Multimedia, 15(7):1628-1637, 2013

Outline

- Feature Selection (Q1)
- Semi-supervised Feature Analysis (Q2)
- Visual Info Helps Text Retrieval (Q3)
- Classifier-specific Representation (Q4)
- Knowledge Adaptation for MED (Q5)

Knowledge Adaptation

QUESTION 5:

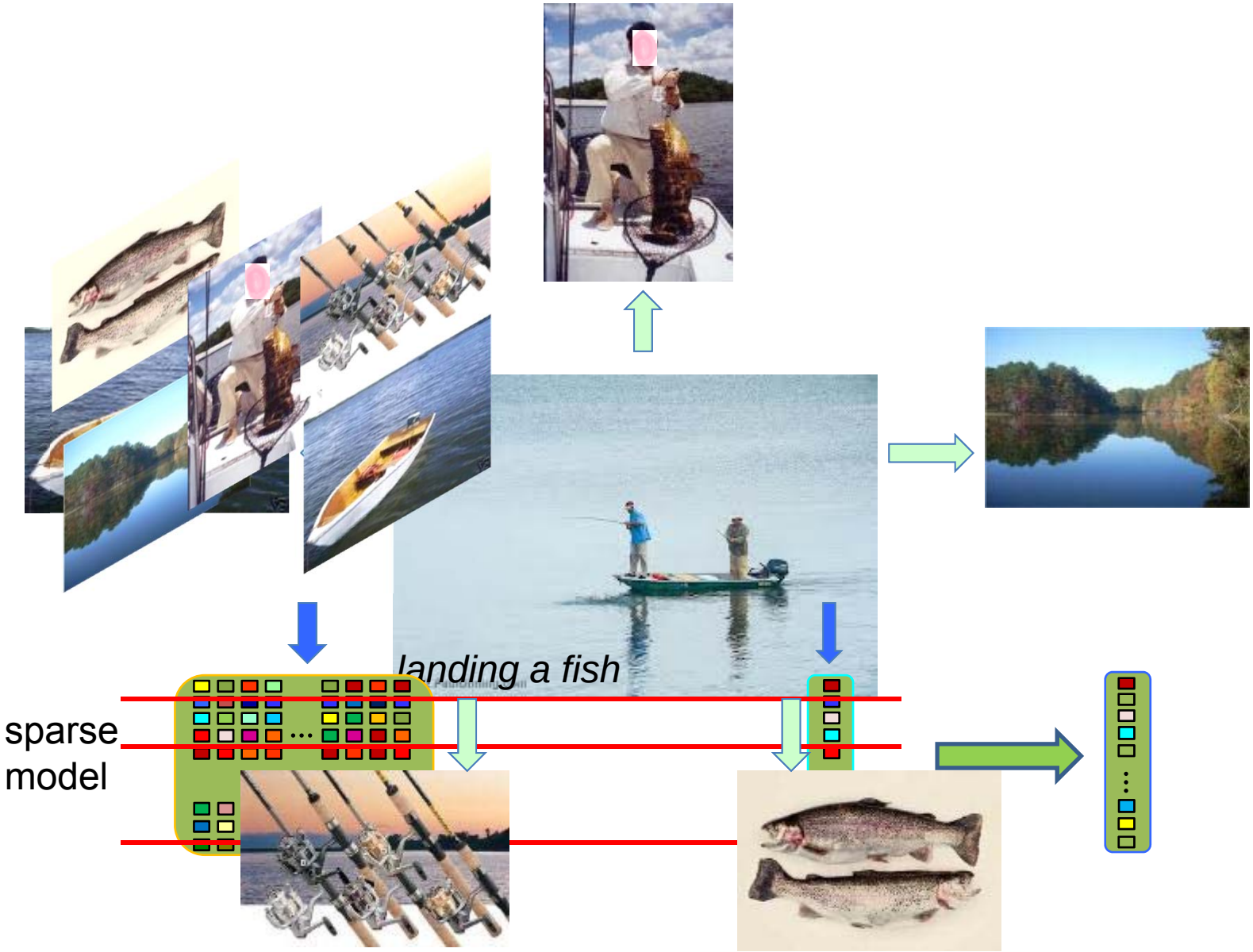
How can we guarantee reasonable multimedia event detection accuracy when only few positive exemplars are provided?

Inspiration

- The information from few positive examples is limited
- Borrow strength from other multimedia resources
- Concepts-based videos are used as auxiliary resource

A Noticeable Difference

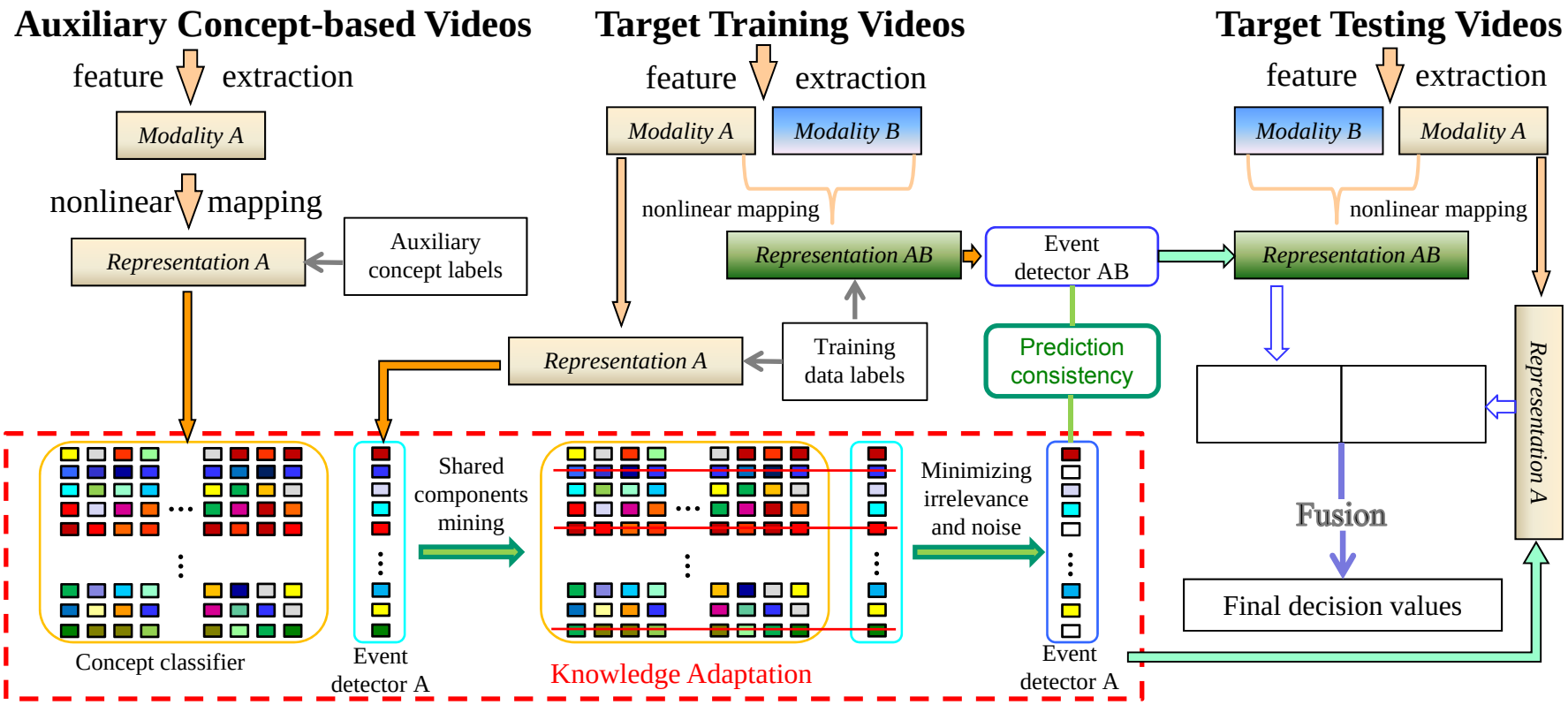
- No requirement for the consistency between the auxiliary and target domains in feature type
- Benefits:
  - Flexible with the situation that data collection platforms change or augment their capabilities


landing a fish
sparse
model

Methodology

- Map the homogeneous features of the auxiliary and target videos (i.e., Modality A) into another space by a nonlinear mapping
- The video concept classifier and the video event detector obtained from the homogeneous features have common components which contain irrelevance and noise: remove by joint optimization
- Another event detector of MED videos is trained based on Modality B
- Integrate the two event detectors for optimization after which the decision values from both are fused for the final prediction

Framework



Z. Ma, Y. Yang, N. Sebe and A. G. Hauptmann: “Knowledge Adaptation with Partially Shared Features for Event Detection Using Few Exemplars". IEEE Transactions on Pattern Analysis and Machine Intelligence, 36(9):1789-1802, 2014

Objective function (1)

1. Associate the low-level representation and high-level semantic concepts

target training videos

event detector (modality A)

$$\min_{P_t, b_t} \left\| \tilde{Z}_t^T P_t + 1_t b_t - y_t \right\|_{2,1}$$

2. Do the same for the auxiliary videos

auxiliary videos

event detector

$$\min_{W_a, b_a} \left\| \tilde{X}_a^T W_a + 1_a b_a - Y_a \right\|_{2,1}$$

3. The predicted labels in modalities A and B are consistent

event detector (modality B)

$$\left\| \tilde{X}_t^T W_t - \tilde{Z}_t^T P_t \right\|_F^2$$

Objective function (2)

$$\min_{P_t, W_t, W_a, b_t, b_a} \left\| \tilde{Z}_t^T P_t + 1_t b_t - y_t \right\|_{2,1} + \left\| \tilde{X}_t^T W_t - \tilde{Z}_t^T P_t \right\|_F^2$$
$$+ \left\| \tilde{X}_a^T W_a + 1_a b_a - Y_a \right\|_{2,1} + \alpha \|W\|_{2,p}^p + \beta(\|W_t\|_F^2 + \|W_a\|_F^2)$$

Auxiliary Concept-based Videos

Target Training Videos

avoids overfitting

Experiments

- Datasets:
  - TRECVID MED 2010
  - TRECVID MED 2011 development set

  } 9746 video clips
- TRECVID 2012 semantic indexing task dev. set
  - auxiliary videos
  - concepts with few positive examples are removed
  - 65 concepts related to human, environment and objects
  - 3244 video frames
- UCF50
  - auxiliary videos
  - 50 actions
  - 6681 video clips

Experiments

- Features
  - Overlapping: SIFT + CSIFT (SIN12 as auxiliary data)
    STIP (UCF50 as auxiliary data)
  - Different: MFCC
- Setting
  - 10 positive example

Experiments

- Comparison algorithms:
  - Heterogenous Features based Structural Adaptive Regression (HF-SAR)
  - Structural Adaptive Regression (SAR), ACM MM
  - Adaptive Multiple Kernel Learning (A-MKL), T-PAMI
  - Multiple Kernel Transfer Learning (MKTL), ICCV
  - SAR&SVM
  - SVM
  - TaylorBoost, CVPR

Notations:

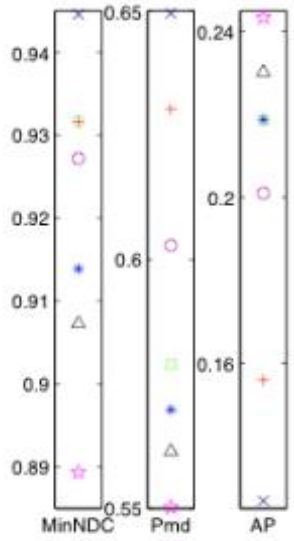
- HF-SAR
- SAR
- A-MKL
- MKTL
- SAR&SVM
- SVM
- TaylorBoost

MinNDC: Minimum NDC
The LOWER, the BETTER.

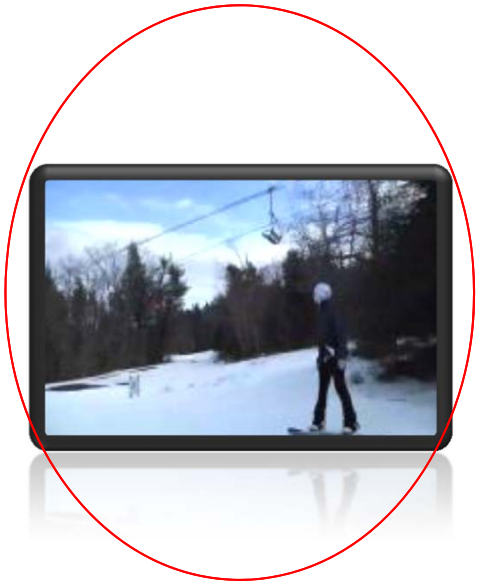
Pmd: Pmd@TER=12.5
The LOWER, the BETTER.

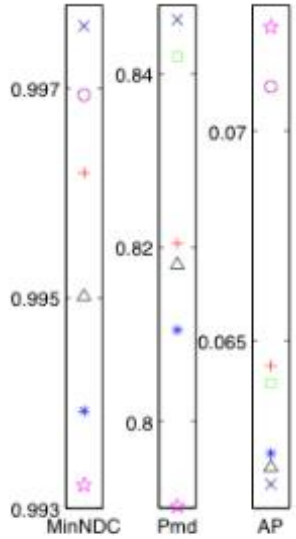
AP: Average Precision
The HIGHER, the BETTER.

(a) Notations

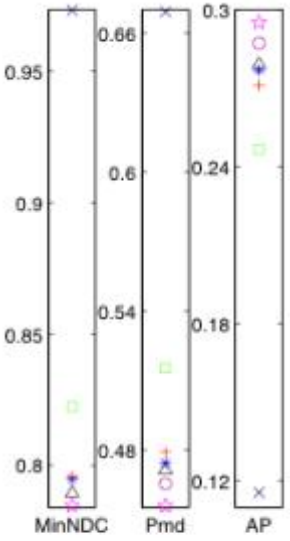


(b) *Attempting a board trick*

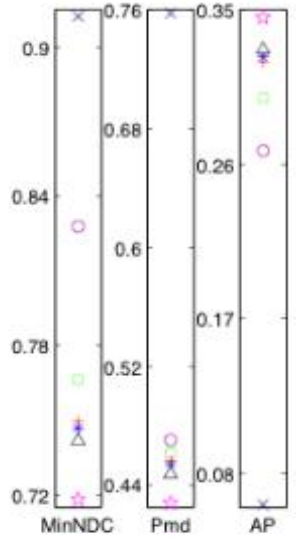




(c) *Feeding an animal*



(d) *Landing a fish*



(e) *Wedding ceremony*

Notations:

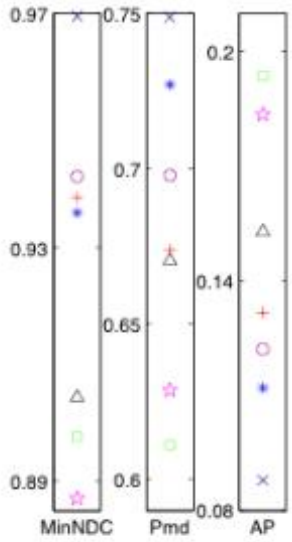
- HF-SAR
- SAR
- A-MKL
- MKTL
- SAR&SVM
- SVM
- TaylorBoost

MinNDC: Minimum NDC
The LOWER, the BETTER.

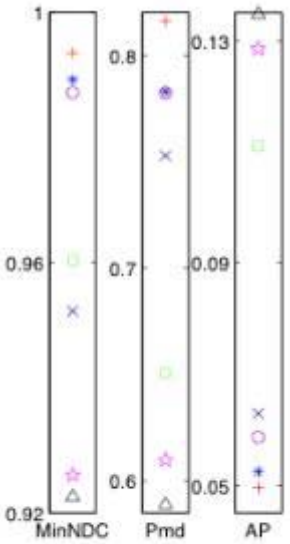
Pmd: Pmd@TER=12.5
The LOWER, the BETTER.

AP: Average Precision
The HIGHER, the BETTER.

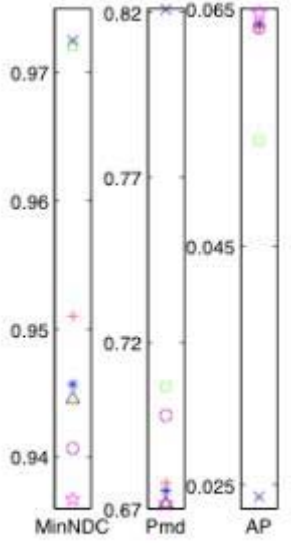
(a) Notations



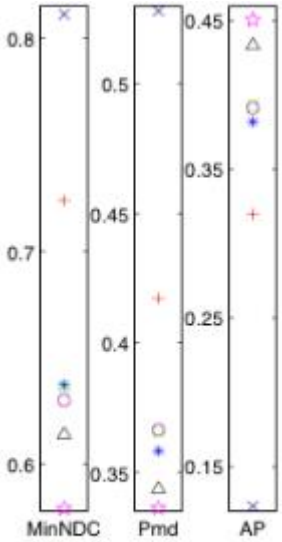
(f) *Working on a wood-working project*



(g) *Birthday party*



(h) *Changing a vehicle tire*



(i) *Flash mob gathering*

Notations:

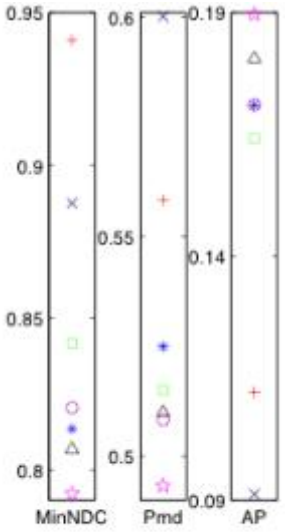
- HF–SAR
- SAR
- A–MKL
- MKTL
- SAR&SVM
- SVM
- TaylorBoost

MinNDC: Minimum NDC
The LOWER, the BETTER.

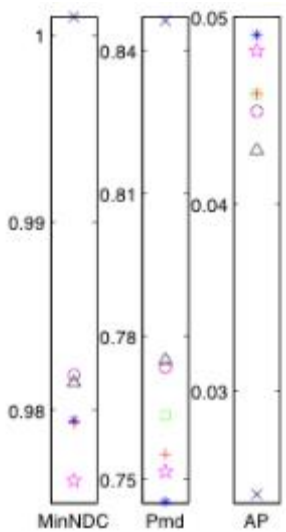
Pmd: Pmd@TER=12.5
The LOWER, the BETTER.

AP: Average Precision
The HIGHER, the BETTER.

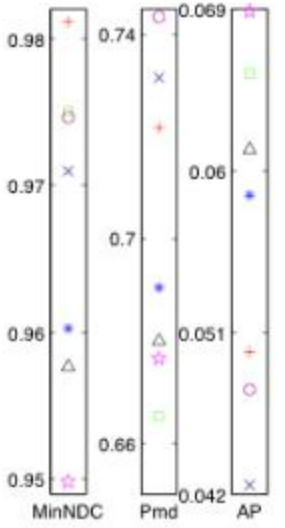
(a) Notations



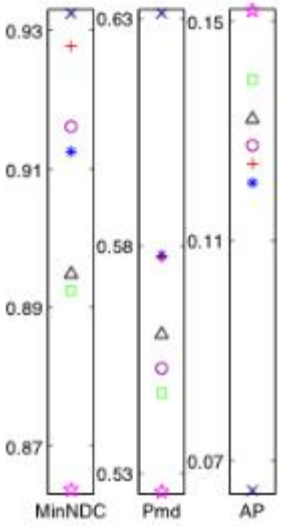
(j) *Getting a vehicle unstuck*



(k) *Grooming an animal*



(l) *Making a sandwich*



(m) *Parade*

Notations:

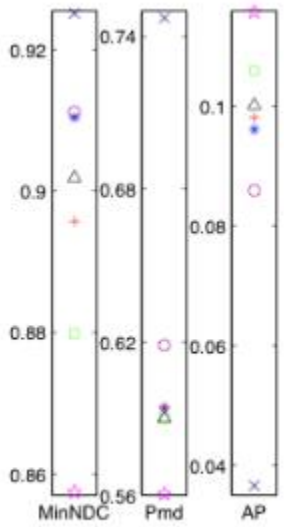
HF−SAR
SAR
A−MKL
MKTL
SAR&SVM
SVM
TaylorBoost

MinNDC: Minimum NDC
The LOWER, the BETTER.

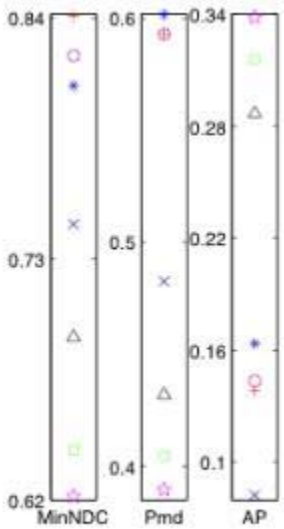
Pmd: Pmd@TER=12.5
The LOWER, the BETTER.

AP: Average Precision
The HIGHER, the BETTER.

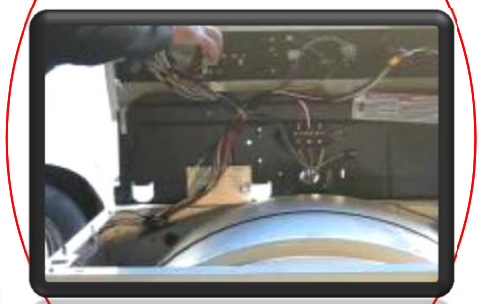
(a) Notations

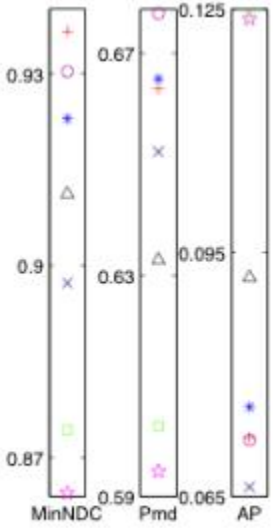


(n) *Parkour*

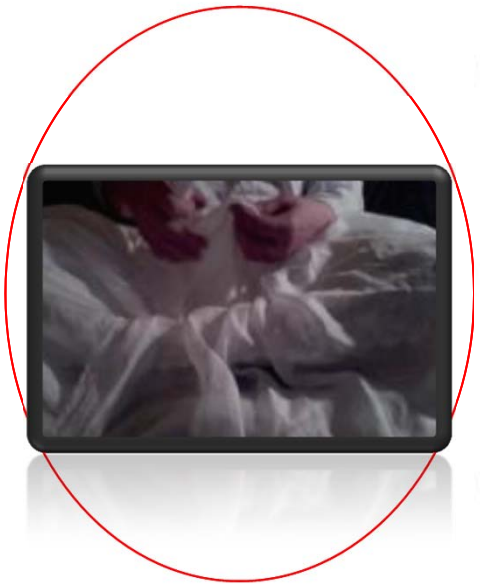


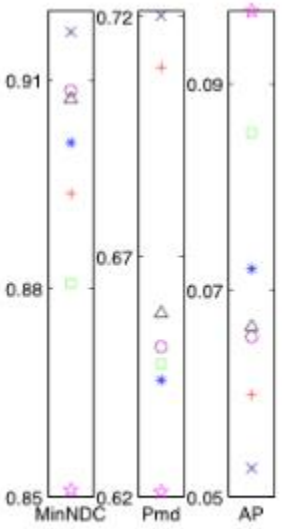
(o) *Repairing an appliance*





(p) *Working on a sewing project*





(q) *Making a cake*

Notations:

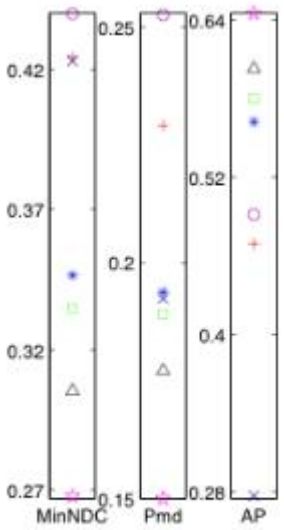
- HF-SAR
- SAR
- A-MKL
- MKTL
- SAR&SVM
- SVM
- TaylorBoost

MinNDC: Minimum NDC
The LOWER, the BETTER.

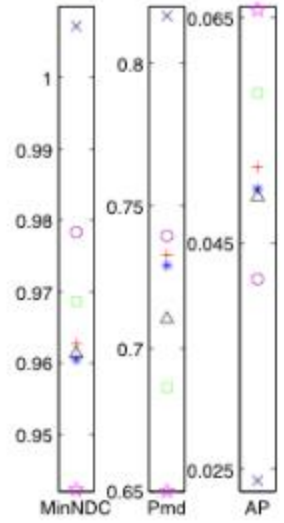
Pmd: Pmd@TER=12.5
The LOWER, the BETTER.

AP: Average Precision
The HIGHER, the BETTER.

(a) Notations



(r) *Batting a run*



(s) *Assembling a shelter*

Experiments

- Average results (MinNDC/Pmd@TER=12.5/AP)

SAR	A-MKL	MKTL	SAR&SVM	SVM	Taylor Boost	HF-SAR
0.860	0.881	0.873	0.841	0.850	0.902	**0.817**
0.601	0.617	0.610	0.572	0.575	0.677	**0.549**
0.162	0.144	0.153	0.183	0.181	0.080	**0.201**

Experiments

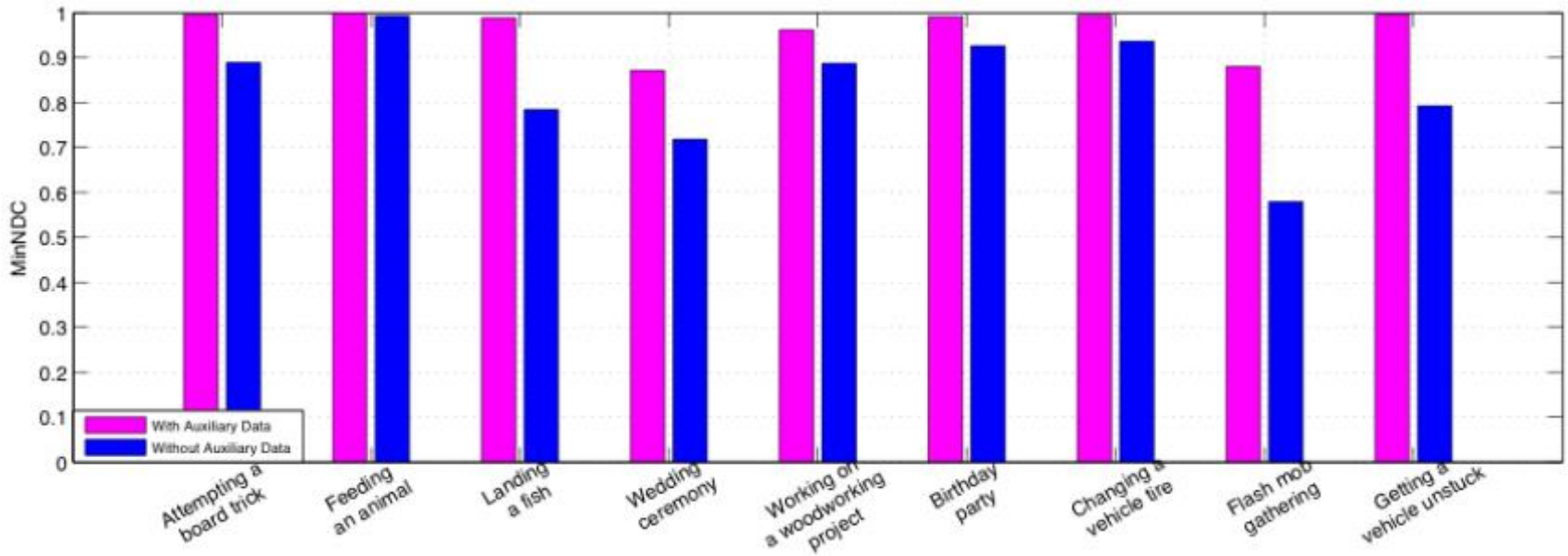
- When using UCF50, HF-SAR is similarly more robust than SVM
- HF-SAR is better than SVM for 17, 17, 15 events with different metrics

Average performance of SVM and SAR

Evaluation Metric	SVM	HF-SAR	Relative Improvement
MinNDC	0.965	0.932	3.5%
Pmd@TER=12.5	0.857	0.764	12.2%
AP	0.069	0.098	42.0%

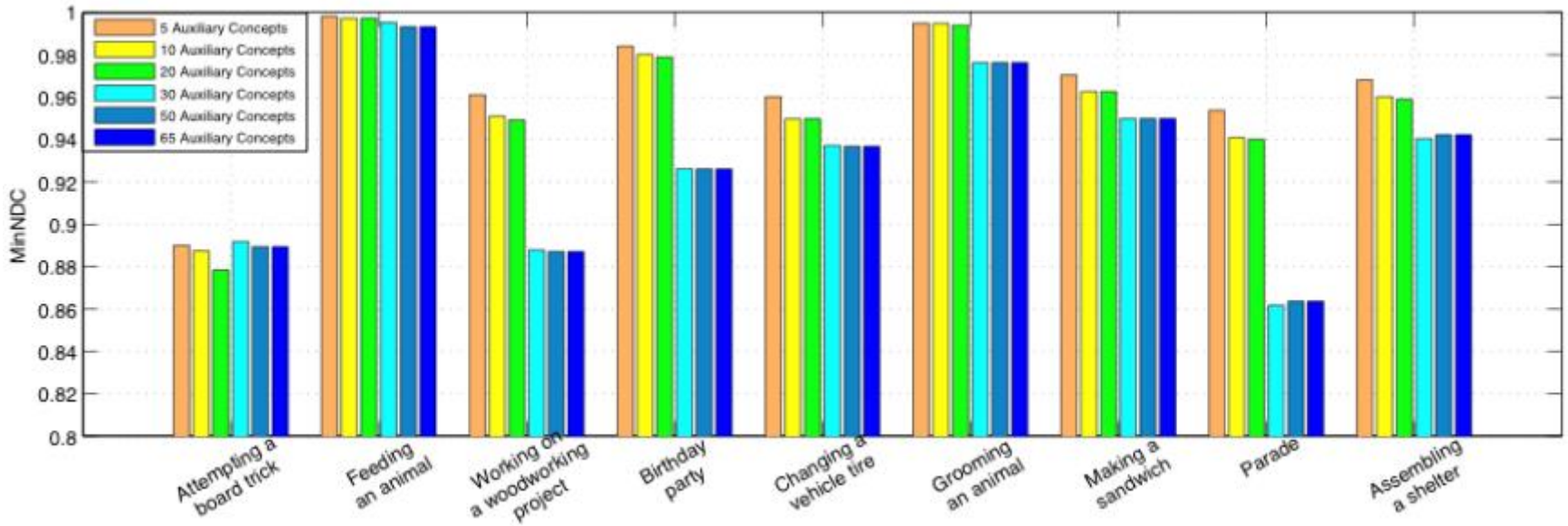
Experiments

- Influence of knowledge adaptation



Experiments

- Influence of auxiliary concepts



Summary

- An attempt on MED with few exemplars
- More generic, complicated and meaningful events
- Knowledge adaptation from concepts-based videos
- Heterogeneous feature type
- Effectiveness on TRECVID MED videos

Conclusion

Focused on image and video annotation and MED

- From algorithm perspective:
  - Feature selection - Solution for Q1: A better representation?
  - Semi-supervised learning - Solution for Q2: With few labels?
  - Multimodal approach - Solution for Q3: multiple modalities?
  - Shared subspace learning - Solution for Q4: Classifier-specific intermediate representation?
  - Transfer learning - Solution for Q5: Handling complex event detection with few exemplars?
- From application perspective:
  - Concepts to events
  - Images to videos

 Progressive Process

Ongoing Work

- Harnessing different features jointly as symbiotic solutions
- Model the importance of negative examples
- Knowledge adaptation that leverages unlabeled data in multiple related domains
- Knowledge adaptation between two domains that have partially shared data features
- User-centric research problems