Object Category Recognition

Christian Leistner,
Tinne Tuytelaars,
Luc Van Gool

Also based on slides by J. Gall, K. Grauman, B. Leibe

Plan

1. Problem definition, classification vs. detection

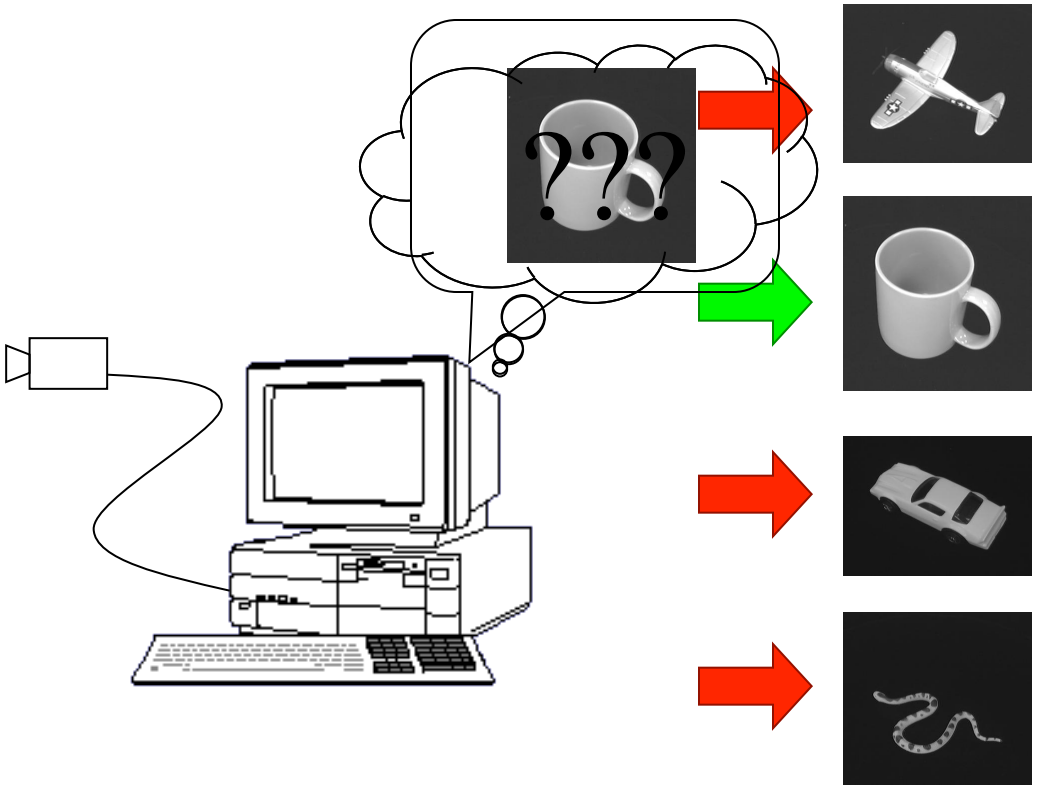
2. bag-of-words
3. sliding-windows (with global or part-based models)
4. Hough approaches (implicit shape model & forests)

5. Open challenges

1. Problem definition

Specific objects vs. object categories





Object Categorization

- How to recognize ANY car







- How to recognize ANY cow




What Could Be Done With Recognition Algorithms?

There is a wide range of applications, including…



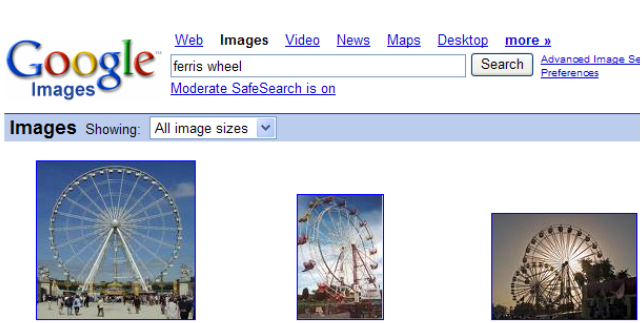
Autonomous robots

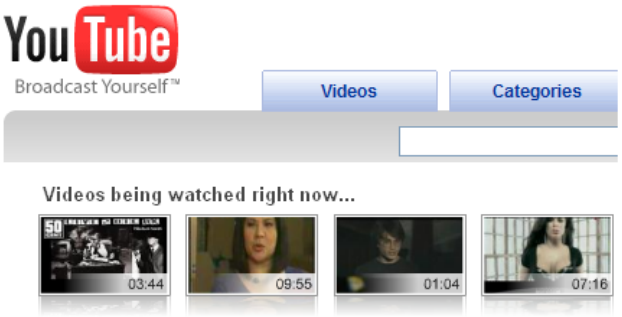


Navigation, driver safety

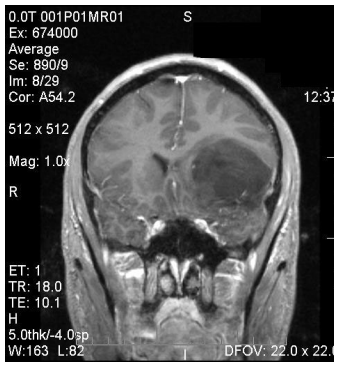


Surveillance





Content-based retrieval and analysis for images and videos



Medical image analysis

Slide credit: Kristen Grauman

Challenges: Robustness



Illumination





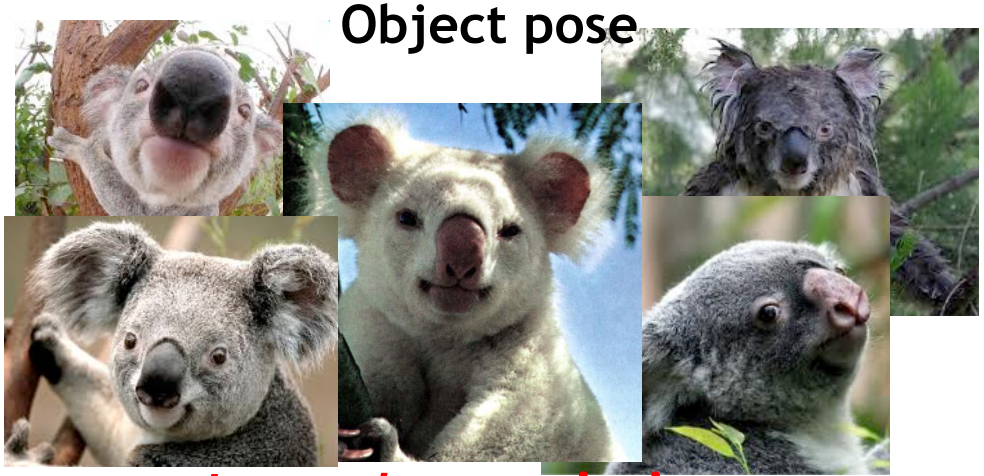
Object pose



Clutter



Occlusions



Intra-class variation



Viewpoint

On top of factors affecting specific object recognition, the added complexity of intra-class variation

Modified from Kristen Grauman

Two main tasks: Classification and Detection



Classification: *is there* a car in this image ? Yes or no suffices

Detection: *where* is the car ? Needs localization

2. Bag-of-Words

Plan

1. Problem definition, classification vs. detection

2. *bag-of-words*
3. sliding-windows (with global or part-based models)
4. Hough approaches (implicit shape model & forests)

5. Open challenges

Analogy to Documents

Of all the sensory impressions proceeding to the brain, the visual experiences are the dominant ones. Our perception of the world around us is based essentially on the messages that [illegible] brain from our eyes. For [illegible] ought that the re[illegible] point by [illegible] brain; t[illegible] screen[illegible] image[illegible] the dis[illegible] now kn[illegible] visual pe[illegible] considerab[illegible] of events. By fo[illegible] ses along their path t[illegible] vers of the optical cortex, Hubel and [illegible] have been able to demonstrate tha[illegible] *message about the image falling on[illegible] retina undergoes a step-wise analysis[illegible] system of nerve cells stored in colum[illegible] In this system each cell has its specific function and is responsible for a specific detail in the pattern of the retinal image.*

sensory, brain, visual, perception, retinal, cerebral cortex, eye, cell, optical nerve, image Hubel, Wiesel

China is forecasting a trade surplus of $90bn (£51bn) to $100bn this year, a threefold increase on 2004's $32bn. The Commerce Ministry said the surplus would be crea[illegible] 30% jump in exports [illegible] a 18% rise [illegible] ures are like[illegible] has lo[illegible] unfair[illegible] under[illegible] surplu[illegible] only on[illegible] Zhou Xia[illegible] needed to [illegible] demand so n[illegible] n the country. China [illegible] f the yuan against the dollar by 2.1% [illegible] and permitted it to trade within a [illegible] band, but the US wants the yuan to [illegible] allowed to trade freely. However, Be[illegible] has made it clear that it will take its ti[illegible] and tread carefully before allowing the yuan to rise further in value.

China, trade, surplus, commerce, exports, imports, US, yuan, bank, domestic, foreign, increase, trade, value

Slide credit: Li Fei-Fei

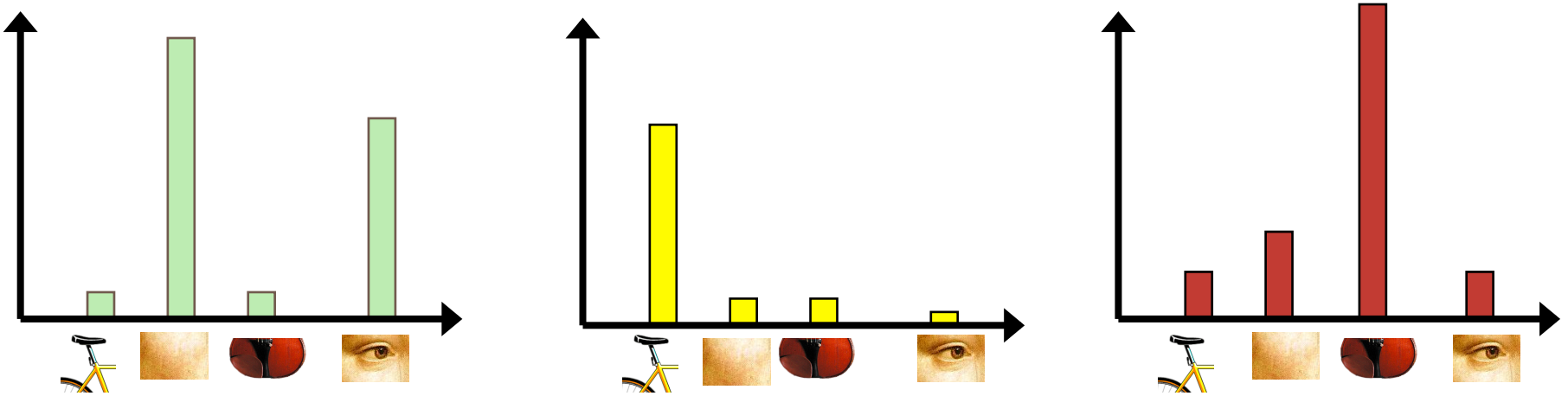
Object → Bag of ‘words’





Source: ICCV 2005 short course, Li Fei-Fei





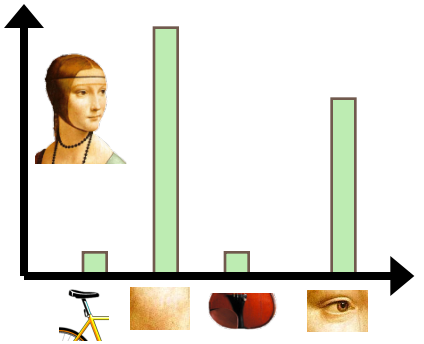


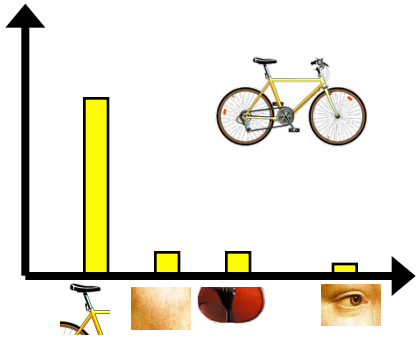
Source: ICCV 2005 short course, Li Fei-Fei

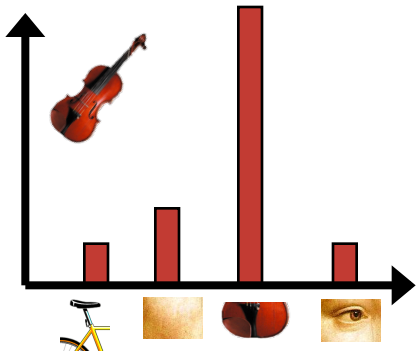
Bags of Visual Words



- Summarize entire image based on its distribution (histogram) of word occurrences.
- Analogous to bag of words representation commonly used for documents.







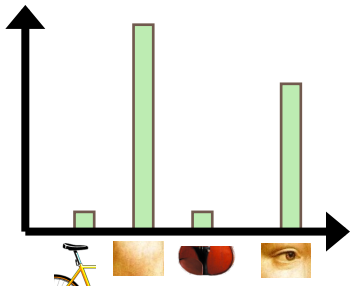
Slide credit: Kristen Grauman

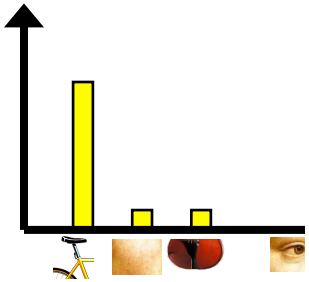
Image credit: Li Fei-Fei

Comparing Bags of Words

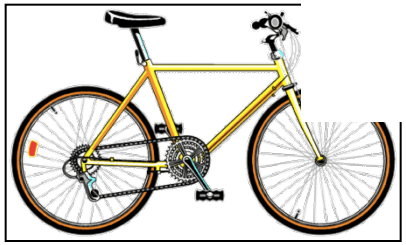
- Any histogram comparison measure can be used here.
- E.g. normalized scalar product between their (possibly weighted) occurrence counts as similarity score
  - *Nearest neighbor* search for similar images.

[1 8 1 4] • [5 1 1 0]









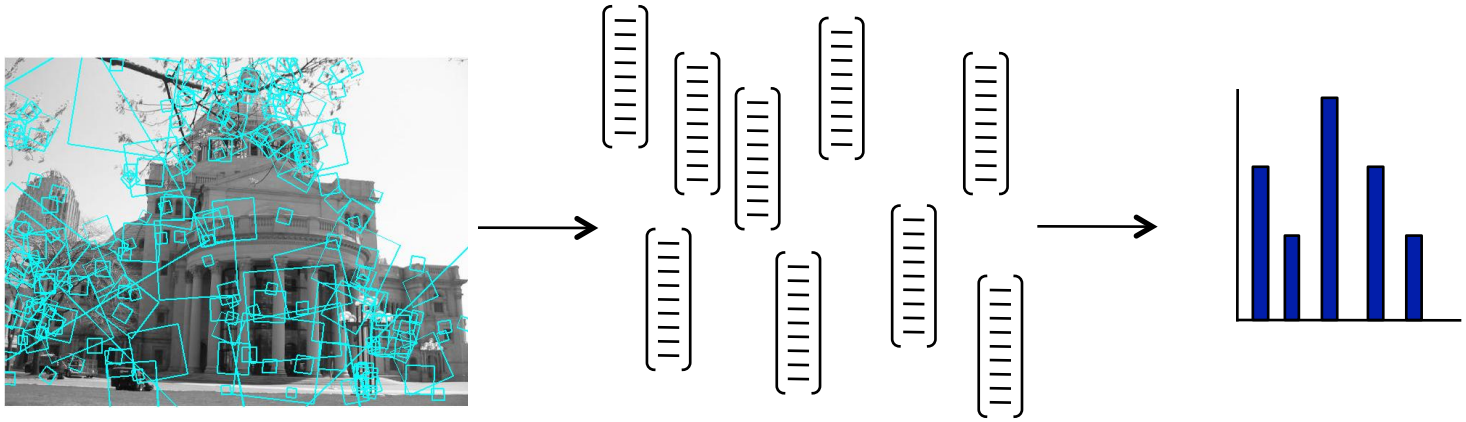
$\vec{d}_j$ $\vec{q}$

$$sim(d_j, q) = \frac{\vec{d}_j \bullet \vec{q}}{|\vec{d}_j| \times |\vec{q}|} = \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{j=1}^{t} w_{i,q}^2}}$$

Slide adapted from Kristen Grauman

Learning/Recognition with BoW Histograms

- Bag of words enable to describe the unordered feature set with a single vector of fixed dimensionality



- Provides easy way to use distribution of feature types with various learning algorithms requiring vector input

Slide credit: Kristen Grauman

BoW for Object Categorization





{face, flowers, building}

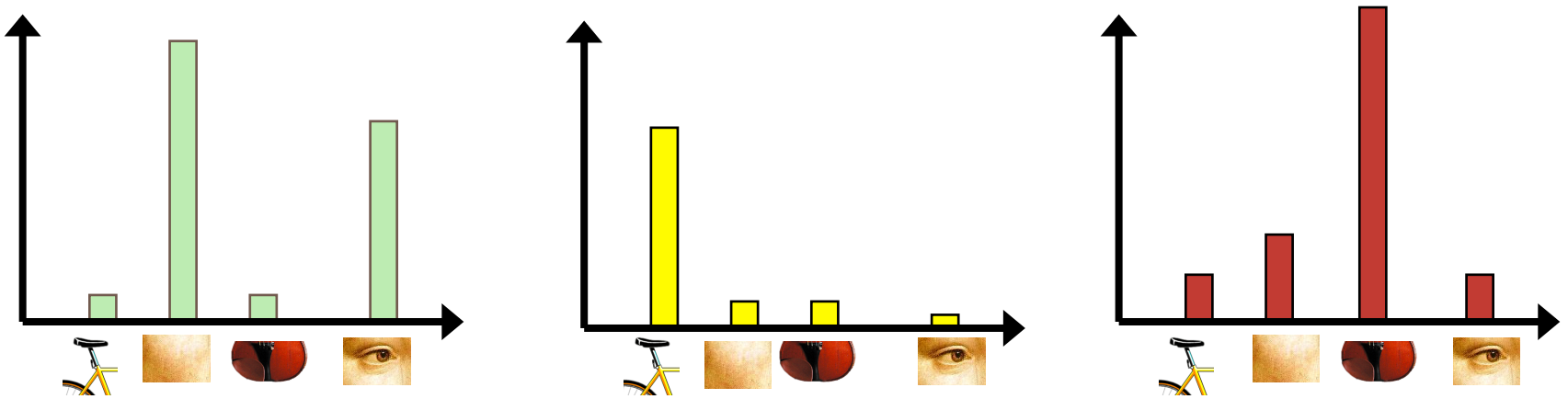
- Works pretty well for whole-image classification

Csurka et al. (2004), Willamowski et al. (2005), Grauman & Darrell (2005), Sivic et al. (2003, 2005)

Slide credit: Svetlana Lazebnik

Image Classification

- Given bag-of-features of images from different classes, how do we learn a model for distinguishing them?



Slide credit: Svetlana Lazebnik

Classifier Choices

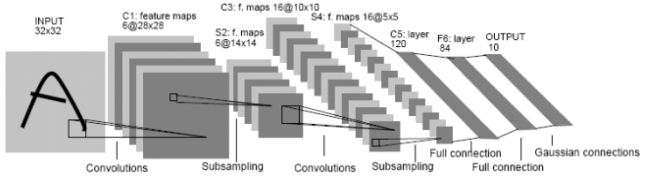
Nearest neighbor



10^6 examples

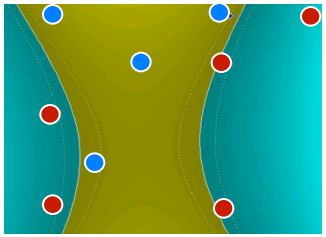
Shakhnarovich, Viola, Darrell 2003
Berg, Berg, Malik 2005...

Neural networks



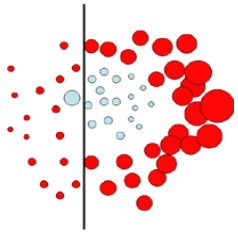
LeCun, Bottou, Bengio, Haffner 1998
Rowley, Baluja, Kanade 1998
…

Support Vector Machines



Guyon, Vapnik
Heisele, Serre, Poggio, 2001,…

Boosting



Viola, Jones 2001,
Torralba et al. 2004,
Opelt et al. 2006,…

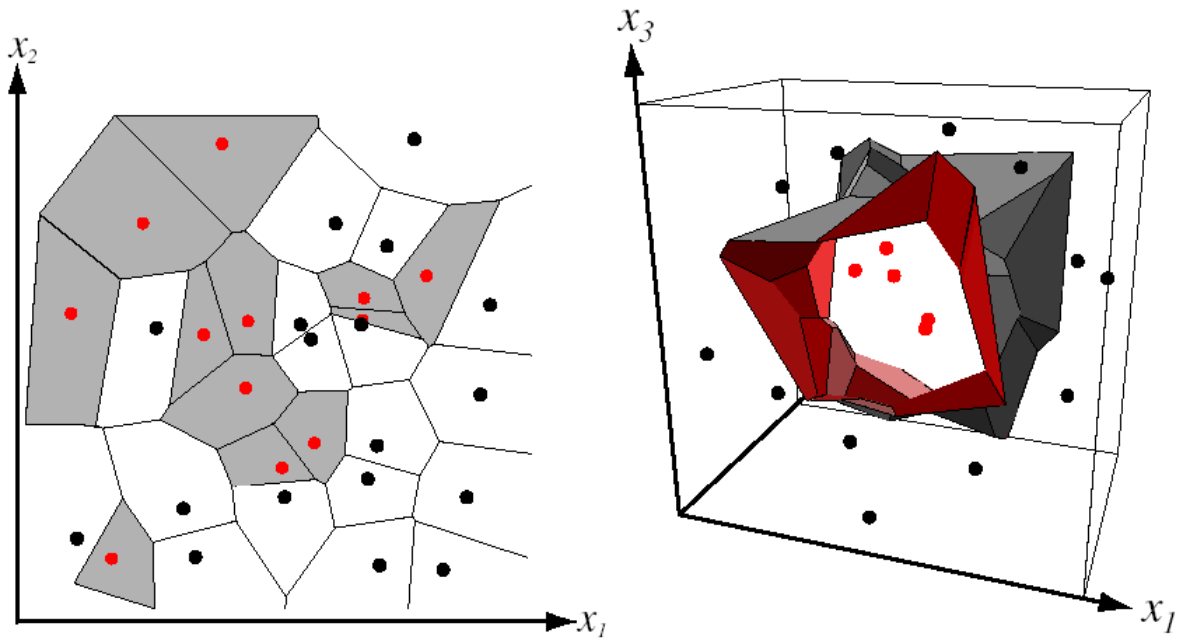
●●●

Slide credit: Kristen Grauman

Slide adapted from Antonio Torralba

Nearest Neighbor Classifier

- Assign label of nearest training data point to each test data point

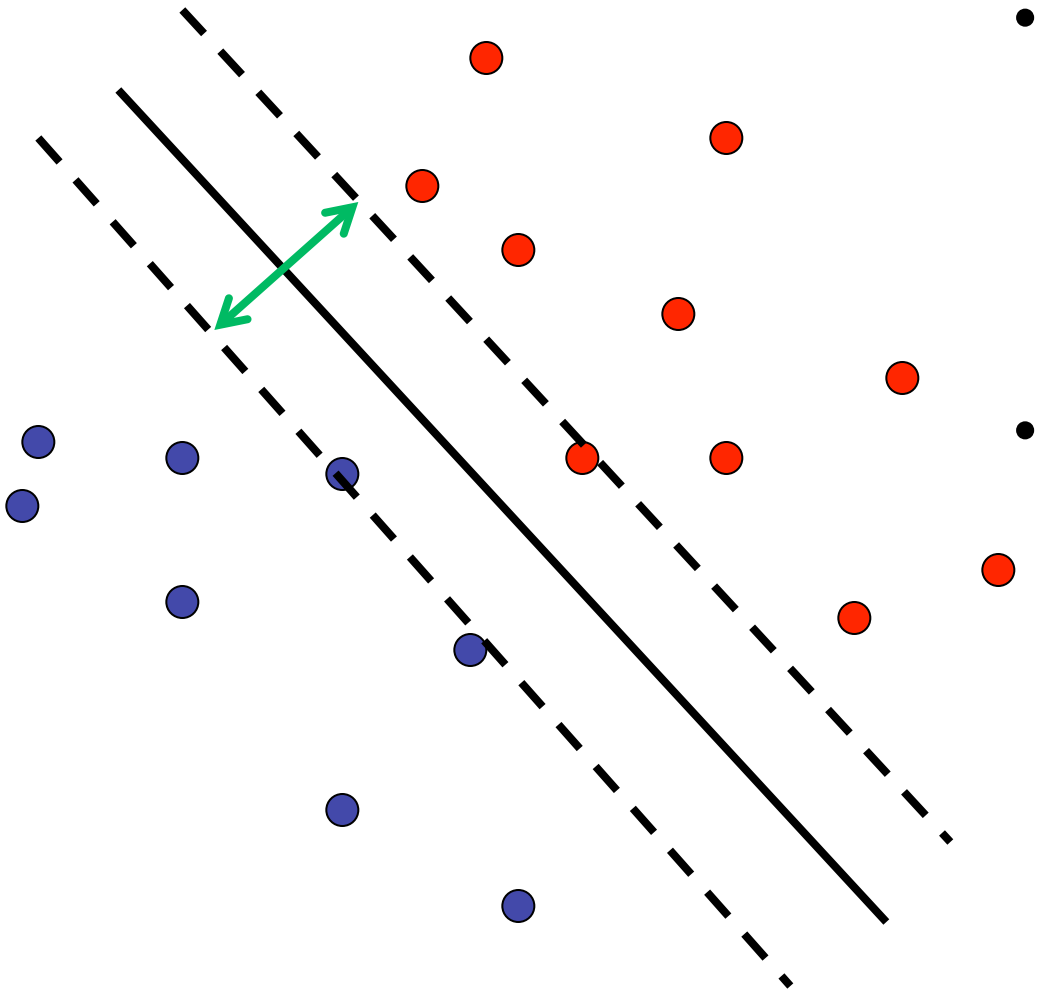


Voronoi partitioning of feature space
for 2-category 2-D and 3-D data

Slide credit: David Lowe

Source: D. Lowe

Support Vector Machines (SVMs)



- Discriminative classifier based on *optimal separating hyperplane (i.e. line for 2D case)*
- Maximize the *margin* between the positive and negative training examples

Slide credit: Kristen Grauman

Boosting Methods

- Are discussed with detection

Neural Networks

- Are not discussed here

Slide credit: Svetlana Lazebnik

BoW for Object Categorization (classification)

Caltech6 dataset







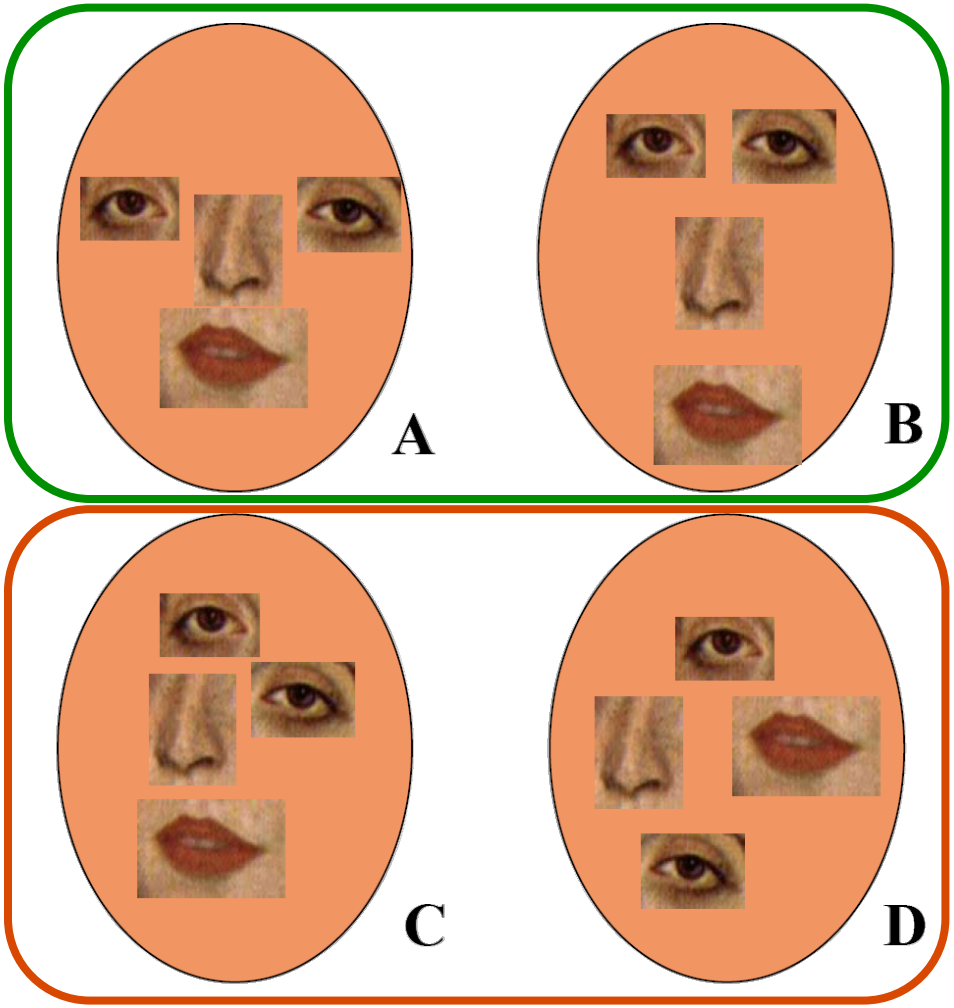

class	bag of words	bag of words	Parts-and-shape model
	Zhang et al. (2005)	Willamowski et al. (2004)	Fergus et al. (2003)
airplanes	**98.8**	97.1	90.2
cars (rear)	98.3	**98.6**	90.3
cars (side)	**95.0**	87.3	88.5
faces	**100**	99.3	96.4
motorbikes	**98.5**	98.0	92.5
spotted cats	**97.0**	—	90.0

- **Good performance for pure classification (object present/absent)**
  - **Better than more elaborate part-based models with spatial constraints…**

Slide credit: Svetlana Lazebnik

Limitations of BoW Representations

- The bag of words removes spatial layout.
- This is both a strength and a weakness.
- *Why a strength?*
- *Why a weakness?*



Slide adapted from Bill Freeman

BoW Representation: Spatial Information

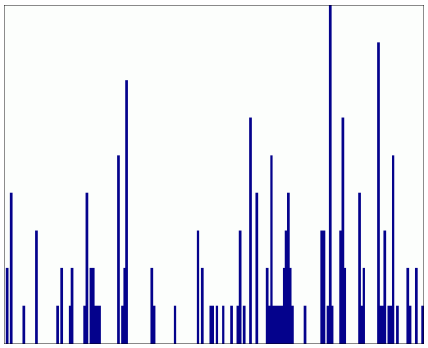
- A bag of words is an *orderless* representation. i.e. fully discarding spatial relationships between features
- Middle ground:
  - Visual "phrases" : frequently co-occurring words (e.g. using `frequent itemsets')
  - Semi-local features : describe configuration, neighborhood
  - Let position be part of each feature descriptor
  - *Count bags of words only within sub-grids of an image (e.g. hierarchically decomposing the image in parts, `pyramids')*

Slide credit: Kristen Grauman

Spatial Pyramid Representation

- Representation between orderless BoW and configuration



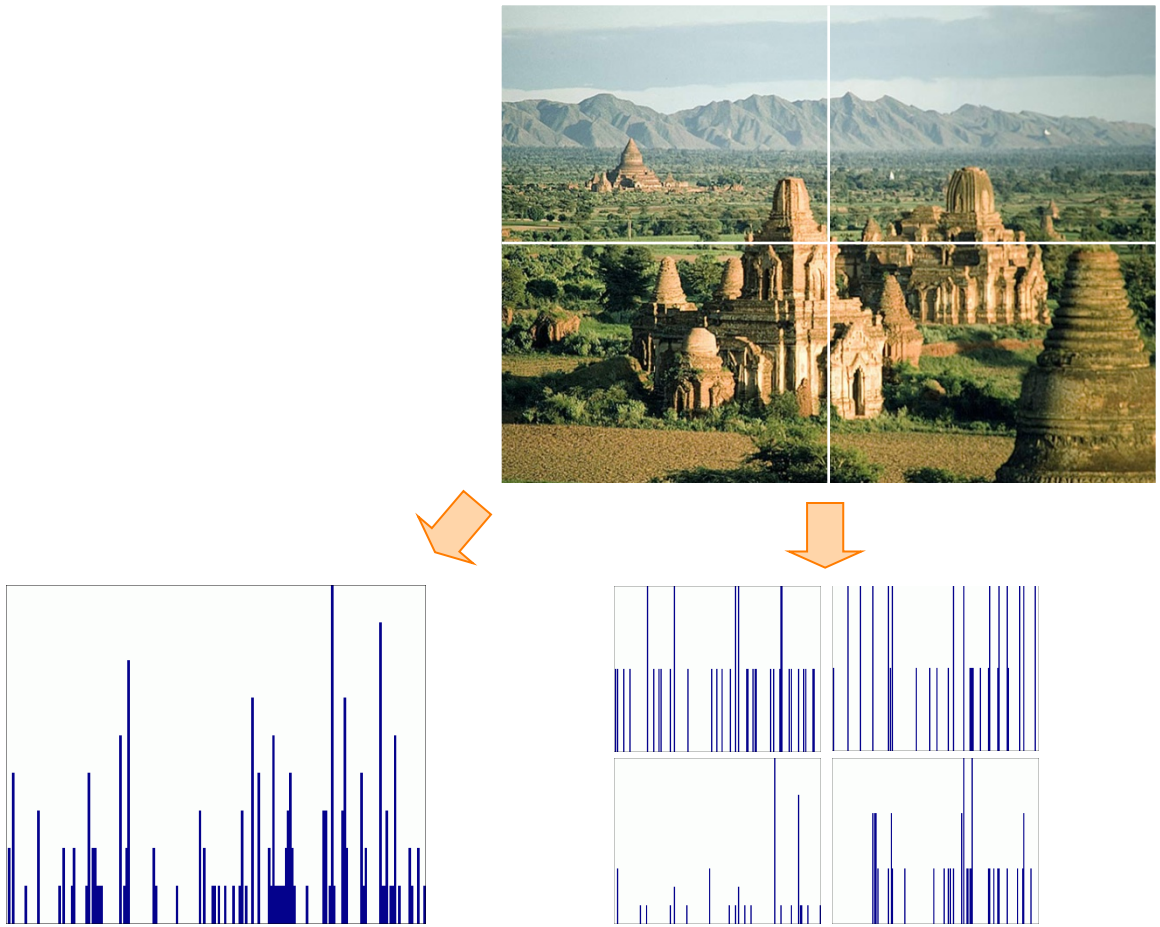


Slide credit: Svetlana Lazebnik

[Lazebnik, Schmid & Ponce, CVPR'06]

Spatial Pyramid Representation

- Representation between orderless BoW and configuration

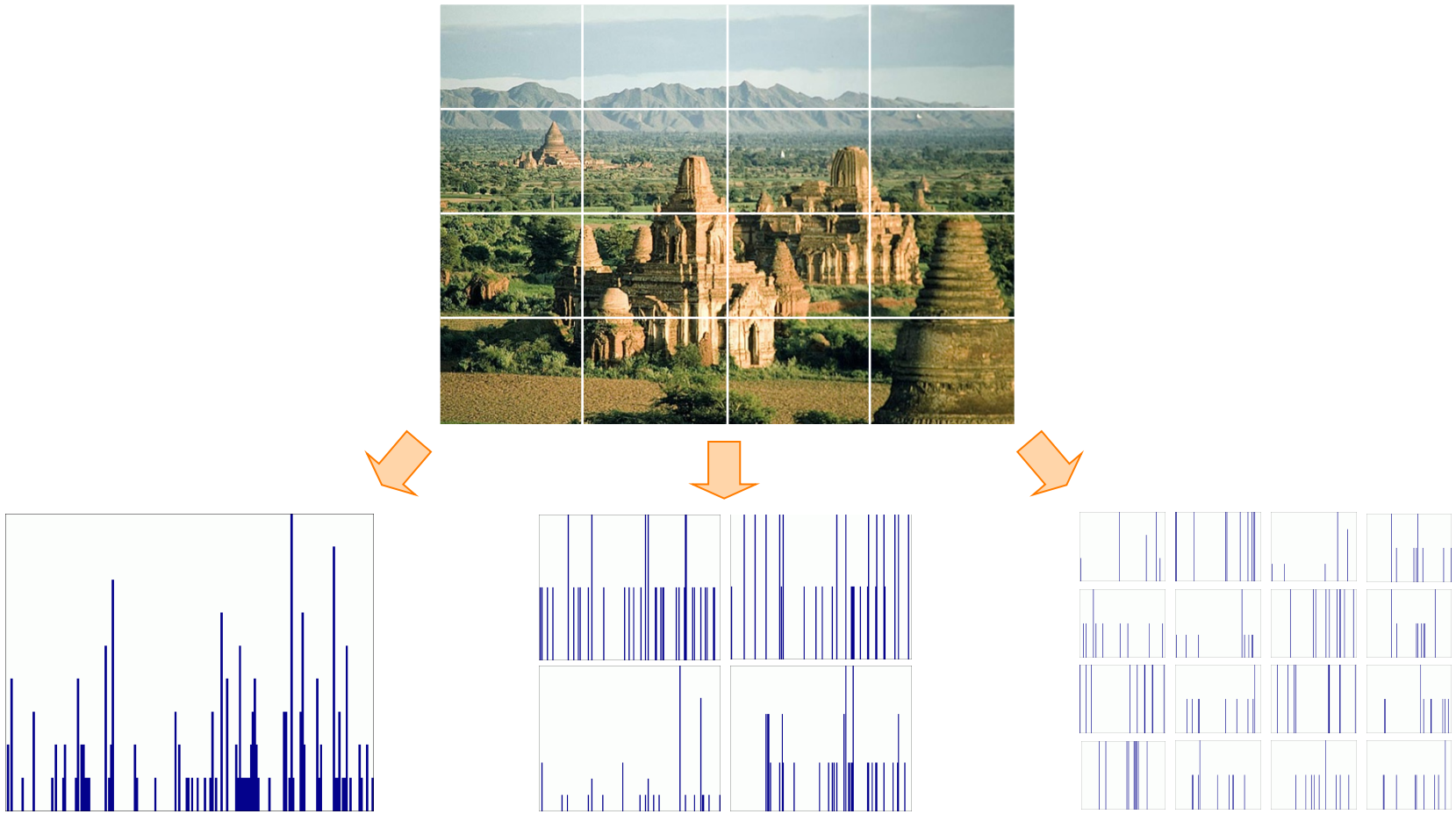


Slide credit: Svetlana Lazebnik

[Lazebnik, Schmid & Ponce, CVPR'06]

Spatial Pyramid Representation

- Representation between orderless BoW and configuration



Slide credit: Svetlana Lazebnik

[Lazebnik, Schmid & Ponce, CVPR'06]

Spatial Pyramid description matching

- Can capture scene categories well



Spatial Pyramid description matching

- Can capture scene categories well---texture-like patterns but with some variability in the positions of all the local pieces.
- Sensitive to global shifts of the view



bedroom



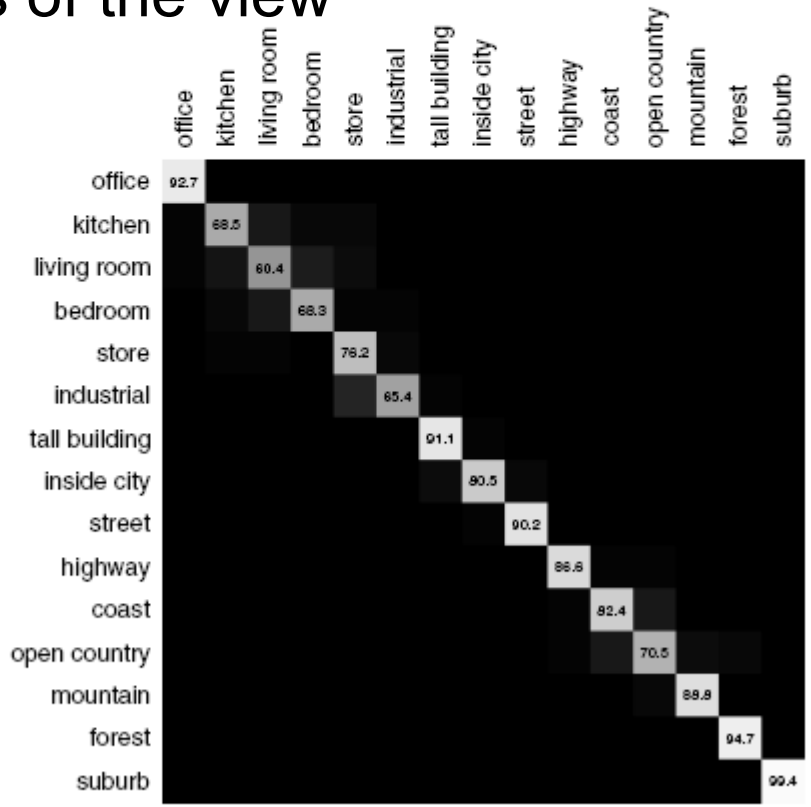
tall building*



highway*



mountain*



Confusion table



Example queries and NN with spatial pyramid matching

Discussion: Bag-of-Words

- **Pros:**
  - Flexible to geometry / deformations / viewpoint
  - Compact summary of image content
  - Provides vector representation for sets
  - Empirically good recognition results in practice

- **Cons:**
  - Basic model ignores geometry – can verify afterwards, or embed within feature descriptors
  - Background and foreground mixed when bag covers whole image
  - Interest points or sampling: no guarantee to capture object-level parts
  - Optimal vocabulary formation remains unclear

Slide credit: Kristen Grauman

3. Sliding Windows

Plan

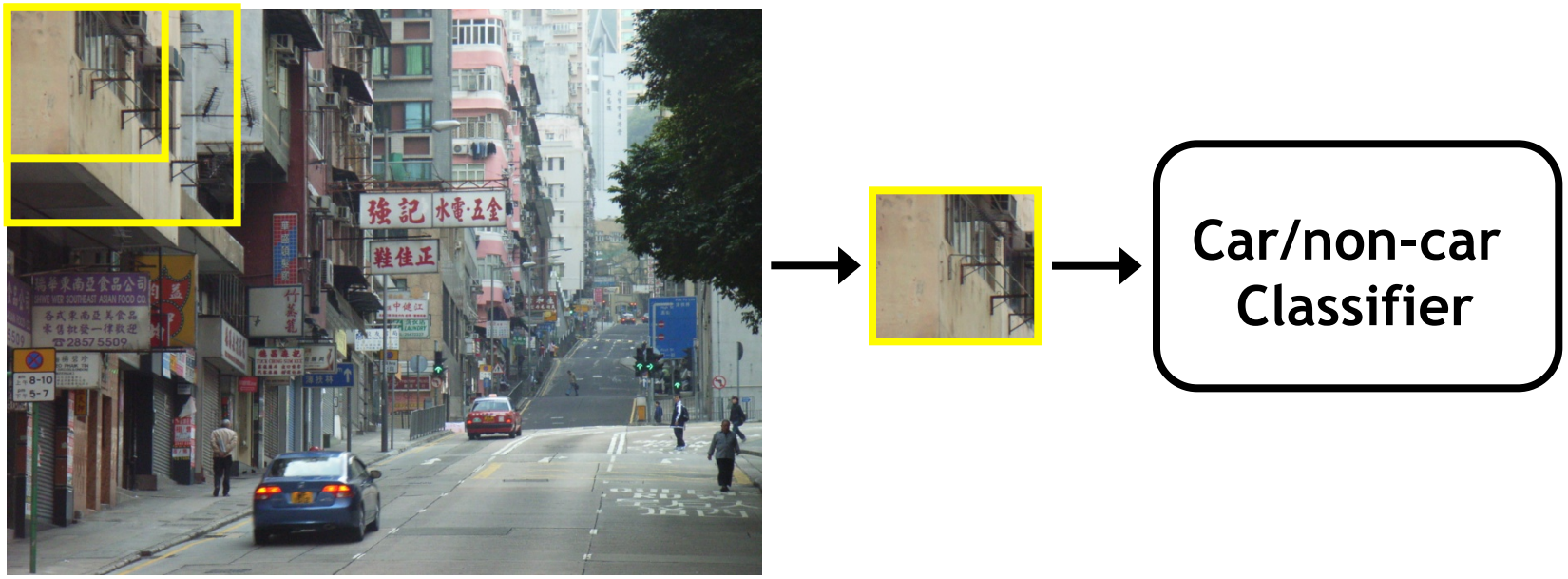
1. Problem definition, classification vs. detection

2. bag-of-words
3. *sliding-windows* (with global or part-based models)
4. Hough approaches (implicit shape model & forests)

5. Open challenges

Detection via Classification: Main Idea

- If object may be in a cluttered scene, slide a window around looking for it.



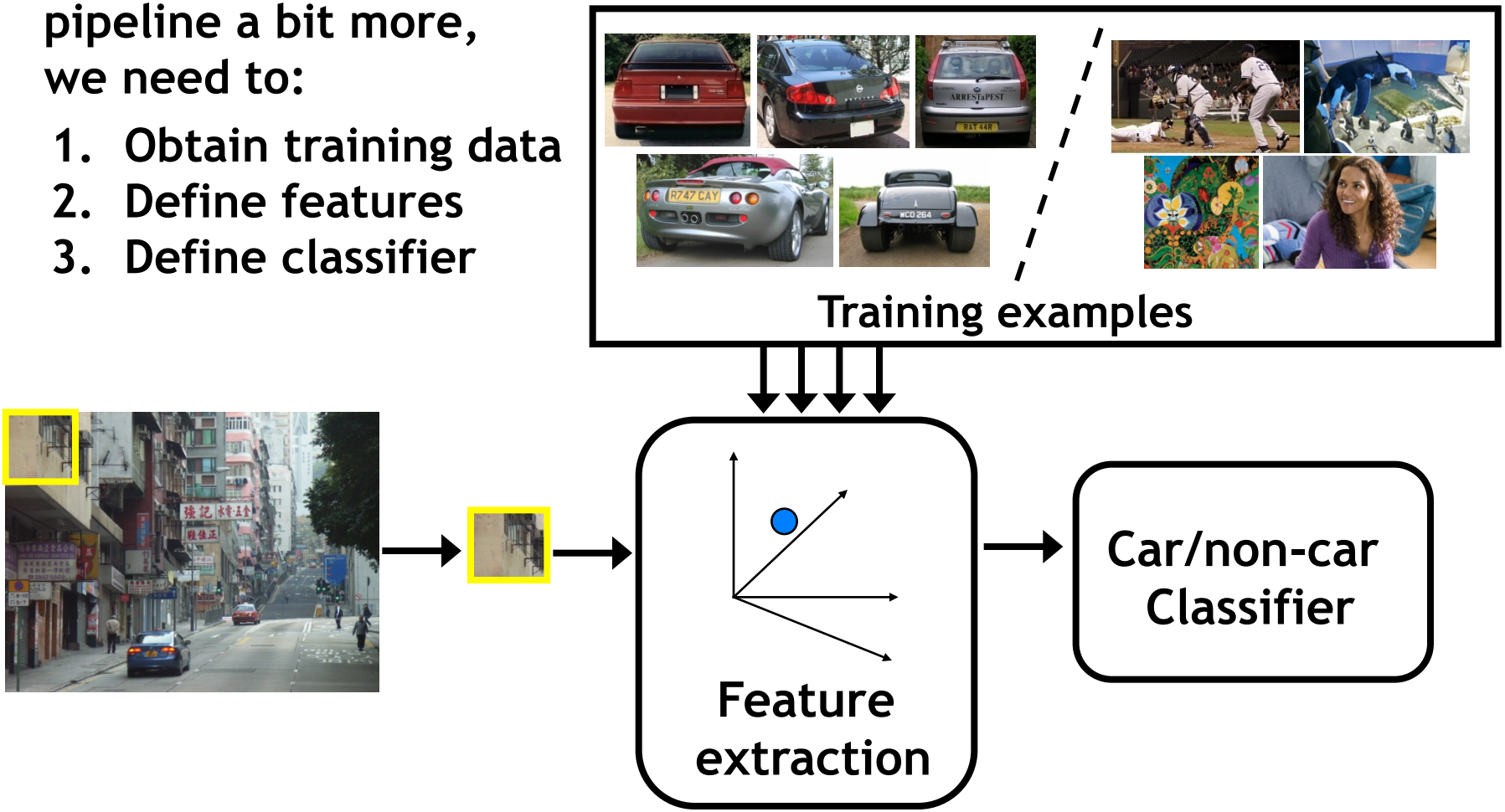
- a brute-force approach with many local decisions.

Slide adapted from Kristen Grauman

Detection via Classification: Main Idea

Fleshing out this pipeline a bit more, we need to:

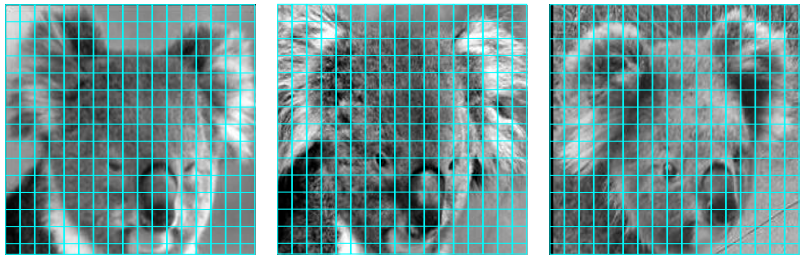
1. Obtain training data
2. Define features
3. Define classifier

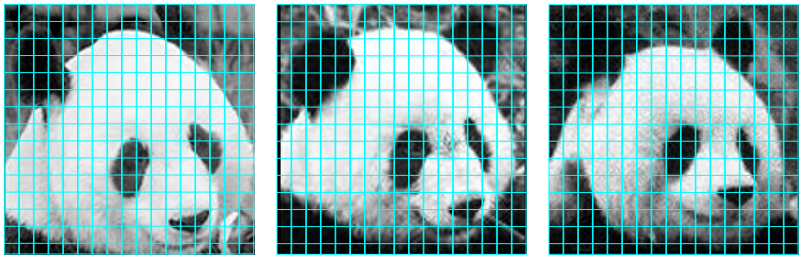


Slide credit: Kristen Grauman

Feature Extraction: Global Appearance

- Pixel-based representations sensitive to small shifts





- Color or grayscale-based appearance description can be sensitive to illumination and intra-class appearance variation

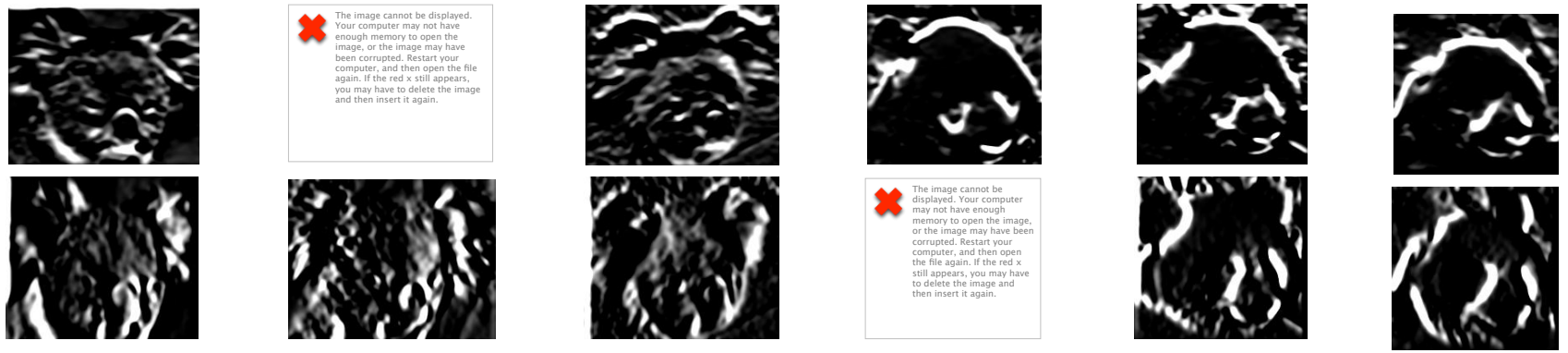


Cartoon example:
an albino koala

Slide credit: Kristen Grauman

Gradient-based Representations

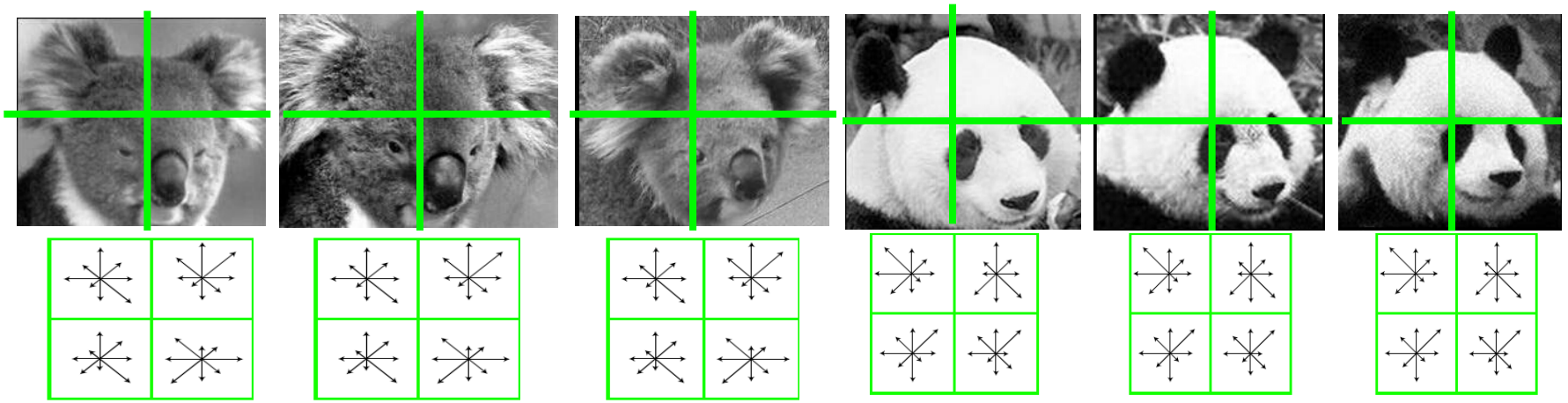
- Consider edges and (oriented) intensity gradients



Slide credit: Kristen Grauman

Gradient-based Representations

- Consider edges and (oriented) intensity gradients



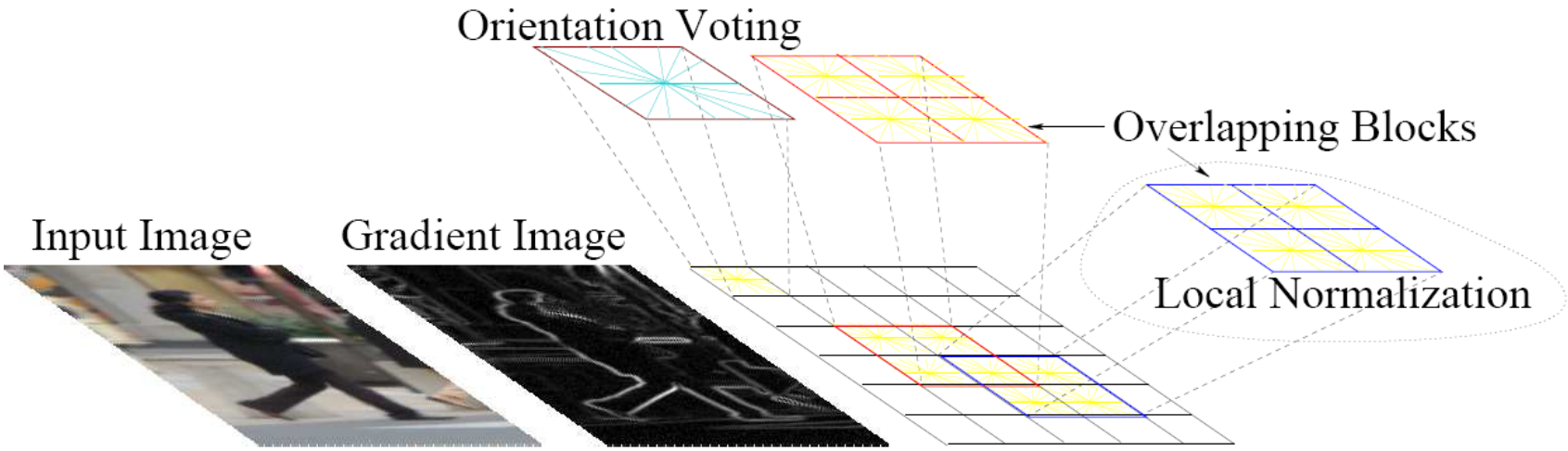
- Summarize local distribution of gradients with histogram
  - Locally orderless: offers invariance to small shifts and rotations
  - Contrast-normalization: try to correct for variable illumination

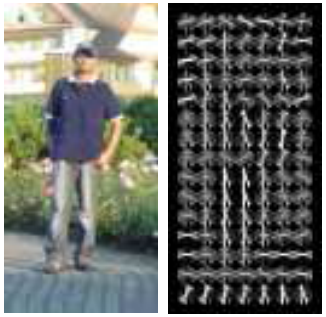
Slide credit: Kristen Grauman

3. Sliding Windows

holistic models

Gradient-based Representations: Histograms of Oriented Gradients (HoG)





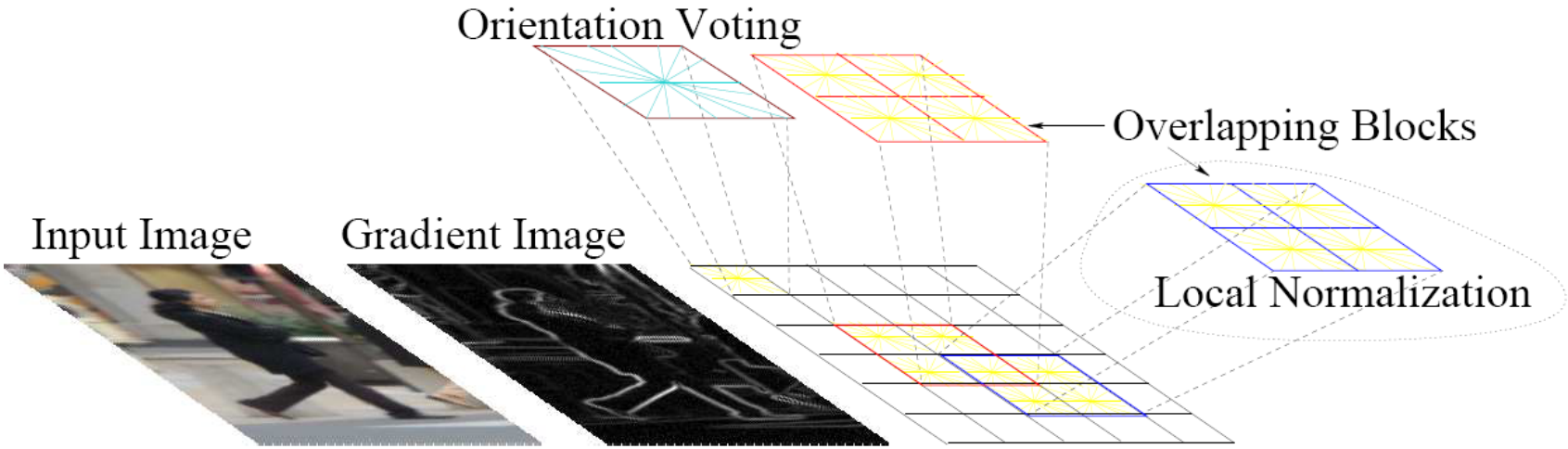
Map each grid cell in the input window to a histogram counting the gradients per orientation.

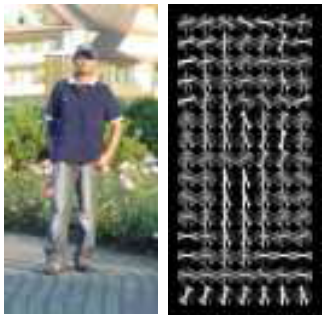
Code available:
http://pascal.inrialpes.fr/soft/olt/

Slide credit: Kristen Grauman

[Dalal & Triggs, CVPR 2005]

Example: Pedestrian Detection with HoG and SVMs





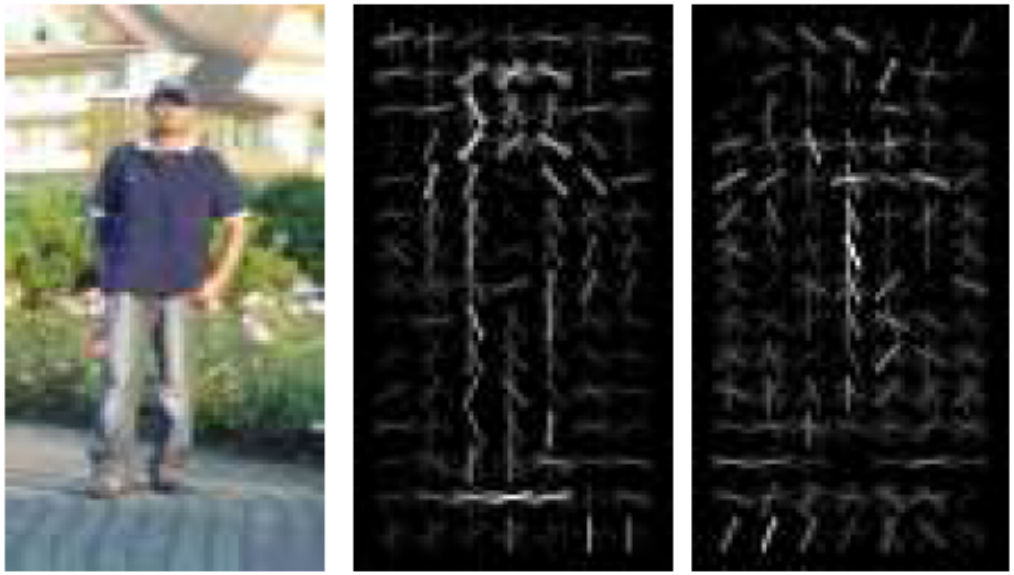
- Map each grid cell in the input window to a histogram counting the gradients per orientation.
- Train a linear SVM using training set of pedestrian vs. non-pedestrian windows.

Slide adapted from: Kristen Grauman

[Dalal & Triggs, CVPR 2005]

Pedestrian Detection

- Detecting standing and walking humans using sliding window's



frontal sideways

SVM with HoGs [Dalal & Triggs, CVPR 2005]

Slide credit: Kristen Grauman

HoG pedestrian detector

Finally, a decision needs to be made for each window…

example result:



Classifier Choices

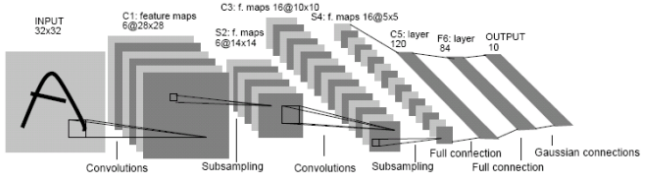
Nearest neighbor



10^6 examples

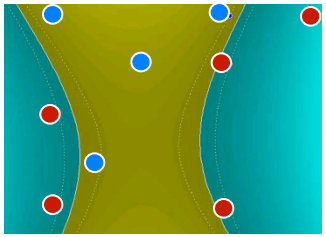
Shakhnarovich, Viola, Darrell 2003
Berg, Berg, Malik 2005...

Neural networks



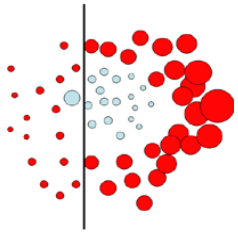
LeCun, Bottou, Bengio, Haffner 1998
Rowley, Baluja, Kanade 1998
…

Support Vector Machines



Guyon, Vapnik
Heisele, Serre, Poggio, 2001,…

Boosting



Viola, Jones 2001,
Torralba et al. 2004,
Opelt et al. 2006,…

●●●

Slide adapted from Antonio Torralba

Boosting

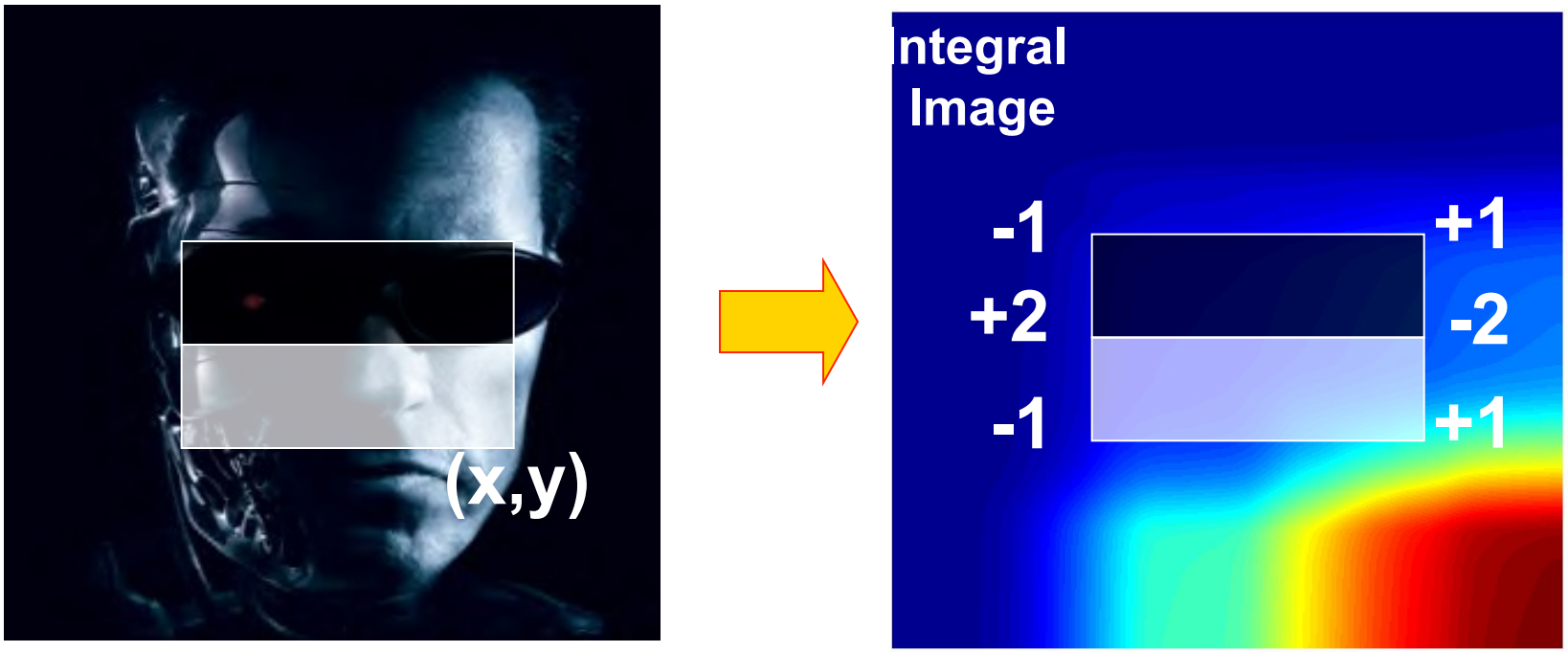
- Build a strong classifier by combining many "weak classifiers", which need only be better than chance
- Sequential learning process: at each iteration, add a weak classifier
- Flexible to choice of weak learner
  - including fast simple classifiers that alone may be inaccurate

- We'll look at Freund & Schapire's AdaBoost algorithm
  - Easy to implement
  - Base learning algorithm for Viola-Jones face detector

Y. Freund and R. Schapire, A short introduction to boosting, *Journal of Japanese Society for Artificial Intelligence*, 14(5):771-780, 1999.

Slide adapted from Kristen Grauman

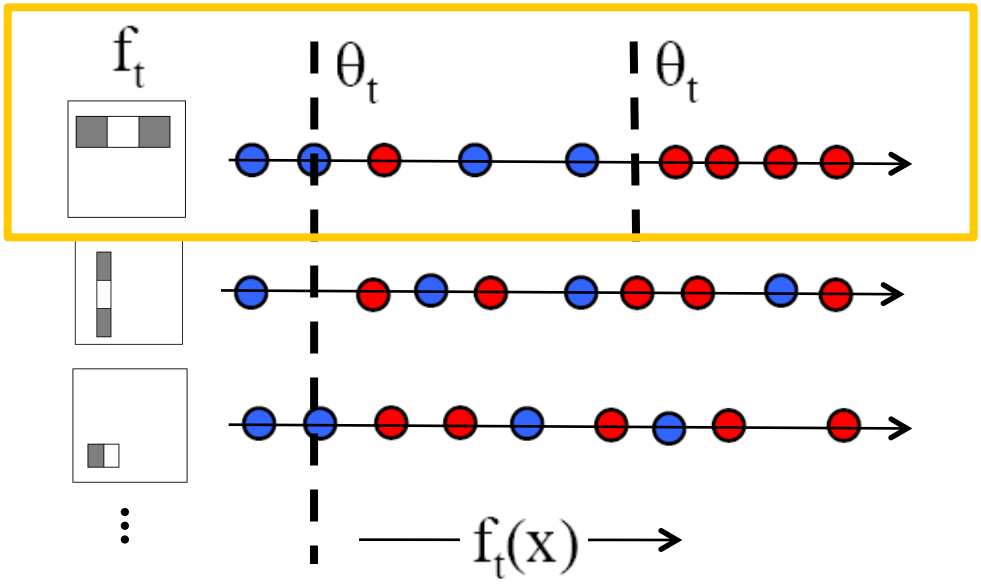
Example



Slide credit: Svetlana Lazebnik

AdaBoost for Feature+Classifier Selection

- Want to select the single feature and threshold that best separates positive (faces) and negative (non-faces) training examples, in terms of *weighted* error.



Outputs of a possible rectangle feature on faces and non-faces.

Resulting weak classifier:



$$h_t(x) = \begin{cases} +1 & \text{if } f_t(x) > \theta_t \\ -1 & \text{otherwise} \end{cases}$$

For next round, reweight the examples according to errors, choose another filter/threshold combo.

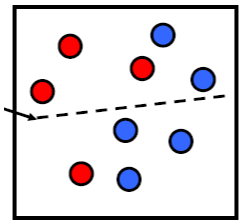
Slide credit: Kristen Grauman

[Viola & Jones, CVPR 2001]

AdaBoost Algorithm

- Given example images $(x_1, y_1), \ldots, (x_n, y_n)$ where $y_i = 0, 1$ for negative and positive examples respectively.
- Initialize weights $w_{1,i} = \frac{1}{2m}, \frac{1}{2l}$ for $y_i = 0, 1$ respectively, where m and l are the number of negatives and positives respectively.
- For $t = 1, \ldots, T$:
  1. Normalize the weights,
  $$w_{t,i} \leftarrow \frac{w_{t,i}}{\sum_{j=1}^{n} w_{t,j}}$$
  so that w_t is a probability distribution.
  2. For each feature, j, train a classifier h_j which is restricted to using a single feature. The error is evaluated with respect to w_t, $\epsilon_j = \sum_i w_i |h_j(x_i) - y_i|$.
  3. Choose the classifier, h_t, with the lowest error ϵ_t.
  4. Update the weights:
  $$w_{t+1,i} = w_{t,i} \beta_t^{1-e_i}$$
  where $e_i = 0$ if example x_i is classified correctly, $e_i = 1$ otherwise, and $\beta_t = \frac{\epsilon_t}{1-\epsilon_t}$.
- The final strong classifier is:
$$h(x) = \begin{cases} 1 & \sum_{t=1}^{T} \alpha_t h_t(x) \geq \frac{1}{2} \sum_{t=1}^{T} \alpha_t \\ 0 & \text{otherwise} \end{cases}$$
where $\alpha_t = \log \frac{1}{\beta_t}$

Start with uniform weights on training examples



$\{x_1, \ldots x_n\}$

For T rounds

Evaluate *weighted* error for each feature, pick best.

Re-weight the examples:
incorrectly classified ⇨ more weight
Correctly classified ⇨ less weight

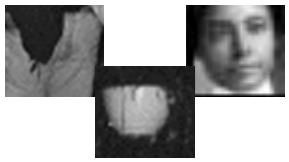
Final classifier is combination of the weak ones, weighted according to the error they had.

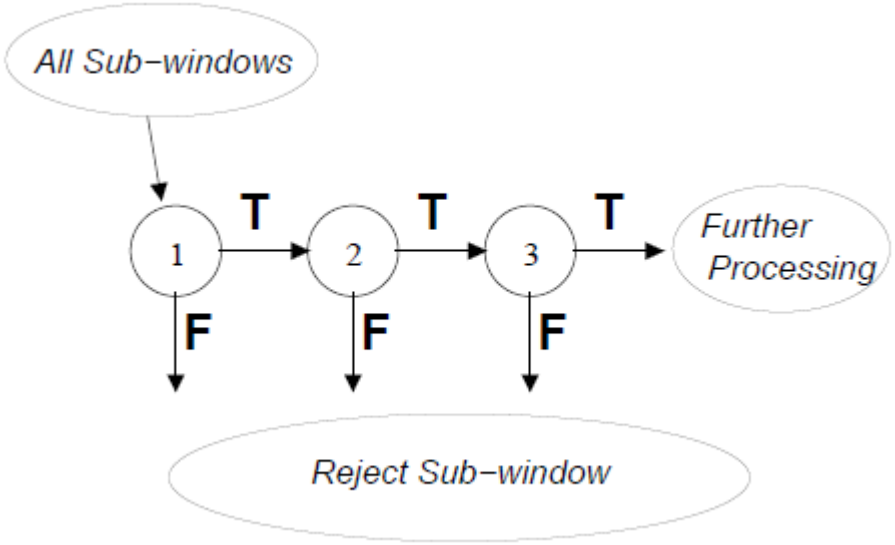
[Freund & Schapire 1995]

Efficiency Considerations

- Even if the filters are fast to compute, each new image has a lot of possible windows to search.

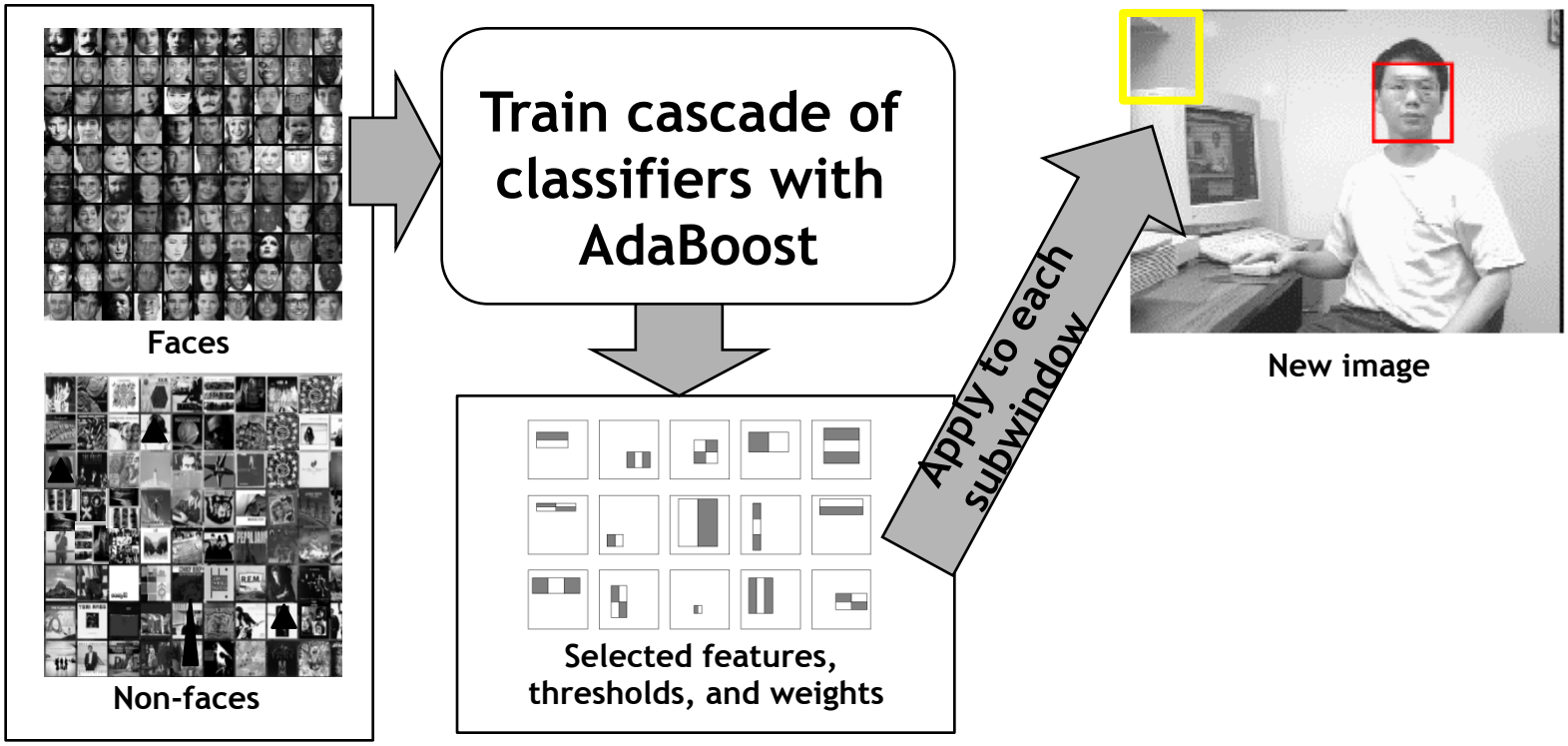
- How to make the detection more efficient?
  → cascade of classifiers of increasing complexity





Slide adapted from Kristen Grauman

Viola-Jones Face Detector: Summary

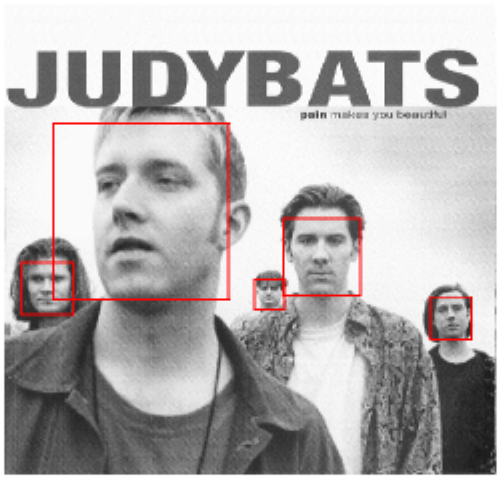


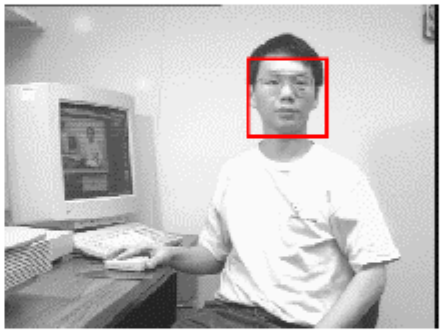
- Trained with 5K positives, 350M negatives
- Real-time detector using 38 layer cascade
- 6061 features in final layer
- [Implementation available in OpenCV: http://sourceforge.net/projects/opencvlibrary/]

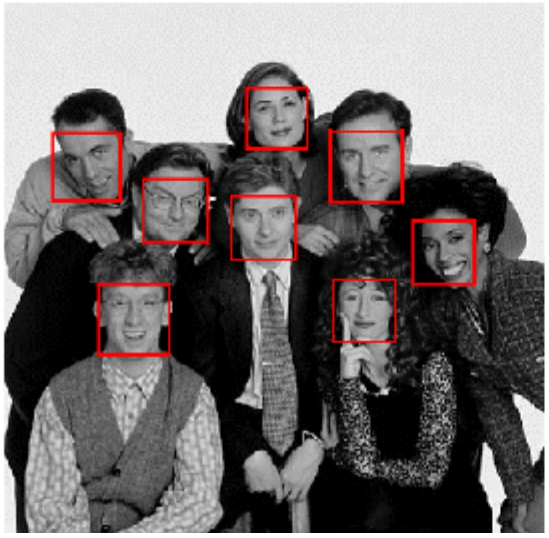
Slide credit: Kristen Grauman

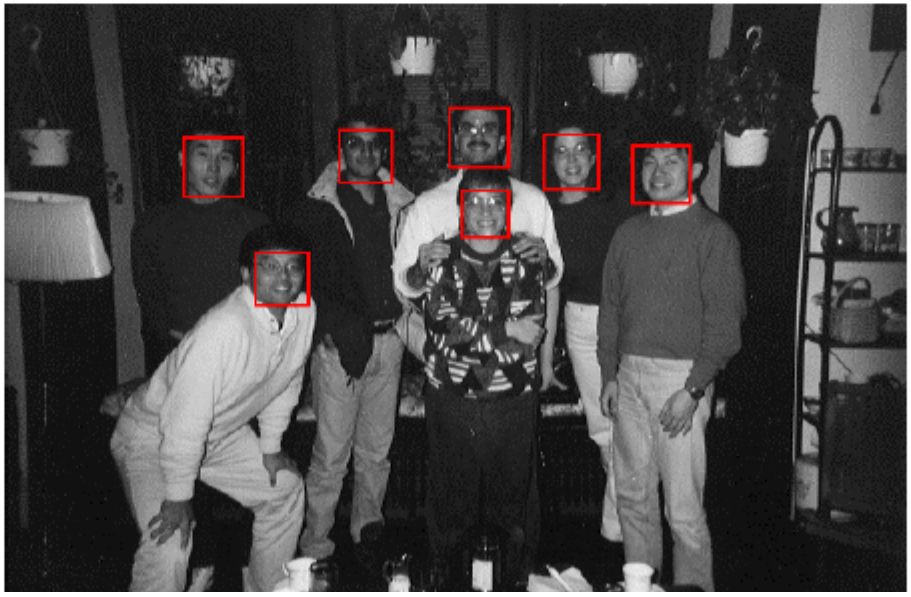
Viola-Jones Face Detector: Results





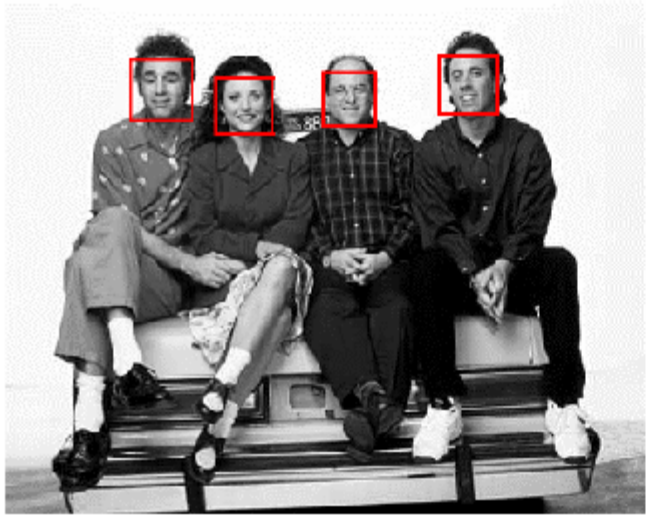


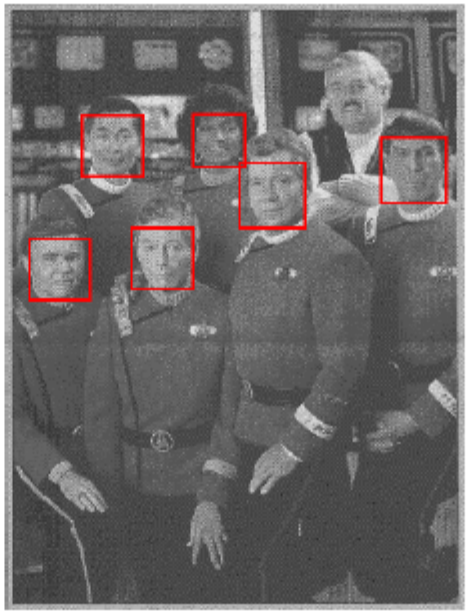


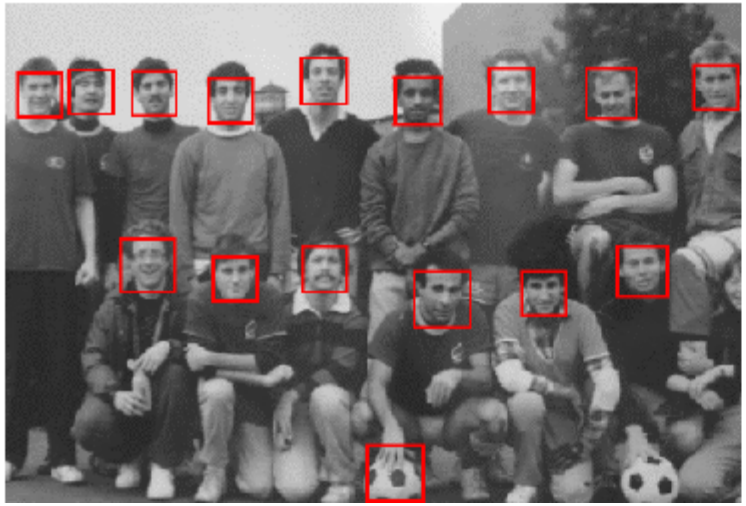


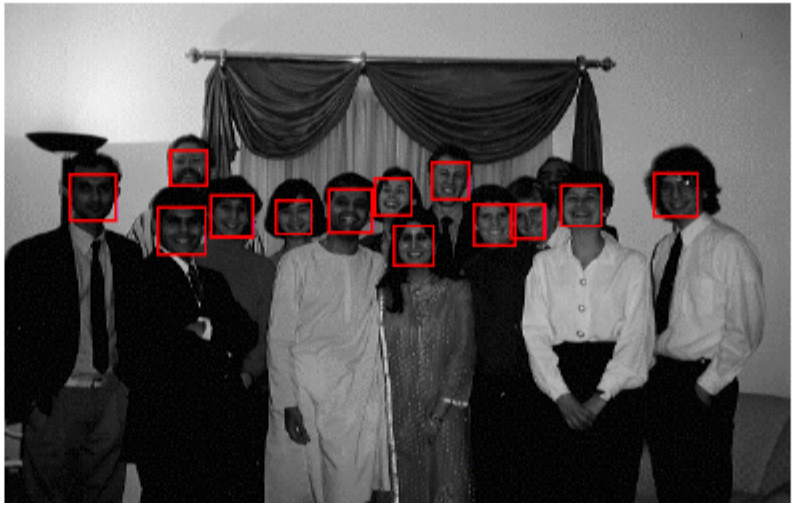
Slide credit: Kristen Grauman

Viola-Jones Face Detector: Results







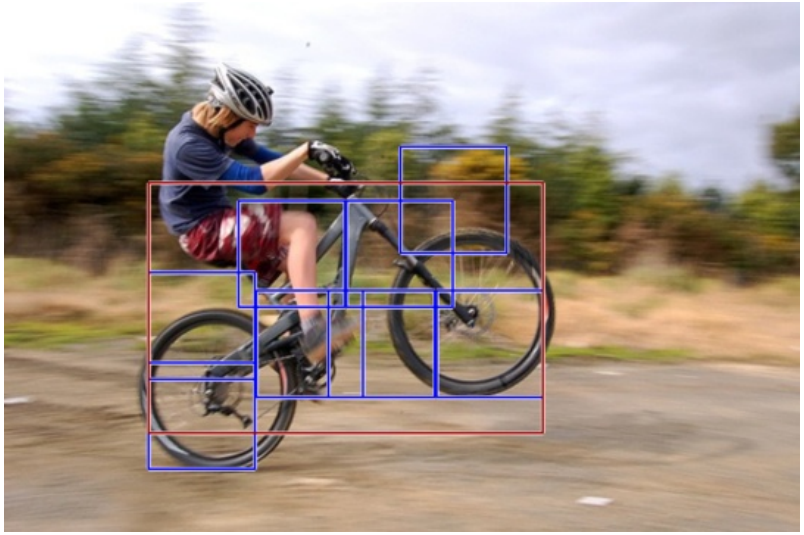


Slide credit: Kristen Grauman

3. Sliding Windows

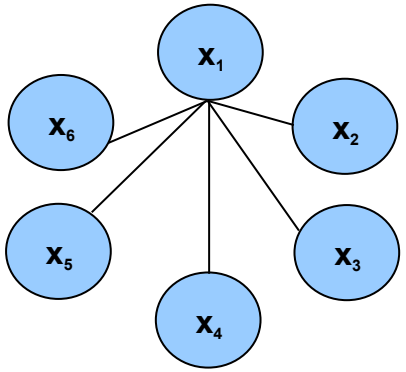
star models

Deformable Part Model



- Sliding Window Approach

- Based on the star model: all parts x are positioned relative to one root node x1:

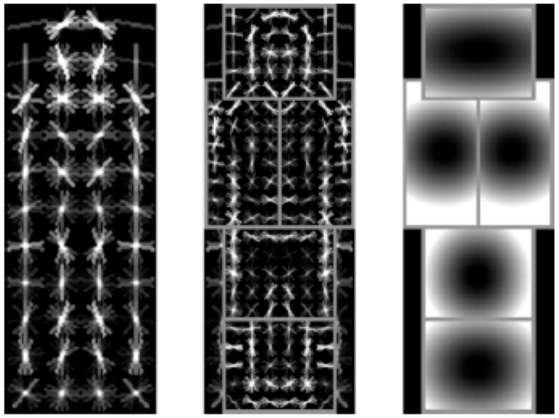


- Built on Dalal & Triggs (HOG Features + SVM)
- Goal: Train a more flexible model using parts:
  - detectors for different viewpoints: several `root nodes'
  - parts at changeable positions relative to root (deform.)

[Felzenszwalb et al., PAMI'10]

Deformable Part Model

- Person





root filters
coarse resolution

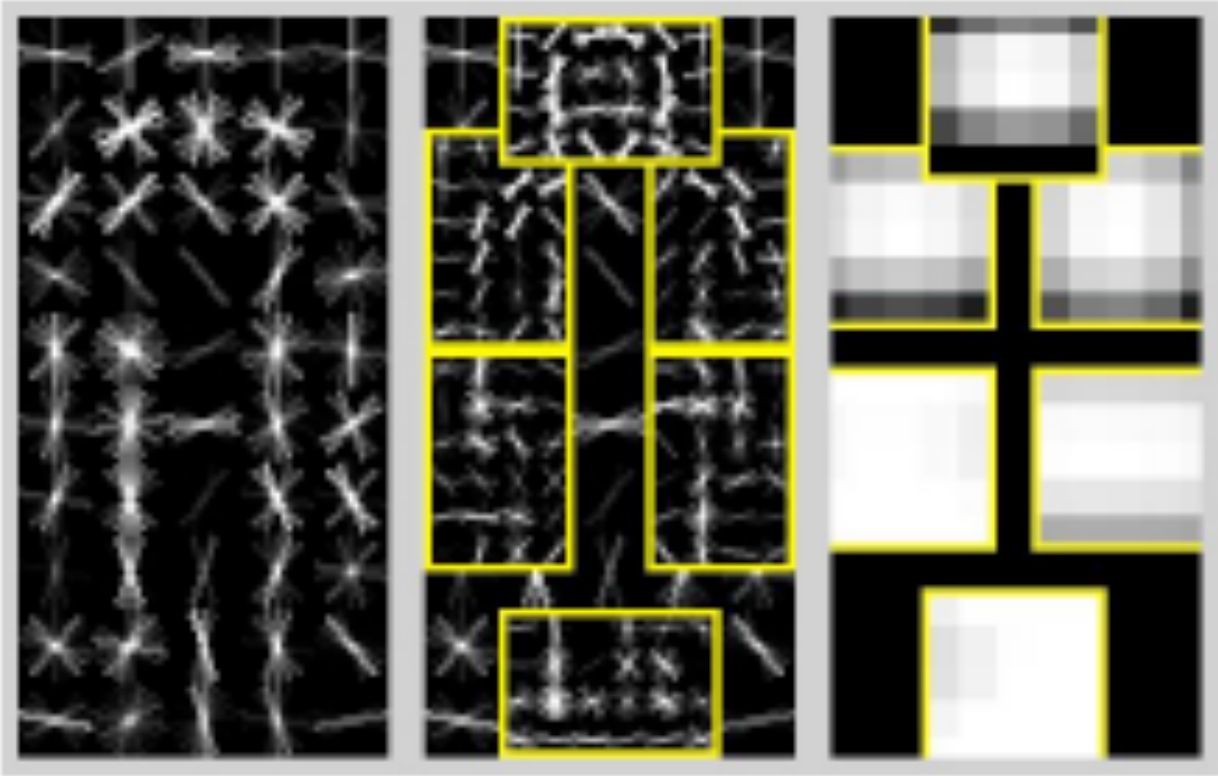
part filters
finer resolution

deformation
models

[Felzenszwalb et al., PAMI'10]

Deformable Part Model

Model learnt with `Latent SVM'

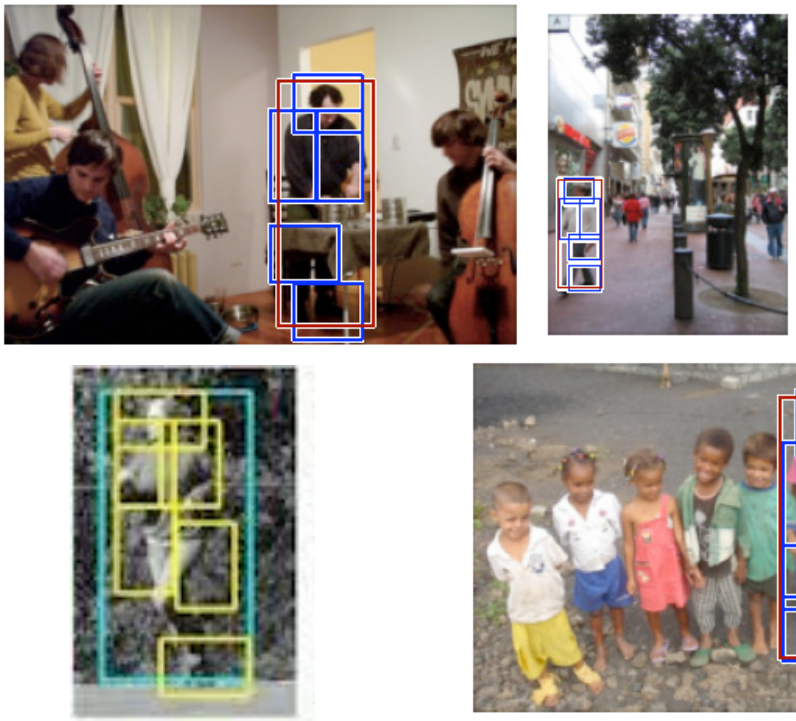


[Felzenszwalb et al., PAMI'10]

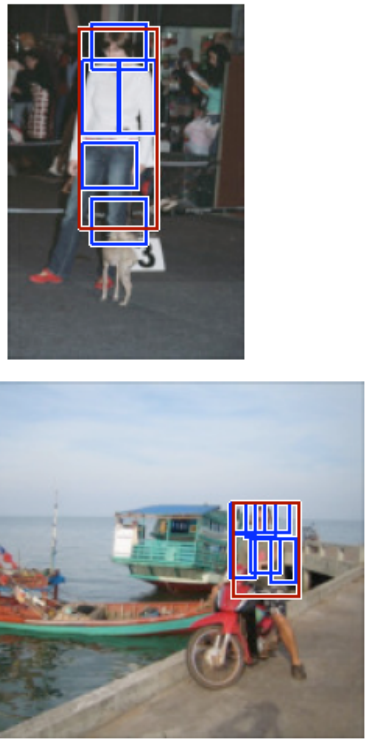
Results

- Persons

high scoring true positives



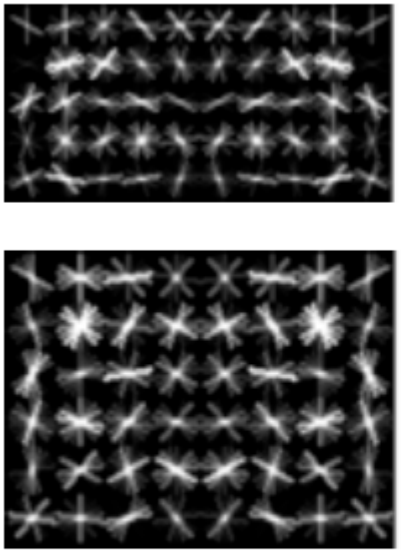
high scoring false positives
(not enough overlap)



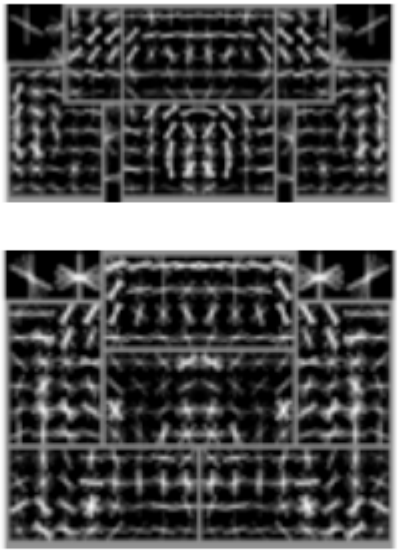
[Felzenszwalb et al., PAMI'10]

Deformable Part Model

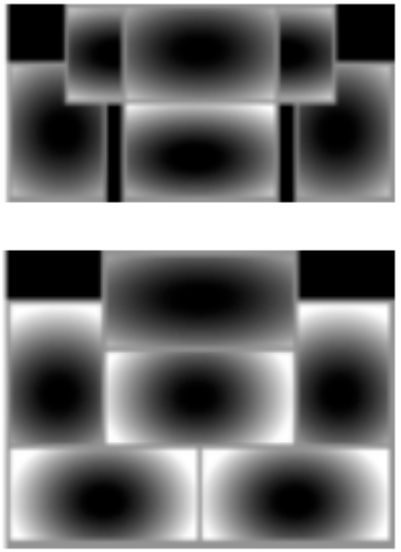
- **Car**



root filters
coarse resolution



part filters
finer resolution



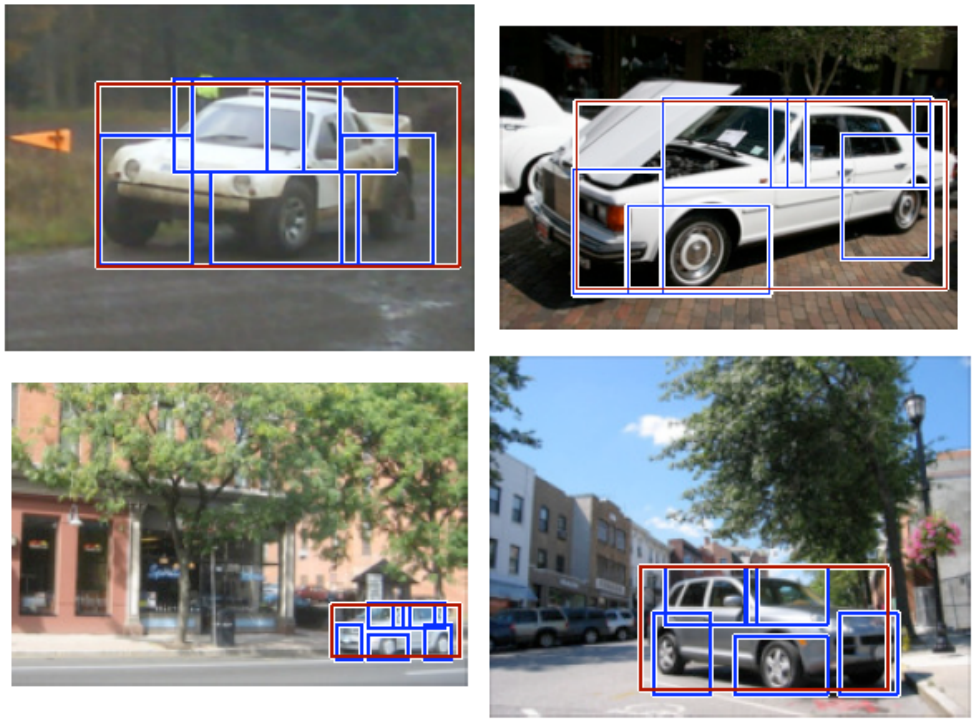
deformation
models

[Felzenszwalb et al., PAMI'10]

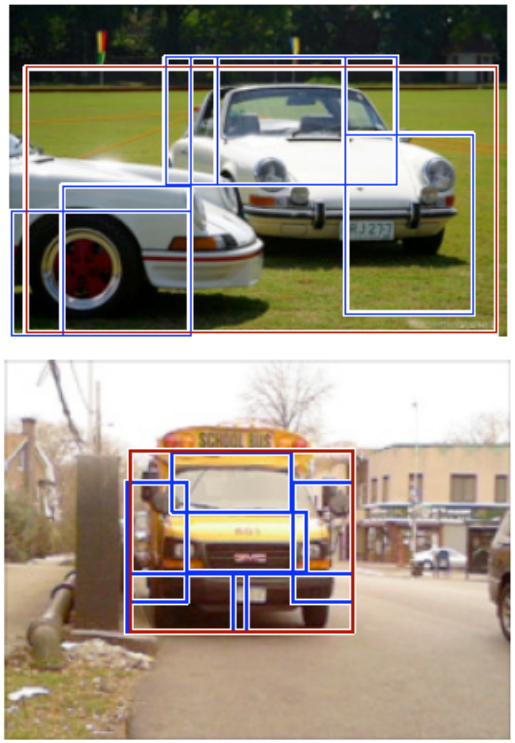
Results

- Cars

high scoring true positives



high scoring false positives

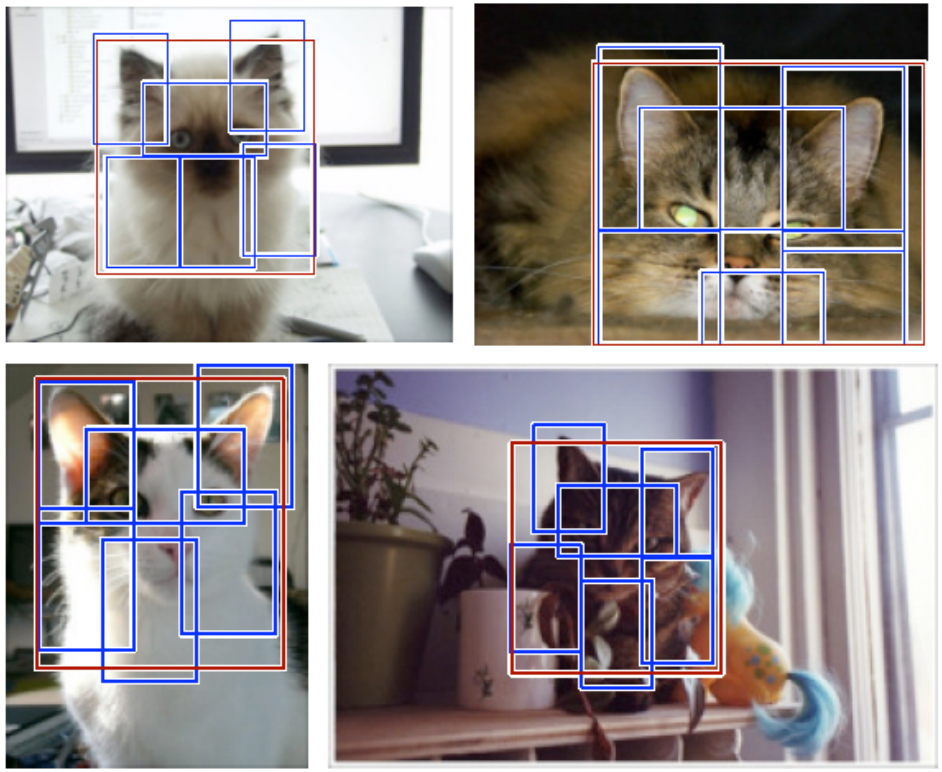


[Felzenszwalb et al., PAMI'10]

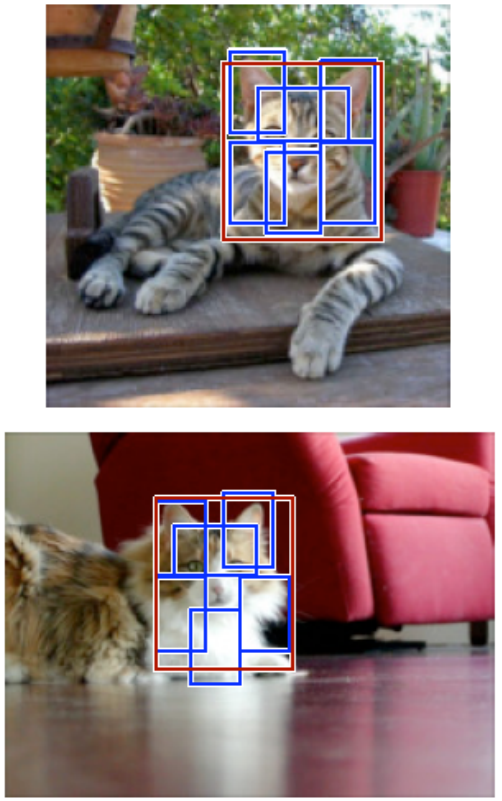
Results

- **Cats**

high scoring true positives



high scoring false positives
(not enough overlap)



[Felzenszwalb et al., PAMI'10]

4. Hough Approaches

Plan

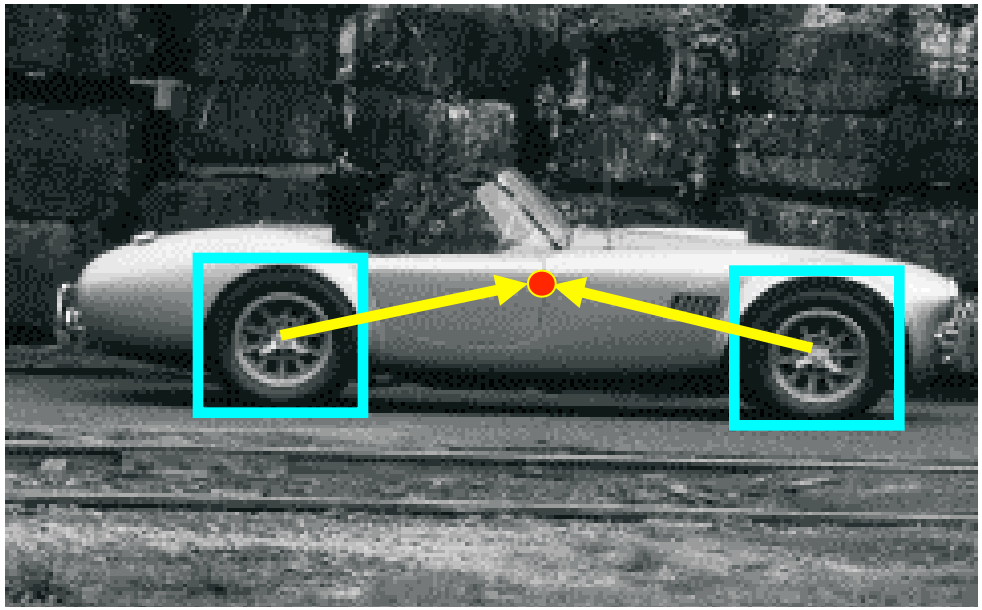
1. Problem definition, classification vs. detection

2. bag-of-words
3. sliding-windows (with global or part-based models)
4. *Hough approaches* (implicit shape model & forests)

5. Open challenges

Implicit Shape Model: Basic Idea

- Visual vocabulary is used to index votes for object position [a visual word = a “part”].



Training image

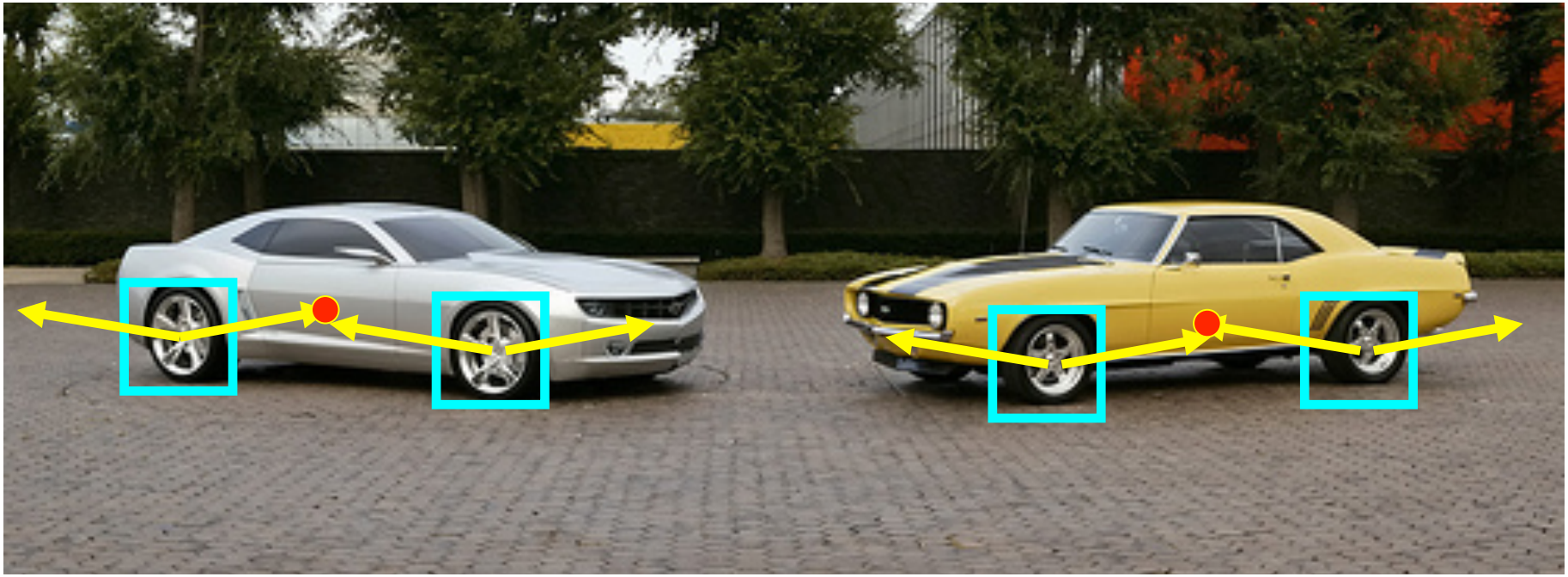


Visual codeword with displacement vectors

B. Leibe, A. Leonardis, and B. Schiele,
Robust Object Detection with Interleaved Categorization and Segmentation, International Journal of Computer Vision, Vol. 77(1-3), 2008.

Implicit Shape Model: Basic Idea

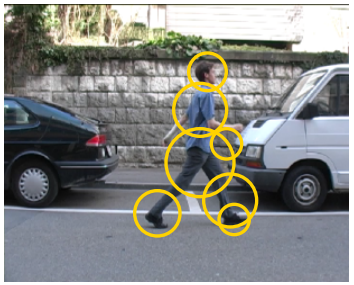
- Objects are detected as consistent configurations of the observed parts (visual words).



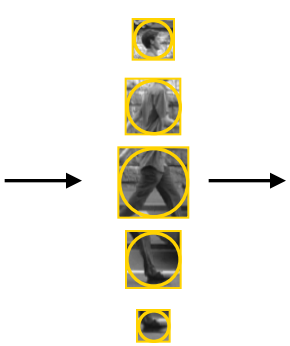
Test image

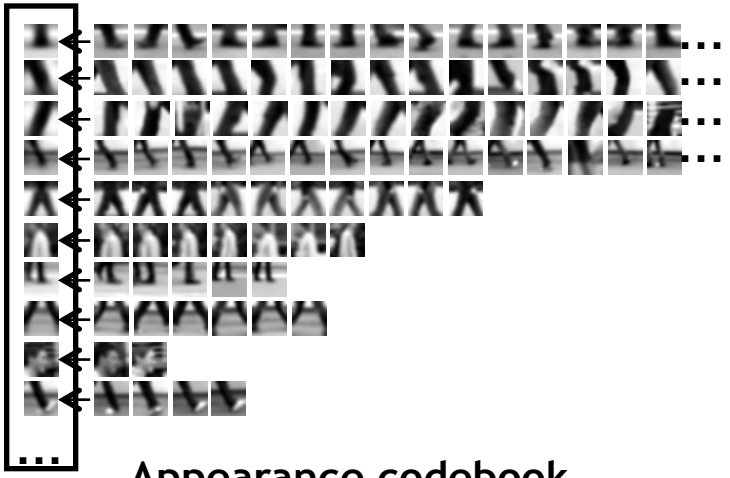
B. Leibe, A. Leonardis, and B. Schiele,
Robust Object Detection with Interleaved Categorization and Segmentation, International Journal of Computer Vision, Vol. 77(1-3), 2008.

Implicit Shape Model - Training



Training images
(+object center and scale)

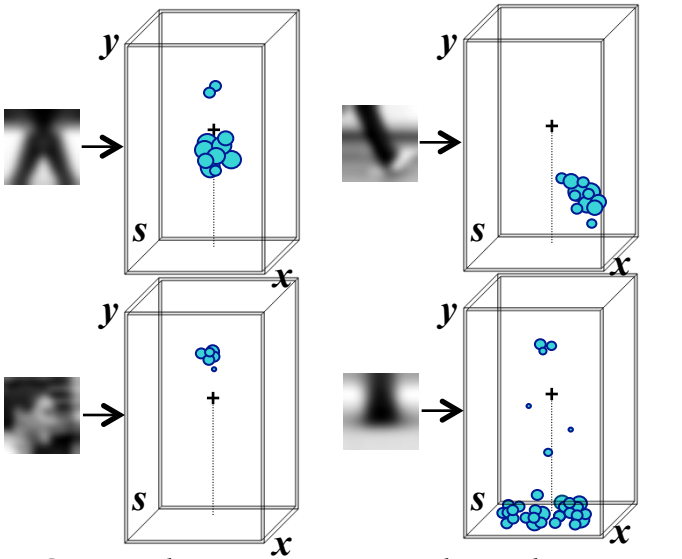




Appearance codebook

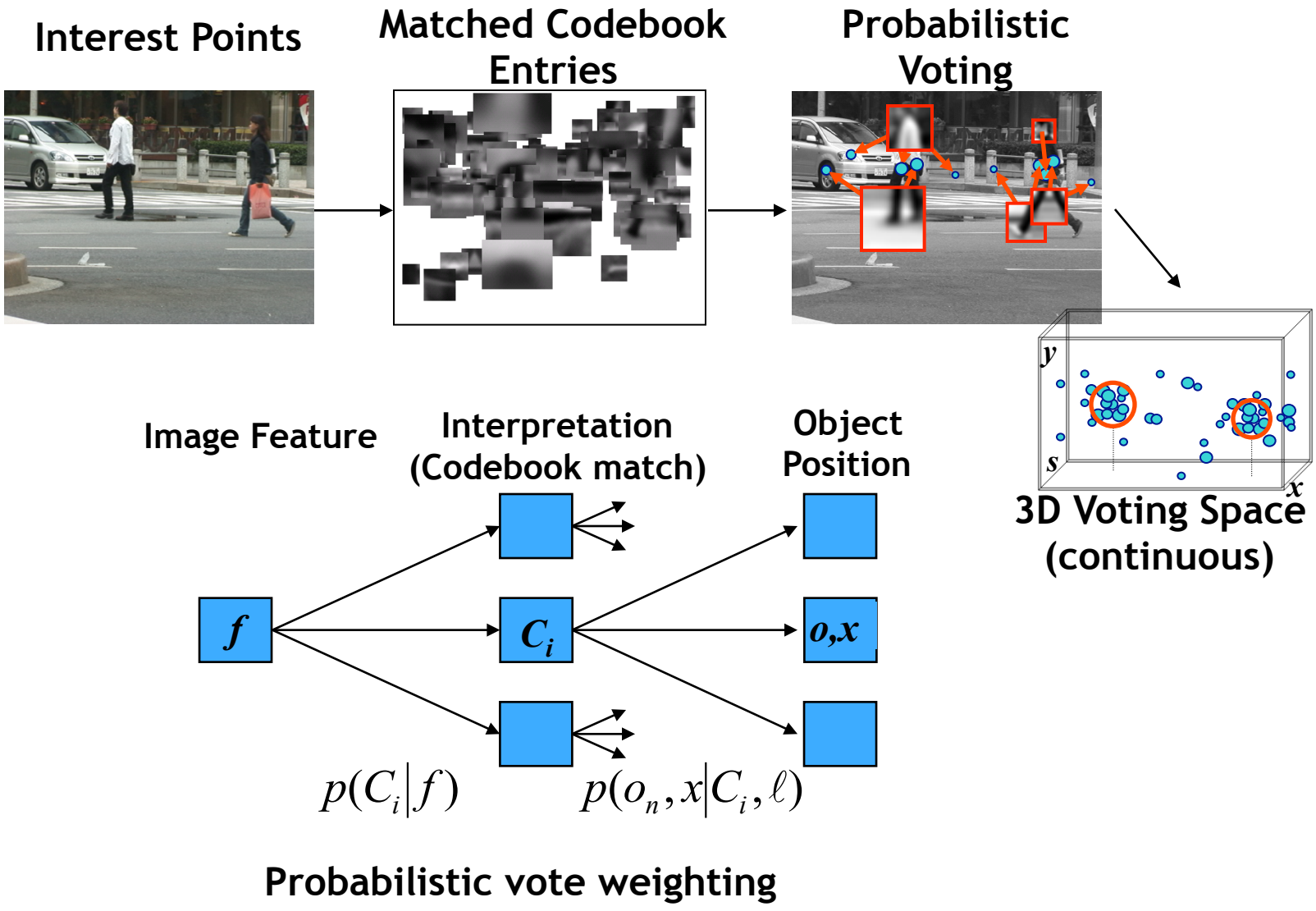
- Learn appearance codebook
  - Extract local features at interest points
  - Clustering descriptors ⇨ codebook of visual words

- Learn spatial distributions
  - Match codebook to training images
  - Record matching positions on object



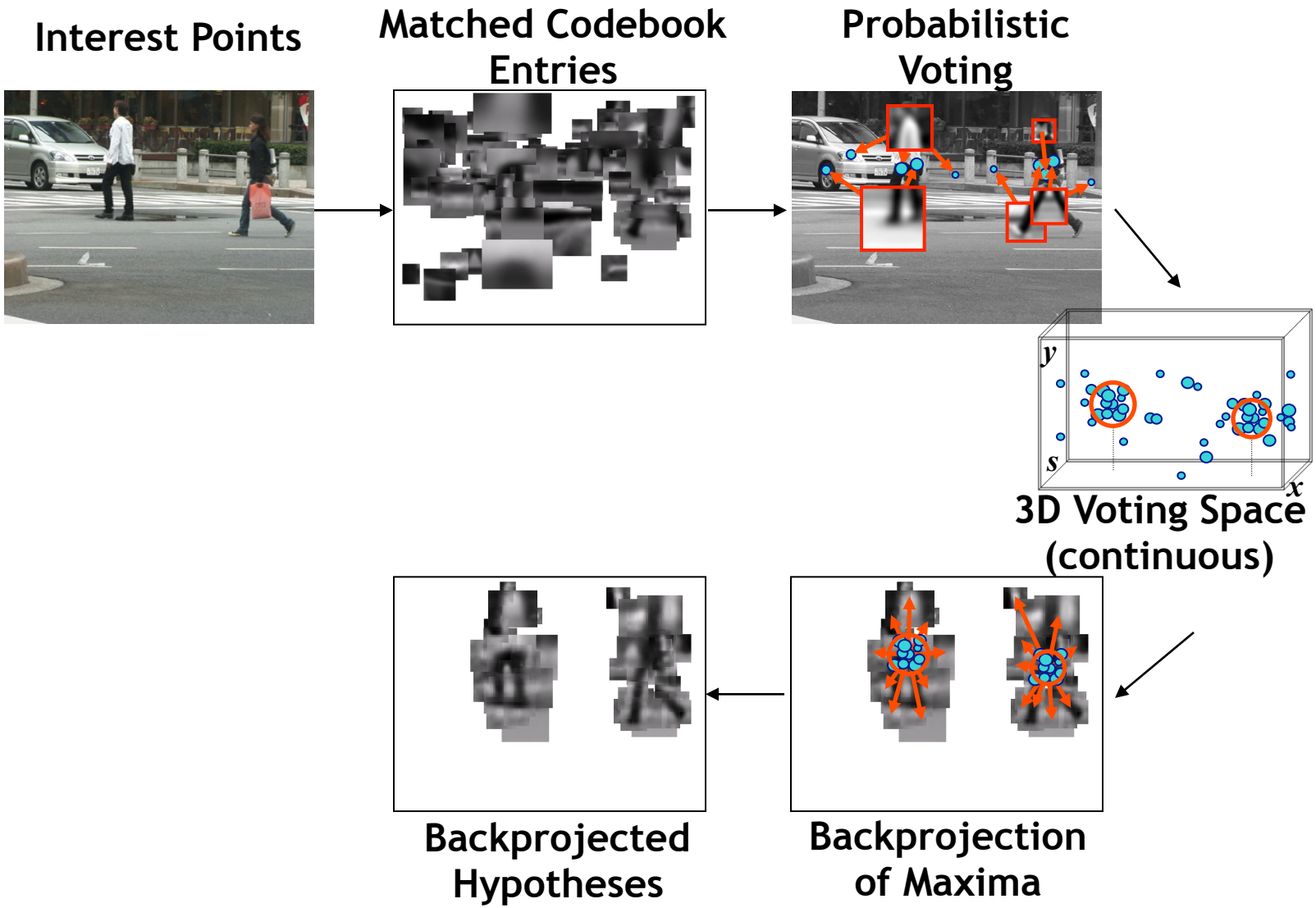
Spatial occurrence distributions

Implicit Shape Model – testing / recognition



[Leibe, Leonardis, Schiele, SLCV'04; IJCV'08]

Implicit Shape Model – testing / recognition



[Leibe, Leonardis, Schiele, SLCV'04; IJCV'08]

Example: Results on Cows



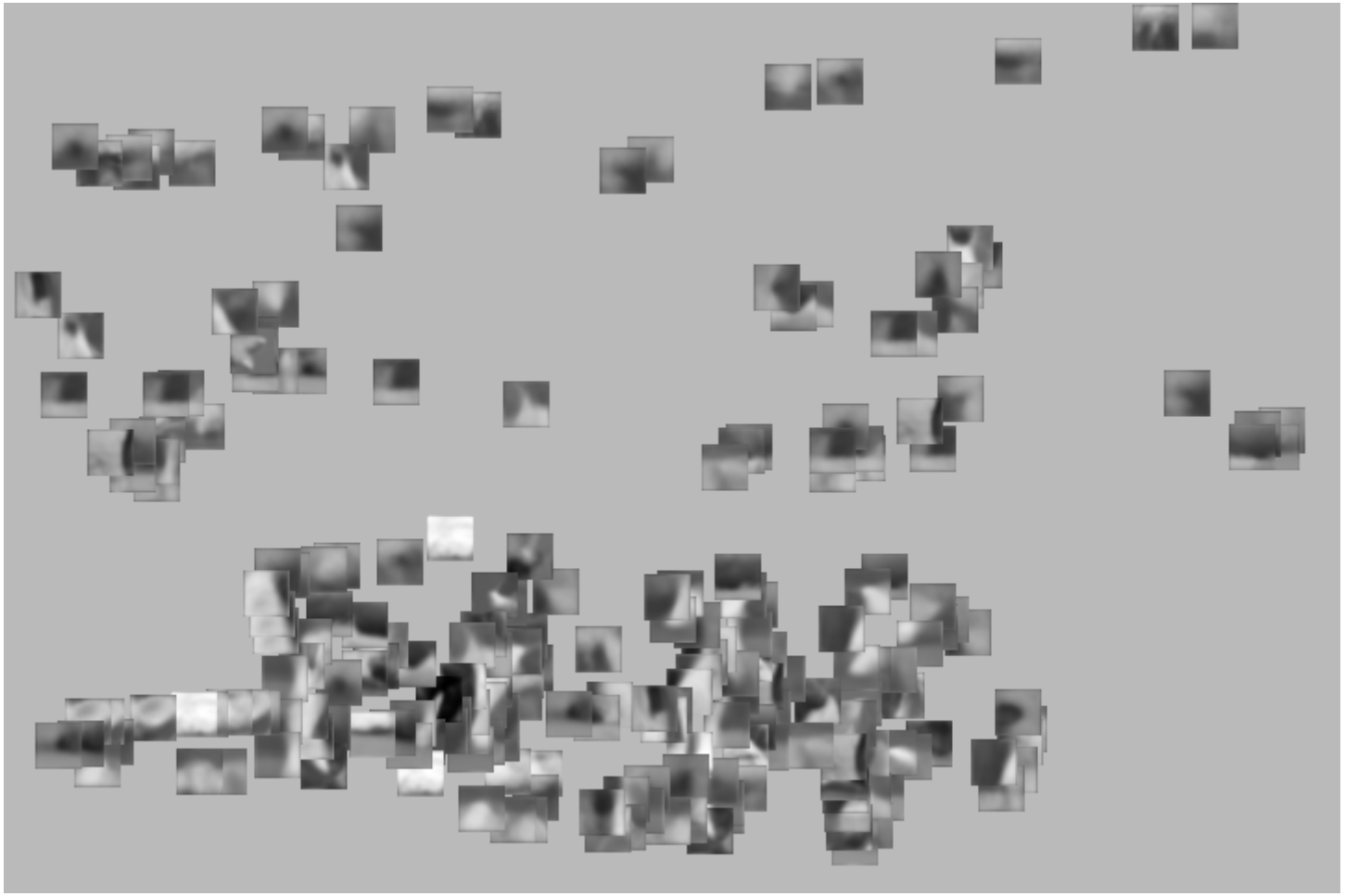
Original image

Example: Results on Cows



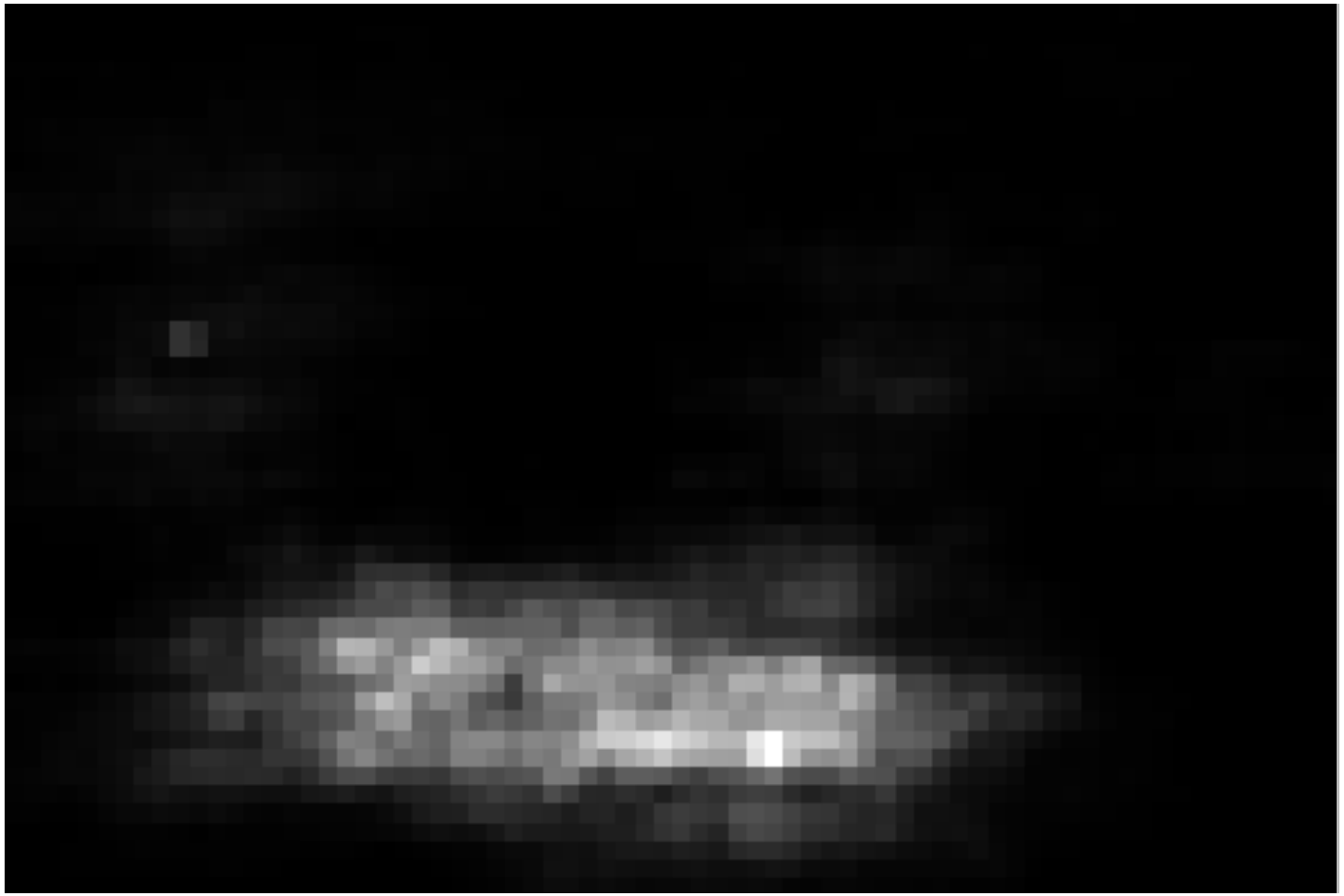
Interest points

Example: Results on Cows



Matched patches

Example: Results on Cows



Prob. Votes

Example: Results on Cows



1st hypothesis

Example: Results on Cows



2nd hypothesis

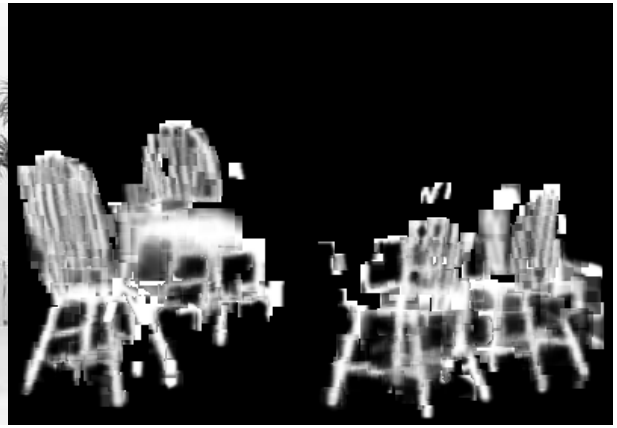
Example: Results on Cows



3rd hypothesis

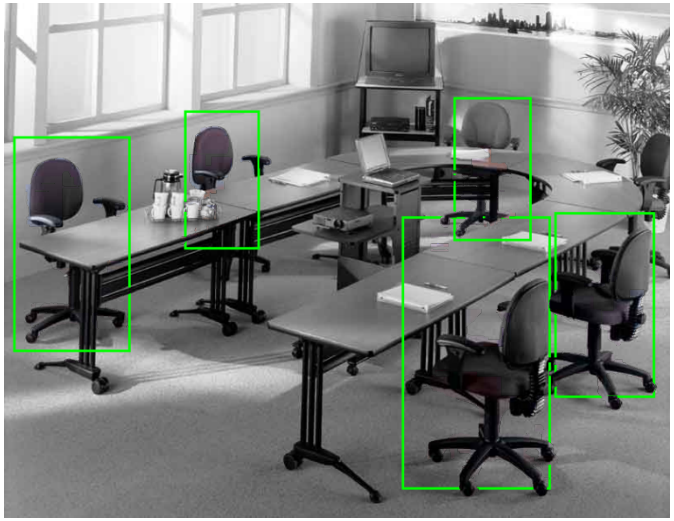
Example Results: Chairs



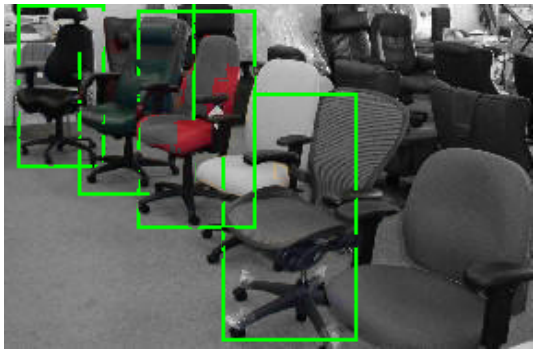


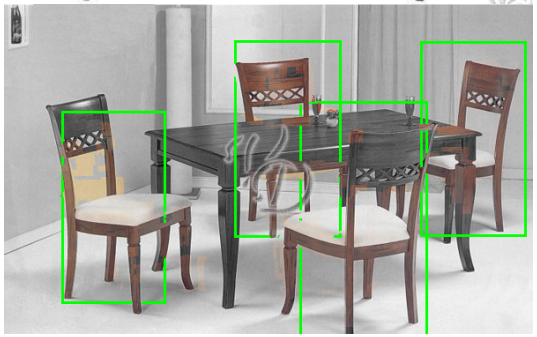
Dining room chairs

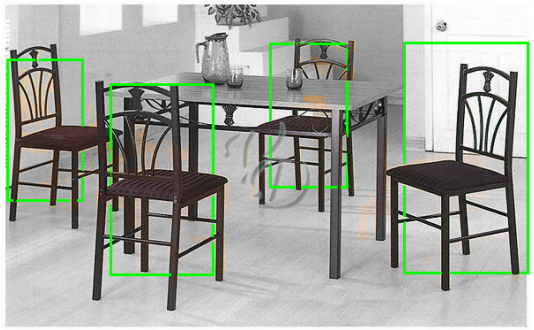












Office chairs

5. Open Challenges



Plan

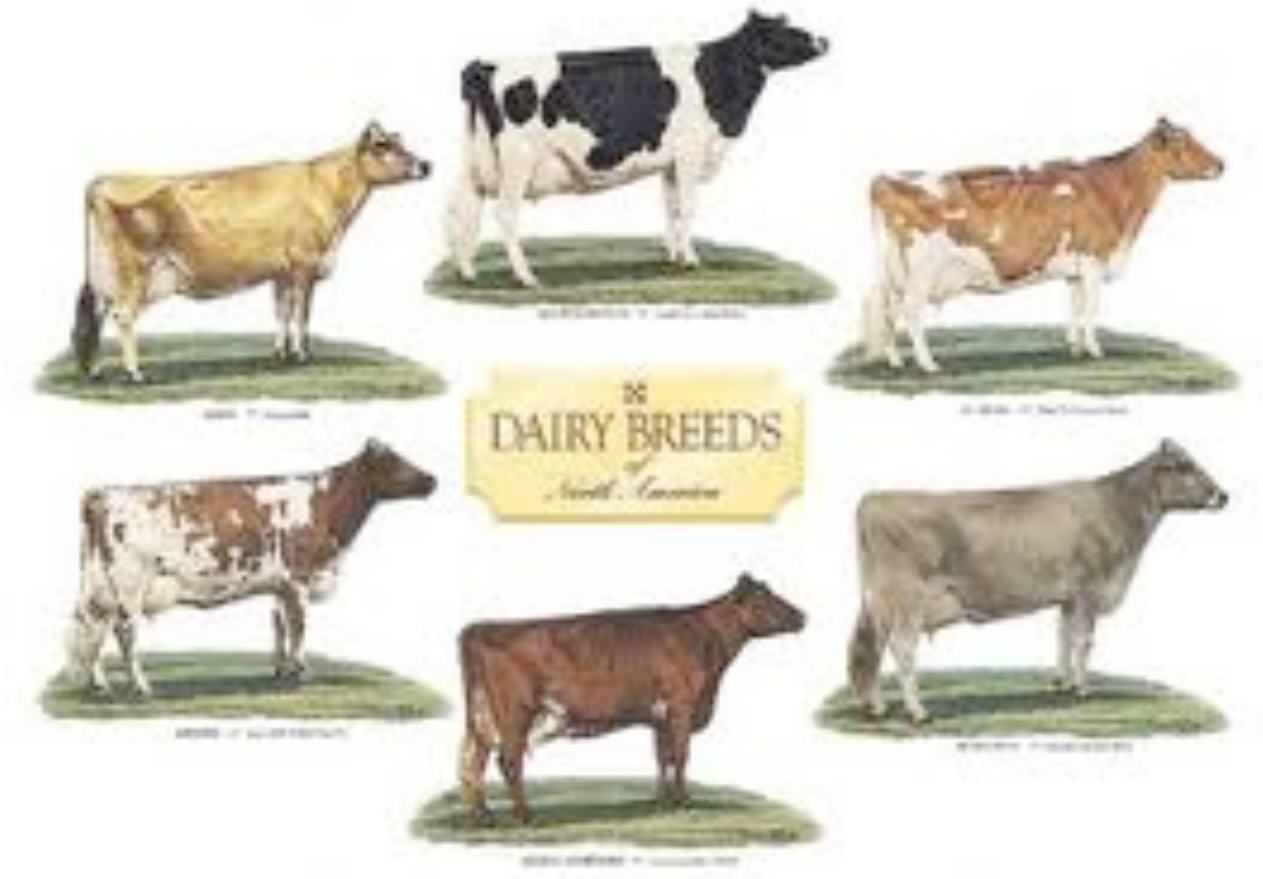
1. Problem definition, classification vs. detection

2. bag-of-words
3. sliding-windows (with global or part-based models)
4. Hough approaches (implicit shape model & forests)

5. *Open challenges*

Multi-Category Discrimination

- Distinguish similar categories.
- Need to look at specific details!



Life-long Autonomous Learning



Systems that automatically notice and learn new classes
able to add class incrementally without global retraining

Integrating Segmentation + Recognition



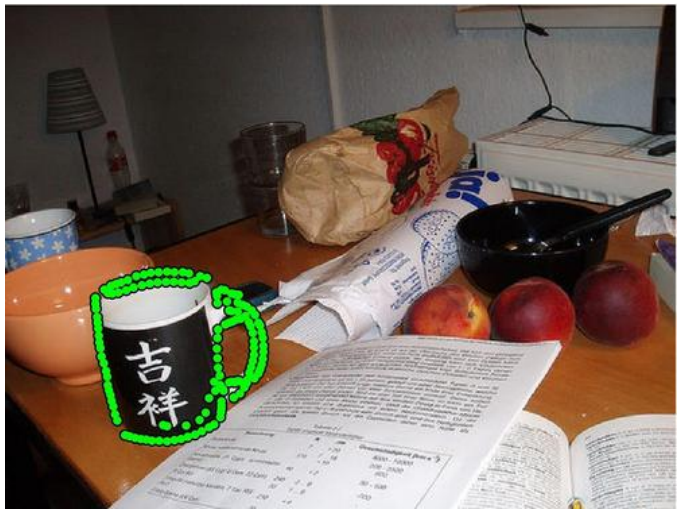












Role for contours in addition to local patches !

Joint recognition and segmentation also possible for object classes

Multi-Aspect Recognition

- Detectors for different viewpoints ⇨ how can this be improved?



- For instance automatically learning about how many viewpoints are relevant

- Sharing examples among viewpoints

Multi-Aspect Recognition

- Video often provides several viewpoints, increasing the chance of success when learning from videos…



Conclusion

- **Object category recognition is …**
  - **… a central subfield of vision → great efforts are put in it still**
  - **… key to much of vision: imagine feedback into other tasks**
  - **… far from solved ! → think of what a person can do**