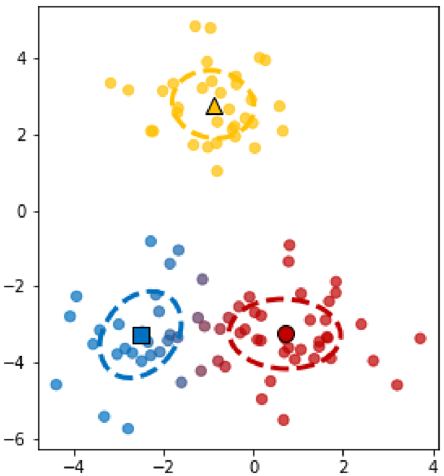
Photogrammetry & Robotics Lab

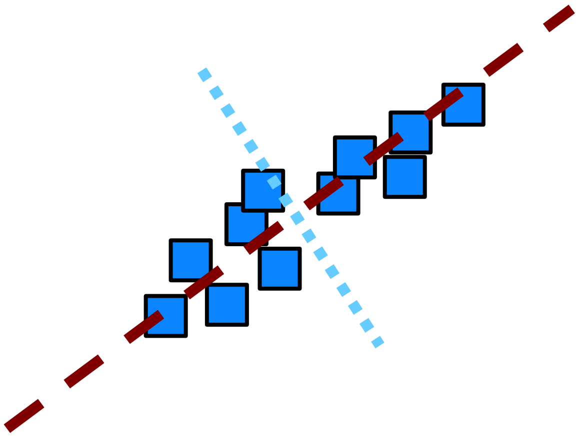
Machine Learning for Robotics and Computer Vision

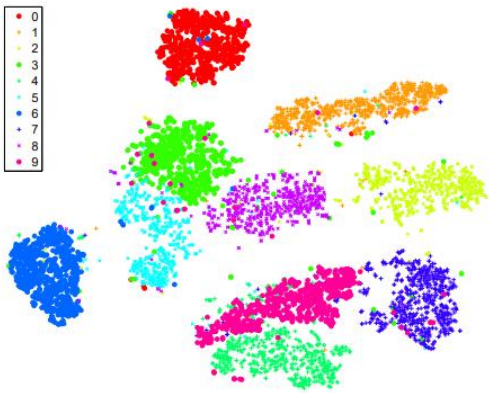
ML for Computer Vision Tasks

Jens Behley

Last Lecture







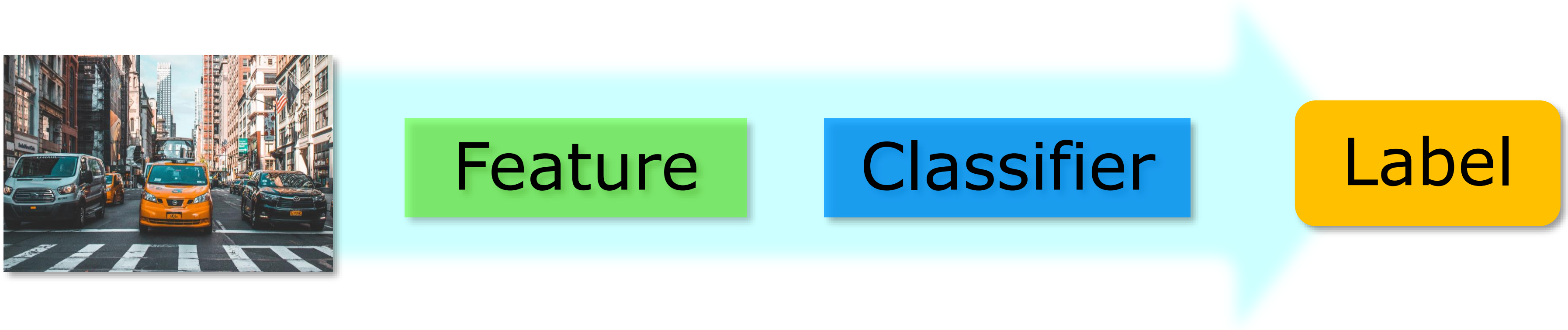
- Discussed several unsupervised learning approaches solving different tasks:
  - Density Estimation (Gaussian Mixture Models)
  - Dimensionality Reduction (PCA)
  - Visualization (t-SNE)

Methods, methods, methods…

- Until now we looked at the core (traditional) methods for supervised & unsupervised learning
  - **Regression:** Linear Regression, Regression Trees
  - **Classification:** k-Nearest Neighbor, Naïve Bayes, Decision Trees, Logistic Regression, Random Forest, AdaBoost, Gradient Boosted Trees
  - **Unsupervised:** GMM, k-means, PCA, t-SNE

- Until now we abstractly talked about the feature vectors $\mathbf{x} \in \mathbb{R}^D$

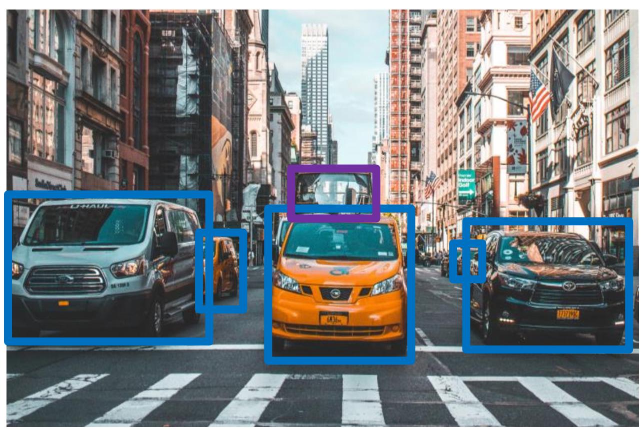
Feature Engineering



- Applications to Computer Vision tasks: Extract features and apply supervised learning methods
- Most of the time: designing task-specific features → **feature engineering**

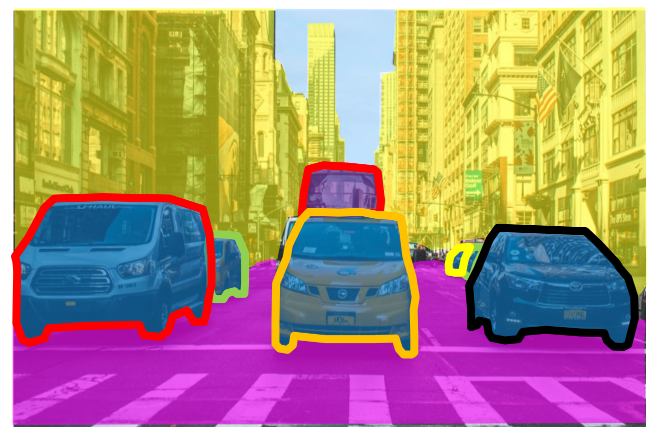
Perception Tasks

Classification



Object Detection

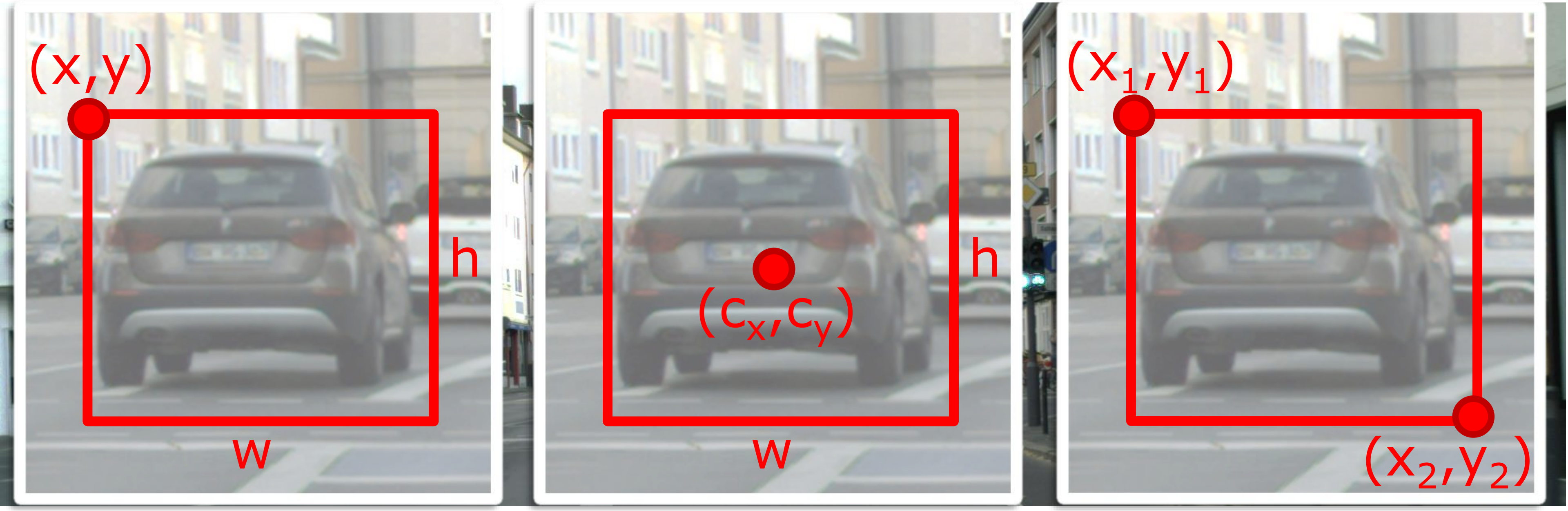
Semantic Segmentation



Panoptic Segmentation

Photo by Marisol Benitez on Unsplash

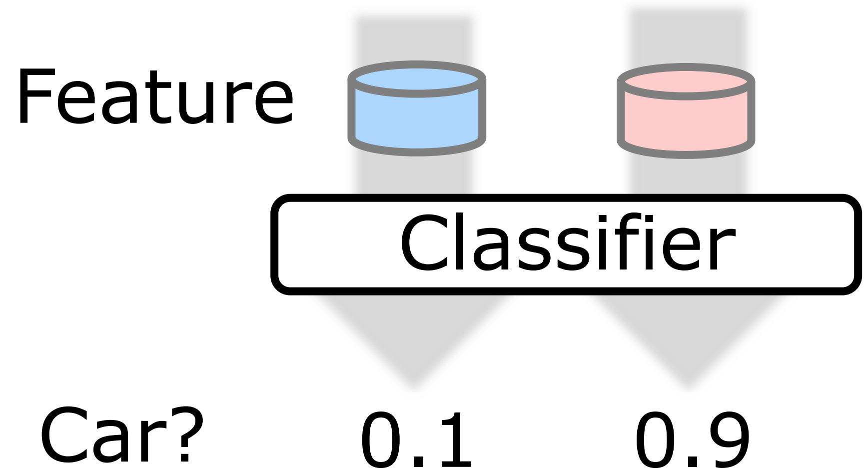
Anatomy of an Object Detector



- Input: RGB Image
- Output:
  - bounding boxes defined by
    (x, y, w, h) or (c_x, c_y, w, h) or (x_1,y_1, x_2,y_2)
  - confidence scores in [0,1]

Image from Cityscapes Dataset

Anatomy of an Object Detector

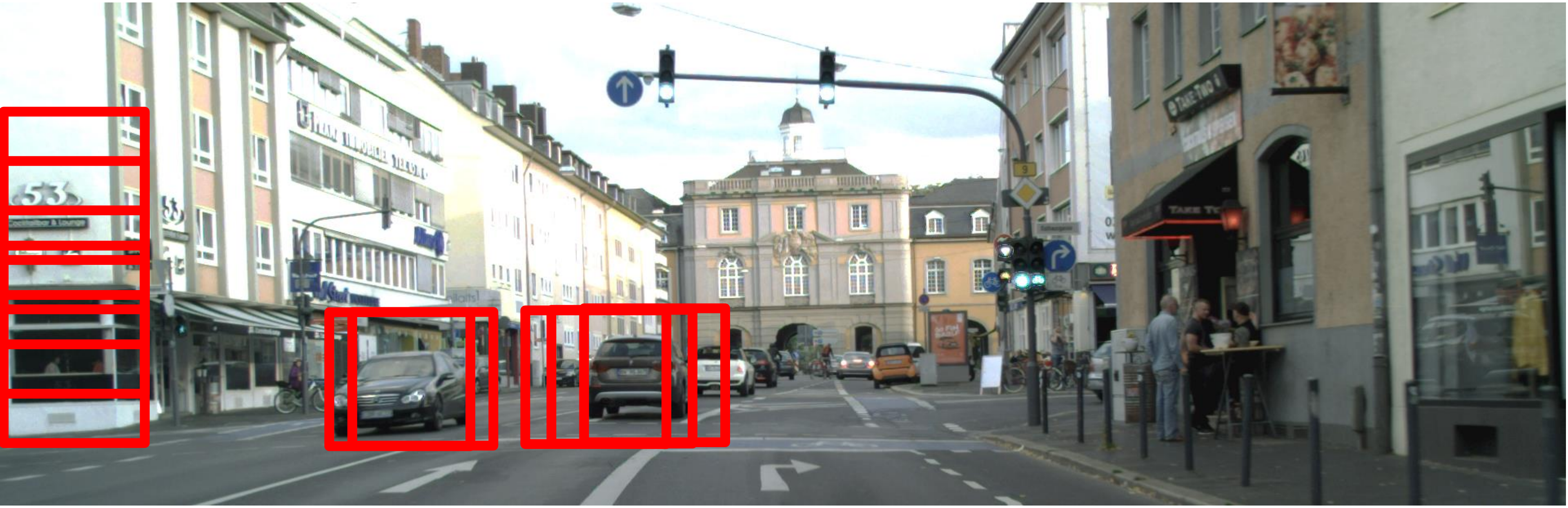


General Approach

1. Extract regions
2. Classify and score regions
3. Keep high scoring regions

Image from Cityscapes Dataset

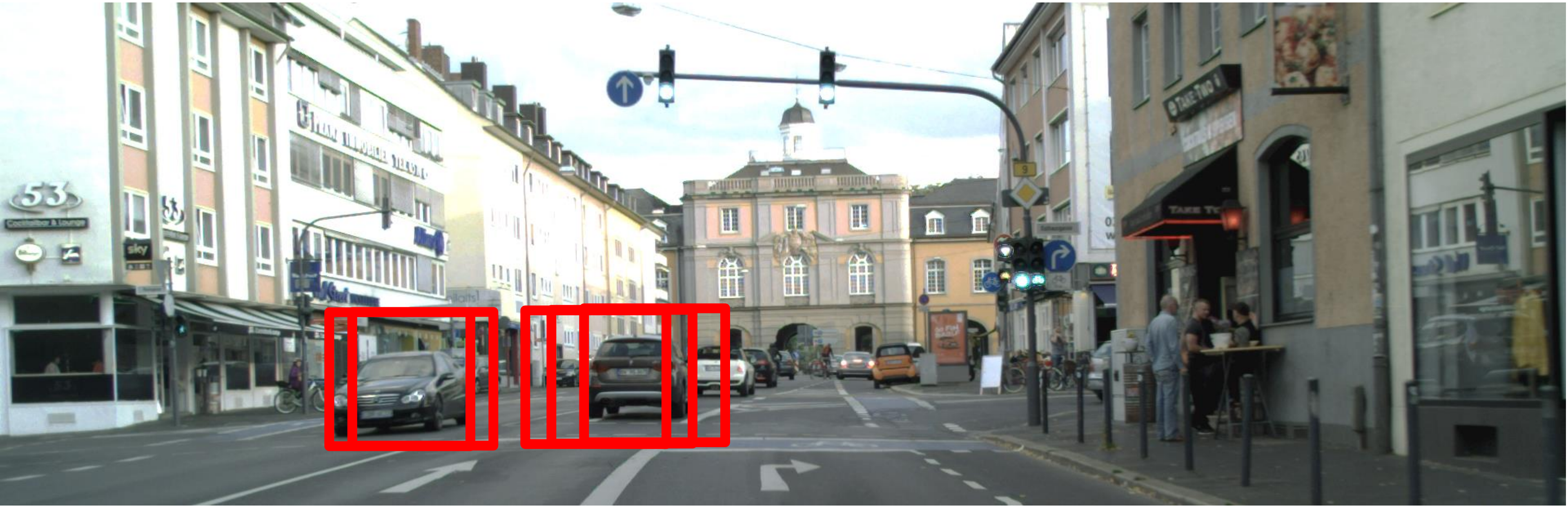
Sliding Window Approaches



- Densely sample regions from image
- Classify image features extracted from the region

Image from Cityscapes Dataset

Non-maximum Suppression (NMS)

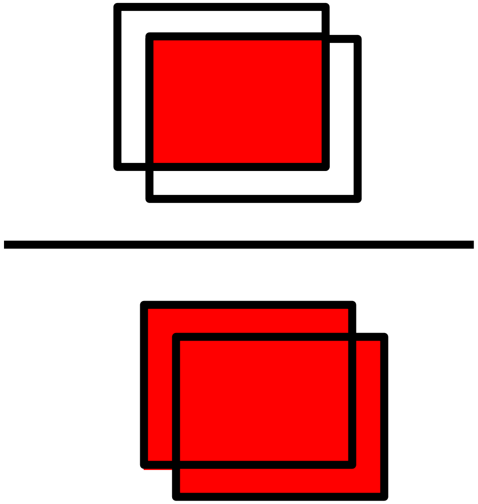


- Keep high confidence detections
- Remove non-maximum bounding boxes with too large **overlap**

Image from Cityscapes Dataset

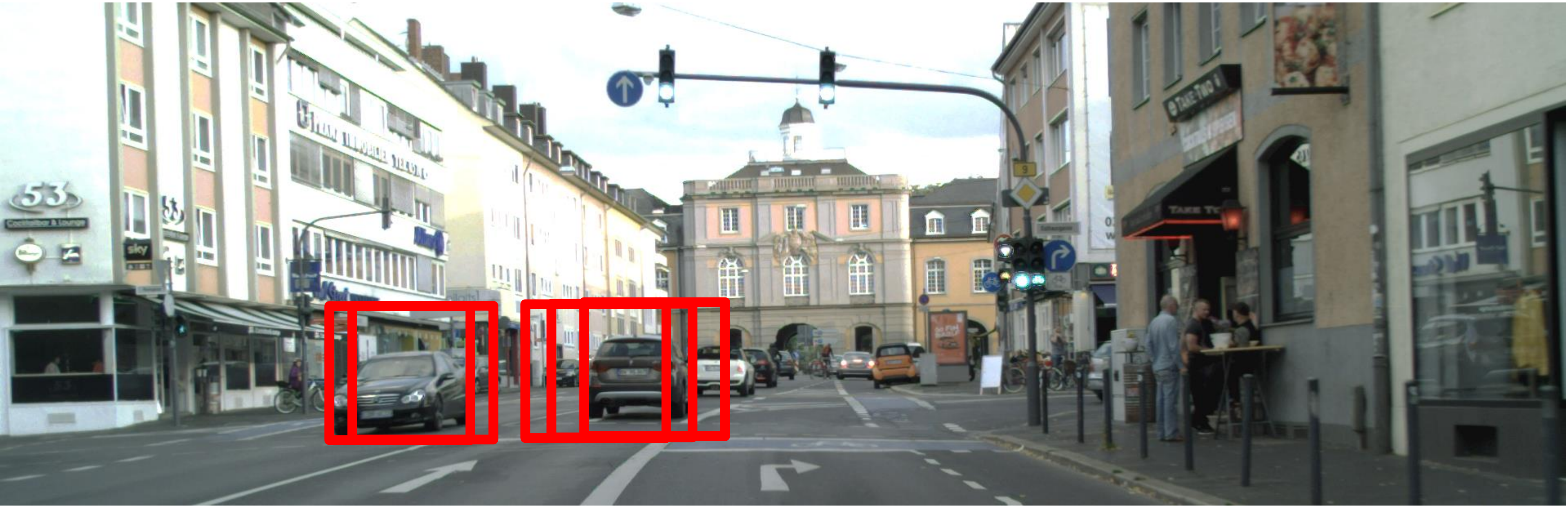
Intersection-over-Union (IoU)

$\mathrm{IoU}(B_1, B_2) =$



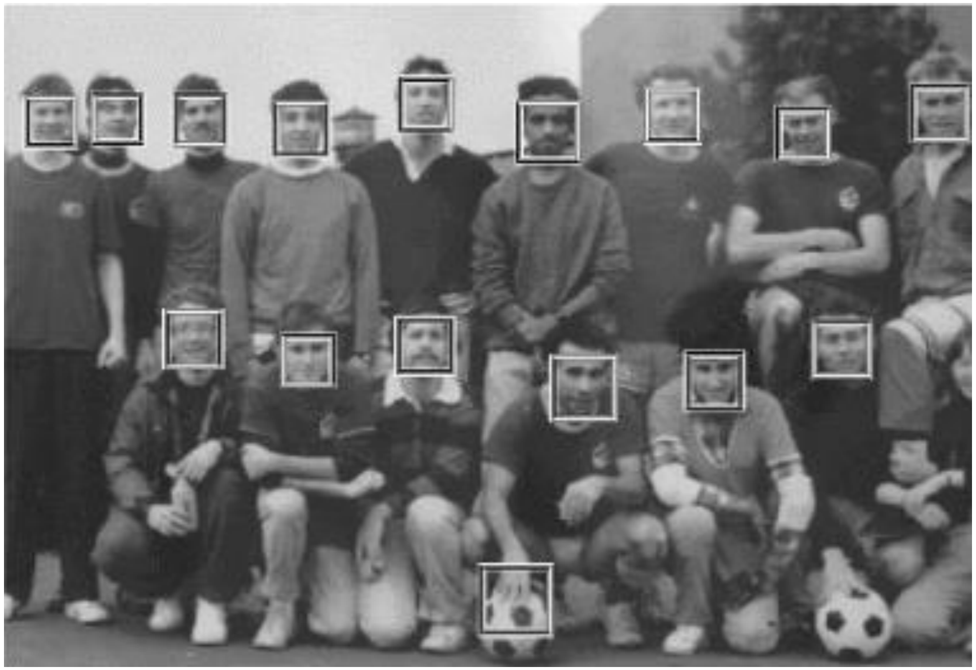
- Area of intersection of B_1 and B_2 divided by area of union of B_1 and B_2

Non-Maximum Suppression



1. Sort boxes by confidence score
2. For each box: If overlap with accepted boxes is larger than threshold ➔ drop box

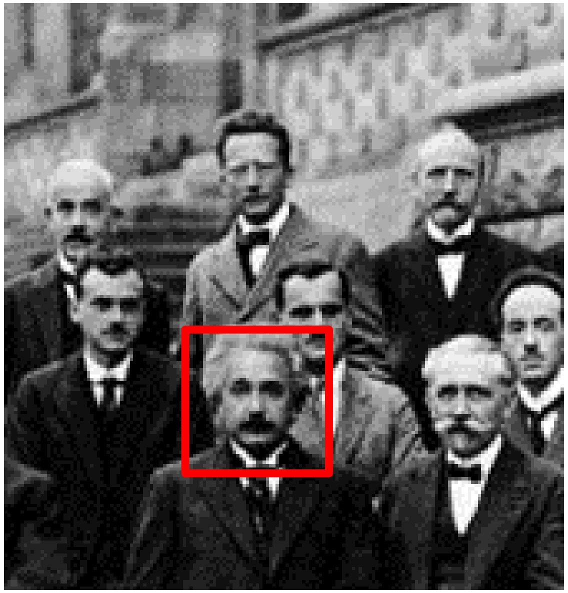
Viola Jones Object Detector



- Main building blocks:
  - Features: Haar-like features
  - Classifier: Decision Stumps with AdaBoost
- Cascade of increasingly complex classifiers

[Viola & Jones, 2001]

Haar-like Features



- Difference of sum over regions located inside bounding box:

$$\Delta = \sum_{(x,y)\in\text{white}} I(x,y) - \sum_{(x,y)\in\text{black}} I(x,y)$$

[Viola & Jones, 2001] Image from Solvay Conference/Wikipedia Commons

Weak classifier

- Find optimal weak classifier that best separates weighted positive and negative examples:

$$h_j(\mathbf{x}) = \begin{cases} 1 & \text{, if } p_j \Delta < p_j \theta_j \\ 0 & \text{, otherwise} \end{cases} \qquad p_j \in \{0, 1\}$$

- In each stage, best Haar feature and parameters determined to classify weighted examples.

[Viola & Jones, 2001]

Fast Implementation

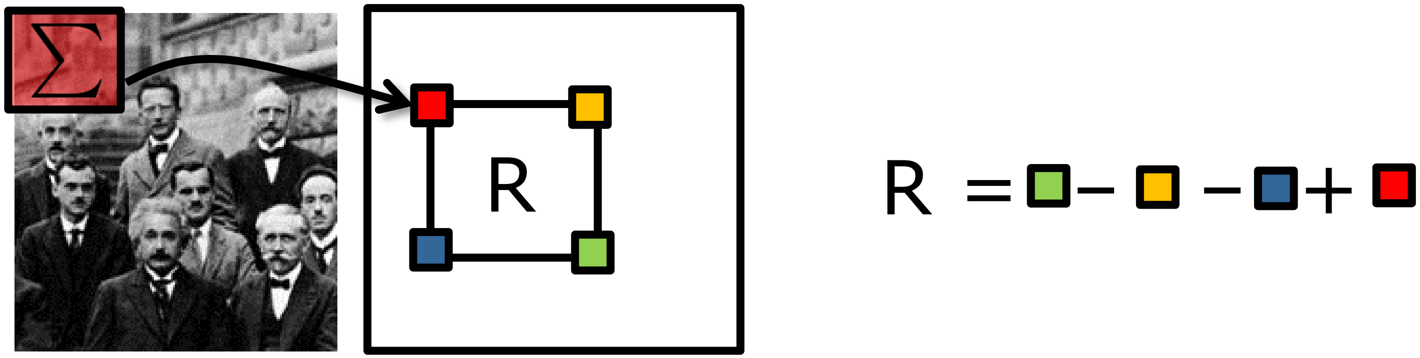
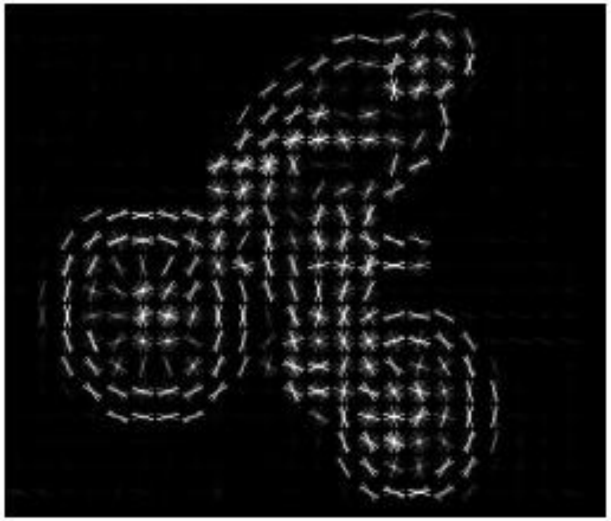


Image Integral Image

- Even on 700 MHz for 384x288 images only 0.067 s per image
- Two tricks that enable fast evaluation:
  1. **Integral images** enable evaluation of Haar Features in constant time
  2. **Cascaded classifiers** that quickly allow to reject negative windows

[Viola & Jones, 2001]

Person Detection with HOG

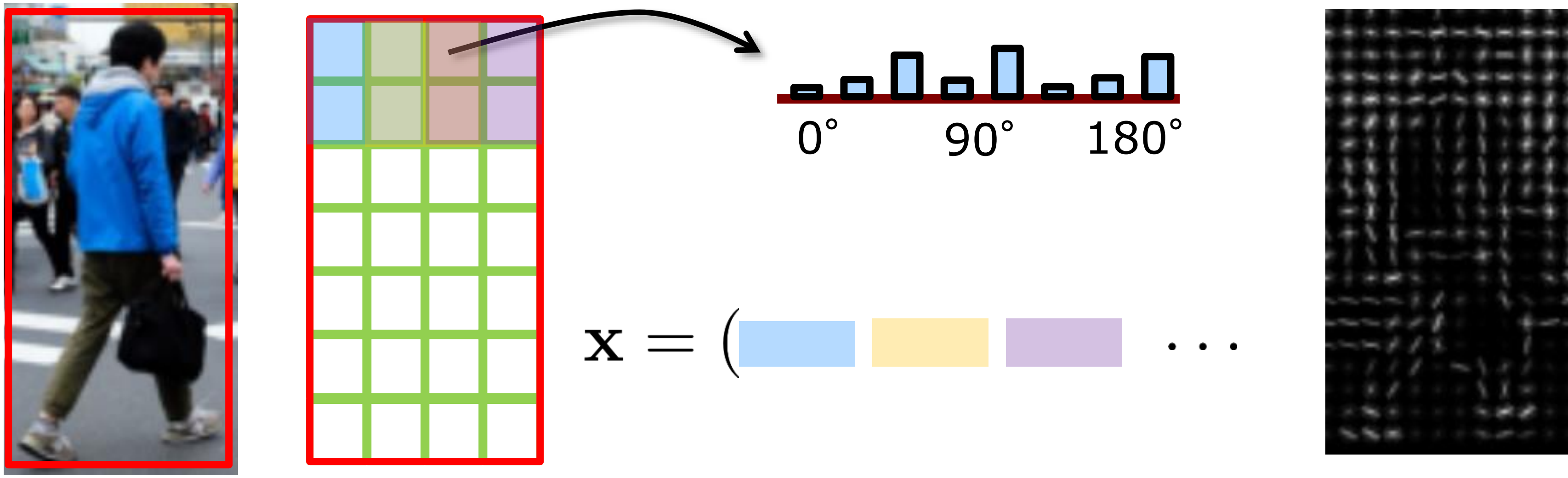



- Main ingredients:
  - Feature: Histogram of Oriented Gradients (HOG)
  - Classifier: Linear SVM (~ Logistic Regression)
- Fine grained feature to capture shape of persons

[Dalal & Triggs, 2005]

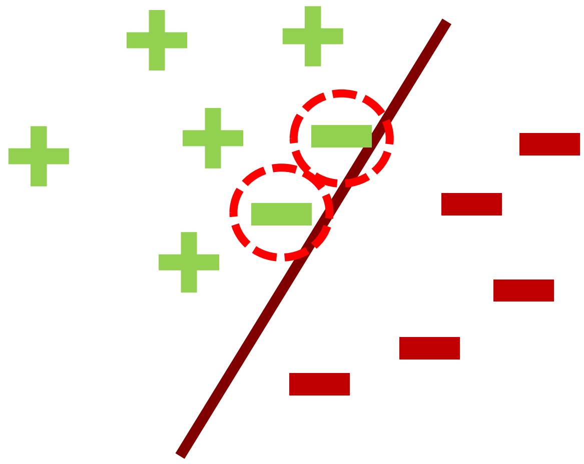
Photo from [Unsplash]()

Histogram of Oriented Gradients



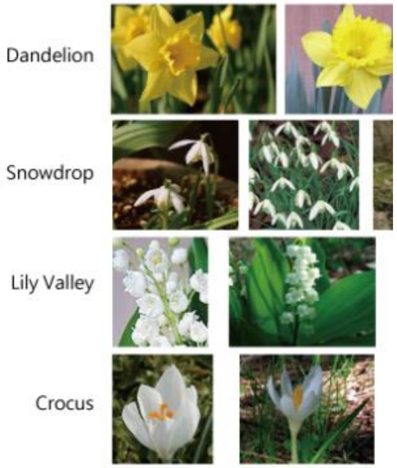
- Subdivide detection window into cells
- For each cell: histogram over gradient orientations weighted by magnitude
- Overlapping blocks of D x D cells
- Final feature vector is concatenated L2-normalized block histograms

Additional Tricks

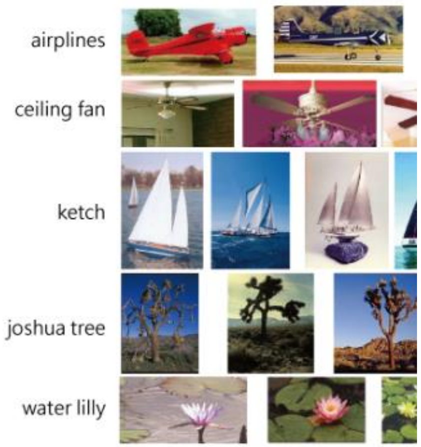


- **Data Augmentation**
  - Horizontal flip/mirroring
  - → More training examples
- **Hard Negative Mining**
  - Enlarge Training set with negative (non-person) examples that are wrongly classified

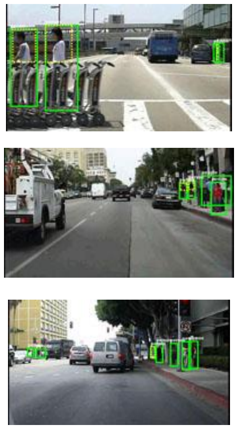
Datasets & Benchmarks



Oxford Flower



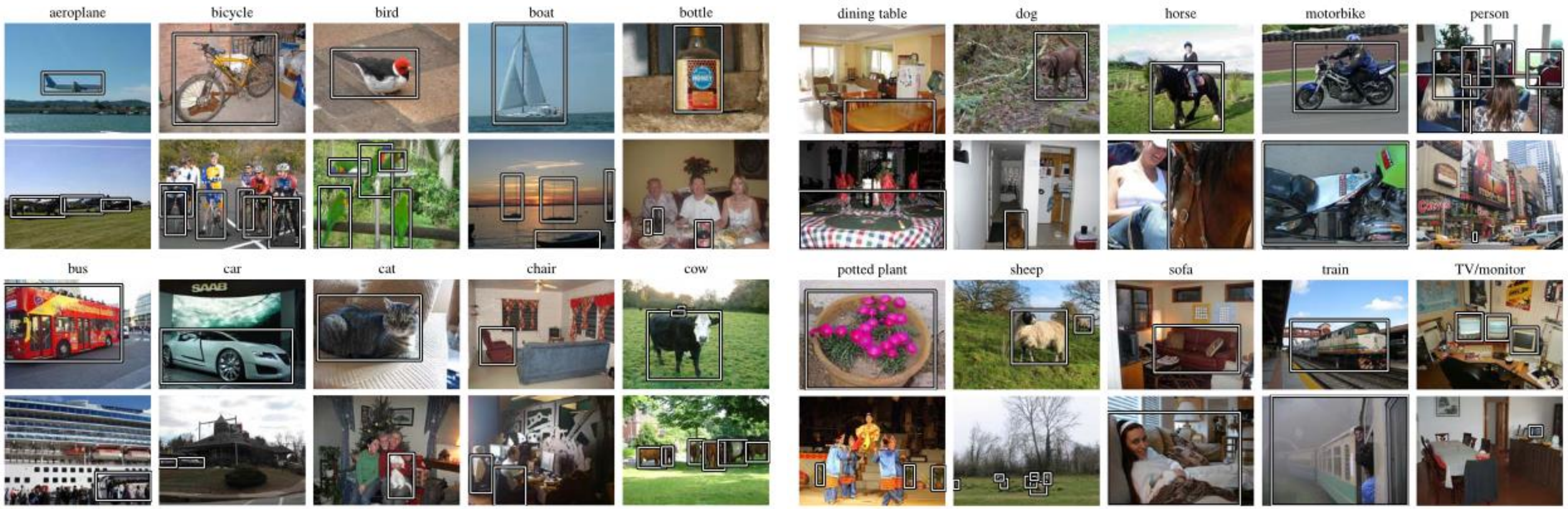
Caltech 101


Caltech Pedestrian


MIT LabelMe

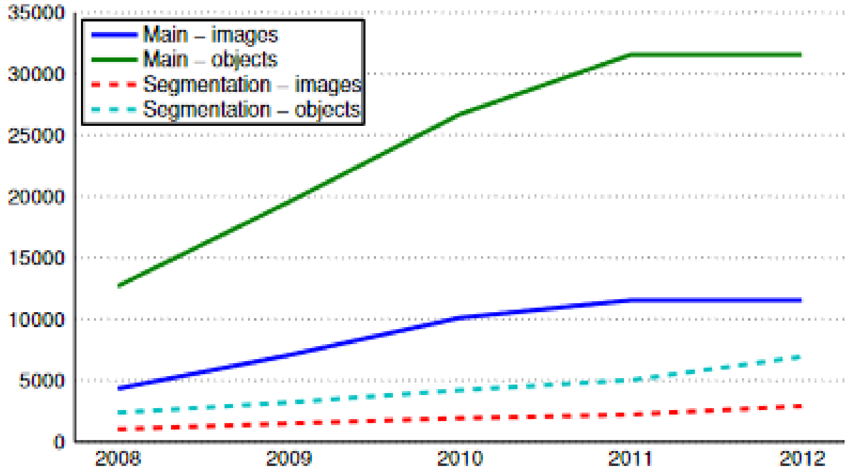
- Key principle of computer vision research: datasets and associated benchmarks
- New datasets provide new challenges
- Incentivize progress by competitions

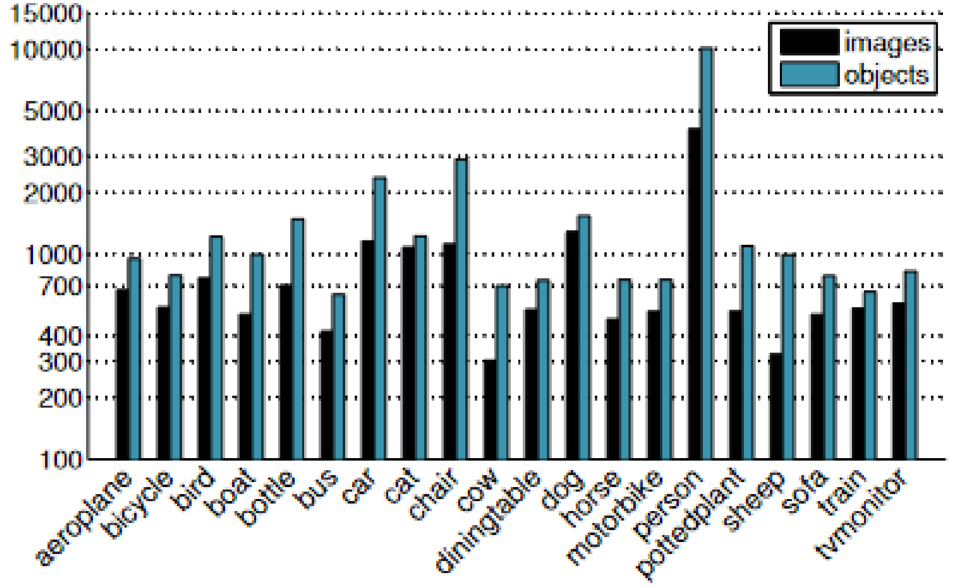
PASCAL Visual Object Classes (VOC)



- Classification and Detection Challenges
  - Collected from Flickr images
  - 11,540 Images (Pascal VOC 2012)
  - 20 classes
- Annual competitions & workshops(2006-12)

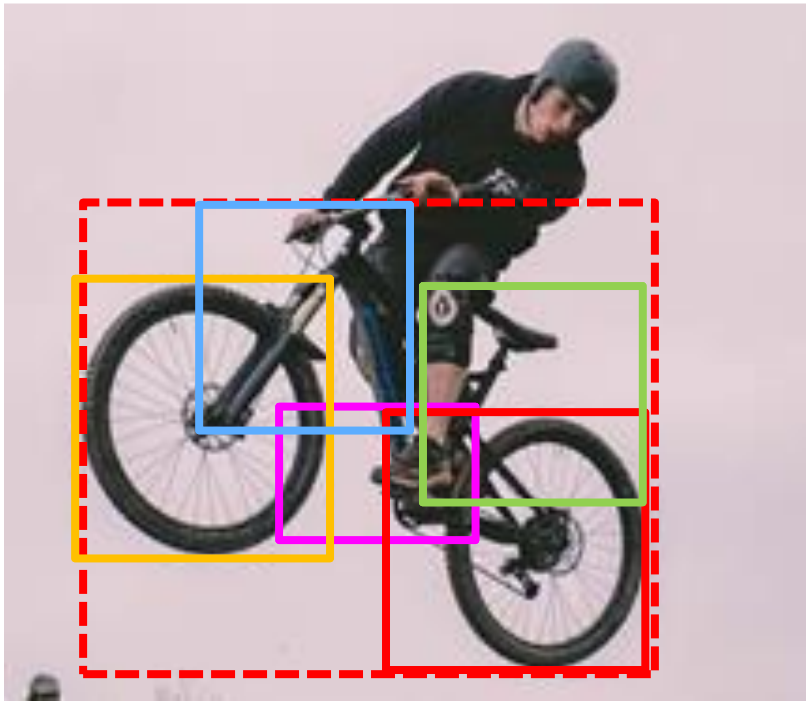
Pascal VOC 2007-2012



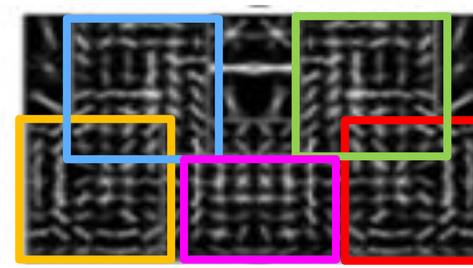


- Number of images grew over the years
- Each class has at least about 300 images
- Diverse mix of rigid and deformable object classes

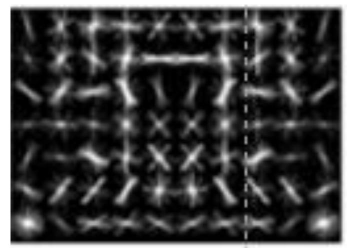
Deformable Part Models



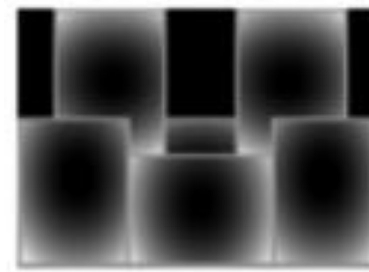
bicycle model



Part filter



root filter



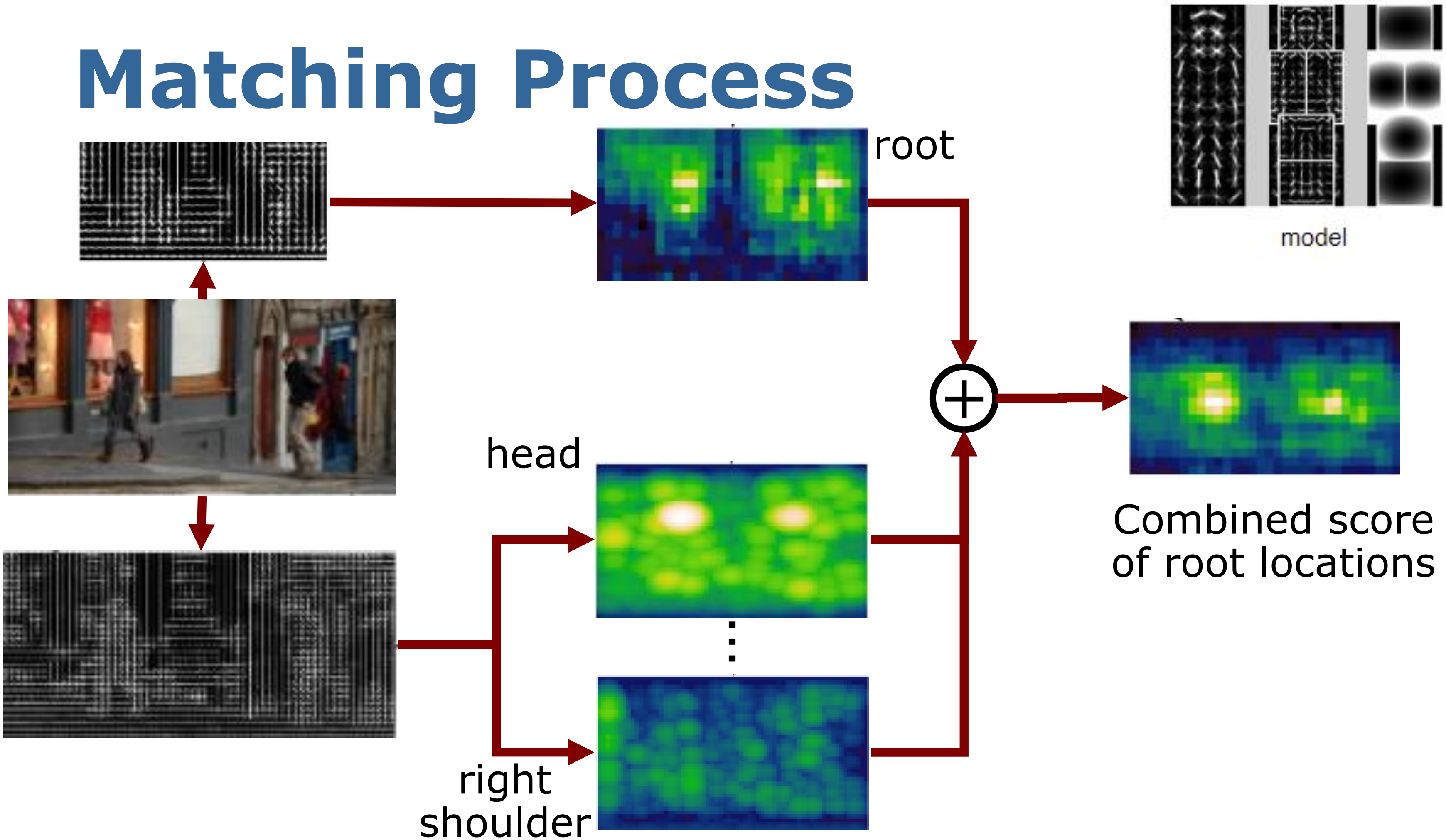
deformation cost

- Coarse root filter and fine part filters
- Features: HOG on different levels of the image pyramid

[Felzenszwalb et al., 2009]

Photo from [Unsplash]()

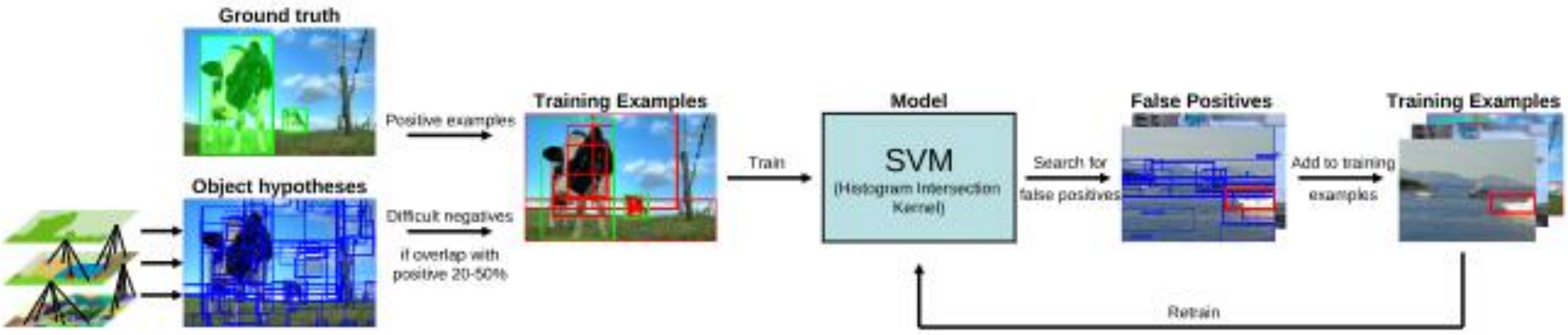
Matching Process



- Root filters are evaluated on coarse images
- Part filters are applied on finer images
- Aggregated votes determine root location

[Felzenszwalb et al., 2009]

Selective Search



Object Proposals

Hard Negative Mining

- Sliding Window approach quite inefficient
  - Need to check/classify many irrelevant windows
- **Main idea:** Extract only regions that corresponding to objects (object proposals)
- Fewer evaluated regions → stronger features

[van de Sande, 2011]

Selective Search

- Fine-to-coarse aggregation of super-pixel regions

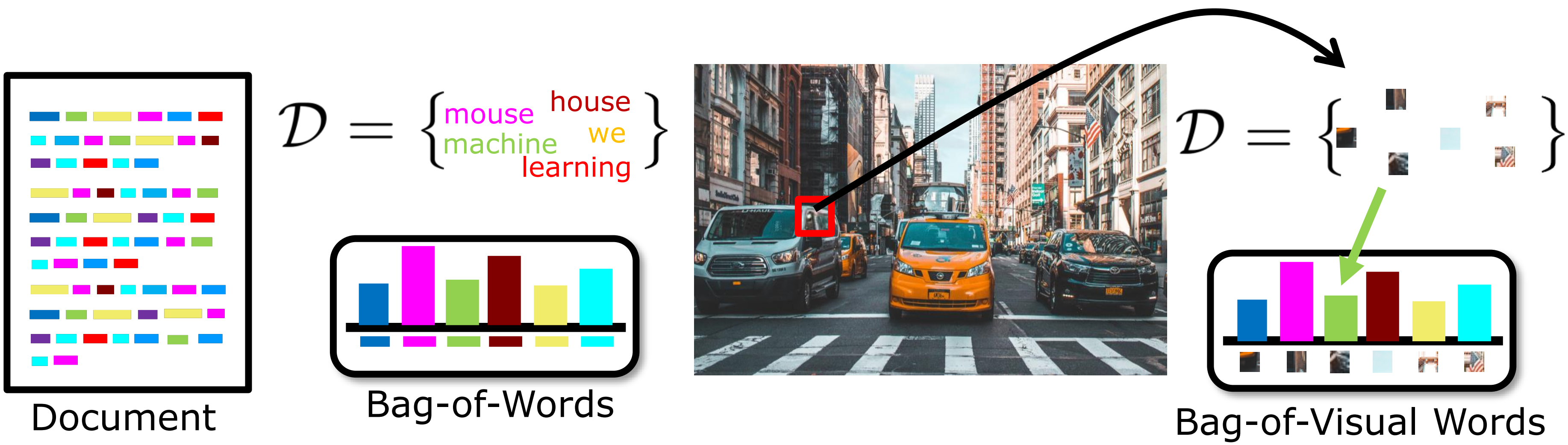
[van de Sande, 2011]

Selective Search

- Fine-to-coarse aggregation of super-pixel regions
- Far less proposals then sliding window
- Includes different scales

[van de Sande, 2011]

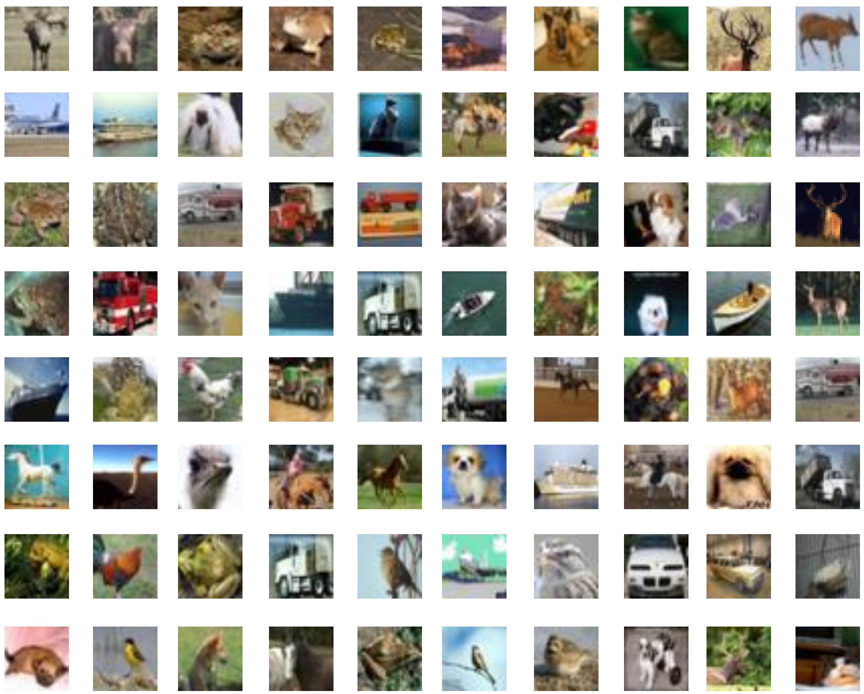
Bag-of-(Visual)-Words

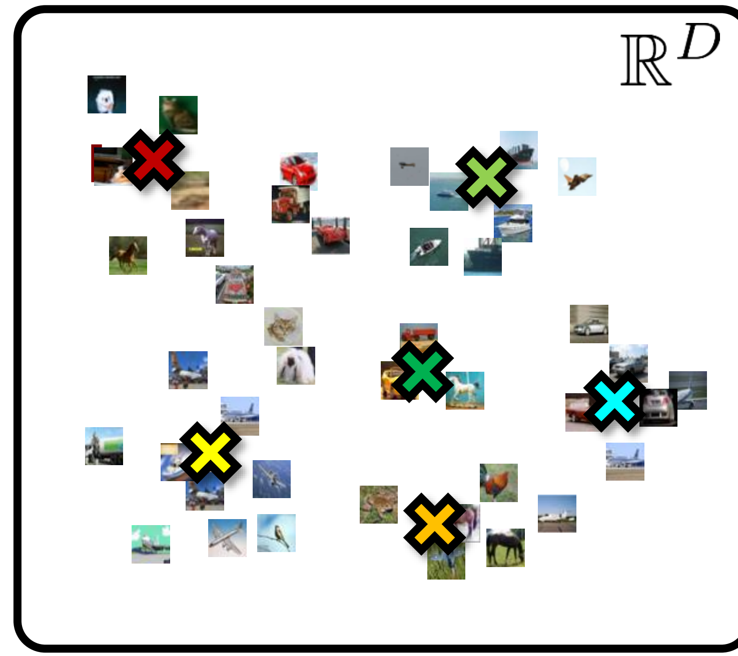


- **Idea:** Histogram of occurrences of words from a **dictionary** in a text document
- Translated to image domain: dictionary is set of **representative** image descriptors, e.g. SIFT descriptors

[Csurka et al., 2004]

Learning a Dictionary

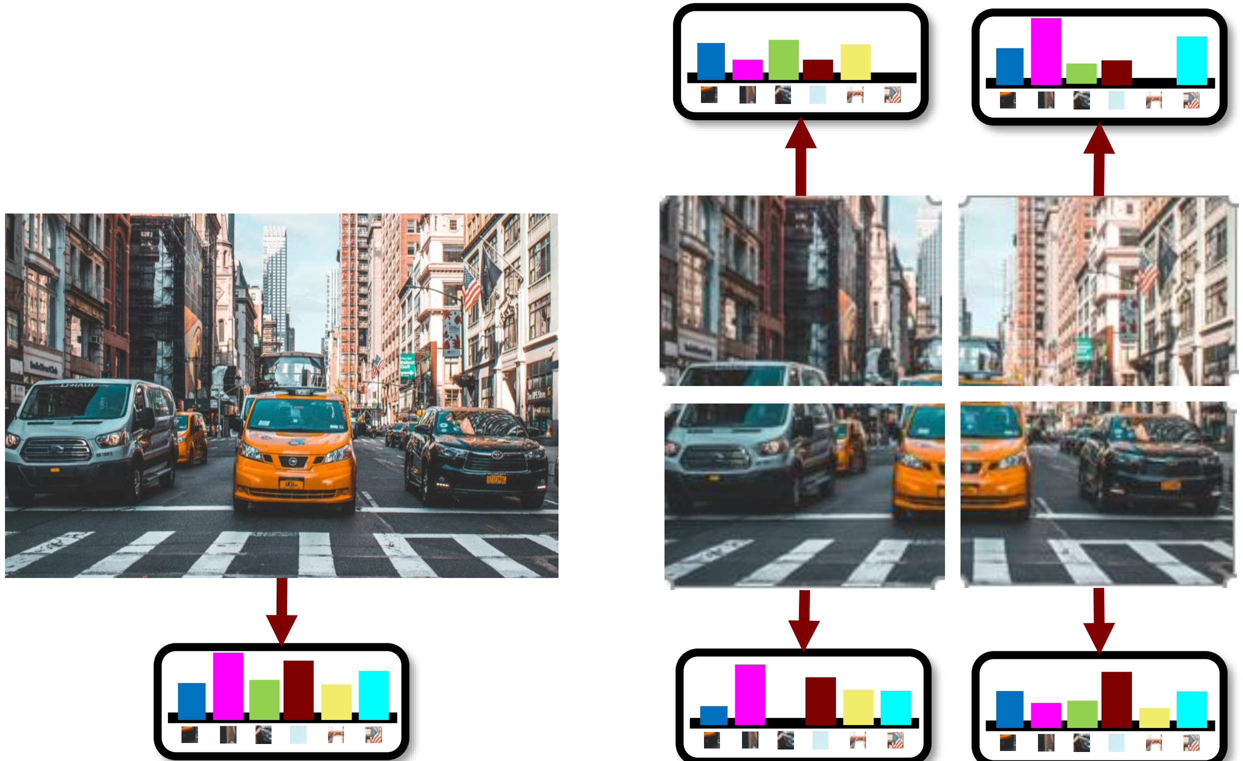




- Extract large set of descriptors/image patches from training set
- K-means on these descriptors results in K dictionary entries (= cluster centers)

Images from CIFAR10

Spatial Pyramid



- Instead only computing bag-of-words for whole region, subdivide region in smaller parts to retain spatial locations.

[Lazebnik et al., 20006]

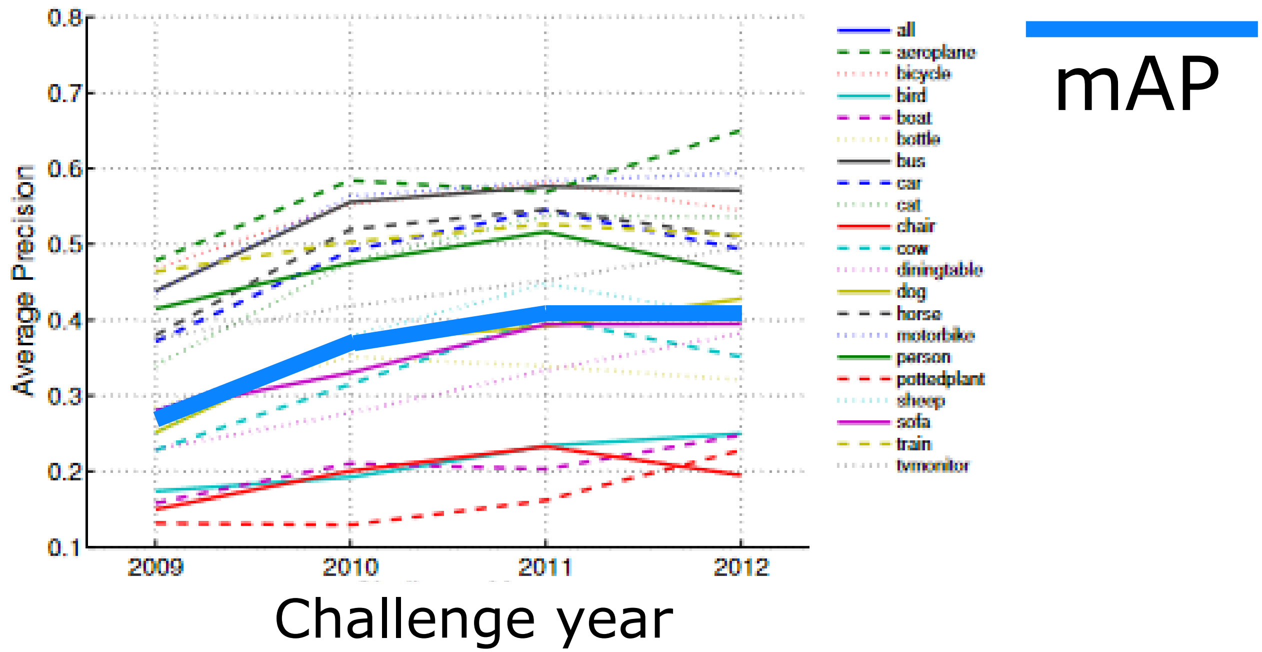
Photo by Marisol Benitez on Unsplash

Selective Search@Pascal VOC

- Descriptor for BoW: Variants of SIFT descriptors on color images
- BoW (K=4000) + Spatial Pyramid
  ➔ feature vectors of length 360,000
- Classifier: Support Vector Machines
- Hard Negative Mining

- Winning entry of Pascal VOC 2012 detection challenge

Pascal VOC Detection



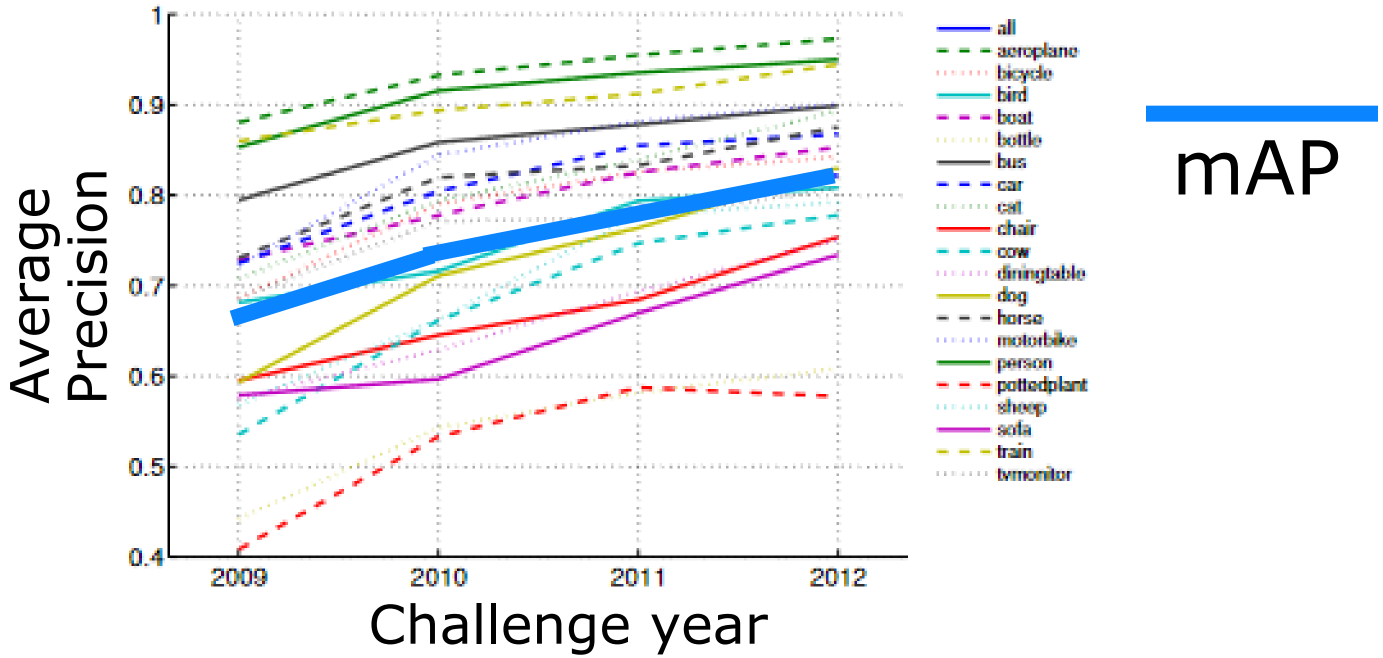
- 2008-2011 dominated by DPM-based methods: other features, re-scoring.
- 2012: Selective search with improved features

[van de Sande, 2011]

Image Classification

- **Task:** Determine label for image; which objects are present in an image
- Categorization of images & image search

Classification on PASCAL VOC



- Bag-of-Visual Words dominant approach
- Combination with Spatial Pyramids and multiple Bag-of-Words
- Classification-by-detection by using output of classifier applied to regions

IMAGENET Dataset

- Based on WordNet hierarchy
  - Semantic hierarchy and taxonomy
  - Large fraction of English nouns
- Crawled using multiple search engines
  - 12M images, 15k categories
- Image categories are verified by Amazon's Mechanical Turk workers

ImageNet Large Scale Visual Recognition Challenge (ILSVRC)

- Evaluate image classification and object detection algorithm at large scale
- Workshops & Competitions from 2010-2017

- Subset of ImageNet data:
  - 1.2M train, 50k validation, 100k hidden test images (732-1300 images per class)
  - 1,000 classes
- The ImageNet-1k data is usually the thing, when people refer to “ImageNet”

Comparison with Pascal VOC

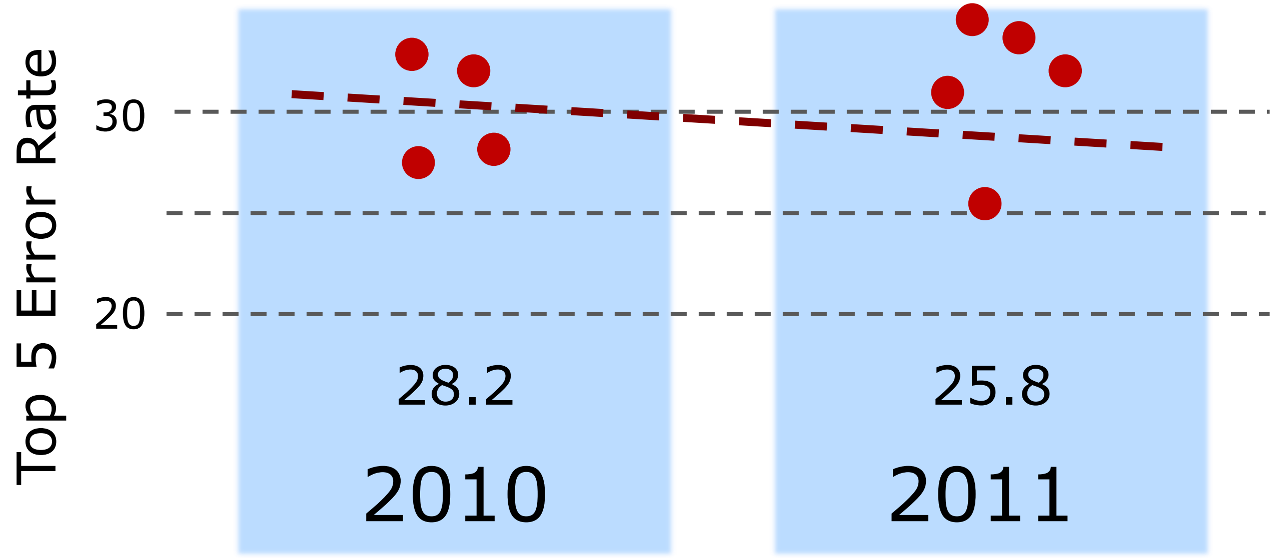
- More fine-grained categorization of classes
- Different birds, cat and dog breeds.

Top-5 error rate

Example images for category ‘paint brush’

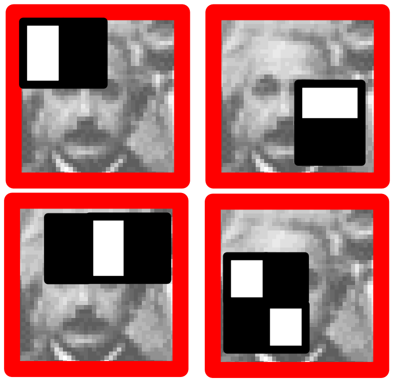
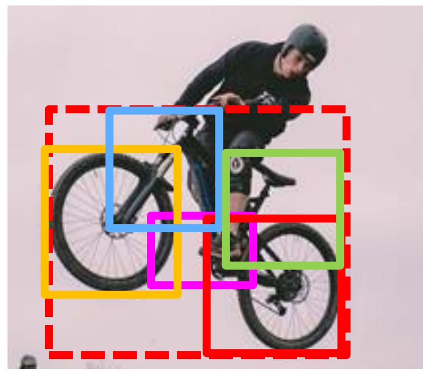
- **Task:** Given an image predict categories of objects that may be present in the image
- Targeted label might be ambiguous
  → Consider top-5 predictions for evaluation
- Is target label under top-5 predictions?

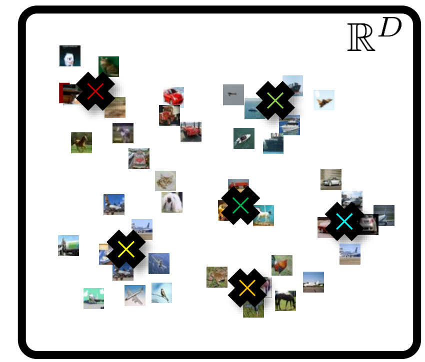
Progress on ImageNet



- Mainly more expressive features: Fisher Vectors (“soft” BoW) → 1M-dimensional fisher vectors (2011) + Compression
- Combinations of different encodings

Summary



- We looked at a couple of applied ML approaches for object detection & image classification
- Designing better features is the main deal

References

- Viola & Jones, “Rapid Object Detection using a Boosted Cascade of Simple Features”, CVPR, 2001.
- Dalal & Triggs, “Histograms of Oriented Gradients for Human Detection”, CVPR, 2005.
- Everingham et al., “The Pascal Visual Object Classes Challenge: A Retrospective”, IJCV, vol. 111, pp. 98-136, 2015.
- Felzenszwalb et al., “Object Detection with Discriminatively Trained Part Based Models”, T-PAMI, Vol. 32(9), pp. 1627-1645, 2009.
- Csurka et al., “Visual categorizationwith bags of keypoints“, ECCV SLCV Workshop, 2004.
- Lazebnik et al., “Beyoind Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories”, CVPR, 2006.
- Russakovsky et al., „ImageNet Large Scale Visual Recognition Challenge“. *IJCV,* 2015.
- Perronnin et al., “Fisher kernels on visual vocabularies for image categorization”, CVPR, 2007.

See you in two weeks!