Learning Compact Visual Attributes for Large-scale Image Classification

Yu Su and Frédéric Jurie

Université de Caen Basse Normandie

GREYC-CNRS UMR 6072

Outline

- Motivation
- Method
- Experiments

Image Classification

- **Assign one or multiple labels to an image based on its semantic content.**

→ ***Motorbike***
Person

Image Classification

- **Assign one or multiple labels to an image based on its semantic content.**

Motorbike

Person

- **Small scale datasets**
  - 15Scene (15 classes, ~5K images)
  - PASCAL VOC (20 classes, ~10K images)
  - Caltech101 (101 classes, ~8K images)
- **Large scale datasets**
  - SUN (397 classes, ~100K images)
  - LSVRC (1K classes, ~1.2M images)
  - ImageNet (10K classes, ~9M images)

Image Representation

- **Fisher Vector** [Perronnin et al., ECCV'10]
  - State-of-the-art image representation

	BoW	LLC	Super Vector	Fisher Vector
PASCAL VOC (20 classes)	56.1%	57.6%	58.2%	61.7%
SUN (397 classes)	27.9%	34.1%	35.5%	41.3%

- Bag of Words (BoW) [Sivic&Zisserman, ICCV'03]
- Locality-constrained Linear Coding (LLC) [Wang et al., CVPR'10]
- Super Vector [Zhou et al., ECCV'10]

Image Representation

- **Fisher Vector** **[Perronnin et al., ECCV'10]**
  - **State-of-the-art image representation**

Large Scale Visual Recognition Challenge (LSVRC, 2011)

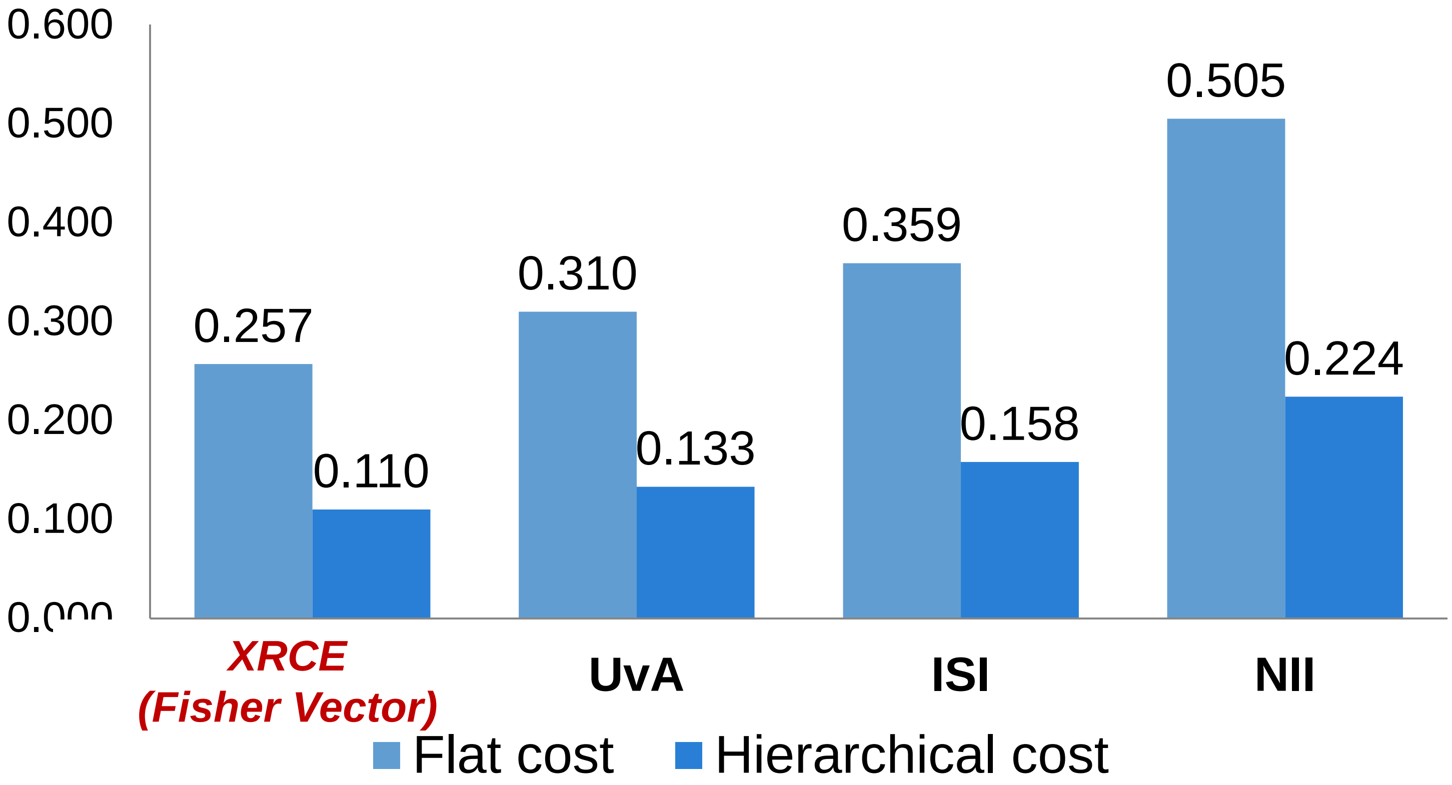
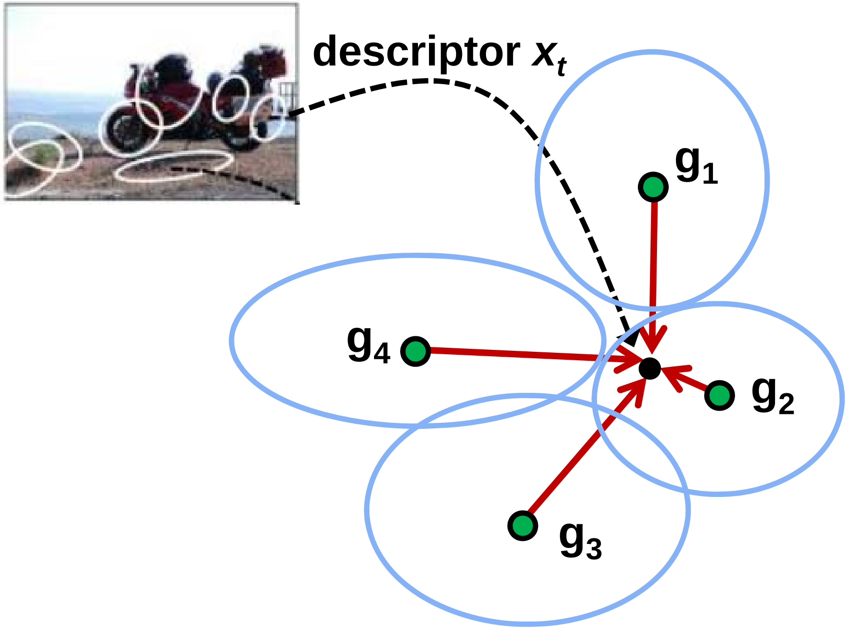


Image Representation

- **Fisher Vector** [Perronnin et al., ECCV'10]
  - **High dimensionality**
    - GMM with 256 components, SIFT reduced to 64-d by PCA
    - Spatial pyramid: 1x1, 2x2, 3x1
    - Fisher Vector: 256x64x2x8 = 262,144-d



Gaussian Mixtures of SIFT

$$G^X_{u,i} = \frac{1}{T\sqrt{w_i}} \sum_{t=1}^{T} \gamma_t(i)\left(\frac{x_t - \mu_i}{\sigma_i}\right)$$

1st order Fisher Vector

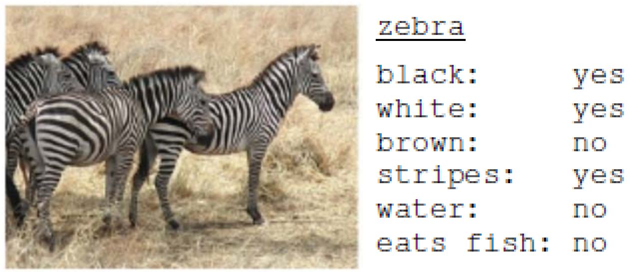
$$G^X_{\sigma,i} = \frac{1}{T\sqrt{2w_i}} \sum_{t=1}^{T} \gamma_t(i)\left[\frac{(x_t - \mu_i)^2}{\sigma_i^2} - 1\right]$$

2nd order Fisher Vector

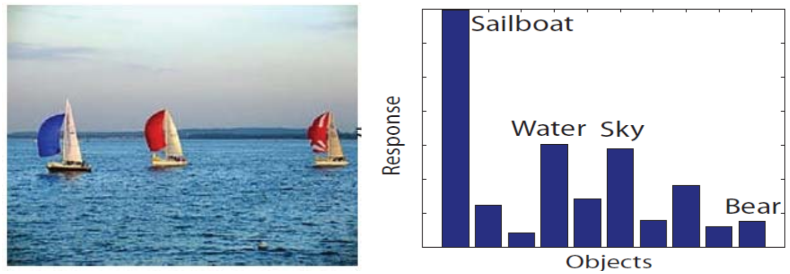
Image Representation

- **Fisher Vector [Perronnin et al., ECCV'10]**
  - **High dimensionality**
    - **GMM with 256 components, SIFT reduced to 64-d by PCA**
    - **Spatial pyramid: 1x1, 2x2, 3x1**
    - **Fisher Vector: 256x64x2x8 = 262,144-d**
  - **Compression**
    - **Product Quantization (PQ)**
    - **Locality-Sensitive Hashing (LSH)**
    - **Principal Component Analysis (PCA)**
    - **Visual Attributes (our work)**

Visual Attributes

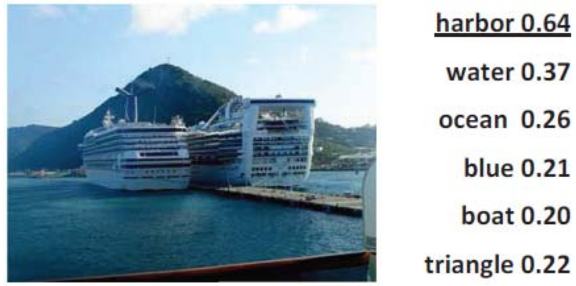


Lampert et al., CVPR'09



Li et al., NIPS'10

Farhadi et al., CVPR'09



Su and Jurie., IJCV'12

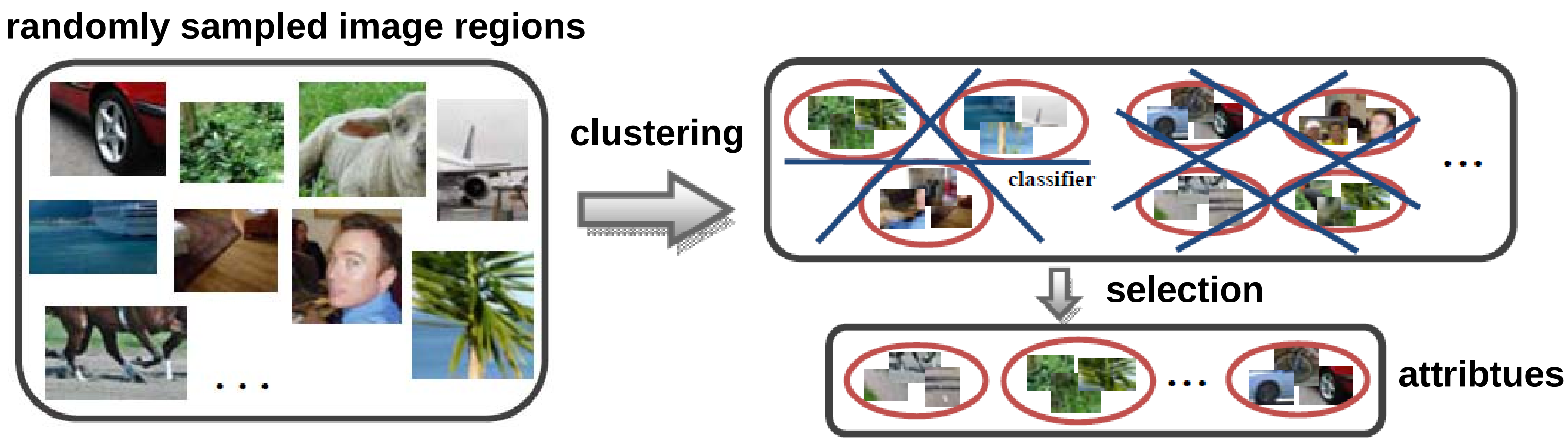
- **Compact image representation**
- **BUT need large amont of human efforts**
  - **Define attributes from expertise or ontology**
  - **Collect and annotate training images**

Outline

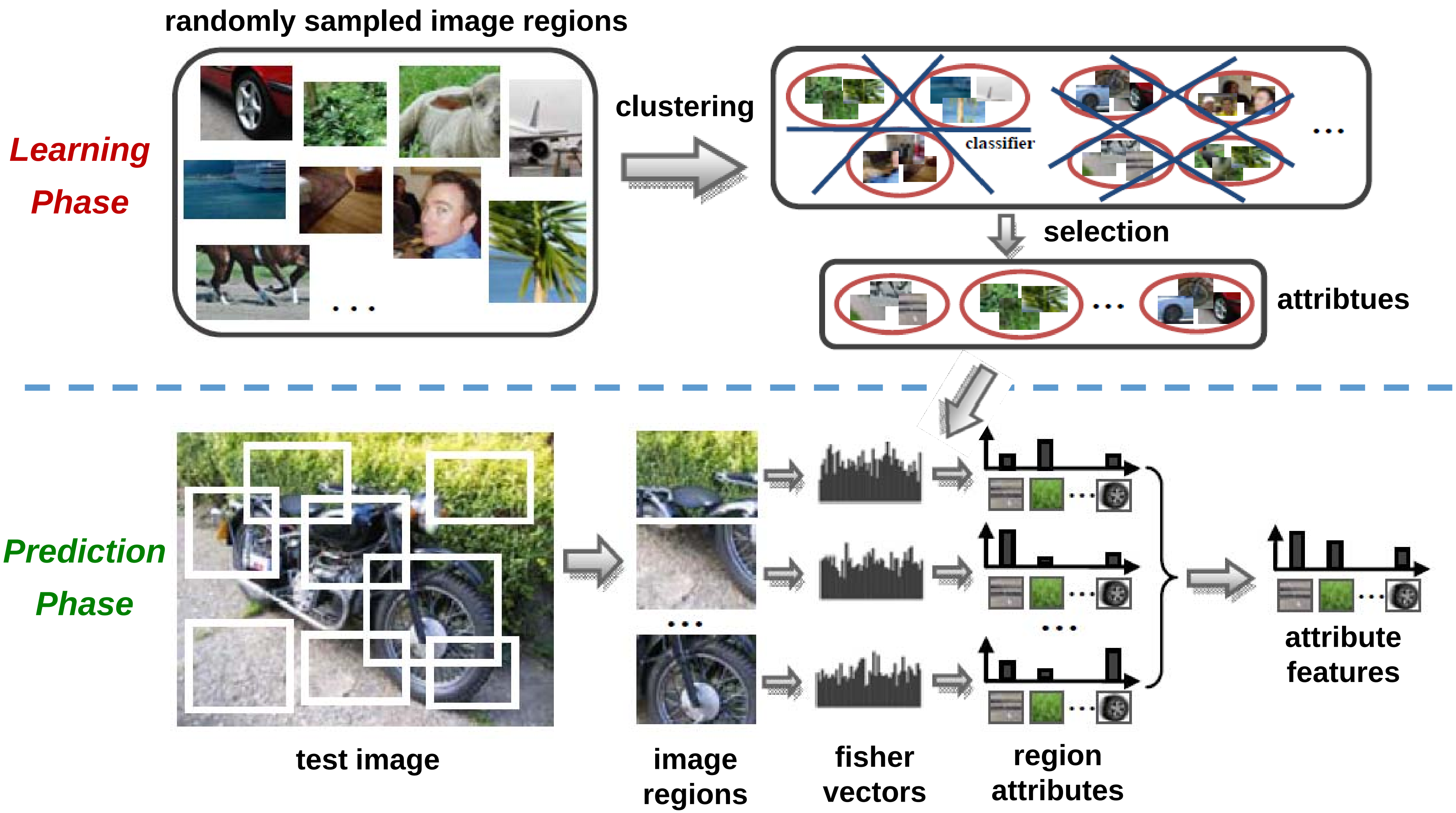
- Motivation
- **Method**
- Experiments

Overview – Region Attributes

Learning Phase

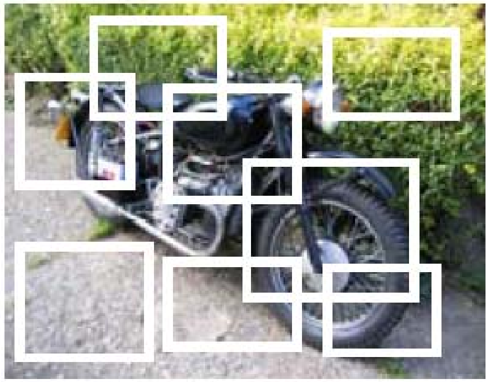


Overview – Region Attributes



Image/Region Representation

- Generate image regions



vs.

Randomly sampling

+ simple, no paras

- less meaningful

Image segmentation

+ semantic meaningful

- many paras, slower

- Fisher Vector

$$G_{u,i}^{X} = \frac{1}{T\sqrt{w_i}} \sum_{t=1}^{T} \gamma_t(i)\left(\frac{x_t - \mu_i}{\sigma_i}\right) \quad G_{\sigma,i}^{X} = \frac{1}{T\sqrt{2w_i}} \sum_{t=1}^{T} \gamma_t(i)\left[\frac{(x_t - \mu_i)^2}{\sigma_i^2} - 1\right]$$

Image/Region Clustering

- Spectral clustering
  - Suits for high-dimensional Fisher Vector (32,768-d)
  - Gaussian kernel as similarity measurement

$$s(f_i, f_j) = \exp(-||f_i - f_j||^2 / 2\sigma^2)$$

- Multi-level clustering
  - # of clusters: 50, 100, ... , 500 (totally 2750 clusters)
- Learn attribute (cluster) classifiers
  - SVM with linear kernel
  - One-vs-rest strategy

Generate Attribute Features

- **Classifier-based soft assignment**

$$\Theta(f, a) = \frac{1}{1 + \exp(-\phi_a(f))}$$

$\Theta(f, a)$ **:** the probability that attribute ***a*** appears in image/region ***f***

$\phi_a(f) = w_a^T f + b_a$ **:** linear classifier (SVM) of attribute ***a***

Image attributes: $\Psi^g(I, a^g) = \Theta(f, a^g)$

Region attributes: $\Psi^l(I, a^l) = \frac{1}{R} \sum_{i=1}^{R} \Theta(f_i, a^l)$

Compact Image Signature

- Attribute selection
  - Objective: compact set of attributes with low redundancy
  - Algorithm: sequential greedy search [Peng et al, PAMI'05]

Compact Image Signature

- **Attribute selection**
  - **Objective: compact set of attributes with low redundancy**
  - **Algorithm: sequential greedy search [Peng et al, PAMI'05]**
- **Binarization**
  - **Locality-Sensitive Hashing: random projection and thresholding.**

$$h_p(x) = \begin{cases} 1, & p^T x \geq 0 \\ 0, & else \end{cases}$$

p : randomly generated projection.

Outline

- Motivation
- Proposed method
- **Experiments**

Examples of Learned Attributes

"horizontal structure"

"vertical structure"

"circular object"

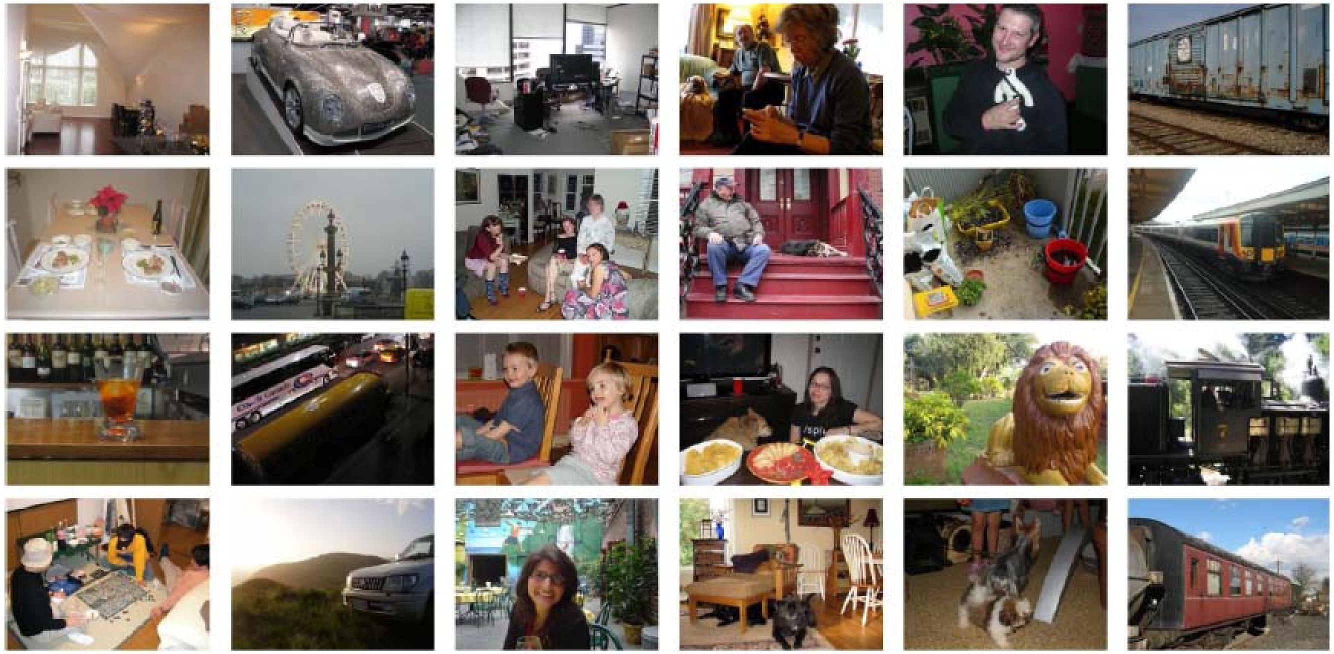
"road/ground"

"group of persons"

"animal in the grass"

Databases

- PASCAL VOC 2007 [Everingham et al., 2007]
  - 20 objects, 9963 images
  - Binary classification
  - Performance measure: mean Average Precision (mAP)

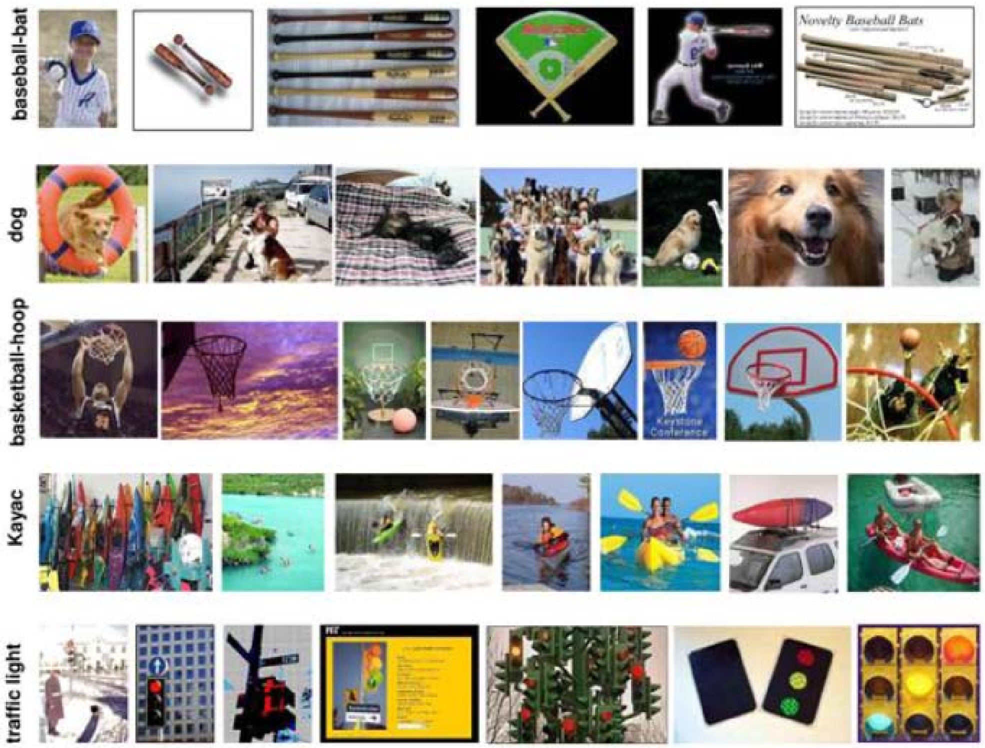


bottle car chair dog plant train

http://raweb.inria.fr/rapportsactivite/RA2007/lear/uid64.html

Databases

- **Caltech-256** **[Griffin et al., CIT-TR, 2007]**
  - **256 objects, ~30K images**
  - **Multi-class classification**
  - **Performance measure: mean accuracy**



http://www.robots.ox.ac.uk/~vgg/software/MKL/

Databases

- **SUN-397 [Xiao et al., CVPR'10]**
  - **397 scenes, ~100K images**
  - **Multi-class classification**
  - **Performance measure: mean accuracy**

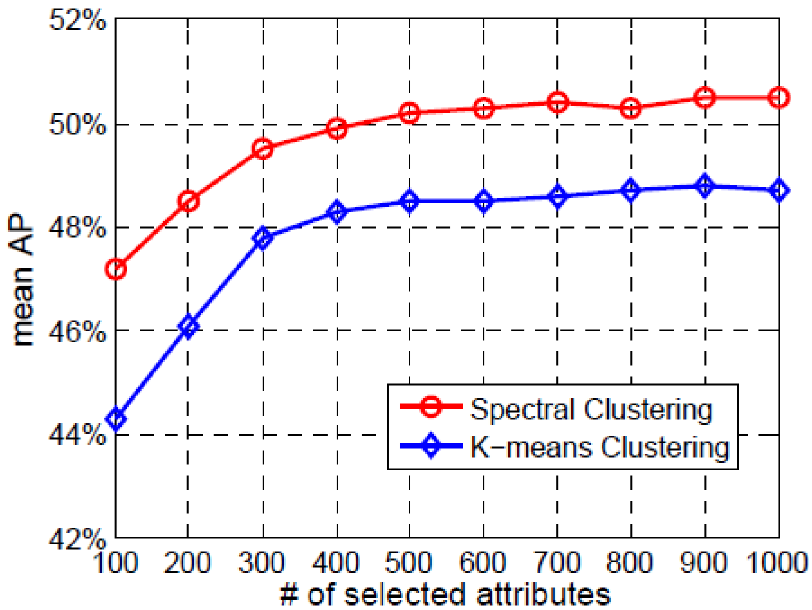
Implementation Details

- SIFT descriptor
  - Densely sampled, reduced to 64-d by PCA
- Fisher Vector
  - GMM with 256 components
  - Dimension: 256x64x2=32,768
- Image classification
  - SVM with linear kernel
  - λ is determined on PASCAL train/val set

Attribute learning (including clustering, feature selection etc.) is ONLY performed on PASCAL train/val set.

Learn & Predict Attribute

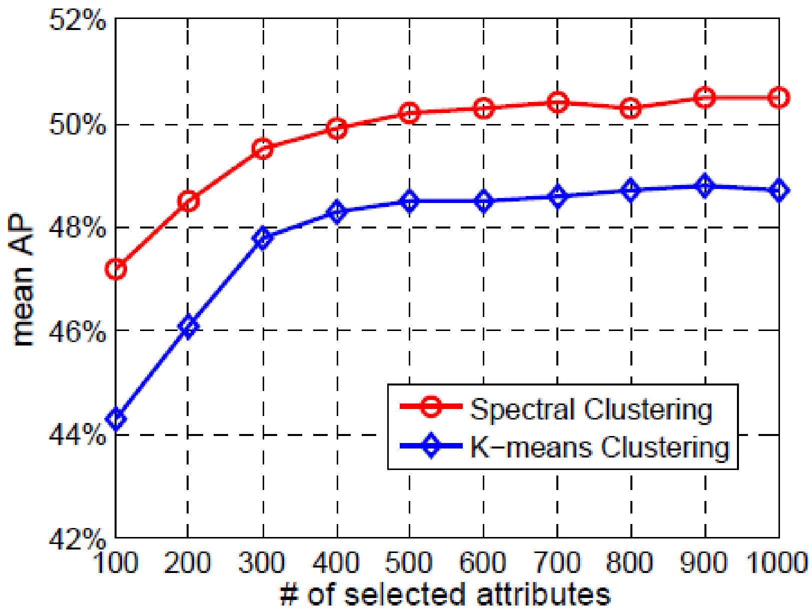
PASCAL VOC 2007 train/validation



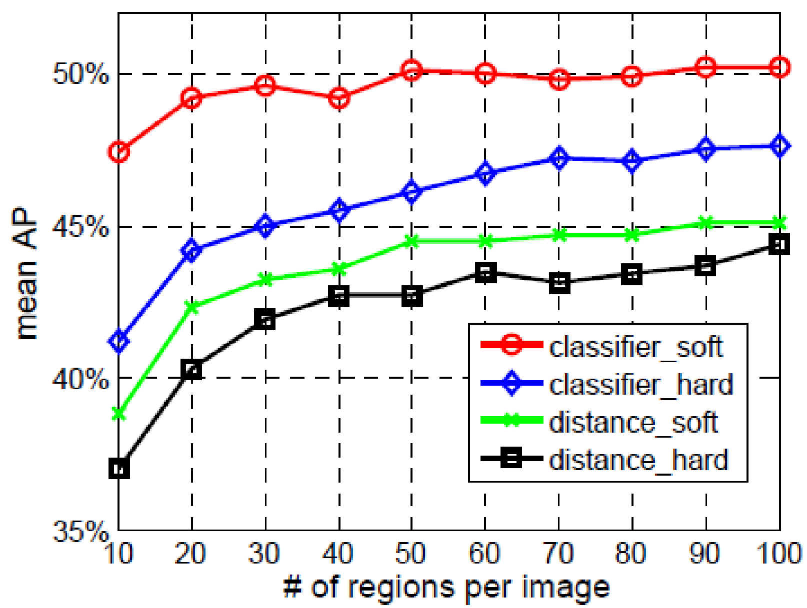
Spectral Clustering vs. K-means

Learn & Predict Attribute

PASCAL VOC 2007 train/validation

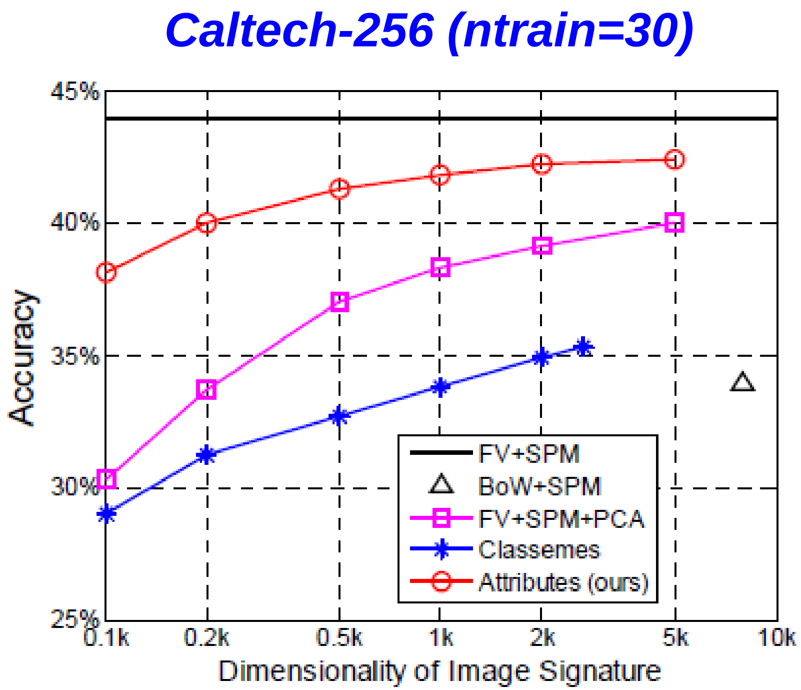


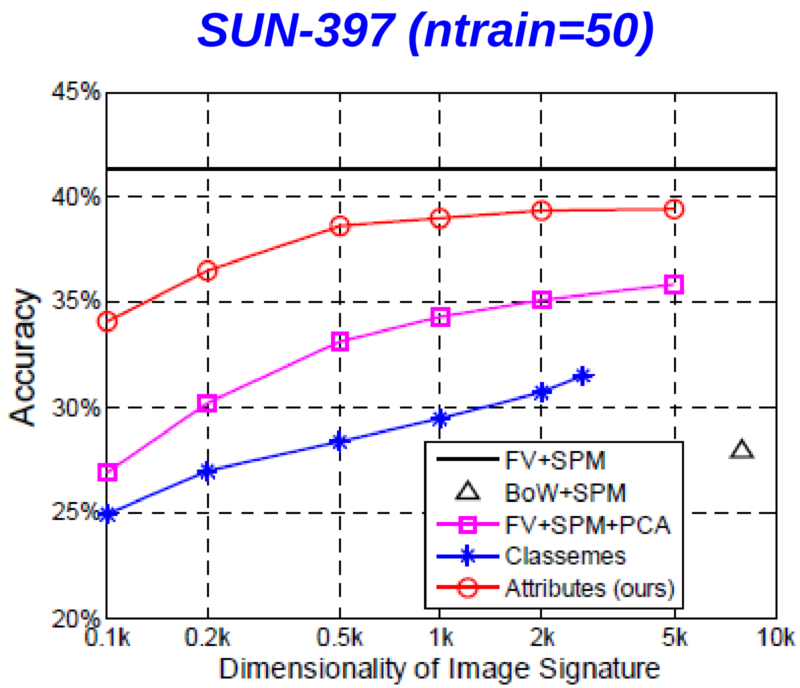
Spectral Clustering vs. K-means



Different Encoding Methods

Real-valued Attribute Feature

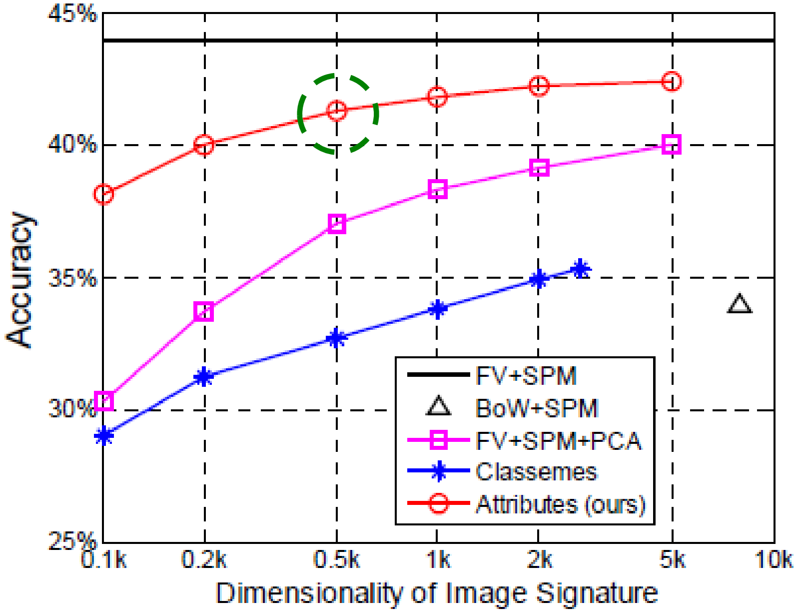




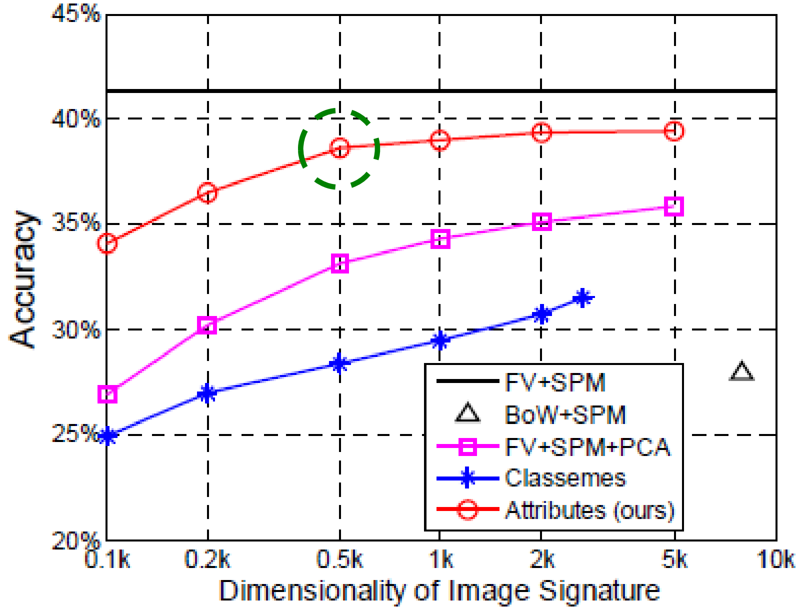
- **FV with SPM (1x1, 2x2, 3x1) : 262,144-d**
- **FV+SPM+PCA: PCA is learnt on PASCAL VOC**
- **Classemes [Torresani, ECCV'10]: multiple low-level features**
- **Our method:**
  (1) 500 times more compact than FV+SPM with 3% performance loss
  (2) better than PCA and Classemes

Real-valued Attribute Feature

Caltech-256 (ntrain=30)

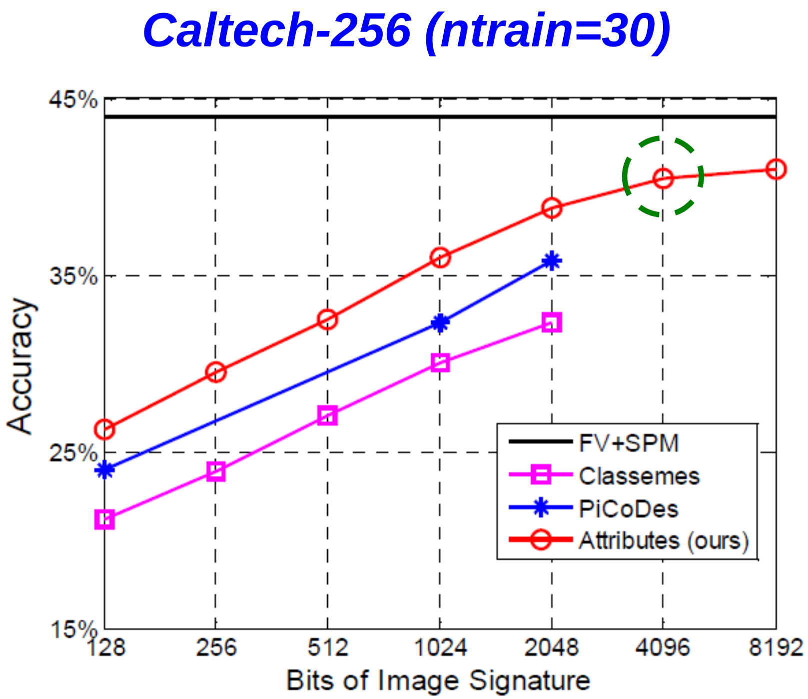


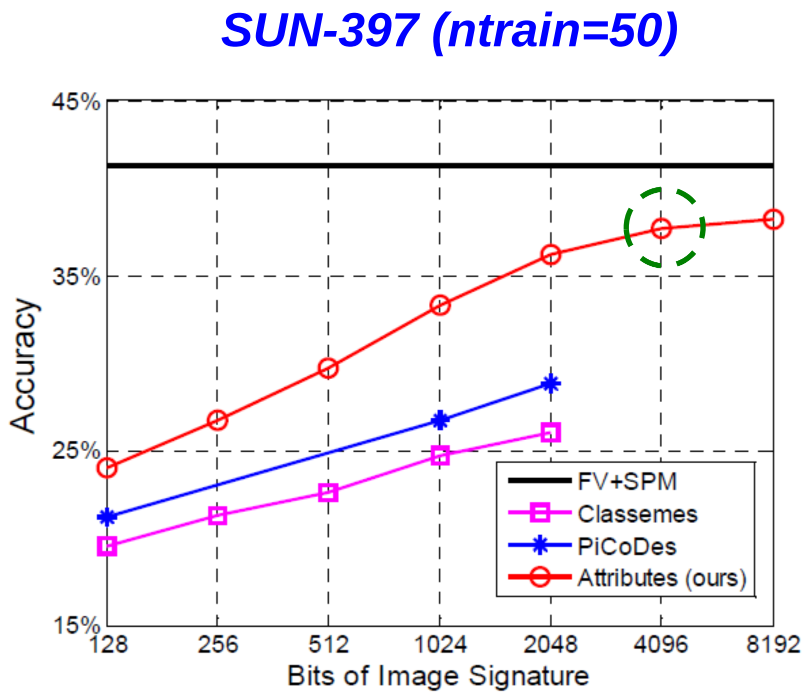
SUN-397 (ntrain=50)



- **FV with SPM (1x1, 2x2, 3x1) : 262,144-d**
- **FV+SPM+PCA: PCA is learnt on PASCAL VOC**
- **Classemes [Torresani, ECCV'10]: multiple low-level features**
- **Our method:**
  (1) 500 times more compact than FV+SPM with 3% performance loss
  (2) better than PCA and Classemes

Binary Attribute Feature





- **FV+SPM: 262,144 x 4 bytes**
- **Classemes [Torresani, ECCV'10] : binarized by thresholding**
- **PiCoDes [Bergamo, NIPS'11]: optimizing an independent classification task**

- **Our method:**
  (1) 2048 times more compact than FV+SPM with 3% performance loss
  (2) better than Classemes and PiCodes

Thanks for your attention !