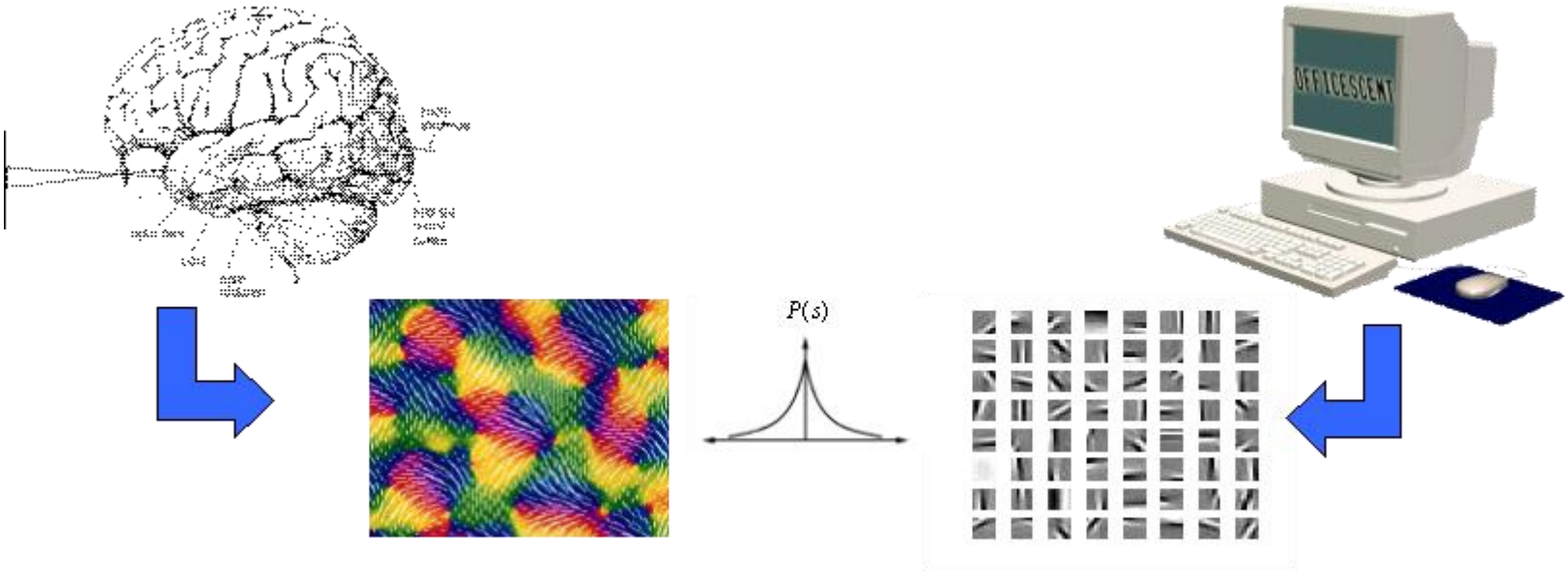
Sparse coding for image classification



Columbia University Electrical Engineering:
Kun Rong(kr2496@columbia.edu)
Yongzhou Xiang(yx2211@columbia.edu)
Yin Cui(yc2776@columbia.edu)

Outline

Background

Introduction to Sparse Coding

Introduction to Image classification

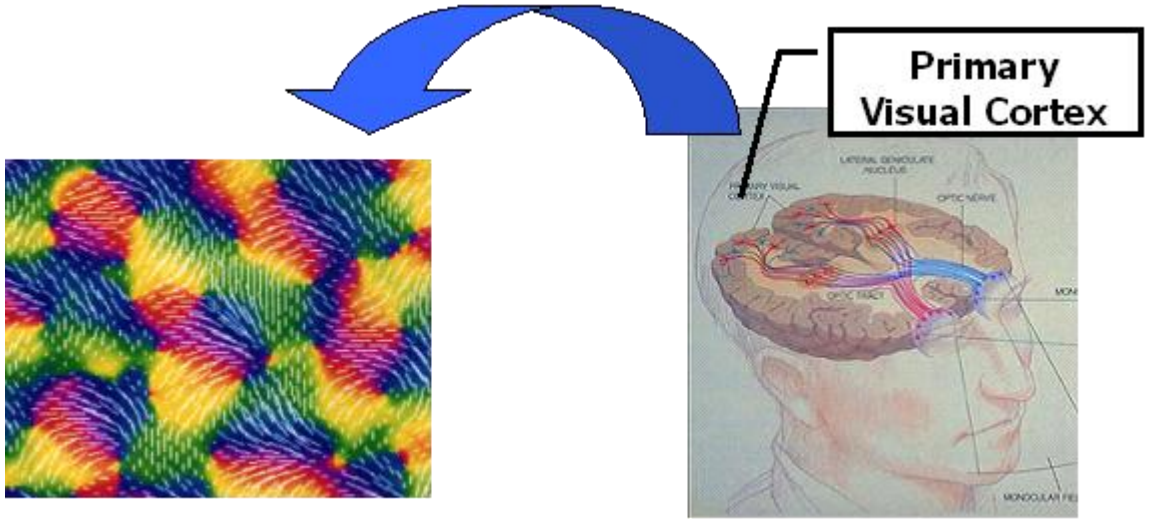
- Feature Extraction
- Building Dictionary
- Coding
- Pooling
- Classification

Sparse Coding for Image Classification

- Vector quantization
- Sparse coding
- Locality-constrained Linear Coding
- Experiment results & Application

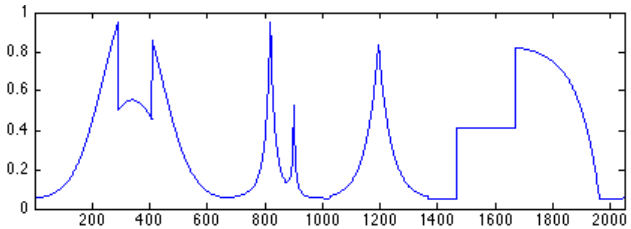
Background

- (Hubel, Weisel, 1962-68) The human visual system, at the primary cortex (V1), has receptive fields that are:
- spatially localized
- oriented
- bandpass

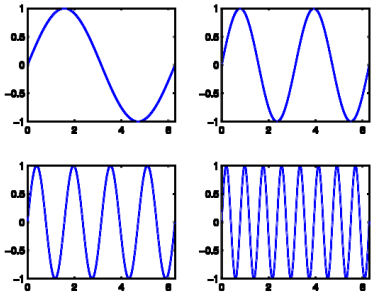


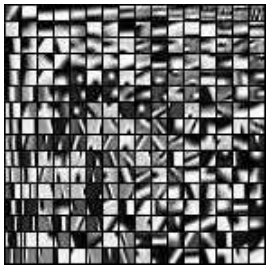
Background

- Let x in Rm be a signal.



- Let D = [d1, . . . , dp] $\in R^{m \times p}$ be a set of normalized "basis vectors"





- D is "adapted" to x if it can represent it with a few basis vectors—that is, there exists a sparse vector in Rp such that x ≈ D . We call the sparse code.

$$\underbrace{\begin{pmatrix} \mathbf{x} \end{pmatrix}}_{\mathbf{x} \in \mathbb{R}^m} \approx \underbrace{\left(\begin{array}{c|c|c|c} \mathbf{d}_1 & \mathbf{d}_2 & \cdots & \mathbf{d}_p \end{array} \right)}_{\mathbf{D} \in \mathbb{R}^{m \times p}} \underbrace{\begin{pmatrix} \alpha[1] \\ \alpha[2] \\ \vdots \\ \alpha[p] \end{pmatrix}}_{\alpha \in \mathbb{R}^p, \text{sparse}}$$

Introduction to Sparse Coding

- Applications
  - Image processing
  - Image classification

Sparse Coding in Image Processing

- Sparse Coding could be used in image processing.
- Two examples will be shown.

Denoising

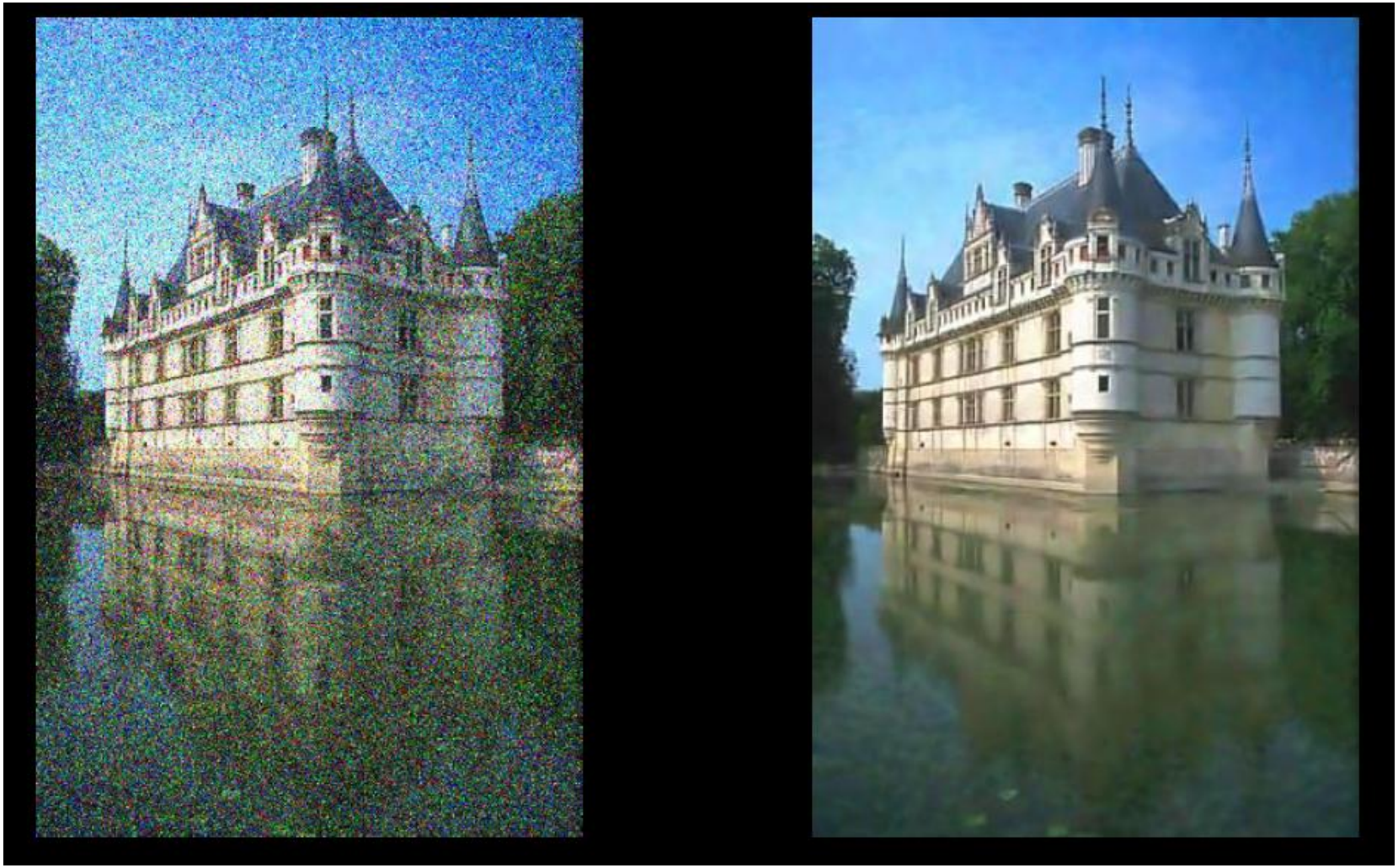
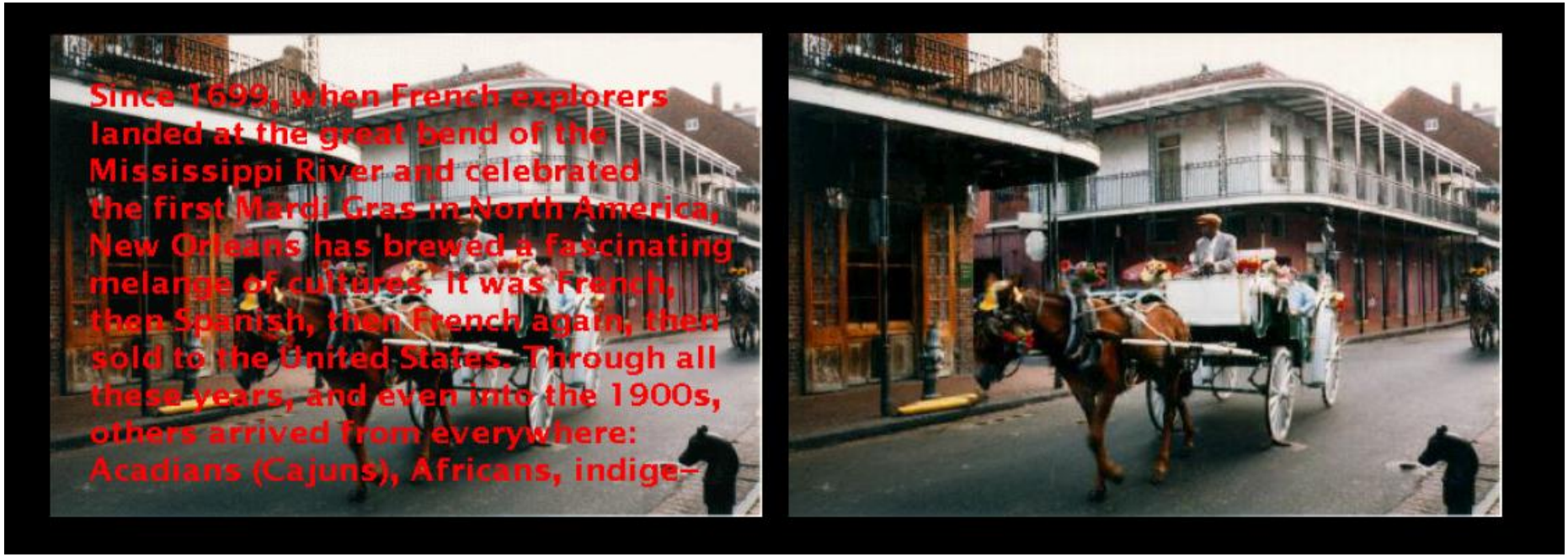
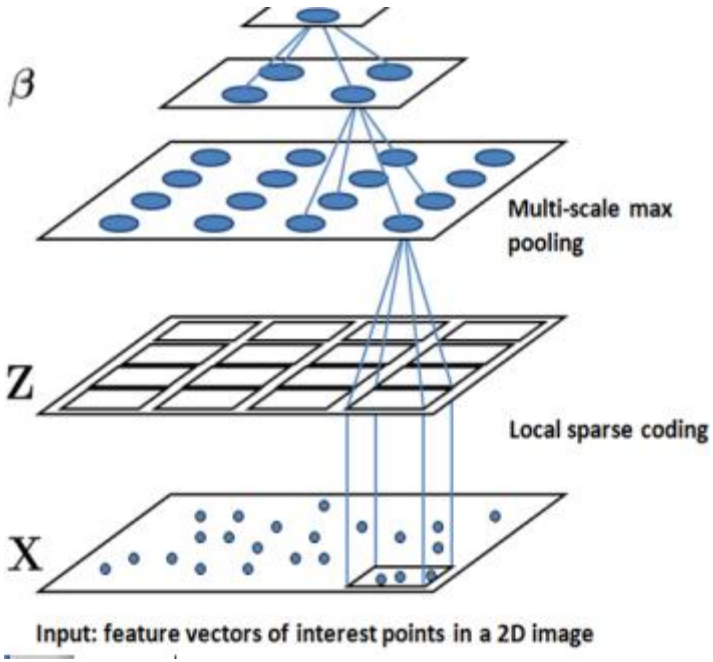


Image Inpainting



Sparse Coding in Image Classification



Introduction to Image Classification

- Tasks: assigning the input image with one of the predefined classes.

Image Classification is difficult!

- Problems: view point changes, lighting changes, high intra-class variations, background noise.

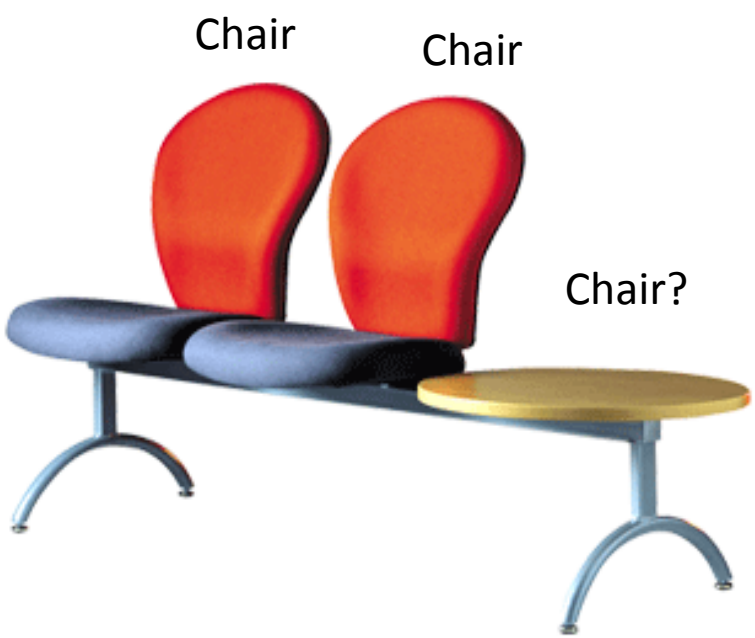
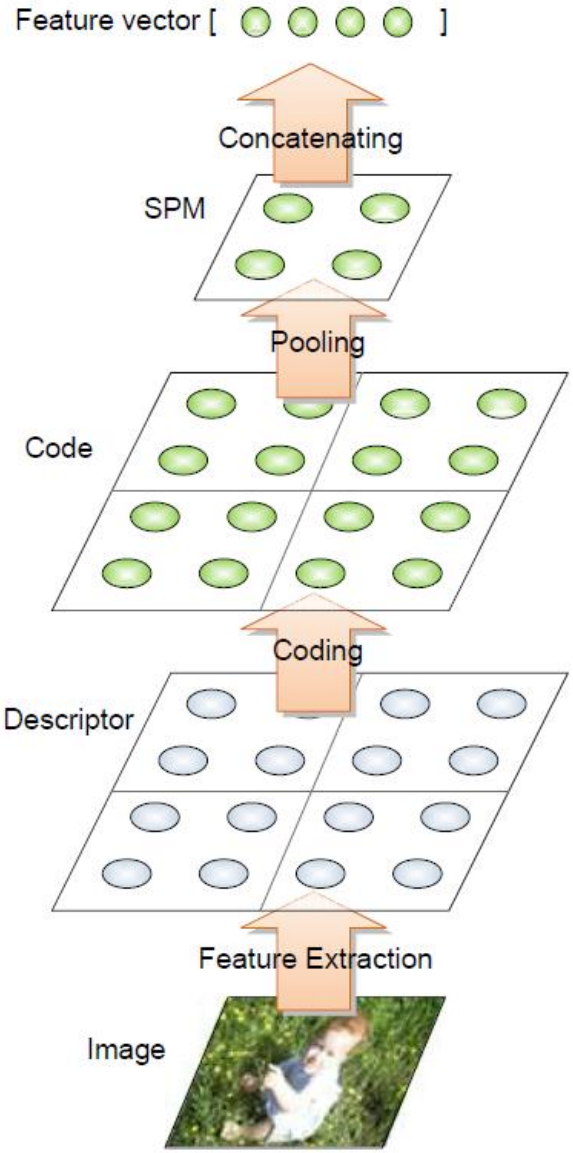


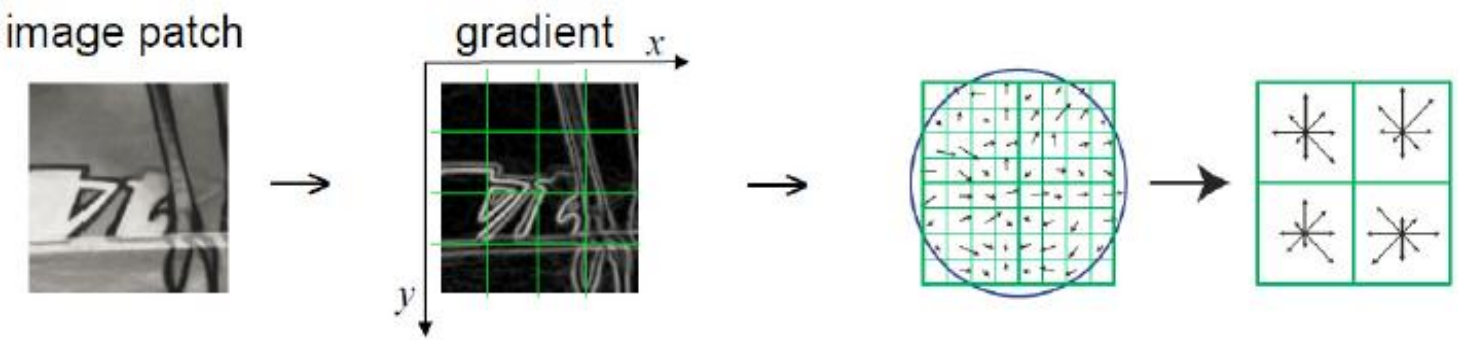
Image Classification Framework

- Feature Extraction
- Building Dictionary
- Coding
- Pooling
- Classification



Feature Extraction

- Local Descriptors-SIFT



Compute image gradients at each location

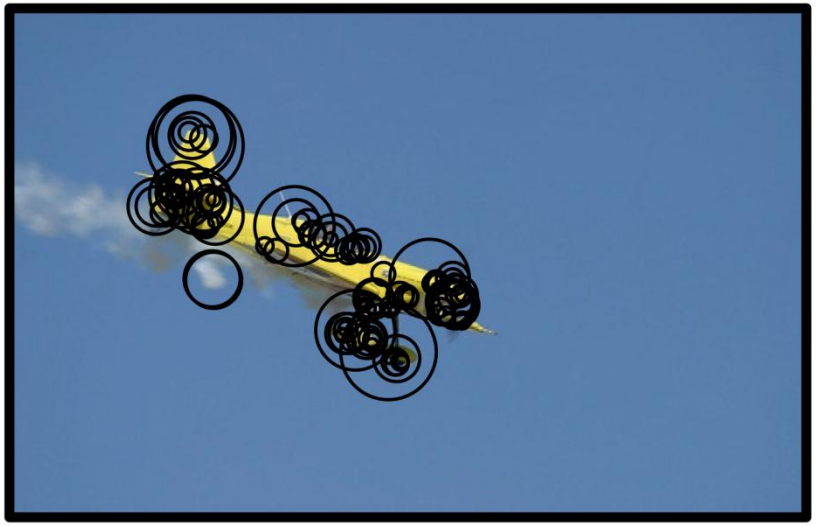
- **Histogram of the gradient orientation (8 bins)**

- **Concatenate histograms of 4 x 4 spatial grid**

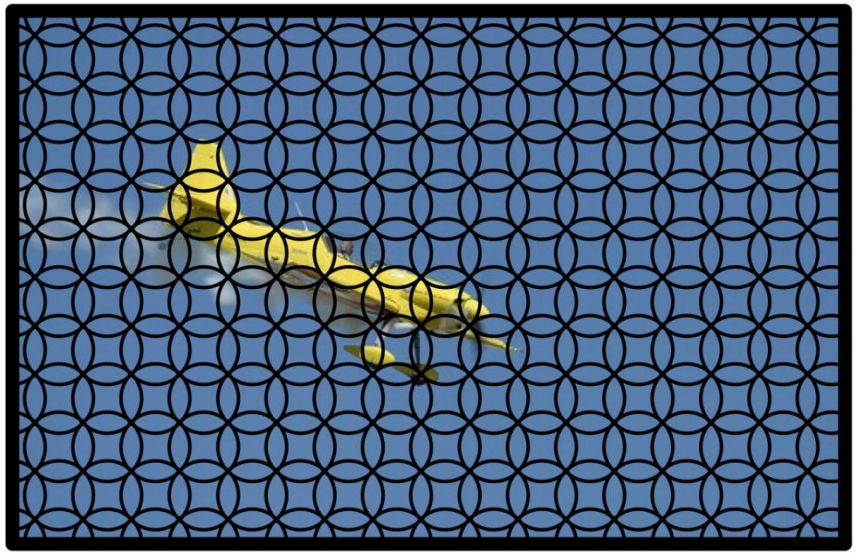
- **Feature Dimension:**
8 orientation x (4x4) spatial grid = 128

Feature Extraction

- **We often use dense sift to capture global image information, as follow:**



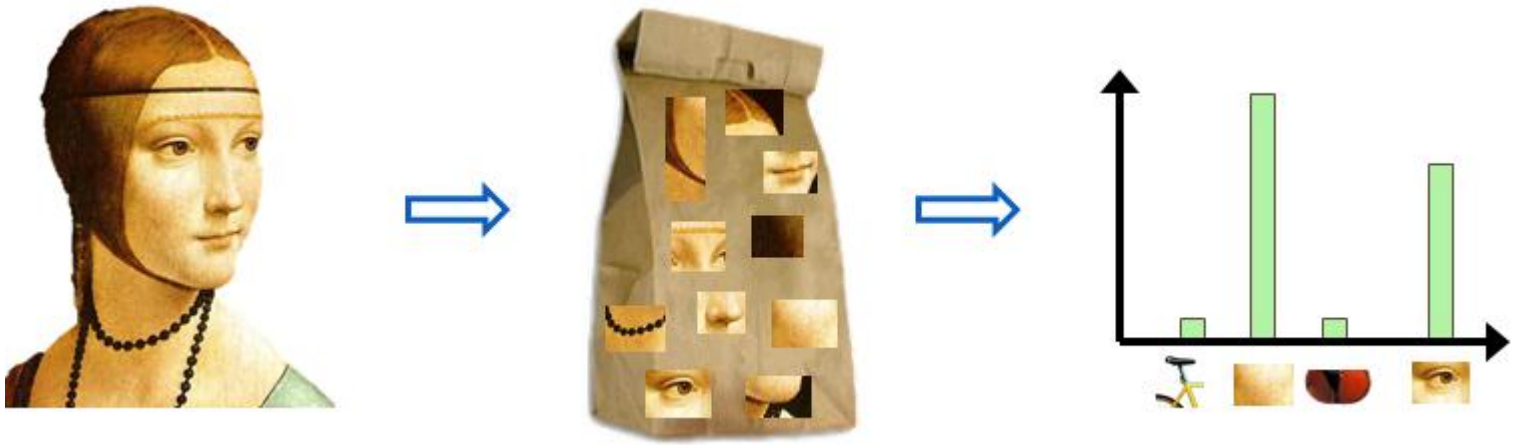
Harris-Laplace detector



Dense sampling detector

Bag-of-feature Model

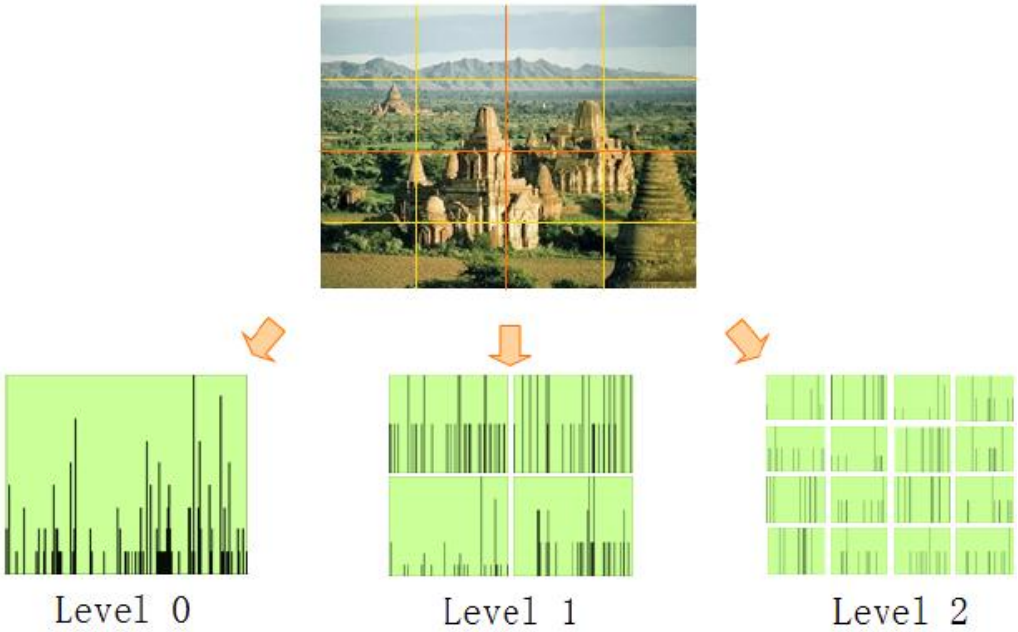
- SIFT + Bag of Word representation.
- Treat image as a collection of local “visual words”
- Vector quantize the “visual words” to get a histogram representation



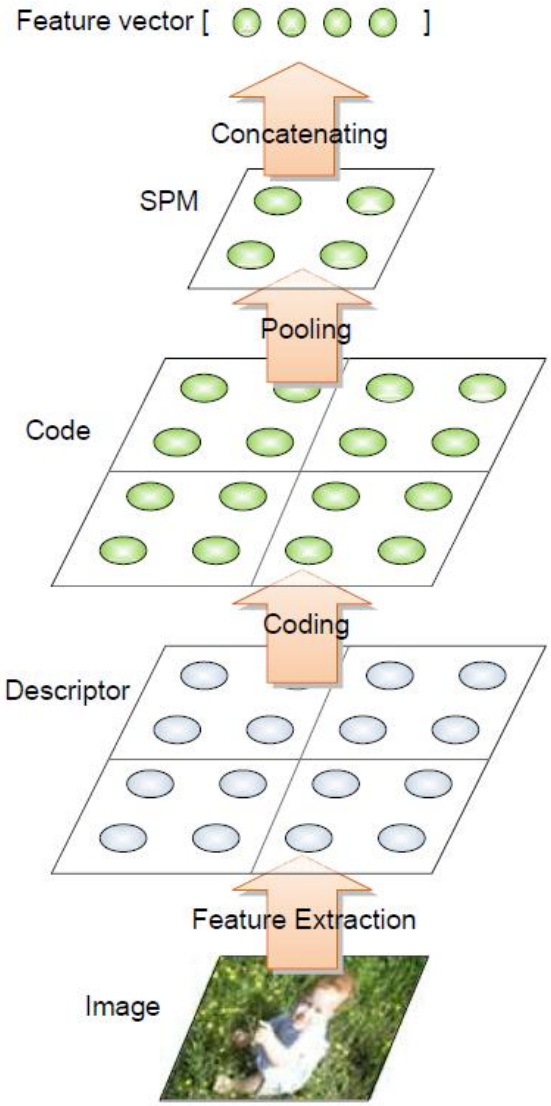
- However, can not capture the spatial information of these local “visual words”.

Spatial Pyramid Matching

- SPM(spatial pyramid matching)
- Pooling histograms from different spatial locations across multiple spatial scales



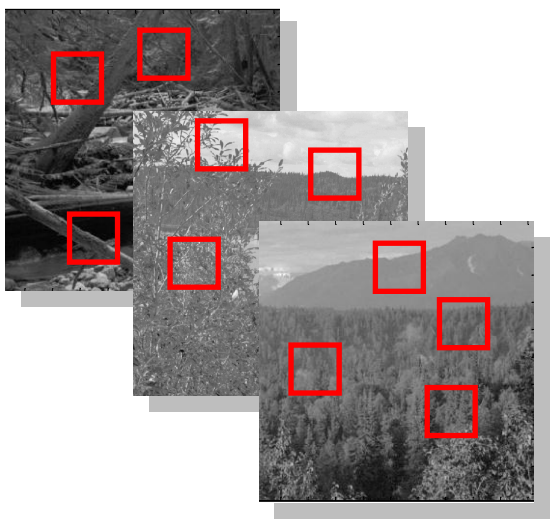
Building Dictionary



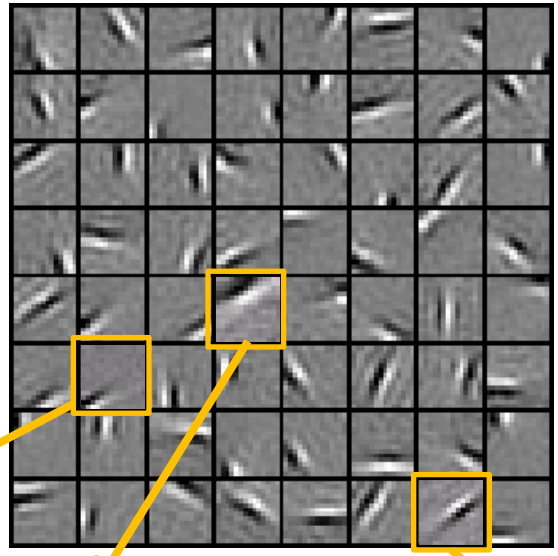
Building Dictionary

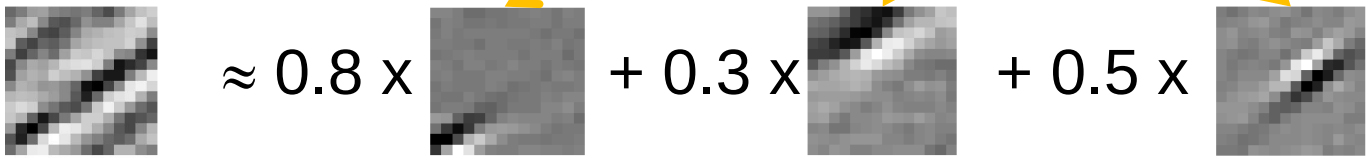
- K-means; Sparse Coding; Structured Dictionary
- Example:

Natural Images



Dictionary with bases ($\phi_1, \ldots, \phi_{64}$)



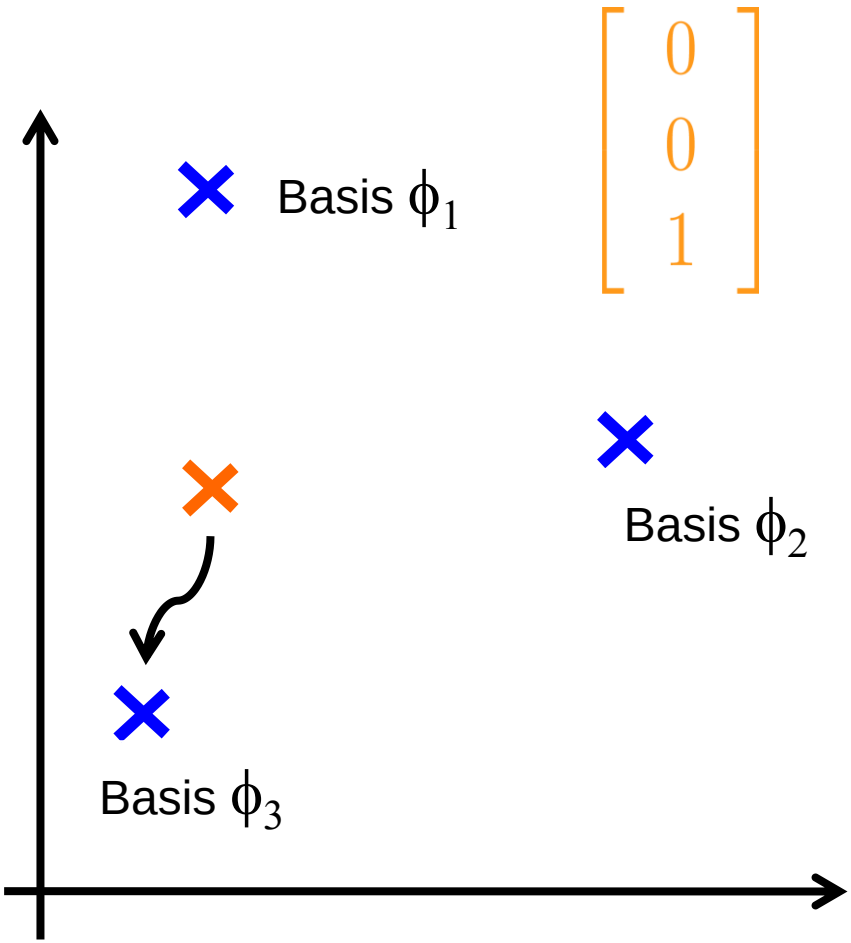


$\approx 0.8 \times$ $+ 0.3 \times$ $+ 0.5 \times$

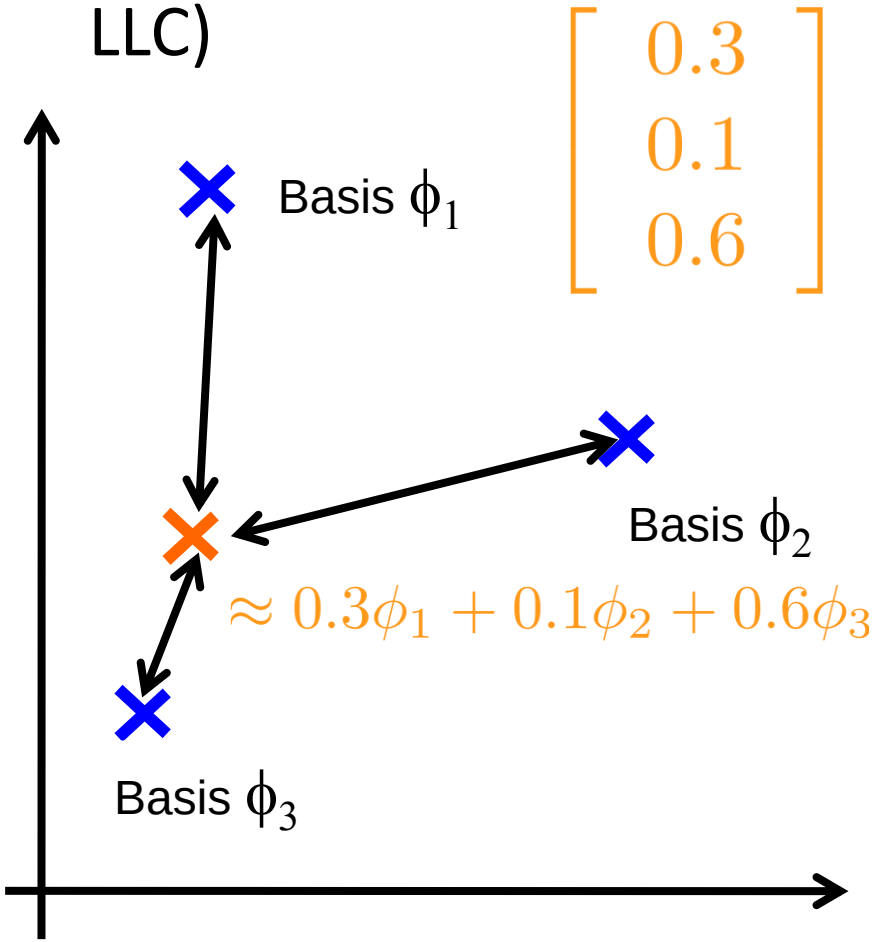
$= [0, \ldots, 0, \mathbf{0.8}, 0, \ldots, 0, \mathbf{0.3}, 0, \ldots, 0, \mathbf{0.5}, \ldots] \times [\phi_1, \ldots, \phi_{64}]^T$

Coding

- Representing descriptors by dictionary bases.
- Hard Quantization (VQ)



- Soft Quantization (SC, LLC)



Pooling

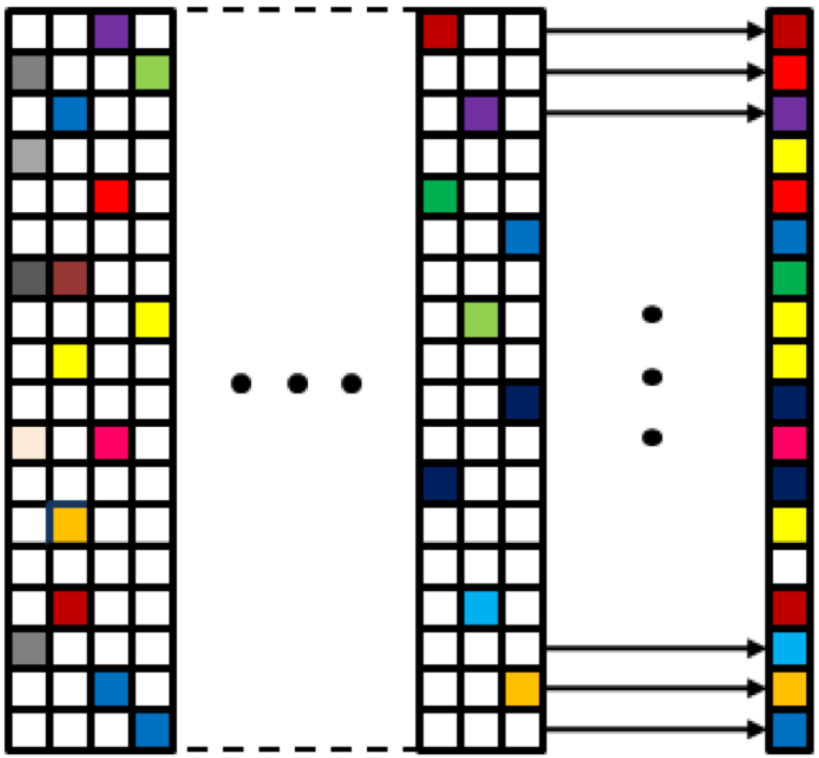
Combining feature detectors into a ‘global feature’ that preserves task-related information while removing irrelevant details.

average pooling:

$$f_a(k) = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{z}_i(k), \;\; k = 1, 2, ..., K.$$

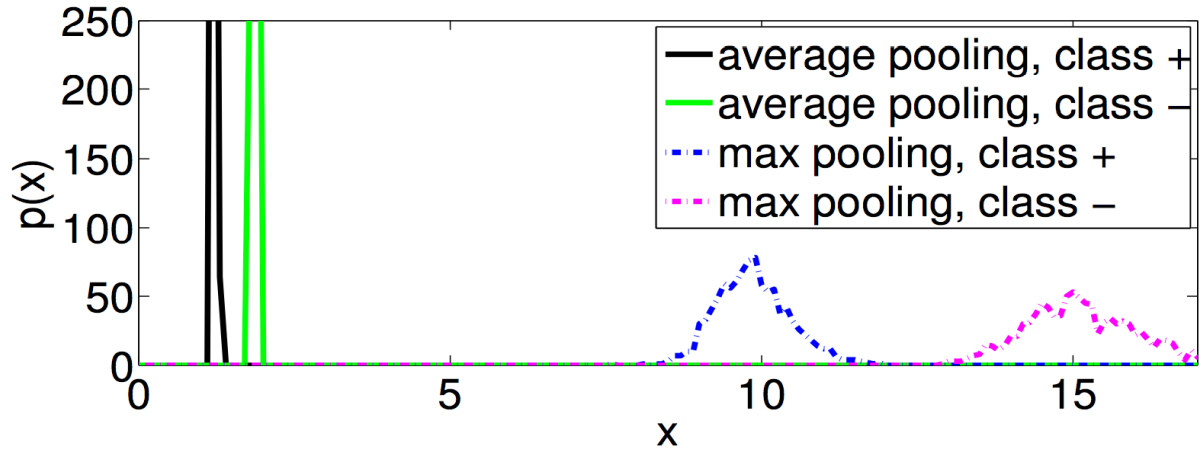
max pooling:

$$f_m(k) = \max_i |\boldsymbol{z}_i(k)|, \;\; k = 1, 2, ..., K.$$

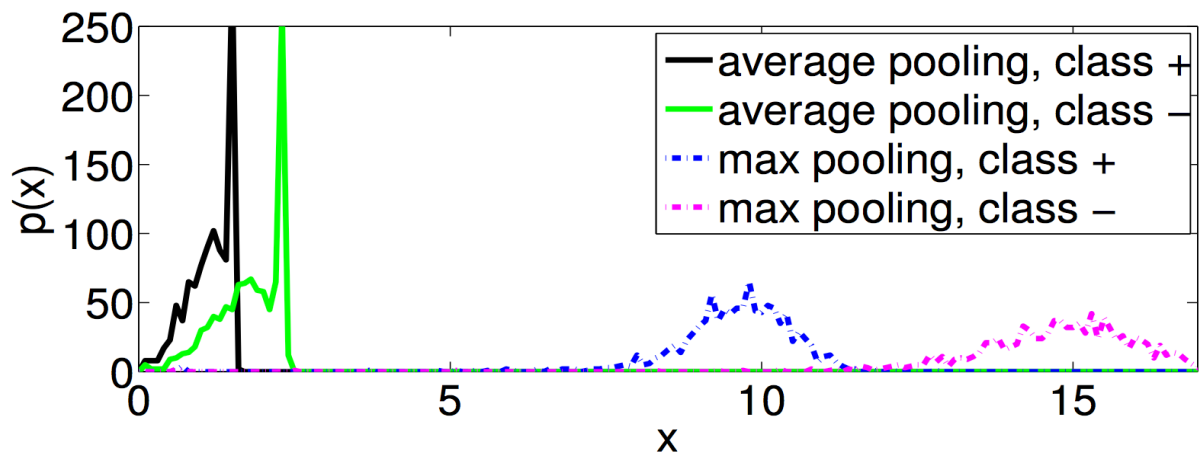


Average Pooling vs. Max Pooling

Clutter is homogeneous across images



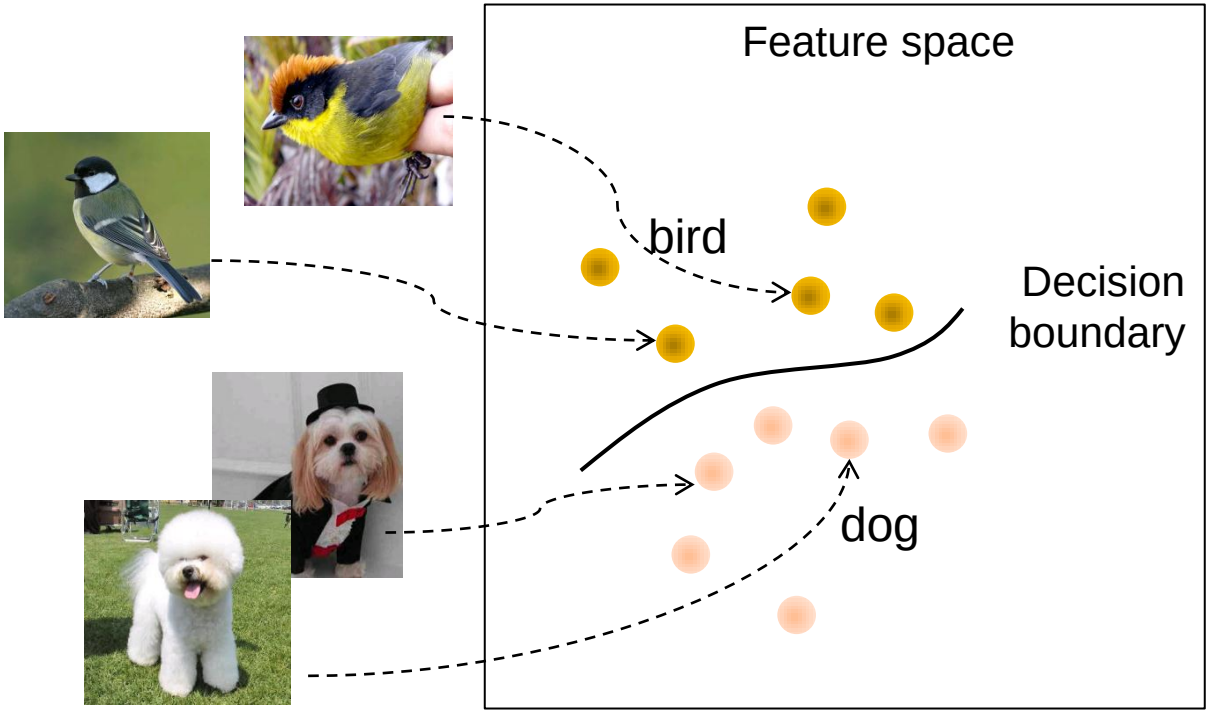
Clutter level has high variance



When the clutter has high variance, distributions remain well separated with max pooling, but have significant overlap with average pooling.

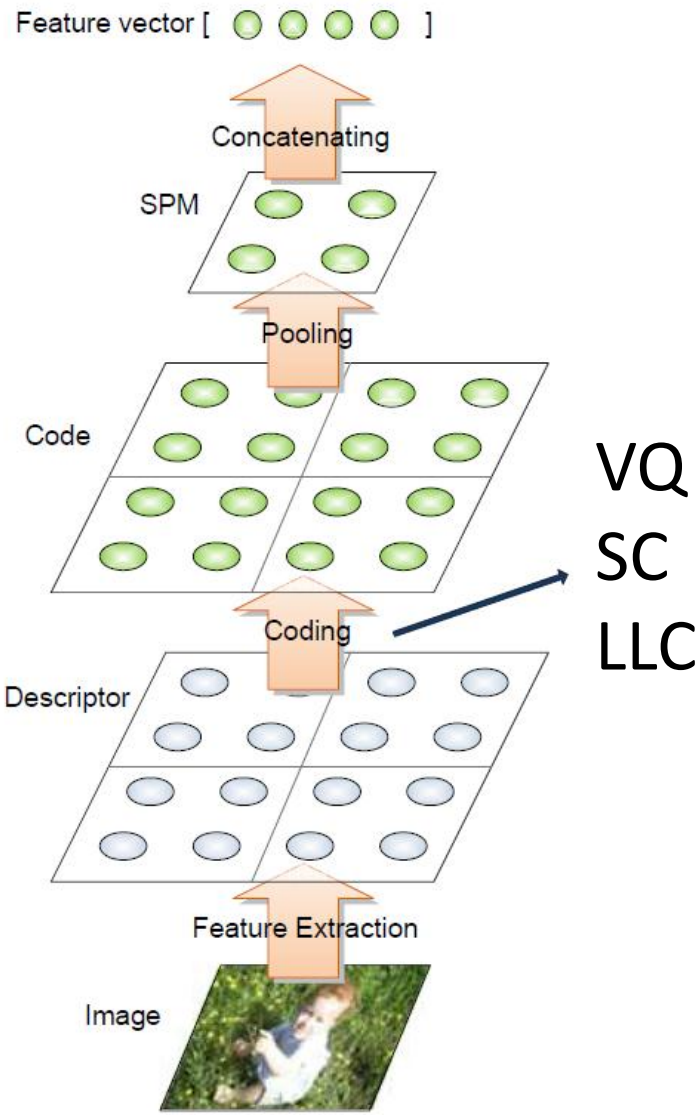
Classification

- Usually SVM with linear or non-linear kernels (intersection, chi-square, etc.)
- One vs all strategy for multiclass

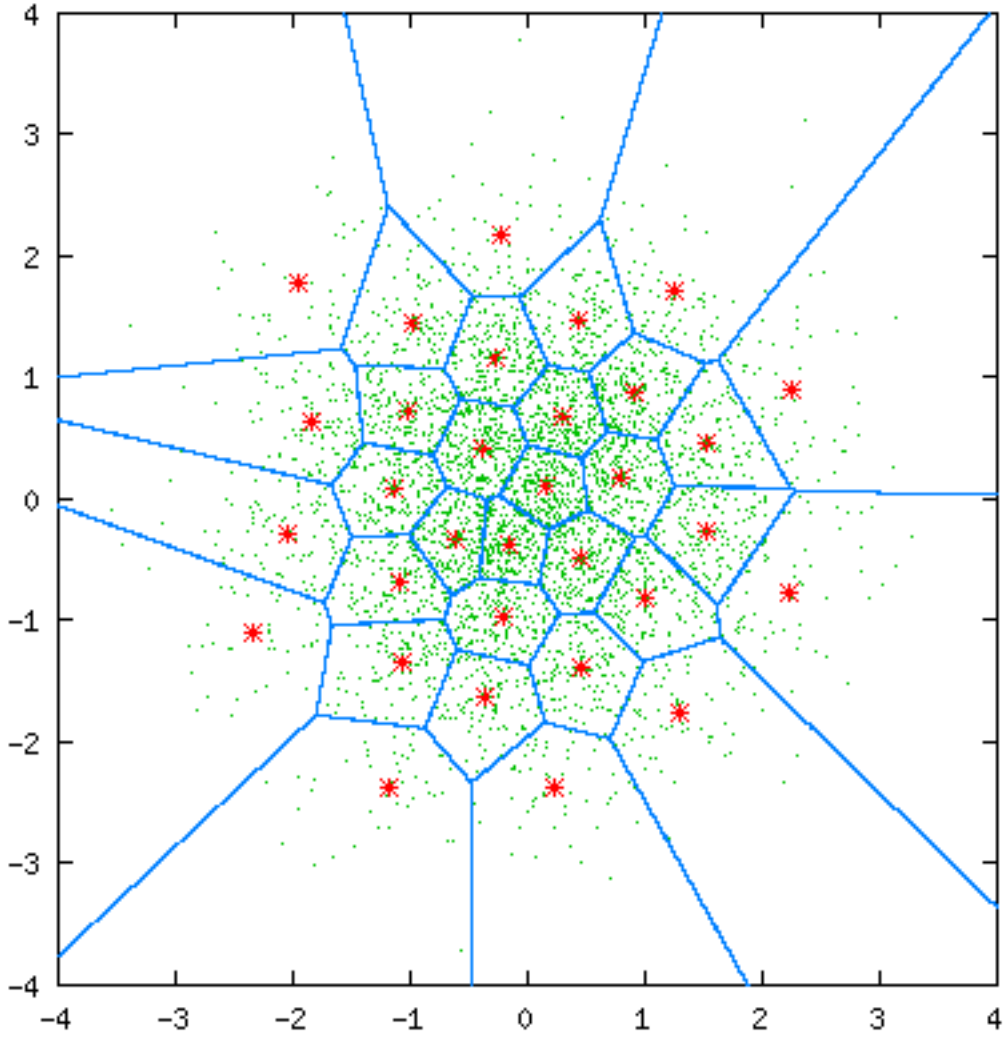


Sparse Coding for Image Classification

- Feature descriptor obtained from dense SIFT and BoF representation.
- To improve image classification accuracy, well-designed coding method is needed.



Vector Quantization



Vector Quantization

- Let X be a set of SIFT appearance descriptors in a D- dimensional feature space, i.e. X = [x1 , . . . , xM]T ∈ RM×D. The vector quantization (VQ) method applies the K-means clustering algorithm to solve the following problem

$$\min_{\mathbf{V}} \quad \sum_{m=1}^{M} \min_{k=1\ldots K} \|\mathbf{x}_m - \mathbf{v}_k\|^2$$

- where V = [v1,...,vK]T are the K cluster centers to be found, called codebook, and || · || denotes the L2-norm of vectors.

Vector Quantization

- The optimization problem can be re-formulated into a matrix factorization problem with cluster member- ship indicators U = [u1,...,uM]T

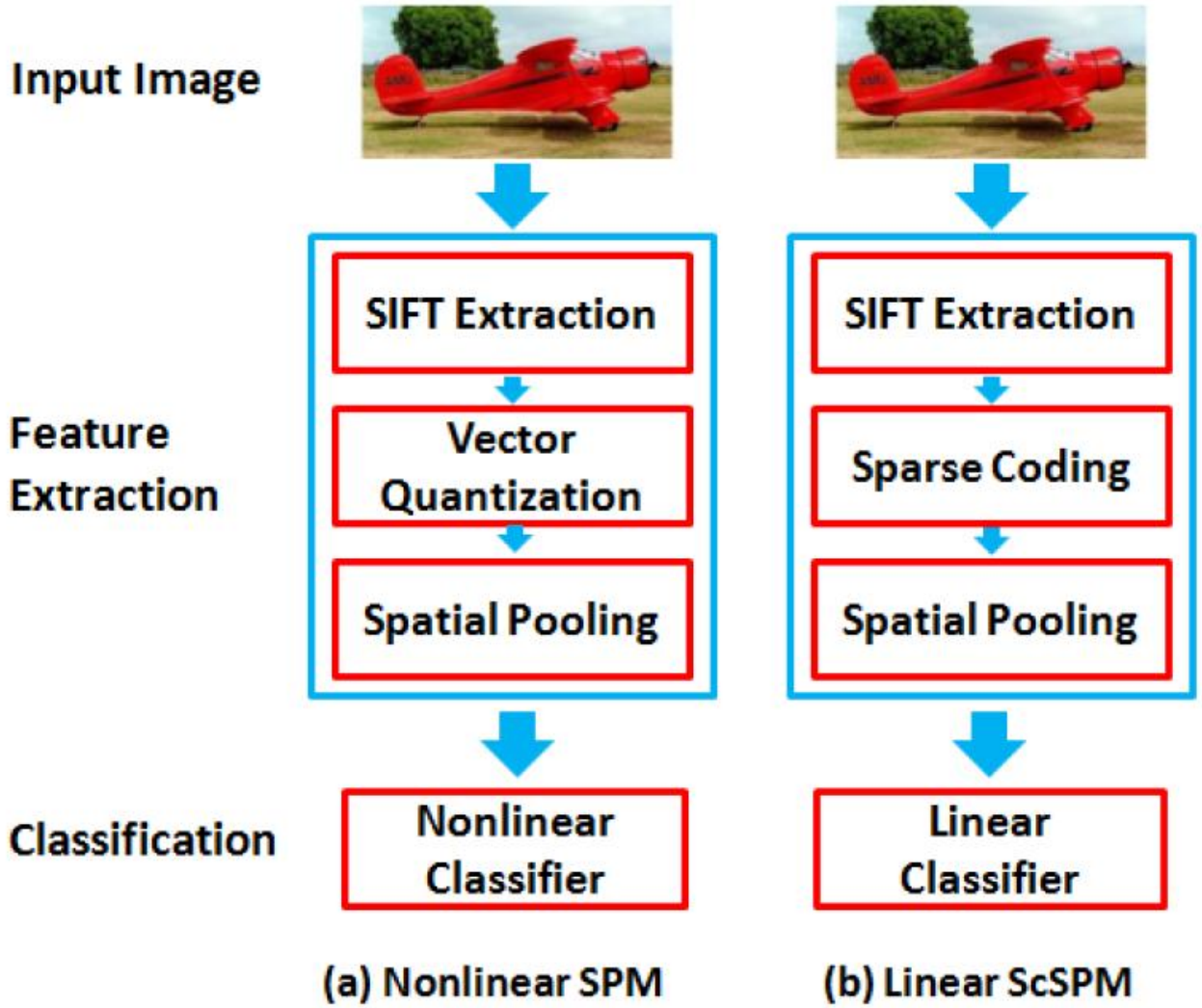
$$\min_{\mathbf{U},\mathbf{V}} \quad \sum_{m=1}^{M} \|\mathbf{x}_m - \mathbf{u}_m \mathbf{V}\|^2$$
$$\text{subject to} \quad Card(\mathbf{u}_m) = 1, |\mathbf{u}_m| = 1, \mathbf{u}_m \succeq 0, \forall m$$

- In training phase: the optimization is solved with respect to both U and V.
- In coding phrase:V will be applied for a new set of X

Vector Quantization

- Drawbacks of VQ:
  - Large quantization error
  - Non-linearly separable (SVM with nonlinear kernel).
- Drawbacks of nonlinear SVM:
  - nonlinear kernels, e.g. the intersection kernel or the chi-square kernel.
  - computational complexity: O(n^3)
  - memory complexity: O(n^2) in the training phase (n is the training size)

Sparse Coding



Sparse Coding

- Sparse Coding is an extension of VQ.

$$\min_{\mathbf{U},\mathbf{V}} \quad \sum_{m=1}^{M} \|\mathbf{x}_m - \mathbf{u}_m \mathbf{V}\|^2 + \lambda |\mathbf{u}_m|$$
$$\text{subject to} \quad \|\mathbf{v}_k\| \leq 1, \quad \forall k = 1, 2, \ldots, K$$

$$\min_{\mathbf{U},\mathbf{V}} \quad \sum_{m=1}^{M} \|\mathbf{x}_m - \mathbf{u}_m \mathbf{V}\|^2$$
$$\text{subject to} \quad Card(\mathbf{u}_m) = 1, |\mathbf{u}_m| = 1, \mathbf{u}_m \succeq 0, \forall m$$

- constraint Card(um) = 1 may be too restrictive
- Not putting a L1-norm regularization on U_m, which enforces U_m to have a small number of nonzero elements.

Sparse Coding

- A unit L2-norm constraint on V_k is applied to avoid trivial solutions.
- Similar to VQ, SC has a training phase and a coding phase.
- Training phase: solving the equation with respect to U (code) and V (dictionary) alternatively.
- Coding phase: for each descriptor X, the SC codes are obtained by optimizing with respect to U only.

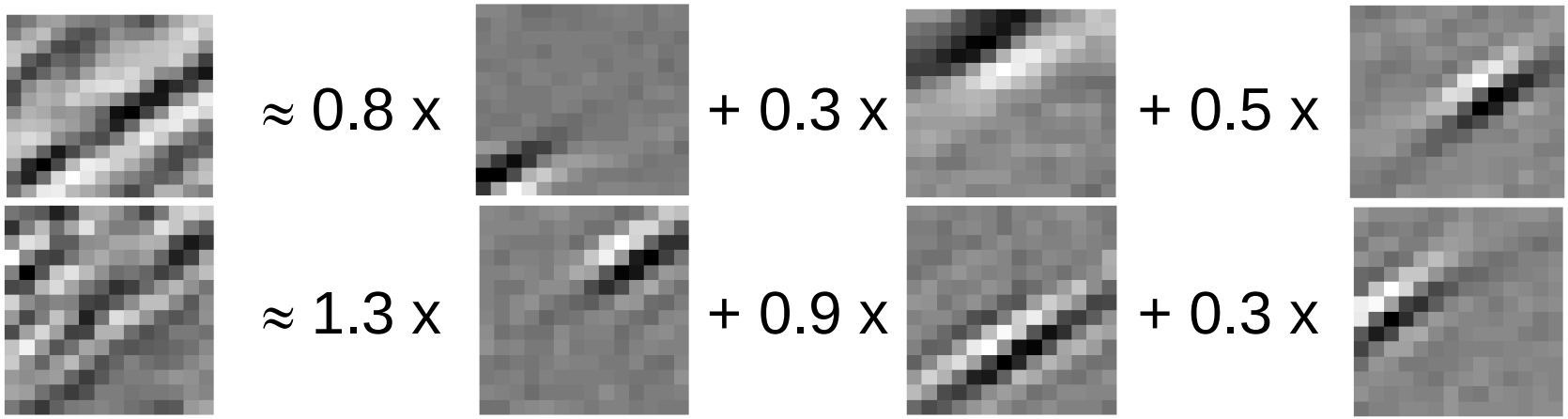
SC Versus VQ

- SC coding can achieve a much lower reconstruction error
- SC more specialize and could capture salient properties of images
- Image patches are sparse signals.
- Reduces the complexity of SVMs to O(n) in training and a constant in testing.

Improving Sparse Coding

Intuitions:

- Sparse coding is slow.
- A descriptor may be represented by very dissimilar bases.
- Similar descriptors may be represented by very different bases and codes.



Adding local constraint

- Locality is more essential than sparsity (K.Yu et al. 2009).
- Locality must lead to sparsity but not vice versa.

$$\min_{\mathbf{U},\mathbf{V}} \quad \sum_{m=1}^{M} \|\mathbf{x}_m - \mathbf{u}_m \mathbf{V}\|^2 + \lambda |\mathbf{u}_m| \qquad \text{subject to} \quad \|\mathbf{v}_k\| \leq 1, \quad \forall k = 1, 2, \ldots, K$$

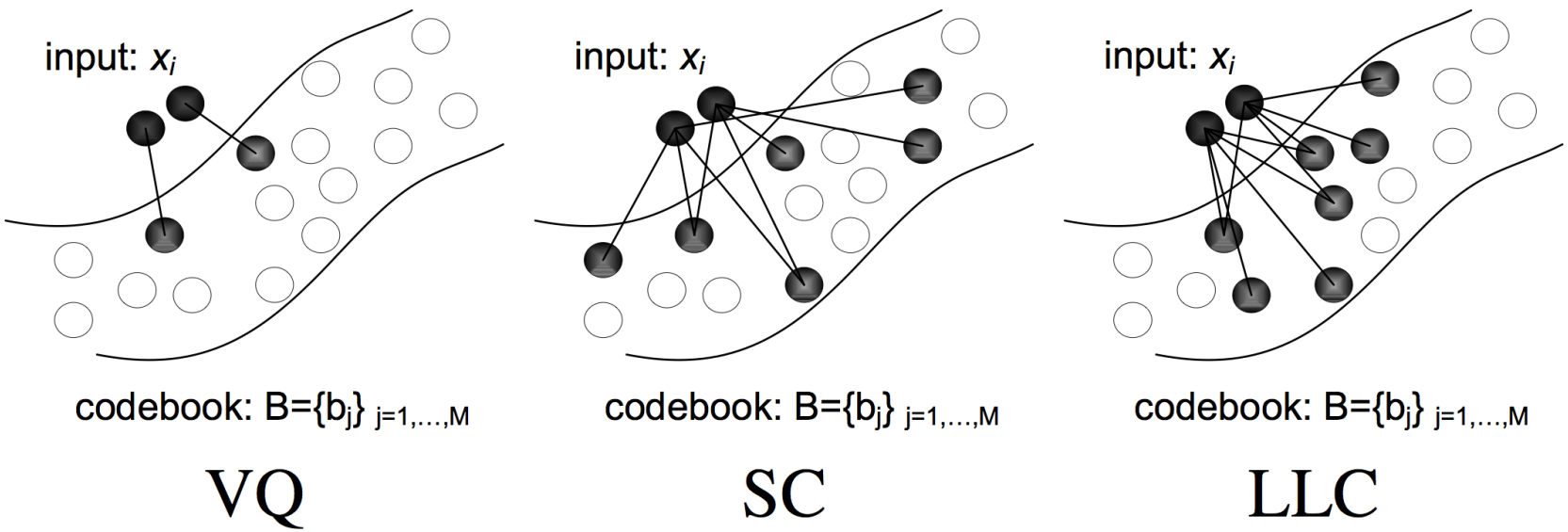
$$\min_{U} \quad \sum_{m=1}^{M} ||x_m - u_m V||^2 + \lambda ||d \odot u_m||^2 \qquad \text{subject to} \quad 1^T U = 1$$

- Where $\odot$ denotes the element-wise multiplication, and dist() is the Euclidean distance.

$$d = exp\left(\frac{dist(x_m, V)}{\sigma}\right)$$

Properties of LLC

- Better reconstruction than VQ
- Local smooth sparsity
- Faster than SC (from O(MK) to O(M+K^2), where K<<M)



Approximated LLC for fast encoding

- Use K-NN of x as the local bases Bi

$$\min_{\tilde{\mathbf{C}}} \sum_{i=1}^{N} ||\mathbf{x}_i - \tilde{\mathbf{c}}_i \mathbf{B_i}||^2$$
$$st.\ \mathbf{1}^\top \tilde{\mathbf{c}}_i = 1,\ \forall i.$$

- Computational Complexity: O(M + K^2)

Applications

- Scene Classification
- Object Recognition
- Action Recognition
- Face Recognition
- Hand Written Digit Recognition

Scene Classification

15 Scenes

The dataset contains 4485 images falling into 15 categories: office, living room, kitchen, bedroom, store, tall building, street, industrial, inside city, highway, coast, open country, mountain, forest, suburb. Each category contains 200 to 400 images, and we randomly take 100 of them as training for this category.

Scene Classification: Examples

Scene Classification: Results

15 Scenes

The dataset contains 4485 images falling into 15 categories: office, living room, kitchen, bedroom, store, tall building, street, industrial, inside city, highway, coast, open country, mountain, forest, suburb. Each category contains 200 to 400 images, and we randomly take 100 of them as training for this category.

Results

Algorithms	KSPM	KC	LLC
Accuracy	81.40 ± 0.50	76.67 ± 0.39	$\mathbf{83.1 \pm 0.60}$

Object Recognition

Caltech101

The Caltech-101 dataset contains 102 classes (including animals, vehicles, faces, flowers, etc.) with high shape variability. The number of images per category varies from 31 to 800. Most images are medium resolution, i.e., about 300 $\times$ 300 pixels.

Example

Object Recognition - Results

Classification results

Algorithms	15 training	30 training
SVM-KNN	59.10 ± 0.60	66.20 ± 0.50
KSPM	56.40	64.40 ± 0.80
NBNN	65.00 ± 1.14	70.40
ML+CORR	61.00	69.60
KC	–	64.14 ± 1.18
LLC	$\mathbf{67.0 \pm 0.45}$	$\mathbf{73.2 \pm 0.54}$

- **KSPM**: nonlinear SPM with histograms.
- **KC**: nonlinear SPM with soft assignment for histograms.
- **LLC:** linear SVM.

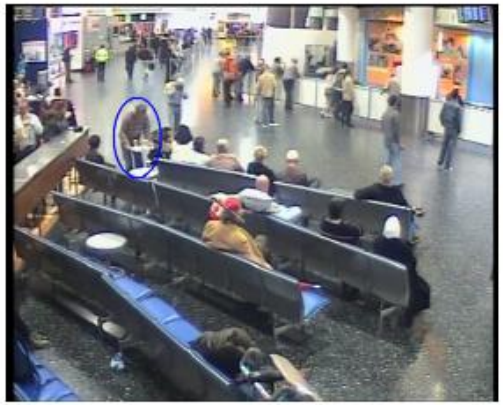
Action Recognition

TRECVID Surveillance Event Detection Evaluation

- 100 hours of surveillance videos, 10 hours each day, from London Gatwick International Airport.
- 50 hours of annotated videos for training, and the rest 50 hours of videos for testing.
- We evaluated on 3 event detection tasks: cellphone to ear, object put, and pointing.
- Very large scale, only linear classifiers applicable.

Action Recognition - examples

Pointing



ObjectPut

CelltoEar

ObjectPut

Action Recognition – Results

TRECVID Surveillance Event Detection Evaluation

- 100 hours of surveillance videos, 10 hours each day, from London Gatwick International Airport.
- 50 hours of annotated videos for training, and the rest 50 hours of videos for testing.
- We evaluated on 3 event detection tasks: cellphone to ear, object put, and pointing.
- Very large scale, only linear classifiers applicable.

Results (AUC score)

Algorithms	CellToEar	ObjectPut	Pointing
LSPM	0.688	0.714	0.744
LLC	**0.744**	**0.773**	**0.769**

Face Recognition

CMU PIE Database

The dataset contains 41368 images of 68 people, each person with under 13 poses, 43 different illumination conditions with 4 different expressions. We use five near frontal poses (C05, C07, C09, C27, C29) for our evaluation.

Face Recognition

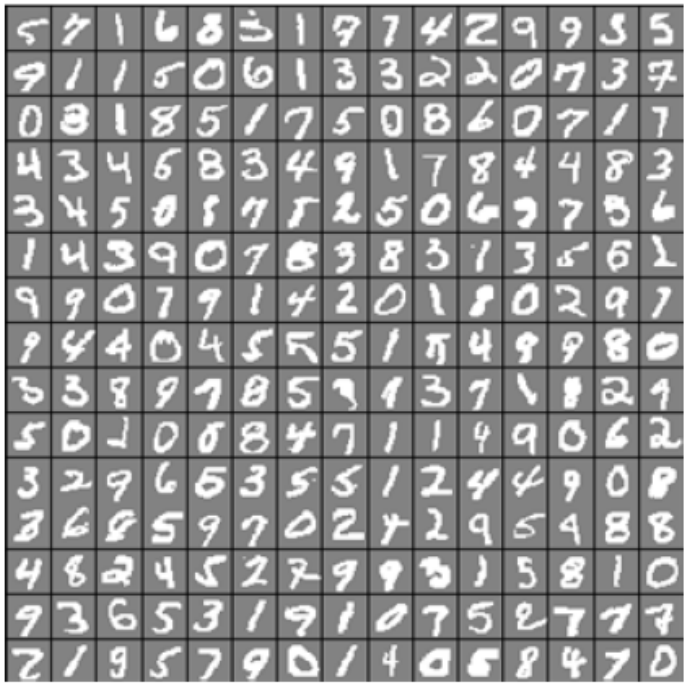
Classification error (%) on CMU PIE.

Training	30	50	70	90	130
LDA	7.9	4.8	4.0	3.4	2.9
R-LDA	5.0	3.9	3.5	3.2	3.0
S-LDA	3.6	2.5	2.1	1.8	1.6
LLC	**0.81**	**0.26**	**0.22**	**0.11**	**0.037**

Hand Written Digit Recognition

MNIST: The dataset consists of 70,000 handwritten digits, of which 60,000 are selected for training and the rest 10,000 for testing.

Algorithms	Error Rate
SVM (RBF)	1.41
L1 sparse coding	2.02
Local coordinate coding	1.90
Deep Belief Network	1.20
CNN	**0.82**
LLC	0.98



Q&A