Locality-constrained Linear Coding for Image Classification

Jinjun Wang, Jianchao Yang, Fengjun Lv, Thomas Huang, Yihong Gong

Ken Chatfield

UNIVERSITY OF OXFORD

Introduction

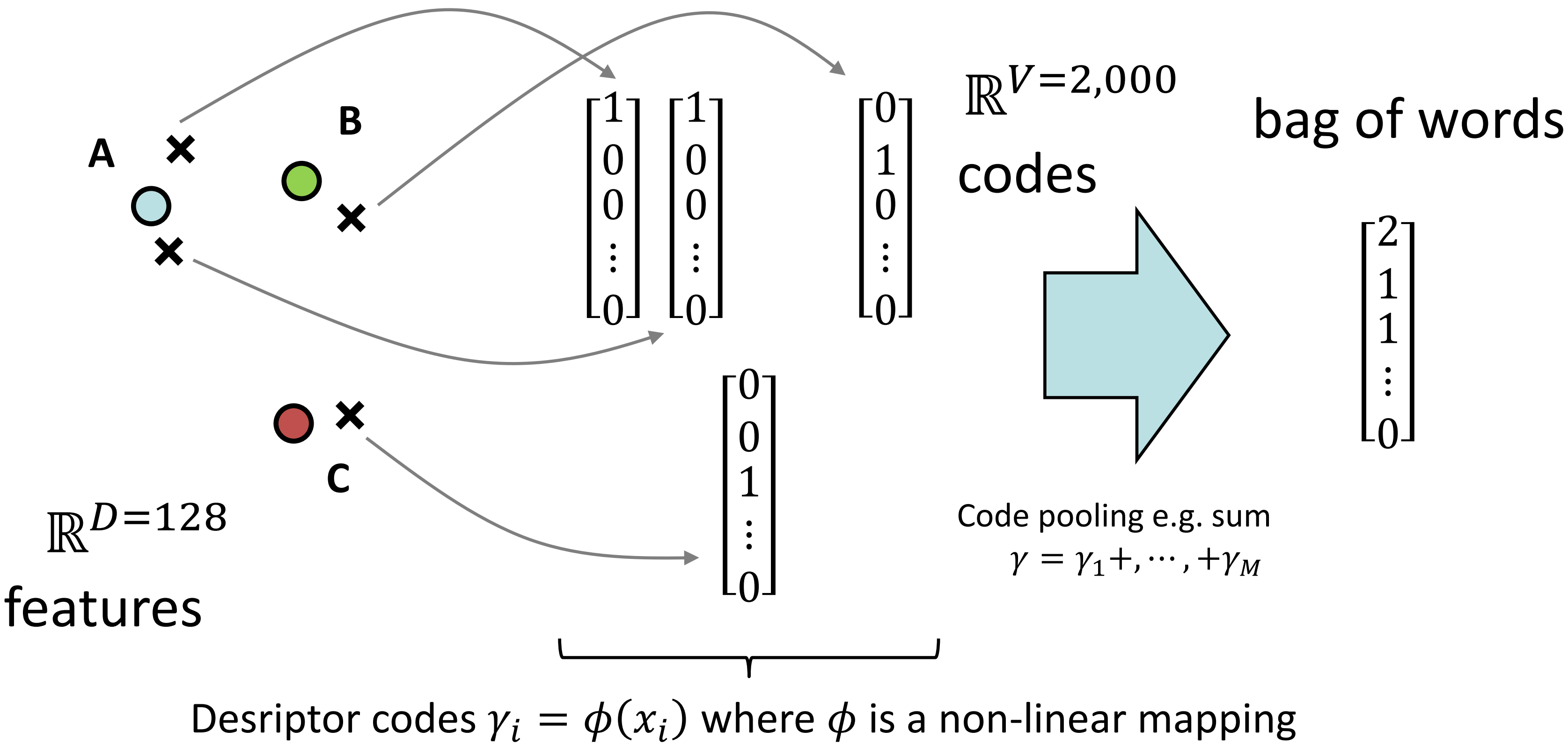
- How do we classify visual object categories?
- Bag of visual words approach highly successful – at the core of winning entries for PASCAL VOC 2007-2010

monkey?

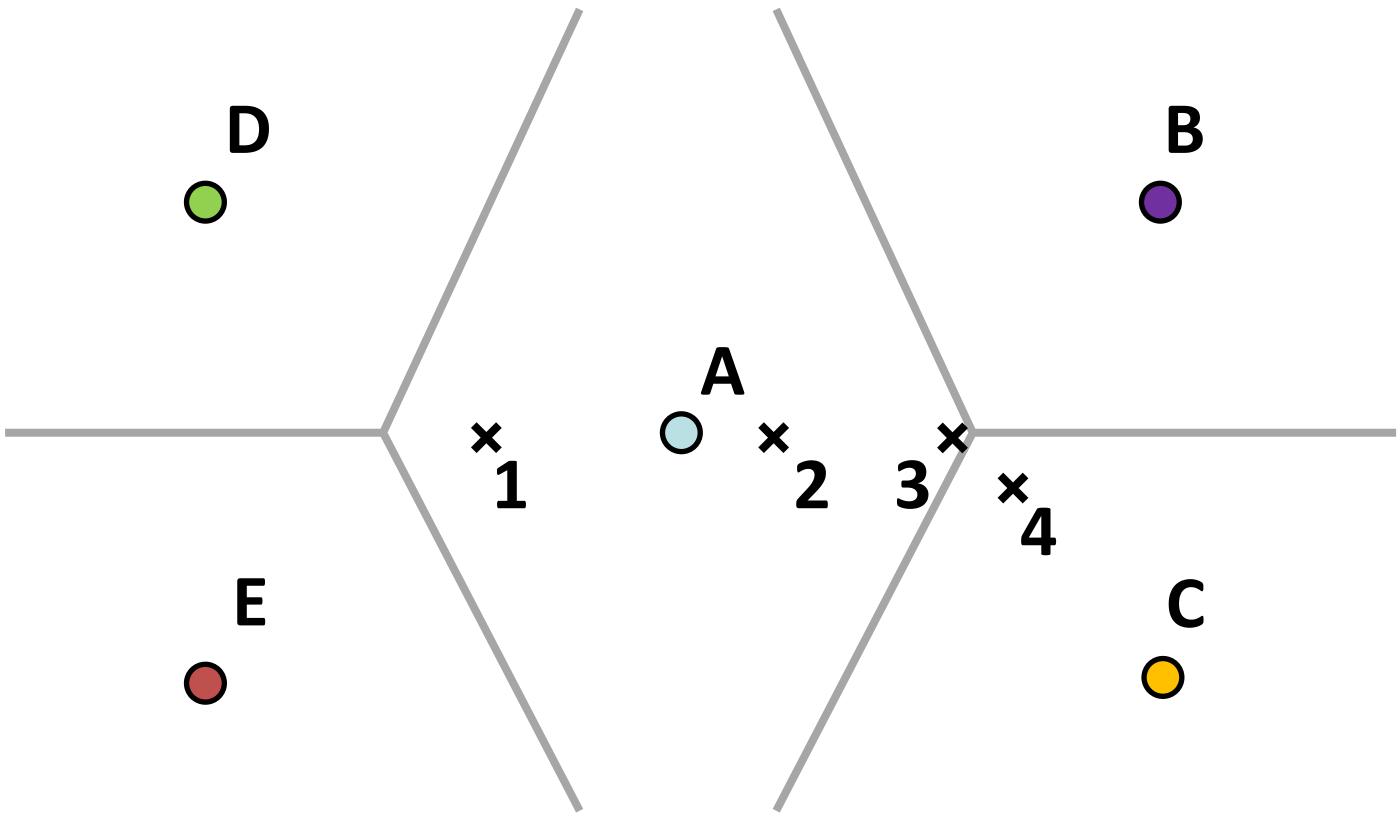
monkey?

Bag of Visual Words as Descriptor Coding

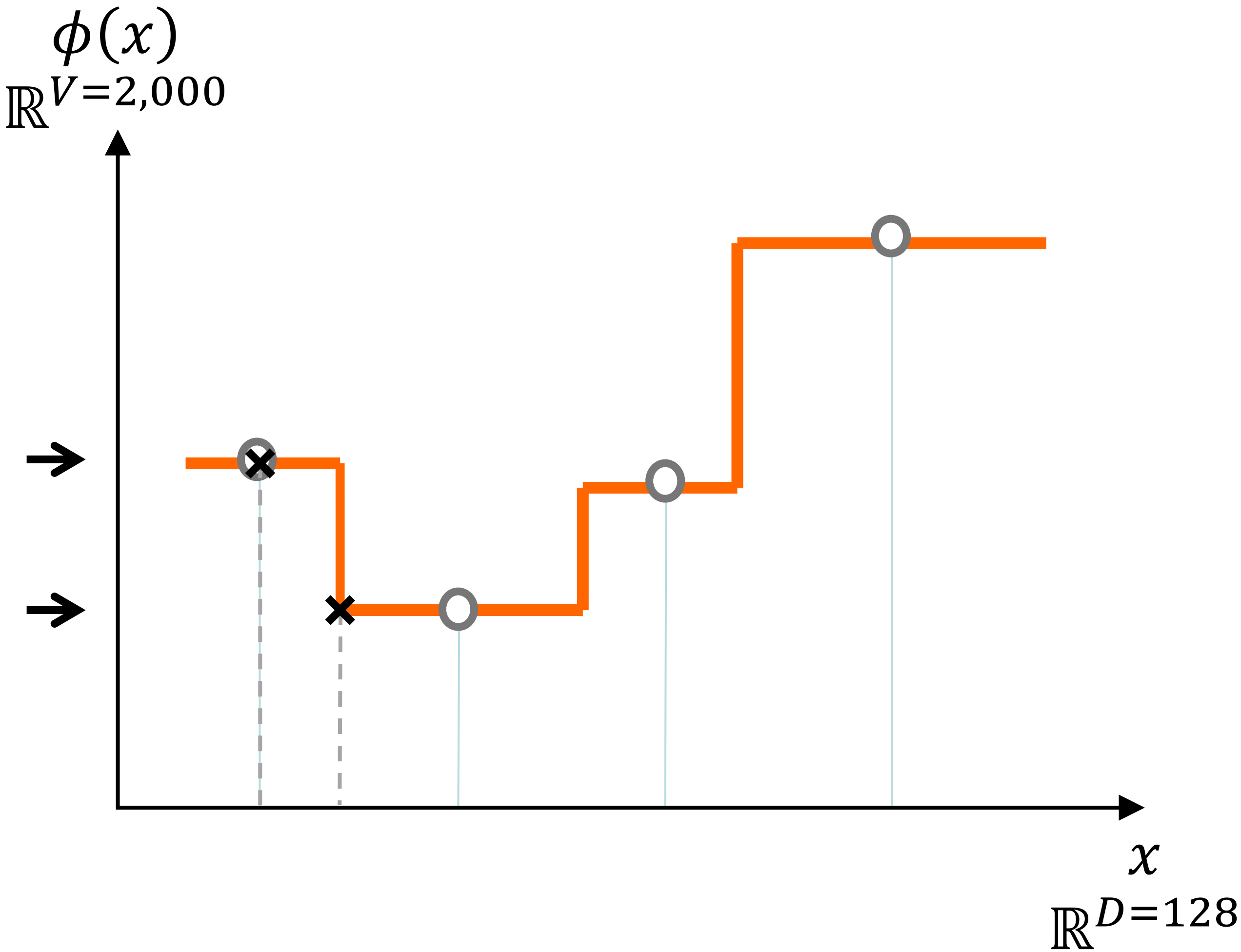
- 'Bag of Visual Words' using vector quantization for visual word assignment can be considered to be a type of **feature coding**
- In VQ each feat. in an image is encoded by assigning to a single visual word
- These codes are **sparse** and **high dimensional**
- Codes are pooled to form a single sparse 'bag of words' to describe the image



The Problem with Vector Quantization

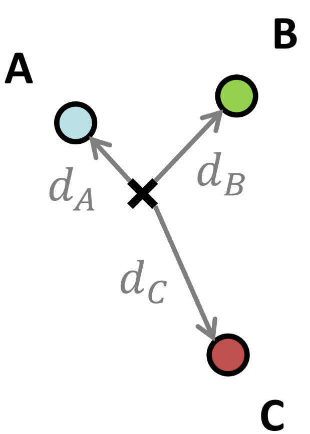


The Problem with Vector Quantization



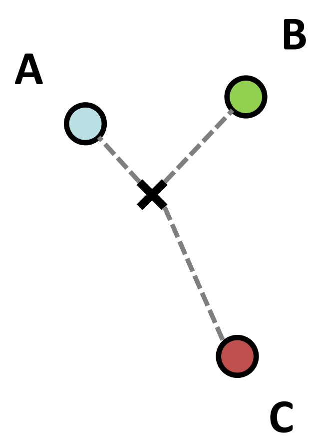
Approaches to Soft-Assignment

- Distance-based soft assignment
- Soft assignment through learning an optimal reconstruction
  - With sparsity regularization → ScSPM (CVPR '09)
  - With locality regularization → LCC (NIPS '09) / LLC (CVPR '10)



$$x \approx \sum_{j=1}^{V} K_\sigma(\|x, v_j\|) \cdot v_j$$

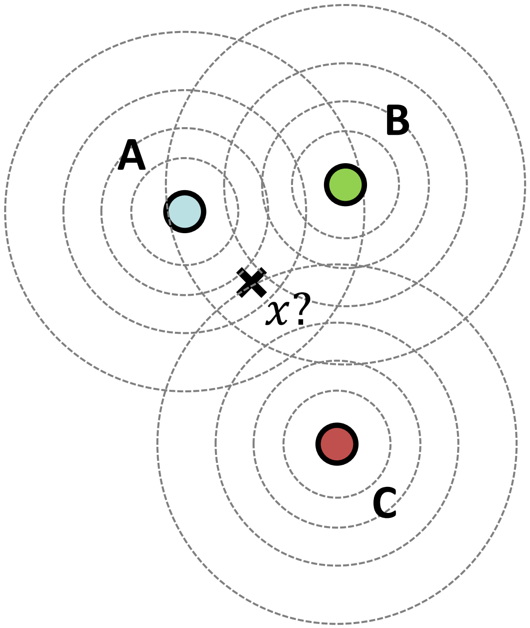
Distance-based



$$x \approx \sum_{j=1}^{V} \gamma_j v_j$$

Reconstruction

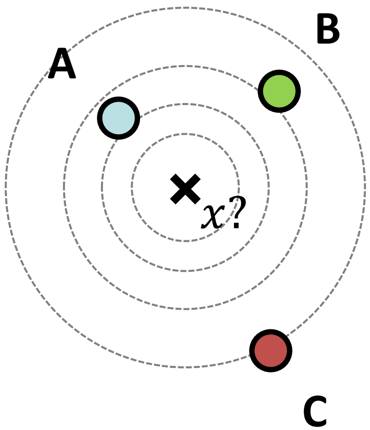
Distance-based Soft Assignment



$$K_\sigma(x - X_i) = K_\sigma(X_i - x)$$

$$K_\sigma(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{x^2}{2\sigma^2})$$

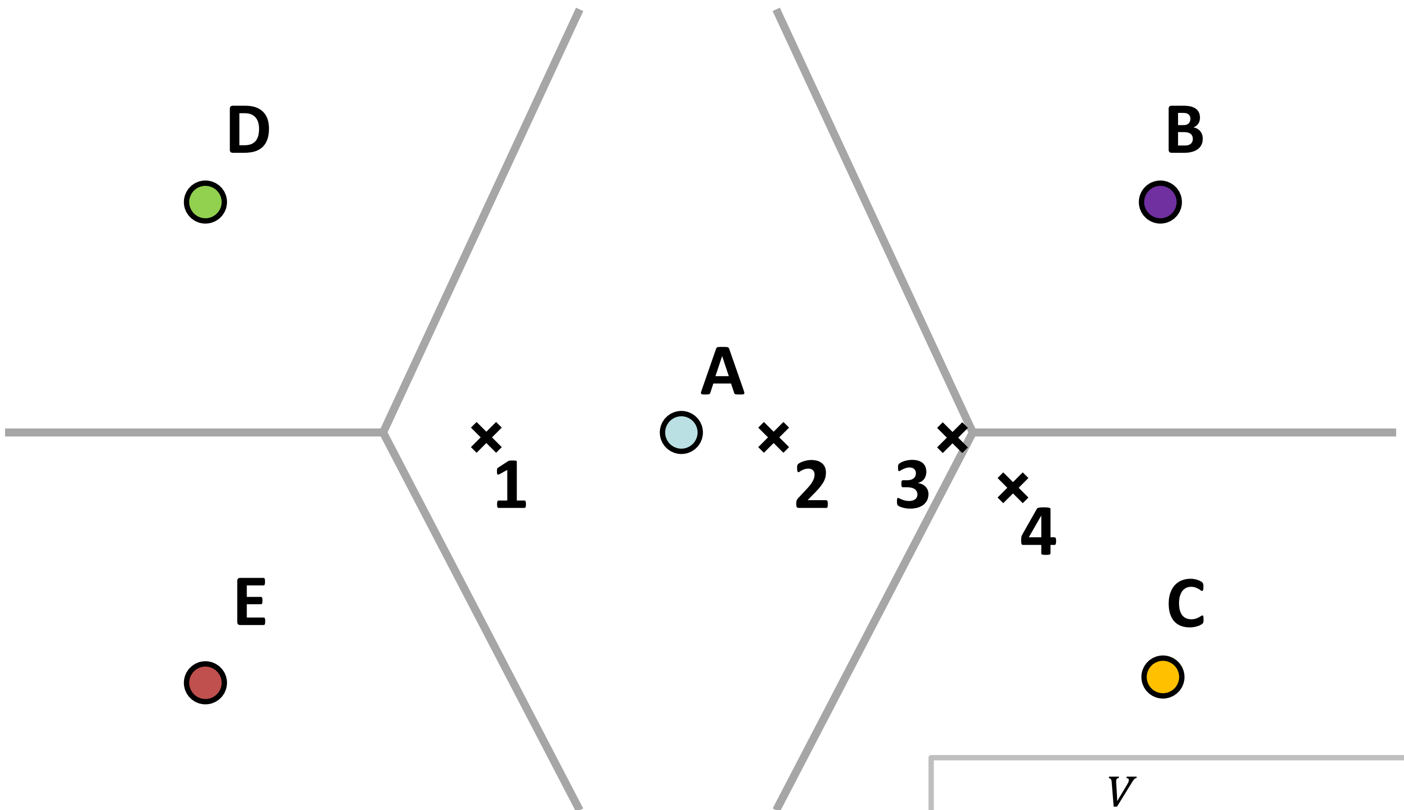
$$x \approx \sum_{j=1}^{V} K_\sigma(\|x, v_j\|) \cdot v_j$$



- Replace **histogram estimator** of the codewords with a **gaussian mixture model**
- However, if the kernel is symmetric, can place kernel on codeword instead
- Choose N nearest neighbour codewords and assign weighted by kernel
- Essentially assigning based on **distances** in feature space $\mathbb{R}^{D=128}$

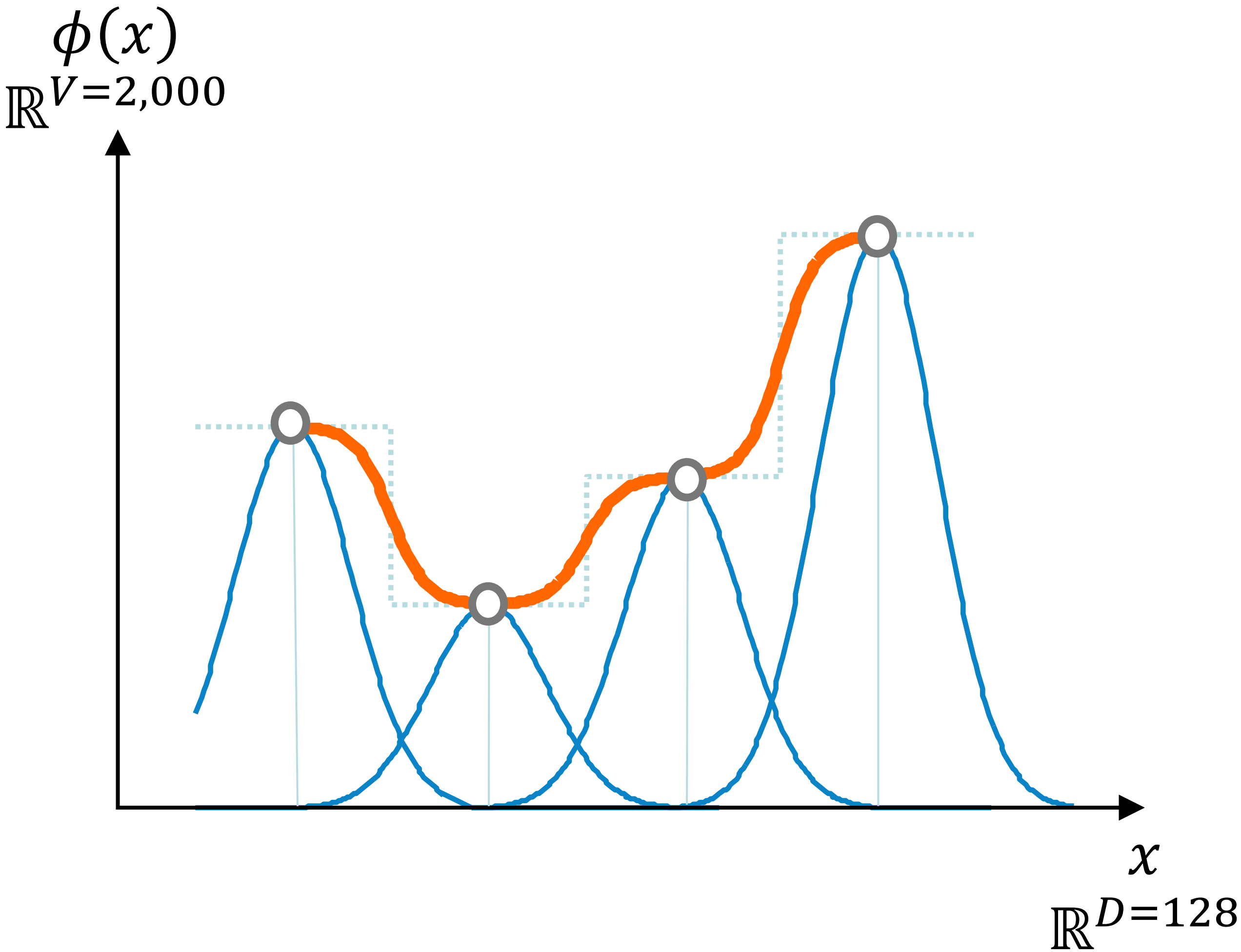
Philbin et al. CVPR 2008
Gemert et al. ECCV 2008

Distance-based Soft Assignment



$$\gamma \approx \sum_{j=1}^{V} K_\sigma\left(\left\|x, v_j\right\|\right)$$

Distance-based Soft Assignment



Encoding using Sparsity Reg. (ScSPM)

- Over all features x_i for $i = 1 \ldots N$ Vector Quantization becomes a constrained least square fitting problem:

$$\arg\min_{\gamma} \sum_{i=1}^{N} \|x_i - \underbrace{\mathrm{N}}_{dxM\text{ matrix codebook}} \overbrace{\gamma_i}^{\text{Encoding for image } i}\|^2$$

s.t. only one element of γ_i is non-zero and equal to 1

(i.e. $\|\gamma_i\|_{\ell^0} = 1$, $\|\gamma_i\|_{\ell^1} = 1$) this non-zero element corresponds to v_j

- But why should the feature be assigned to only one codebook entry?
- Ameliorate the quantization loss of VQ by removing the constraint $\|\gamma_i\|_{\ell^0} = 1$ and instead using a sparsity regularization term to restrict the number of non-zero bases:

$$\arg\min_{\gamma} \sum_{i=1}^{N} \|x_i - \mathrm{N}\gamma_i\|^2 + \lambda\|\gamma_i\|_{\ell^1}$$

Encoding using Sparsity Reg. (ScSPM)

$$\arg\min_{\gamma} \sum_{i=1}^{N} \|x_i - \mathrm{N}\gamma_i\|^2 + \lambda \|\gamma_i\|_{\ell^1}$$

- This is the sparse coding scheme ScSPM (Yang et al. CVPR '09)
- ℓ^1 regularization required as codebook N is usually overcomplete (i.e. $M > d$)
- By assigning to multiple bases we overcome the quantization errors introduced by VQ
- Over Caltech-101 using dense SIFT yields 10% improvement over VQ, and 5~6% improvement over soft-assignment using kernel codebooks using a **linear SVM** (see results later)

Coding Provides Non-linearity

Considering general case and a typical classification framework:

↓ feature extraction

Features

↓ non-linear coding

Codes

↓ linear pooling

Bag of Words Vector

↓ linear SVM

Classification

$$\mathbf{X} = [x_1, x_2, \cdots, x_N] \in \mathbb{R}^{D=128}$$

where D is # feature dimensions e.g. SIFT = 128
and N is the number of features (DxN matrix)

$$\phi\,(\mathbf{X}) = [\gamma_1, \gamma_2, \cdots, \gamma_N] \in \mathbb{R}^{V}$$

where V is the codebook size (MxN matrix)

$$\gamma = \sum_{i=1}^{N} \gamma_i$$

$$f_c(\gamma) = w^{\mathrm{T}}\gamma$$

$$f_c(\gamma) = w^{\mathrm{T}}\gamma = \sum_{i=1}^{N} w^{\mathrm{T}}\gamma_i = \sum_{i=1}^{N} \overbrace{w^{\mathrm{T}}}^{\text{linear classifier}} \underbrace{\phi(x_i)}_{\text{non-linear coding}}$$

Encoding using Distance Reg. (LCC/LLC)

- Using ScSPM soft-assignment is formulated as a least squares fitting problem using an ℓ^1 sparsity regularization
- However, the effectiveness of distance-based soft-assignment suggests that the locality of the visual words used to describe any feature is also important
- We can account for this by replacing the sparsity regularization with a **locality constraint**:

$$\arg\min_{\gamma} \sum_{i=1}^{N} \|x_i - \mathrm{N}\gamma_i\|^2 + \lambda \|d_i \odot \gamma_i\|^2$$

$$d_i = \exp\left(\frac{\mathrm{dist}(x_i, \mathrm{N})}{\sigma}\right)$$

- This is not sparse in sense of ℓ^1 norm, but in practice has few significant values – those values below a certain threshold can be set to zero

Approximated LLC for Fast Encoding

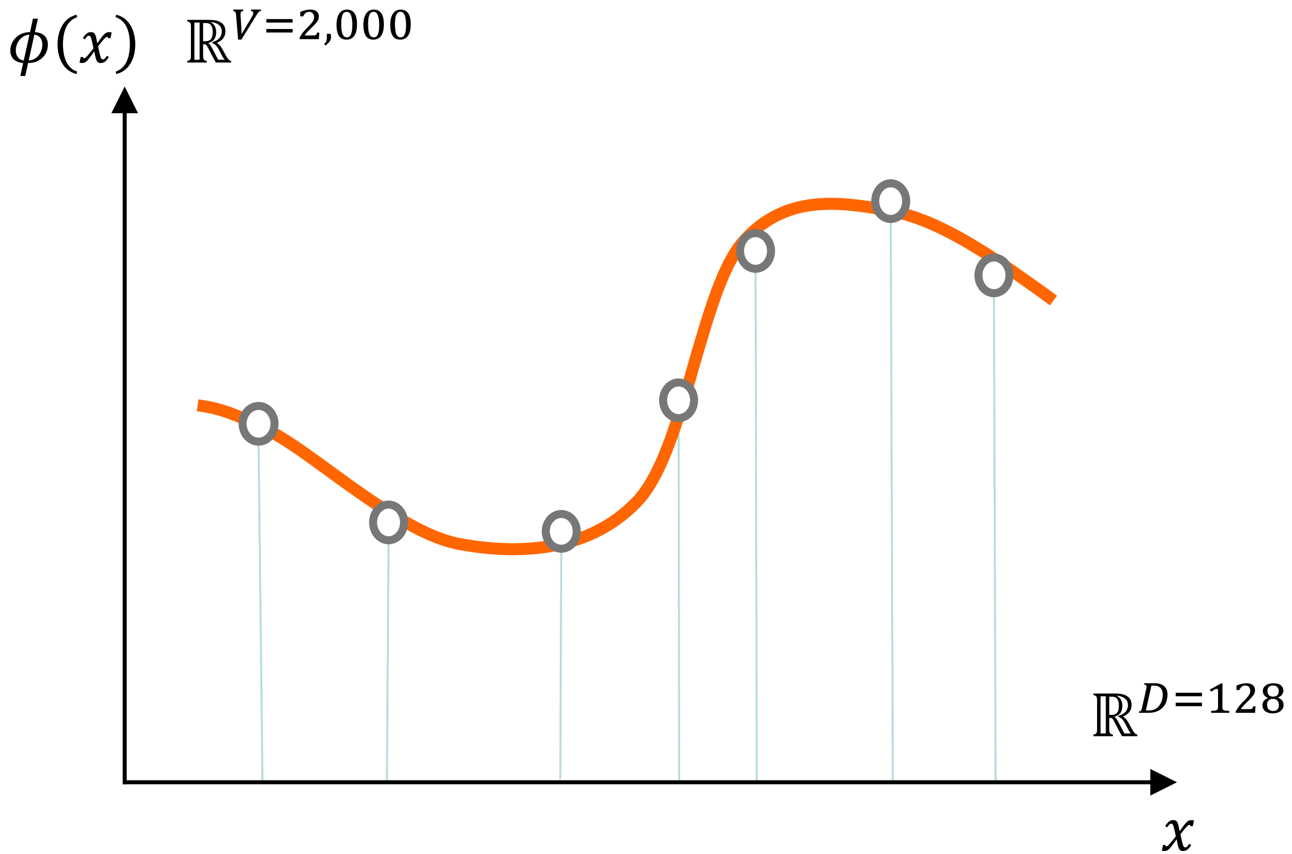
$$\arg\min_{\gamma} \sum_{i=1}^{N} \|x_i - \mathrm{N}\gamma_i\|^2 + \lambda \|d_i \odot \gamma_i\|^2$$

- The distance regularization of LLC effectively performs **feature selection**, and in practice only those bases close to x_i in feature space have non-zero coefficients
- This suggests we can develop a fast approximation of LLC by removing the regularization completely and instead using the **K nearest neighbours** of x_i ($K < D < V$ and in the paper $K = 5$) as a set of local bases N_i:

$$\arg\min_{\tilde{\gamma}} \sum_{i=1}^{N} \|x_i - \mathrm{N}_i\tilde{\gamma}_i\|^2 \qquad st. \|\tilde{\gamma}_i\|_{\ell^1} = 1, \forall i$$

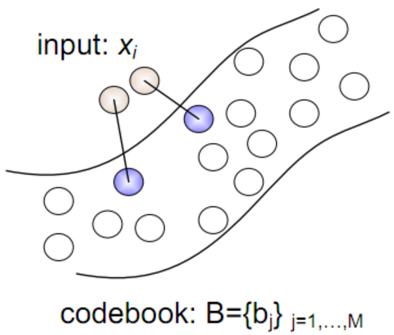
- This reduces the computation complexity from $\mathcal{O}(V^2)$ to $\mathcal{O}(V + K^2)$ and the nearest neighbours can be found using ANN methods such as kd-trees

Locally-constrained Linear Coding



- A smooth function is fitted between visual words and assignment is optimized to **minimize reconstruction error** unlike purely distance-based assignment
- For LLC only the K nearest neighbours (=5) are used → equivalent of **V-dimensional spline interpolation across intervals of K**

Soft Assignment Methods Comparison

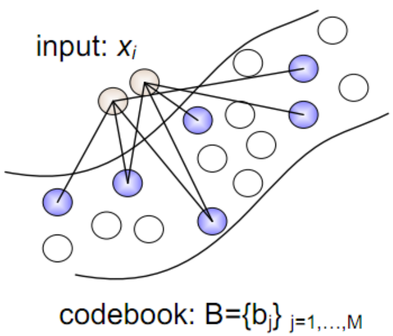


Vector Quantization

- ✓ Fast
- ✗ Quantization a problem

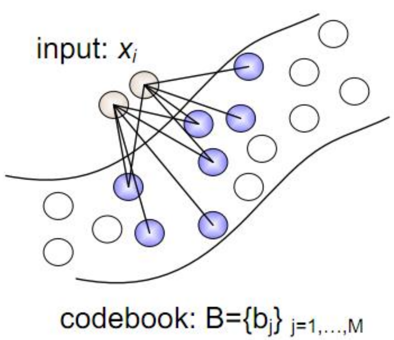
Distance-based Soft-Assignment

- ✓ Assigns features to multiple visual words based on locality
- ✗ Does not minimize reconstruction error



ScSPM (sparsity regularization)

- ✓ Minimizes reconstruction error $\sum_{i=1}^{N} \|x_i - \mathrm{N}\gamma_i\|^2$
- ✗ Optimization is computationally expensive
- ✗ Regularization term is not smooth



LLC (locality regularization)

- ✓ Minimizes reconstruction error $\sum_{i=1}^{N} \|x_i - \mathrm{N}\gamma_i\|^2$
- ✓ Local smooth sparsity
- ✓ Fast computation through approximated LLC

Results

Algorithm	15 training	30 training
SVM-KNN (Zhang CVPR '06)	59.10	66.20
KSPM (Lazebnik CVPR '06)	56.40	64.40
NBNN (Boiman CVPR '08)	65.00	70.40
ML+CORR (Jain CVPR '08)	61.00	69.60
Hard Assignment	--	62.00
KC (Gemert ECCV '08)	--	64.14
ScSPM (Yang CVPR '09)	**67.00**	73.20
LLC	65.43	**73.44**

↑ Results over Caltech-101 dataset

↓ Results over Caltech-256

Algorithm	15 training	30 training
Hard Assignment	--	25.54
KC (Gemert ECCV '08)	--	27.17
ScSPM (Yang CVPR '09)	27.73	34.02
LLC	**34.36**	**41.19**