Bag of Visual Features

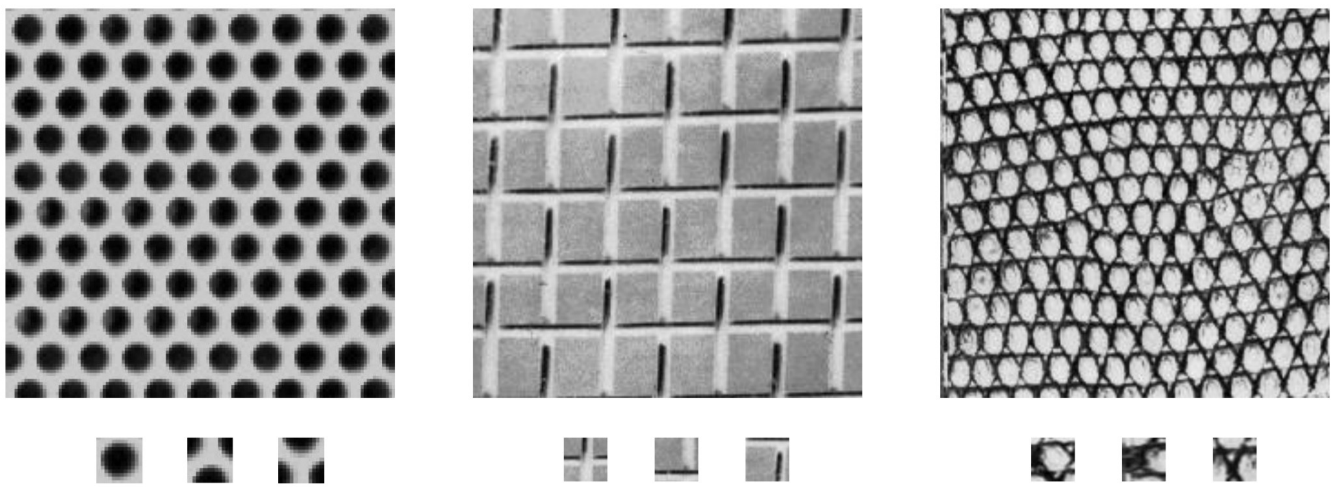
Bag of Features

Overview

- Bag of features framework
- Examples of feature encodings

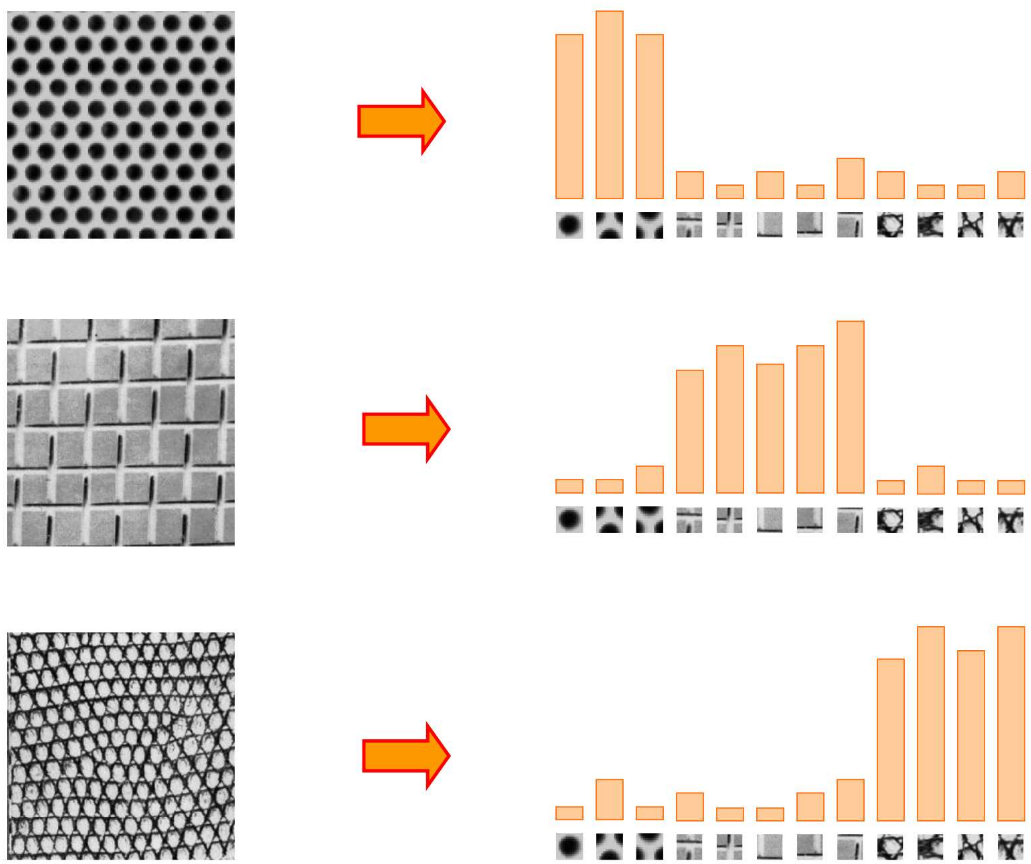
Example: Texture Recognition

- Texture is characterized by the repetition of basic elements or textons
- For stochastic textures, it is the identity of the textons, not their spatial arrangement, that matters



Julesz, 1981; Cula & Dana, 2001; Leung & Malik 2001; Mori, Belongie & Malik, 2001; Schmid 2001; Varma & Zisserman, 2002, 2003; Lazebnik, Schmid & Ponce, 2003

Example: Texture Recognition



General Steps for Bag of Features

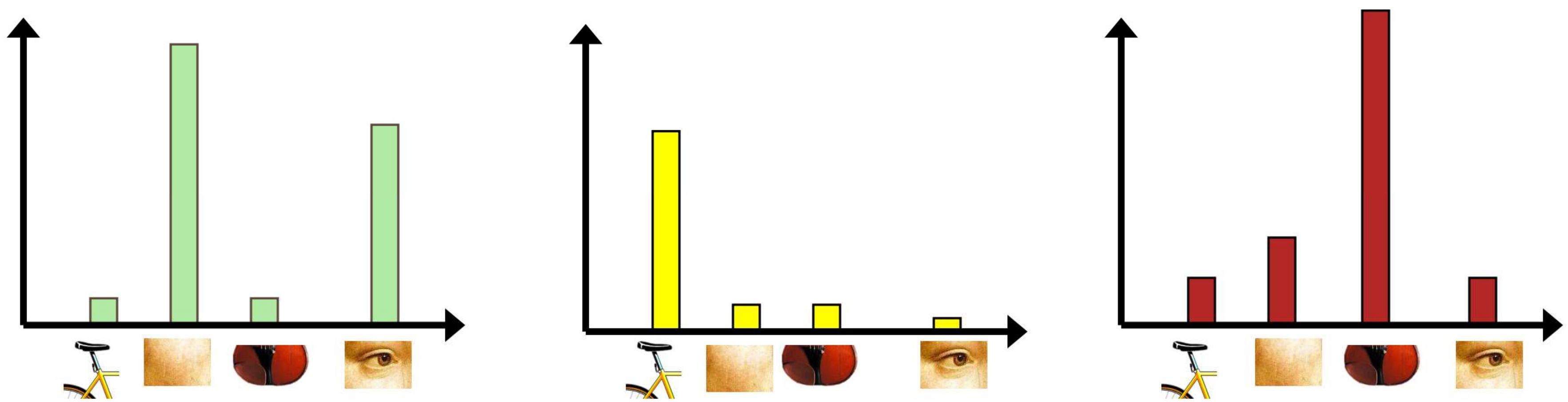
- 1. Extract features

General Steps for Bag of Features

- 2. Learn “visual vocabulary”

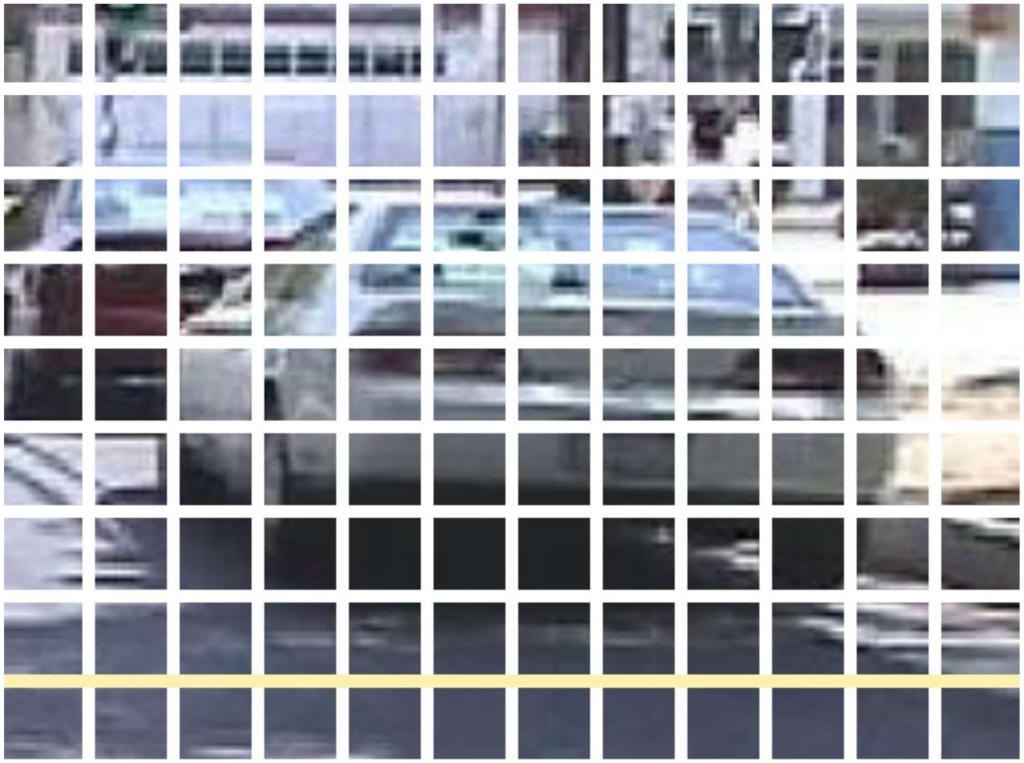
General Steps for Bag of Features

- 3. Quantize features using visual vocabulary
- 4. Represent images by frequencies of “visual words”



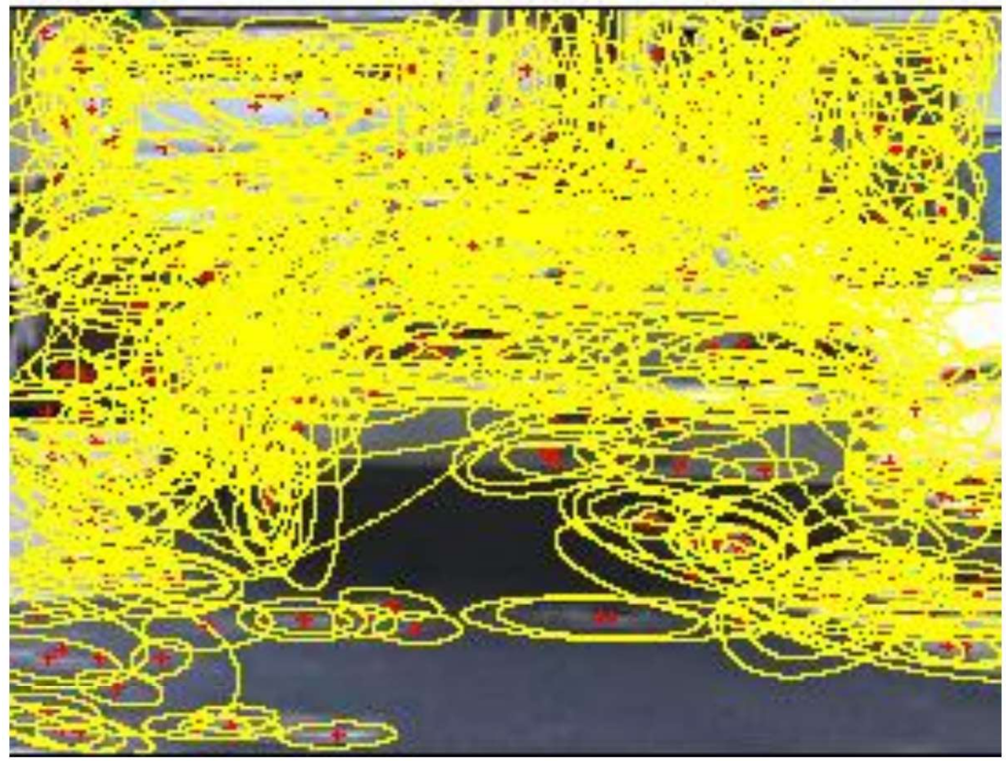
Extract Features

- Regular grid
  - Vogel & Schiele, 2003
  - Fei-Fei & Perona, 2005



Extract Features

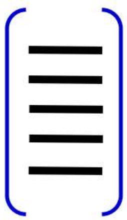
- Interest point detector
  - Csurka et al. 2004
  - Fei-Fei & Perona, 2005
  - Sivic et al. 2005



Extract Features

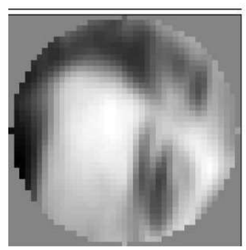
- Regular grid
  - Vogel & Schiele, 2003
  - Fei-Fei & Perona, 2005
- Interest point detector
  - Csurka et al. 2004
  - Fei-Fei & Perona, 2005
  - Sivic et al. 2005
- Other methods
  - Random sampling (Vidal-Naquet & Ullman, 2002)
  - Segmentation-based patches (Barnard et al. 2003)

Examples

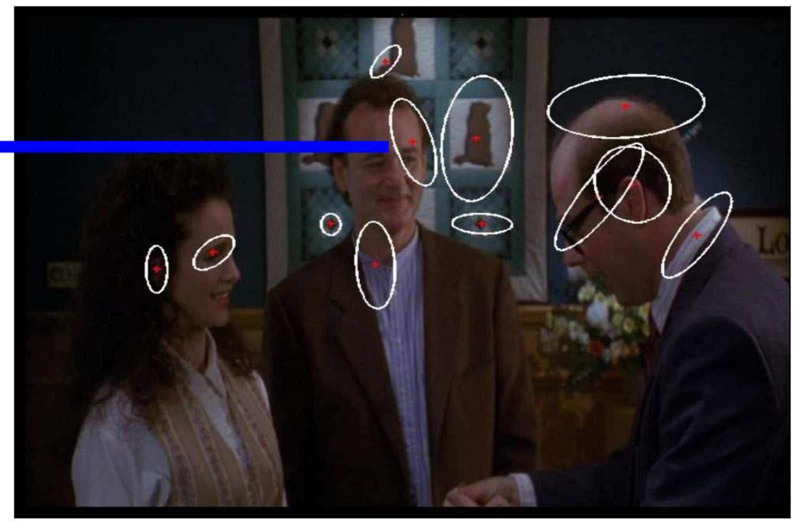


Compute SIFT descriptor

[Lowe'99]



Normalize patch



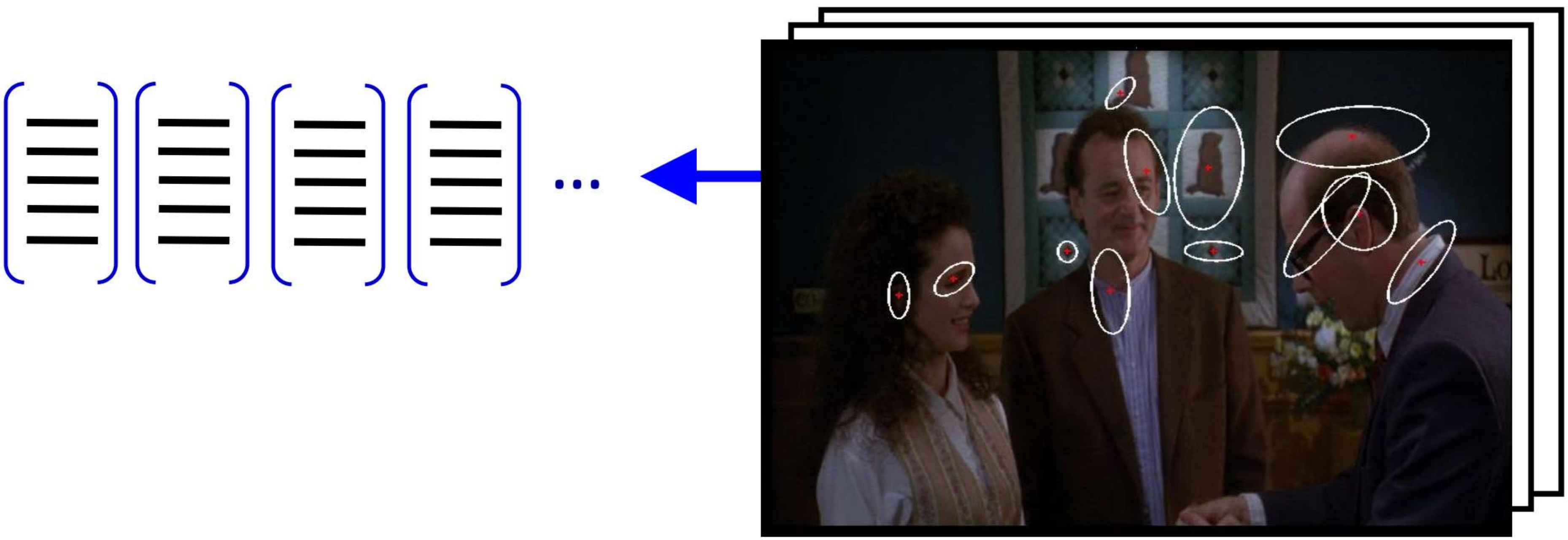
Detect patches

[Mikojaczyk and Schmid '02]

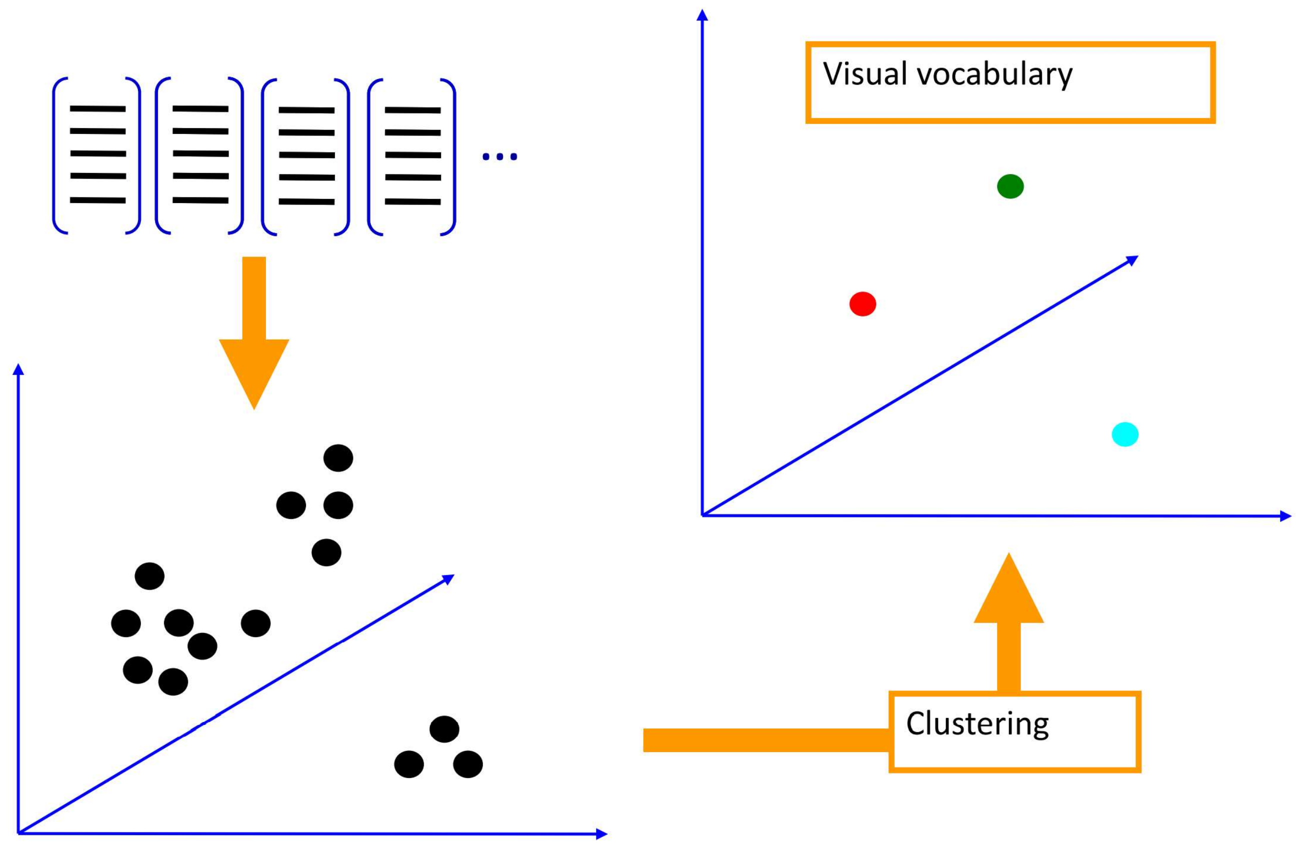
[Mata, Chum, Urban & Pajdla, '02]

[Sivic & Zisserman, '03]

Examples



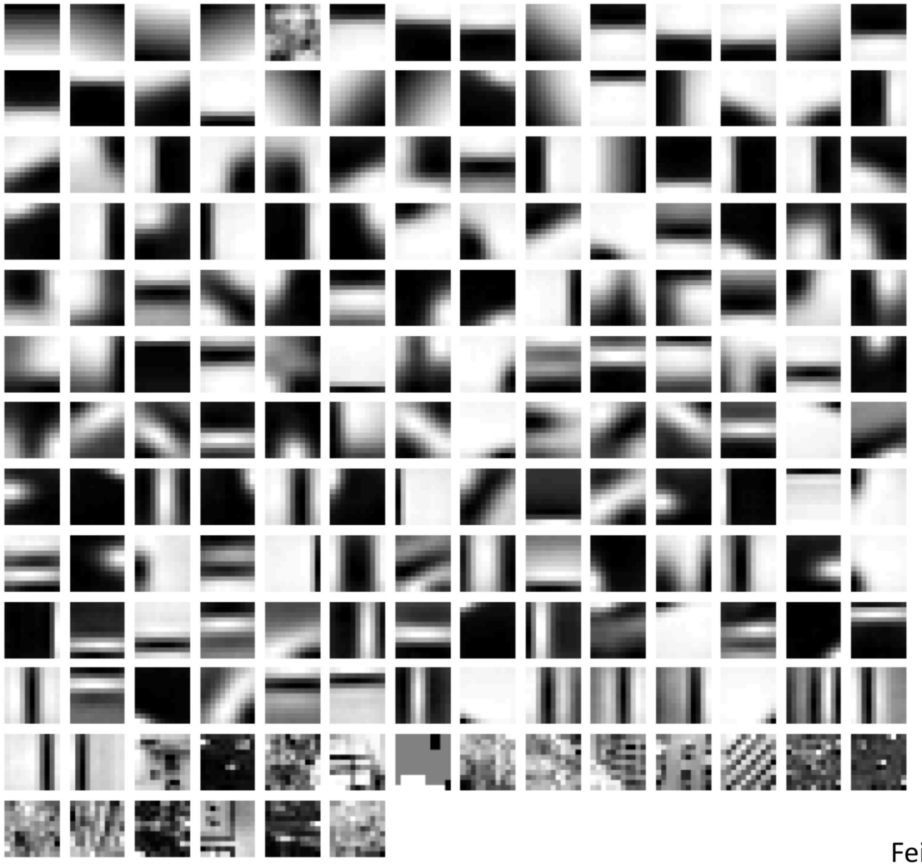
Learn the Visual Vocabulary



From Clustering to Vector Quantization

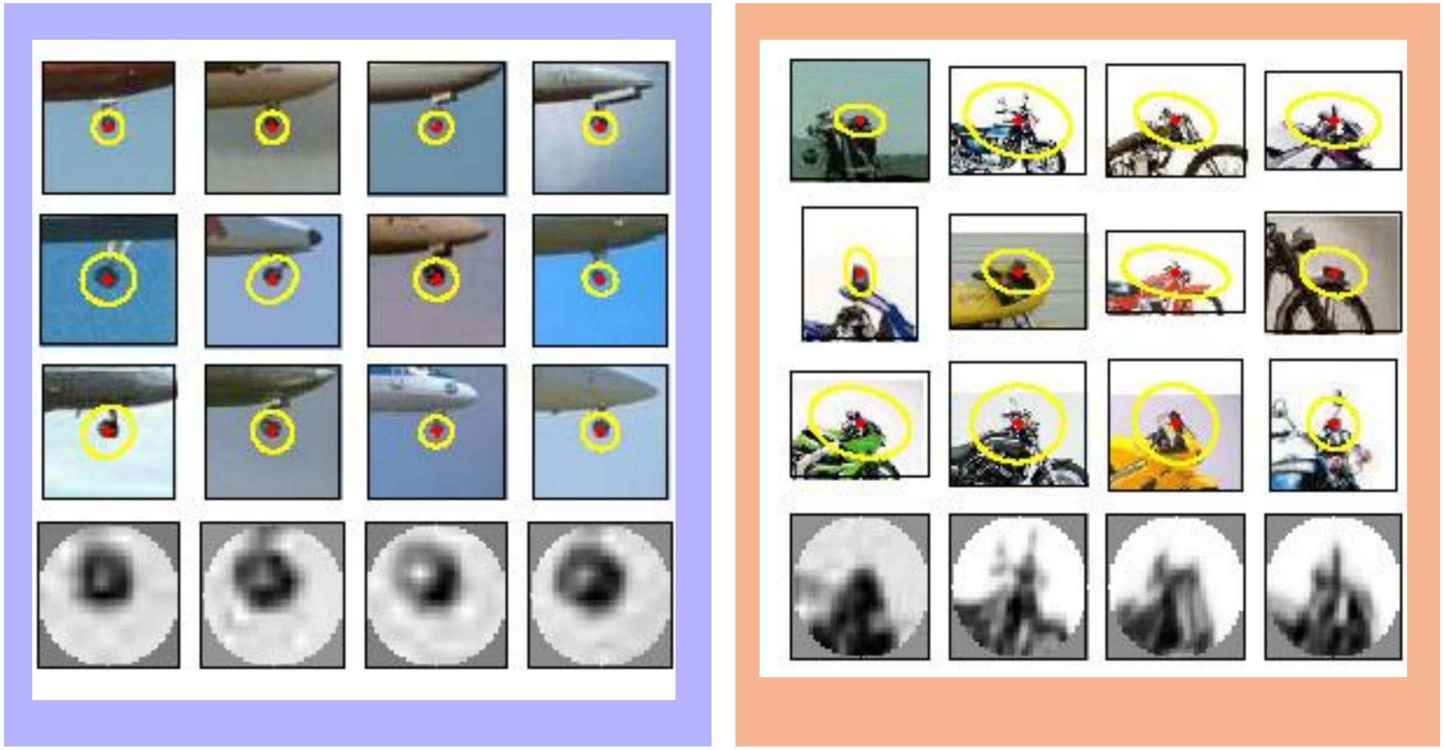
- Clustering is a common method for learning a visual vocabulary or codebook
  - Unsupervised learning process
  - Each cluster center produced by k-means becomes a codevector
  - Codebook can be learned on separate training set
  - Provided the training set is sufficiently representative, the codebook will be "universal"
- The codebook is used for quantizing features
  - A vector quantizer takes a feature vector and maps it to the index of the nearest codevector in a codebook
  - Codebook = visual vocabulary
  - Codevector = visual word

Examples: Visual Vocabulary



Fei-Fei et al. 2005

Image Patch Examples



Sivic et al. 2005

Image Representation

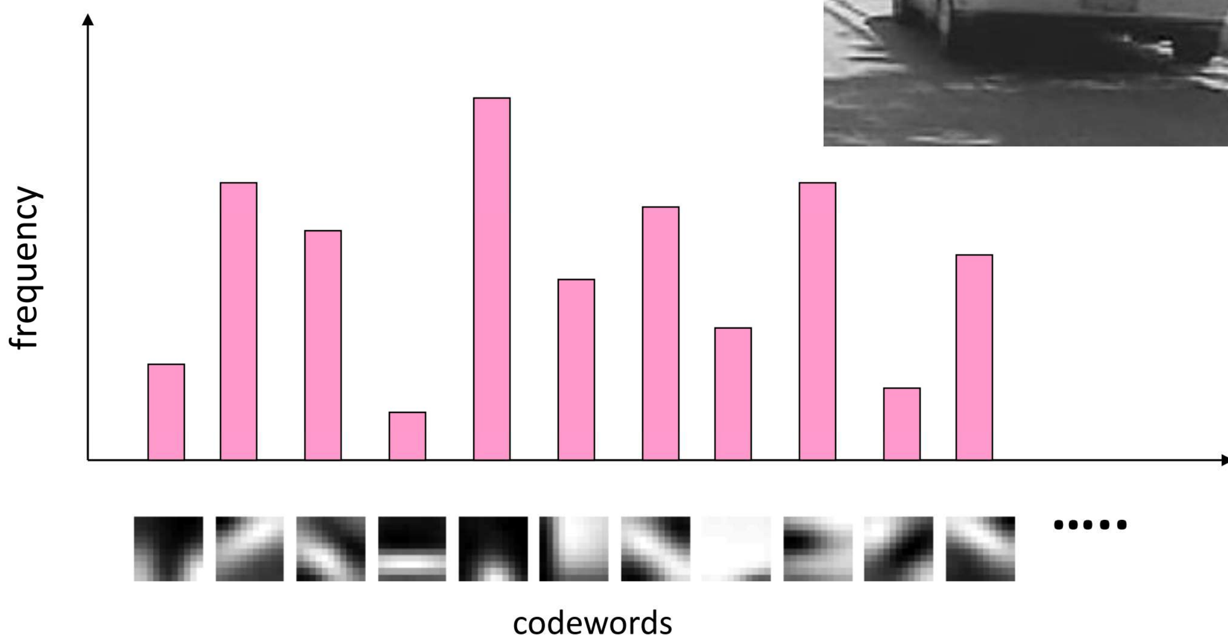


Image Classification

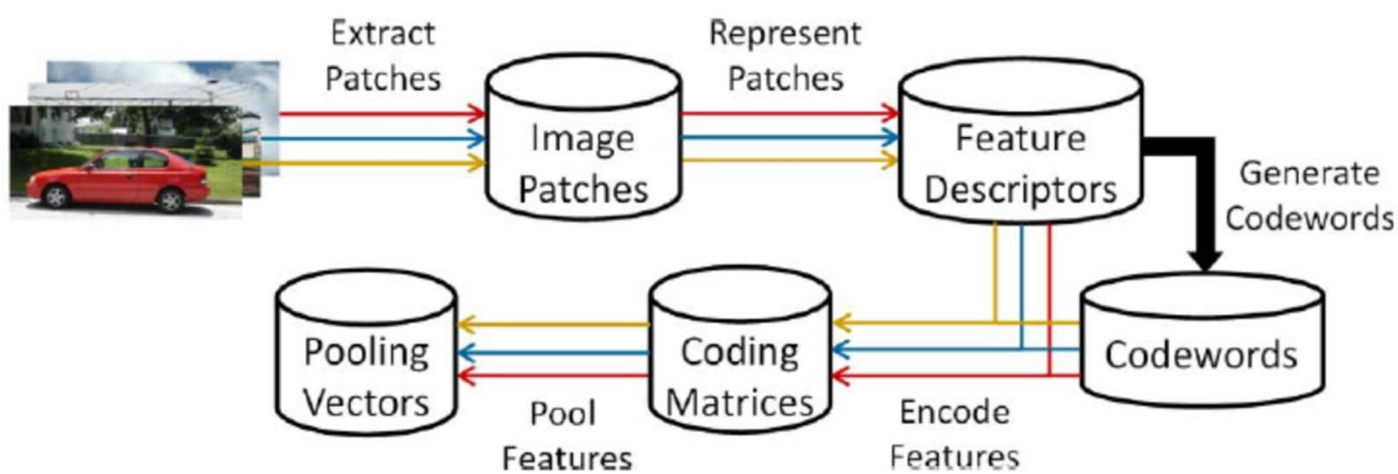


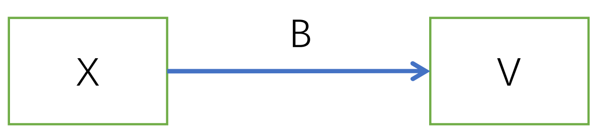
Fig. 1. The general pipeline of the BoF framework for image classification.

Feature Coding

- Voting-Based Coding
- Reconstruction-Based Coding
- Saliency-Based Coding

Notations

- Let $X = [x_1, x_2, ..., x_N] \in R^{D \times N}$ be N D-dimensional features extracted from an image.
- Let $B = [b_1, b_2,..., b_M] \in R^{D \times M}$ be a codebook with M codewords (typically obtained by clustering over features).
- Let $V = [v_1, v_2,...,v_N] \in R^{M \times N}$ be the corresponding representation of these N features.



Voting-Based Coding

- Hard voting

$$v(i) = \begin{cases} 1, & if\ i = \arg\min_j(\|x - b_j\|_2) \\ 0, & otherwise \end{cases}, i = 1, 2, \ldots, M.$$

- For example

$$x = \begin{pmatrix} 0.5 \\ 0 \\ 0 \end{pmatrix}, B = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, v = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$$

Voting-Based Coding

- Soft voting

$$v(i) = \frac{\exp(-\| x - b_i \|_2^2 / \sigma)}{\sum_{k=1}^{K} \exp(-\| x - b_k \|_2^2 / \sigma)}, i = 1,2,\ldots,M,$$

- K is set to a smaller number and accordingly $[b_1, . . . , b_K]$ denote the K closest codewords of x.

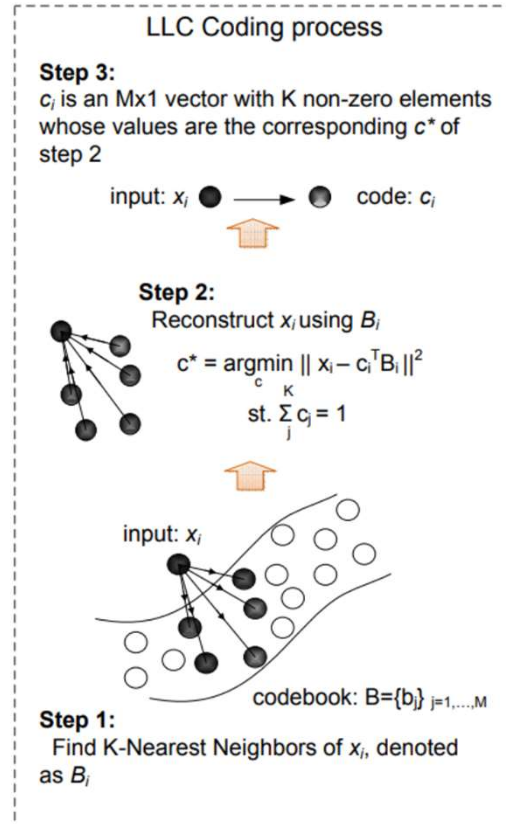
Reconstruction-Based Coding

- Sparse Coding

$$\arg\min_{\mathbf{C}} \sum_{i=1}^{N} \|\mathbf{x}_i - \mathbf{B}\mathbf{c}_i\|^2 + \lambda \|\mathbf{c}_i\|_{\ell^1}$$

- Locality-constrained Linear Coding (LLC)

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



Saliency-Based Coding

- The saliency coding employs the difference between the closest codeword and the other K-1 closest codewords to reflect saliency.

$$v(i) = \begin{cases} \psi(x), & if\ i = \arg\min_j (\|x - b_j\|_2) \\ 0, & otherwise, \end{cases}$$

$$\psi(x) = \sum_{j=2}^{K} (\|x - \tilde{b}_j\|_2 - \|x - \tilde{b}_1\|_2)/\|x - \tilde{b}_j\|_2,$$

- Where $\psi(x)$ denotes the saliency degree and $[\tilde{b}_1, \tilde{b}_2, \ldots, \tilde{b}_K]$ are the K closest codewords to x.

Spatial Pyramid Matching (Pooling)