Patch Descriptors -2

ECE P 596

Linda Shapiro

Bag-of-words models

- Orderless document representation: frequencies of words from a dictionary Salton & McGill (1983)
- Retrieve documents based on matching words in a query to frequencies of words in the document

- Computer vision people grabbed this idea and invented the idea of visual words: little subimages that were representative of an image
- So an image can be retrieved by the frequency of its important subimages

Example in Video:

Sivic's "Video Google" work to retrieve frames from videos

- Two types of viewpoint covariant regions computed for each frame
  - Shape Adapted (SA) Mikolajczyk & Schmid
  - Maximally Stable (MSER) Matas *et al.*
- Detect different kinds of image areas
- Provide complimentary representations of frame
- Computed at twice originally detected region size to be more discriminating

Examples of Harris-Affine Operator

(Shape Adapted Regions)

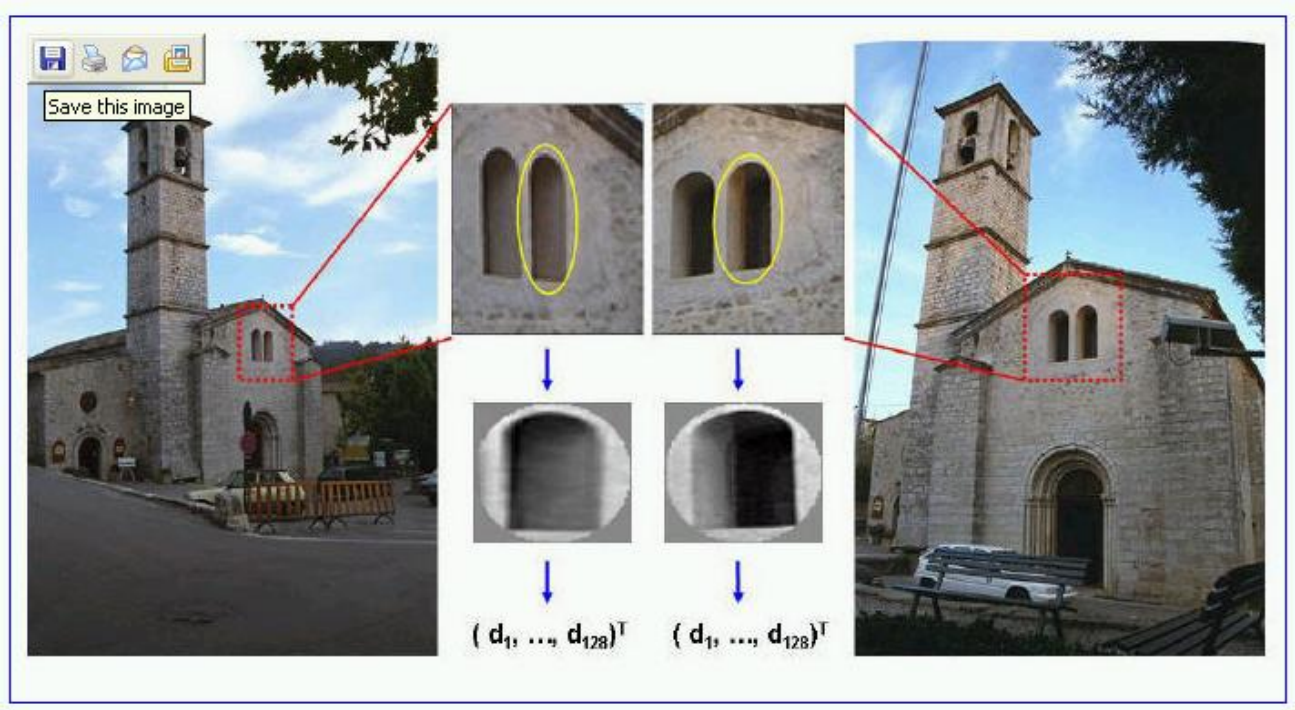
(a) b)

Fig. 6. (a) Example of a 3D scene observed from significantly different viewpoints. There are 14 inliers to a robustly estimated fundamental matrix, all of them correct. (b) An image pairs for which our method fails. There exist, however, corresponding points which we have selected manually.

Examples of Maximally Stable Regions

Feature Descriptor

- Each region represented by 128 dimensional vector using SIFT descriptor

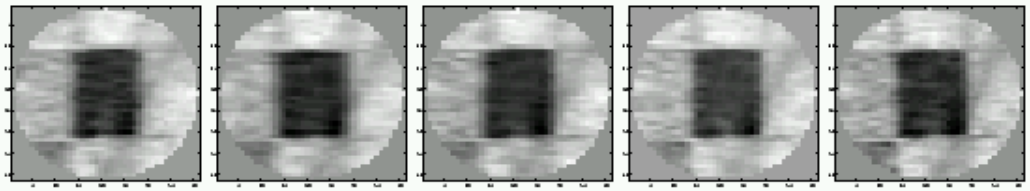


Noise Removal

- **Tracking region over 70 frames (must track over at least 3)**

First (left) and last (right) frame of the track.

Close-up of the 1st, 20th, 40th, 55th, 70th frame.



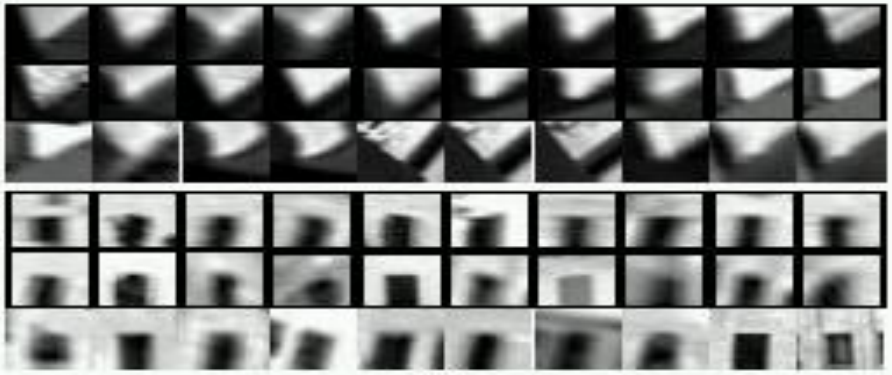
Visual Vocabulary for Sivic's Work

- Implementation: K-Means clustering
- Regions tracked through contiguous frames and average description computed
- 10% of tracks with highest variance eliminated, leaving about 1000 regions per frame
- Subset of 48 shots (~10%) selected for clustering
- Distance function: Mahalanobis
- 6000 SA clusters and 10000 MS clusters

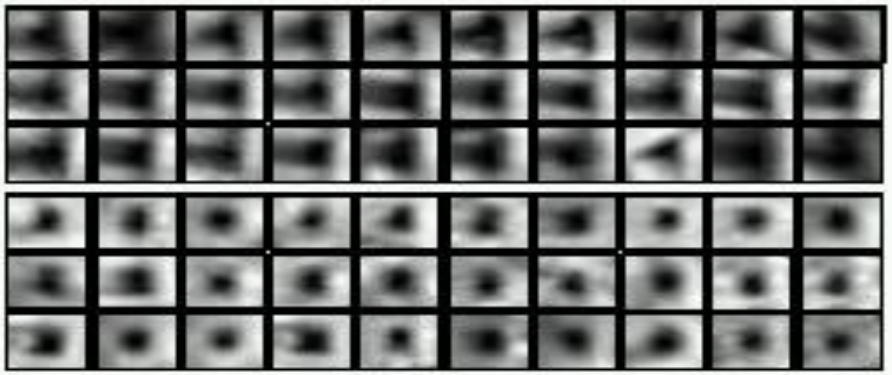
Visual Vocabulary

Shape-Adapted

Maximally Stable



(a)



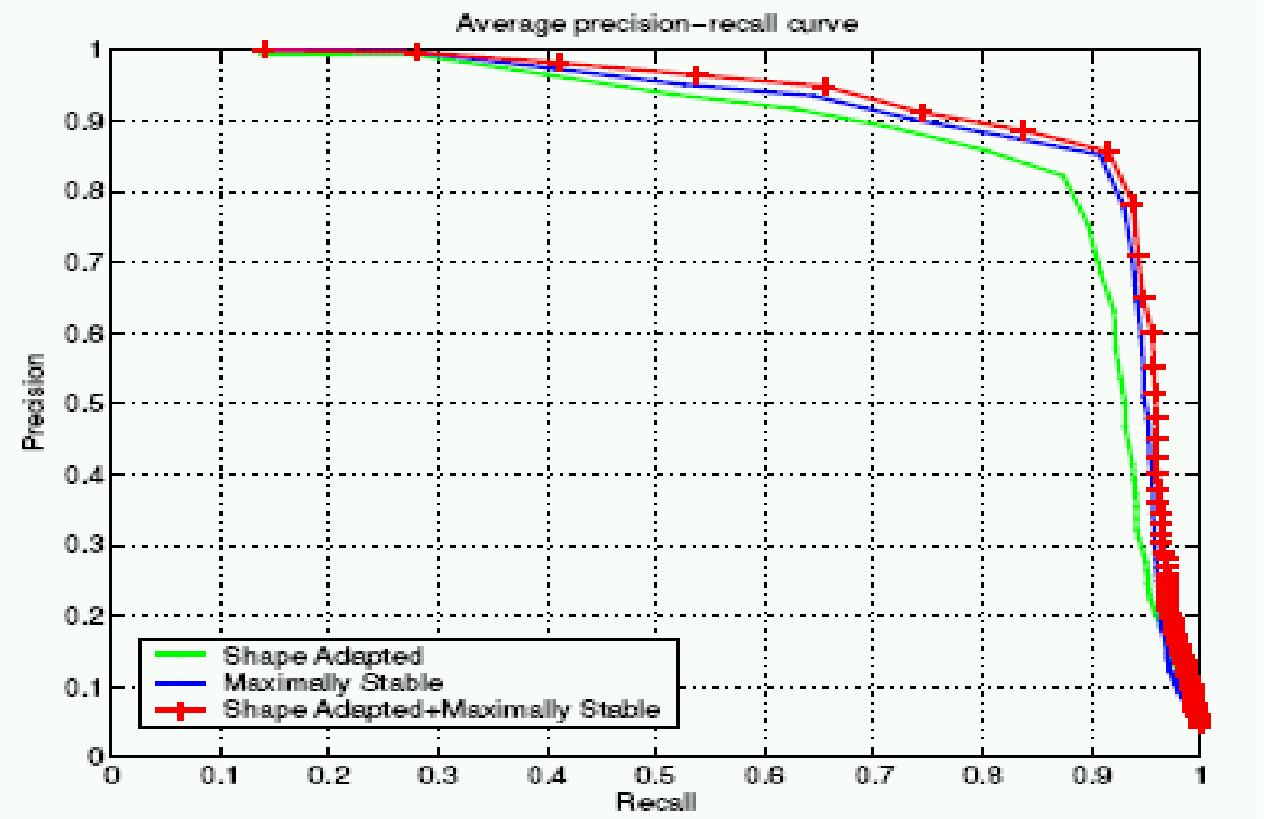
(b)

Figure 2: Samples from the clusters corresponding to a single visual word. (a) Two examples of clusters of Shape Adapted regions. (b) Two examples of clusters of Maximally Stable regions.

Sivic's Experiments on Video Shot Retrieval

- Goal: match scene locations within closed world of shots
- Data:164 frames from 48 shots taken at 19 different 3D locations; 4-9 frames from each location

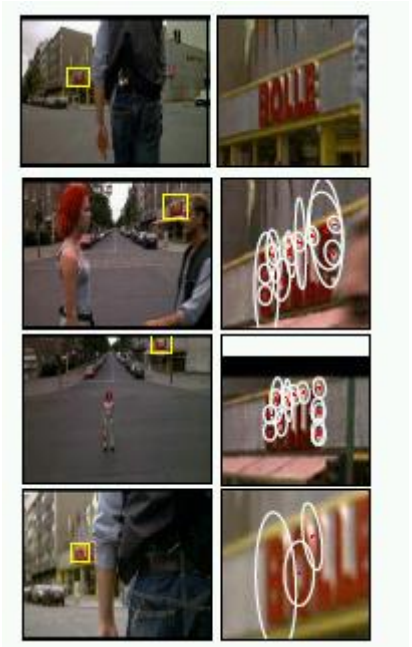
Experiments - Results



Precision = # relevant images/total # of frames retrieved

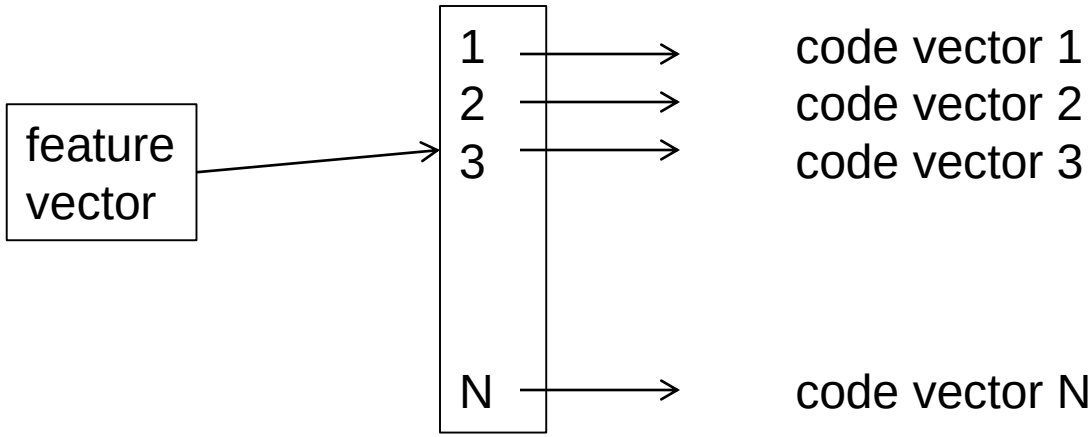
Recall = # correctly retrieved frames/ # relevant frames

More Pictorial Results

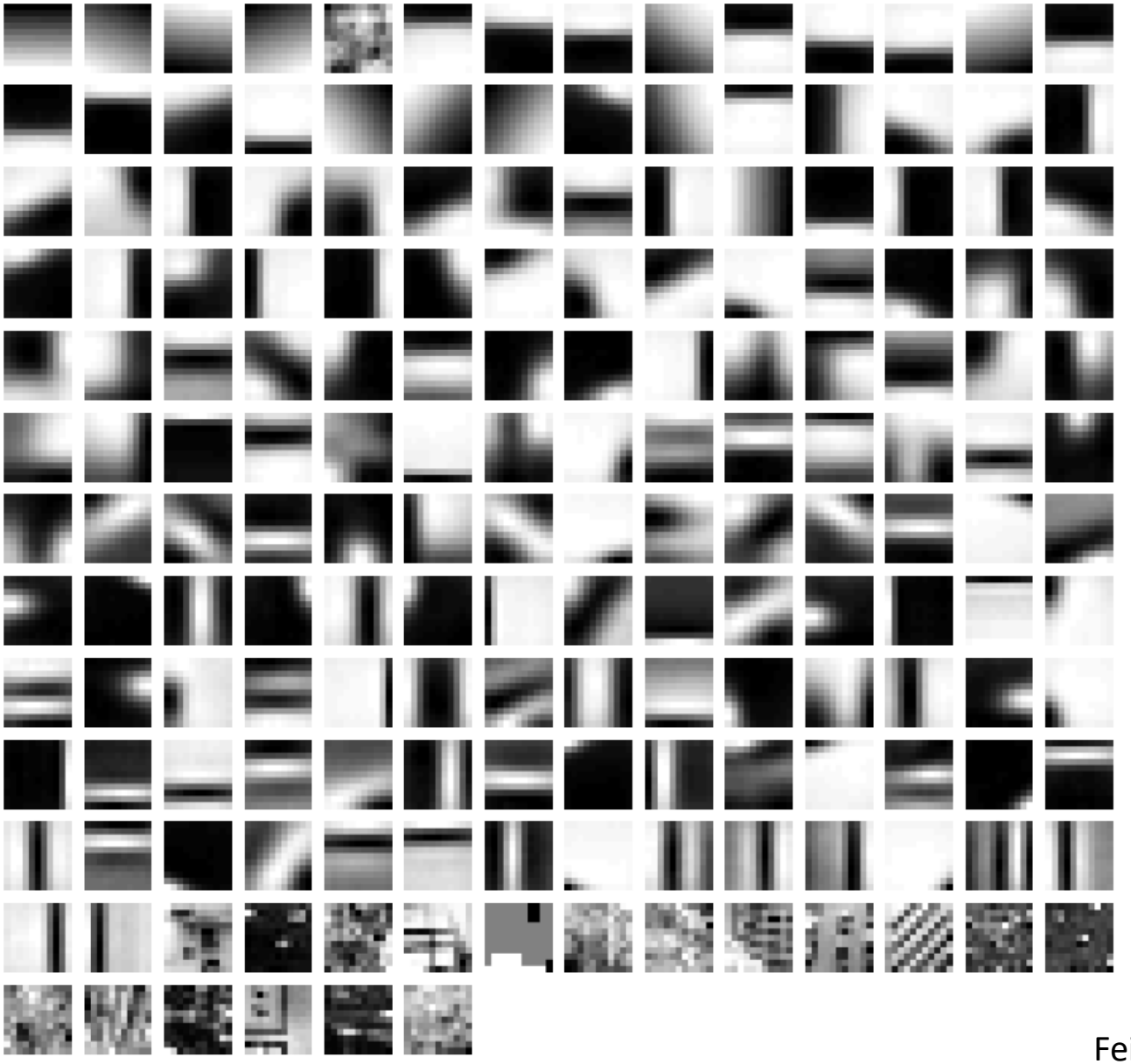


Clustering and vector quantization

- Clustering is a common method for learning a visual vocabulary or codebook
  - Each cluster center produced by k-means becomes a codevector
  - Codebook can be learned on separate training set
- The codebook is used for quantizing features
  - A *vector quantizer* takes a feature vector and maps it to the index of the nearest code vector in a codebook
  - Codebook = visual vocabulary
  - Code vector = visual word

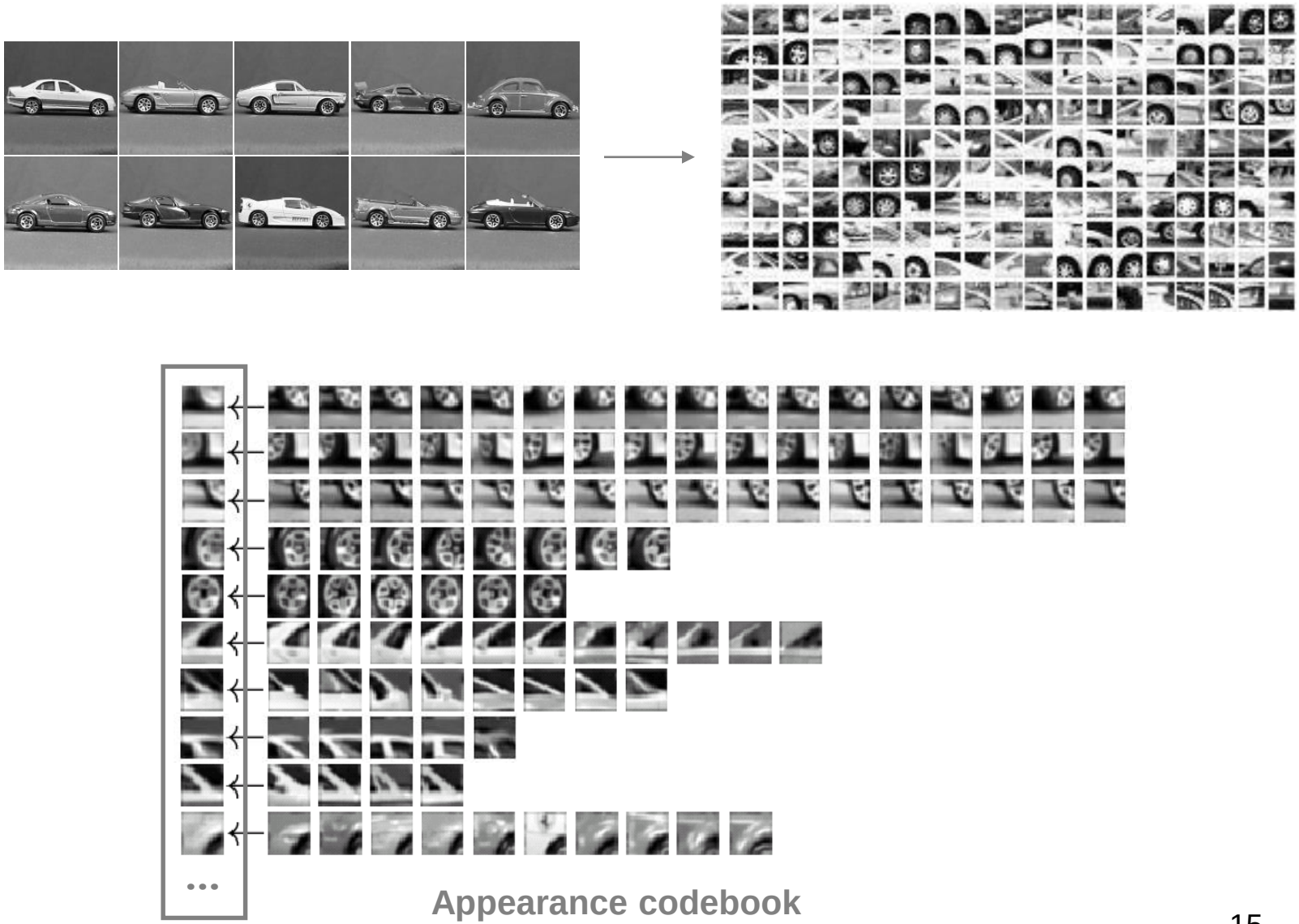


Another example visual vocabulary



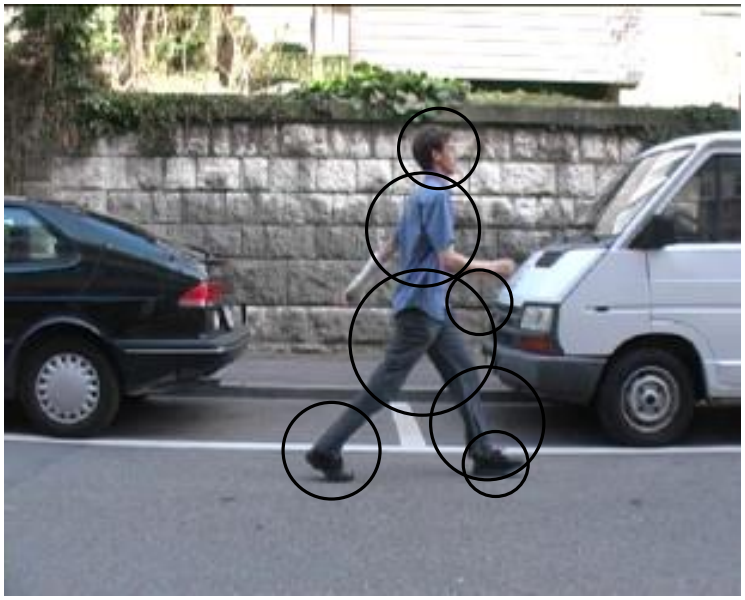
Fei-Fei et al. 2005

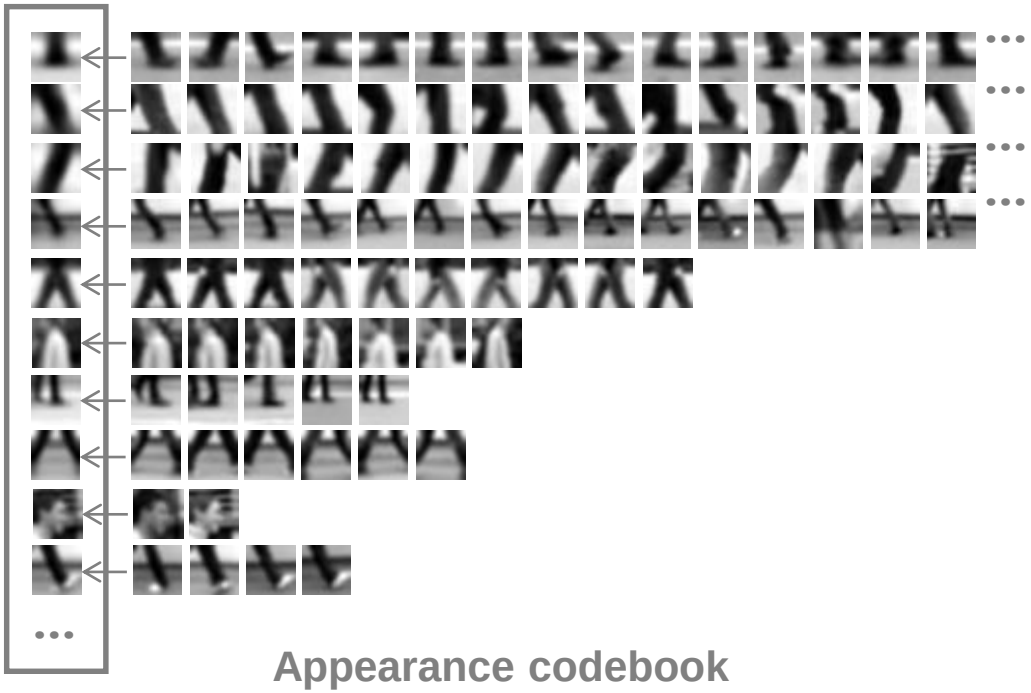
Example codebook



Source: B. Leibe

Another codebook





Appearance codebook

Source: B. Leibe

Image representation: histogram of codewords

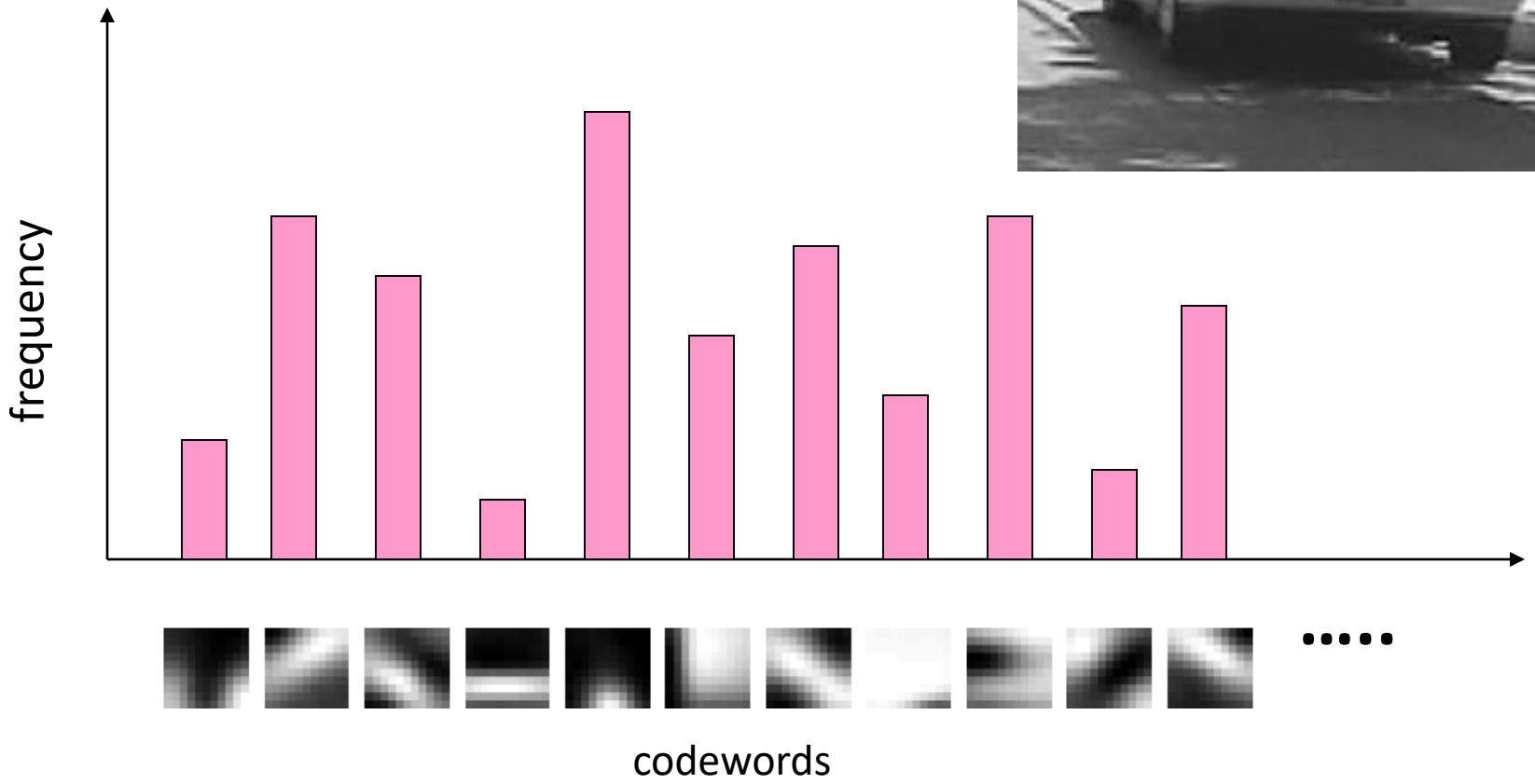
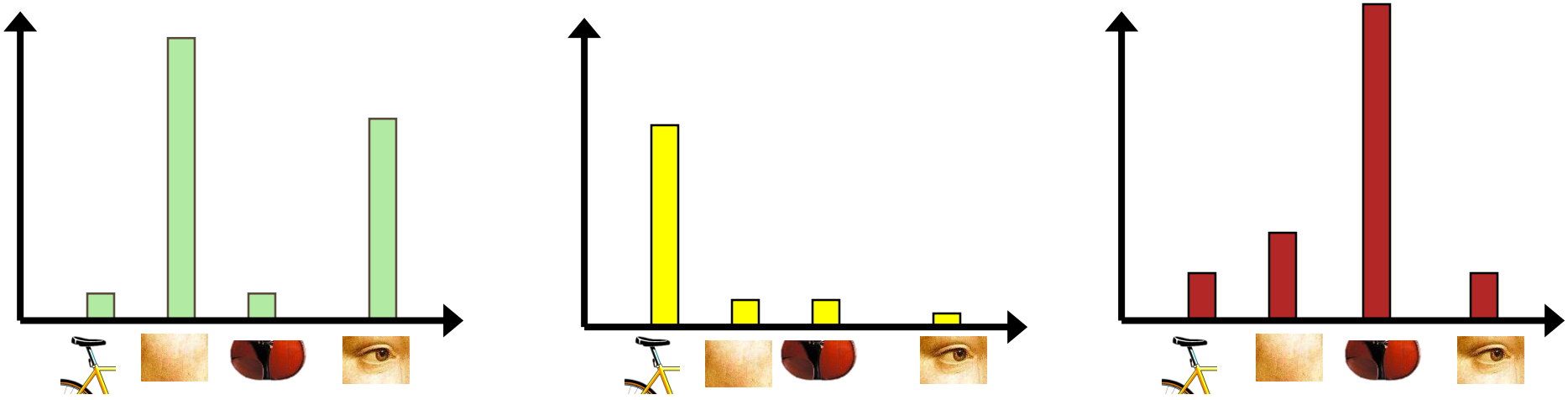
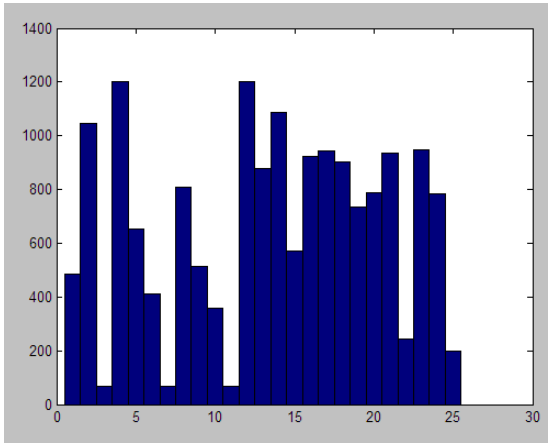


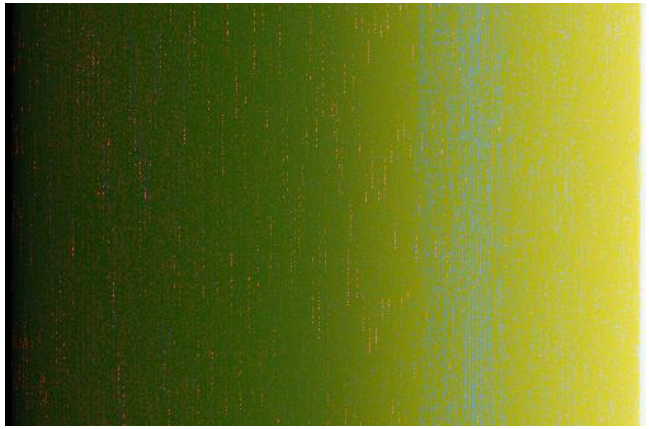
Image classification (later in course)

- Given the bag-of-features representations of images from different classes, learn a classifier using machine learning



But what about layout?

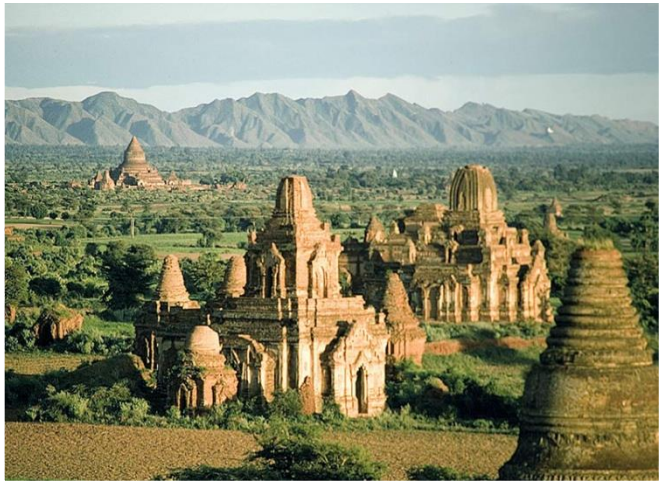


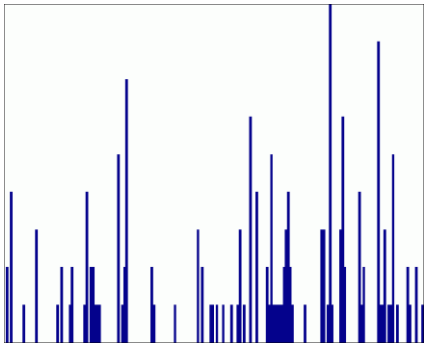


All of these images have the same color histogram

Spatial pyramid representation

- Extension of a bag of features
- Locally orderless representation at several levels of resolution

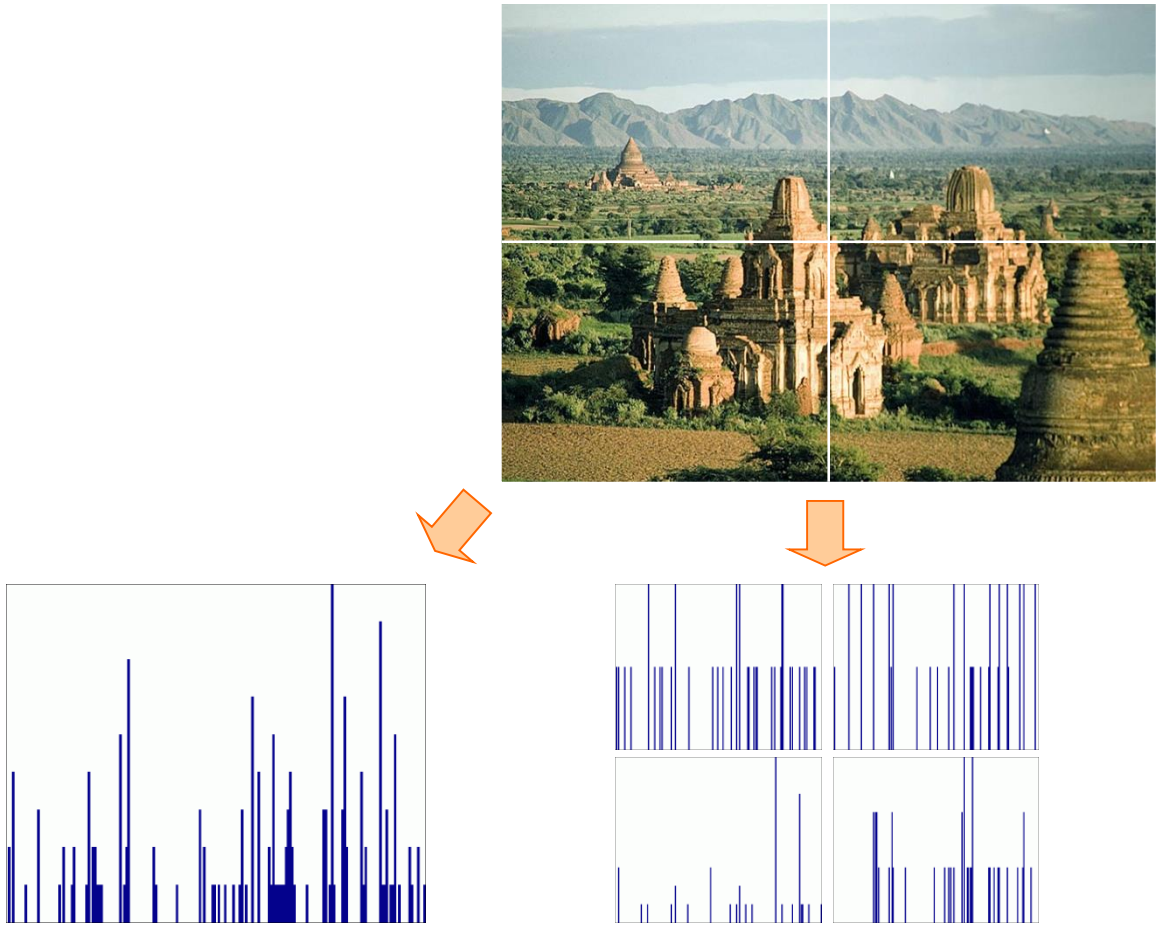




Lazebnik, Schmid & Ponce (CVPR 2006)

Spatial pyramid representation

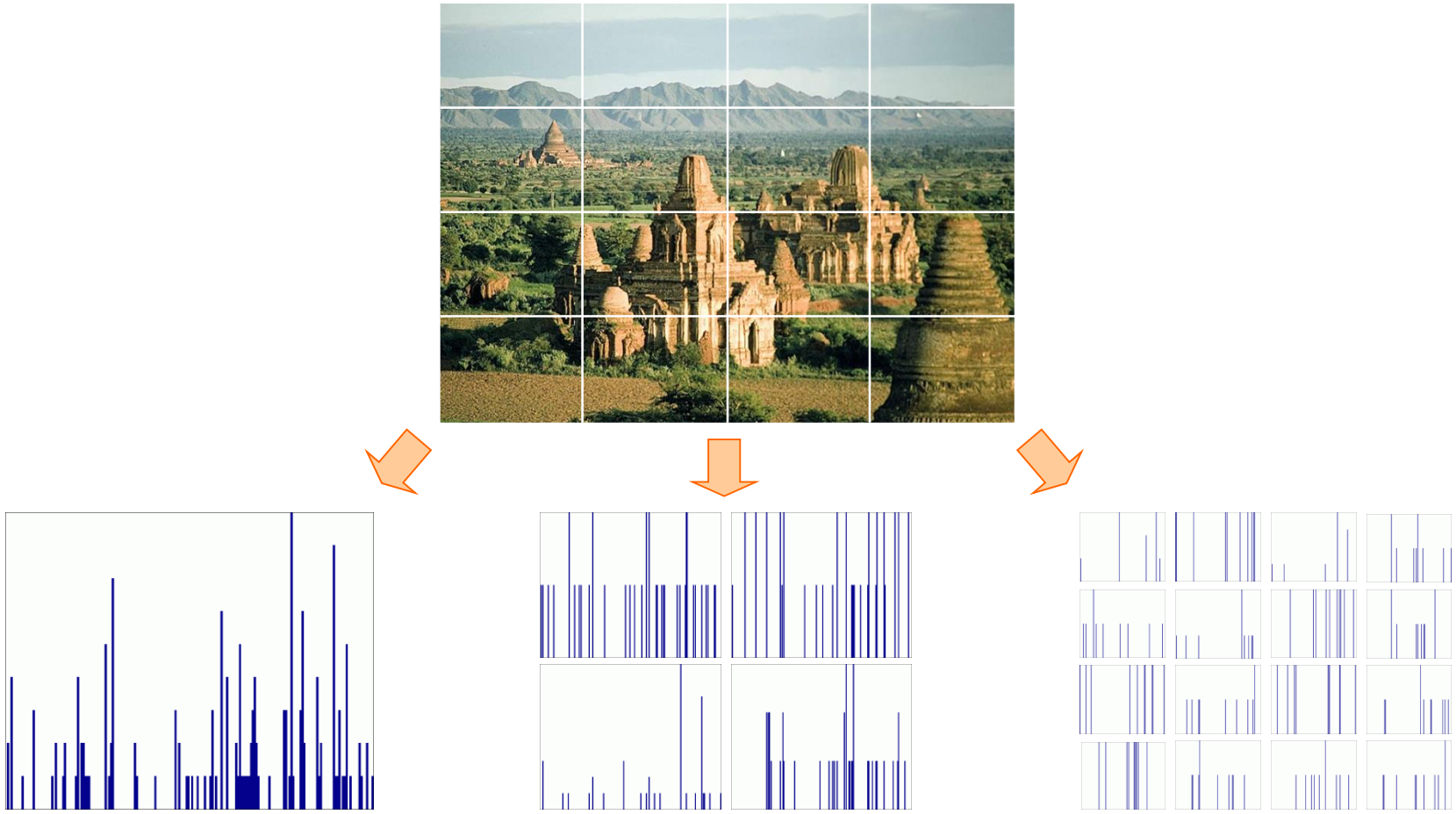
- Extension of a bag of features
- Locally orderless representation at several levels of resolution



Lazebnik, Schmid & Ponce (CVPR 2006)

Spatial pyramid representation

- Extension of a bag of features
- Locally orderless representation at several levels of resolution



Lazebnik, Schmid & Ponce (CVPR 2006)

Finale

- Describing images or image patches is very important for matching and recognition
- Texture descriptors are also useful.
- Bag-of-words is a handy technique borrowed from text retrieval. Lots of people use it to compare images or regions.
- Sivic developed a video frame retrieval system using this method, called it Video Google.
- The spatial pyramid allows us to describe an image as a whole and over its parts at multiple levels.