
BIG collections of images and videos
1-100 million images
Thousands hours of video


query
image
Find all images that depict visual
content similar to the query
(same scene, location, etc)
→ image retrieval/annotation


query
image
Find all images that depict visual
content similar to the query
(same scene, location, etc)
→ image retrieval

Find some particular object for which you have an example

→ **particular object retrieval**


Find some particular object for which you have an example
→ particular object retrieval


Find all images where an object of a "class" appear:
Find all "trees"
→ image classification

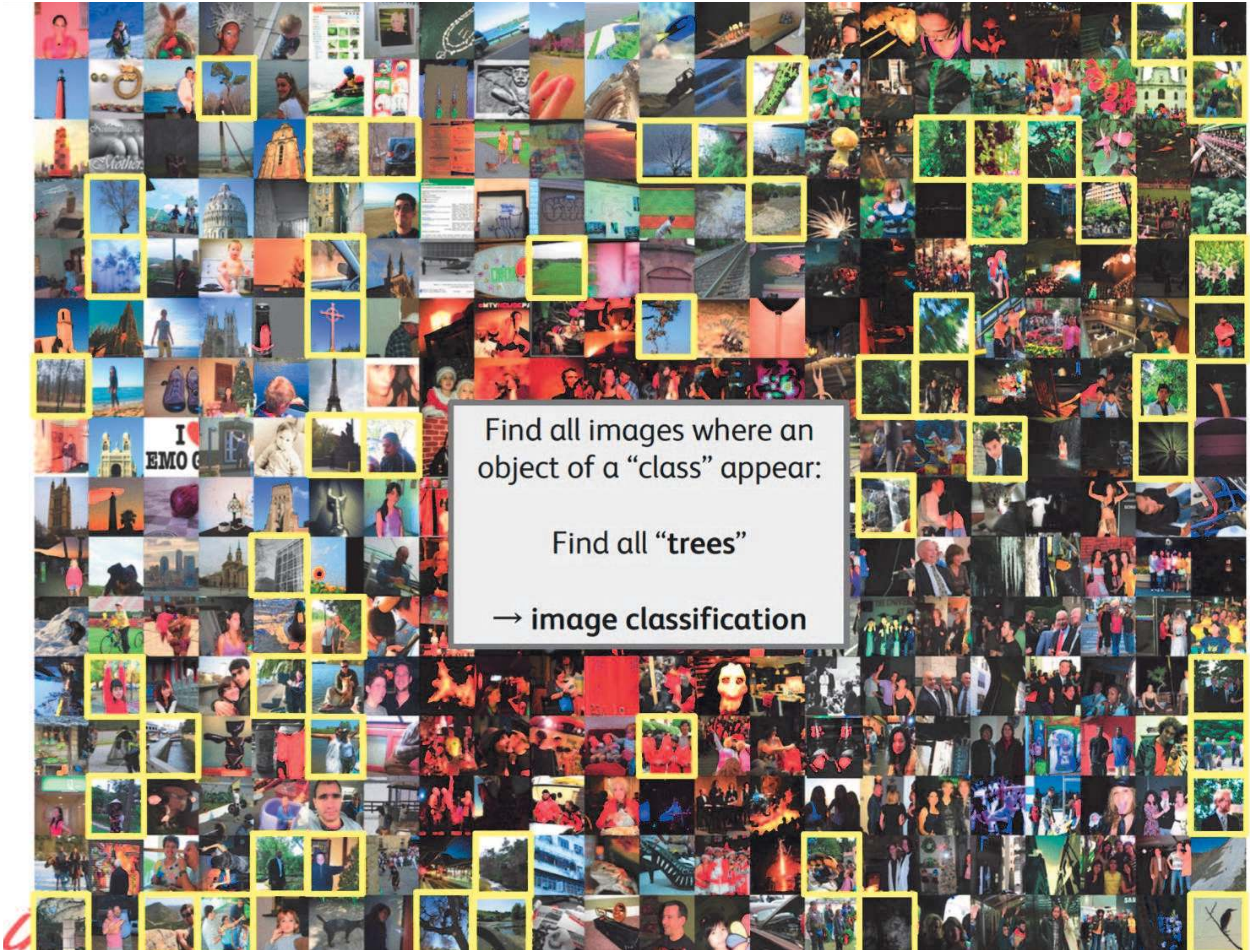

Find all images where an object of a "class" appear:
Find all "trees"
→ image classification

Image retrieval challenges

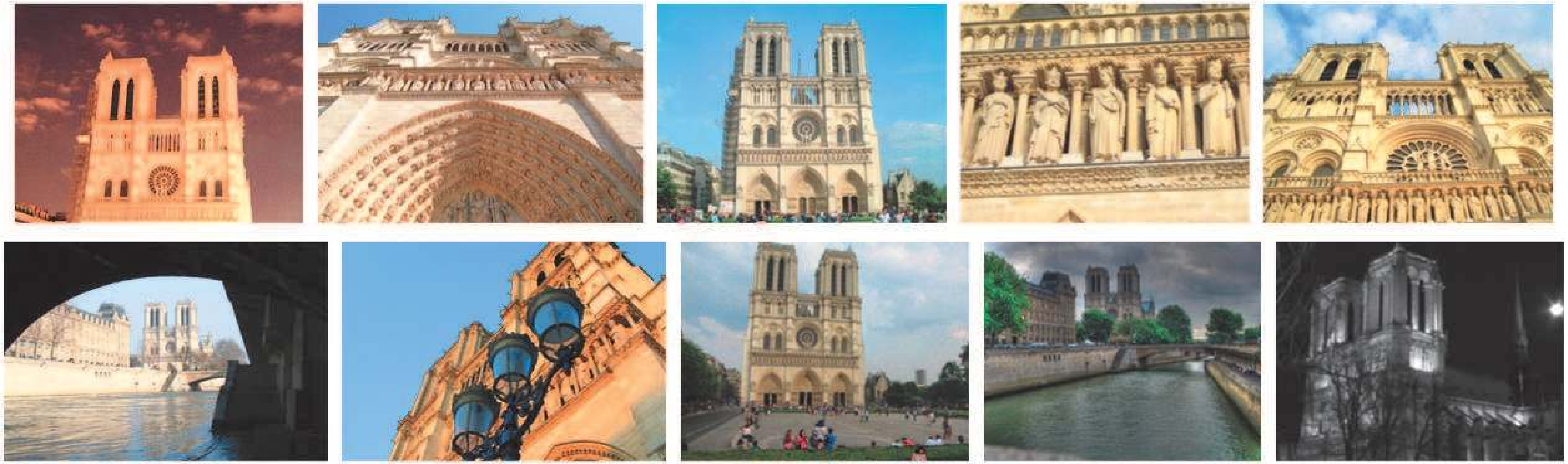
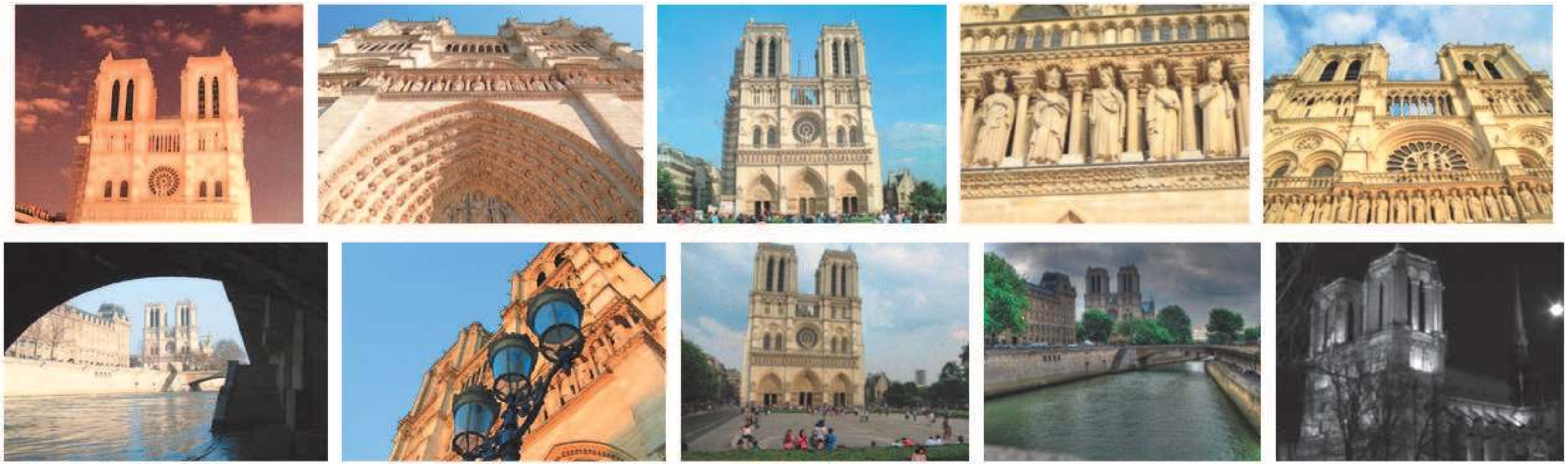


Image retrieval challenges



- scale
- viewpoint
- occlusion
- clutter
- lighting
- distinctiveness
- distractors

Image classification challenges

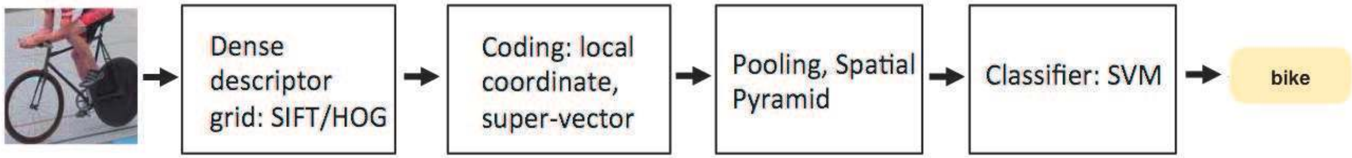
Image classification challenges

- scale
- viewpoint
- occlusion
- clutter
- lighting
- number of instances
- texture/color
- pose
- deformability
- intra-class variability

Visual descriptors

Visual descriptors

Pre-deep pipeline



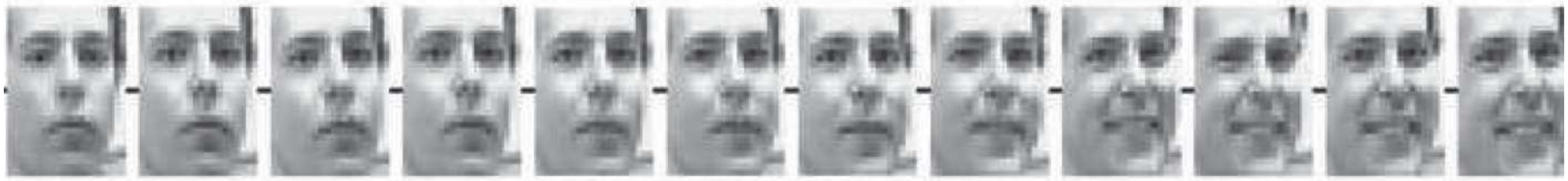
Visual descriptors

Concatenation of pixels into 1D descriptors

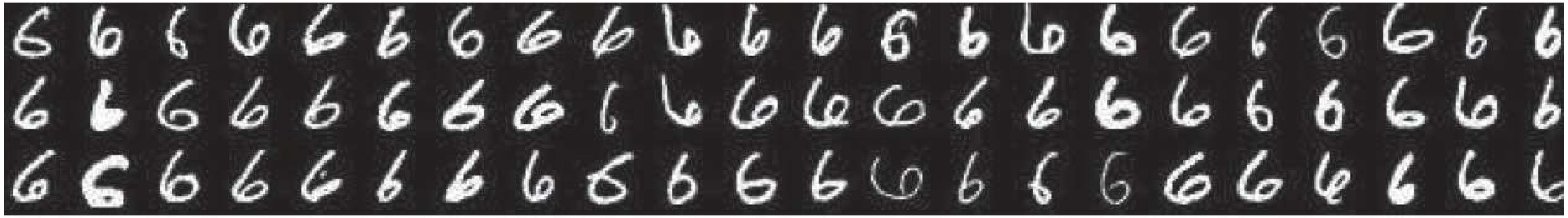
Global descriptors

Concatenation of pixels into 1D descriptors

- face recognition



- digit recognition



Global descriptors

Tiny images

- resize images to 32×32 pixels ($3072d$ vectors)

- high speed, limited accuracy
- used for scene recognition

Global descriptors

Color histogram

- Histogram is a summary of the data describing image statistics (here color)

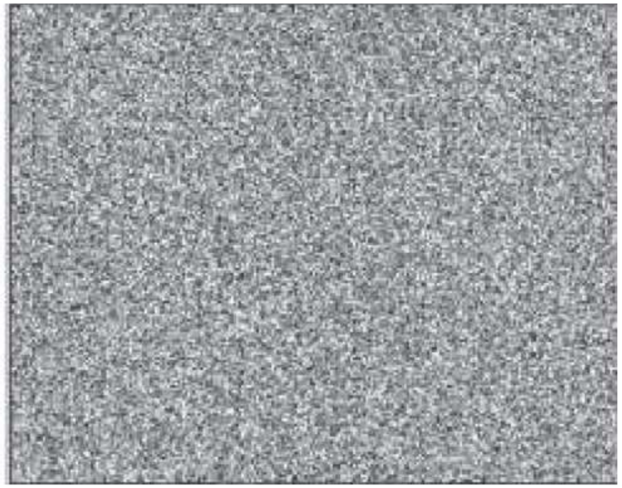
Global descriptors

Color histogram

- Histogram is a summary of the data describing image statistics (here color)

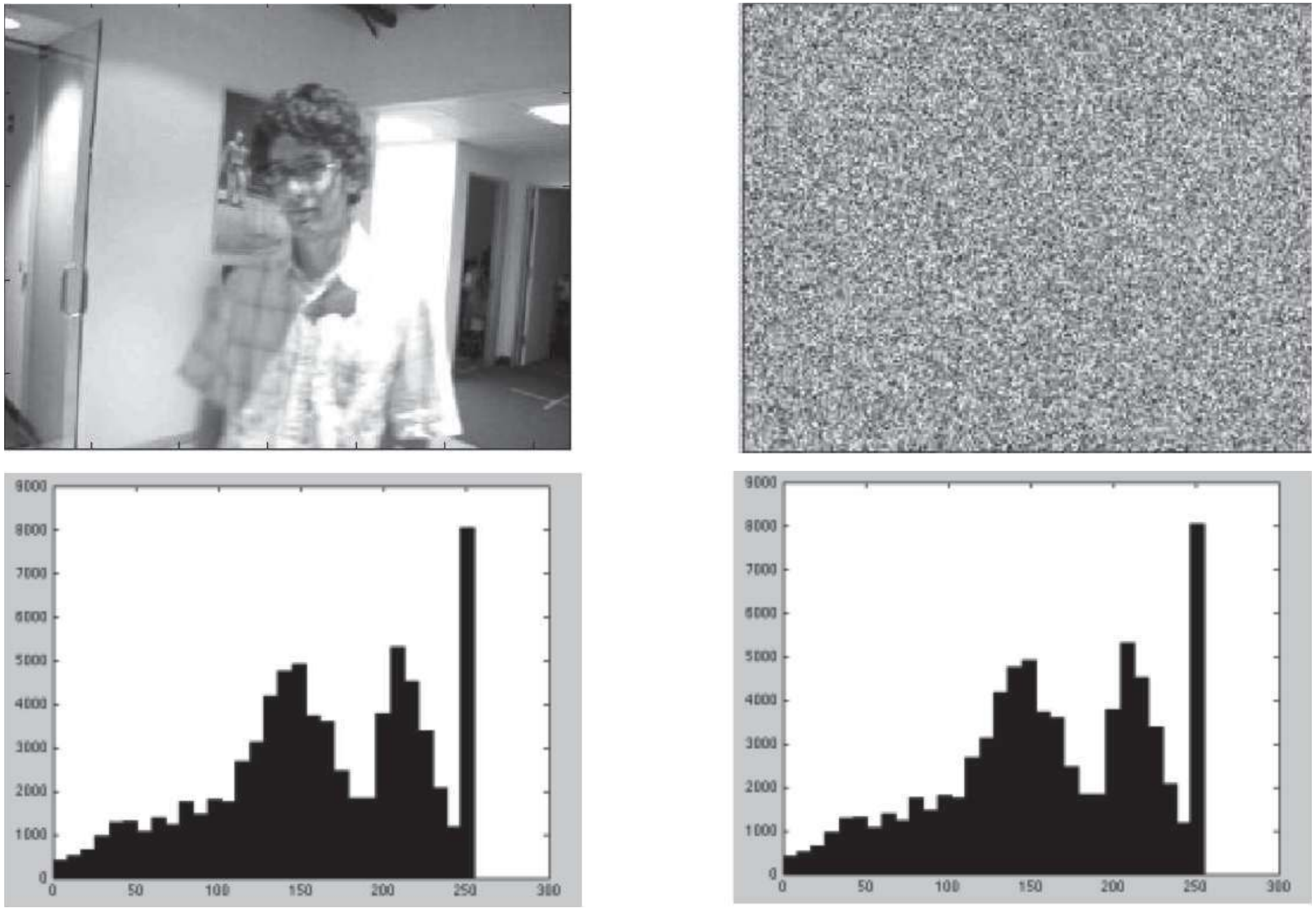
Global descriptors

Color histogram

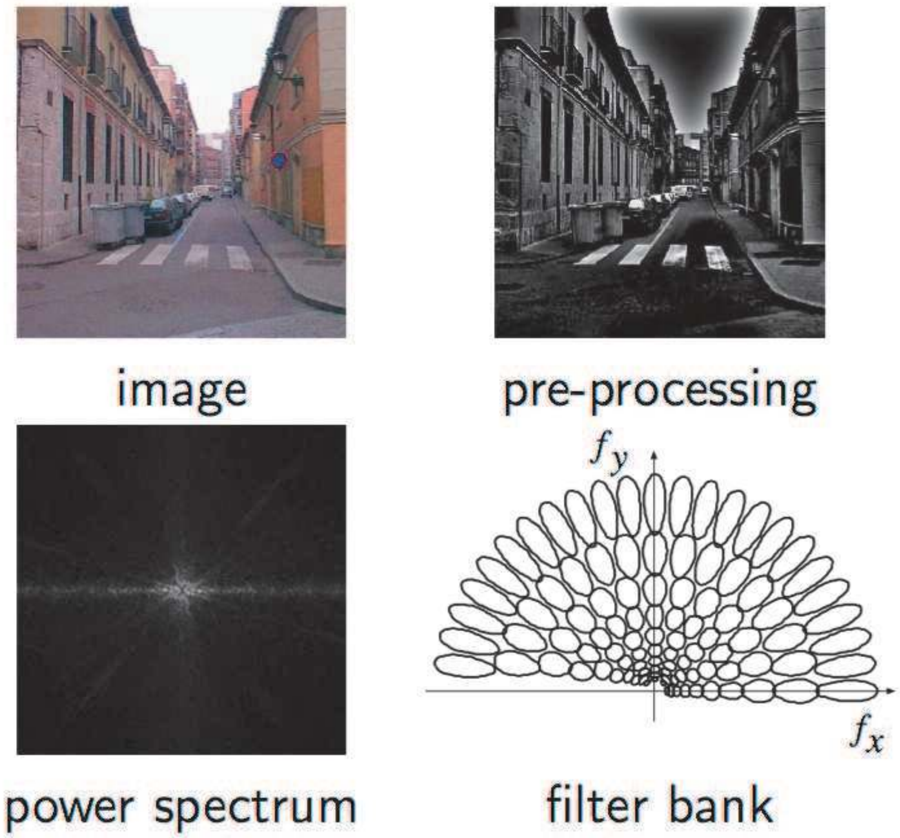


Global descriptors

Color histogram

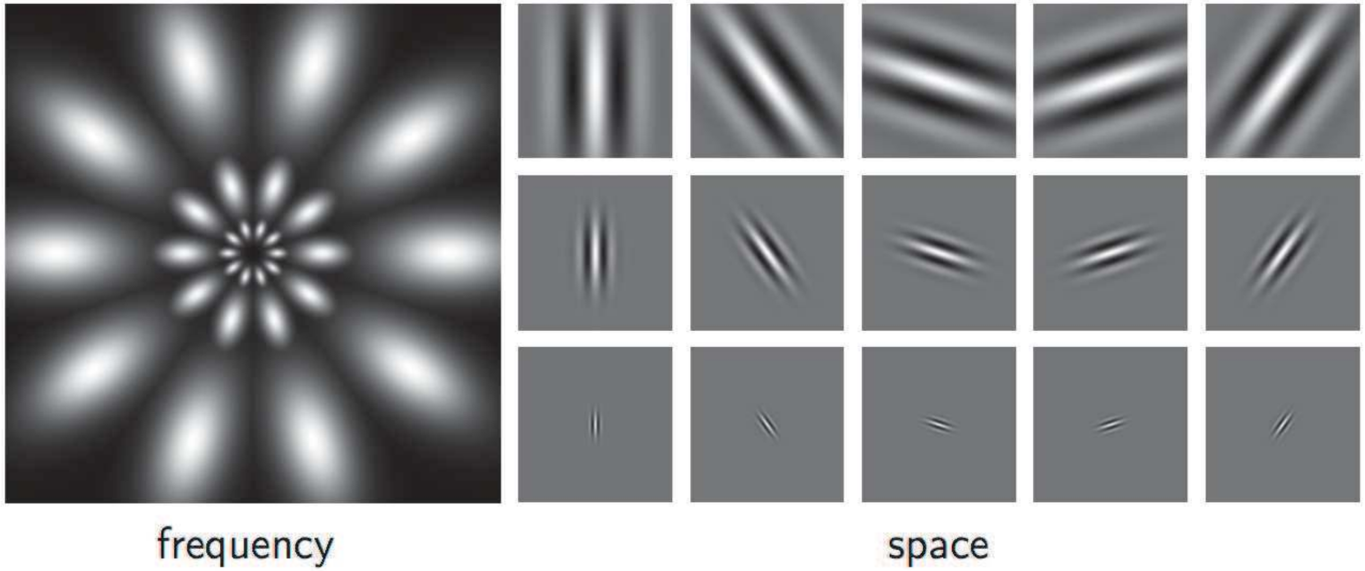


Global descriptors



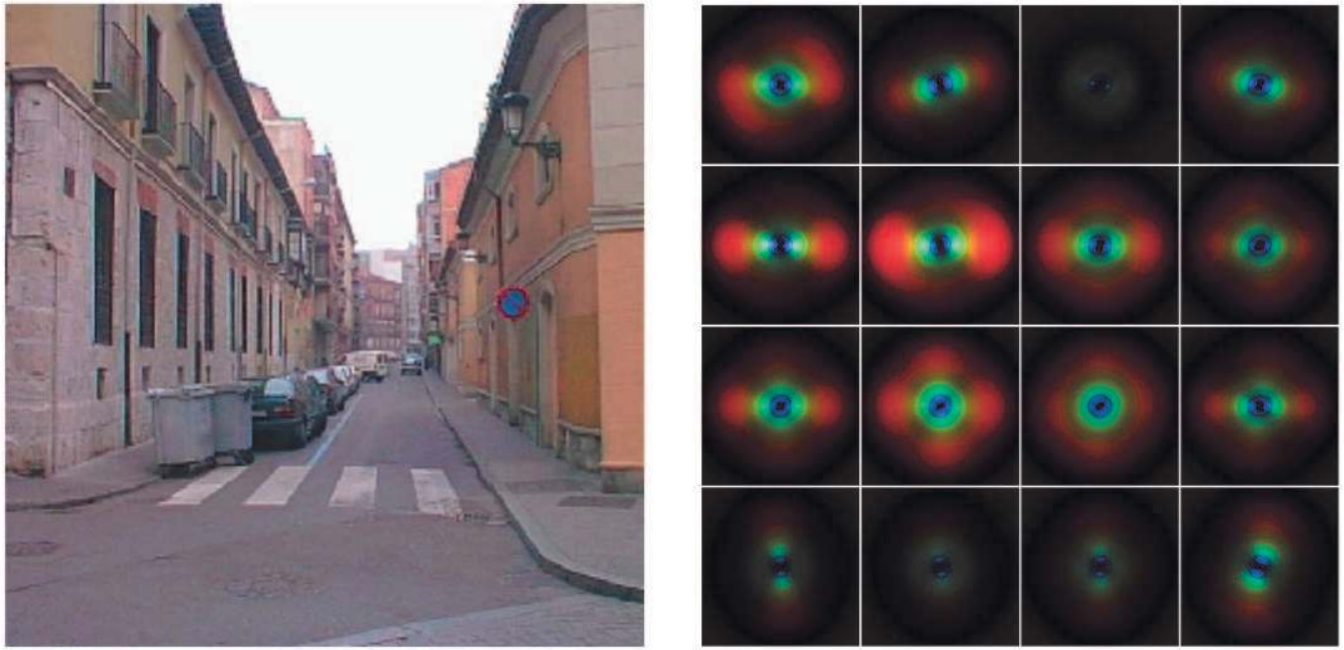
- sampling scheme adapted to power spectrum statistics
- filtering and global pooling in frequency domain

Global descriptors



Global descriptors

The gist descriptor

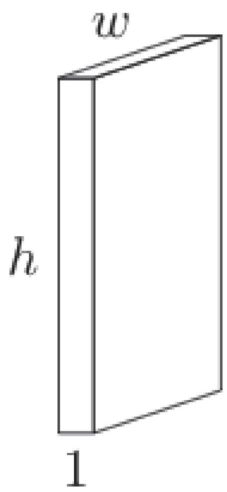


- apply filter bank to entire image in frequency domain
- partition image in 4×4 cells
- average pooling of filter responses per cell

Building the Gist of a Scene: the Role of Global Image Features in Recognition; Oliva and Torralba; VP 2006

Global descriptors

gist pipeline

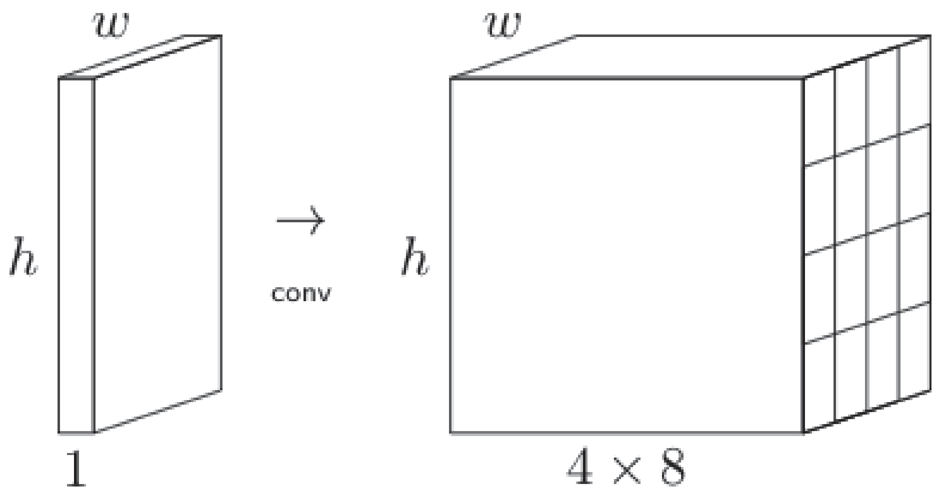


- 3-channel RGB input $\rightarrow$ 1-channel gray-scale

Building the Gist of a Scene: the Role of Global Image Features in Recognition; Oliva and Torralba; VP 2006

Global descriptors

gist pipeline

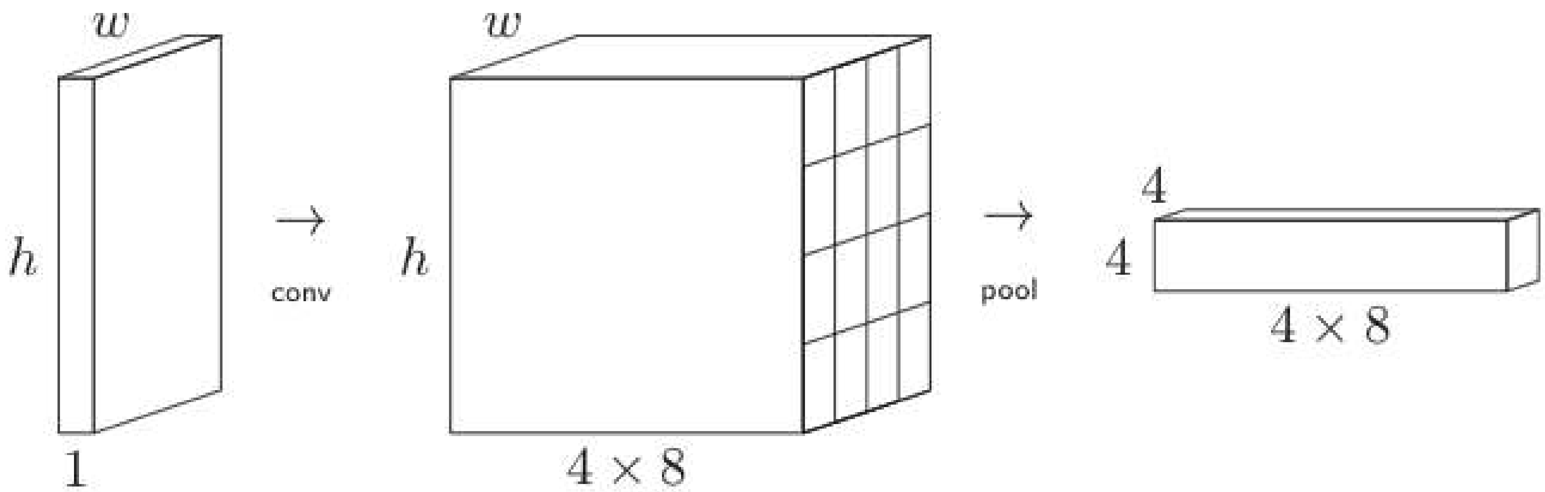


- 3-channel RGB input $\rightarrow$ 1-channel gray-scale
- apply filters at 4 scales $\times$ 8 orientations

Building the Gist of a Scene: the Role of Global Image Features in Recognition; Oliva and Torralba; VP 2006

Global descriptors

gist pipeline

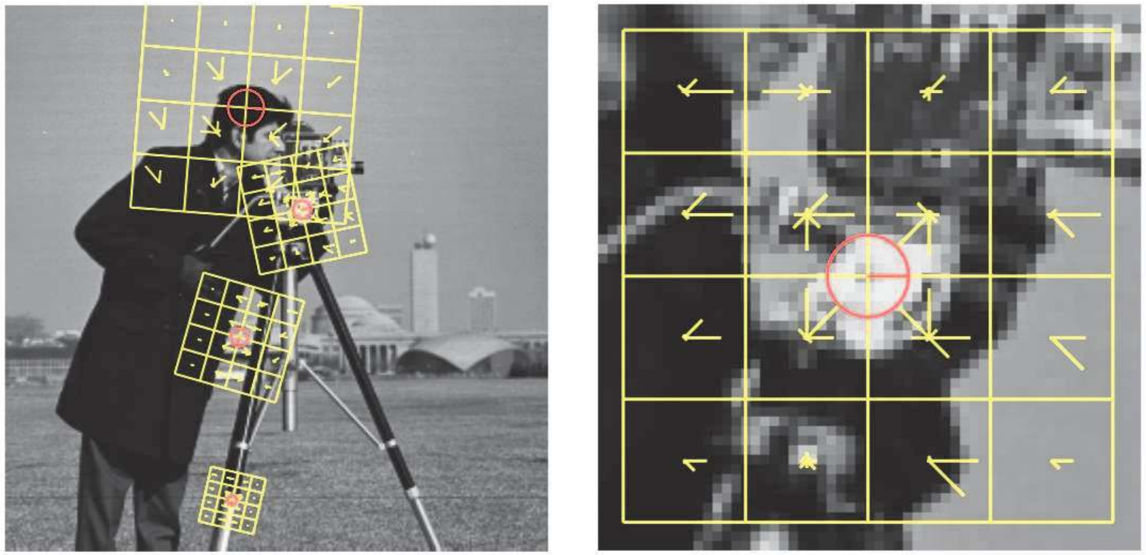


- 3-channel RGB input $\to$ 1-channel gray-scale
- apply filters at 4 scales $\times$ 8 orientations
- average pooling on 4×4 cells $\to$ descriptor of length 512

Building the Gist of a Scene: the Role of Global Image Features in Recognition; Oliva and Torralba; VP 2006

Local descriptors

scale-invariant feature transform (SIFT)

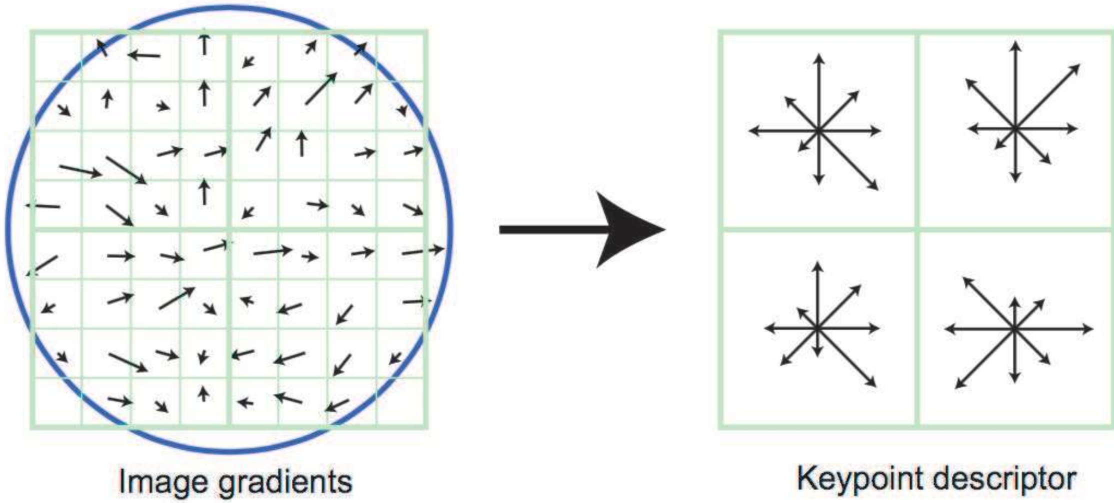


- detect a sparse set of "stable" features (rectangular patches) equivariant to translation, scale and rotation
- for each patch:
  - normalize with respect to scale and orientation
  - construct a histogram of gradient orientations

Object recognition from local scale-invariant features.; Lowe; ICCV 1999

Local descriptors

scale-invariant feature transform (SIFT)

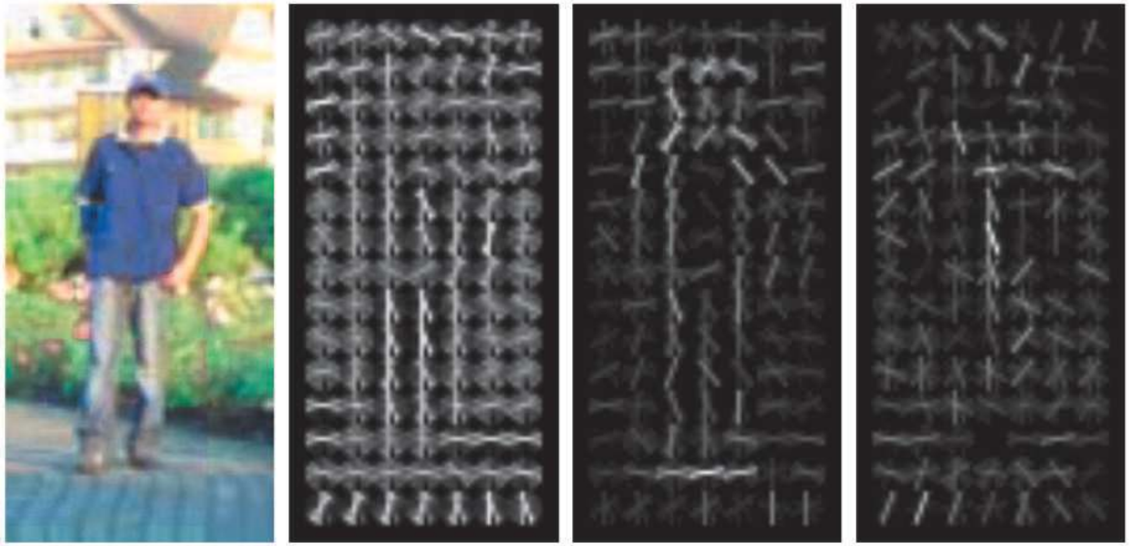


- votes in $8-$bin orientation histograms weighted by magnitude and by weighted by a Gaussian window,
- histograms pooled over 4×4 cells,
- 128-dimensional descriptor, normalized, clipped at 0.2, normalized

Object recognition from local scale-invariant features.; Lowe; ICCV 1999

Local descriptors

Histogram of Oriented Gradients (HoG)



- applied to person detection by sliding window and SVM
- classifier learns positive and negative weights on positions and orientations
- switch focus back to dense features for classification

Histogram of Oriented Gradients for Human Detection; Dalal and Triggs; CVPR 2005

Local descriptors

HOG descriptor

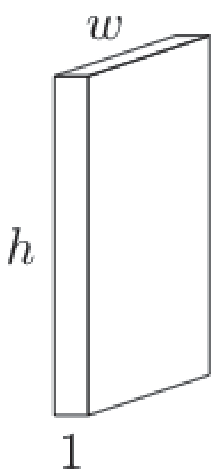
 

- applied densely to adjacent cells of 8×8 pixels
- no scale or orientation normalization; only single-scale
- normalized by overlapping blocks of 3×3 cells -- redundant

Histogram of Oriented Gradients for Human Detection; Dalal and Triggs; CVPR 2005

Local descriptors

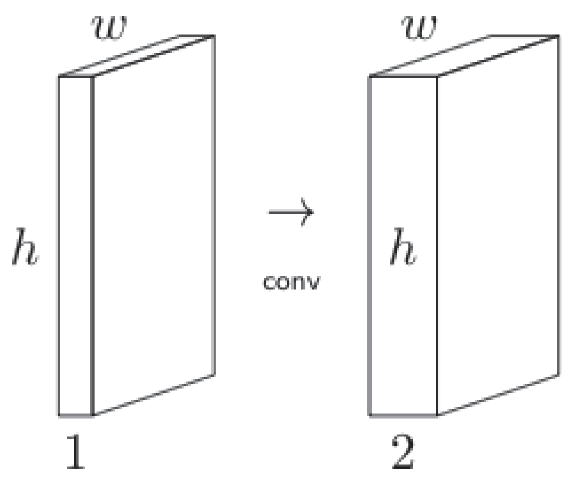
SIFT/HOG pipeline



- 3-channel patch (image) RGB input $\rightarrow$ 1-channel gray-scale

Local descriptors

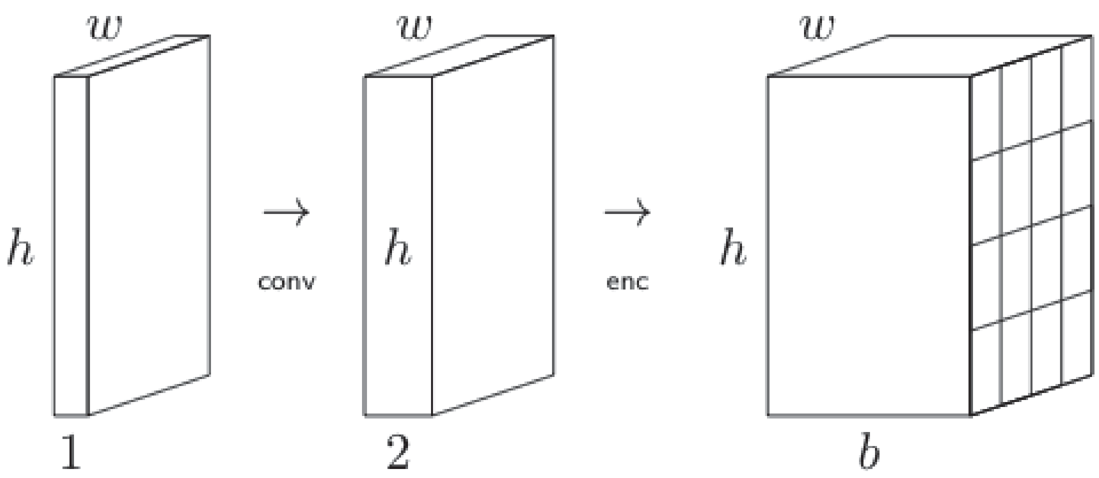
SIFT/HOG pipeline



- 3-channel patch (image) RGB input $\rightarrow$ 1-channel gray-scale
- compute gradient magnitude and orientation

Local descriptors

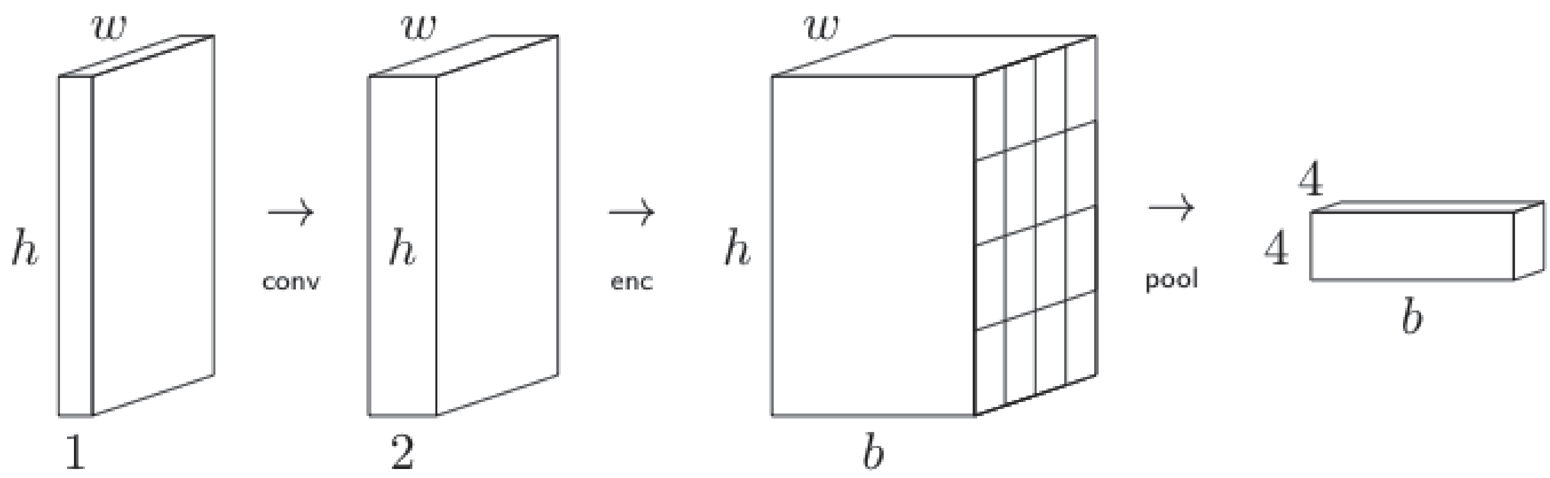
SIFT/HOG pipeline



- 3-channel patch (image) RGB input $\rightarrow$ 1-channel gray-scale
- compute gradient magnitude and orientation
- encode into $b = 8$ orientation bins

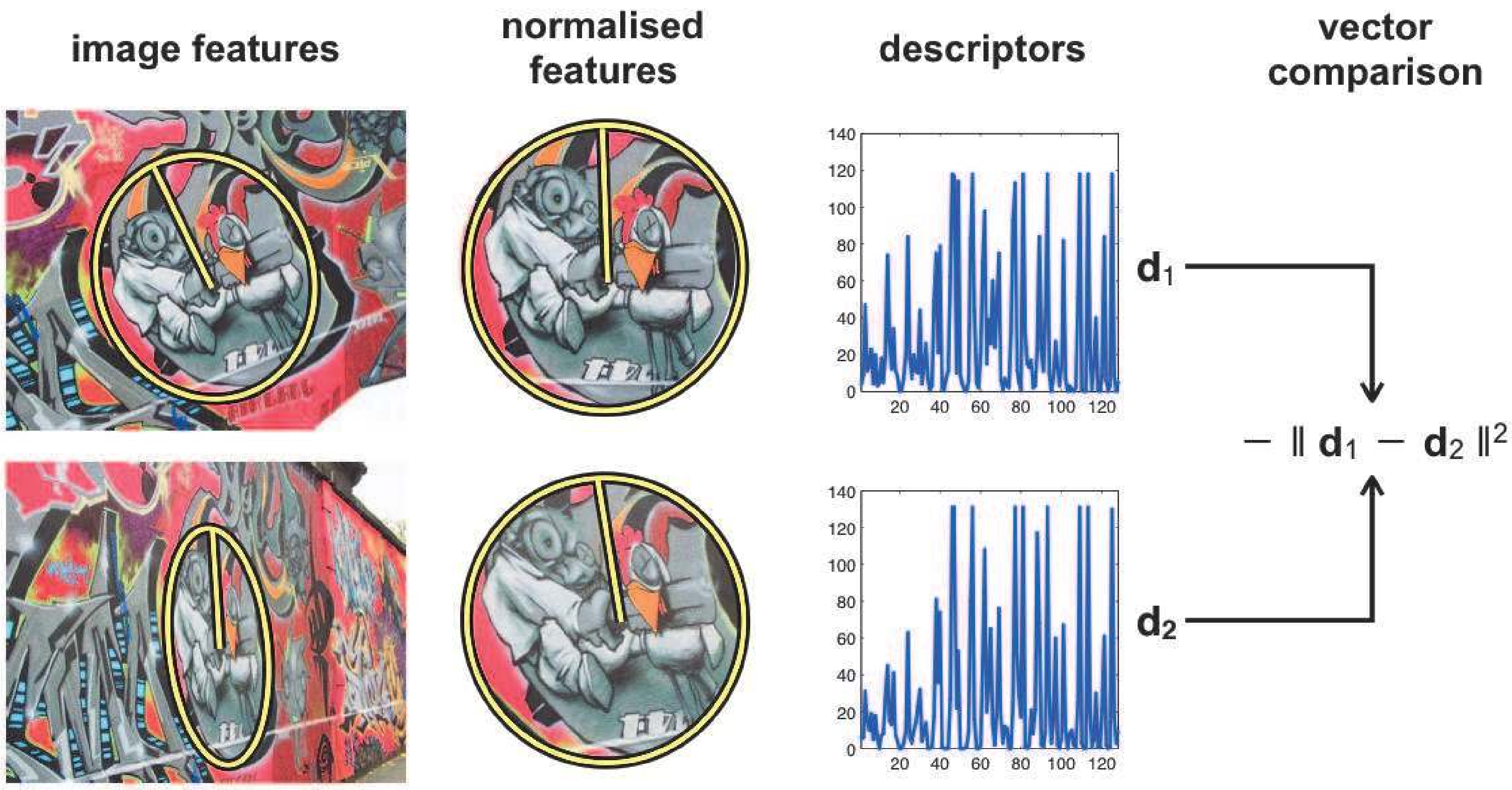
Local descriptors

SIFT/HOG pipeline



- 3-channel patch (image) RGB input $\rightarrow$ 1-channel gray-scale
- compute gradient magnitude and orientation
- encode into $b = 8(9)$ orientation bins
- average pooling on $c = 4 \times 4$ cells
- descriptor of length $c \times b = 128$

Local descriptors



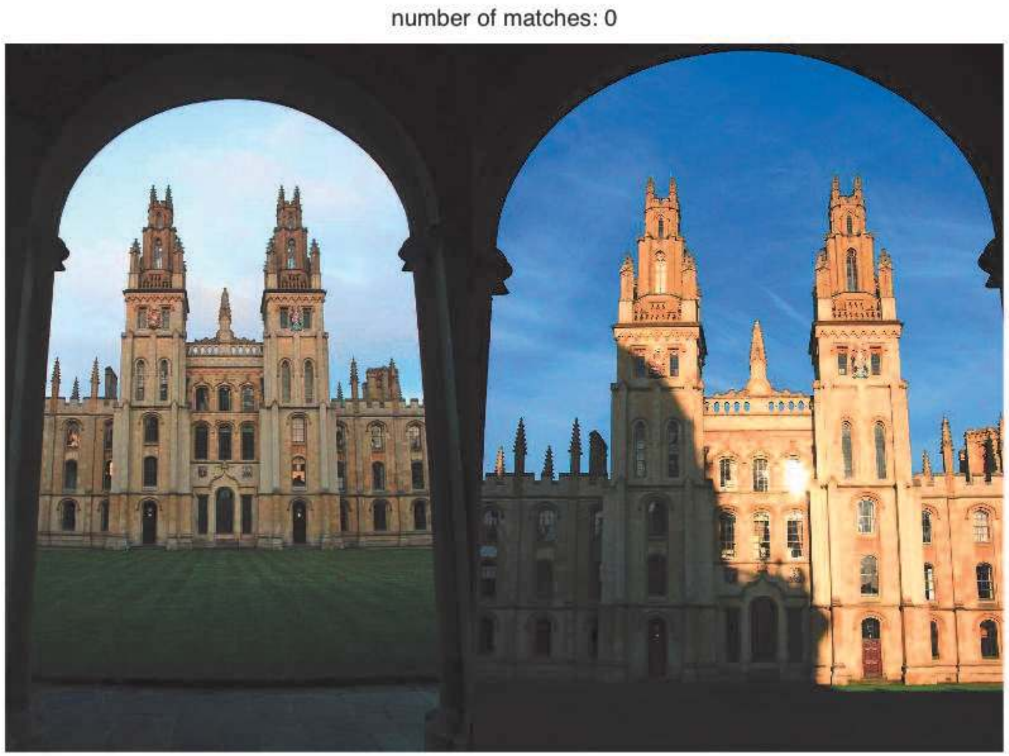
Local descriptors

- matching everything with everything

Local descriptors

Exhaustive matching

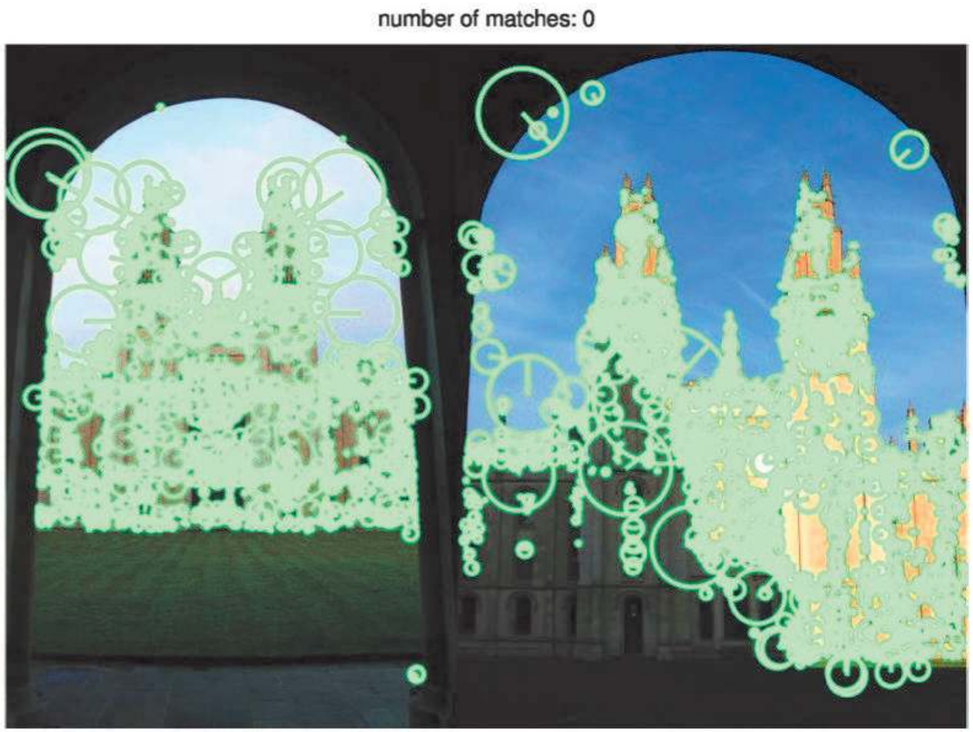
Step 0: get an image pair



Local descriptors

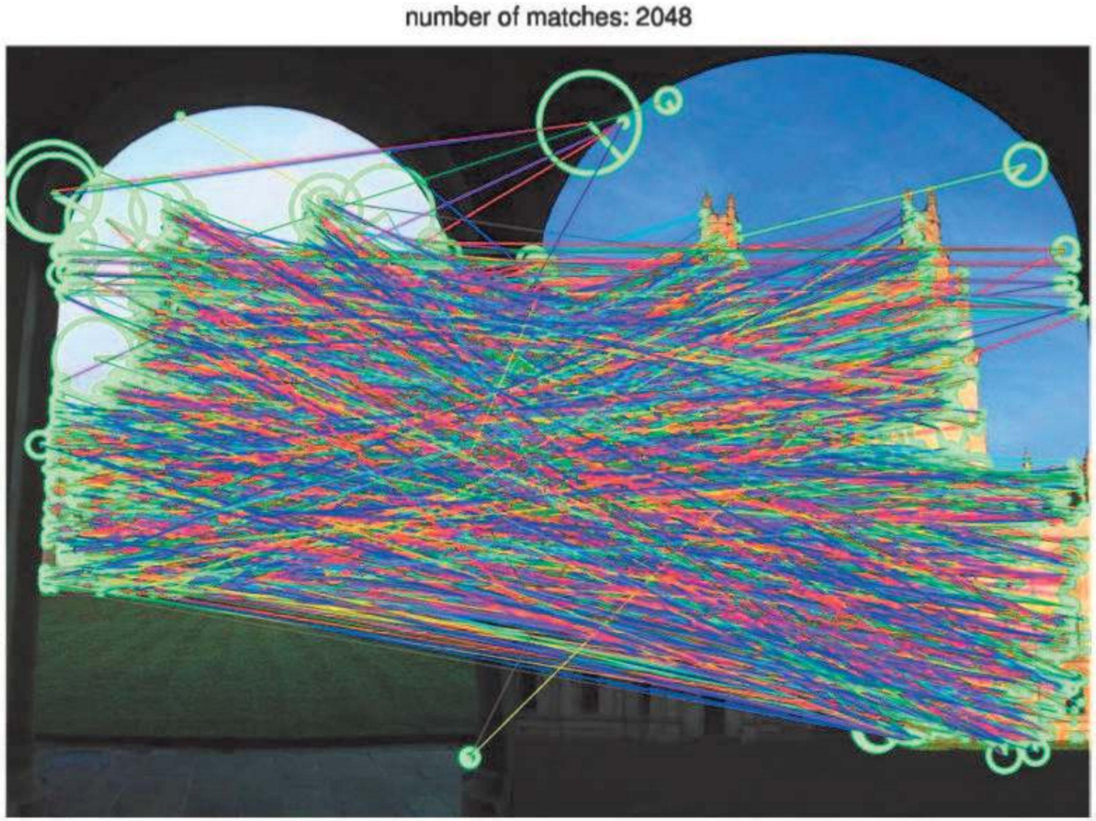
Exhaustive matching

Step 1: detect local features **f** and extract descriptors **d**

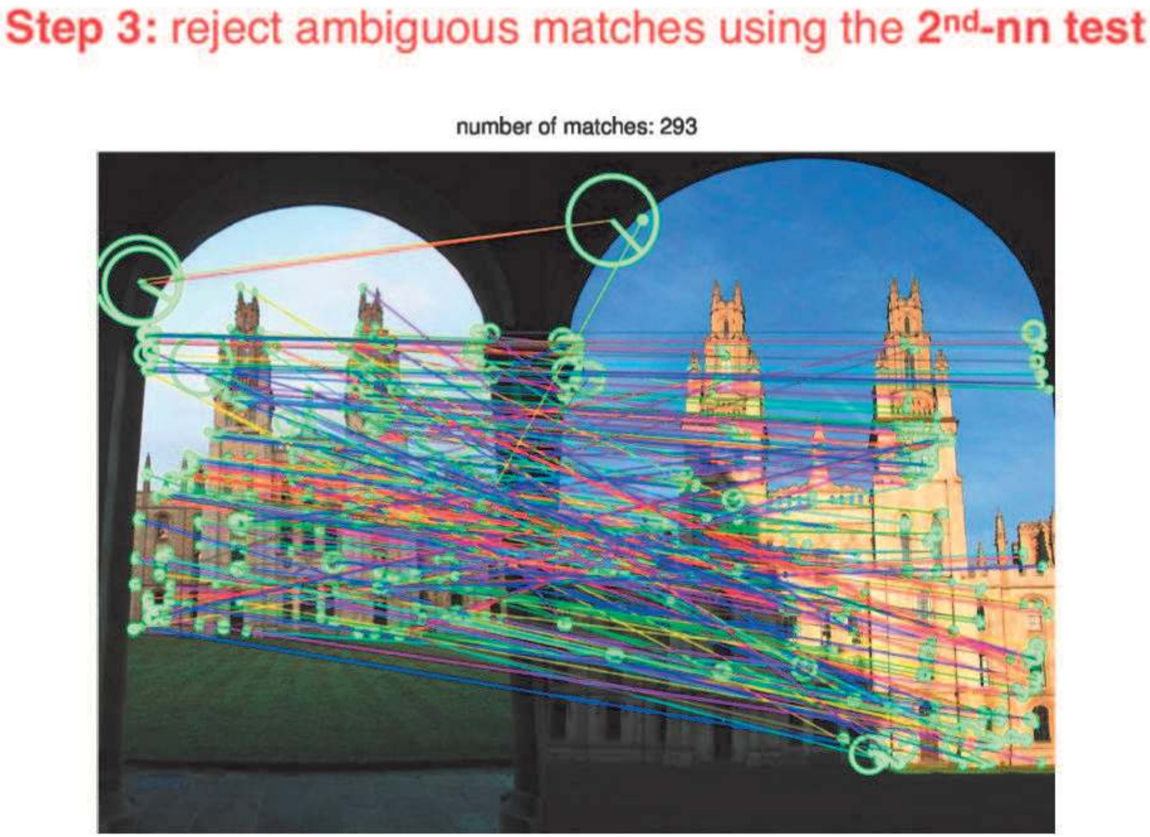


Local descriptors

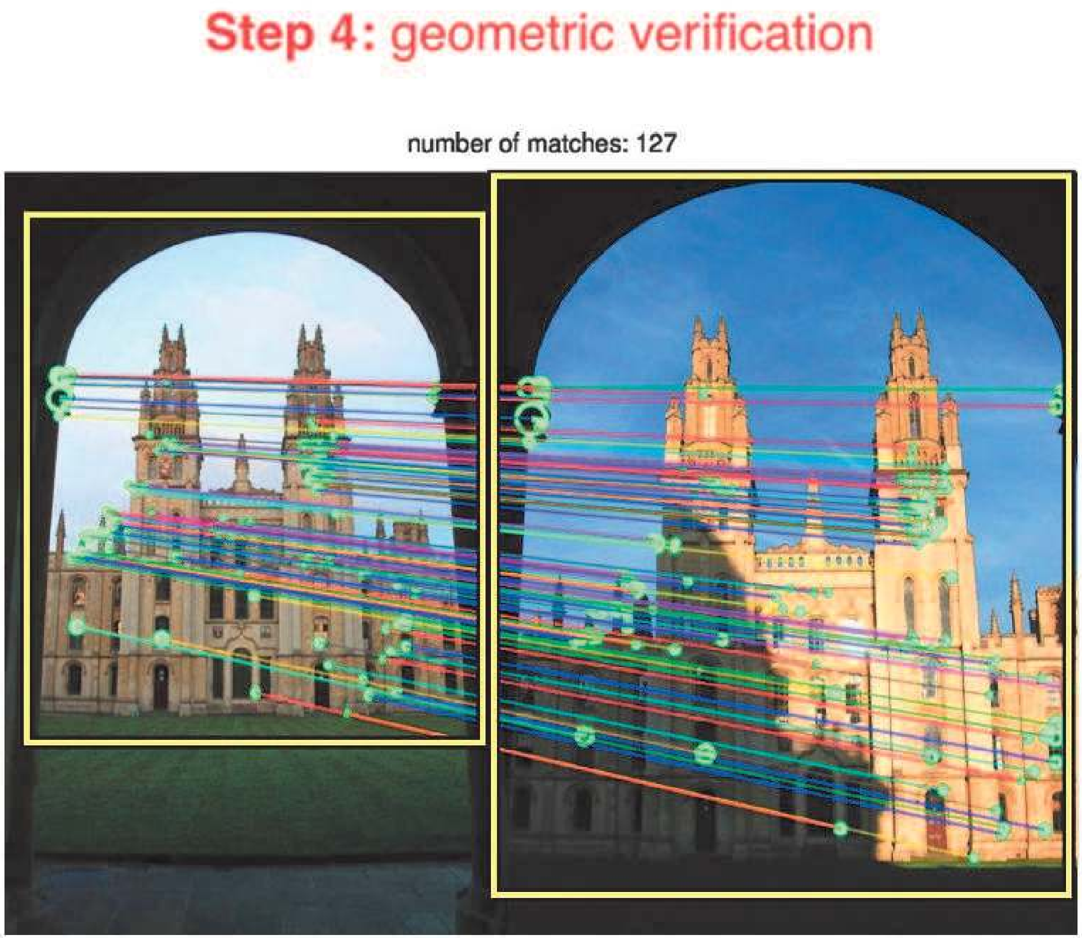
Step 2: match each descriptor to its closets one



Local descriptors



Local descriptors



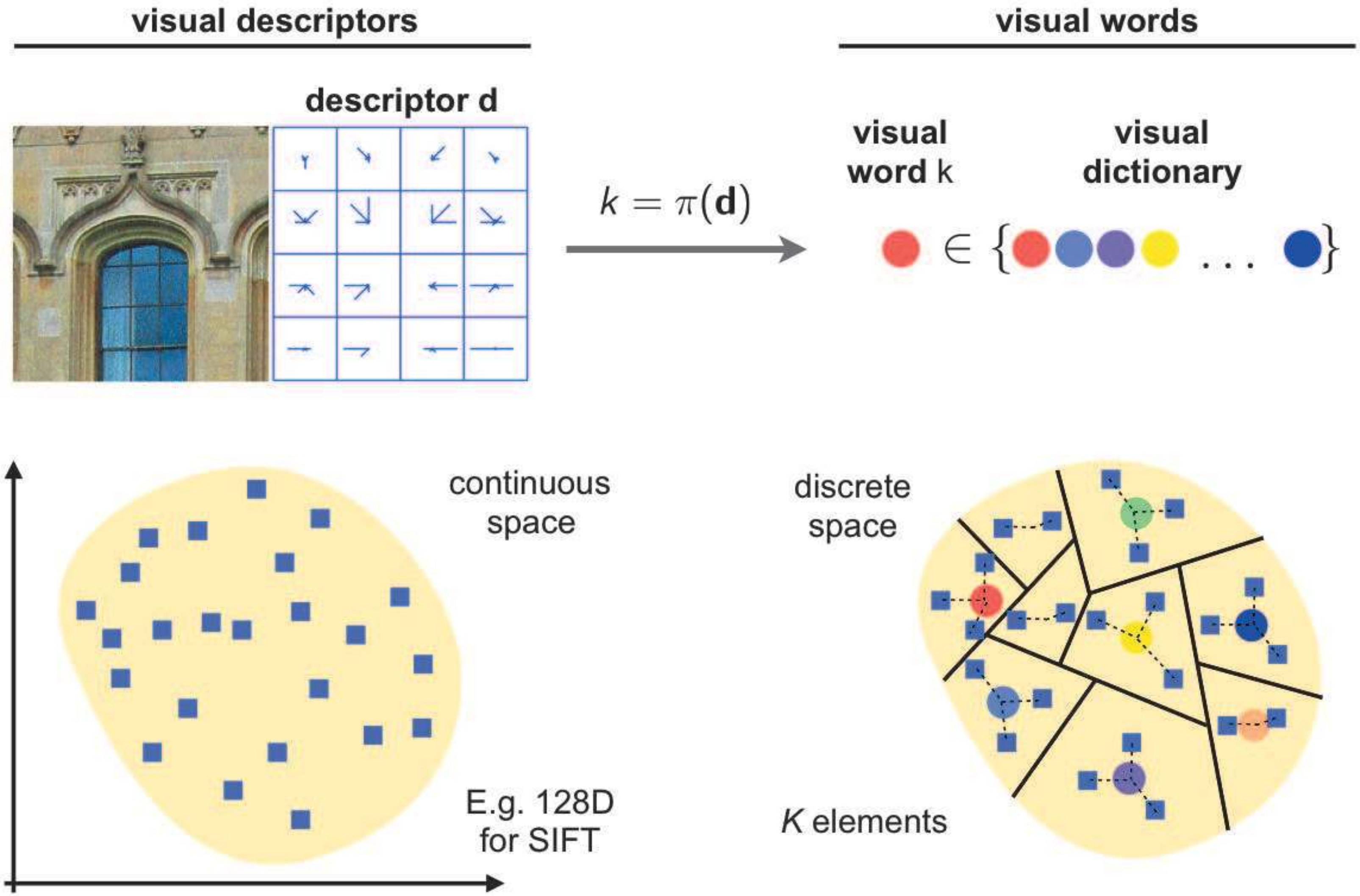
- the final step is to test whether matches are consistent with an overall image transformation
- inconsistent matches are rejected

From image matching to image search

- This matching strategy can be used to search a few images exhaustively
- However this is far too slow to search a large database
- Example:

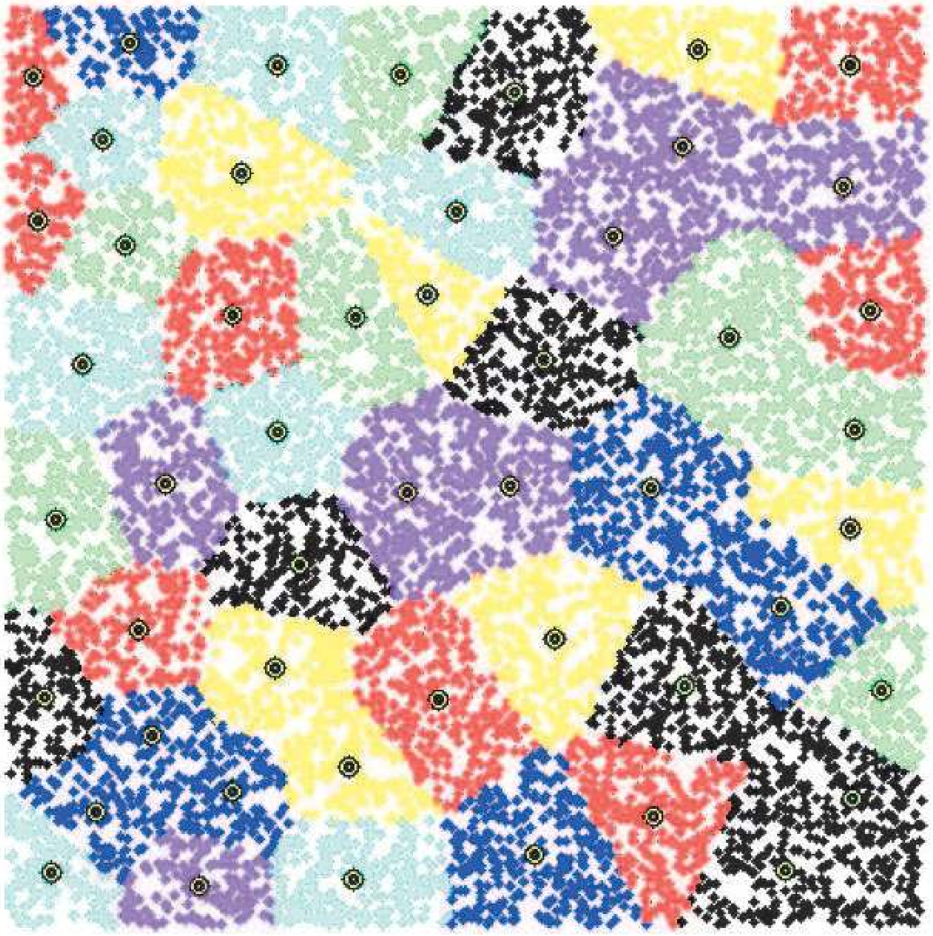
L images in the database	e.g. 10^6 - 10^{10} (FaceBook)
N features per image (incl. query)	e.g. 10^3 (~ SIFT detector)
D dimensional feature descriptor	e.g. 10^2 (~ SIFT descriptor)
Exhaustive search cost: $O(N^2 L D)$	10^{11} - 10^{15} ops = 100 days - 300 years
Memory footprint: $O(NLD)$	1TB - 1PB

Visual words



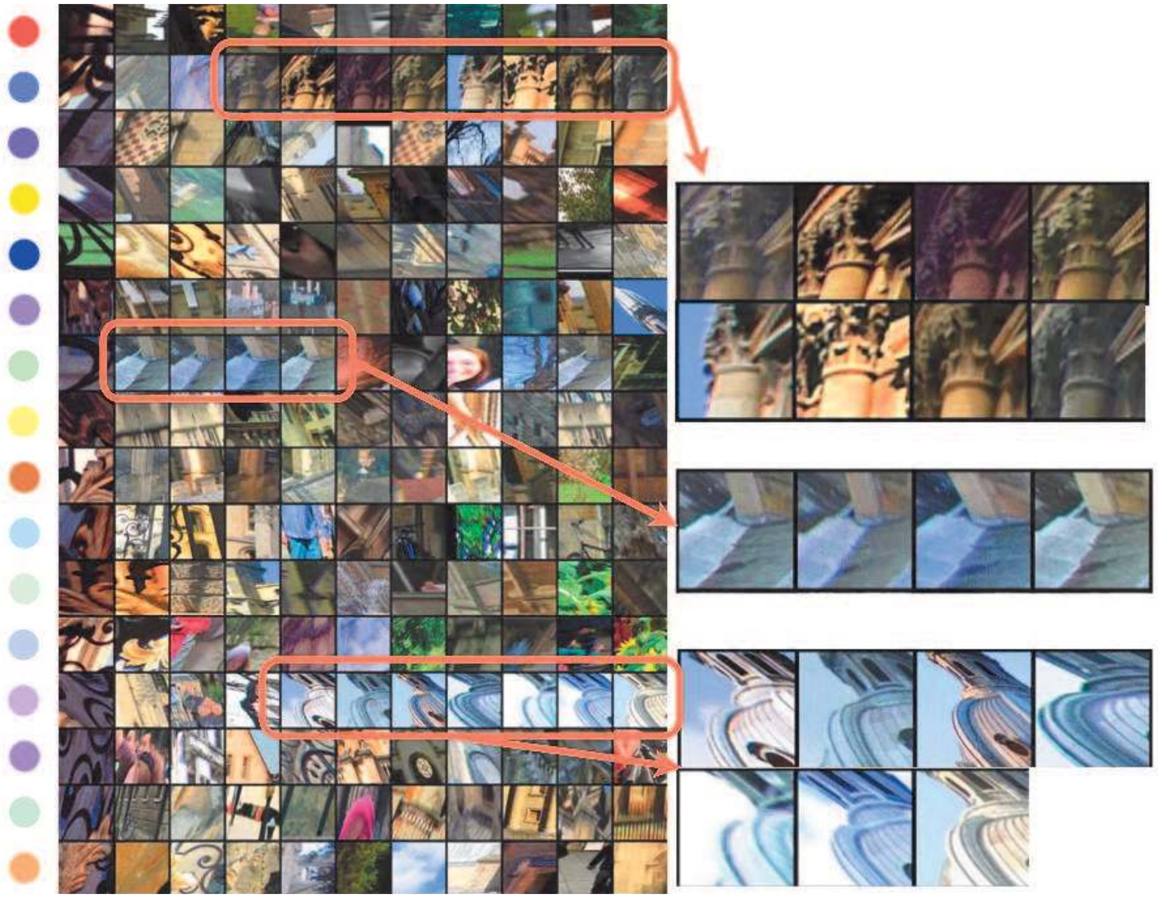
Visual words

- Dictionary is typically learned using k-means
- Value of k depends on the task: from 8 to $16M$



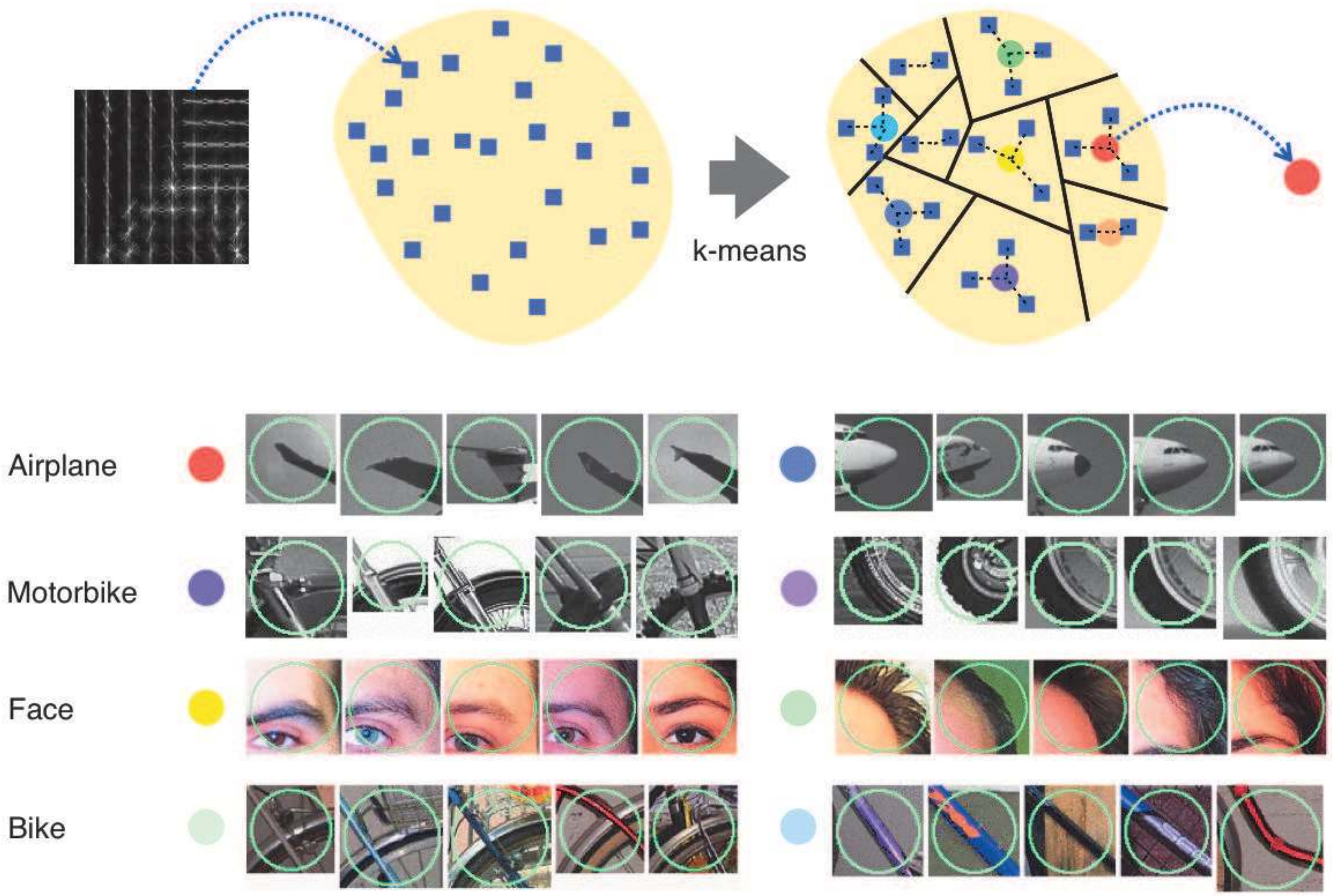
Visual words

- Visual word examples: each row is an equivalence class of patches mapped to the same cluster by k-means
- Visual words = iconic image fragments



Visual words

Quantization



Visual words

- Two steps:
  - Extraction: extract local features and compute corresponding descriptors
  - Quantization: map the descriptors to k-means cluster centroids to obtain the corresponding visual words

Histogram of visual words

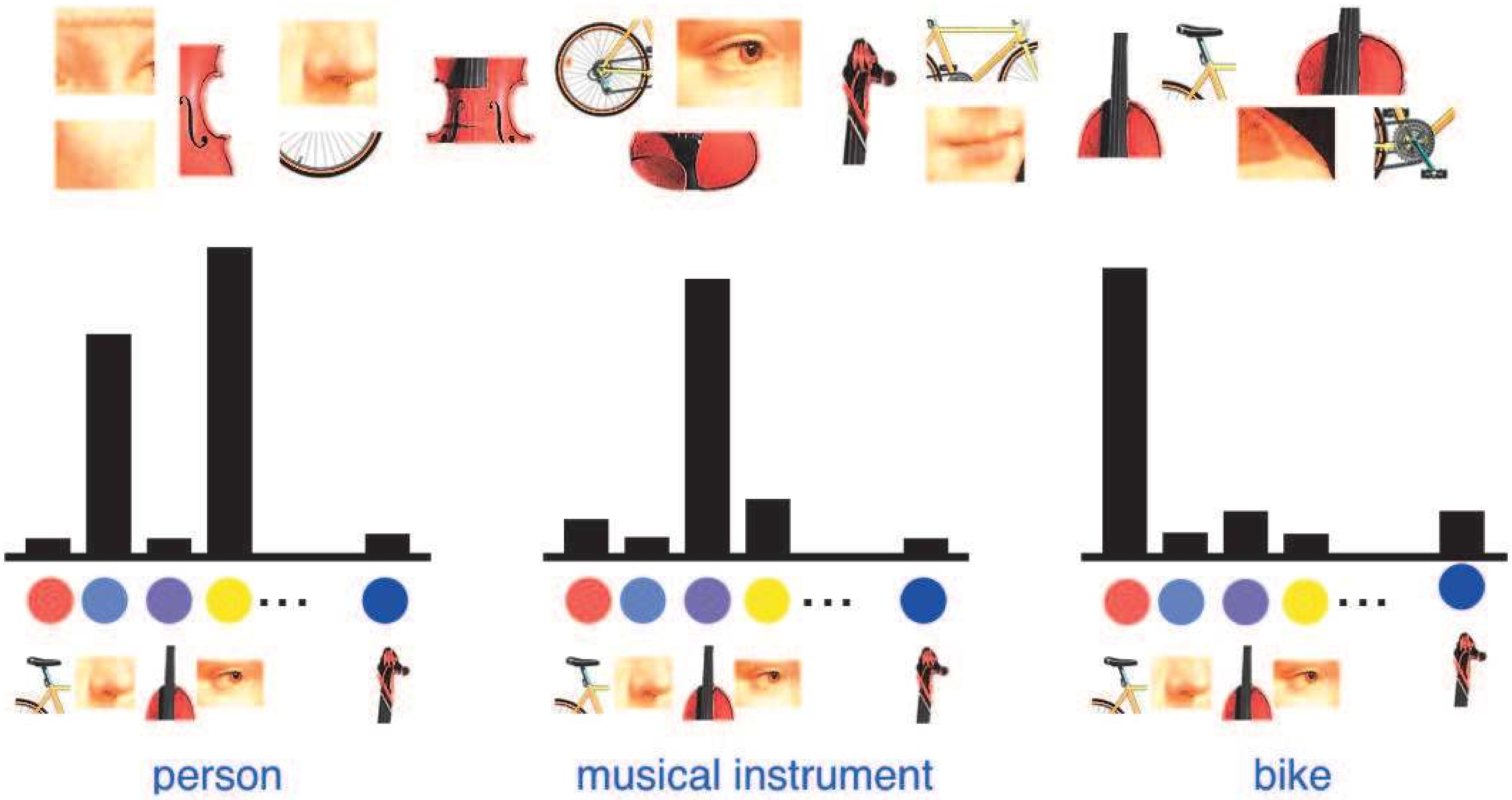
- A simple but efficient global image descriptor
- Vector of the number of occurrences of the k visual words in the image
- If there are k visual words, then $h \in \mathbb{R}^k$
- The vector h is a global image descriptor

Histogram of visual words

- This is also called a bag of (visual) words - BOW because it does not remember the relative positions of the features, just the number of occurrences
- h discards spatial information
- Pros: more invariant to viewpoint changes and other nuisance factors
- Cons: less discriminative

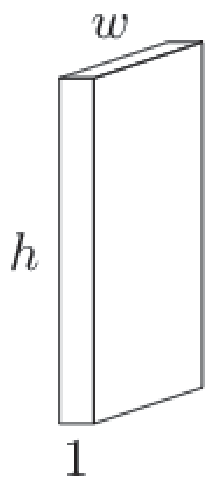
Histogram of visual words

Intuition



Global descriptors

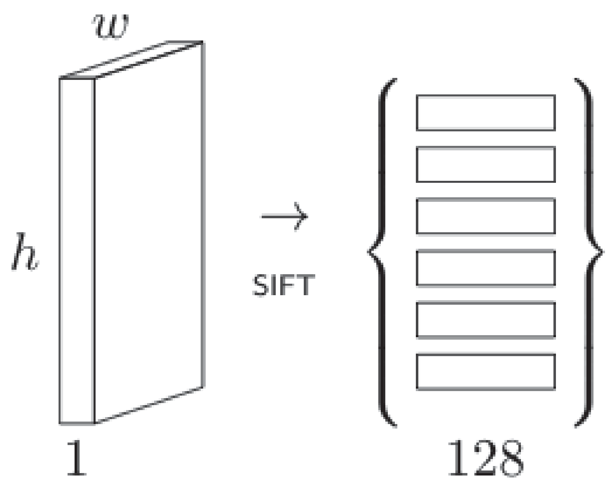
Bag-of-Words pipeline



- 3-channel patch RGB input $\rightarrow$ 1-channel gray-scale

Global descriptors

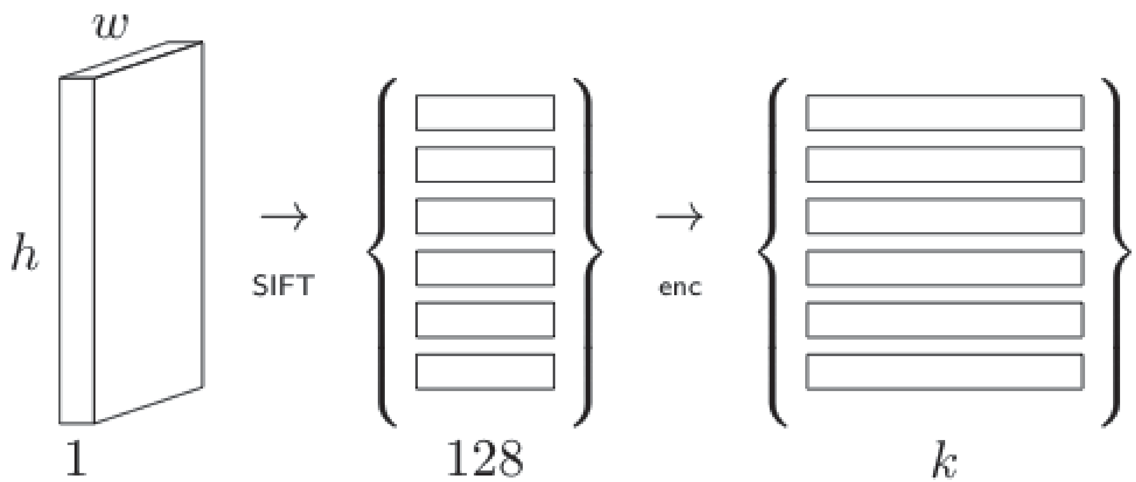
Bag-of-Words pipeline



- 3-channel patch RGB input $\rightarrow$ 1-channel gray-scale
- set of ~1000 features $\times$ 128-dim SIFT descriptors

Global descriptors

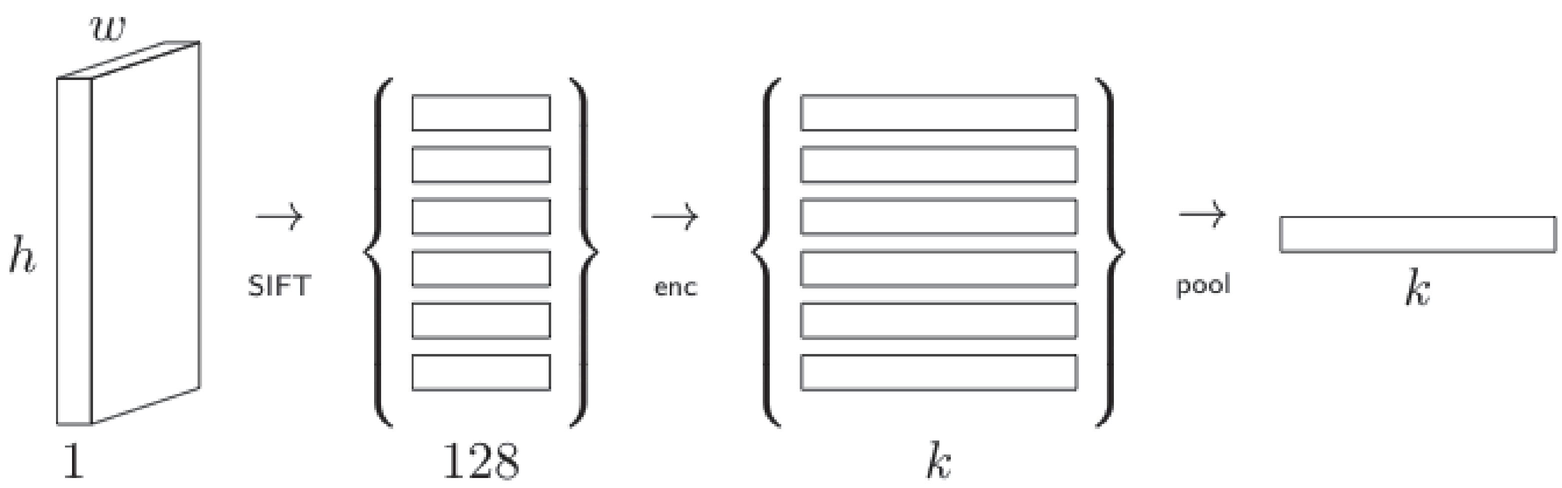
Bag-of-Words pipeline



- 3-channel patch RGB input $\rightarrow$ 1-channel gray-scale
- set of ~1000 features $\times$ 128-dim SIFT descriptors
- element-wise encoding of $k = 10^4$ visual words

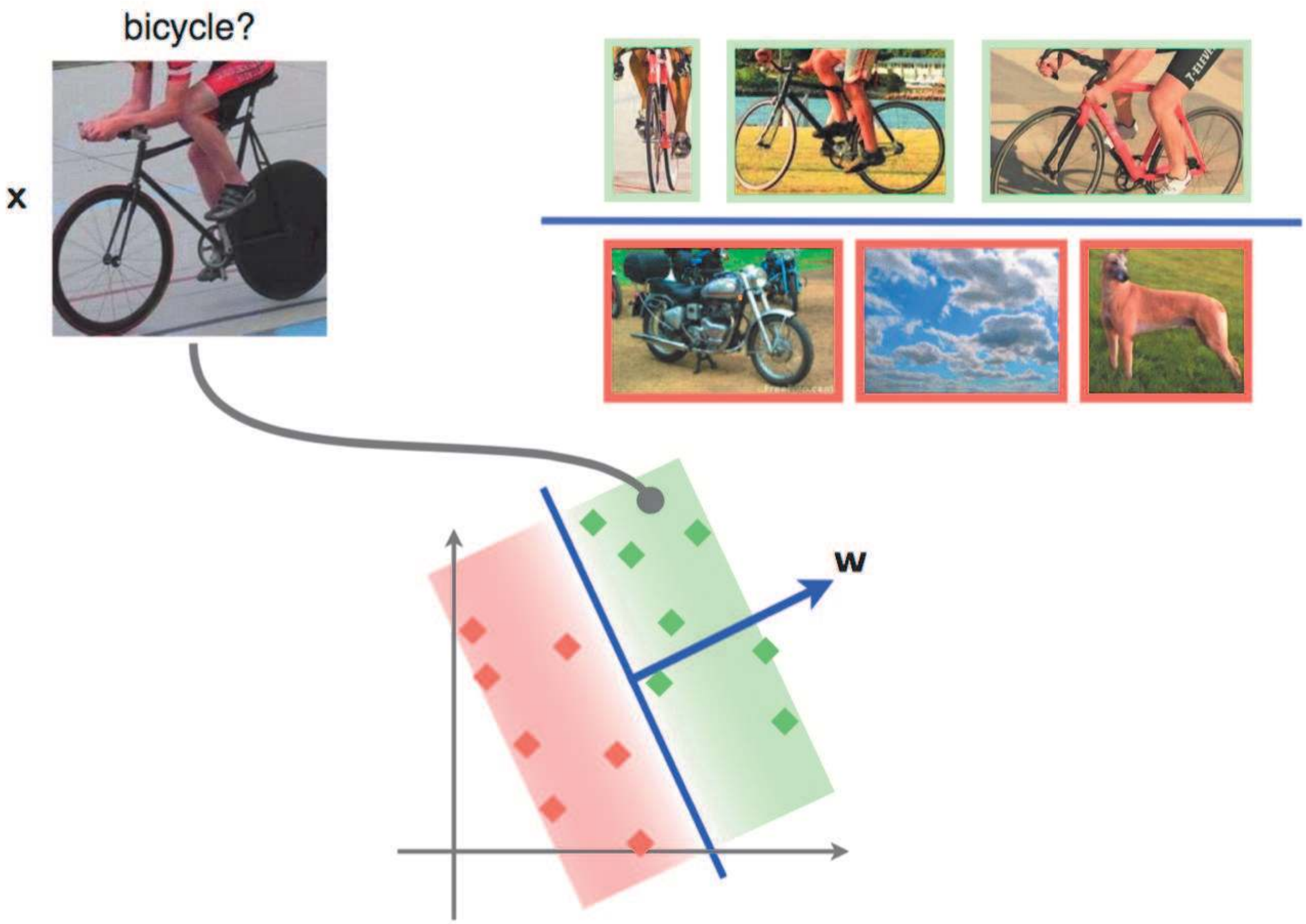
Global descriptors

Bag-of-Words pipeline



- 3-channel patch RGB input $\rightarrow$ 1-channel gray-scale
- set of ~1000 features $\times$ 128-dim SIFT descriptors
- element-wise encoding of $k = 10^4$ visual words
- global sum pooling, ℓ^2 normalization

Linear predictor

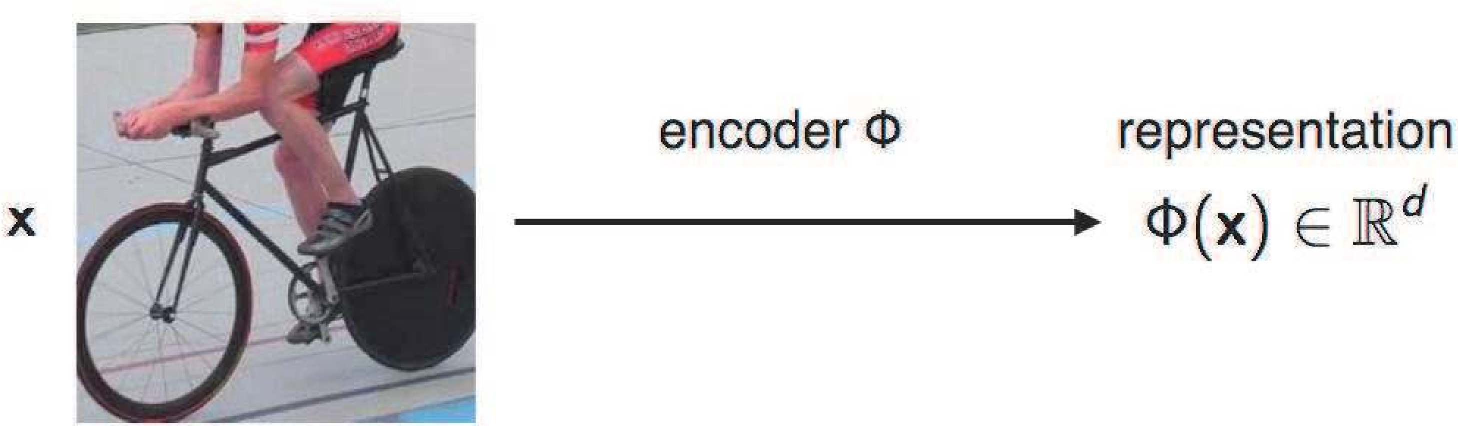


$$F(x) = \langle w, x \rangle$$

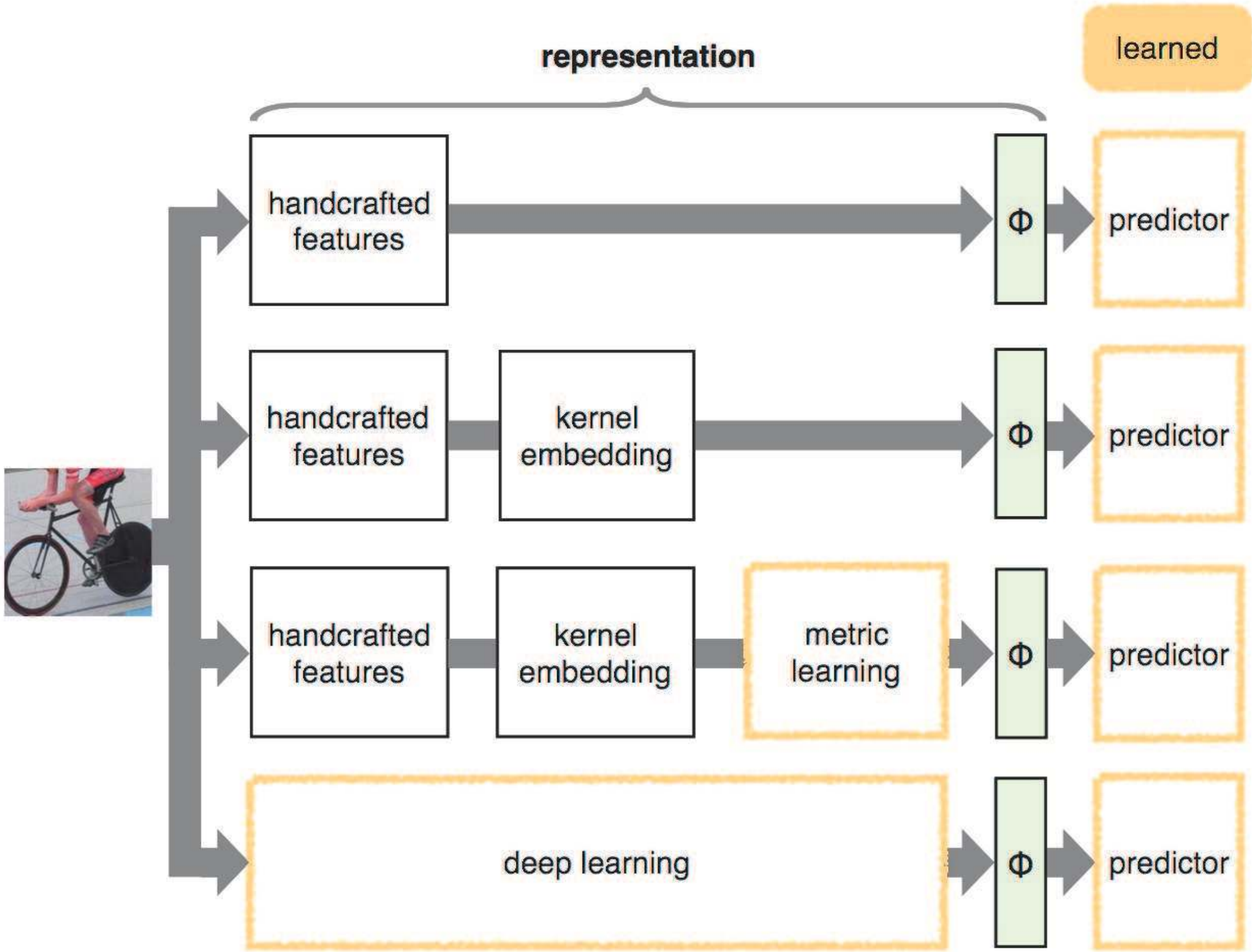
Data representations

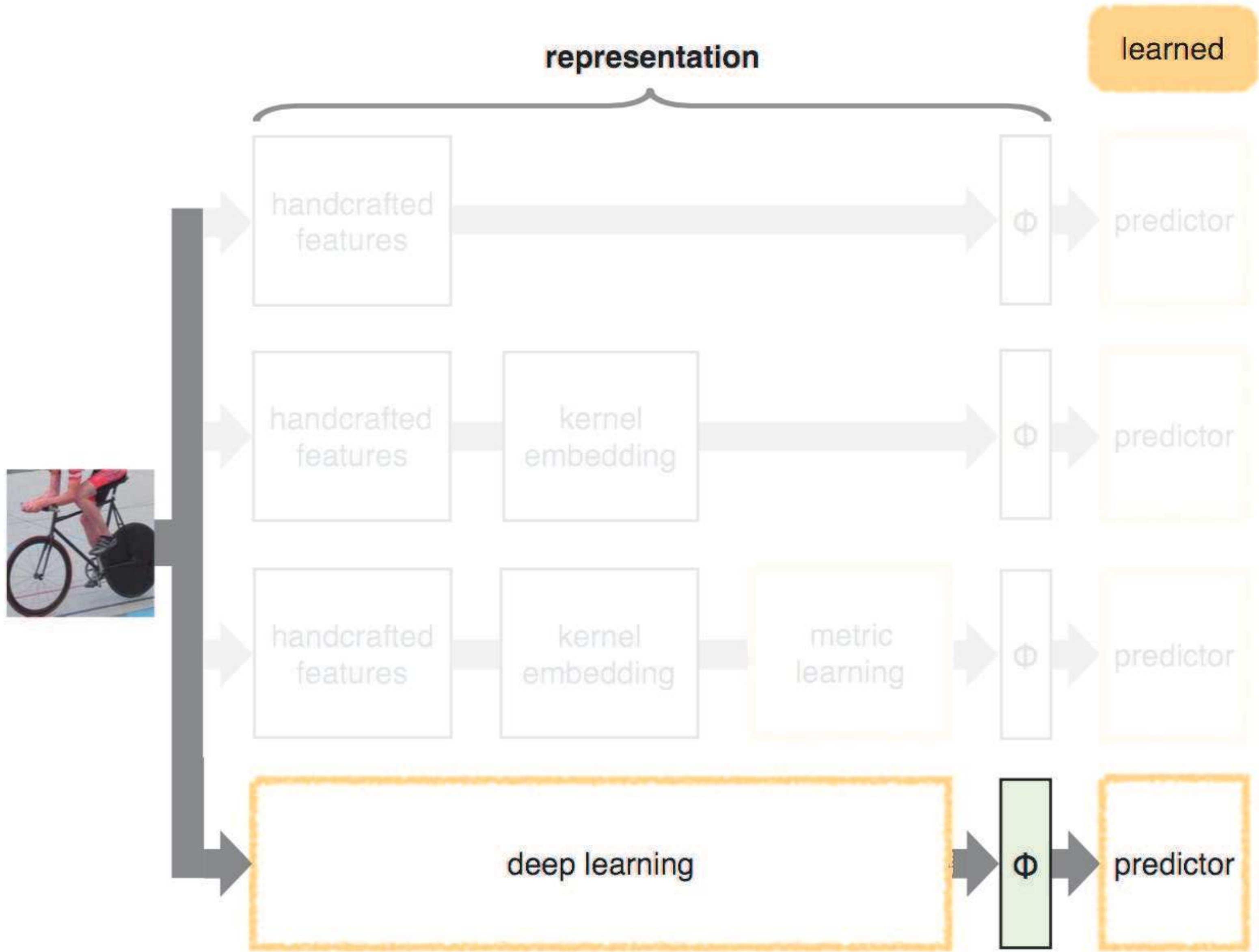
A linear predictor can be used to classify vector data. The question is how such a predictor can be applied to images, text, videos, or sounds.

This is solved by an encoder, which maps the data to a vectorial representation



$$F(x) = \langle w, \Phi(x) \rangle$$


representation
learned
handcrafted features
Φ
predictor
handcrafted features
kernel embedding
Φ
predictor
handcrafted features
kernel embedding
metric learning
Φ
predictor
deep learning
Φ
predictor


representation
learned
handcrafted features
Φ
predictor
handcrafted features
kernel embedding
Φ
predictor
handcrafted features
kernel embedding
metric learning
Φ
predictor
deep learning
Φ
predictor