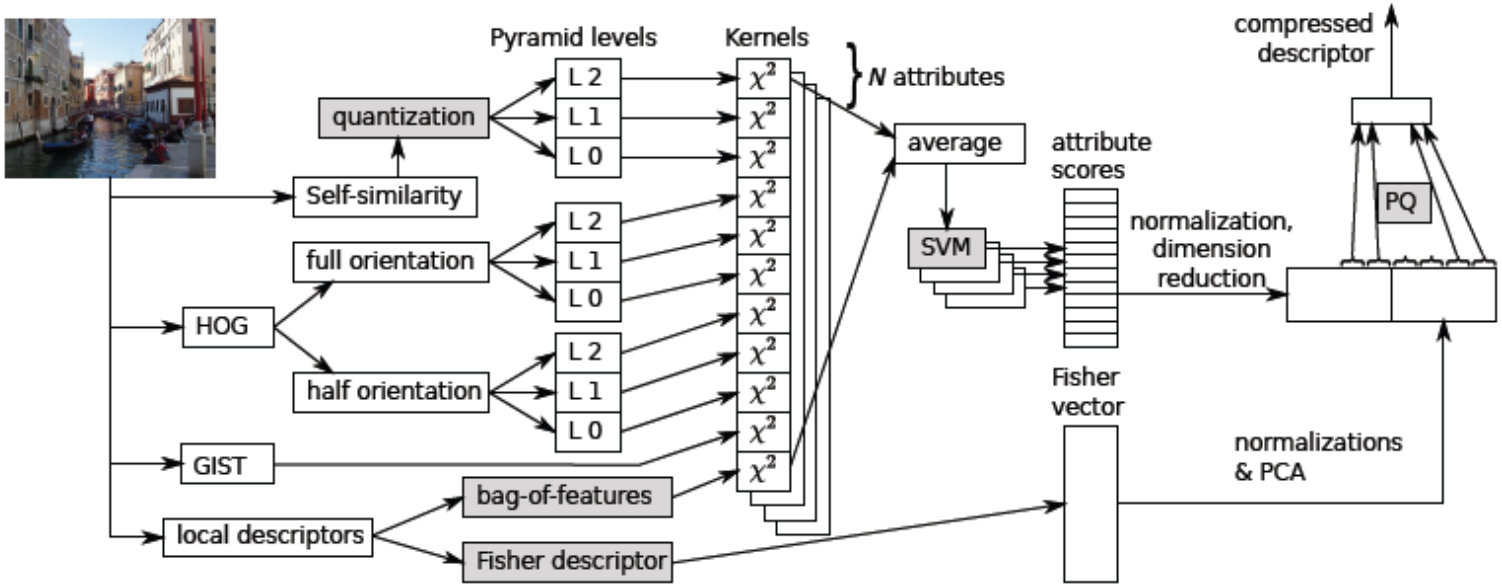
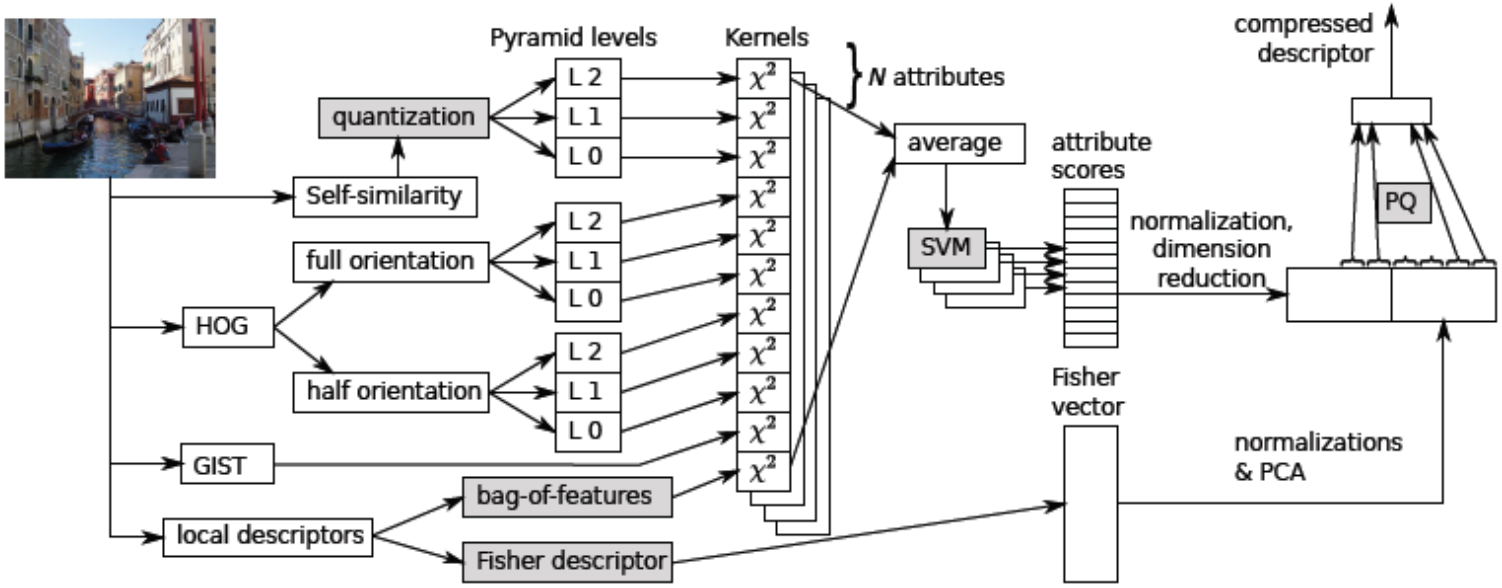
Matthijs Douze, Arnau Ramisa, Cordelia Schmid

Combining Attributes and Fisher Vectors for Efficient **Image Retrieval**

"Norah Jones" query





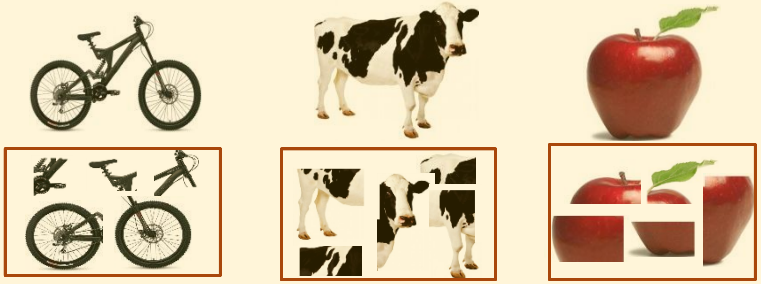
Retrieving **images** of objects from **large datasets**

- Recently received increasing attention
  - Most approaches build on the bag-of features**(BOF)**

Retrieving **images** of objects from **large datasets**

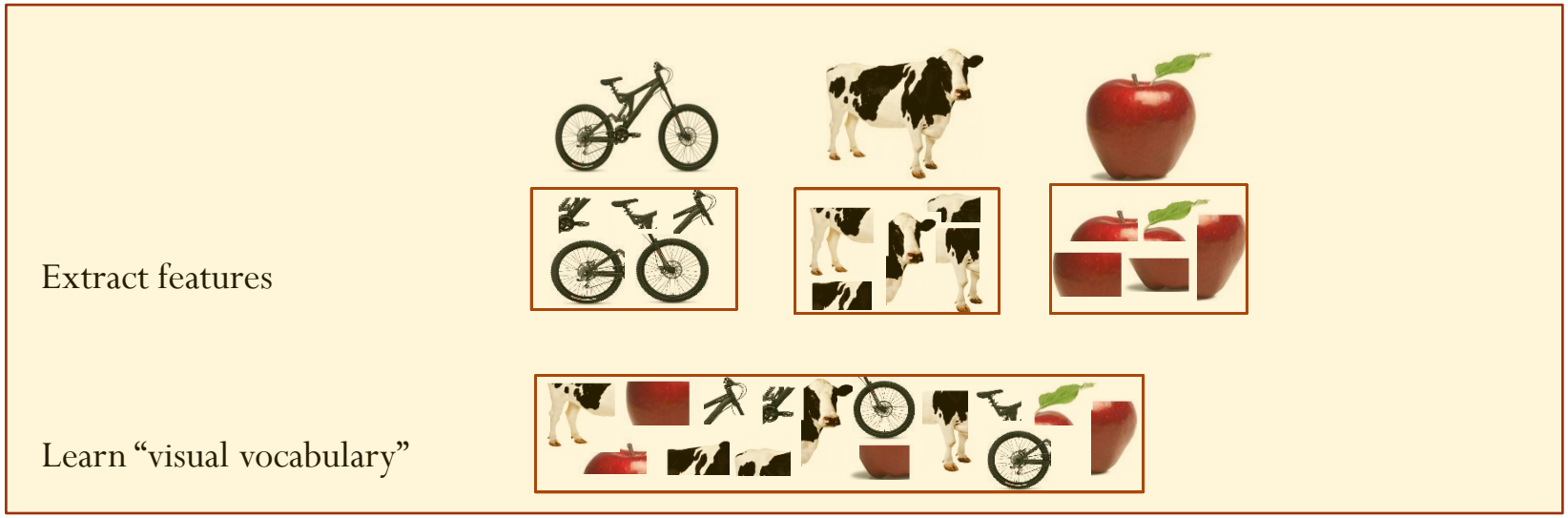
- Recently received increasing attention
  - Most approaches build on the bag-of features**(BOF)**

Extract features



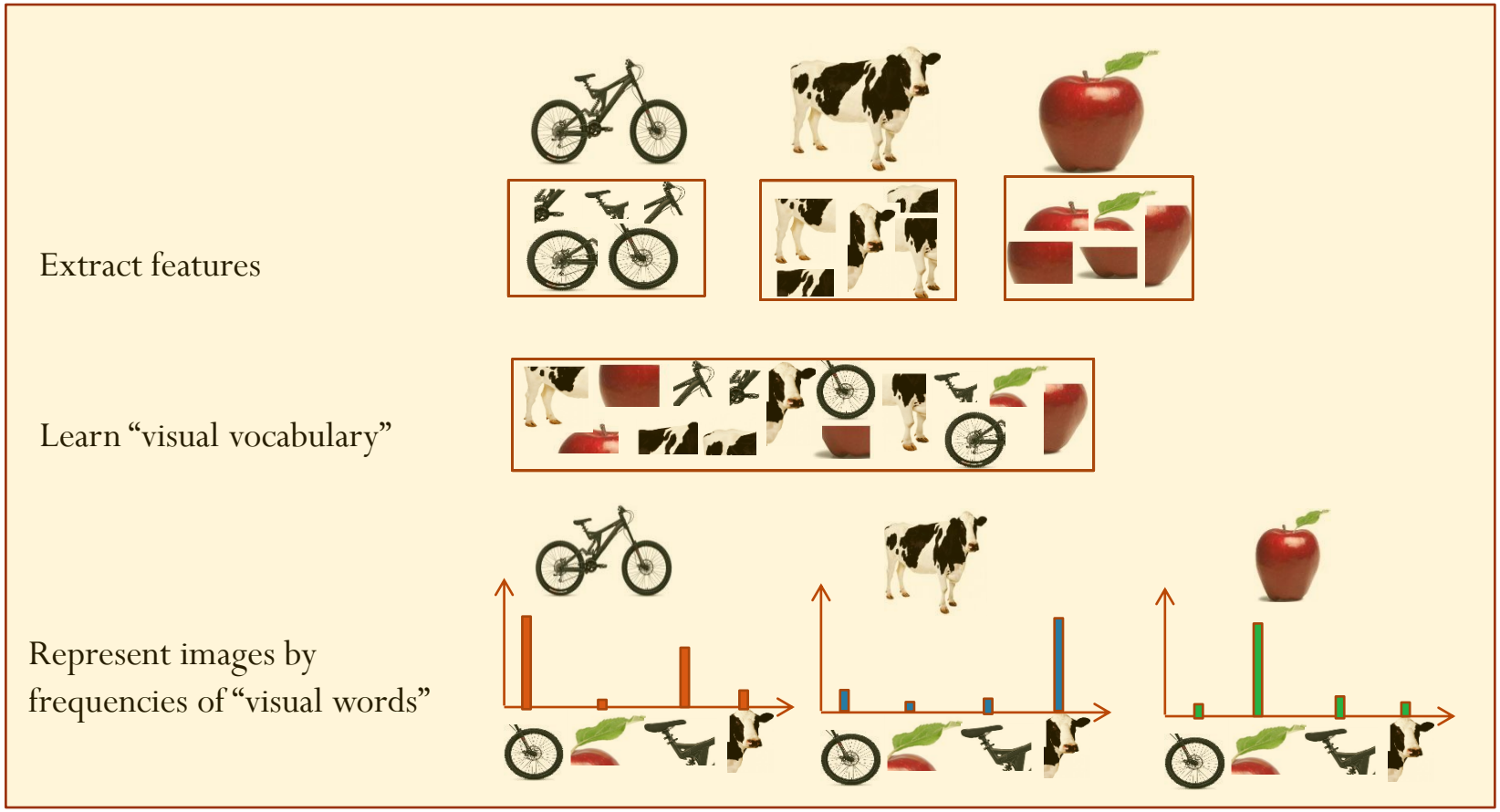
Retrieving **images** of objects from **large datasets**

- Recently received increasing attention
  - Most approaches build on the bag-of features**(BOF)**



Retrieving **images** of objects from **large datasets**

- Recently received increasing attention
  - Most approaches build on the bag-of features**(BOF)**



Inverted File Indexing

- Simple and effective
- Faster than searching every image.

Recent extensions

- Speed up the assignment of individual descriptors to visual words
  - **[Nister et al., CVPR'06]** and **[Philbin et al. , CVPR'07.]**

- Improve the accuracy by complementing the visual word index for a given descriptor
  - a binary vector **[Jegou et al., ECCV'08]**
  - by learning descriptor projections [**Philbin et al., ECCV'10]**

Recent extensions

- Speed up the assignment of individual descriptors to visual words
  - **[Nister et al., CVPR'06]** and **[Philbin et al. , CVPR'07.]**

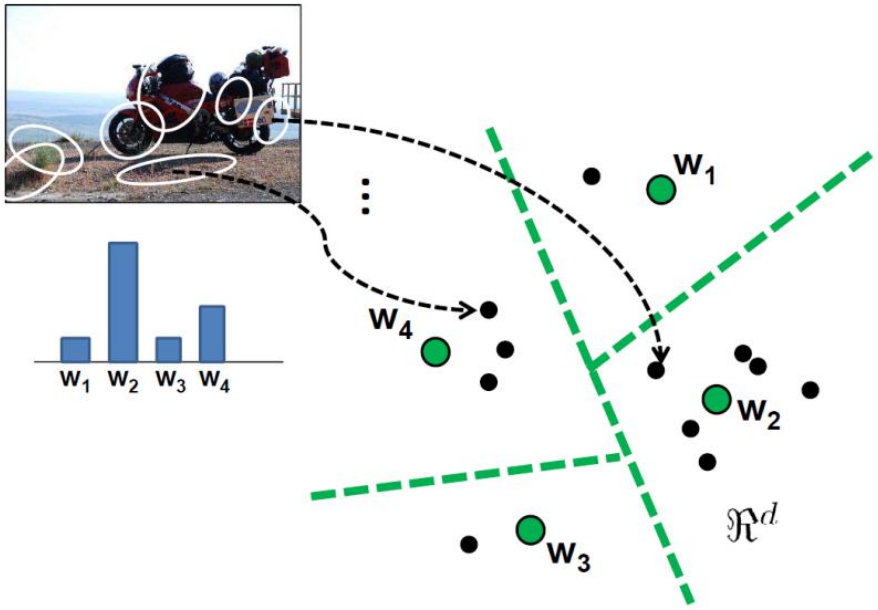
- Improve the accuracy by complementing the visual word index for a given descriptor
  - a binary vector **[Jegou et al., ECCV'08]**
  - by learning descriptor projections **[Philbin et al., ECCV'10]**

Store one index per local images descriptor

prohibitive for very large datasets.

Large-scale visual recognition

BOV answer to the problem : increase visual vocabulary size



How to increase amount of information without increasing the visual vocabulary size?

http://www.cs.utexas.edu/~grauman/courses/fall2009/papers/bag_of_visual_words.pdf

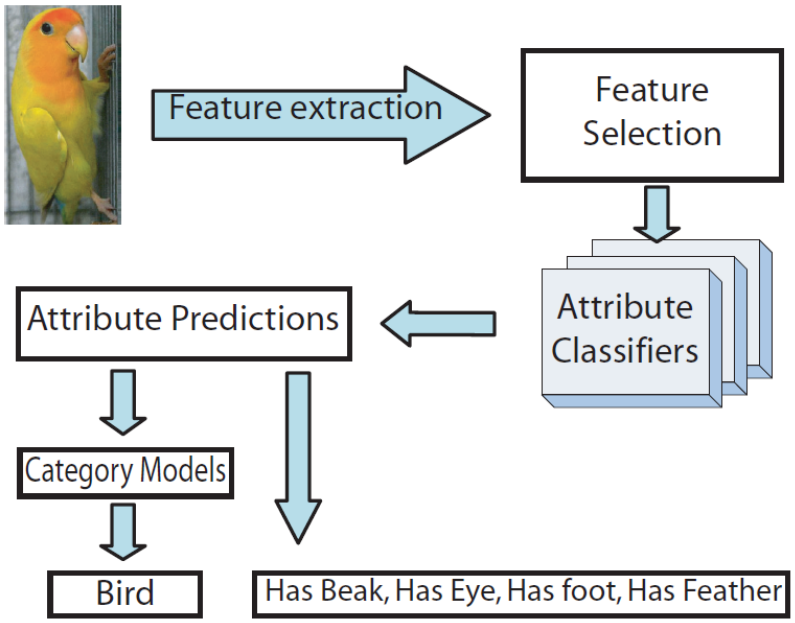
Low dimensional representations

- Obtain more compact representations ;
  - Compress to obtain very compact codes.
    - Bag-of-features with a small number of visual words **[Jegou et al., CVPR'10]**
    - GIST descriptors **[Oliva et al., IJCV'01]**
  - Some approaches compress GIST descriptors by converting them to compact binary vectors.
    - **[Douze et al., CIVR'2009], [Torralba et al., CVPR'2008], [Weiss et al., NIPS'2008]**
  - In some : local descriptors are aggregated into low-dimensional vectors and compressed to small codes.
    - **[Jegou et al., CIVR'2009], [Jegou et al., CVPR'2010]**

- **[Perronnin et al., CVPR'2010] :**
  - Image description based on **Fisher vectors**
  - Outperform the bag-of-features representation for the same dimensionality.

A different representations

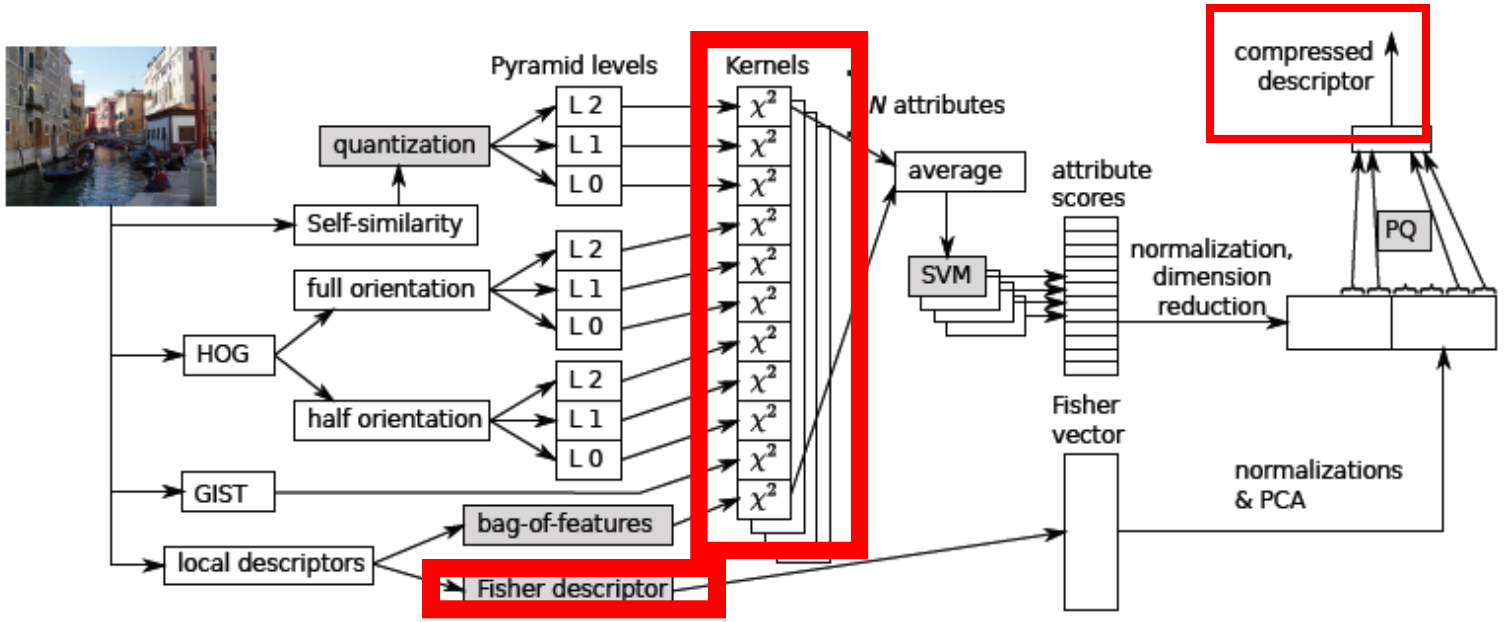
- Torresani et al. [1] learn a set of classifiers and use the scores of these classifiers to obtain a low dimensional description of the image.
- The classifiers are trained on an independent dataset obtained automatically from the Bing image search engine for categories from the LSCOM ontology [2].



[1] L. Torresani, M. Summer, and A. Fitzgibbon. Efficient object category recognition using classemes. In ECCV, 2010.
[2]M. Naphade, J. Smith, J. Tesic, S. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis. Large-scale concept ontology for multimedia. IEEE Multimedia, 13(3):86– 91, 2006.

Authors Do

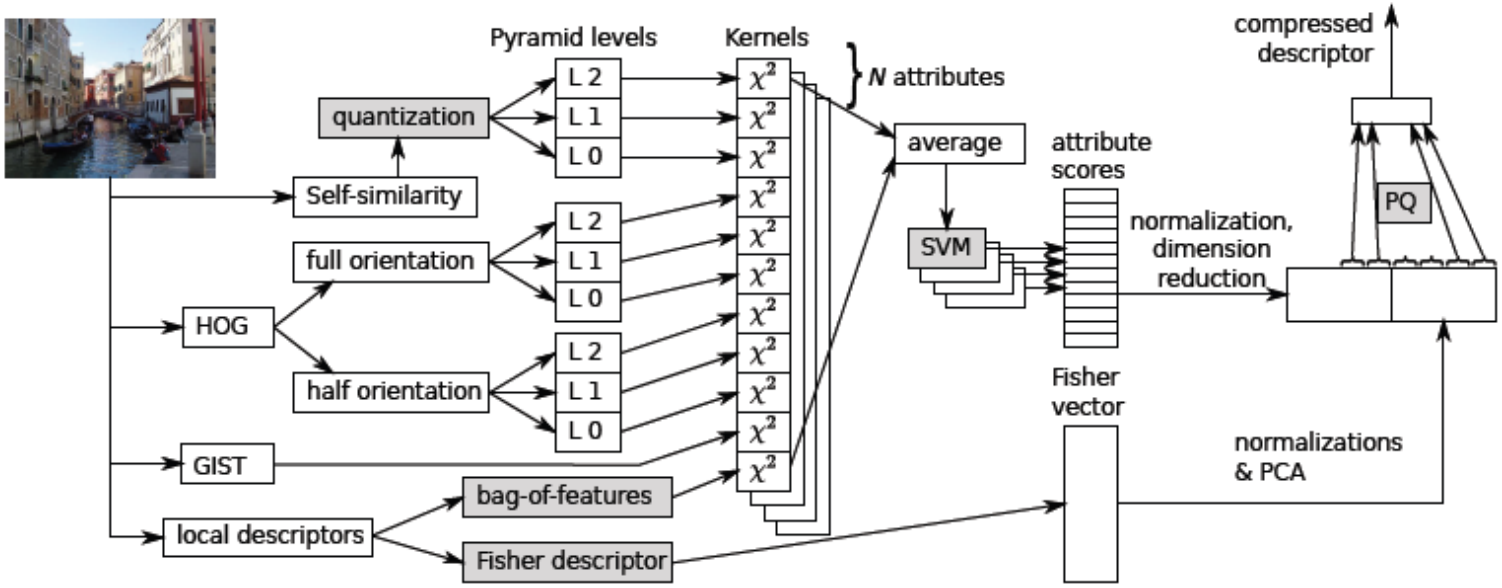
- Demonstrate that attributes of [1] give excellent results for retrieval
- Combination of attribute features with Fisher vector[2] significantly outperforms state of the art
- Implement an efficient technique for compressing descriptor, based on dimensionality reduction and product quantization [7] .



[1] L. Torresani, M. Summer, and A. Fitzgibbon. Efficient object category recognition using classemes. In ECCV, 2010.

[2] F. Perronnin, Y. Liu, J. Sanchez, and H. Poirier. Large-scale image retrieval with compressed Fisher vectors. In CVPR,2010.

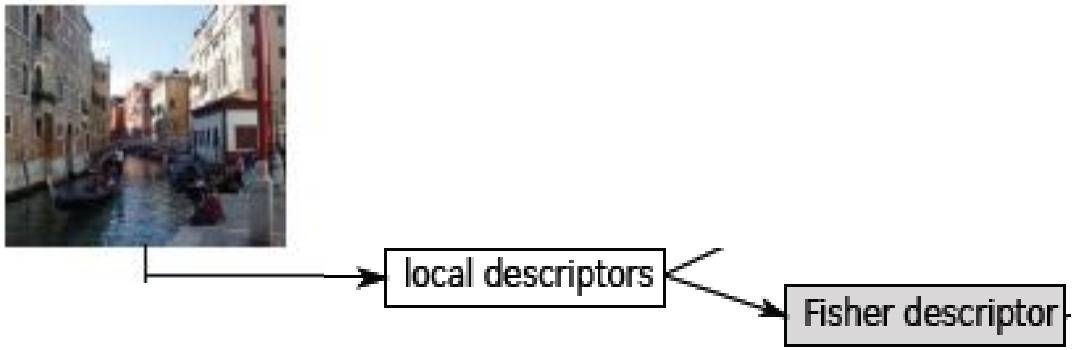
[3] H. J´egou, M. Douze, and C. Schmid. Product quantization for nearest neighbor search. IEEE Transactions on Pattern Analysis & Machine Intelligence, 33(1):117–128, jan 2011.



Three image descriptors used in this work

2.1. Fisher vector

- Fisher vectors [16] are a means of aggregating local descriptors into a global descriptor.



- Fisher descriptors outperform BOF as a global descriptor for image classification **[Perronnin et al., CVPR'2006]** and retrieval **[Perronnin et al., CVPR'2010]**

2.1. Fisher vector

- Let p be a likelihood function u_λ with parameters λ. The score function of given samples $X = \{x_t, t = 1 \dots T\}$ with the following gradient vector:

$$\boldsymbol{G_\lambda^X = \nabla_\lambda log u_\lambda X}$$

- *The gradient of the log-likelihood describes the direction in which parameters should be modified to best fit the data.*

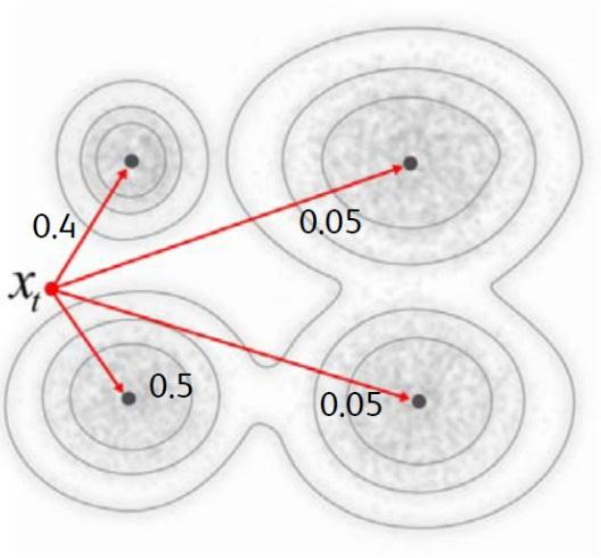
2.1. Fisher vector

The Fisher vector Application to images

- $X = \{x_t, t = 1 \dots T\}$ is the set of T i.i.d. D-dim local descriptors (e.g. SIFT) extracted from an image: average pooling is a direct consequence of independence assumption

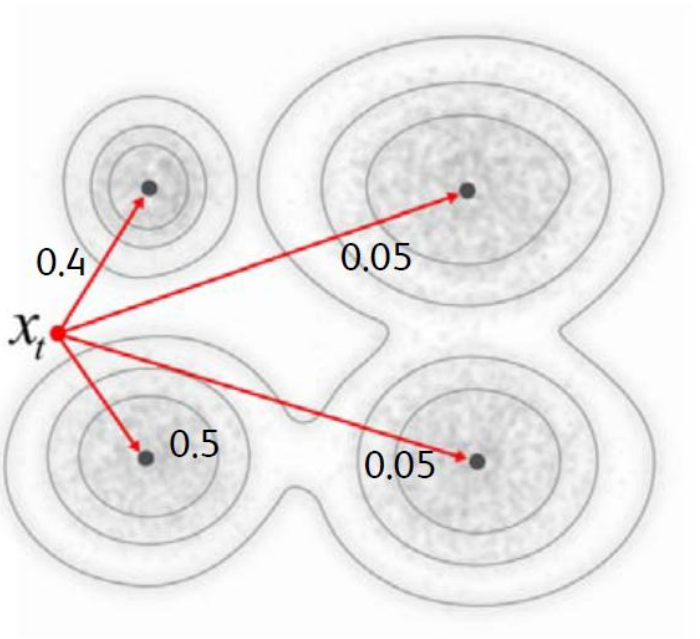
$$G_\lambda^X = \frac{1}{T} \sum_{t=1}^{T} \nabla_\lambda \, log u_\lambda(x_t)$$

- $u_\lambda(x) = \sum_{i=1}^{K} w_i u_i(x)$ is a Gaussian Mixture Model (GMM) with parameters $\lambda = \{w_i, u_i, \sum_i, i = 1 \dots N\}$ trained on a large set of local descriptors → **a probabilistic visual vocabulary**



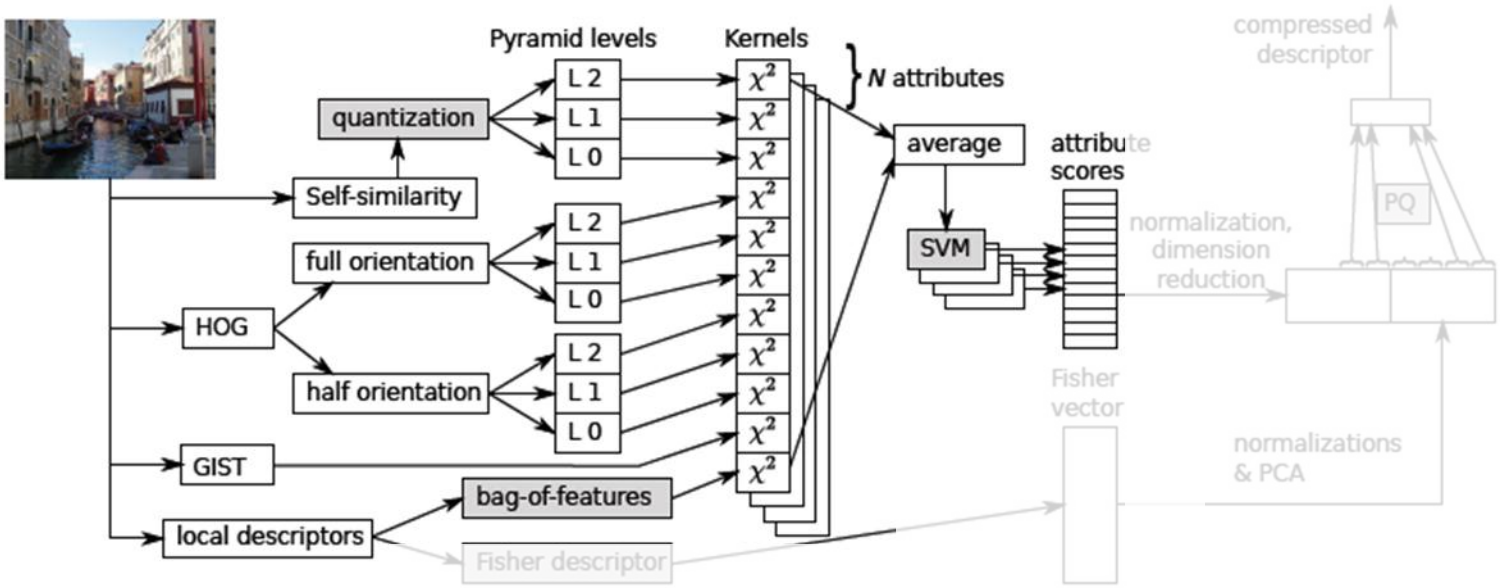
2.1. Fisher vector

- A **64-centroid Gaussian mixture model (GMM)** .
  - The Fisher descriptor is the derivative of this likelihood with respect to the GMM parameters.
  - descriptor has 64X64 = 4096 dimensions.

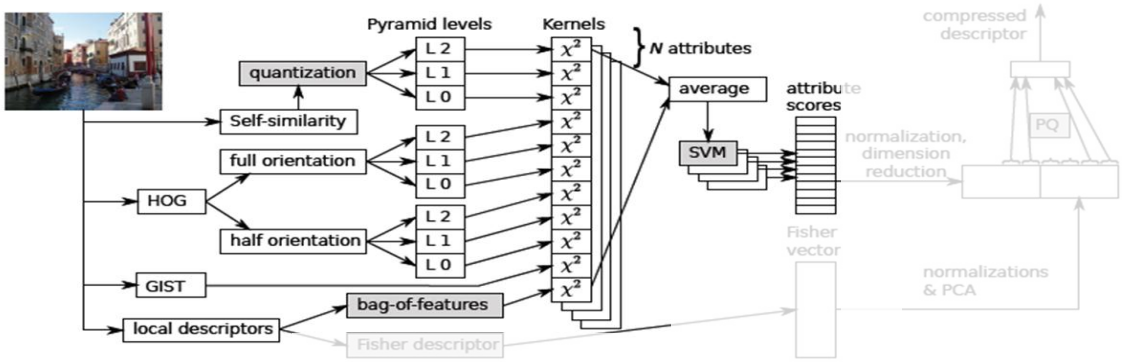


2.2. Attribute features

- Each attribute corresponds to a term from a vocabulary.
- Attribute descriptor encodes how relevant each term that image.
- Attribute descriptors are computed from image classifiers built for each of the terms



2.2. Attribute features



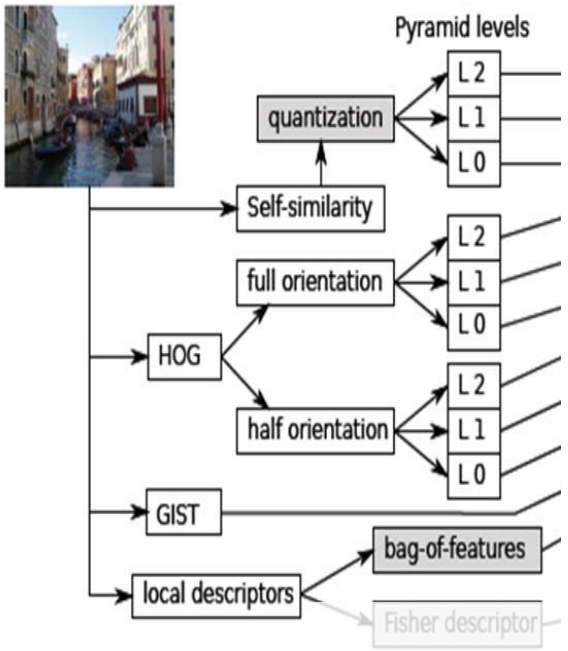
- **The vocabulary and learning set.**
  - Vocabulary from Torresani et al. [1], which contains C = 2659 attributes
  - **names for object classes** ("warplane", "logo", "hu jintao"),
  - **terms that are less related to visual representations** ("democratic national conventions", "group of tangible things", "indoors isolated from outside")
  - a **few abstract concepts** ("attempting", "elevated", "temporal thing").
  - Top 150 images returned by the bing.com image search engine

[1] L. Torresani, M. Summer, and A. Fitzgibbon. Efficient object category recognition using classemes. In ECCV, 2010.

2.2. Attribute features

Low-level image features.

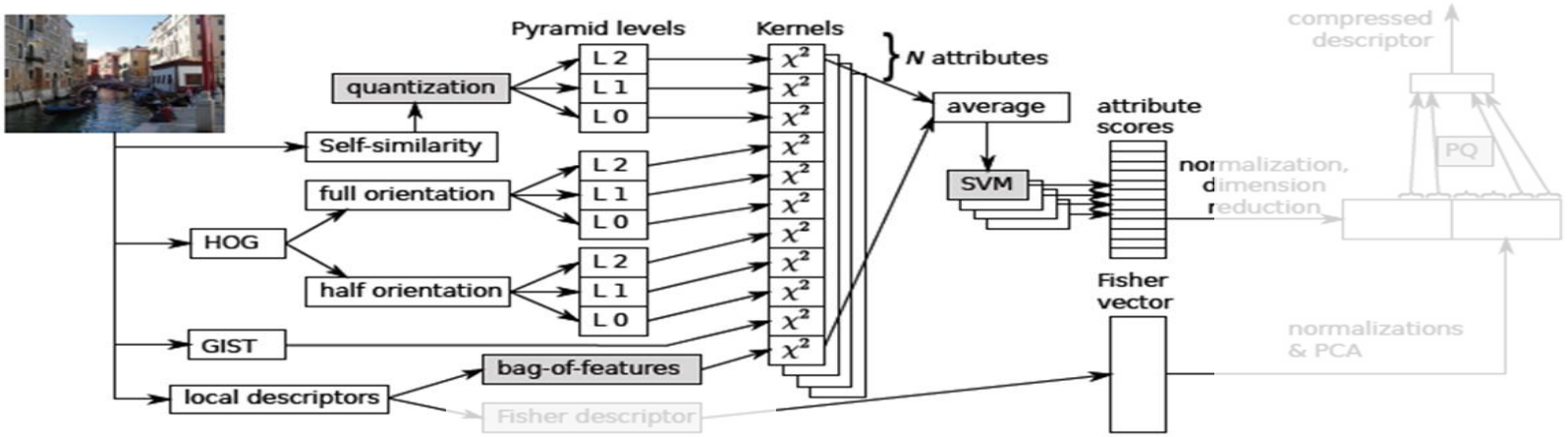
- Color GIST descriptor.
- Pyramid of self-similarity descriptors [21].
- Pyramid of histograms of oriented gradients (PHOG) .
  - Half orientation:
  - Full orientation :
- A bag-of-features with vocabulary of dimensionality 4000.



[1] L. Torresani, M. Summer, and A. Fitzgibbon. Efficient object category recognition using classemes. In ECCV, 2010.

2.2. Attribute features

Image classifiers.

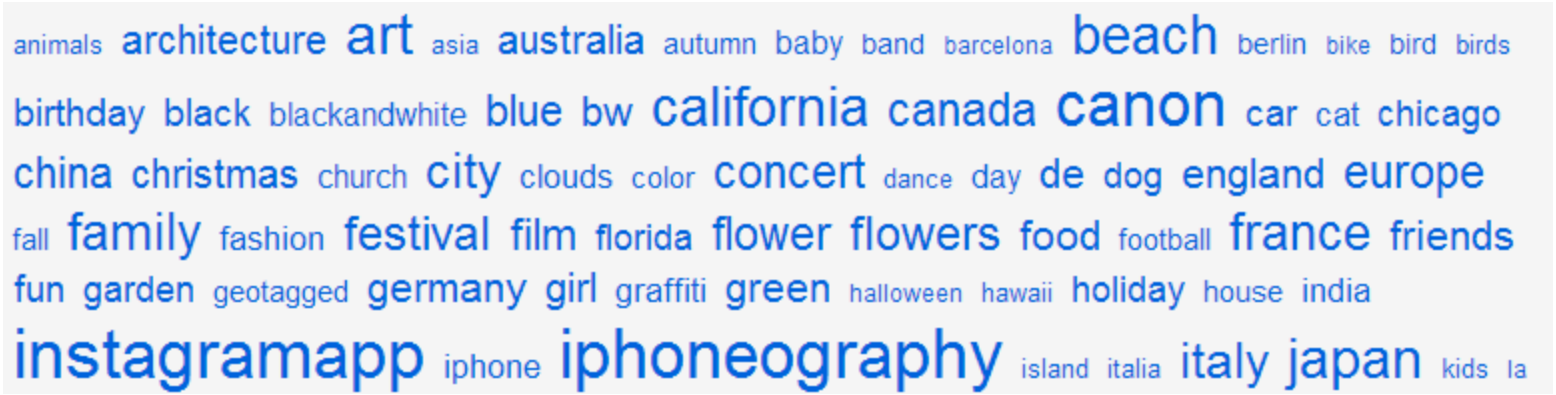


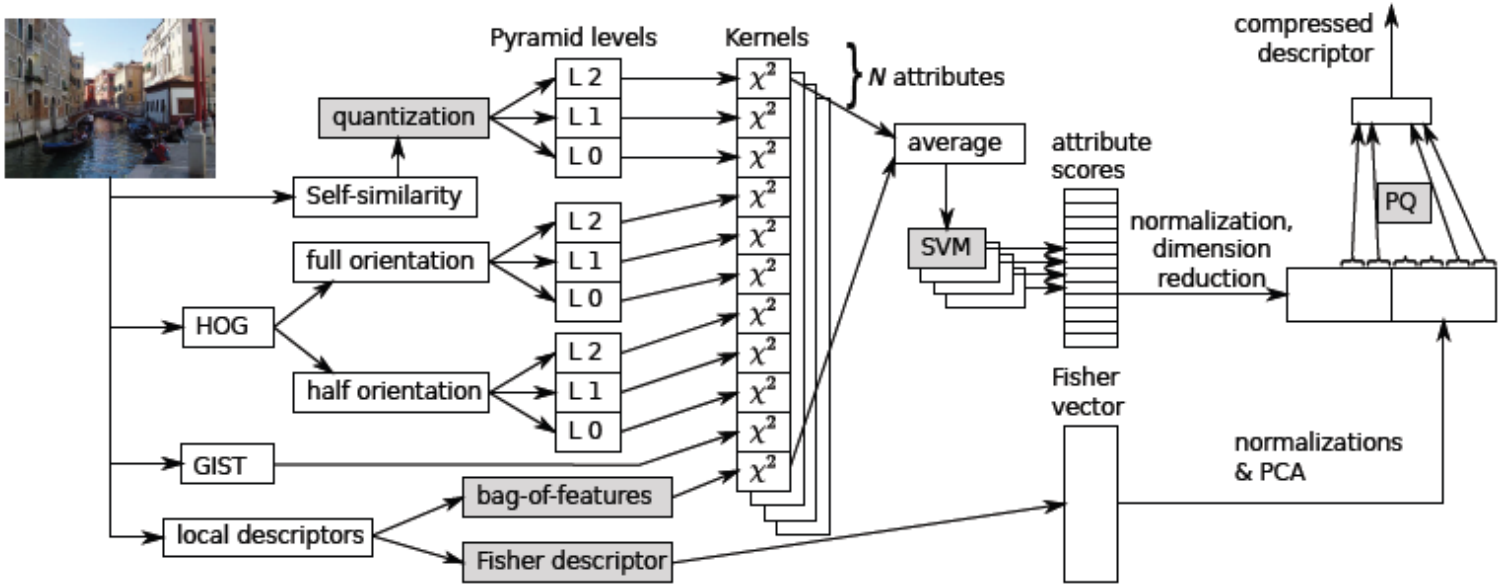
- Classifiers based on a standard SVM using LIBSVM with **2-RBF averaged kernels**
- Average of the classifiers ; make little difference
- Negative data includes one random image from each class

- **The attributes.**
  - **The image descriptor is a "class coding" of an image,** obtained as the concatenation of the scores of the attribute classifiers.

2.3 Textual features

- Images are often associated with text. (tags and user comments)
- A basic text descriptor from these annotations can be built.
  - Remove punctuation and convert all text to lowercase
  - Tokenize it into words and build a dictionary from all the words found in the corpus.
  - Remove stop words and words that are too rare in the corpus.
  - Describe each image with a (sparse) histogram of the words appearing in its annotations.





3. Indexing descriptors

- Images are represented **by global descriptors**, i.e., Fisher vectors and attribute features.
- Retrieval consists in **finding the nearest neighbors** in a high-dimensional descriptor space.
- **Describe how to combine descriptors and, then, how to search nearest neighbors efficiently.**

3.1. Combining descriptors

- **To combine Fisher vectors and attribute features, each of them should be normalized.**
  - Use the power normalization ($\alpha = 0{:}5$) [1] and normalize the vectors with the L2 norm.

- **Attribute vectors contain SVM classification scores.**
  - Normalizing the vectors with L2 or L1 norm decreases the retrieval performance.
  - For normalization rely on the distribution of the descriptors extracted from n training images: $A = [a_1 \cdots a_n]$
  - The mean description vector is significantly different from 0. $a'_i = a_i - \frac{1}{n}\sum_j a_j$
  - Compute the average vector norm on the training set, $\alpha = \frac{1}{n}\sum_j \|a'_j\| \qquad a^*_i = \frac{a'_i}{\alpha}$
  - The normalized description matrix is then $A^* = [a^*_1 \cdots a^*_n].$

[1] F. Perronnin, J. Sanchez, and T. Mensink. Improving the Fisher kernel for large-scale image classification. In ECCV, 2010.

3.1. Combining descriptors

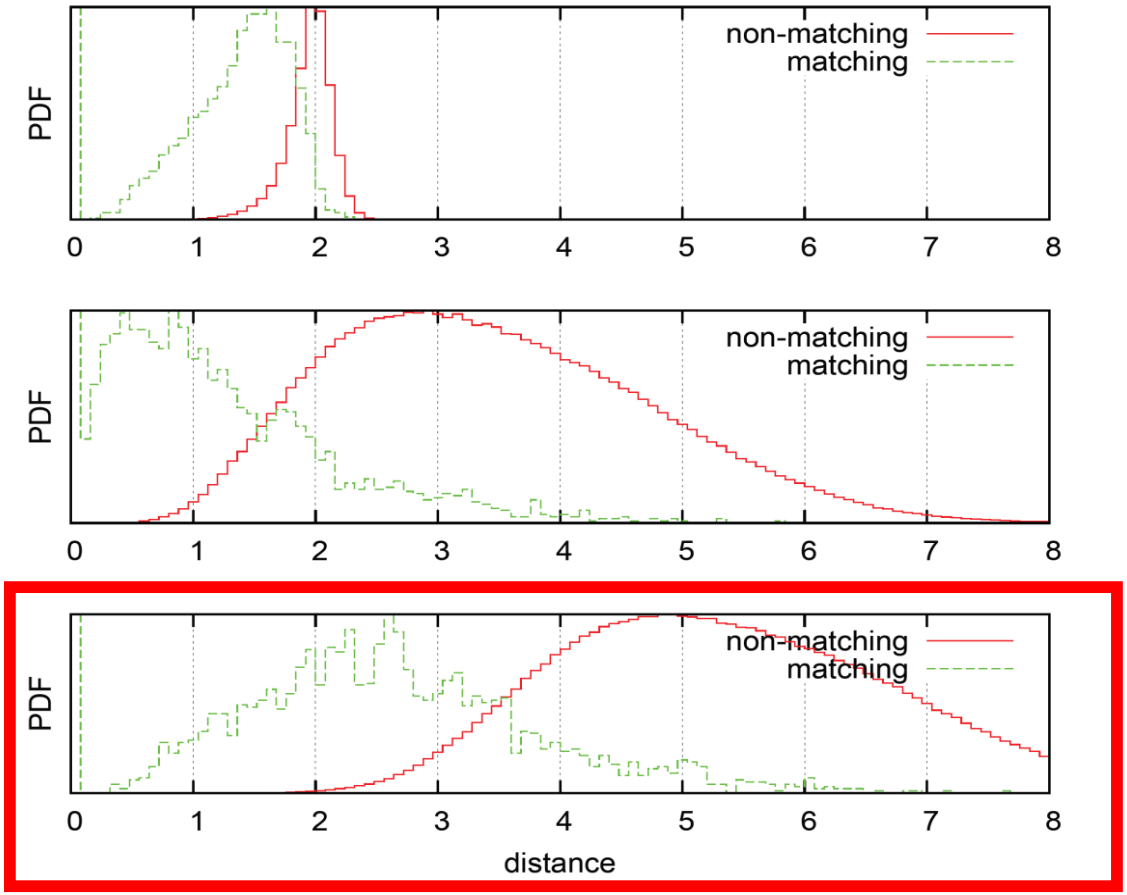


Figure 2. Probability densities for the squared distances between descriptors of matching and non-matching images in the Holidays dataset. Descriptors are: normalized Fisher descriptors (top), normalized attribute descriptors (middle), the A+F combined descriptor, without weighting (bottom).

3.1. Combining descriptors

Descriptor	dimension	mAP
BOF k=1000 [6]	1000	41.1
Fisher k=64 [17]	4096	$\approx$ 60
Fisher k=4096 [17]	262144	70.5
VLAD k=64 [8]	8192	52.6
Fisher (F), k=64, L2 dist.	4096	59.5
Attributes (A), L2 dist.	2659	55.0
A + F, F-weight $\times$1	6755	64.5
A + F, F-weight $\times$2	6755	69.5
A + F, F-weight $\times$2.3	6755	69.9

Table 1. Comparison of the different descriptors and their combination on the Holidays dataset.

- To combine them, add up the squared L2 distances.
- Performance can be improved by using a weighting factor **to increase the contribution of the Fisher vector.**

3.2. Dimension reduction

- **To accelerate retrieval, project the vectors to a lower dimension**.
  - A good choice is to apply a PCA transform.
  - Random selection and
  - Selection based on the cross-validation error of the classifiers

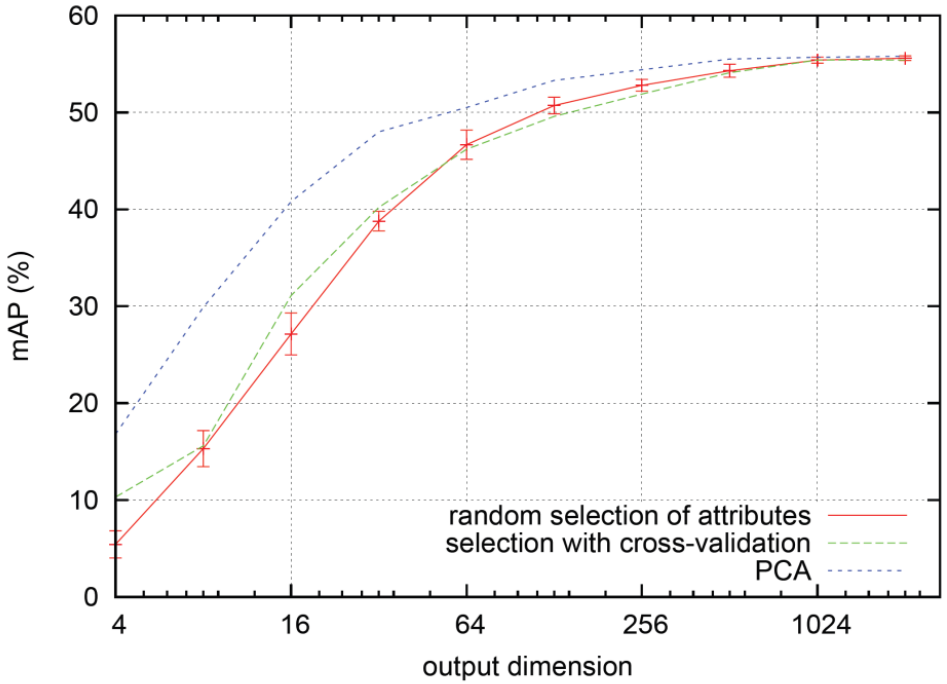


Figure 4. Comparison of different dimensionality reduction techniques for the attribute features on the Holidays dataset. mAP performance is displayed as a function of the dimension.

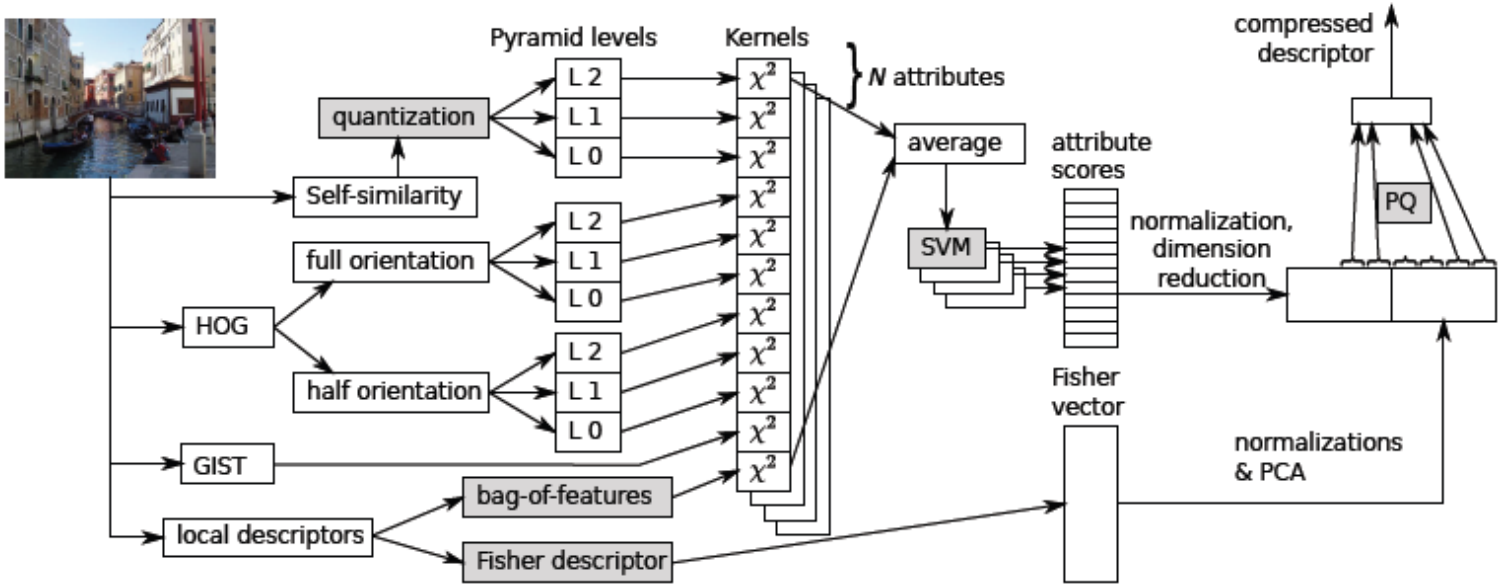
3.3. Coding and searching

- An additional improvement of efficiency and compactness can be obtained by **encoding the image descriptors.**

- To encode dimensionality reduced vectors, use product quantization method of Jegou & al. [7].

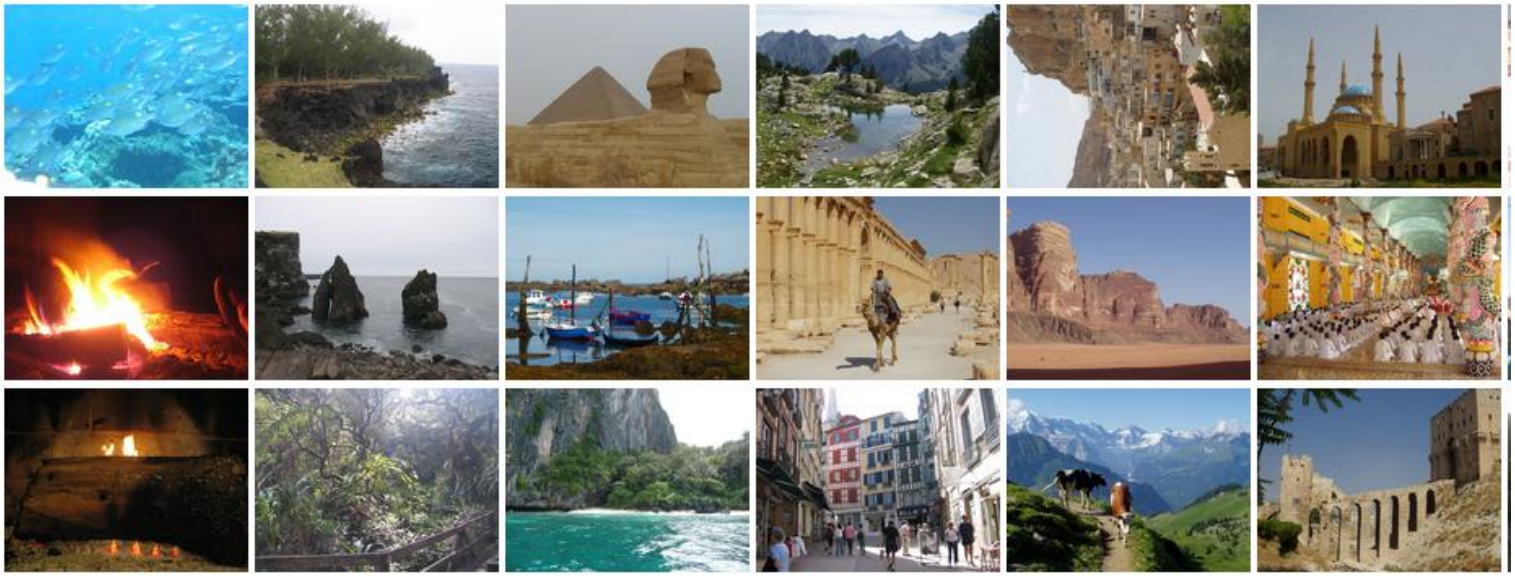
- This method was shown to be very efficient for approximate nearest neighbor search with the L2 distance in high dimensional spaces for large datasets .

[7] H. J´egou, M. Douze, and C. Schmid. Product quantization for nearest neighbor search. IEEE Transactions on Pattern Analysis & Machine Intelligence, 33(1):117–128, jan 2011.



4.1. Image retrieval of particular objects

- The retrieval of particular objects on the INRIA Holidays dataset
  - A collection of 1491 holiday images,
  - 500 of them being used as queries, each of which represents a distinct scene.
  - The accuracy is measured by mean Average Precision (mAP).



4.1. Image retrieval of particular objects

Descriptor	dimension	mAP
BOF k=1000 [6]	1000	41.1
Fisher k=64 [17]	4096	≈ 60
Fisher k=4096 [17]	262144	70.5
VLAD k=64 [8]	8192	52.6
Fisher (F), k=64, L2 dist.	4096	59.5
Attributes (A), L2 dist.	2659	55.0
A + F, F-weight $\times 1$	6755	64.5
A + F, F-weight $\times 2$	6755	69.5
A + F, F-weight $\times 2.3$	6755	69.9

Table 1. Comparison of the different descriptors and their combination on the Holidays dataset.

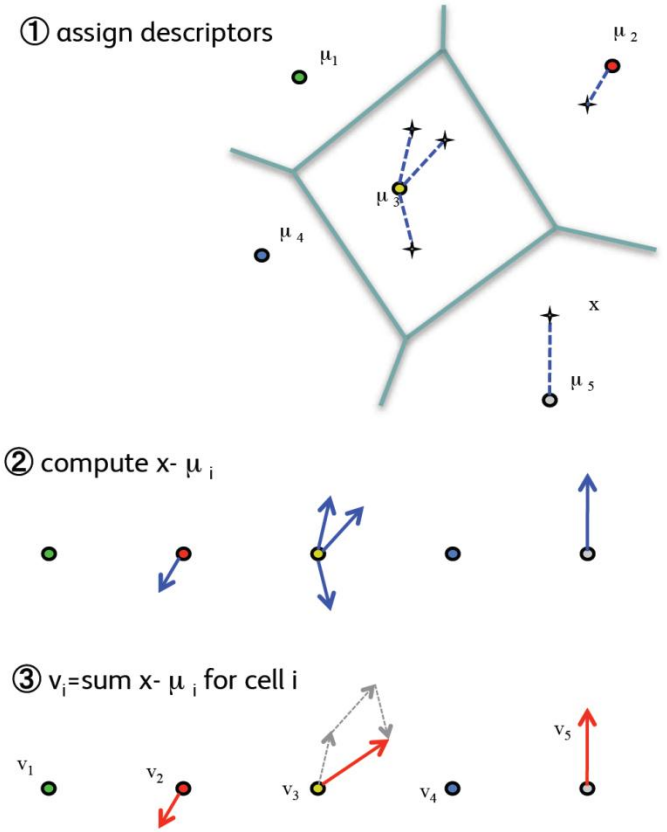
[6] H. Jegou, M. Douze, and C. Schmid. Packing bag-of-features. In ICCV, 2009.
[8] H. J´egou, M. Douze, C. Schmid, and P. Perez. Aggregating local descriptors into a compact image representation. In CVPR, 2010.
[17] F. Perronnin,Y. Liu, J. Sanchez, and H. Poirier. Large-scale image retrieval with compressed Fisher vectors. In CVPR, 2010.

4.1. Image retrieval of particular objects

A first example: the VLAD

Given a codebook $\{\mu_i, i = 1 \dots N\}$,
e.g. learned with K-means, and a set of local descriptors $X = \{x_t, t = 1 \dots T\}$:

- ① assign: $\mathrm{NN}(x_t) = \arg\min_{\mu_i} ||x_t - \mu_i||$
- ②③ compute: $v_i = \sum_{x_t : \mathrm{NN}(x_t) = \mu_i} x_t - \mu_i$
- concatenate v_i's + ℓ_2 normalize



Jégou, Douze, Schmid and Pérez, "Aggregating local descriptors into a compact image representation", CVPR'10.

4.1. Image retrieval of particular objects

Descriptor	dimension	mAP
BOF k=1000 [6]	1000	41.1
Fisher k=64 [17]	4096	$\approx$ 60
Fisher k=4096 [17]	262144	70.5
VLAD k=64 [8]	8192	52.6
Fisher (F), k=64, L2 dist.	4096	59.5
Attributes (A), L2 dist.	2659	55.0
A + F, F-weight $\times$1	6755	64.5
A + F, F-weight $\times$2	6755	69.5
A + F, F-weight $\times$2.3	6755	69.9

Table 1. Comparison of the different descriptors and their combination on the Holidays dataset.

Implementation of the Fisher descriptor performs similarly to implementation [17].

Fisher and VLAD descriptors **with a somewhat lower dimensionality.**

[6] H. Jegou, M. Douze, and C. Schmid. Packing bag-of-features. In ICCV, 2009.
[8] H. J´egou, M. Douze, C. Schmid, and P. Perez. Aggregating local descriptors into a compact image representation. In CVPR, 2010.
[17] F. Perronnin,Y. Liu, J. Sanchez, and H. Poirier. Large-scale image retrieval with compressed Fisher vectors. In CVPR, 2010.

4.1. Image retrieval of particular objects

Descriptor	dimension	mAP
BOF k=1000 [6]	1000	41.1
Fisher k=64 [17]	4096	$\approx$ 60
Fisher k=4096 [17]	262144	70.5
VLAD k=64 [8]	8192	52.6
Fisher (F), k=64, L2 dist.	4096	59.5
Attributes (A), L2 dist.	2659	55.0
A + F, F-weight $\times$1	6755	64.5
A + F, F-weight $\times$2	6755	69.5
A + F, F-weight $\times$2.3	6755	69.9

Table 1. Comparison of the different descriptors and their combination on the Holidays dataset.

Combination of the attribute features with Fisher descriptor improves the performance

Similar dimensionality authors' descriptor outperforms the **state of the art.**

[6] H. Jegou, M. Douze, and C. Schmid. Packing bag-of-features. In ICCV, 2009.
[8] H. J´egou, M. Douze, C. Schmid, and P. Perez. Aggregating local descriptors into a compact image representation. In CVPR, 2010.
[17] F. Perronnin,Y. Liu, J. Sanchez, and H. Poirier. Large-scale image retrieval with compressed Fisher vectors. In CVPR, 2010.

4.1. Image retrieval of particular objects

Descriptor	dimension	mAP
BOF k=1000 [6]	1000	41.1
Fisher k=64 [17]	4096	≈ 60
Fisher k=4096 [17]	262144	70.5
VLAD k=64 [8]	8192	52.6
Fisher (F), k=64, L2 dist.	4096	59.5
Attributes (A), L2 dist.	2659	55.0
A + F, F-weight $\times 1$	6755	64.5
A + F, F-weight $\times 2$	6755	69.5
A + F, F-weight $\times 2.3$	6755	69.9

Table 1. Comparison of the different descriptors and their combination on the Holidays dataset.

- The influence of the weighting factor is not critical.
- Values in the range between 1.5 and 2.5 produce very similar results.
- In the following, they always use a weight of 2.3.

[6] H. Jegou, M. Douze, and C. Schmid. Packing bag-of-features. In ICCV, 2009.
[8] H. Jegou, M. Douze, C. Schmid, and P. Perez. Aggregating local descriptors into a compact image representation. In CVPR, 2010.
[17] F. Perronnin,Y. Liu, J. Sanchez, and H. Poirier. Large-scale image retrieval with compressed Fisher vectors. In CVPR, 2010.

4.1. Image retrieval of particular objects

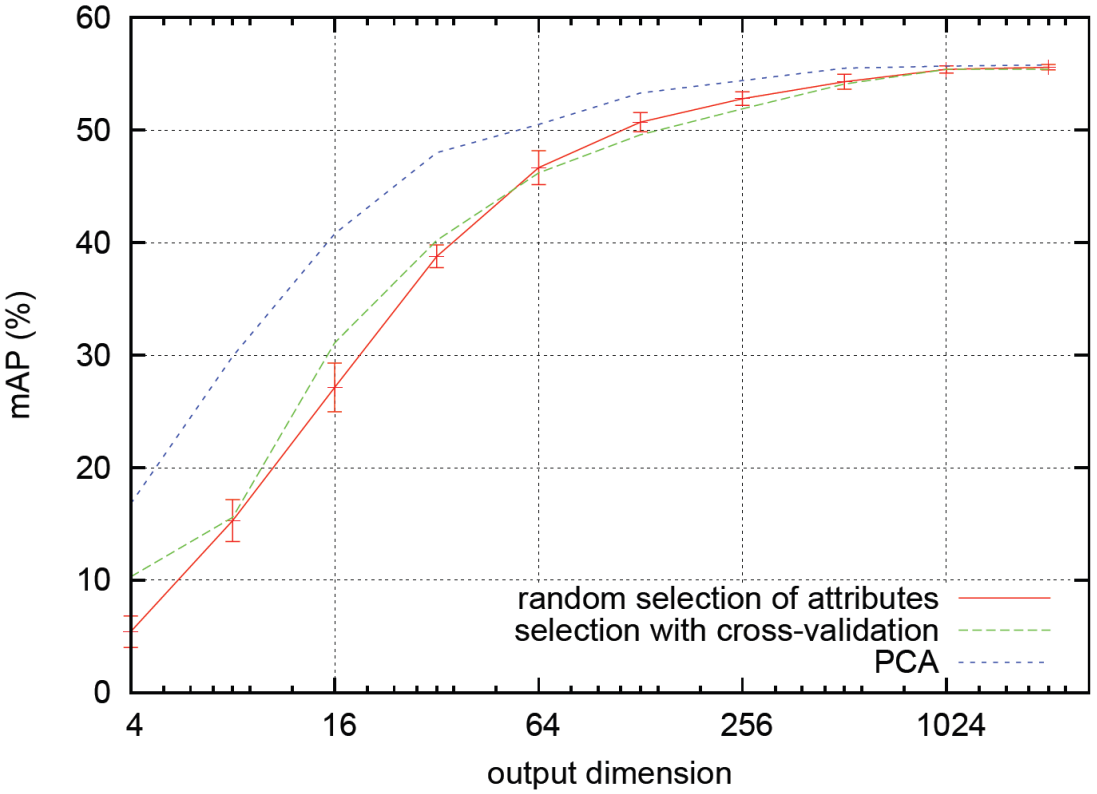
Comparison of the retrieval results obtained with the

- Fisher vector,
- Attribute features,
- Their combination.

The images retrieved with attribute features are more likely to represent similar categories.

4.2. Compression and indexing

- Dimension reduction.

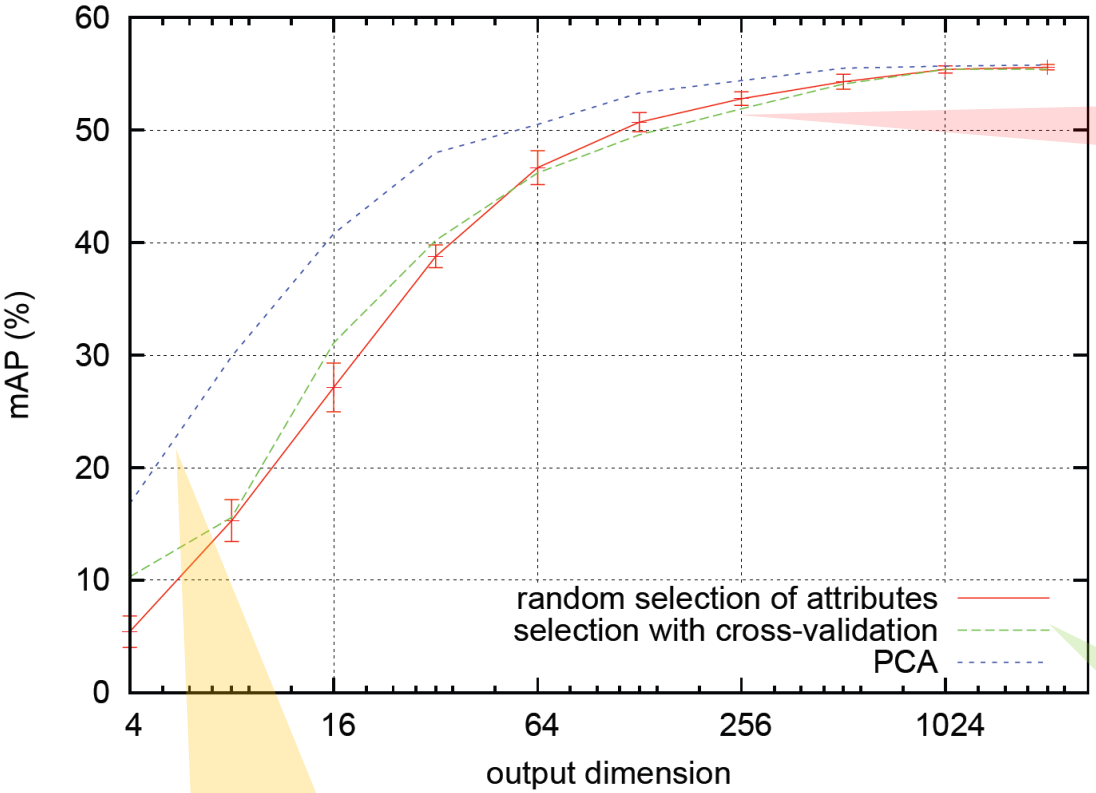


Comparison of different dimensionality reduction techniques for the **attribute features** on the Holidays dataset.

mAP performance is displayed as a function of the dimension.

4.2. Compression and indexing

- Dimension reduction.



All methods obtain excellent performance if a dimension of 256 or higher is used.

selection with cross-validation does not improve over random selection.

The curve for PCA saturates rapidly.

4.2. Compression and indexing

Evaluates the impact of the dimension on the combined A + F descriptor.

Dimension reduction			
Attributes	Fisher	dimension	mAP
PCA 512	PCA 512	1024	69.3
PCA 256	PCA 256	512	68.2
PCA 64	PCA 64	128	63.3
PCA 16	PCA 16	32	54.0
select random 256	PCA 256	512	67.9

Table 2. Dimensionality reduction for the combined descriptor on the Holidays dataset.

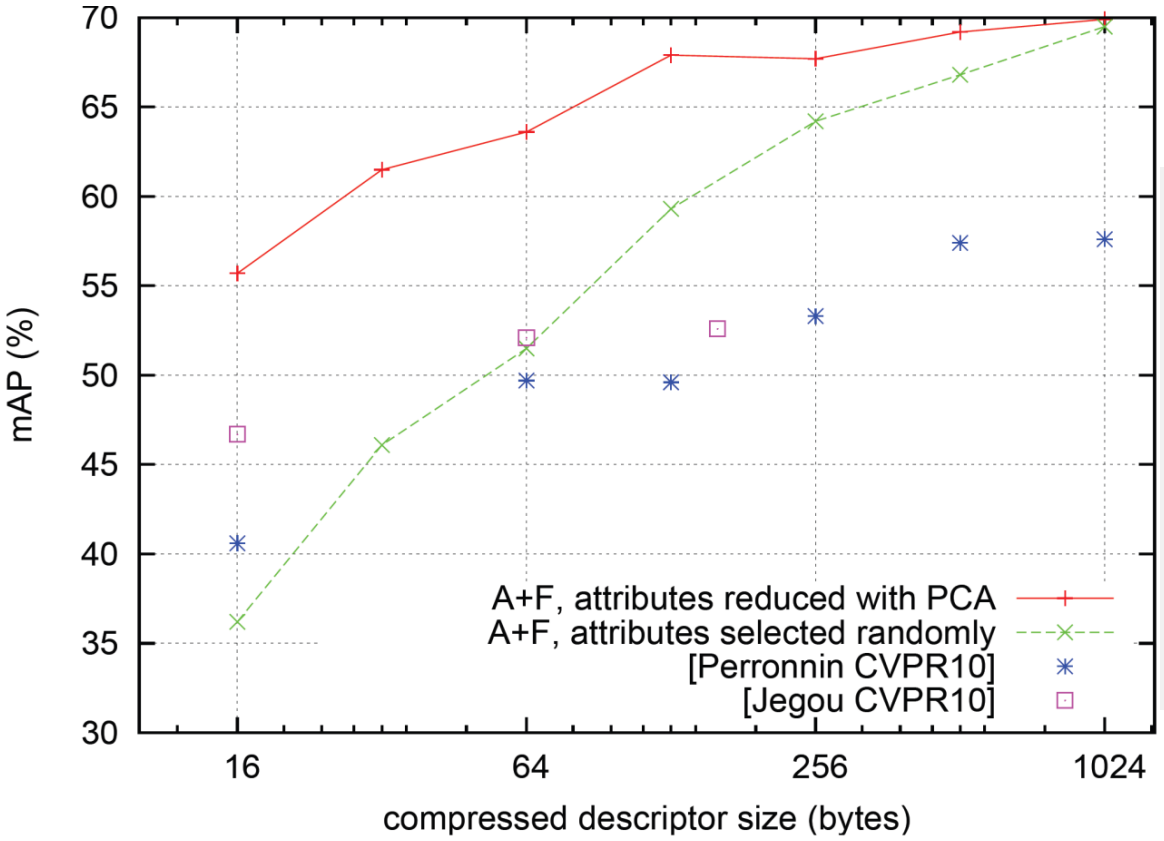
Reducing the number of dimensions to 1024, with PCA has almost no impact on the results.

A + F, F-weight ×2.3	6755	69.9

the reduction with randomly selected features gives almost as good results as PCA for 256 dimensions.

4.2. Compression and indexing

- Encoding.

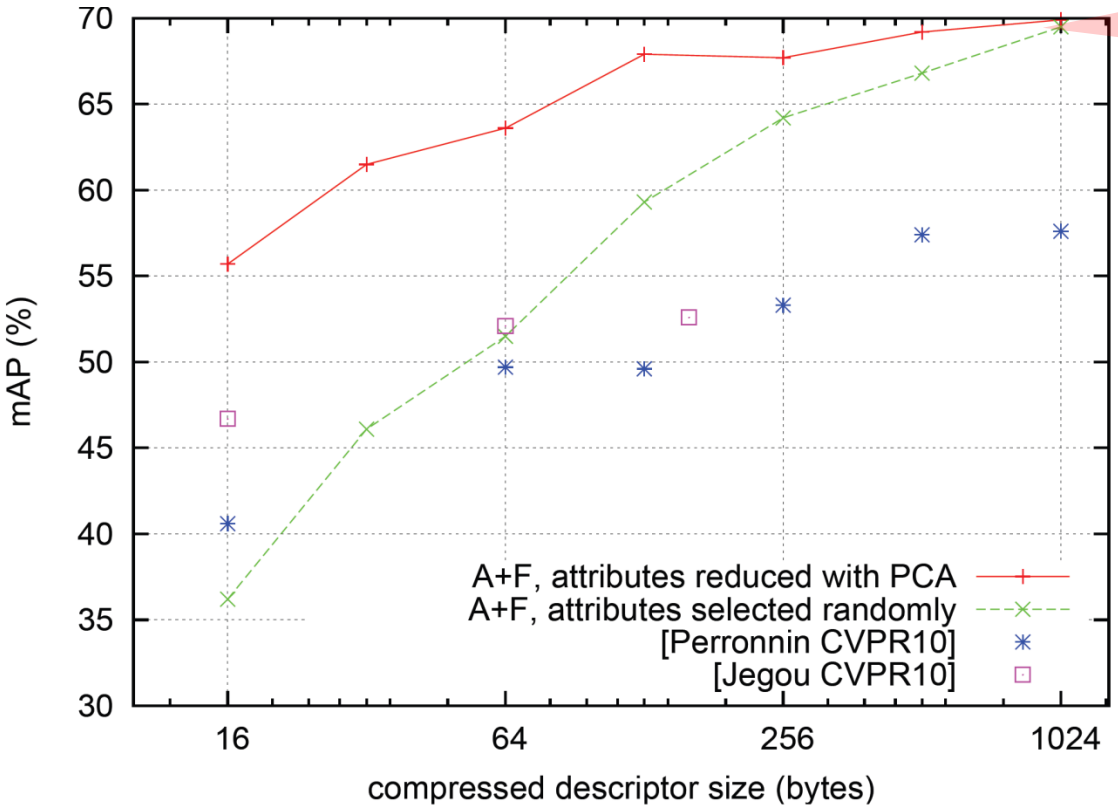


- Performance of the A+F descriptor after dimension reduction
- Encoding of descriptors in addition to dimension reduction
- The Fisher vectors reduced with PCA.

F. Perronnin, Y. Liu, J. Sanchez, and H. Poirier. Large-scale image retrieval with compressed Fisher vectors. In CVPR, 2010.
H. Jegou, M. Douze, C. Schmid, and P. Perez. Aggregating local descriptors into a compact image representation. In CVPR, 2010.

4.2. Compression and indexing

- Encoding.



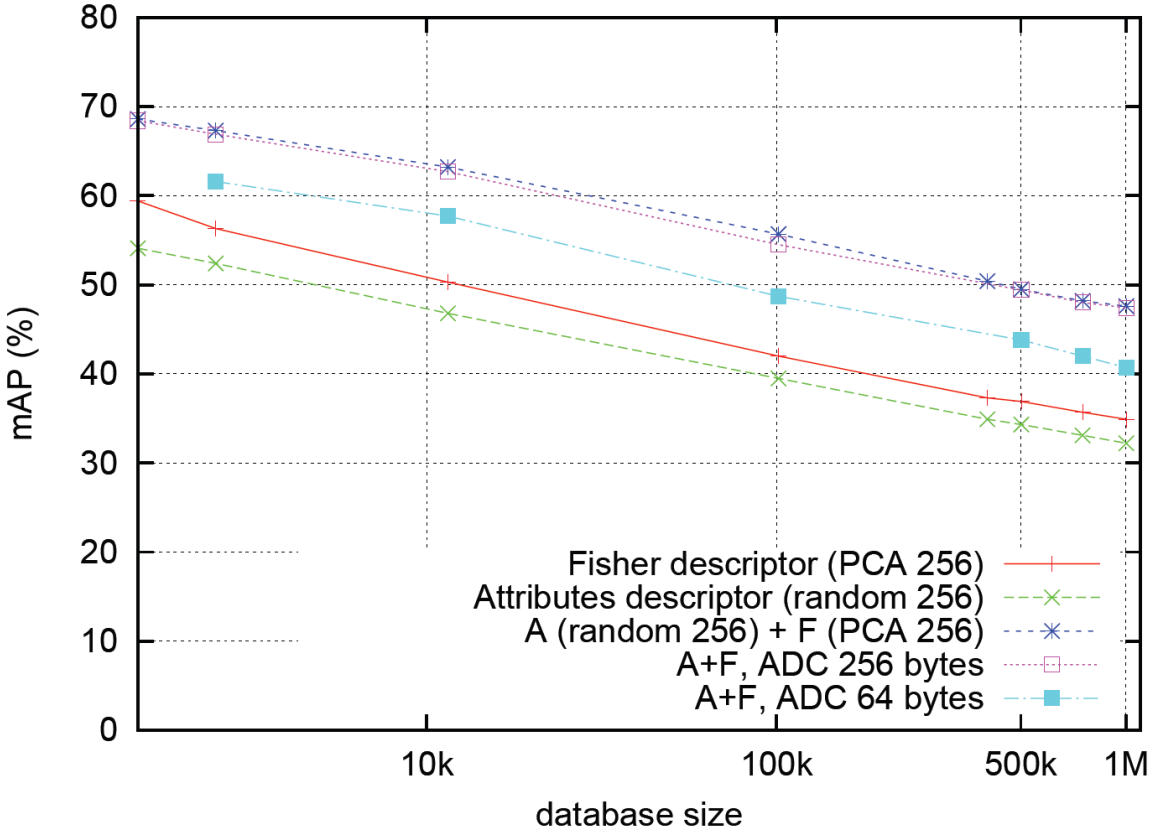
Author's approach significantly outperforms that state of the art.

Encoding increases performance gap between dimension reduction by PCA and random attribute selection.

F. Perronnin, Y. Liu, J. Sanchez, and H. Poirier. Large-scale image retrieval with compressed Fisher vectors. In CVPR, 2010.
H. Jegou, M. Douze, C. Schmid, and P. Perez. Aggregating local descriptors into a compact image representation. In CVPR, 2010.

4.2. Compression and indexing

- Large-scale experiments.

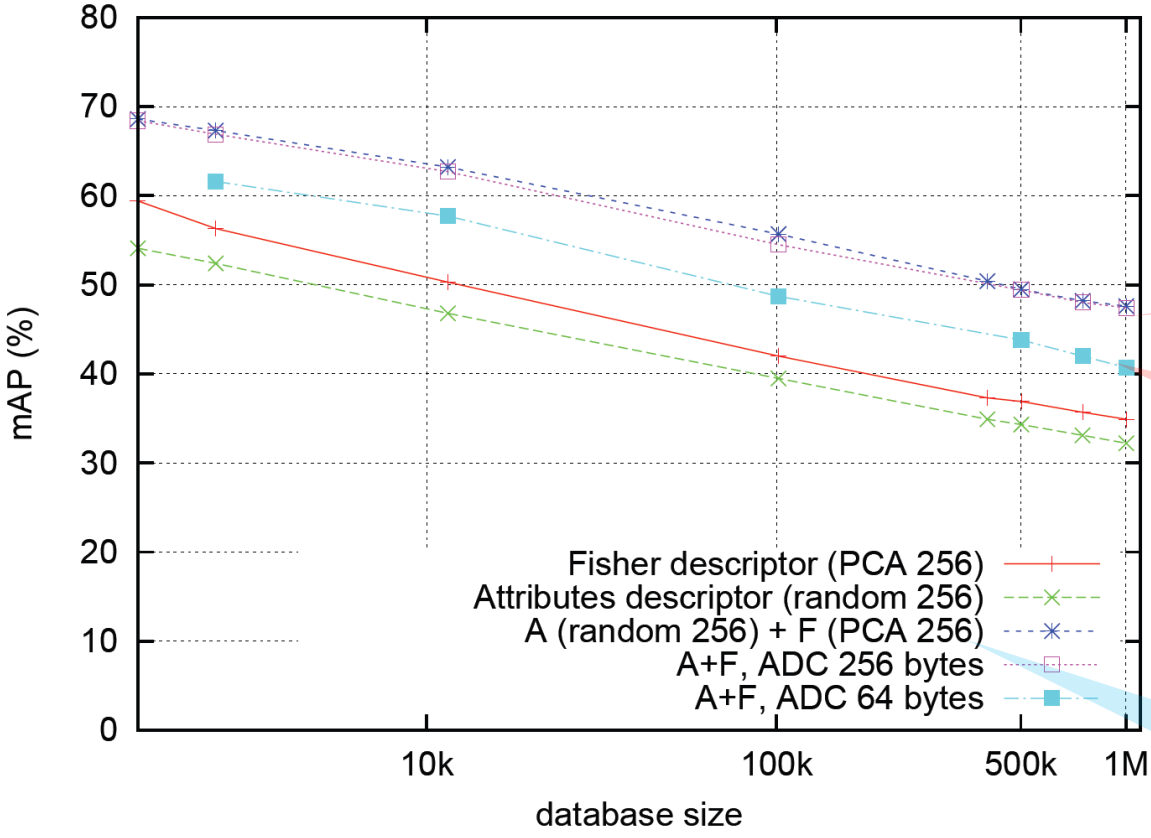


Performance on the Holidays dataset combined with one million distractor

images from Flickr.

4.2. Compression and indexing

- Large-scale experiments.



A compression to 256 bytes : almost no loss in performance

results for a compression to 64 excellent.

Combination improves significantly over the individual descriptors

Attributes, random subset of 256 attributes, to speed u computation.
Project Fisher vectors to 256 dimensions with PCA.

4.3. Image retrieval of categories

Authors evaluate retrieval of categories on the "webqueries" dataset

The images **also contain text**.

Evaluate the precision@10.

group	concepts		nb.
	examples	nb.	images
Person/people	Madonna, Spiderman	104	8721
Object/animal	violin, shark	64	4892
Landmark	Machu Picchu, Big Ben	55	5087
Specific image	Tux Linux, Guernica	54	3086
Cartoon	B. Simpson, Snoopy	18	1817
Scene	tennis court, forest	16	982
Type of image	painting, clipart	16	1543
Event	Cannes festival, race	11	242
Other	meal, family, GTA	15	1586
Total		353	27956

Table 3. The "web-queries" dataset. The 353 concepts are split in 9 groups. The number of concepts as well as relevant/positive images are indicated for each group.

4.3. Image retrieval of categories

Concepts vs. attributes.

concept	3 strongest attributes
george clooney	actor on TV, celebrity, actor in musicals
spider	insect, pit, invertebrate
dolphin	warplane, submarine, rowboat
forest	forest, garden, broadleaf forest
tower	bell tower, observation tower, minaret
mont blanc	glacier, snow skier, mountain
opera sydney	boat ship, warplane, patrol boat

Table 4. Relationship between web-queries concepts and attributes.

- **The attributes give relevant semantic information about the concepts.**

4.3. Image retrieval of categories

The precision@10 results for the web-queries dataset.

group	Fisher (F)	Att. (A)	Combination A+F	Combination A+F+T	Text (T)
Person/people	11.9	8.5	13.2	61.1	51.8
Object/animal	4.1	6.6	6.6	45.7	37.8
Landmark	18.2	20.8	27.8	72.4	62.6
Specific image	35.4	33.5	37.4	53.9	33.2
Cartoon	16.3	14.0	19.4	64.4	54.8
Scene	4.3	6.1	6.9	37.9	28.3
Type of image	9.0	10.2	12.2	45.7	31.7
Event	20.5	13.8	21.1	30.2	18.0
Other	30.1	27.3	32.5	65.6	50.2
Total	15.2	14.6	18.7	58.2	47.1

Table 5. Precision@10 for different descriptors (Fisher, attribute and text as well their combination) on the web-queries dataset.

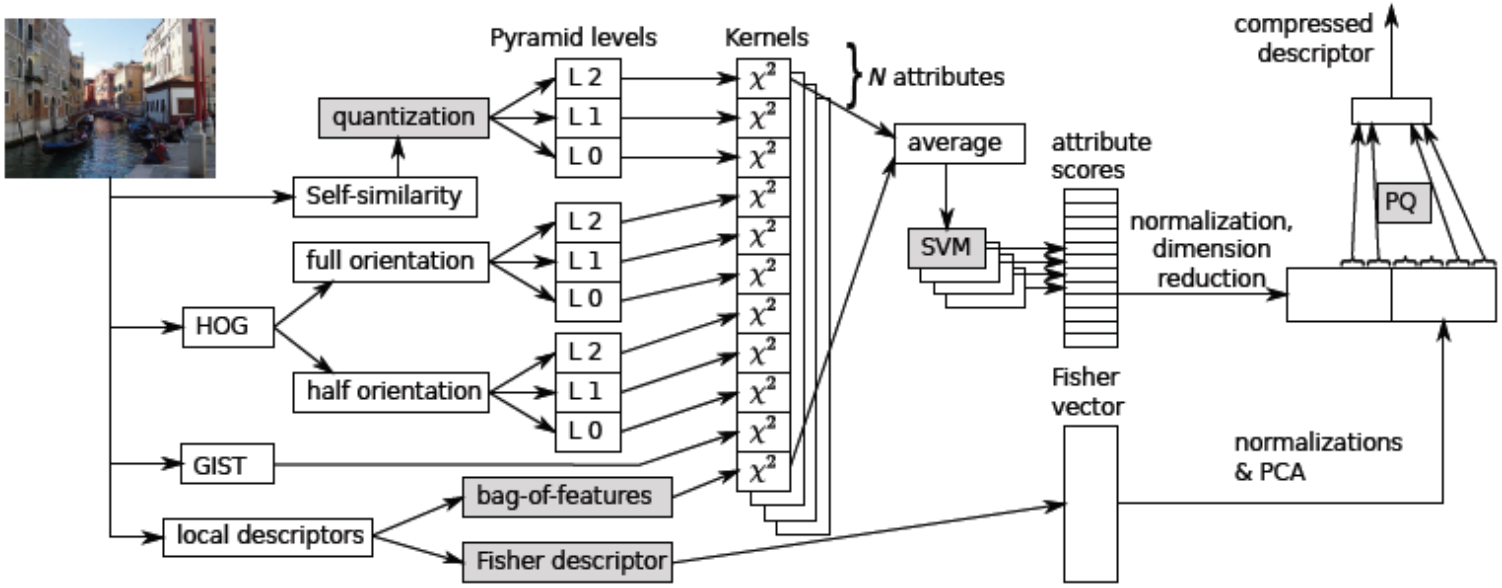
Results for image-based are low.

- the variety of the images
- image descriptors are far from sufficient for category search on a web-scale.

4.3. Image retrieval of categories

- Retrieval examples for the web-queries dataset with our A+F descriptor.

Figure 7. Example queries for "place de la concorde" (first two queries), "moon", "guitar", and "Norah Jones" concepts. We use the image-only A+F descriptor. Retrieval results are displayed in order. True positives are marked with a box.



5. Conclusion

- **Attribute features, a** high-level classification-based image representation, **contribute to the task of image retrieval.**
- **Combining state-of-the-art Fisher vectors with attribute features improves the performance significantly.**
- **Combination of image and text improves significantly**

[1] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In CVPR, 2009.

Fisher Kernels on Visual Vocabularies for Image Categorization Florent Perronnin and Christopher Dance
http://www.cedar.buffalo.edu/~srihari/CSE574/index.html
Large / larger-scale image search Introduction, Hervé Jégou, INRIA