Chapter 4
Spatial and multi-resolution context in visual indexing

Jenny Benois-Pineau, Aurélie Bugeau, Svebor Karaman, Rémi Mégret

Abstract Recent trends in visual indexing make appear a large family of methods which use a local image representation via descriptors associated to the interest points, see chapter 2. Such approaches mostly ”forget” any structure in the image considering unordered sets of descriptors or their histograms as image model. Hence, more advanced approaches try to overcome this drawback by adding spatial arrangements to the interest points. In this chapter we will present two trends in incorporation of spatial context into visual description, such as considering spatial context in the process of matching of signatures on one hand and design of structural descriptors which are then used in a global Bag-of-Visual-Words (BoVW) approach on the other hand. As images and video are mainly available in a compressed form, we shortly review global descriptors extracted from compressed stream and hence less sensible to compression artifacts. Furthermore, on the basis of scalable, multi-resolution/multi-scale visual content representation in modern compression standards, we study how this multi-resolution context can be efficiently incorporated into a BoVW approach.

4.1 Introduction

If one can try today to trace the origins of approaches for indexing and retrieval of visual information such as images, videos and visual objects in them, tree main sources could be identified. They are i) text indexing and retrieval approaches, ii) visual coding by vector quantization, iii) structural pattern recognition.

The first two families of methods together with local image analysis inspired the Bag-of-Visual-Words (BoVW) approach which has been exhaustively presented in chapters 2 and 3. In this approach the visual content of an image, a video frame or an object of interest is characterized by a global signature. The latter represent histograms of quantized visual descriptors obtained by the analysis of local neighbourhoods. Hence the spatial relations in image plane between regions in images and object parts are lost.

Conversely, spatial relations are at the core of structural pattern recognition, where spatial graph models are used to describe the visual content. In this case the nodes of the graph represent the visual primitives extracted from the images during an analysis step, and encode elements such as homogeneous regions, linear primitives from contour and skeleton or points of interest. The graph edges encode the spatial relations of the primitives in the image plane. Hence matching and similarity search between visual entities can be formulated as a graph-matching problem [195, 177].

In the case of large graphs covering the whole image content the graph matching becomes computationnaly heavy and cannot be deployed at a large scale for visual indexing and retrieval. Therefore, given the need of analysis of video content practically "on-the-fly" in broadcast applications, the tremendous volumes of visual information in image and video databases make classical methods of structural pattern recognition inapplicable.

Hence, spatial context has been incorporated in visual indexing approaches by relating visual signatures to more local areas in image plane, and matching characteristic points for similarity refining, etc. Furthermore, for the sake of scalability, it is very much seducing to incorporate multi-resolution and multi-scale representation of visual content in image and video retrieval in order to realise a "progressive" indexing and retrieval which allows to fasten the search operation. Indeed, to recognise a visual scene, humans hardly need any fine details available at full resolution, but successfully fulfil the visual recognition task on under-scaled and degraded versions of the content. This was the subject of "Rough indexing" paradigm we developed for indexing and retrieval of HD video [139]. The incorporation of a multi-resolution representation not only at the level of images, but also at the level of salient region extraction (SIFT) or characteristic points, seems also a promising way in visual indexing.

Nevertheless all these approaches comprise the same fundamental step which consists in detecting features in the raw image domain. With regard to the properties of visual content to analyse this detection may be unstable. Indeed, image and video content items are stored in repositories and exchanged via heterogeneous channels of visual information in a compressed form. They are practically never available in a raw pixel domain. Yet image and video compression artefacts affect interest point detectors or other differential features. On the other hand, modern compression standards such as JPEG2000 [2] already incorporate multi-resolution/multi-scale representations.

Hence in this chapter we are interested in these two aspects: incorporation of spatial context in visual content indexing and retrieval and multi-resolution/multi-scale content description. In section 4.2 we will review methods which incorporate spatial context into indexing and retrieval of visual content and present our recent works on the border of BoVW and structural pattern recognition approaches we call "GraphWords". In section 4.3 we are interested in multi-resolution/multi-scale strategies and, in the follow up of our research under the Rough Indexing paradigm, propose visual indexing approaches based on the wavelet pyramids of the JPEG2000 standard. The conclusions and perspectives are drawn in section 4.4

4.2 Incorporating spatial context

The standard Bag-of-Visual-Words approach, presented in chapter 3, represents an image by a global histogram of visual words distribution. This representation does not cover one important part of an image or an object: the spatial organization.

In the past few years, several methods have tried to overcome the lack of spatial information and relations between interest regions in the BoVW framework. For example, a method for integrating the spatial information was presented in [166], where after applying a BoVW approach for retrieval the top ranked images where re-ranked by applying a LO-RANSAC [45] algorithm with affine transformations. This method can be seen as a post-processing. We will focus in this section on methods where the spatial organization is fully integrated in the approaches.

We will first report approaches that use local histograms instead of global histograms. Then we review approaches which introduce spatial context during the matching process. Structural pattern have been widely represented by graphs, we will show that graph matching is an efficient approach for object recognition but can hardly be applied to large databases retrieval. Finally we will introduce our research which tries to overcome the ambiguity of the visual words relying on very local features and the lack of spatial organization in the BoVW framework.

4.2.1 Local histograms of visual words

One popular and successful approach to overcome the lack of spatial information within the BoVW framework is the Spatial Pyramid Matching Kernel (SPMK) approach introduced in [126] and referenced in chapter 3 . The method uses the Pyramid Match Kernel [86] in order to compare image signatures according to a visual vocabulary but applying the pyramid construction to the coordinates of the features in the image space.

The image plane is successively partitioned into blocks according to the ”levels” of pyramid. At level $l = 0$ the only block is the whole image. At all other levels up to $l = L$ the image is partitioned into $2^l x 2^l$ blocks. The features are quantized into K discrete classes according to a visual vocabulary C obtained by traditional clustering techniques in feature space. Only features of the same class k can be matched. For a pair of images X and Y to compare, each class k gives two sets of two-dimensional vectors, X_k and Y_k, representing the coordinates of features of class k found in images X and Y respectively. Let us denote $H^l(X_k)$ the histogram of features of class k in image X according to the fixed partitioning of the pyramid at level l. Using the histogram intersection similarity measure $\mathcal{I}$ introduced in chapter 3, the SPMK for features of class k is defined as:

$$K^L(X_k, Y_k) = \frac{1}{2^L}\mathcal{I}(H^0(X_k), H^0(Y_k)) + \sum_{l=1}^{L} \frac{1}{2^{L-l+1}}\mathcal{I}(H^l(X_k), H^l(Y_k)) \qquad (4.1)$$

The final kernel (4.2) is the sum of the kernels associated to each feature class:

$$K^L(X,Y) = \sum_{k=1}^{K} K^L(X_k, Y_k) \tag{4.2}$$

For $L = 0$ the approach reduces to a standard BoVW.

Experiments on three publicly available image databases show significant improvements using the Spatial Pyramid Matching approach compared to BoVW. However, since locations are expressed in absolute coordinates, the representation is unsuitable in the case of spatial displacement of the object of interest unless exhaustive search is done using a spatial sub-window.

In [6], Albatal et al. note two important limitations of the BoVW framework as visual words are much more ambiguous than text words and that in the global histogram representation, all information related to topological organization of the regions of interest in the image is lost. They have proposed a method to create groups of regions in the image to form areas which are spatially larger than the individual regions and have the same robust visual properties. As shown in [231], grouping several regions is more discriminative for object classification than individual regions.

Albatal et al. uses a "Single Link" clustering function, with a topological proximity criterion based on the Euclidean distance between the regions of interest. This criterion defines two regions as close if the Euclidean distance between their centres is less or equal than the sum of their radii. This type of clustering does not depend on the starting point and ensure that the created groups are disjoint. Each cluster corresponds to a set of regions that defines a "visual phrase".

Each Visual Phrase is then represented as a BoVW with regard to a visual dictionary C. By construction, resulting visual phrases are invariant to scale, rotation, and translation transformations and to brightness changes.

The approach is evaluated on an automatic annotation task on the VOC2009 collection. The evaluation shows that using Visual Phrases only yields poorer results than the baseline (BoVW on the whole images). According to the authors this is mainly because Visual Phrases account only for the description of the objects while the baseline approach integrates information about the background as well. A late fusion of recognition score for each image enhances the performance above baseline's initial results. This approach builds local BoVWs but does not take into account the spatial organization for description of the visual phrases.

4.2.2 Context-matching kernels

The previous approaches have defined local histograms by a fixed partitioning or data adaptive partitioning. However, they still miss description of spatial configuration of features.

A spatial weighting approach was introduced in [140] for object recognition on

cluttered background. The hypothetical object mask is estimated in test images by matching quantized features with those in training images. During the matching affine transformations are applied accordingly to the scale and orientation of feature points. The final hypothesis segmentation mask is a weighted sum of the transformed masks. The test features are then weighted according to this mask thus giving lower weights to background features.
In [183] and [182], Sahbi et al. have introduced a kernel which takes into account both feature similarity "alignment quality" and spatial alignment in a "neighbourhood" criteria. Let us denote two sets of interest regions $S_A = \{r_1^A, \ldots, r_n^A\}$ and $S_B = \{r_1^B, \ldots, r_m^B\}$ extracted from two images A and B respectively, where a region r_i^I of image I is defined by its coordinates (x_i^I, y_i^I) and a feature f_i^I: $r_i^I = (x_i^I, y_i^I, f_i^I)$. Considering any pair of regions (r_i^I, r_j^J) of two images I and J, let us denote D the matrix of dissimilarity in the feature space: $D_{r_i^I, r_j^J} = d(r_i^I, r_j^J) = \left\| f_i^I - f_j^J \right\|_2$. Let $\mathcal{N}(r_i^I)$ be the set of neighbours of r_i^I. Let us denote P the proximity matrix defined according to the neighbourhood criterion:

$$P_{r_i^I, r_j^J} = \begin{cases} 1 \text{ if I} = \text{J and r}_\text{j}^\text{J} \in \mathcal{N}(r_i^I) \\ 0 \text{ otherwise} \end{cases} \tag{4.3}$$

The Context-Dependent Kernel K is the unique solution of the energy function minimization problem and is the limit of $K^{(t)}$ defined according to the following equations:

$$\begin{aligned} K^{(t)} &= \frac{G(K^{(t-1)})}{\left\| G(K^{(t-1)}) \right\|_1} \\ G(K) &= exp(-\frac{D}{\beta} + \frac{\alpha}{\beta} P K^{(t-1)} P) \\ K^{(0)} &= \frac{exp(\frac{-D}{\beta})}{\left\| exp(\frac{-D}{\beta}) \right\|_1} \end{aligned} \tag{4.4}$$

Where exp represents the coefficient-wise exponential and $\|M\|_1 = \sum_{ij} |M_{ij}|$ represents the L_1 matrix norm. The two parameters β and α can be seen respectively as weights for features distance and spatial consistency propagation. The CDK convergence is fast, in [182] only one iteration was applied. Then the authors use the kernel values thus obtained for classification with SVM. The CDK was evaluated on the Olivetti face database, the Smithsonian leaf set, the MNIST digit database and ImageClef@ICPR set showing significant improvements of equal error rate (ERR) compared to Context-Free Kernels.

4.2.3 Graph-matching

At the other end of the spectrum of methods addressing the problem of object recognition, the spatial information has often been incorporated through a graph representation. The most common idea is to build a graph model of an object, the recognition process consisting in matching the prototype to a candidate one.

In [177], a pseudo-hierarchical graph matching has been introduced. Using local interest points, the pseudo-hierarchical aspect relies on progressively incorporating "smaller" model features (in terms of scale) as the hierarchy increases. The edges of the graph were defined accordingly to a scale-normalized proximity criterion. The model graph is matched to a new scene by a relaxation process starting from a graph model including only points of highest scale and adding smaller model features during the matching process. In [128], the graph model was defined according to locally affine-invariant geometric constraint. Each point is represented as an affine combination of its neighboring points. Defining an objective function taking into account both feature and geometric matching costs, the matching is solved by linear programming. These approaches are efficient for object matching, however when dealing with a large amount of image candidates, the matching process becomes too costly.

The comparison of graphs can be also expressed under the form of graph kernels [220], that allows to consider graphs as belonging to a RKHS, and apply standard tools such as SVM classifiers. In particular, random walk kernels are defined by considering a simultaneous walk on the two graphs to compare, with corresponds to a random walk on their direct product. Other approach transforms a graph into a set of paths [204], and apply a minor kernel to the obtained set of simpler features. Such measures rely on the extraction of meaningful sets of features, or on the exhaustive evaluation of edge matching possibilities, which scales at least quadratically with the number of node of the graphs, or requires a very sparse structures, thus limiting the size of the considered graphs in practice.

Based on the previous discussion, we believe that integrating spatial information with local interest points into a BoVW can be an elegant approach to overcome the limitations of both the BoVW framework and object matching in the case of large scale retrieval. Therefore, we will present a new semi-structural approach for content description, by a "Bag-of-Graph-Words", and study its application to object recognition.

4.2.4 Graph Words

The idea of the method consists in describing image content by a set of "small" graphs with good properties of invariance and then in fitting these features to a BoVW approach. Hence the spatial context is taken into account at the feature level.

Another property we seek is a structural multi-resolution: the graphs will be of increasing size with a nested topology of nodes.

Graph feature construction

Let us consider a graph $G = (X, E)$ where X is a set of nodes corresponding to some feature points $x_{k,k=1,..,K}$, in image plane (we take SURF points) and $E = \{e_{kl}\}_{k=1,..,K,l=1,..,K}$ is a set of edges $e_{kl} = (x_k, x_l)$ connecting these points. We call such a graph a "graph feature". We will build these features upon sets of neighbouring feature points in image plane. In order to build such graphs two questions have to be addressed:

- the choice of a feature point set X;
- the design of the connectivity edges E.

To define the feature point sets X we first select the "seeds". Around them, other feature points will be selected to build each graph feature. Selected seeds have to form a set of SURF points which are more likely to be detected in various instances of the same object. SURF points are detected where local maxima of the response of the approximated Hessian determinant are reached [16]. The points with higher response correspond to more salient visual structures and are therefore more likely to be repeatable. Hence, we select them as seeds. Considering a fixed number of seeds N_{Seeds}, we can define the set of seeds $S = \{s_1, \ldots, s_{N_{Seeds}}\}$.

Given S, our aim is to add partial structural information of the object while keeping the discriminative power of SURF key points. We will therefore define graphs over the seeds and their neighboring SURF points. Finding the k spatial nearest SURF neighbors of each seed s_i defines the set of neighbors $P_i = \{p_1, \ldots, p_k\}$.

Hence the set of nodes X^{G_i} for each graph G_i is defined as the seed s_i and the neighbours P_i. For the edges we use the Delaunay triangulation [198] of the set $\{s_i\} \bigcup P_i$, which is invariant with regard to affine transformations of image plane preserving angles: translation, rotation and scaling.

The nested layered approach

The choice of the number of nodes in a graph feature obviously depends on various factors such as image resolution, complexity of visual scene or its sharpness... This choice is difficult a priori. Instead we propose a hierarchy of "nested" graphs for the same image, capturing structural information of increasingly higher order and illustrate it in Figure 4.1. Let us introduce a set of L "layers". We say that the graph G_i^l at layer l and the graph G_i^{l+1} at layer $l+1$ are nested if the set of nodes of graph G_i^l is included in the set of nodes of graph G_i^{l+1}: $X_i^l \subset X_i^{l+1}$. Note that, so defined, the number of graphs at each layer is the same. Furthermore, in the definition (by construction) of graph features a node can belong to more than one graph of the same layer. We still consider these graph features as separate graphs.

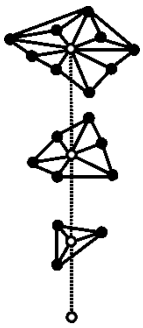


Fig. 4.1 The nested approach. Bottom to top: SURF seed depicted as the white node, 3 neighbours graph where neighbours are in black, 6 neighbours graph and 9 neighbours graph at the top level.

Introducing this layered approach, where each layer adds more structural information, we can define graphs of increasing size when moving from one layer to the next one. Each layer has its own set of neighbours around each seed s_i and the Delaunay triangulation is performed on each layer separately. To avoid a large number of layers, the number of nodes added at each layer should induce a significant change of structural information. To build a Delaunay triangulation, at least two points have to be added to a single seed. Adding one more node may yield three triangles instead of just one, resulting in a more complete local pattern. Therefore, the number of nodes added from one layer to the upper one is fixed to three. We define four layers, the bottom one containing only one SURF point, the seed, and the top one containing a graph built upon the seed and its 9 nearest neighbours.

Graph comparison

In order to integrate these new graph features in a BoVW framework a dissimilarity measure and a clustering method have to be defined. In this section, we define the dissimilarity measure. We are dealing with attributed graphs, where nodes can be compared with respect to their visual appearance. Although it could be possible to take into account similarities of node features only or the topology of the graph only, more information can be obtained by combining both for defining a dissimilarity measure between local graphs. To achieve this we will investigate the use of the Context Dependent Kernel (CDK), see 4.2.2.

The definition of the CDK relies on two matrices: D which contains the distances between node features, and T which contains the topology of the graphs being compared. Considering two graphs A and B with respective number of nodes m and n, let us denote C the union of the two graphs:

$$C = A \oplus B$$
$$\text{with} \begin{cases} x_i^C = x_i^A & \text{for } i \in [1..m] = I_A \\ x_i^C = x_{i-m}^B & \text{for } i \in [m+1..m+n] = I_B \end{cases} \tag{4.5}$$

with I_A and I_B, the sets of indices of each graph nodes.

The feature correspondence square matrix D of size $(m+n) \times (m+n)$ contains the "entry-wise" L2-norm of the difference between SURF features:

$$D = (d_{ij}) \text{ where } d_{ij} = \left\| x_i^C - x_j^C \right\|_2 \tag{4.6}$$

The square topology matrix T (corresponding to the proximity matrix P in 4.2.2) of size $(m+n) \times (m+n)$ defines the connectivity between two vertices $x_i^C$and x_j^C. In this work we define a crisp connectivity as we set T_{ij} to one if an edge connects the vertices x_i^C and x_j^C and 0 otherwise. Hence, only sub matrices with both lines and columns in I_A or I_B are not entirely null. More precisely, we can define sub matrices T_{AA} and T_{BB} corresponding to the topology of each graph A and B respectively, while sub matrices T_{AB} and T_{BA} are entirely null, vertices of graphs A and B are not connected.

$$T = (T_{ij}) \text{ where } T_{ij} = \begin{cases} 1 \text{ if edge } (x_i^C, x_j^C) \text{ belongs to A or B} \\ 0 \text{ otherwise} \end{cases} \tag{4.7}$$

The CDK denoted K is computed by an iterative process consisting of the propagation of the similarity in the description space according to the topology matrix as detailed in 4.2.2.

Similarly to the definition of sub matrices in topology matrix T we can define sub matrices in the kernel matrix K. The sub matrix $K_{AB}^{(t)}$ represents the strength of the inter-graph links between graphs A and B once the topology has been taken into account. We can therefore define a kernel-wise similarity γ between graphs A and B as:

$$\gamma(A,B) = \sum\nolimits_{\{i \in I_A,\, j \in I_b\}} K_{ij}^{(t)} \in [0,1] \tag{4.8}$$

and induce the dissimilarity as standard kernel distance [188] by evaluating the sum of self-similarity measures of graphs A and B minus twice the cross-similarity between graphs:

$$\rho(A,B) = \gamma(A,A) + \gamma(B,B) - 2\gamma(A,B) \in [0,1] \tag{4.9}$$

This dissimilarity measure will be applied separately on each layer. However, for the bottom layer, since there is no topology to take into account for isolated points we will use directly the "entrywise" L2-norm (4.6). This corresponds to an

approximation of the dissimilarity measure used for graphs features by considering a graph with a single point.

Visual dictionaries and signatures

The state-of-the-art approach for computing the visual dictionary C of a set of features is the use of the K-means clustering algorithm [201] with a large number of clusters, often several thousands, where the code-word is usually the center of a cluster. This approach is not suitable for the graph-features because using the K means clustering algorithm implies iteratively moving the cluster centers with interpolation. Therefore, we use a hierarchical agglomerative (HAG) clustering [190] which does not require graph-interpolation. The median graph G of each cluster V, defined as $median = \underset{G \in V}{argmin} \sum_{i=1}^{m} \|v_i - G\|$, i.e. the graph minimizing the distance to all the graphs v_i of cluster V, represents a code-word.

When targeting object classification on a large database, it can be interesting to use a two pass clustering approach as proposed in [84], as it enables a gain in terms of computational cost. Here, the first pass of the HAG clustering will be run on all the features extracted from training images of one object. The second pass is applied on the centers of clusters generated by the first pass on all objects of the database.

Finally, the usual representation of an image in a BoVW with the dictionary C is built. The BoVW are normalized to sum to one by dividing each value by the number of features extracted from the image. The distance between two images is defined as the L_1 distance between BoVWs (as defined in chapter 3).

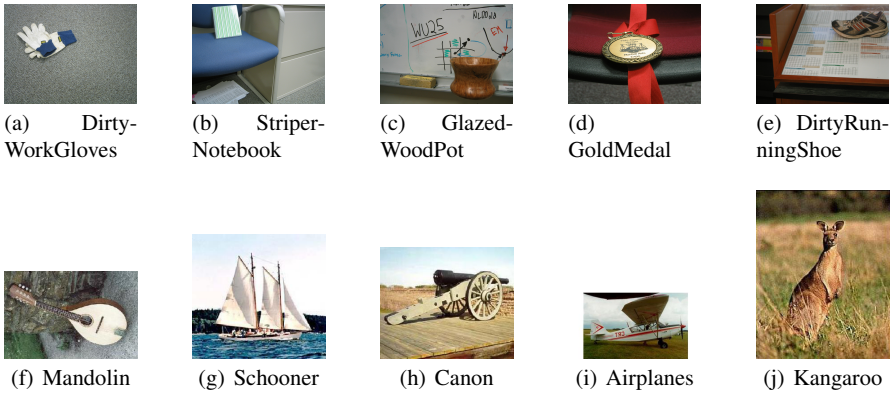
Experiments

The approach is evaluated on publicly available data sets in the problem of object retrieval. The choice of the data sets are guided by the need of annotated objects. We thus chose two datasets.

The SIVAL (Spatially Independent, Variable Area, and Lighting) data set [173] includes 25 objects, each of them being present in 60 images taken in 10 various environment and different poses yielding a total of 1500 images. This data set is quite challenging as the objects are depicted in various lighting conditions and poses. The second one is the well known Caltech-101 [67] data set, composed of 101 object categories. The categories are different types of animals, plants or objects. A snippet of both data sets is shown in Figure 4.2.

Evaluation protocol

We separate learning and testing images by a random selection. On each data set, 30 images of each category are selected as learning images for building the visual



(a) DirtyWorkGloves (b) StriperNotebook (c) GlazedWoodPot (d) GoldMedal (e) DirtyRunningShoe

(f) Mandolin (g) Schooner (h) Canon (i) Airplanes (j) Kangaroo

Fig. 4.2 Excerpts from image data sets. SIVAL (a)-(e), Caltech-101 (f)-(j)

dictionaries and for the retrieval task. Some categories of Caltech-101 have several hundred of images when others have only a few. The testing images are therefore a random selection of the remaining images up to 50. The first pass clustering yields 500 clusters from all the features of all learning images of each object. The final dictionary size varies in the range 50-5000. Details on the experimental setup can be found in [117]. Each layer of graph-features will yield its own dictionary. We compare our method with standard BoVW approach. For that purpose, we use all the SURF features available on all images of the learning database to build the BoVW dictionary by k-means clustering.

The graph features are built only on a selected subset of all SURF points detected in an image. To analyse the influence of this selection, signatures are computed for the set of SURF which have been selected to build the different layers of graphs. These configurations will be referred to as SURF3NN, SURF6NN and SURF9NN corresponding respectively to all the points upon which graphs with 3, 6 and 9 nearest neighbours have been defined.

For each query image and each database image, the signatures are computed for isolated SURF and the different layers of graphs. We have investigated the combination of isolated SURF and the different layers of graphs by an early fusion of signatures i.e. concatenating the BoVWs. For SIVAL this concatenation has been done with the signature from the selected SURF corresponding to the highest level whereas for Caltech-101 we used the classical BoW SURF signature. Finally, the L_1-distance between histograms is computed to compare two images.

The performance is evaluated by the Mean Average Precision (MAP) measure. Here, the average precision metric is evaluated for each test image of an object, and the MAP is the mean of these values for all the images of an object in the test set. For all categories, we measure the performance by the average value of the MAP of objects.

SURF based BoW vs Graphs Words

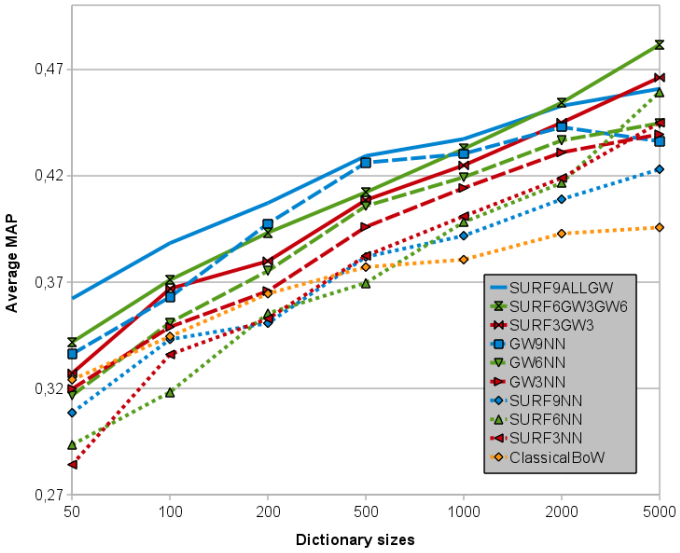


Fig. 4.3 Average MAP on the whole SIVAL data set. Isolated SURF features are the dotted curves, single layer Graphs Words are drawn as dashed curves and the multilayer approach in solid curves.

First of all, it is interesting to analyse if the graph words approach obtains similar performances compared to the classical BoVW approach using only SURF features. This is depicted in Figure 4.3, Figure 4.5 and 4.6 where isolated SURF points are depicted as dotted lines and single layer of graph words are dashed lines. At first glance, we can see that for SIVAL isolated SURF features perform the poorest, separated layers of graphs perform better. Our clustering approach seems to give worse results for very small size of dictionaries but better results for dictionaries larger than 500 visual words, which are the commonly used configurations in BoVW approaches. Each layer of graph words performs much better than the SURF upon which they are built. The introduction of the topology in our features have a significant impact on the recognition performance using the same set of SURF features.

The average performance hides however differences in the performance on some specific objects. To illustrate this we select two object categories where graph features and SURF features give different performances in Figure 4.5 and Figure 4.6. For the object “banana” from SIVAL, the isolated SURF features outperform the graph approach, see Figure 4.5. This can be explained as the “banana” object represents a small part of the bounding box and is poorly textured. In some environments the background is highly textured, this characteristics induce many SURF points detected in it and these SURF points may have a higher response than those detected on the object. This will lead to the construction of many “noisy” graph features on

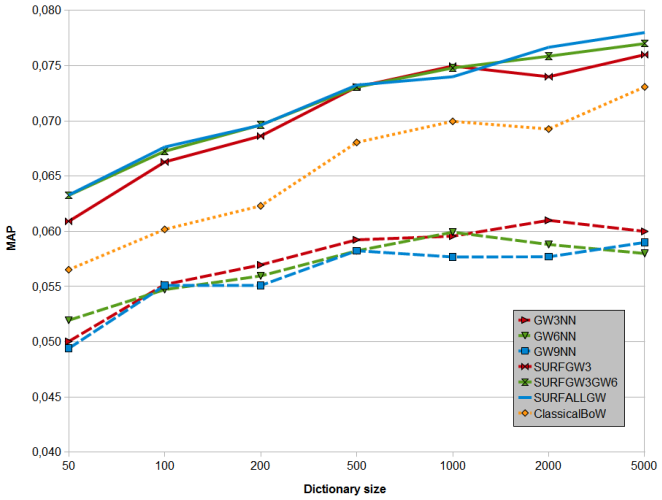


Fig. 4.4 Average MAP on the whole Caltech-101 data set. Isolated SURF features are the dotted curves, single layer Graphs Words are drawn as dashed curves and the multilayer approach in solid curves.

the background and less on the object. On the other hand, for the “Faces” category from Caltech-101 the graph features perform better, see Figure 4.6. Here, the object covers most of the bounding box and many SURF points are detected. In this situation, the graph features capture a larger part of the object than isolated SURF points, making them more discriminative.

This unequal discriminative power of each layer leads naturally to the use of the combination of the different layers in a single visual signature.

The multilayer approach

The combination of graphs and SURF features upon which the graphs have been built is done by the concatenation of the signatures of each layer. The three curves in solid lines in Figure 4.3 correspond to the multilayer approach using only the two bottom layers (SURF + 3 nearest neighbours graphs) depicted with double ”horizontal” triangles, the three bottom layers (SURF + 3 nearest neighbours graphs + 6 nearest neighbours) depicted with double ”vertical” triangles and all the layers depicted by a simple poly-line. For SIVAL, the improvement in the average MAP is clear, and each addition of layer improves the results. The average performance of the combination always outperforms the performance of each layer taken separately.

For Caltech-101, see Figure 4.4, the average MAP values of all methods are much lower which is not surprising as there are much more categories and images. Single layer of graphs gives lower results than the classical BoVW framework on SURF features. However, the combination of all layers outperforms here again SURF or

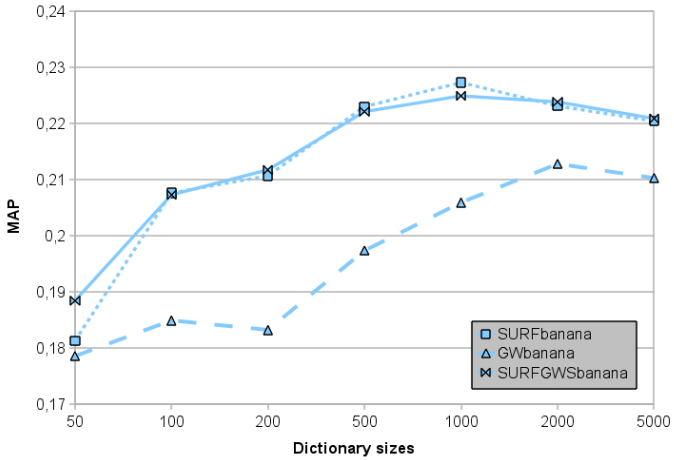


Fig. 4.5 MAP for the object "banana" from SIVAL where isolated SURF features (dotted curves) outperforms graphs (dashed curves). The multilayer approach is the solid curve.

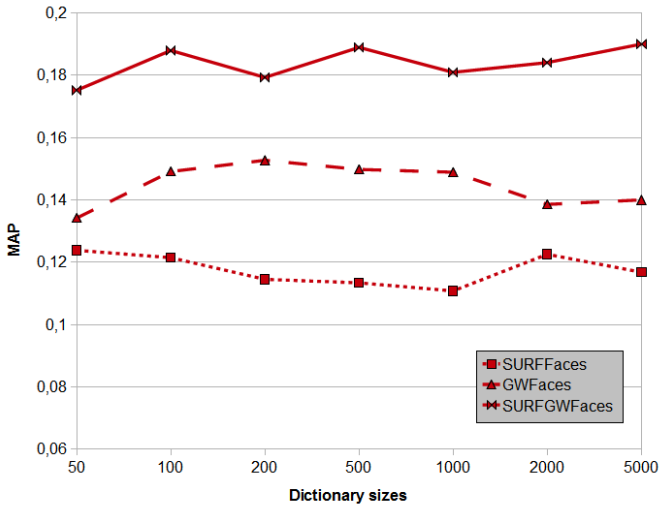


Fig. 4.6 MAP for category "Faces" from Caltech-101 where graphs (dashed curves) outperforms isolated SURF features (dotted curves). The multilayer approach is the solid curves.

graphs used separately. The performance of single layers of graphs can be explained as the fixed number (300) of seeds selection induces for Caltech-101 a strong overlapping of graphs as the average number of SURF points within the bounding box is much lower than for SIVAL. This may give less discriminant graph words as it will be harder to determine separable clusters in the clustering process.

The detailed results presented in Figure 4.5 and Figure 4.6 show that the combination, depicted as a solid line, of the visual signatures computed on each layer separately performs better or at least as well as the best isolated feature.

4.3 Multi-resolution in visual indexing

The images available today on the web are rarely in their raw form as it would require a too huge amount of space to store them. Before being processed, the encoded data is generally decoded. Instead of decoding it completely, some approaches intend to take advantage of the data available in the compressed stream with only partial decoding, thus working in the rough indexing paradigm [139]. Modern standards of visual coding are ”scalable”, which means that in the same code-stream multi-resolution versions of the same content are available. This gives a tremendous opportunity to follow multi-resolution strategy in visual indexing directly using the new low-level features available in code streams. The advantage is obvious. On one hand, the signal will not be deteriorated by double resolution reduction (one when encoding and one when building multi-resolution pyramids on decoded images and videos). On the other hand, computational time savings will be achieved. JPEG2000 standard for images and MJPEG2000 standard for videos have this seducing property of scalability. In this section, we aim at performing image indexing on images encoded in JPEG2000. Following the line of research on rough indexing paradigm, we propose to make use of the multi-resolution information from the wavelet basis and study different techniques to perform indexing in this context.

4.3.1 Low resolution and Rough Indexing Paradigm

When aiming at multi-resolution indexing of visual content, the natural step is to get robust results on the lowest available resolution. The rough indexing paradigm has been introduced in [139] for foreground object extraction purposes. It enables a fast and approximate analysis of multimedia content at a poor resolution. Indeed, it takes advantages of the content directly available in the compressed streams (*e.g.* DC coefficients, motion vectors from video streams, region-based colour segmentation...). It has been used for different applications in the recent years: shot boundary detection [155], object retrieval [43, 44] or video indexing [149]. More specifically, in [149], the HD video is only partially decoded to perform video indexing after detecting moving objects. A similar methodology is presented in [20]. These meth-

ods are all designed for content having the property of scalable representation. In the same line of research, Adami et al. studied a scalable joint data and descriptor encoding of image collections [3]. This work has been extended to videos in [4].

Other works that can be seen as closely related to the rough data processing are those focusing on the analysis of tiny images. Namely, Torralba et al. [208] propose a framework that allows performing object recognition in a huge database (tens of millions images). To that end, they directly process low-resolution 32×32 color images. This low resolution could correspond to the coarsest level of coded images.

In this chapter, we only focus on image indexing and do not address the video indexing problem. The rough data then only consists of partially decoded colour/intensity information.

4.3.2 Multi-resolution and multiscale in image indexing.

For raw data the multi-resolution comes from the construction of image pyramids (Gaussian pyramids for instance), whereas for encoded data (such as JPEG2000 images) it is directly available from the wavelet decomposition. As always, these techniques rely on the computation of local or global descriptors on the resulting multi-resolution multi-scale pyramids.

Global descriptors are generally based on computation of histograms. The use of multi-resolution histograms for recognition was first proposed in [90]. The multi-resolution decomposition is computed with Gaussian filtering. A filtered image $I * \mathcal{G}(l)$ is the result of the convolution of the image I with the Gaussian filter:

$$\mathcal{G}(l) = \frac{1}{2\pi l\sigma^2} exp\left(-\frac{x^2+y^2}{2l\sigma^2}\right),$$

where σ is the standard deviation of the filter and l is the resolution.

An example of the use of histograms in the rough indexing paradigm can be found in [149]. This paper addresses the problem of scalable indexing of HD videos encoded in the MJPEG2000 standard. After detecting moving objects, their indexing is performed. A global descriptor is built for the object. It consists of a pair of two histograms:

$$H = \{h^k_{LL}, h^k_{HF}, k = 1 \ldots K\},$$

where K is the number of levels in pyramid defined in JPEG2000. The first histogram h_{LL} is the YUV joint histogram of LL coefficients. The second one, h_{HF}, is computed from the High Frequency (*i.e.* HL, LH and HH) sub-band. Each sub-band represents a different orientation: HL horizontal, LH vertical and HH diagonal. The histogram h_{HF} is finally the histogram of mean absolute values of coefficients LH, HL and HH. Hence the invariance to rotation to multiple of 45° is obtained.

Amongst few multi-resolution approaches existing today, the most known is the so-called spatial pyramid matching (SPM) [126], also referenced in chapter 3. BoVW are built on nested partitions of image plane from coarse-to-fine. Neverthe-

less, this approach cannot be qualified as a truly "multi-resolution", as the features and level image descriptors (sparse SIFT) are built only on the full resolution image. The adaptation of SPM to Gaussian scale space has been proposed in [193]. Spatial pyramids are computed at different scales which allow combining different levels of details.

In our approach, on the contrary, we aim to incorporate multi-scale representation of image content in the whole feature extraction, quantization and matching process.

4.3.3 Multi-resolution features and visual dictionaries

Based on the rationale we expressed before we present and analyze some approaches for indexing images encoded in JPEG2000 using its "natural" multi-scale representation of visual content. From the Daubechies pyramid, we only use the LL sub-band at all decomposition levels. All the methods described could also directly be used on uncompressed data by computing a multi-resolution pyramid by standard Gaussian filtering and sub-sampling. Here, we focus only on methods inspired from the BoVW and the SPM approaches.

Multi-resolution approach on wavelet pyramids

From the colour Daubechies 9/7 pyramid as defined in JPEG2000, we extract only the Y component of the LL sub-band at $K = 3$ levels of the pyramid. In the following, we denote the different levels of pyramid by k, $k = 1 \dots K$.

Our approach follows the BoVW scheme with

- level features which are SURF points and descriptors extracted at each level Y_{LL}^k in the wavelet pyramid. We denote a set of features at level k by $\mathcal{D}^k$.
- visual dictionaries we build per level, denoted by C^k, and for all levels together denoted by C. The number of visual words varies from 50 to 5000. Every visual dictionary we refer to is constructed by applying the k-means++ algorithm [8] on the training set (see section 4.3 for more details on training sets).
- image signature which is a histogram of visual words from C^k, denoted by H^k, and built for Y_{LL}^k or a histogram H built for all levels together Y_{LL}^k, $k = 1 \dots K$ with the global dictionary C.

Hence, the descriptor of an image is the histogram of visual words from appropriate dictionary. To compare the images at different resolution levels in wavelet domain, we use the histogram intersection kernel as a similarity measure. For the BoVWs of two images at level k, this function is given by:

$$\mathcal{I}(H_1^k, H_2^k) = \sum_{i=1}^{N} \min(H_1^k(i), H_2^k(i)). \tag{4.10}$$

N being the number of visual words (vocabulary size).

The proposed description schemes are used for image retrieval and classification. For the simple retrieval scenario, images are ranked according to histogram intersection similarity (4.10) with regard to a query image. The mean average precision is then computed to evaluate the methods. Classification of images is performed with a supervised learning framework with multi-class support vector machine (SVM) using the libSVM library [36] and a *one versus all* rule. For each scheme, a kernel inspired by the pyramid match kernel [86] is provided to the SVM.

As in previous section, all the results presented here are on the two datasets: Caltech-101 and SIVAL. The SURF features are only extracted in a bounding box around the object of interest. The coordinates of the bounding box can be downloaded together with the set of images. The mean average precisions and classification rates on both datasets and for all the methods we present here are presented in Table 4.1, 4.3 and 4.2, 4.4 respectively.

Merging the information at different resolutions

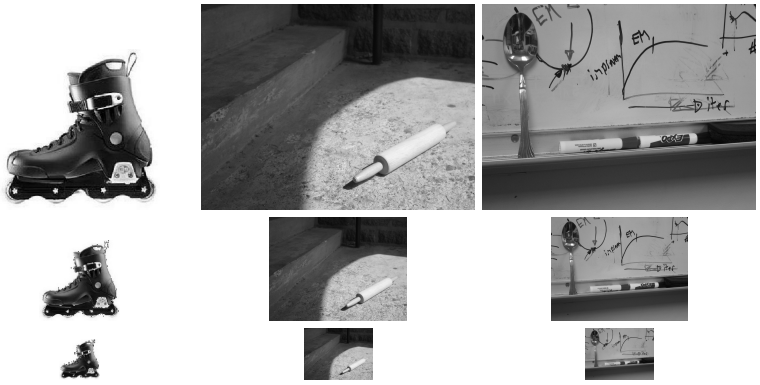
The direct application of BoVW in the context of the rough indexing paradigm consists in applying the BoVW method at the coarsest level ($k = K$) of the wavelet pyramid. Nevertheless the image of low frequency coefficients at this level, Y_{LL}^K, is obviously very blurry and does not contain many interest points. Many important details are lost. The induced visual dictionary C^K and the corresponding signatures H^K are therefore not enough informative. We have tested the application of BoVW at each level independently. The results are visible in the second, third and fourth columns of the four tables at the end of this section. At the finest scale ($k = 1$) it corresponds to applying the BoVW method to the original full-resolution gray-scale image. The classification rates in Table 4.2 and 4.4 come from the following kernel:

$$\kappa_{BoVW}^k(X,Y) = \mathcal{I}(H_X^k, H_Y^k). \tag{4.11}$$

These results permit to confirm that working at the finest scale is more efficient than working at coarsest scales. In particular the number of relevant documents retrieved significantly decreases when using only the information at level $k = 3$. When looking more into details at the precision values, we observed that for some images, processing the coarsest level could improve the results. For instance, for the class *inline_skate* of Caltech 101 the map is 0.09 at level $k = 3$ against 0.02 at level $k = 1$. Similarly for the class *woodrollingpin* of SIVAL, the map is 0.19 at level $k = 3$ against 0.13 at level $k = 1$. Examples of images for which the same conclusion can be drawn are presented in Figure 4.7. A natural extension of mono-level approach is to try to combine information from different levels of a multi-resolution pyramid in the same way we combine structural graph words in the the "early fusion" manner.

Our first attempt has then been to concatenate the histograms at different resolutions H^k, $k = 1 \dots K$ into a unique signature, $\tilde{H}$:

$$\tilde{H} = \cup_{k=1\dots K} H^k.$$



inline_skate, Caltech-101 *woodrollingpin*, SIVAL *largespoon* , SIVAL

Fig. 4.7 Images leading to better results for BoVW at coasest scales. First row: One image from each class at level $k = 1$. Second row: level $k = 2$. Third row: level $k = 3$.

The dimensionality of this vector is $(K.N)$. The same importance is given to all the resolutions so that the concatenation is not weighted. As a reference to BoVW method, we will now refer to this approach as mBoVW (multi-resolution BoVW). The kernel used for classification is given by:

$$\kappa_{mBoVW}(X,Y) = \sum_{k=1}^{K} \mathcal{I}(H_X^k, H_Y^k) = \mathcal{I}(\tilde{H}_X, \tilde{H}_Y). \tag{4.12}$$

It is worth mentioning that this kernel is not related to the pyramid match kernel [86]. Indeed, at each multi-resolution level, the dictionary is different. Results of this method are presented in the sixth column of the different tables. It can be seen that, even if we can find several classes for which it does improve the results compared to the BoVW at level $k = 1$, it globally deteriorates the classification rates and the mean average precision for both databases.

A comparison to SPM [126] is provided in the fifth column. This method has been implemented with three scales (spatial resolution), $L = 3$: 21 histograms, $\mathcal{H}^l$, $l = 0 \ldots \sum_{l=0}^{L-1} 4^l$ are representing each image. The kernel is:

$$\kappa_{SPM}(X,Y) = \sum_{i=1}^{N} \left(\frac{1}{2^L} \mathcal{I}(\mathcal{H}_{X_i}^0, \mathcal{H}_{Y_i}^0) + \sum_{l=1}^{L-1} \frac{1}{2^{L-l+1}} \mathcal{I}(\mathcal{H}_{X_i}^l, \mathcal{H}_{Y_i}^l) \right). \tag{4.13}$$

The application of SPM to the two datasets we are studying lead to opposite conclusions. While it deteriorates the result on the SIVAL database, compared to the standard BoVW at level $k = 1$, an improvement appears on Caltech-101. The main reason for this is that the different objects of SIVAL have been acquired with the same background. It means that SPM is not the best choice to differentiate objects

in the same environment, especially when an object is not in its usual environment (loss of context).

To be complete we also merged the two previous methods (mBoVW and SPM) into a common framework called multi-resolution spatial pyramid matching (mSPM). At each resolution of the wavelet pyramid, a spatial pyramid is built. K histograms of dimension $(N\sum_{l=0}^{L-1}4^l)$ are indeed computed for each image (see figure 4.8). Once again, each resolution is considered independent:

$$\kappa_{mSPM}(X,Y)=\sum_{k=1}^{K}\sum_{i=1}^{N}\left(\frac{1}{2^L}\mathcal{I}(H_{X_i}^{0,k},H_{Y_i}^{0,k})+\sum_{l=1}^{L-1}\frac{1}{2^{L-l+1}}\mathcal{I}(H_{X_i}^{l,k},H_{Y_i}^{l,k})\right). \quad (4.14)$$

As mBoVW was degrading the results of BoVW, it is not surprising to observe that mSPM also deteriorates the results of SPM.

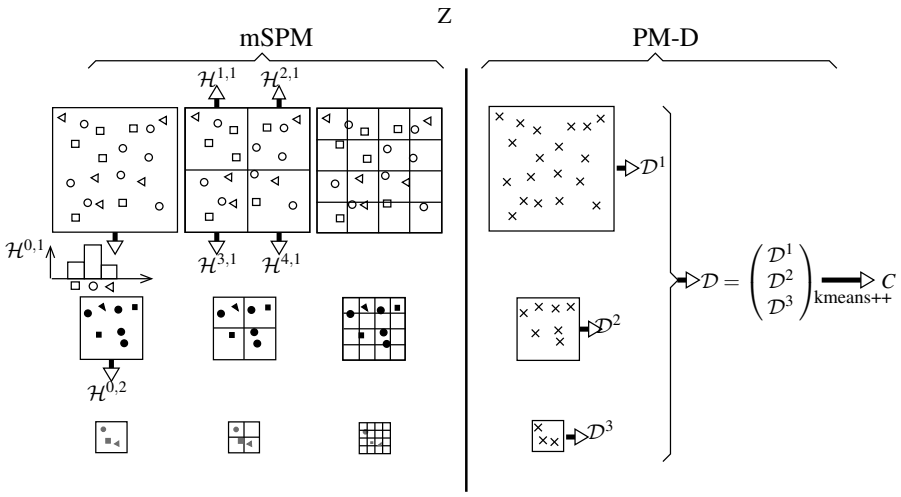


Fig. 4.8 Illustration of the different multi-resolution methods.

Adding the multi-resolution by early fusion of BoVW histograms from each level k is not optimal. Therefore, we elaborated a different strategy for merging the information at different resolutions. Until now, each level k was assigned one dictionary C^k. Here we propose to consider one dictionnary C that is common to all levels. It is computed by using all sets of descriptors $\{\mathcal{D}^k\}_{k=1\ldots K}$. By taking into account all the descriptors at all levels together a more complete vocabulary can be obtained. For each image, the set of all available features is considered: $\mathcal{D}=\cup_{k=1\ldots K}\mathcal{D}^k$. The unique dictionary C is then obtained by clustering this unique set. Each image is finally represented by a unique signature H that incorporates directly the information from all levels. We call this approach Pyramid Matching with Descriptors (PM-D).

The kernel is the same as the one used for the classical BoVW method (equation (4.11)).

Its extension to spatial pyramid matching (SPM-D) is also shown in the tables. In this case, the position of the points at coarsest levels are projected to the finest level before partitioning the space. The results obtained using the combination of the descriptors with these two last methods (PM-D and SPM-D) are the most promising ones on both datasets.

N	BoVW k=0	BoVW k=1	BoVW k=2	SPM [126]	mBoVW	mSPM	SPM-D	PM-D
50	**0.326**	0.279	0.208	0.088	0.268	0.084	0.087	0.323
100	**0.346**	0.299	0.218	0.095	0.281	0.091	0.094	0.345
200	0.354	0.31	0.236	0.104	0.3	0.099	0.102	**0.355**
500	0.364	0.334	0.254	0.12	0.33	0.113	0.118	**0.366**
1000	**0.373**	0.36	0.262	0.135	0.35	0.124	0.132	0.37
2000	0.397	0.383	0.28	0.153	0.377	0.137	0.152	**0.4**
5000	**0.439**	0.413	0.289	0.192	0.401	0.16	0.199	0.436

Table 4.1 Mean Average Precision for the SIVAL dataset

N	BoVW k=0	BoVW k=1	BoVW k=2	SPM [126]	mBoVW	mSPM	SPM-D	PM-D
50	**81.2**	77.33	65.73	56.8	78.53	49.33	56.13	80
100	86.67	85.33	75.07	61.2	82.67	54.27	61.6	**89.46**
200	**91.6**	88.4	80.27	63.47	85.47	56.8	65.47	91.07
500	93.07	92	86.67	68.8	90.53	60	69.73	**93.73**
1000	94.67	93.47	88.67	71.2	90.8	61.07	72.4	**96**
2000	95.73	94.67	91.47	73.87	91.2	60.53	74	**96**
5000	96.67	95.07	94.27	74.4	93.33	57.73	75.47	**97.33**

Table 4.2 Classification rates (%) for the SIVAL dataset

N	BoVW k=0	BoVW k=1	BoVW k=2	SPM [126]	mBoVW	mSPM	SPM-D	PM-D
50	0.057	0.041	0.024	0.077	0.038	0.076	**0.08**	0.059
100	0.064	0.047	0.025	0.082	0.043	0.081	**0.086**	0.065
200	0.068	0.05	0.025	0.086	0.045	0.085	**0.09**	0.07
500	0.073	0.052	0.026	0.088	0.048	0.086	**0.094**	0.076
1000	0.077	0.054	0.026	0.088	0.047	0.084	**0.094**	0.081
2000	0.055	0.052	0.027	0.086	0.046	0.081	**0.094**	0.084
5000	0.08	0.052	0.029	0.081	0.045	0.079	**0.089**	0.086

Table 4.3 Mean Average Precision for the Caltech-101 dataset

N	BoVW k=0	BoVW k=1	BoVW k=2	SPM [126]	mBoVW	mSPM	SPM-D	PM-D
50	24.96	21.93	11.04	45.26	20.92	43.09	**45.74**	24.52
100	28.28	25.16	11.41	45.87	23.5	43.9	**47.2**	29.78
200	31.17	27.33	11.85	46.76	24.69	44.58	**47.5**	31.95
500	34.84	28.56	13	45.94	26.49	42.58	**47.23**	35.38
1000	36.26	28.9	12.66	42.78	26.59	38.34	**45.06**	37.35
2000	37.52	28.56	12.56	38.91	26.79	30.46	**41.09**	37.72
5000	37.52	28.59	13.62	27.77	25.29	22.99	31.38	**38.61**

Table 4.4 Classification rates (%) for the Caltech-101 dataset

4.4 Conclusion

In this chapter we were interested in two aspects in visual indexing: the incorporation of spatial context and of multi-resolution/multi-scale strategies in the state-of-the art BoVW approaches. Analysis of the performance of the methods on publicly available databases in both approaches converge to the same conclusions: incorporating information from spatial neighbourhood or from the multi-resolution pyramids into visual content description improves performances. Indeed, in both cases fusion of information coming from different nested layers of local graphs or from different layers of content resolution does bring an improvement in terms of Mean Average Precision (MAP) and classification rates. Obviously the visual scenes/objects to recognize have to be sufficiently rich in terms of quantity of potential characteristic points to ensure a statistical soundness of built visual dictionaries. In the GraphWords approach mixing all BoVWs from singular interest points and local graphs with increasing number of nodes in one description space shows better performances than a "single layer" BoVWs. In the multi-resolution/multi-scale approach building only one dictionary for all levels together is better than building one dictionary per level. In other words, combining the features extracted at different levels of resolution gives the most promising results.

These approaches are far from being totally exhausted. In the GraphWords approach a promising perspective for handling structural deformations of graphs due to occlusions is in the spatial weighting of node features. In a multi-resolution context, intelligent weighting schemes are also needed to tune the importance of local saliencies at different resolution levels. Another perspective is in the use of colour. Indeed the descriptors considered, such as the SURF features, reflect only the "textural" content in the vicinity of characteristic points. The colour has not been considered yet. An interesting way to do it in our vision is to make usage of the local support related to the graphs or to the SURF points themselves. One of the possibilities is in the use of dense features as done in [126]. Furthermore, a direct way of combining both spatial context and multi-resolution would be in a definition of a strategy of combining the layers in graphs with resolution levels in pyramids. Hence the visual content can be indexed with the degree of detail in structure corresponding to its spatial resolution. Furthermore, the use of the high frequency coefficients in the

wavelet pyramid can yield computationnaly interesting alternatives to the state-of-the-art SURF and SIFT descriptors in the combined global framework.

Acknowledgements This work was partially supported by ANR 09 BLAN 0165 IMMED grant