Nested Graph Words for Object Recognition

Svebor Karaman, Jenny Benois-Pineau, Rémi Mégret

Nested Graph Words for Object Recognition

Svebor Karaman
LaBRI, University of Bordeaux
351, Cours de la Libération
33405 Talence Cedex
+33(0)5 4000 3880
svebor.karaman@labri.fr

Jenny Benois-Pineau
LaBRI, University of Bordeaux
351, Cours de la Libération
33405 Talence Cedex
+33(0)5 4000 8424
jenny.benois@labri.fr

Rémi Mégret
IMS, University of Bordeaux
351, Cours de la Libération
33405 Talence Cedex
+33(0)5 4000 3621
remi.megret@ims-bordeaux.fr

ABSTRACT
In this paper, we propose a new, scalable approach for the task of object based image search or object recognition. Despite the very large literature existing on the scalability issues in CBIR in the sense of retrieval approaches, the scalability of media and scalability of features remain an issue. In our work we tackle the problem of scalability and structural organization of features. The proposed features are nested local graphs built upon sets of SURF feature points with Delaunay triangulation. A Bag-of-Visual-Words (BoVW) framework is applied on these graphs, giving birth to a Bag-of-Graph-Words representation. The nested nature of the descriptors consists in scaling from trivial Delaunay graphs - isolated feature points - by increasing the number of nodes layer by layer up to graphs with maximal number of nodes. For each layer of graphs its proper visual dictionary is built. The experiments conducted on the SIVAL data set reveal that the graph features at different layers exhibit complementary performances on the same content. The nested approach, the combination of all existing layers, yields significant improvement of the object recognition performance compared to single level approaches.

Categories and Subject Descriptors
I.4.7, I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis – *Object recognition,* Feature Measurement – *Feature representation.*

General Terms
Algorithms, Measurement, Design, Experimentation.

Keywords
Nested features, Bag-of-Visual-Words, Graph Words, Delaunay triangulation, Context Dependent Kernel.

1. INTRODUCTION
Visual object retrieval in images and videos is one of the most active field of research in the community. One important aspect of the most popular techniques addressing this task are the use of local features, SIFT (Scale Invariant Feature Transform) of Lowe [1] or SURF (Speed-Up Robust Features) of Bay [3] key points for instance. SIFT and SURF key points are robust and discriminative local features. SIFT points are detected on local minima/maxima of a Difference of Gaussians (DoG) image computed at different scales. The SIFT key point feature is a orientation histogram in a close spatial neighborhood of the key point. SURF is based on sums of approximated Haar wavelet responses and use integral images in order to speed-up key points extraction.

In the trending approach of Bag-of-Visual-Words [2], the features are quantized in visual dictionaries by clustering and images are depicted as a distribution of the visual words within them. The Bag-of-Visual-Words approach is an adaptation of the text retrieval approach Bag-of-Words (BoW) to images. The BoVW operates on local key points when BoW operates on words, the semantic power of a word is much higher than one's of a local key point. A visual word is also much more ambiguous than a text word. Moreover, the BoVW approach discard all spatial information about the relations between key points. Having a similar neighborhood of key points in two images indicates stronger similarity of content than sparse isolated key points.

To overcome this limitation of the BoVW, some approaches have been presented in the past few years. The spatial pyramid matching proposed in [4] and used in [5] and [6] for its application to object or scene recognition enables to compare area generated from arbitrary splitting instead of the whole image. In [7] an approach called "Visual Phrases" is used to group visual words according to their proximity in image plane. The visual phrases are represented by a histogram containing the distribution of the visual words in the phrase. In these works, the common idea is to build local signature according to a visual dictionaries from an arbitrary splitting for the spatial pyramid matching or on a set built by a proximity criterion for visual phrases. Our approach is introducing the local topological information within the visual features.

In this paper we propose a spatial embedding of features with local Delaunay graphs. The motivation for building such graphs comes from the invariance of Delaunay triangulation with regard to affine transformations of image plane: rotation, translation and scale. Hence, with invariant key point features such as SURF the global invariance of graphs is maintained. We plunge these graph features into a Bag-of-Visual-Words framework, building visual dictionaries by clustering graphs. Then the state-of-the-art visual signatures are used for object retrieval. Increasing the number of nodes in graphs yields a layered approach where each layer induce a stronger spatial embedding within graph features. We call this approach "nested" .It combines visual signatures of all graphs from trivial graphs which are isolated SURF points to the largest graphs we work with.

The paper is organized as follows, in section 2 we discuss the process of building these graphs and introduce their nested construction. In section 3, we introduce the dissimilarity metric used to compare graphs and built visual dictionaries by clustering. The latter are presented in section 4. Experiments with these new features are presented in section 5. Conclusions and perspectives are given in section 6.

2. GRAPH FEATURE CONSTRUCTION
Let us consider a graph $G=(X,E)$ with X a set of nodes corresponding to some feature points $x_{k,k=1,..,K}$, in image plane and $E=\{e_{kl}\}_{,k=1,..,K,l=1,..,K}$, where $e_{kl}=(x_k,x_l)$, a set of edges connecting these points. We call such a graph a "graph feature". We will build these features upon sets of neighboring feature points in

image plane. Hence we propose a spatial embedding of local features with graphs. To build such graphs two questions have to be addressed : i) the choice of feature points sets $\boldsymbol{X}$ and ii) the design of connectivity as edges $\boldsymbol{E}$.

To define the feature points sets $\boldsymbol{X}$ upon which graphs will be built we are looking for a set of feature points that we call the "seeds". Around them other feature points will be selected to build each graph feature. One could consider all detected SURF points as seeds, however that would lead to a strong overlapping of graphs and would induce high redundancy in the set of graph features created. Therefore, the selected seeds have to form a set of SURF points which are more likely to be detected in various instances of the same object. This selection can be done using a criterion on the "size" of the key point or on the octave in which it has been detected. This would mean giving priority to SURF points detected in higher scale or lower scale which has no real significant impact on their repeatability. The best criterion for this task is the response of the SURF key points, which is the approximated determinant of the Hessian. SURF points are detected where local maxima of the approximated Hessian determinant are reached [3]. SURF points with higher response are then more likely to be more repeatable. Hence, the seeds considered for building the graphs will be the SURF points with highest responses. Considering a fixed number of seeds $\boldsymbol{N_{Seeds}}$, we can define the set of seeds $\boldsymbol{S}$:

$$S = \left\{ s_1, \ldots, s_{N_{Seeds}} \right\}$$

Given $\boldsymbol{S}$, 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 $\boldsymbol{k}$ spatial nearest SURF neighbors of each seed $\boldsymbol{s_i}$ gives the set of neighbors $\boldsymbol{P_i}$:

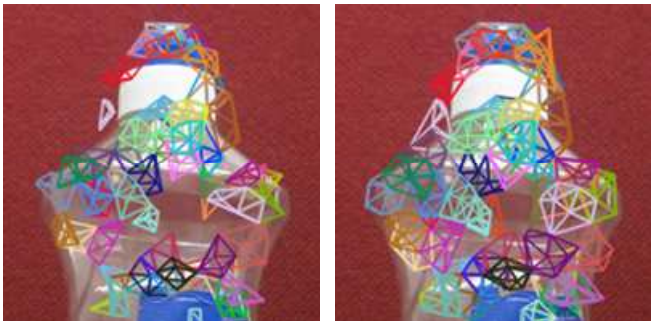
$$P_i = \left\{ p_1, \ldots, p_k \right\}$$

Hence the set of nodes for each graph upon a seed point is built. For the edges we use the Delaunay triangulation which is invariant with regard to affine transformations of image plane preserving angles: translation, rotation and zoom. Furthermore, regarding the future extensions of this work to video, the choice of Delaunay triangulation is also profitable for its good properties in tracking of structures [8]. The set of all vertices used for building the graph $\boldsymbol{G_i}$ is $\boldsymbol{X^{Gi}}$, the union of the seed and its neighborhood:

$$X^{Gi} = \left\{ x_1^{Gi}, \ldots, x_k^{Gi} \right\} = P_i \cup \{ s_i \}$$

A Delaunay triangulation is computed on the points of $\boldsymbol{X^{Gi}}$, building triangles according to the Delaunay constraint. An edge $\boldsymbol{e_{ij}=(x_i^{Gi},x_j^{Gi})}$ is defined between two vertices of the graph $\boldsymbol{G_i}$ if an edge of a triangle connects these two vertices.

(a) SURF Features (b)3-nearest neighbor graphs



(c) 6-nearest neighbor graphs (d) 9-nearest neighbor graphs

Figure 1: SURF and graph features on a cropped image of the object ajaxorange from SIVAL database.

Introducing a layered approach, where each layer adds more structural information we can define graphs of increasing size while moving from one layer to the upper one. To avoid a large number of layers, the number of nodes added at each layer should induce a significant change of structural information. To built Delaunay triangulation, at least two points have to be added to the seed at the second layer. 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 neighbors, see examples in Figure 1. One layer will always contain the points of all the lower layers, hence we call this approach "nested" and illustrate it in Figure 2.

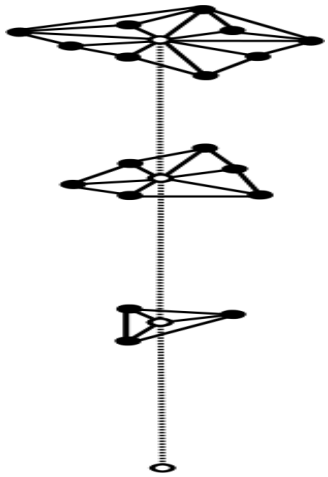


Figure 2: The nested approach. Bottom to top: SURF seed depicted as the white node, 3 neighbors graph where neighbours are in black, 6 neighbors graph and 9 neighbors graph.

3. GRAPH COMPARISON

In order to integrate these new graph features in a Bag-of-Visual-Words framework a dissimilarity measure and a clustering method have to be defined. In this section, we define the dissimilarity measure.

When defining such a measure, the (dis)similarity of node features only or only topology could be considered. However, it should be much more interesting to take into account both SURF

descriptors – node features- and graph topology. To achieve this we will investigate the use of the Context Dependent Kernel (CDK) presented in [9]. The definition of the CDK relies on two matrices: $\boldsymbol{D}$ which contains the distances between node features, and $\boldsymbol{T}$ which contains the topology of the graphs being compared.

Considering two graphs $\boldsymbol{A}$ and $\boldsymbol{B}$ with respective number of nodes $\boldsymbol{m}$ and $\boldsymbol{n}$, let us denote $\boldsymbol{C}$ the union of the two graphs:

$$C = A \cup B$$

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

The feature correspondence square matrix $\boldsymbol{D}$ of size $(m+n)x(m+n)$ contains the "entrywise" L_2-norm (i.e., the sum of the square values of vector coefficients) of the difference between SURF features:

$$D = (d_{ij})_{ij}$$

$$where \ d_{ij} = \left\| x_i^C - x_j^C \right\|_2$$

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

$$T = (T_{ij})_{ij}$$

$$where \ T_{ij} = \begin{cases} 1 \ if \ \exists e(x_i^C, x_j^C) \\ 0 \ otherwise \end{cases}$$

The CDK denoted $\boldsymbol{K}$ is computed by an iterative process consisting of the propagation of the similarity in the description space according to the topology matrix.

$$K^{(0)} = \frac{\exp(-\frac{D}{\beta})}{\left\| \exp(-\frac{D}{\beta}) \right\|_1} \quad , \quad K^{(t)} = \frac{G(K^{(t-1)})}{\left\| G(K^{(t-1)}) \right\|_1}$$

$$G(K) = \exp(-\frac{D}{\beta} + \frac{\alpha}{\beta} T K^{(t-1)} T)$$

Where exp represents the coefficient-wise exponential and $\boldsymbol{||M||_1 = \Sigma_{ij}|M_{ij}|}$ represents the L1 matrix norm. Similarly to the definition of sub matrices in topology matrix $\boldsymbol{T}$ we can define sub matrices in the kernel matrix $\boldsymbol{K}$. The sub matrix $\boldsymbol{K_{AB}^{(t)}}$ represents the strength of the inter-graph links between graphs $\boldsymbol{A}$ and $\boldsymbol{B}$ once the topology has been taken into account. We can therefore define the dissimilarity measure that will be used for clustering:

$$s(A,B) = \sum_{\{i \in I_F, j \in I_G\}} K_{ij}^{(t)} \in [0,1]$$

$$\rho(A,B) = s(A,A) + s(B,B) - 2s(A,B) \in [0,1]$$

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" L_2-norm of the difference between SURF features as an approximation. This approximation is valid if we consider high enough values for $\boldsymbol{\beta}$, see Appendix for details.

4. VISUAL DICTIONNARIES

The state-of-the-art approach for computing the visual dictionary of a set of features is the use of the K-means clustering algorithm with a large number of clusters, often several thousands. The code-word is either the center of a cluster or a non-parametric representation like a K-Nearest Neighbors (K-NN) voting approach.

Both of these approaches are not suitable for the graph-features as using the K-means clustering algorithm implies iteratively moving the cluster centers with interpolation whereas defining a mean graph is a difficult task; and a fast K-NN needs an indexing structure which is not available in our graph feature space since it is not a vector space. Therefore, we present in the following section the selected method which is a two pass agglomerative hierarchical clustering. The model of a cluster is chosen as its median instead of the mean.

4.1 Clustering method

In order to quantize a very large database, it can be interesting to use a two pass clustering approach as proposed in [10], as it enables gain in terms of computational cost. Here, the first pass of the agglomerative hierarchical clustering will be run on all the features extracted from training images of one object. The second pass is fulfilled on clusters generated by the first pass on all objects of the database. To represent a cluster, we use the following definition of the median:

$$\text{median} = \underset{G \in V}{\text{argmin}} \sum_{i=1}^{m} \left\| v_i - G \right\|$$

With $\boldsymbol{V}$ – a cluster and $\boldsymbol{v_i}$ – members of a cluster, G the candidate median and $\| \ \|$ is a distance or dissimilarity measure in our case.

For the first pass, the dissimilarities between all the features, of the same layer, extracted on all the images of an object are computed. For the second pass, only the dissimilarities between all the medians of all object clusters are computed. Each layer being processed independently, we obtain a visual dictionary per layers of graphs with 1, 3,…, N_{max} nodes.

4.2 Visual signatures

The usual representation of an image in a BoVW approach is to compute a histogram of all the visual words of the dictionary within the image. We use this representation without rejection, a feature is always assigned to the closest word in the dictionary.

Once the visual signatures of images have been computed, one can define the distance between two images as the distance between their visual signatures. In preliminary experiments we have compared results when using Hamming distance, Euclidean distance and L_1 distance for this task. The L_1 distance giving better results, final results are presented using this measure only.

5. EXPERIMENTS

The SIVAL (Spatially Independent, Variable Area, and Lighting) data set [11] includes 25 objects, each of them being present in 60 images taken in 10 various environment and different poses. This data set is quite challenging as the objects are depicted in various lighting conditions and poses, see a snippet of the data set in Figure 3. It has also been chosen as the future of this works is the recognition of objects of the daily living that may appear in different places of a house, like a hover that may be moved in all the rooms in one's house.

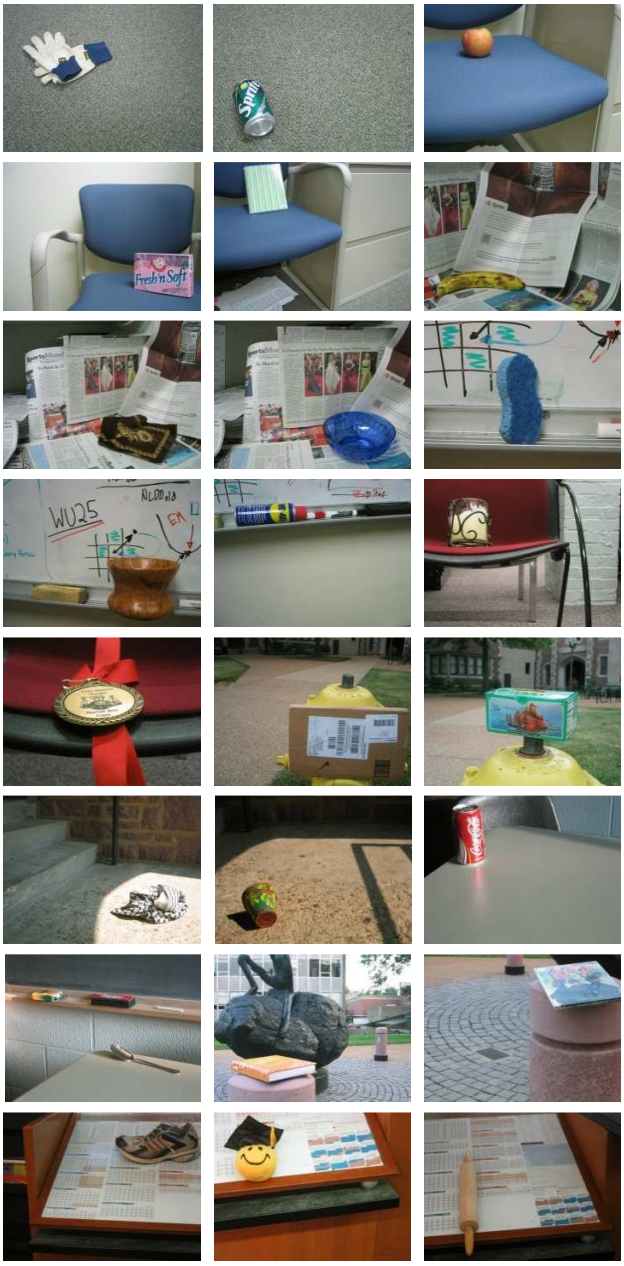


Figure 3: The SIVAL data set.

We separate learning and testing images by a random selection of half of the images for each object. We only take into account the content of a bounding box of each object as the aim of this paper is only object recognition and not yet localization. SURF key points are extracted within the bounding box, the numbers of seeds for the graphs building process is fixed to 300. The second layer corresponds to graphs built upon the seeds and their 3 nearest neighbors, the third layer with the 6 nearest neighbors and the fourth and last layer with the 9 nearest neighbors. For the CDK, α is set to 0.0001, β to 0.1 (ensuring **K** is a proper kernel) and the number of iterations is fixed to 2, H. Sahbi [9] has shown that the convergence of the CDK is fast. The first pass clustering compute 500 clusters for each object. The final dictionary size varies in the range 50-5000. Each layer will yield its own dictionary.

The performance is evaluated by the Mean Average Precision (MAP) measure. For each test images, all images in the learning set are ranked from the closest (in terms of L_1 distance between visual signatures) to the furthest. The average precision 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.

5.1 SURF vs Graphs

First of all, it is interesting to analyze if the graph words approach where each layer is taken into consideration separately obtains similar performances compared to the classical BoVW approach using only SURF features. This is depicted in Figure 4. Here we can observe a really similar behavior between isolated SURF features (dotted lines) compared to a single layer of graph words (dashed lines) even if the approach using 9 nearest neighbor graphs seems to slightly outperform the others. The decrease in performance on the right side of the curve can be imputed to the two pass agglomerative clustering. Indeed as we selected 500 clusters for each object in the first pass, the second pass has only 12500 points to build 5000 clusters in our experimental set-up. This will conduct to an over segmentation of the dictionary where similar features are not grouped in the same clusters. The decrease on the left side is due to a too small dictionary size where not enough words are discriminative between objects.

This similar average performance hides however some very important differences in the performance of each feature on some specific objects. To illustrate this we select four objects categories where graphs features and SURF give different performances in Figure 5 and Figure 6. For the objects "apple" and "spritecan", the isolated SURF features outperform the graph approach, see Figure 5, whereas for the "wd40can" and "feltflowerrug" objects the graphs features perform better, see Figure 6. This unequal discriminative power of each layer leads naturally to the use of the combination of the different layers in a single visual signature.

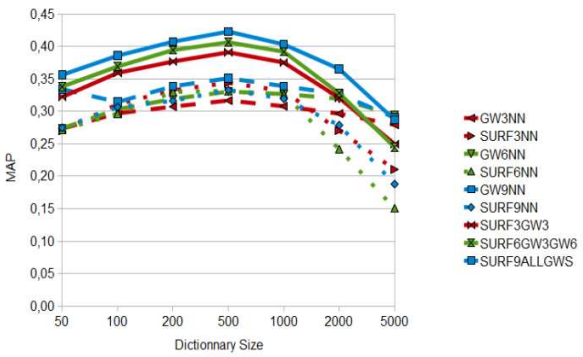


Figure 4: Average MAP on the whole SIVAL data set. Isolated SURF features are the dotted curves, single layer Graphs

Words are drawn as dashed curves and the nested approach in solid curves.

5.2 The nested approach

The combination of graphs and SURF features is done by the concatenation of the signatures of each layer. The three curves in solid lines in Figure 4 correspond to the nested approach using only the two bottom layers (SURF + 3 nearest neighbors graphs) in red, the three bottom layers (SURF + 3 nearest neighbors graphs + 6 nearest neighbors) in green and all the layers in blue. The improvement in the average MAP is really significant, and each addition of layer improves the results.

The detailed results presented in Figure 5 and Figure 6, where the nested approach results are the solid curves, show that the combination of the visual signatures computed on each layer separately performs better or at least as well as the best isolated feature.

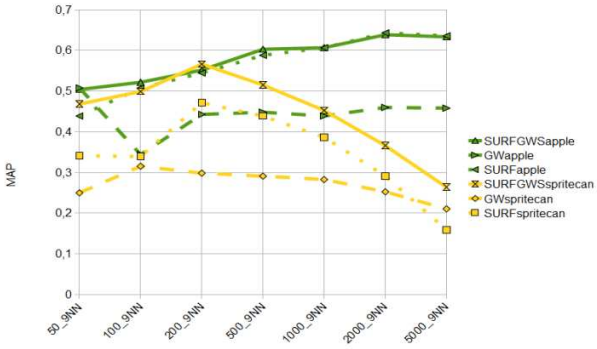


Figure 5: Detailed MAP for a selection of objects where isolated SURF features (dotted curves) outperforms graphs (dashed curves). The nested approach is depicted as solid curves. For better readability only performances of graphs built with 9 nearest neighbors and corresponding SURF features are presented.

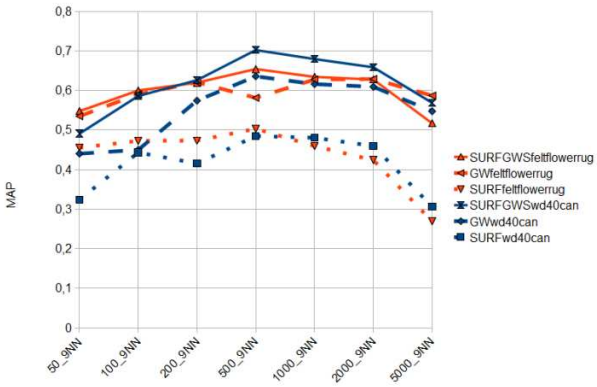


Figure 6: Detailed MAP for a selection of objects where graphs (dashed curves) outperforms isolated SURF features (dotted curves). The nested approach is depicted as solid curves. For better readability only performances of graphs built with 9 nearest neighbors and corresponding SURF features are presented here.

6. CONCLUSION AND PERSPECTIVES

In this paper, we have presented new graph features built upon SURF points as nodes and expressing spatial relations between local key points. The nested approach of using growing neighborhoods in several layers enables to capture the most discriminative visual information for different types of objects. Using growing spatial neighborhood clearly improves the results while each layer taken separately yields similar results.

The future of this work is the application of the method to the recognition of daily living objects in videos. The approach could be enhanced by refining some steps of the graphs construction and comparison. For instance, the selection of seeds could be performed by an adaptive method and the topology matrix be defined with a soft connectivity. In order to be efficient when processing a large amount of images, i.e. in videos, a graph embedding procedure could be applied to use an indexing structure that would speed up the recognition process.

7. ACKNOWLEDGMENTS

This work is partly supported by a grant from the ANR (Agence Nationale de la Recherche) with reference ANR-09-BLAN-0165-02, within the IMMED project.

8. APPENDIX

Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be two SURF points. The application of the same dissimilarity measure as the one used for the graphs cannot be iterate since no topology can be defined for an isolated point. The dissimilarity measure can only be estimated on $\boldsymbol{K^{(0)}}$.

Let us denote $\boldsymbol{d}$ the "entrywise" L_2-norm of the difference between SURF features of $\boldsymbol{A}$ and $\boldsymbol{B}$. If the parameter $\boldsymbol{\beta}$ of the CDK is set to be high enough with regards to the values of $\boldsymbol{d}$, the dissimilarity ρ defined in section 3 can be approximated by $\boldsymbol{d}$.

$$\mathrm{K}^{(0)} = \frac{\exp(-\frac{\mathrm{D}}{\beta})}{\left\|\exp(-\frac{\mathrm{D}}{\beta})\right\|_1} \quad where \quad D = \begin{pmatrix} 0 & d \\ d & 0 \end{pmatrix}$$

$$Let\ \mathrm{E} = \exp(-\frac{\mathrm{D}}{\beta}) = \begin{pmatrix} 1 & e^{\frac{-d}{\beta}} \\ e^{\frac{-d}{\beta}} & 1 \end{pmatrix}$$

$$\mathrm{K}^{(0)} = \frac{E}{\|E\|_1} = \frac{E}{2(1+e^{\frac{-d}{\beta}})}$$

$$\rho(A,B) = s(A,A) + s(B,B) - 2s(A,B)$$

$$= \frac{2-2e^{\frac{-d}{\beta}}}{2(1+e^{\frac{-d}{\beta}})} = \frac{1-e^{\frac{-d}{\beta}}}{1+e^{\frac{-d}{\beta}}}$$

$$Let\ t = \frac{d}{\beta},\ if\ \ \beta >> d:$$

$$\rho(A,B) = \frac{2e^{-t}}{\left(1+e^{-t}\right)^2}\Big|_{t=0} t + o(t) \approx \frac{d}{2\beta}$$

9. REFERENCES

[1] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60(2), pp. 91-110, 2004.

[2] J. Sivic and A. Zisserman, "Video google: a text retrieval approach to object matching in videos," *ICCV'2003*, vol. 2, pp. 1470-1477, 2003.

[3] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, "Surf:Speeded up robust features," *Computer Vision and Image Understanding*, vol. 110, pp. 346-359, 2008.

[4] K. Grauman and T. Darrell, "The pyramid match kernel: Discriminative classification with sets of image features," *In Proc. ICCV*, 2005.

[5] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories,". *In Proc. CVPR*, 2006.

[6] A. Bosch, A. Zisserman, and X. Munoz, "Representing shape with a spatial pyramid kernel,". In Proceedings of the 6th ACM international conference on Image and video retrieval (CIVR '07). ACM, New York, NY, USA, 401-408. http://doi.acm.org/10.1145/1282280.1282340

[7] R. Albatal, P. Mulhem, Y. Chiaramella, "Visual Phrases for automatic images annotation," *CBMI'10*, Grenoble, France, 2010.

[8] A. Mahboubi, J. Benois-Pineau, D. Barba,"Joint tracking of polygonal and triangulated meshes of objects in moving sequences with time varying content," *IEEE International Conference on Image Processing*, vol. 2, pp. 403-406, 2001.

[9] H. Sahbi, J.-Y. Audibert, J. Rabarisoa, R. Keriven, "Robust matching and recognition using context-dependent kernels," *Proceedings of the 25th International Conference on Machine Learning*, pp. 856-863, 2008.

[10] P. H. Gosselin, M. Cord, S. Philipp-Foliguet, "Combining visual dictionary, kernel-based similarity and learning strategy for image category retrieval," *Computer Vision and Image Understanding*, Vol. 100(3), June 2008.

[11] SIVAL Data set: http://accio.cse.wustl.edu/sg-accio/SIVAL.html