IMAGE RETRIEVAL BASED ON SPATIAL CONTEXT WITH RELAXED GABRIEL GRAPH PYRAMID

Xiaomeng Wu, Kunio Kashino

NTT Communication Science Laboratories
3-1, Morinosato Wakamiya Atsugi-shi, Kanagawa, Japan 243-0198
Email: {wu.xiaomeng, kashino.kunio}@lab.ntt.co.jp

ABSTRACT

Imposing the coherence of the spatial context on local features is becoming a necessity for object retrieval and recognition. Motivated by the success of proximity graphs in topological decomposition, clustering, and gradient estimation, we introduce a variation on and a generalization of Delaunay Triangulation, called a Relaxed Gabriel Graph (RGG), as the apex of spatial neighborhood association and design a Centrality-Sensitive Pyramid (CSP) model for hierarchical spatial context modeling. RGG is parameterized, and so allows the tuning of various applications and datasets. CSP achieves better neighborhood association and is more robust as regards feature description error than other related work. Our method is evaluated on Flickr Logos 32, Holiday, and Oxford Buildings benchmarks. Experimental results and comparisons demonstrate the superiority of our method in an image retrieval scenario.

Index Terms Image retrieval, spatial context, proximity graph

1. INTRODUCTION

The Bag-Of-Words (BOW) model based on local features has been shown to be successful in object retrieval and recognition along with its extensions [1, 2, 3, 4, 5, 6]. The standard BOW model suffers from limited discriminative power because of its ignorance of the spatial coherence between images. Many posterior spatial context models [1, 7, 8, 9] have been proposed that impose a spatial coherence on local features matched beforehand. For example, RANdom SAmple Consensus (RANSAC) [1] extracts matched local features at the retrieval stage and approximates the homography between images based on matched features at the ranking stage. Despite their high discriminative power, posterior spatial context models are usually computationally expensive.

To address this issue, some prior spatial context models [10, 11, 12, 13] have been proposed to express the spatial context of local features in the inverted index before matching. A Spatial Co-occurrence Kernel [12] uses Fixed-Radius Near Neighbors (FRNN) to compute a local feature co-occurrence matrix and then uses it for image representation and classification. Liu et al. [13] explore the two-order spatial structure of local features with k-Nearest Neighbor (k-NN), and embed the scale and orientation differential between local features in the inverted index to achieve a much lower time complexity. Although these methods successfully reduce the time cost for retrieval and classification, the FRNN and k-NN used for neighborhood association impose a high computational burden as regards indexing. Kalantidis et al. [11] propose the use of a Multi-Scale Delaunay Triangulation (MSDT) model, which is far faster than FRNN and k-NN, to model the elastic spatial context of images. However, MSDT is sensitive to feature description error and unconsciously ignores useful local features with different scales.

In this paper, we employ the promising direction of prior spatial context modeling and propose a novel hierarchical proximity graph model for the construction, representation and matching of image geometry. The main contributions include: 1. the use of a parameterized proximity graph called a Relaxed Gabriel Graph (RGG) that allows the tuning of various applications and datasets; 2. the proposal of a Centrality-Sensitive Pyramid (CSP) model that achieves better neighborhood association and is more robust as regards feature description error than MSDT; 3. the use of RGG-CSP that imposes a lighter computational burden than greedy algorithms, e.g. k-NN and β-skeleton.

In this study, we regard an image as a bag of local feature doublets that are spatially adjacent to each other and are obtained by RGG-CSP. After quantizing local features into visual words, we quantize local feature doublets into visual phrases by combining their visual words and geometric features related to Liu's method [13]. In consequence, an image is represented by a set of visual phrases for indexing and matching. The rest of the paper is organized as follows. After developing our main contribution to spatial context modeling by RGG-CSP in Sect. 2, we describe our adaptation of the method to content-based image indexing and retrieval in Sect. 3. We then report our experiments and results in Sect. 4 and discuss future directions in Sect. 5.

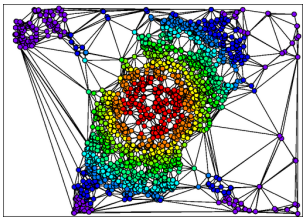
2. SPATIAL CONTEXT MODELING

2.1. Relaxed Gabriel Graph

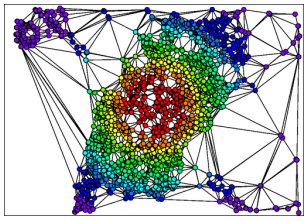
Proximity graphs, e.g. Delaunay Triangulation (DT) [14], Gabriel Graph (GG) [15], and β-skeleton [16], have been shown to be successful for neighborhood association in topological decomposition [17], clustering [18], and gradient estimation [17]. DT is one of the most widely used proximity graphs. Figure 1b shows an example of DT built from the local features detected in the image in Fig. 1a. Each edge in this graph corresponds to a pair of neighboring points. Given a number of points n, the complexity of DT is $O(n \log n)$. DT is highly efficient but in some cases extracts spurious connections between distant points, e.g. the long edges located on the left in Fig. 1b, leading to false matches of unrelated coherence. Since DT is unique, it is naturally hard to tune it to adapt optimally to various applications and datasets. Here, we propose the use of a variation of DT, called a Relaxed Gabriel Graph (RGG), which is defined as follows:

Definition 1. (Relaxed Gabriel Graph) *Given a set S of points in general position and a real number $\alpha \in [0, \pi]$, the Relaxed Gabriel*

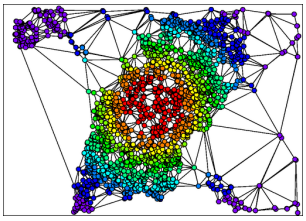
(a) Initial image.



(b) $\alpha = 0$ (DT).



(c) $\alpha = \pi/3$ (RGG).



(d) $\alpha = 2\pi/3$ (RGG).

Fig. 1. RGGs with various α.

Graph $RGG(S,\alpha)$ of S is a graph that has an edge between two vertices x and y if and only if there exists a closed disk D with center c such that:

1. x and y are on the boundary of D;

2. $D \cap S \setminus \{x,y\} = \varnothing$;

3. The absolute angle $\angle xcy \in [0,\pi]$ is at least α.

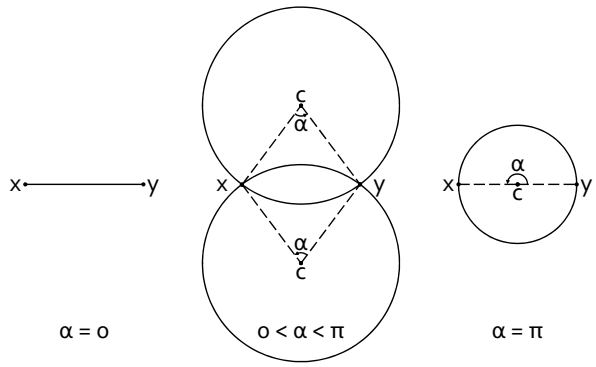
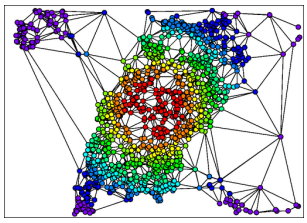


Fig. 2. Definition of RGG.

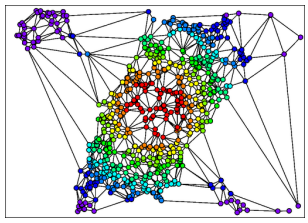
Figure 2 shows this definition. The center figure corresponds to the generalized definition with $\alpha \in (0,\pi)$. x and y have an edge between them if and only if there exists no point $z \in S \setminus \{x,y\}$ inside the intersection of the two disks. RGG imposes the additional Condition 3 in Definition 1 on DT such that spurious connections can be avoided. Figure 1c and 1d serve as two examples of RGG that extract fewer connections of distant points than the DT in Fig. 1b. Choosing $\alpha = 0$ corresponds to removing Condition 3 from Definition 1, and RGG becomes DT. In contrast, RGG equals GG for $\alpha = \pi$. For any $\alpha \in [0,\pi]$, RGG equals the intersection between DT and β-skeleton for $\beta = \sin(\pi - \alpha/2)$. RGG can be found in linear time $O(n)$ if DT is given. The total complexity is dominated by that of DT, i.e. $O(n \log n)$, and allows us to obtain an efficient algorithm for spatial neighborhood association.

2.2. Centrality-Sensitive Pyramid

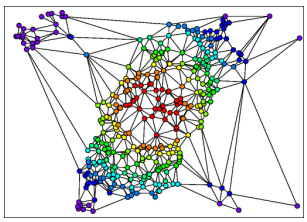
RGG is a planar graph, i.e. no edges cross each other. This planarity is inherited from DT and limits its capacity for neighborhood association when the spatial context is complicated. Multi-Scale Delaunay Triangulation (MSDT) [11] tolerates this problem by employing range partitioning, which regards the local feature scale as the partitioning key. MSDT ignores different-scaled local features, and so is more sensitive to scale variations between images.



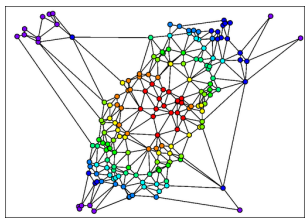
(a) One partition at level 2.



(b) One partition at level 4.



(c) One partition at level 8.



(d) One partition at level 16.

Fig. 3. RGGs with $\alpha = 2\pi/3$ at various pyramid levels. Colors indicate various centralities, e.g. red points show the highest centralities.

In this paper, we propose a hierarchical systematic sampling method to address the planarity issue. We construct a hierarchical partition $\mathscr{B} = \{B_1,\ldots,B_L\}$ that divides the point set S in L different ways. Each $B_l \in \mathscr{B}$ divides S into l partitions with $l \in [1,L]$. The l partitions are obtained by systematic sampling, in which we sort the points in S in a pre-defined order as explained below and the sampling starts by selecting a point from the ordered set at the i^{th} position with $i \in [1,l]$. Every l^{th} point after the i^{th} point in the set is selected individually, i.e. the sampling interval equals l. B_1 is at the densest level and has a single partition equaling S; B_L is at the most widely distributed level. The total number of partitions is $L(L+1)/2$. All we do then is to build a proximity graph from the points in each partition and combine all associated neighborhoods for further indexing. Figure 3 shows the RGGs in a multi-level space for the local features detected from the image in Fig. 1a.

So far we have not defined the pre-defined order for sorting. Suppose we have extracted the connections between each point $x \in S$ and its neighboring points $N(x)$ at the l^{th} level. At the $(l+1)^{\text{th}}$ level, we want to avoid the extraction of duplicate connections $\{x,y\}$ with $y \in N(x)$ for each x. One solution is known as graph coloring, where each point is colored such that no two neighboring points at the l^{th} level share the same color, but is NP-hard. An alternative is to let the points in each partition in B_{l+1} be very widely distributed in the Euclidean space such that $\{x,N(x)\}$ can be separated into different partitions. This can be achieved in a way that is counter to clustering by maximizing the intra-partition distance and minimizing the inter-partition distance. It corresponds to minimizing the distance between the centroids of each partition, and the minimum becomes zero when all centroids converge with the centroid of S.

Motivated by this idea, we propose the Centrality-Sensitive Pyramid (CSP) model. It computes the centroid c_S of S and defines

the inverse of the Euclidean distance between c_S and each $x \in S$ as the centrality of x. All points in S are sorted in descending order of centrality and systematic sampling is performed such that the intra-partition distance is maximized. The compuation of this Euclidean centrality is much more efficient than that of other centralities, e.g. closeness centrality and eigenvector centrality. CSP is sensitive to the distribution of local features in the Euclidean space, and so achieves better neighborhood association than MSDT. From Fig. 3, we can see how CSP allows the points in a certain partition at various levels to be the most widely distributed.

2.3. Complexity

RGG-CSP can be built in two phases: the construction of DT-CSP and the extraction of RGG. Given the number n of points in S and the number L of levels in CSP, the former takes $O(nL\log n - n\log L!)$ time and the latter takes $O(nL)$. Given a fixed L, $O(nL\log n - n\log L!)$ becomes linearithmic and $O(nL)$ becomes linear. We can see that RGG-CSP is much less complex than greedy algorithms, e.g. $O(n^2)$ for k-Nearest Neighbor (k-NN) and approximated β-skeleton. It is comparable to $O(n\log n)$ for approximated k-NN. In Sect. 4.2, we compare the performance of approximated k-NN with that of RGG-CSP.

3. INDEXING

In this study, an image is regarded as a bag of local feature doublets associated by RGG-CSP. We adapt the geometric model proposed by Liu et al. [13] to describe each doublet. Given two local features x and y from the same image, we use $\vec{\cdot}$, s, and θ to denote their Euclidean coordinate, scale, and orientation. Suppose that x is a central feature and y is a satellite feature. Two geometric features are defined as follows:

$$D_{x,y} = \frac{\|\vec{y}-\vec{x}\|_2}{s_x} \quad (1)$$

$$H_{x,y} = \Delta(\arctan(\vec{y}-\vec{x}), \theta_x) \quad (2)$$

where $\Delta(\cdot,\cdot) \in [0, 2\pi)$ computes the principal angle. $D_{x,y}$ indicates the relative distance between x and y; $H_{x,y}$ indicates the heading from x to y. Combining them with the visual word w assigned to each feature, we have an asymmetric vector $(w_x, w_y, D_{x,y}, H_{x,y})$ describing the co-occurrence and geometry of the doublet. A Hough transform is used to cluster reliable hypotheses and thus enable efficient searching with an inverted index. $D_{x,y}$ is transformed into two bins by a threshold 1; $H_{x,y}$ is transformed into four bins by an equal division of $[0, 2\pi)$.

Here, we add two modifications: 1. we use the same large-scale visual vocabulary to quantize the central and satellite features; 2. we propose using the Ochiai index instead of the TF-IDF cosine similarity for image matching. The first modification is designed to improve the discriminative power of the method, and the second modification makes it possible to avoid the computation and storage of the less helpful TF-IDF. IDF was designed to reduce the negative effect of confusing visual words. In our case, the visual phrase describes the co-occurrence and geometry in addition to the appearance and so is highly discriminative and rarely creates confusion. In our experiments, we found that the Ochiai index even slightly outperformed the TF-IDF cosine similarity. Given two bags (A and B) of visual phrases and $n(\cdot)$ denoting the number of elements, the Ochiai index is defined as follows.

$$K = \frac{n(A \cap B)}{\sqrt{n(A)n(B)}} \quad (3)$$

4. EXPERIMENTATION

4.1. Setting

For our evaluation, we use 3 datasets: Flickr Logos 32 (FL32) [19], Holiday (HD) [2], and Oxford Buildings (OB) [1], which are compared in Table 1. We compare our method in an image retrieval scenario with the BOW and the other spatial context models including MSDT [11] and approximated k-NN [13]. The other methods, e.g. query expansion [4], Hamming embedding [2], and soft assignment [3], are not tested here, but are compatible with our method. We measure the retrieval performance by using Mean Average Precision (MAP) [1, 2, 19]. For MSDT, the partition size is varied from 10% to 100% and the overlap ratio is varied from 0% to 90%; for approximated k-NN, the parameter k is varied from 10 to 100; for CSP, the number of levels L is varied from 10 to 100. To achieve a comprehensive comparison, we use the same geometric model described in Sect. 3 for all methods. In the reference [11], a triangle-based co-occurrence model without geometric constraints is used for MSDT. The modifications of Liu's method [13] have been discussed in Sect. 3.

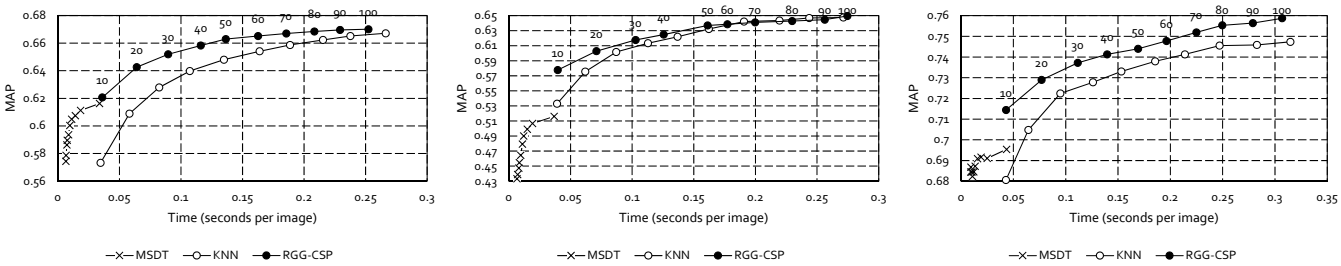
Table 1. Dataset comparison. NOI: Number Of Images.

Dataset	**FL32**	**HD**	**OB**
Category	Logo	Scenery	Building
Number of Queries	960	500	55
Number of Images	4.3K	1.5K	5.1K
Number of Features	12.7M	4.5M	17.9M
Number of Clusters	1M	0.2M	1M
Detector	Hessian	Hessian	Hessian
Descriptor	Root SIFT	SIFT	Root SIFT
Quantization	Self	Stand-Alone	Self
NOI in Quantization	4.3K	60K	5.1K

4.2. Comparison

Figure 4 compares the relationships between the MAP for retrieval and the time for spatial neighborhood association obtained using various datasets. In general, the MAPs of spatial context models are superior to that of BOW. An exception is MSDT. Its poor performance may be because: 1. we use a geometric model that is different from that in the reference [11]; 2. the scale-based partitioning ignores useful different-scaled local features; 3. the range partitioning has a limited capacity for neighborhood association. MSDT might be more suitable for dealing with planar objects, e.g. logos, because it achieved good performance for FL32. Also, MSDT is the fastest among all spatial context models.

Both approximated k-NN and CSP-based methods exhibited greatly improved performance gain compared with BOW. *To the best of our knowledge, our method's MAP of 67.0% for FL32 is the highest yet reported for the retrieval protocol of this dataset, and is around 8% higher than the second highest reported value [20].* On the other hand, Romberg et al. [20] have reported the MAPs of RANSAC [1] obtained using FL32 (56.8%) and OB (72.9%) under the same setting as ours, both of which are inferior to those obtained with our approach (Fig. 4). From Fig. 4, we can see that RGG-CSP almost always obtains a higher MAP than approximated k-NN and MSDT for the same time, which demonstrates the higher efficiency



(a) Flickr Logos 32. BOW-MAP: 54.3%. (b) Holiday. BOW-MAP: 54.7%. (c) Oxford Buildings. BOW-MAP: 71.0%.

Fig. 4. Relationship between MAP and time in seconds per image for spatial neighborhood association. Numbers above curves of RGG-CSP indicate numbers of levels L.

of our method. Figure 4 also shows that a larger number of levels L results in a longer processing time because, as discussed in Sect. 2.3, the complexity of CSP is linear to L given the fixed number of points n. A larger L also results in better neighborhood association, and so results in a higher MAP, especially for OB. For FL32 and HD, RGG-CSP outperforms k-NN with a large MAP gain when L and k are small, but the differential degrades when we enlarge the number of local feature doublets. This is because the image resolution, and in consequence the object scale, of FL32 and HD are lower than those of OB such that the doublets of distant local features become less useful. Basically, RGG-CSP is more suitable for dealing with larger images and objects. On the other hand, we also tested the MAP of RGG-CSP with various α values, but did not find a comprehensive relationship between them. In our experiments, we found that a larger number of doublets does not always correspond to a higher MAP. In the future, we will examine what kind of local feature doublets are more useful and discriminative to enhance the selectivity of neighborhood association.

Figure 5 shows some examples of the local feature doublets matched by RGG-CSP. RGG-CSP successfully extracted true responses even when the viewpoints varied significantly. Notice the small size of the object in the first and third examples, the large scale variation in the second example, the right-angled rotation in the fourth example, and the different products with the same logo in the fifth example. Also, we can see that the matched doublets have coherent shapes from various viewpoints, which also demonstrates the high discriminative power of our method.

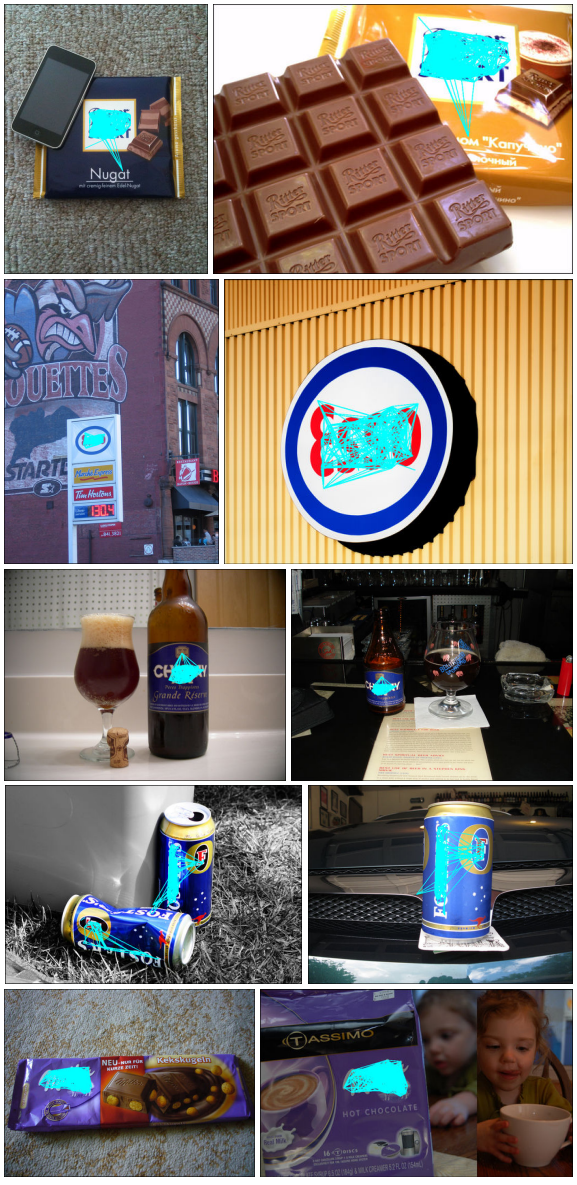


Fig. 5. Local feature doublets matched by RGG-CSP in cyan.

5. CONCLUSION

We have presented a tunable proximity graph (RGG) for spatial neighborhood association and have proposed a hierarchical systematic sampling method (CSP) for sufficient spatial context analysis. We evaluated our method in an image retrieval scenario on various types of benchmarks. RGG-CSP achieves far higher effectiveness than MSDT [11] in neighborhood association. It tends to detect more true responses than the other algorithms, e.g. k-NN [13], for image matching. We have not yet tested the scalability of our method on a large scale. We regard this experiment as future work. Also, we intend to examine whether we can distinguish useful local feature doublets from redandunt ones to enhance the selectivity of spatial neighborhood association.

6. REFERENCES

[1] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *CVPR*, 2007.

[2] Herve Jegou, Matthijs Douze, and Cordelia Schmid, "Hamming embedding and weak geometric consistency for large scale image search," in *ECCV (1)*, 2008, pp. 304–317.

[3] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman, "Lost in quantization: Improving particular object retrieval in large scale image databases," in *CVPR*, 2008.

[4] Ondrej Chum, Andrej Mikulík, Michal Perdoch, and Jiri Matas, "Total recall II: Query expansion revisited," in *CVPR*, 2011, pp. 889–896.

[5] Pierre Letessier, Olivier Buisson, and Alexis Joly, "Scalable mining of small visual objects," in *ACM Multimedia*, 2012, pp. 599–608.

[6] Giorgos Tolias, Yannis Kalantidis, Yannis Avrithis, and Stefanos Kollias, "Towards large-scale geometry indexing by feature selection," *Computer Vision and Image Understanding*, vol. 120, no. 0, pp. 31 – 45, 2014.

[7] Zhong Wu, Qifa Ke, Michael Isard, and Jian Sun, "Bundling features for large scale partial-duplicate web image search," in *CVPR*, 2009, pp. 25–32.

[8] Xiaoyu Wang, Ming Yang, Timothée Cour, Shenghuo Zhu, Kai Yu, and Tony X. Han, "Contextual weighting for vocabulary tree based image retrieval," in *ICCV*, 2011, pp. 209–216.

[9] Yimeng Zhang, Zhaoyin Jia, and Tsuhan Chen, "Image retrieval with geometry-preserving visual phrases," in *CVPR*, 2011, pp. 809–816.

[10] Herve Jegou, Matthijs Douze, and Cordelia Schmid, "Improving bag-of-features for large scale image search," *International Journal of Computer Vision*, vol. 87, no. 3, pp. 316–336, 2010.

[11] Yannis Kalantidis, Lluis Garcia Pueyo, Michele Trevisiol, Roelof van Zwol, and Yannis S. Avrithis, "Scalable triangulation-based logo recognition," in *ICMR*, 2011, p. 20.

[12] Yi Yang and Shawn Newsam, "Spatial pyramid co-occurrence for image classification," in *ICCV*, 2011, pp. 1465–1472.

[13] Zhen Liu, Houqiang Li, Wengang Zhou, and Qi Tian, "Embedding spatial context information into inverted file for large-scale image retrieval," in *ACM Multimedia*, 2012, pp. 199–208.

[14] Mark de Berg, Otfried Cheong, Marc van Kreveld, and Mark Overmars, *Computational Geometry: Algorithms and Applications*, Springer-Verlag TELOS, Santa Clara, CA, USA, 3rd edition, 2008.

[15] David W. Matula and Robert R. Sokal, "Properties of Gabriel graphs relevant to geographic variation research and the clustering of points in the plane," *Geographical Analysis*, vol. 12, no. 3, pp. 205–222, 1980.

[16] Jean Cardinal, Sébastien Collette, and Stefan Langerman, "Empty region graphs," *Comput. Geom.*, vol. 42, no. 3, pp. 183–195, 2009.

[17] Carlos D. Correa and Peter Lindstrom, "Towards robust topology of sparsely sampled data," *IEEE Trans. Vis. Comput. Graph.*, vol. 17, no. 12, pp. 1852–1861, 2011.

[18] Carlos D. Correa and Peter Lindstrom, "Locally-scaled spectral clustering using empty region graphs," in *KDD*, 2012, pp. 1330–1338.

[19] Stefan Romberg, Lluis Garcia Pueyo, Rainer Lienhart, and Roelof van Zwol, "Scalable logo recognition in real-world images," in *ICMR*, 2011, p. 25.

[20] Stefan Romberg and Rainer Lienhart, "Bundle min-hashing for logo recognition," in *ICMR*, 2013, pp. 113–120.