Second-Order Configuration of Local Features for Geometrically Stable Image Matching and Retrieval

Xiaomeng Wu, *Member, IEEE,* and Kunio Kashino, *Senior Member, IEEE*

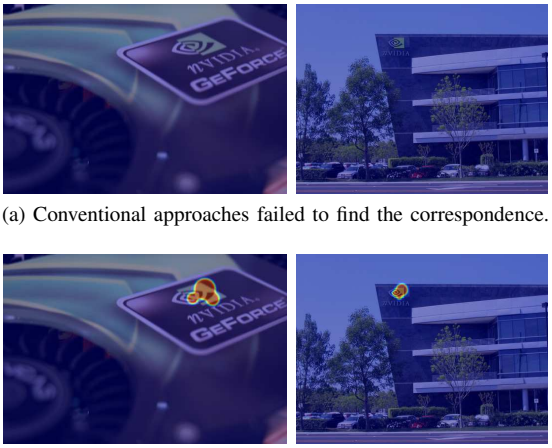
***Abstract* – Local features offer high repeatability, which supports efficient matching between images, but they do not provide sufficient discriminative power. Imposing a geometric coherence constraint on local features improves the discriminative power but makes the matching sensitive to anisotropic transformations. We propose a novel feature representation approach to solve the latter problem. Each image is abstracted by a set of tuples of local features. We revisit affine shape adaptation and extend its conclusion to characterize the geometrically stable feature of each tuple. The representation thus provides higher repeatability with anisotropic scaling and shearing than previous research. We develop a simple matching model by voting in the geometrically stable feature space, where votes arise from tuple correspondences. To make the required index space linear as regards the number of features, we propose a second approach called a Centrality-Sensitive Pyramid to select potentially meaningful tuples of local features on the basis of their spatial neighborhood information. It achieves faster neighborhood association and has a greater robustness to errors in interest point detection and description. We comprehensively evaluated our approach using Flickr Logos 32, Holiday, Oxford Buildings and Flickr 100K benchmarks. Extensive experiments and comparisons with advanced approaches demonstrate the superiority of our approach in image retrieval tasks.**

***Index Terms* – Feature Extraction, Geometry, Graph Theory, Image Retrieval.**

I. Introduction

THE bag-of-visual-words (BOVW) representation of local features [1] has been shown to be successful in image retrieval. When an image is represented using BOVW it can be treated as a document. BOVW includes several steps for defining visual words in this document: interest point detection, local feature description, and visual vocabulary generation. After interest point detection, each image is abstracted by a set of local patches. Feature representation methods, e.g. scale-invariant feature transform (SIFT) [2], represent the patches as numerical vectors called local feature descriptors. The final step is clustering, e.g. approximated k-means [3], over all the vectors to produce a visual vocabulary. After the clustering, a classification approach such as 1-nearest neighbor (1-NN) is performed to associate each patch with a visual word, and the image can be represented by a histogram of the visual words.

Current local features offer high repeatability but do not provide enough discriminative power. The direct matching of these descriptors results in massive mismatches [4]. As a result, spatial coherence models have been used to help BOVW filter out the mismatches. Kalantidis et al. [5] and Yang et al. [6] demonstrated that matching the intra-image co-occurrence of local features could achieve a higher discriminative power than matching BOVW histograms. Another group of researchers [4], [7]–[10] demonstrated that imposing a geometric constraint on co-occurrence could further improve the ability of the matching algorithm to reject mismatches. These approaches introduced a new problem called affine invariance: they make the matching sensitive to anisotropic transformations. Fig. 1 shows an example in which conventional approaches failed to find the correspondence due to the large viewpoint difference. This constitutes one of the main problems that we tackle in this paper.



(a) Conventional approaches failed to find the correspondence.

(b) Responses obtained using our approach.

Fig. 1. Images with significant anisotropic transformations that are hard to deal with using conventional geometric models.

Spatial coherence models can also be categorized in terms of prior configuration [5], [6], [8], [10] and posterior filtering [3], [4], [7], [9]: the former determines a spatial configuration of local features before matching; the latter rejects mismatches online. Approaches based on posterior filtering place an added computational burden on the online phase, which is undesirable in real applications. In contrast, prior configuration moves this burden to the feature representation phase thus making it unrelated to retrieval. The matching of local features, which reduces the redundancy of interest points, is not available here, and so the burden of image indexing is much larger than that in the posterior case. Another main objective of this paper is to improve the efficiency of image indexing to make the required indexing space linear as regards the number of features.

X. Wu and K. Kashino are with NTT Communication Science Laboratories, 3-1, Morinosato Wakamiya Atsugi-shi, Kanagawa, Japan 243-0198. E-mail: {wu.xiaomeng, kashino.kunio}@lab.ntt.co.jp.

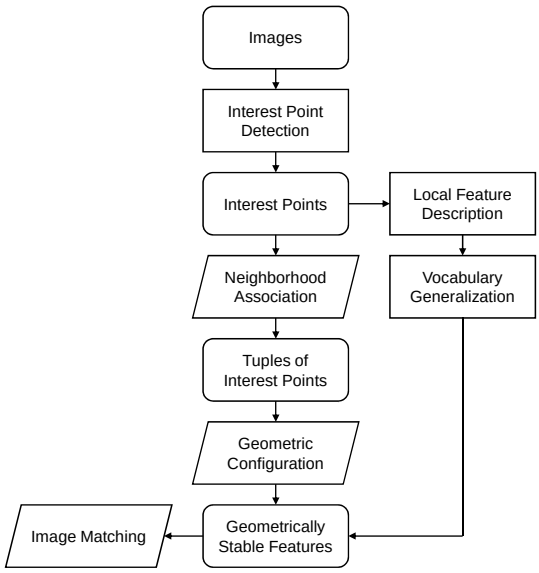


Fig. 2. Sequential diagram of geometrically stable feature representation.

In this paper, *feature representation* indicates the method used for describing the visual features of a certain entity related to the image. It corresponds to local feature description if the entity is an interest point, and corresponds to *spatial configuration* explained later if the entity is a tuple of interest points. *Image representation* indicates the method used for describing the visual features of the image, which can be a BOVW representation if only local features are taken into account. In the proposed approach, *image representation* is based on a *spatial configuration* of local features. A *spatial configuration* or *geometric configuration* is a particular layout of interest points in the image space, which contains the co-occurrence information of local features and/or the geometric relationship between interest points. *Spatial neighborhood association* is the problem of finding neighboring interest points that are close to one another in the image space.

In this paper, we develop a feature representation and image matching approach based on a second-order configuration of local features for geometrically stable retrieval. The approach follows the standard prior configuration-based spatial coherence models as shown in Fig. 2. After interest point detection, interest points that appear spatially close to one another are detected, and each image is abstracted by a set of tuples of neighboring interest points. We keep track of the semi-local geometric information of each tuple in a normalized space and develop our geometrically stable feature. The descriptors capturing both the co-occurrence and geometric characteristics of local features are embedded in an inverted index. A visual vocabulary and a Hough transform are used to enable fast searching. We develop a simple matching model by voting in the geometrically stable feature space, where votes arise from tuple correspondences. Compared with existing researches, our approach contributes the following to the state of the art:

- We revisit affine shape adaptation and extend its conclusion to characterize the second-order geometric coherence of local features. The feature representation thus provides higher repeatability with anisotropic scaling and shearing than previous research focusing simply on translation, rotation, and isotropic scaling. The matching model tailored to the geometrically stable feature significantly mitigates the risk of generating mismatch errors.
- We propose a novel Centrality-Sensitive Pyramid (CSP) model based on Delaunay triangulation for the faster association of spatially neighboring interest points. Given the same time for indexing, the retrieval achieves a higher sensitivity to true responses and a higher robustness to errors in interest point detection and description.

The remainder of the paper is organized as follows. After describing related studies in Section II, we review affine shape adaptation in Section III. In Section IV-A, we describe one of our main contributions, namely the second-order configuration of local features, and explain how the conclusion of affine shape adaptation can be extended to make the configuration invariant to anisotropic transformations. The application of our approach to image retrieval is circumstantially explained in Sections IV-B and IV-C. Section V describes our proposed CSP model. We then present our experiments in Sections VI and VII, and discuss future directions in Section VIII.

II. Literature Review

BOVW has been improved in several ways in recent years. These improvements include query expansion [11], [12], spatial coherence models [3], Hamming embedding [13], [14], soft assignment [15], query adaptation [16], [17], and database augmentation [18], [19]. All these topics concern the field of image retrieval, but we focus on the state of the art in the context of spatial coherence models because the other topics are not related to the main objective of this research.

A. Matching Based on Spatial Configuration of Local Features

Ma et al. [20] represented images with a subspace learning method that preserves spatial correlations of image pixels at the expense of losing affine invariance. In the field of spatial coherence models, a spatial configuration is a particular layout of interest points in the image space. Approaches based on spatial configurations extract the appropriate information from the spatial distribution of the interest points for image representation. Poullot et al. [21] and Kalantidis et al. [5] demonstrated that the third-order intra-image co-occurrence of neighboring local features provided a higher discriminative power than a non-spatial configuration. Poullot et al. [21] proposed grouping interest points into triangles using a nearest neighbor search (NNS) and compressing the co-occurrence information into a compact binary signature. Multi-scale Delaunay triangulation [5] replaces NNS with a graph model and treats each image as a bag of triplets of visual words. Third-order neighborhood co-occurrence is more discriminating than its second-order counterpart, but has difficulty taking full advantage of the strengths because of the large cost of computation and memory usage. In contrast, a spatial co-occurrence kernel [6] with NNS abstracts each image by using a second-order co-occurrence matrix. Zhang et al. [22] adopted the graph model [5] but used a representation based on pairs of local features.

Another group of studies [3], [4], [7]–[10], [13], [14] demonstrated that imposing a geometric constraint on co-occurrence can further improve the ability of the matching algorithm to reject mismatches but introduces the affine invariance problem (Section I). Zhang et al. [9] defined the coherence of the Euclidean distances in the image space between inter-image feature correspondences as geometric coherence. It is robust to translation but varies with rotation and scaling [9]. Wu et al. [8] presented a descriptor based on the geometric order of intra-image neighboring interest points sorted on Cartesian coordinate axes. It achieves scale invariance but still varies with rotation [8]. Weak geometrical consistency (WGC) [13], [14] defines the coherence of the distances, in scale and orientation spaces, between inter-image correspondences. It is robust to translation, rotation, and isotropic scaling but is less discriminating because it ignores the neighborhood constraint. Carneiro et al. [4] proposed a similar second-order configuration approach imposing both the neighborhood and geometric constraints on matching. It still varies with anisotropic transformations but achieves higher tolerance than WGC [13], [14]. The same idea has been adapted to different vision tasks by Gao et al. [7] and Liu et al. [10].

Among posterior filtering-based methods, Shen et al. [23] incorporated the spatial configuration of local features into an inter-image similarity measure by affine transformation simulation. Specifically, each image is transformed by rotation and isotropic scaling for a finite number of scale factors and angles, and a voting map is generated according to the relative positions of feature correspondences in the simulated image space. Instead of considering a regularized simulation, Tolias et al. [24] proposed transforming matched interest points by local transformations estimated from single feature correspondences. The same authors also proposed a novel feature selection approach based on database augmentation to make the index space linear in the number of features. Avrithis and Tolias [25] disregarded the simulation of affine transformation and focused on a transformation space spanned by the parameters of translation, rotation and isotropic scaling. Local transformations between matched features are projected onto this space, and a voting map is generated, where votes arise from feature correspondences. These approaches are efficient and robust as regards isotropic transformations, but are still sensitive to anisotropic scaling and shearing in theory.

To the best of our knowledge, the locally optimized random sample consensus (LO-RANSAC) [3] later improved by Lebeda et al. [26] is the only geometric model that does not vary with full 6 degree of freedom (6DOF) affine transformations. It repeatedly and randomly selects the inter-image correspondences and computes the parameters of a geometric model that fits the sample. It then finds all inliers to the model and defines the quality of the parameters as the spatial coherence. LO-RANSAC is much more computationally expensive than the other approaches because of the iterative model adaptation [10]. It is also known to perform poorly when the percentage of true inliers falls much below 50% [2].

B. Spatial Neighborhood Association

In the field of spatial coherence models, spatial neighborhood association is the problem of finding neighboring interest points that are close to one another in the image space. In consequence, the image is abstracted by a set of n-tuples of neighborhoods. This abstraction is usually called an n-th-order representation [7], [27]. Ascribing the spatial configuration to the n-tuples imposes a neighborhood constraint on the measure of inter-image coherence, and at the same time, significantly reduces the computation cost and memory usage.

A solution based on a nearest neighbor search (NNS) [4], [6], [8], [10], [21] is the most popular choice for this purpose. k-nearest neighbor (k-NN) [4], [10], [21] identifies the top k closest neighbors to each interest point. Spatial co-occurrence kernel [6] uses an alternative technique called fixed-radius near neighbors (FRNN), which finds all the neighbors within a given radius from each interest point. Wu et al. [8] presented a solution similar to FRNN but made the radius adaptive to the characteristic scale of each interest point. Given m as the number of interest points and k as the number of neighbors, the complexity of NNS-based solutions is close to $O(km^2)$. Approximate NNS based on a tree structure is much less complex and uses $O(m \log m)$ for building the tree and $O(km \log m)$ for searching.

Since the interest points are given in a 2D Euclidean space, neighborhood association can also be formulated as a problem of computational geometry. Kalantidis et al. [5] proposed the use of Delaunay triangulation (DT) for this purpose. To tolerate the planarity problem that leads to sensitivity to errors in interest point detection, a multi-scale Delaunay triangulation (MSDT) scheme is proposed, which divides the set of interest points into overlapped partitions according to the characteristic scale and constructs a DT from each partition. Zhang et al. [22] adapted the same model to posterior filtering in an instance search scenario. The complexity of MSDT with a divide and conquer implementation is also linearithmic as regards m, but the speed is faster than NNS because each single operation for distance computation in NNS is less efficient than the comparison operation in DT. However, MSDT achieves less complete neighborhood association because it unconsciously ignores useful neighborhoods with different scales.

III. Background

The computer vision community has made many attempts to improve the robustness of local features to affine transformations. One solution is called affine shape adaptation [28], [29]. It iteratively adapts the shape of a smoothing kernel to the region near an image point such that, provided that this iteration converges, the fixed point will be invariant to affine transformations. For an image I_L, let $\mathbf{x}_L = (x_L, y_L)^T$ be an image point in I_L. Introducing an affine transformation

$$\xi_R = A\xi_L \tag{1}$$

where A is a 2×2 matrix, we define a transformed image I_R:

$$I_L(\mathbf{x}_L + \xi_L) = I_R(\mathbf{x}_R + \xi_R) \tag{2}$$

TABLE I
EXPLANATION OF NOTATIONS USED IN SECTIONS III AND IV.

Notation	Explanation
I	Image
L, R	Indexes of images
$\mathbf{x}$, $\mathbf{y}$	Interest points
x, y	Cartesian coordinates of interest points
ξ	Position vector heading from one point to another
A	Affine transformation matrix
M	Second-moment matrix
$M^{1/2}$	Anisotropic normalization matrix
P	Rotation matrix
I_φ	Affine-normalized image
$u(\cdot)$	Local feature descriptor
ξ_φ	Affine-normalized position vector
a, b, c, d	Coefficients of anisotropic normalization matrix
θ	Characteristic orientation of interest point
t	Tuple of interest points
$h(\cdot)$	Raw descriptor of interest point tuple
$\hat{u}(\cdot)$	Visual word
ρ_φ, α_φ	Log-polar coordinates of ξ_φ
ϵ_ρ	Threshold of Hough transform for ρ_φ
$\hat{f}(\cdot)$	Visual phrase
$\hat{\rho}_\varphi$, $\hat{\alpha}_\varphi$	Geometric words
$\hat{h}(\cdot)$	Refined descriptor of interest point tuple
$\mathcal{N}(\cdot)$	Nearest neighbors of interest point
N	Set of tuples of interest points in image
$C_{\cdot,\cdot}$	Set of tuple correspondences between images
$\kappa(\cdot,\cdot)$	Kernel function
$S(\cdot,\cdot)$	Similarity between images
σ	Parameter of RBF kernel
Π	Penalty function

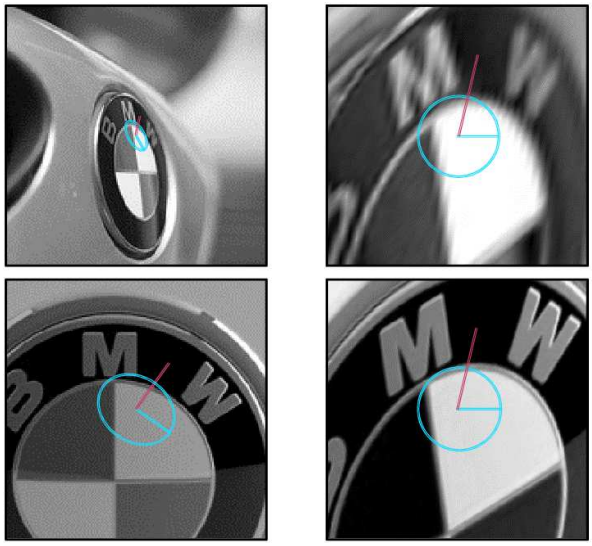
Equation (2) indicates the correspondence between each pixel $\mathbf{x}_L + \xi_L$ in I_L and each pixel $\mathbf{x}_R + \xi_R$ in I_R. Given two images that are related as regards an unknown transformation A due to a viewpoint change, if we can estimate A, then we can successfully match the images regardless of the viewpoint change. Lindeberg [28], [29] demonstrated that given a correspondence between two image points $\mathbf{x}_L$ and $\mathbf{x}_R$, A can be estimated from measurements of the affine-adapted second-moment matrices M_L and M_R. Mikolajczyk et al. [30] continued with this formulation and showed that $M^{1/2}$ can transform the original anisotropic regions around $\mathbf{x}_L$ and $\mathbf{x}_R$ into isotropic regions that are related through a rotation matrix P. The eigenvalues and eigenvectors of M characterize the curvature and shape of the ellipsoid known as an affine region (AR). The isotropic regions can be thought of as a normalized reference (NR). A can then be estimated by:

$$A = A_R^{-1} A_L \tag{3}$$

$$A_{(\cdot)} = P_{(\cdot)} M_{(\cdot)}^{1/2} \tag{4}$$

Mikolajczyk et al. [30] proposed an iterative algorithm for estimating M. P can be recovered using gradient methods.

Fig. 3 shows an example of affine normalization. Fig. 3a corresponds to the images I_L and I_R, and Fig. 3b corresponds to the normalized images $I_\varphi^{(L)}$ and $I_\varphi^{(R)}$. ARs and corresponding NRs are highlighted in blue. Although the viewpoints of the initial images vary significantly, the NRs are almost identical. Any feature computed from the NRs, usually known as a local feature, becomes invariant to the transformation A. Given $u(\mathbf{x})$ as such a feature descriptor and $\mathbf{x}$ as the image point, we denote the correspondence as



(a) Initial images I_L and I_R. (b) Normalized images.

Fig. 3. Affine normalization. Blue ellipsoids are ARs and blue circles NRs. Blue lines heading from each center show the orientation with the largest gradient. Pink lines are position vectors heading from the center of each blue point to that of a neighboring point. ARs and NRs of each neighboring point are not shown here for better viewability.

$$u(\mathbf{x}_L) = u(\mathbf{x}_R) \tag{5}$$

This equality indicates first-order feature coherence. In Section IV-A, we explain how we can model the images' second-order geometric coherence, given that (5) is satisfied, by extending the conclusion of affine shape adaptation.

IV. GEOMETRICALLY STABLE FEATURE REPRESENTATION

A. Second-Order Geometric Coherence

Given feature coherence between the image points $\mathbf{x}_L$ and $\mathbf{x}_R$ detected from two images satisfying (5), Equation (2) holds provided that the position vectors ξ_L and ξ_R are related according to (1). Given that (5) and (1) are satisfied, consider the image points at the positions $\mathbf{y}_L = \mathbf{x}_L + \xi_L$ and $\mathbf{y}_R = \mathbf{x}_R + \xi_R$. It is reasonable to expect that

$$u(\mathbf{y}_L) = u(\mathbf{y}_R) \tag{6}$$

Equation (6) naturally holds if the ARs of $\mathbf{y}_L$ and $\mathbf{y}_R$ are inside those of $\mathbf{x}_L$ and $\mathbf{x}_R$. Otherwise, the equality depends on a few assumptions: 1. we can temporarily ignore non-affine transformations, e.g. illumination changes, 2. both $\mathbf{x}$ and $\mathbf{y}$ are inside the same object, and 3. we can confine our attention to near-planar and rigid objects. Note that these assumptions are common in the field of particular object retrieval based on local features. With this configuration, Equation (6) should hold and there should be coherence between $\mathbf{y}_L$ and $\mathbf{y}_R$. Needless to say, we assume that I_L and I_R include the same object.

Note that in practice, our approach does not rely heavily on the above assumptions except for the rigidity. Illumination changes are dealt with by a local normalization of intensities before the SIFT description. Some non-planar objects such as buildings are polyhedrons, while the others such as jars are curved objects. The latter can be approximated by a polygon mesh, i.e. a collection of small *planar* faces. Our approach may fail if the curvilinear surface is too complex, otherwise it can successfully deal with this issue. Note that these assumptions are common in the field of particular object retrieval.

We now invert the problem and assume that we have two pairs of image points $(\mathbf{x}_L, \mathbf{y}_L)$ and $(\mathbf{x}_R, \mathbf{y}_R)$ that satisfy (5) and (6). Let $\xi_L = \mathbf{y}_L - \mathbf{x}_L$ and $\xi_R = \mathbf{y}_R - \mathbf{x}_R$ be the position vectors heading from $\mathbf{x}$ to $\mathbf{y}$. The normalized position vectors ξ_φ can be computed by projecting ξ_L and ξ_R such that:

$$\xi_\varphi^{(\cdot)} = P_{(\cdot)} M_{(\cdot)}^{1/2} \xi_{(\cdot)} \tag{7}$$

where $(\cdot)$ denotes L or R. If it is observed that

$$\xi_\varphi^L = \xi_\varphi^R \tag{8}$$

then $(\mathbf{x}_L, \mathbf{y}_L)$ and $(\mathbf{x}_R, \mathbf{y}_R)$ provide evidence for the belief that I_L and I_R include the same object under a certain transformation. Given the collection of all such evidence detected across I_L and I_R, a similarity can be formulated by aggregating the semi-local geometric coherence into a global coherence between the images. This similarity is naturally more discriminative than the feature coherence alone because a co-occurrence constraint is imposed by (5) and (6) and a geometric constraint is imposed by (8). It is also robust as regards viewpoint changes because the constraint does not vary with the transformation A as discussed above.

Fig. 3 shows an example where the central point $\mathbf{x}$ is highlighted in blue and the pink lines are position vectors heading from $\mathbf{x}$ to a satellite point $\mathbf{y}$. The ARs and NRs of $\mathbf{y}$ are not shown here for better viewability. The position vectors show a noticeable declination in the initial images, but become almost identical after being projected onto the normalized space. We adapt this constraint to image matching in Section IV-C.

The affine transformation $A_{(\cdot)}$ in (4) has a unique decomposition $P_{(\cdot)} M_{(\cdot)}^{1/2}$. $M^{1/2}$ transforms the anisotropic ARs into isotropic NRs, and is described by a 2×2 real matrix

$$M^{1/2} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \tag{9}$$

$M^{1/2}$ expresses an isotropic scaling if and only if $a = d$ and it is anisotropic otherwise. It also expresses a shear parallel to the horizontal axis if $b \neq 0$ and a shear parallel to the vertical axis if $c \neq 0$. In consequence, the geometric constraint proposed in this section can deal with more significant affine transformations than those proposed in previous studies [4], [7]–[10], which focus simply on translation, rotation, and isotropic scaling. We use Mikolajczyk's method [30] to estimate M. In contrast, the rotation matrix P has the following form

$$P = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \tag{10}$$

Input: $\mathbf{t} = (\mathbf{x}, \mathbf{y})$, $M(\mathbf{x})$, $P(\mathbf{x})$, $u(\mathbf{x})$, $u(\mathbf{y})$
$\xi(\mathbf{t}) \leftarrow \mathbf{y} - \mathbf{x}$
$\xi_\varphi(\mathbf{t}) \leftarrow P(\mathbf{x})\mathtt{sqrt}(M(\mathbf{x}))\xi(\mathbf{t})$
$\hat{u}(\mathbf{x}) \leftarrow \mathtt{assign}(u(\mathbf{x}))$ ▷ visual word assignment
$\hat{u}(\mathbf{y}) \leftarrow \mathtt{assign}(u(\mathbf{y}))$ ▷ visual word assignment
$(x_\varphi, y_\varphi)^T \leftarrow \xi_\varphi(\mathbf{t})$ ▷ Cartesian coordinates
$\rho_\varphi \leftarrow \mathtt{log}(\mathtt{sqrt}(x_\varphi^2 + y_\varphi^2))$
$\epsilon_\rho \leftarrow \mathtt{log}(\mathtt{sqrt}(\mathtt{det}(M(\mathbf{x}))))$
if $\rho_\varphi < \epsilon_\rho$ **then**
  $\hat{\rho}_\varphi \leftarrow 0$
else
  $\hat{\rho}_\varphi \leftarrow 1$
end if
$\alpha_\varphi \leftarrow \mathtt{atan2}(y_\varphi, x_\varphi)$
$\hat{\alpha}_\varphi \leftarrow \mathtt{floor}(2\alpha_\varphi/\pi)$
$\hat{f}(\mathbf{t}) \leftarrow \hat{u}(\mathbf{x})||\hat{u}(\mathbf{y})||\hat{\rho}_\varphi(\mathbf{t})||\hat{\alpha}_\varphi(\mathbf{t})$ ▷ concatenation operation
$\hat{h}(\mathbf{t}) \leftarrow \{\hat{f}(\mathbf{t}), \xi_\varphi(\mathbf{t})\}$ ▷ pair
Output: $\hat{h}(\mathbf{t})$ ▷ descriptor of tuple $\mathbf{t}$

Fig. 4. Pseudo-code of interest point tuple description. $\mathbf{x}$ is the central point and $\mathbf{y}$ is the satellite point in this case. $\mathtt{assign}(\cdot)$ is a function that assigns a word to each point using a vocabulary generated beforehand.

where the orientation θ of a point $\mathbf{x}$ can be obtained as the maximum value in a histogram of oriented gradients within the NR around $\mathbf{x}$. We compare our geometric constraint with previous reports in more detail in Section VI.

B. Image Representation

Given an image I, a 2-tuple $\mathbf{t} = (\mathbf{x}, \mathbf{y})$ is a pair of points $\mathbf{x} \in I$ and $\mathbf{y} \in I$. Given a position vector heading from $\mathbf{x}$ to $\mathbf{y}$ of $\xi(\mathbf{t}) = \mathbf{y} - \mathbf{x}$, a normalized position vector $\xi_\varphi(\mathbf{t})$ can be computed by projecting $\xi(\mathbf{t})$:

$$\xi_\varphi(\mathbf{t}) = P(\mathbf{x}) M(\mathbf{x})^{1/2} \xi(\mathbf{t}) \tag{11}$$

The appearance and geometric characteristics of $\mathbf{t}$ are thus:

$$h(\mathbf{t}) = \langle u(\mathbf{x}), u(\mathbf{y}), \xi_\varphi(\mathbf{t}) \rangle \tag{12}$$

where $u(\cdot)$ is a feature descriptor. Given two images I_L and I_R, every geometrically stable correspondence can be found by thresholding the similarity between each $h(\mathbf{t}_L)$ and each $h(\mathbf{t}_R)$. Inspired by the success of descriptor quantization, we consider a visual vocabulary and a Hough transform to assign certain visual and geometric words to a tuple. Hough transform is also helpful for rejecting mismatched feature correspondences. We quantize $u(\cdot)$ into a visual word $\hat{u}(\cdot)$ using a visual vocabulary with 1M visual words constructed with an approximated k-means. We also transform ξ_φ from Cartesian coordinates to log-polar coordinates $(\rho_\varphi, \alpha_\varphi)$ with ρ_φ being the log radius and α_φ being the log-polar angle. ρ_φ is further transformed into two bins by a threshold $\epsilon_\rho = \log \sqrt{\det M(\mathbf{x})}$, which is the AR scale of $\mathbf{x}$. α_φ is transformed into four bins by an equal division of $[0, 2\pi)$. We have tried a number of configurations and the performance stabilized with the above setting. We thus have an asymmetric visual phrase

$$\hat{f}(\mathbf{t}) = \langle \hat{u}(\mathbf{x}), \hat{u}(\mathbf{y}), \hat{\rho}_\varphi(\mathbf{t}), \hat{\alpha}_\varphi(\mathbf{t}) \rangle \tag{13}$$

describing the co-occurrence and geometric characteristics of the tuple. Depending on the parameterization, quantization such as a Hough transform usually incurs information loss

Input: $\kappa(\cdot,\cdot)$, Π
1: $F \leftarrow \varnothing$ ▷ inverted index
2: **for all** I **do**
3: **for** $\mathbf{t} \in I$ **do**
4: $F(\hat{f}(\mathbf{t})) \leftarrow F(\hat{f}(\mathbf{t})) \cup \{I, \xi_\varphi(\mathbf{t})\}$
5: **end for**
6: **end for**
7: $\mathbf{S} \leftarrow \varnothing$ ▷ set of ranking lists
8: **for all** I_Q **do** ▷ query
9: **for all** I **do**
10: $S(I_Q, I) \leftarrow 0$
11: **end for**
12: **for** $\mathbf{t}_Q \in I_Q$ **do**
13: **for** $\{I, \xi_\varphi(\mathbf{t})\} \in F(\hat{f}(\mathbf{t}_Q))$ **do**
14: $S(I_Q, I) \leftarrow S(I_Q, I) + \kappa(\xi_\varphi(\mathbf{t}_Q), \xi_\varphi(\mathbf{t}))$
15: **end for**
16: **end for**
17: $\mathbf{S}(I_Q) \leftarrow \varnothing$ ▷ ranking list
18: **for all** I **do**
19: $S(I_Q, I) \leftarrow S(I_Q, I)/\Pi$
20: $\mathbf{S}(I_Q) \leftarrow \mathbf{S}(I_Q) \cup \{I, S(I_Q, I)\}$
21: **end for**
22: $\mathbf{S} \leftarrow \mathbf{S} \cup \mathbf{S}(I_Q)$
23: **end for**
Output: $\mathbf{S}$

Fig. 5. Pseudo-code of image indexing and retrieval. The output is a set of ranking lists, each of which corresponds to a query. Π is actually image-specific but here we denote it as a constant for simplicity.

[31]. Since ξ_φ has only two dimensions, we tap into its descriptor space by preserving ξ_φ in the tuple's representation:

$$\hat{h}(\mathbf{t}) = \langle \hat{f}(\mathbf{t}), \xi_\varphi(\mathbf{t}) \rangle \tag{14}$$

Given two tuples $\mathbf{t}_L$ and $\mathbf{t}_R$, the correspondence can be determined by imposing $\hat{f}(\mathbf{t}_L) = \hat{f}(\mathbf{t}_R)$ on them and thresholding the similarity between $\xi_\varphi(\mathbf{t}_L)$ and $\xi_\varphi(\mathbf{t}_R)$. This enables us to perform an efficient search with an inverted index in which each key is a visual phrase $\hat{f}(\mathbf{t})$ and each mapped value is a pair consisting of image ID I and $\xi_\varphi(\mathbf{t})$. Fig. 4 summarizes the scheme for describing a tuple of interest points.

An image is regarded as a document of tuples. In theory, the total number of tuples is quadratic, but in practice, we can restrict the tuples to the nearest neighbors of each interest point in the image space. In this paper, we propose a Centrality-Sensitive Pyramid approach based on Delaunay triangulation for this purpose, which is described in Section V. Given an interest point $\mathbf{x}$, let its nearest neighbors be $\mathcal{N}(\mathbf{x})$. The set of tuples in I contains all pairs of $\mathbf{x} \in I$ with their neighbors:

$$N = \{\mathbf{t} \in I^2, \mathbf{t} = (\mathbf{x}, \mathbf{y}) : \mathbf{y} \in \mathcal{N}(\mathbf{x})\} \tag{15}$$

C. Image Matching

Given two images I_L and I_R, a set of candidate correspondences is obtained based on the visual phrase:

$$C_{L,R} = \{(\mathbf{t}_L, \mathbf{t}_R) \in N_L \times N_R : \hat{f}(\mathbf{t}_L) = \hat{f}(\mathbf{t}_R)\} \tag{16}$$

It follows that each correspondence should contribute to the similarity score according to how far apart $\xi_\varphi(\mathbf{t}_L)$ and $\xi_\varphi(\mathbf{t}_R)$ are. We define this contribution using a kernel function $\kappa(\xi_\varphi(\mathbf{t}_L), \xi_\varphi(\mathbf{t}_R))$, which gives rise to the image similarity:

$$S(I_L, I_R) = \frac{\sum_{(\mathbf{t}_L, \mathbf{t}_R) \in C_{L,R}} \kappa(\xi_\varphi(\mathbf{t}_L), \xi_\varphi(\mathbf{t}_R))}{\Pi} \tag{17}$$

where Π is a penalty function. It is more reasonable to choose a kernel related to Euclidean distance since ξ_φ is basically a position vector. We define κ as a radial basis function (RBF) kernel with a parameter σ. Π penalizes images with very many features that cause confusion between unrelated images. We compare various functions handling Π in Section VI-C. The choice of Π is inspired by the similarity functions defined in information retrieval [32]. Fig. 5 summarizes the proposal we describe in this section. Lines 1 to 6 correspond to image indexing, and Lines 7 to 23 correspond to retrieval.

Note that the high discriminative power of spatial coherence models [5], [6], [10] usually leads to low robustness such that no responses can be found between related images in certain cases. As a consequence this leads to zero similarities. A simple but effective solution is to first rank the retrieved images according to (17) and then rank the images with zero similarities according to the cosine similarity between the term frequency-inverse document frequency (TF-IDF) histograms of visual words. The latter criterion is exactly the same as that used in standard BOVW, i.e. we fuse the ranking list of BOVW and that of spatial coherence models. This solution leads to additional computation but makes the retrieval more reliable. In Section VI, all the previous studies [5], [6], [10] used for comparison are implemented in the same manner.

Note that (17) does not take the inverse document frequency (IDF) into account. This indirectly avoids the computation and storage costs, but the actual motivation lies in the fact that IDF is less helpful for spatial coherence models. IDF was designed to reduce the negative effect of confusing local features, e.g. those deriving from finely-textured patterns. In our approach, the visual phrase describes both the co-occurrence and the geometric characteristics of semi-local regions and so is highly discriminative and rarely creates confusion. We have conducted a preliminary experiment using an Oxford Buildings dataset for testing and mean average precision (MAP) for evaluation. We found that (17) even slightly exceeded the cosine similarity between the TF-IDF histograms of visual phrases (0.5% MAP improvement). Further details regarding the dataset and the configuration of our experiments can be found in Section VI-A.

V. Graph-Based Neighborhood Association

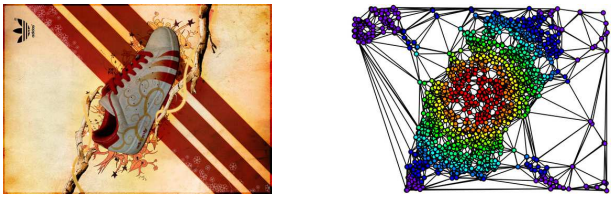
A. Proximity Graph

Proximity graphs including Delaunay triangulations (DT) [33], Gabriel Graphs (GG) [34], and β-skeletons [35] have been shown to be successful for neighborhood association in topological decomposition [36], clustering [37], and gradient estimation [36]. DT [33], which is one of the most widely used, is defined as follows:

Definition 1. (Delaunay Triangulation) *Given a set S of points in a general position, the Delaunay triangulation $DT(S)$ of S is a graph that has an edge between two vertices $\mathbf{x}$ and $\mathbf{y}$ if and only if there exists a closed disk D such that:*

TABLE II
EXPLANATION OF NOTATIONS USED IN SECTION V.

Notation	Explanation
S	Set of interest points
$DT(\cdot)$	Delaunay triangulation (DT)
$\mathbf{x}$, $\mathbf{y}$, $\mathbf{c}$	Points in Euclidean space
D	Closed disk used for DT
α	Parameter of relaxed Gabriel graph (RGG)
$RGG(\cdot,\cdot)$	RGG
$\mathcal{B}$	Hierarchy of sets of partitions in CSP
B_l	Set of partitions at l-th level
L	Number of levels
$\mathcal{N}_l(\cdot)$	Nearest neighbors of interest point at l-th level
$\mathbf{c}_S$	Centroid of interest points in S



(a) Initial image. (b) $\alpha = 0$ (DT). (c) $\alpha = \pi/2$ (RGG). (d) $\alpha = \pi$ (GG).

Fig. 6. RGGs with various α values.

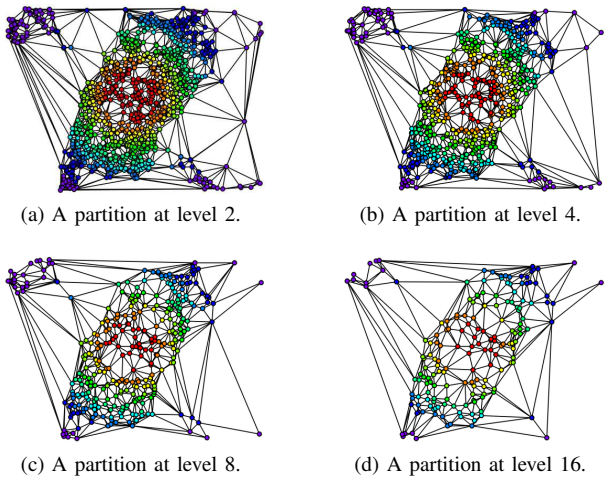
1) $\mathbf{x}$ *and* $\mathbf{y}$ *are on the boundary of* D*;*
2) $D \cap S \setminus \{\mathbf{x}, \mathbf{y}\} = \varnothing$.

We choose DT as the base for neighborhood association to generate the set of tuples N defined in (15) because of its higher completeness and higher efficiency. We also compare DT to a generation of DT and its subcomplexes, called a relaxed Gabriel graph (RGG) [38], proposed in computational geometry. RGG is parameterized and so allows optimal adaptation to various applications and datasets. It is defined as:

Definition 2. (Relaxed Gabriel Graph) *Given a set* S *of points in a general position and a real number* $\alpha \in [0, \pi]$*, the Relaxed Gabriel Graph* $RGG(S, \alpha)$ *of* S *is a graph that has an edge between two vertices* $\mathbf{x}$ *and* $\mathbf{y}$ *if and only if there exists a closed disk* D *with center* $\mathbf{c}$ *such that:*

1) $\mathbf{x}$ *and* $\mathbf{y}$ *are on the boundary of* D*;*
2) $D \cap S \setminus \{\mathbf{x}, \mathbf{y}\} = \varnothing$*;*
3) *The absolute angle* $\angle\mathbf{xcy} \in [0, \pi]$ *is at least* α.

RGG imposes the additional Condition 3 in Definition 2 on DT such that connections between distant points can be avoided. Fig. 6c serves as an example of RGG. Choosing $\alpha = 0$ corresponds to removing Condition 3 from Definition 2, and RGG becomes DT (Fig. 6b). In contrast, RGG equals GG for $\alpha = \pi$ (Fig. 6d). For any $\alpha \in [0, \pi]$, RGG equals the intersection between DT and β-skeleton for $\beta = \sin(\pi - \alpha/2)$.



(a) A partition at level 2. (b) A partition at level 4. (c) A partition at level 8. (d) A partition at level 16.

Fig. 7. DTs at various pyramid levels. Colors indicate various centralities, e.g. red points show the highest centralities.

B. Centrality-Sensitive Pyramid

DT is a planar graph, i.e. no edges cross each other. This planarity limits its capacity for neighborhood association when the interest point context is complicated. Multi-scale Delaunay triangulation (MSDT) [5] tolerates this problem by employing range partitioning, which regards the scale of each interest point as a partitioning key. MSDT ignores different-scaled interest points, and so is more sensitive to scale variations.

We propose a hierarchical systematic sampling scheme to address the planarity issue. We construct a hierarchical partition $\mathcal{B} = \{B_1, \ldots, B_L\}$ that divides the point set S in L different ways. Each $B_l \in \mathcal{B}$ 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 uniformly 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 spatial configuration. Fig. 7 shows the DTs in a multi-level space for the interest points detected from the image in Fig. 6a.

So far we have not mentioned the pre-defined order for sorting. Suppose we have extracted the connections between each point $\mathbf{x} \in S$ and its neighbors $\mathcal{N}_l(\mathbf{x})$ at the l-th level. At level $l+1$, we want to avoid the extraction of duplicate connections $\{\mathbf{x}, \mathbf{y}\}$ with $\mathbf{y} \in \mathcal{N}_l(\mathbf{x})$ for each $\mathbf{x}$. One solution is known as graph coloring, where each point is colored such that no two neighborhoods at level $l+1$ share the same color, but the problem is NP-hard. An alternative is for the points in each partition in B_{l+1} to be very uniformly distributed in the Euclidean space such that $\{\mathbf{x}, \mathcal{N}_l(\mathbf{x})\}$ can be separated

Input:
1: $\mathbf{c}_S \leftarrow \texttt{mean}(S)$
2: $\Delta \leftarrow \varnothing$ ▷ list of points with centralities
3: **for** $\mathbf{x} \in S$ **do**
4: $c(\mathbf{x}) \leftarrow 1/\texttt{deviation}(\mathbf{x}, \mathbf{c}_S)$ ▷ centrality
5: $\Delta \leftarrow \Delta \cup \{\mathbf{x}, c(\mathbf{x})\}$
6: **end for**
7: $\Delta \leftarrow \texttt{sort}(\Delta)$ ▷ sorted in descending order
8: $N \leftarrow \varnothing$ ▷ set of neighborhoods
9: **for** $l \in [1, L]$ **do**
10: $N_l \leftarrow \varnothing$ ▷ set of neighborhoods at l-th level
11: **for** $i \in [1, l]$ **do**
12: $\Lambda(i) \leftarrow \varnothing$ ▷ i-th partition at l-th level
13: **end for**
14: $n \leftarrow |S|$
15: **for** $j \in [1, n]$ **do**
16: $\{\mathbf{x}, c(\mathbf{x})\} \leftarrow \Delta(j)$
17: $i \leftarrow j \bmod l$
18: $\Lambda(i) \leftarrow \Lambda(i) \cup \mathbf{x}$
19: **end for**
20: **for** $i \in [1, l]$ **do**
21: $N_l \leftarrow N_l \cup \texttt{DT}(\Lambda(i))$ ▷ Delaunay triangulation
22: **end for**
23: $N \leftarrow N \cup N_l$
24: **end for**
Output: N

Fig. 8. Pseudo-code of DT-CSP. The output is a set of interest point tuples.

into different partitions at level $l+1$. This can be achieved in a way that is the opposite of clustering, namely maximizing the intra-partition distance and minimizing the inter-partition distance. This 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. This model computes the centroid $\mathbf{c}_S$ of S and defines the inverse of the Euclidean distance between $\mathbf{c}_S$ and each $\mathbf{x} \in S$ as the centrality of $\mathbf{x}$. All points in S are sorted in descending order of centrality and systematic sampling is conducted such that the intra-partition distance is maximized. The computation 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 interest points in the Euclidean space, and so achieves better neighborhood association than MSDT. From Fig. 7, we can see how CSP allows the points in a certain partition at various levels to be the most uniformly distributed. Fig. 8 summarizes the scheme of DT-CSP. Lines 1 to 7 correspond to centrality computation and sorting, and Lines 8 to 24 correspond to hierarchical neighborhood association.

VI. Experimentation

A. Setting

For our evaluation, we use the following datasets: Flickr Logos 32 (FL32) [39], Holiday (HD) [13], and Oxford Buildings (OB) [3], which are compared in Table III. Note that these datasets contain very different types of images. FL32 is a logo dataset, and images of the same class share very small visually similar regions. HD is not an object dataset but a scenery dataset. OB is a building dataset, and the mean object size is much larger than that of FL32 but still smaller than 40%. We employ the same detector [30] based on affine shape adaptation [28], [29] for all datasets to extract interest points. We measure the performance using mean average precision (MAP) [3], [13], [39], [40] and mean precision at top-4 (MP@4) [40].

TABLE III
Dataset comparison.

Dataset	**FL32 [39]**	**HD [13]**	**OB [3]**
Category	Logo	Scenery	Building
#Probe	960	500	55
#Gallery	4.3K	1.5K	5.1K
#Feature	12.7M	4.5M	17.9M
#Cluster	1M	0.2M	1M
Descriptor	Root SIFT	SIFT	Root SIFT
Quantization	Self	Stand-Alone	Self
Object Size (Mean)	9%		38%
Object Size (Median)	5%		28%

B. Advanced Approaches for Comparison

We compare our approach in an image retrieval scenario with the BOVW and other spatial coherence models including multi-scale Delaunay triangulation (MSDT) [5], the spatial co-occurrence kernel (SCK) [6], and Liu's approach [10]. MSDT [5] is chosen because it is the only one that uses a graph-based neighborhood association scheme, and also, it is representative of the few studies that have considered a third-order co-occurrence constraint. SCK [6] is chosen as a representative of second-order co-occurrence models that do not consider intra-feature geometric characteristics. Liu's approach [10] is chosen because it obeys the most types of affine invariance among the geometric models introduced in Section II-A. The others, including query expansion [11], [12], Hamming embedding [13], [14], soft assignment [15], query-adaptive similarity measure [16], [17] and database augmentation [18], [19], are not tested but are compatible with our approach.

Table IV qualitatively compares the approaches implemented in our experiments where ASA2 denotes our geometric model, which extends affine shape adaptation to model the second-order geometric coherence. To demonstrate the superiority of the CSP model proposed in Section V, we also combined traditional neighborhood association models with ASA2 for comparison.

Because Liu's approach [10] is the most closely related to our approach, we provide more detail here on the geometric descriptors it adopts. Given a 2-tuple $\mathbf{t} = (\mathbf{x}, \mathbf{y})$ where $\mathbf{x}$ is the central point and $\mathbf{y}$ is the satellite point, the geometric descriptors are defined as

$$\rho = \frac{||\mathbf{y} - \mathbf{x}||_2}{s(\mathbf{x})} \tag{18}$$

$$\alpha = \Delta_\theta(\arctan(\mathbf{y} - \mathbf{x}) - \theta(\mathbf{x})) \tag{19}$$

where $s(\cdot)$ and $\theta(\cdot)$ denote the scale and orientation of an interest point and $\Delta_\theta(\cdot) \in [-\pi, \pi]$ calculates the principal angle. A Hough transform is applied to these descriptors, and the image is abstracted by a bag of visual phrases similar to

TABLE IV
APPROACH COMPARISON[1].

Approach	BOVW	MSDT [5]	SCK [6]	Liu et al. [10]	MSDT-ASA2	kNN-ASA2	CSP-ASA2
Co-occurrence	×	√	√	√	√	√	√
Geometric Characteristics	×	×	×	√	√	√	√
Order[2]	1	3	2	2	2	2	2
Neighborhood Association		MSDT	FRNN	k-NN	MSDT	k-NN	DT-CSP
Translation				√	√	√	√
Rotation				√	√	√	√
Isotropic Scaling				√	√	√	√
Anisotropic Scaling				×	√	√	√
Shearing				×	√	√	√

[1] √ indicates that co-occurrence or geometric characteristics are taken into account or the approach is invariant to the corresponding transformations, and × the reverse.
[2] The order indicates the number of elements in each tuple.

(13). A TF-IDF histogram of visual phrases is used for image representation. Taking the scale and orientation of the central point into account makes the descriptors robust to isotropic transformations. However, because both $s(\mathbf{x})$ and $\theta(\mathbf{x})$ are variant under anisotropic scaling and shearing, the approach achieves limited invariance to anisotropic transformations.

C. Parameter Examination

In our approach, the following parameters influence the retrieval performance: the L used in CSP, the parameter σ of the RBF kernel, and the penalty function Π in (17). We vary L from 10 to 100, and the results are compared in Section VI-D. We tested the MAP with various σ^2 values. A larger value corresponds to greater robustness as regards noise but less discriminative power and vice versa. The best MAP stabilized within $\sigma^2 \in [4, 6]$ for all datasets, but the retrieval performance is very insensitive to this parameter. We hence chose $\sigma^2 = 5$ for all subsequent experiments. We also compared various functions handling Π shown in Table V where N is defined in (15). These choices were inspired by the similarity functions defined in information retrieval [32]. Since our approach is based on a second-order structure, it is seemingly reasonable to choose quadratic functions, e.g. Function 4 or Function 5. However, the best MAP was stabilized with linear functions, e.g. Function 2 or Function 3. Quadratic functions tend to be too rigid because of the high discernment of the visual phrase defined in (13). We therefore chose $||N_L|| + ||N_R||$ for all later experiments. The parameters of the compared approaches are examined as follows. The partition size for MSDT [5] is varied from 0.1 to 1 and the overlap ratio is varied from 0 to 0.9; the parameter k used in k-NN is varied from 10 to 100 for SCK [6] and Liu's approach [10].

D. Comparison

Table VI compares the best performance of various runs. In general, spatial coherence models are superior to BOVW with the exception being MSDT [5]. The malfunction of MSDT is because the graph model has a lower capacity for neighborhood association than k-NN and CSP and the third-order co-occurrence constraint is too sensitive to errors in feature description. In contrast, SCK, Liu's approach, and our approaches based on ASA2 exhibited greatly improved performance compared with BOVW.

TABLE V
PENALTY FUNCTION COMPARISON (%)[1].

ID	Function	FL32	HD	OB
1	1	67.0	65.1	**76.3**
2	$\sqrt{\|\|N_L\|\|\|N_R\|\|}$	67.4	65.9	76.2
3	$\|\|N_L\|\| + \|\|N_R\|\|$	**67.5**	**67.4**	76.1
4	$\|\|N_L\|\|\|\|N_R\|\|$	66.8	56.2	75.5
5	$\|\|N_L\|\|^2 + \|\|N_R\|\|^2$	67.0	66.6	75.0

[1] kNN-ASA2 with $k = 100$ and $\sigma^2 = 5$ is used in this experiment.

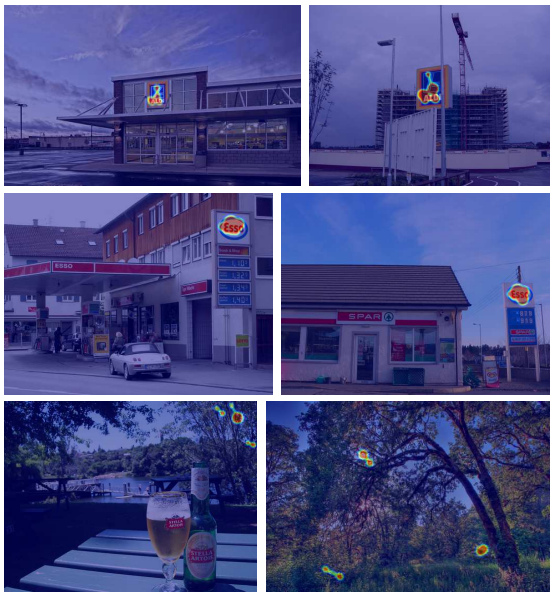
TABLE VI
BEST PERFORMANCE COMPARISON (%)[1].

	FL32		HD[2]	OB	
	MAP	MP@4	MAP	MAP	MP@4
BOVW	54.3	79.8	54.7	70.9	94.1
MSDT [5]	54.8	81.1	49.5	70.1	94.1
MSDT-ASA2	61.9	87.2	55.9	70.6	94.1
SCK [6]	63.4	87.5	63.0	72.8	95.5
Liu et al. [10]	65.3	89.5	66.2	73.6	95.9
kNN-ASA2	67.5	90.9	**67.4**	76.1	**96.4**
CSP-ASA2	**68.0**	**91.2**	66.9	**76.9**	**96.4**

[1] The highest performance was obtained with $k = 100$ for SCK [6], Liu's approach [10], and kNN-ASA2 and with $L = 100$ for CSP-ASA2.
[2] MP@4 is not evaluated here because there are fewer than four ground truth images.

To the best of our knowledge, CSP-ASA2's MAP of 68.0% for FL32 is the highest yet reported for the retrieval protocol of this dataset, and is more than 8% higher than the second highest reported value [40]. Romberg et al. [40] have reported the MAPs of LO-RANSAC [15] obtained using FL32 and OB under the same setting as ours. The best reported MAPs are 56.8% for FL32 and 72.9% for OB, both of which are inferior to those obtained with approaches based on ASA2. RANSAC is known to perform poorly when the percentage of true inliers falls much below 50% [2]. This situation commonly occurs with datasets shown in Table III, in which the object

(a) Responses obtained using BOVW.



(b) Responses obtained using kNN-ASA2 with $k = 100$.

Fig. 9. Comparison of standard BOVW and kNN-ASA2.

size is very small, and so may explain the surprisingly poor performance of LO-RANSAC.

1) BOVW: Fig. 9 shows examples of the features matched by BOVW and the tuples matched by kNN-ASA2. Both approaches extracted true responses corresponding to the logos in the two examples at the top. BOVW also provided many false responses. In contrast, kNN-ASA2 precisely matched the objects and rejected every false response. Please note the small size of the logos and the coherent shapes of the matches in Fig. 9b. The example at the bottom shows the false responses detected between confusing unrelated images. We can see that BOVW is very sensitive to finely-textured patterns, e.g. foliage. Although kNN-ASA2 also detected a few false responses, most false responses were successfully rejected. The second-order configuration of interest points enables us

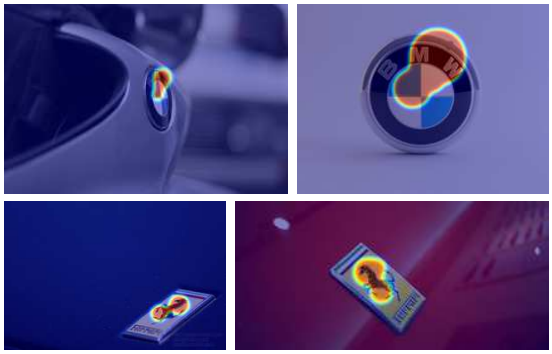
Fig. 10. False responses obtained using SCK [6] with $k = 100$. All these responses were rejected successfully by kNN-ASA2 with the same k.

to impose additional co-occurrence and geometric constraints on the matching such that the massive mismatches obtained by BOVW can be effectively avoided.

2) MSDT [5]: By comparing MSDT [5] and MSDT-ASA2 (Table VI), we can see that given the same set of neighborhoods associated by MSDT, the second-order geometric model greatly outperforms the third-order co-occurrence model. This is mainly because in MSDT, all three visual words must be identical in order to match a single 3-tuple. This leads to rejections of true responses, especially when the visual vocabulary is large. Although the second-order occurrence model may be less discriminating than MSDT, imposing a geometric constraint on it reinforces this potential shortcoming. On the other hand, MSDT-ASA2 did not perform as well as kNN-ASA2 and CSP-ASA2. It supports our claim in Section V-B: MSDT [5] tolerates the planarity problem of DT to some extent but still achieves less complete neighborhood association and unconsciously ignores useful neighborhoods with different scales. As a result, MSDT is more sensitive to errors in interest point detection and description. Kalantidis et al. [5] used a small vocabulary with only 5K words and reported a 10% MAP improvement over BOVW. We believe that this improvement is because of the small vocabulary size.

3) SCK [6]: Fig. 10 shows the false responses detected by SCK [6]. It is known that local features such as SIFT are indiscriminating as regards pictures containing dense characters, e.g. the newspaper in the second example. Although SCK took the local feature co-occurrence into consideration, it still detected some false responses. In contrast, our approach rejected all false responses because we enforced an additional geometric constraint over the confusing interest points.

4) Liu's Approach [10]: Fig. 11 compares Liu's approach [10] and kNN-ASA2. kNN-ASA2 successfully extracted true responses in Fig. 11a even when the viewpoints varied significantly. Liu's approach obtained zero responses from these image pairs. Fig. 1 serves as another example, in which Fig. 1a corresponds to Liu's approach and Fig. 1b to kNN-ASA2. Our approach is not only more robust but also achieves higher discernment than Liu's approach. Liu's approach provided many false responses in Fig. 11b, and all the false responses were rejected by kNN-ASA2.



(a) Responses obtained using kNN-ASA2.

(b) False responses obtained using Liu's approach [10].

Fig. 11. Comparison of Liu's approach [10] and kNN-ASA2 ($k = 100$).

Fig. 12 shows the relationship between the MAP and the k used in k-NN. In all cases kNN-ASA2 obtains a higher MAP than SCK [6] and Liu's approach [10] for the same k. This provides further evidence that, given the same set of neighborhoods associated by k-NN, our geometric model achieves superior effectiveness to the second-order co-occurrence constraint alone and the geometric descriptors defined in (18) and (19) by Liu et al [10]. A larger number k leads to a higher MAP for all these approaches. A larger k allows the model to capture the spatial characteristics of larger objects, but may also cause more confusion between unrelated objects. The degradation of robustness has a negative impact on retrieval when the visual vocabulary is small, as reported by Liu et al. [10]. In contrast, we used a vocabulary with 1M visual words and so avoided the impact of false responses. Hence the curves in Fig. 12 became monotonic.

5) CSP: We examined the relationship between the retrieval performance and the parameter α of RGG [38]. RGG with $\alpha = \pi/6$ achieved the best MAP and increased MAP by about 0.05% compared with DT [33] ($\alpha = 0$), which is a very trivial improvement. It can be concluded that although the parameterized RGG allows optimal adaptation to various applications and datasets, at least for neighborhood association in a spatial coherence model, it is not superior to DT in effectiveness. On the other hand, with GG [34] ($\alpha = \pi$) there was a MAP decrease of around 1% compared with DT. Although the subcomplexes of DT may outperform DT in the other tasks, DT is still the most stable choice for our model.

Fig. 13 shows the relationship between the MAP and the time for neighborhood association. We can see that CSP-ASA2 almost always obtains a higher MAP than MSDT-ASA2 and kNN-ASA2 for the same time, which demonstrates the higher efficiency of CSP. Fig. 13 also shows that a larger number of levels L results in a longer processing time. Given the number n of points in an image, the complexity of DT is $O(n \log n)$. Given the number L of levels in a CSP, our approach takes $O(nL \log n - n \log L!)$ time for neighborhood association. If we fix L, $O(nL \log n - n \log L!)$ becomes linearithmic. We can see that CSP-ASA2 is much less complex than greedy algorithms, e.g. $O(n^2)$ for original k-NN and approximated β-skeleton. It is comparable to $O(n \log n)$ for the approximated k-NN used in kNN-ASA2, but the speed is usually faster because the single operation of distance computation in k-NN is less efficient than the comparison operation in DT. MSDT-ASA2 is much faster than both kNN-ASA2 and CSP-ASA2, but the MAP could not come up to the others.

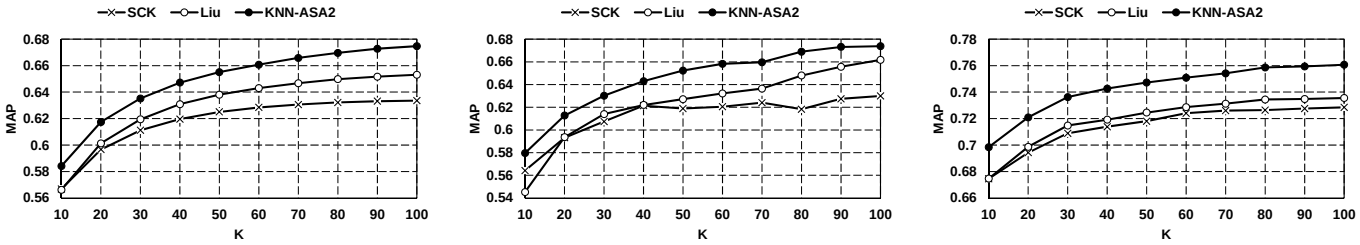
A larger L also results in better neighborhood association, and so results in a higher MAP, especially for FL32 and OB. For HD, CSP-ASA2 outperforms kNN-ASA2 with a large MAP gain when L and k are small, but the differential degrades when we enlarge the number of local feature tuples. This is because HD is a scenery dataset and in certain cases, two images may correspond to the same concept but do not contain the same object. If the object is small or two images have no object in common, CSP-ASA2 tends to obtain more false matches than kNN-ASA2. Basically, CSP-ASA2 with a large L is more suitable for dealing with large objects.

Fig. 14 shows the true responses obtained by CSP-ASA2 but wrongly rejected by kNN-ASA2. We can see that compared with Fig. 9b and Fig. 11a, the images in Fig. 14 have greater scale variation. kNN-ASA2 achieves the same level of scale invariance as CSP-ASA2 because they adopt the same geometric constraint. However, kNN-ASA2 may fail to detect distant interest points inside the same object as neighbors if the object is too large. In other words, for the same period of neighborhood association, the neighborhood constraint of kNN-ASA2 is too strict to cover these distant but useful interest points. Spatial neighborhood association may be sensitive to the error or, more precisely, the inconsistency of interest point detection. This inconsistency is usually because of the severe scale variation between images. Our approach may fail if the variation is too severe, otherwise it can successfully deal with this issue if a sufficient number of neighborhoods are taken into account. The first two examples in Fig. 14 are good examples, in which the object in the left image is much larger than the same object in the right image. CSP-ASA2 may lead to more false responses, e.g. the long edge in the third example, than kNN-ASA2, but the discriminating geometric constraint successfully tolerated this problem.

VII. Larger-Scale Experimentation

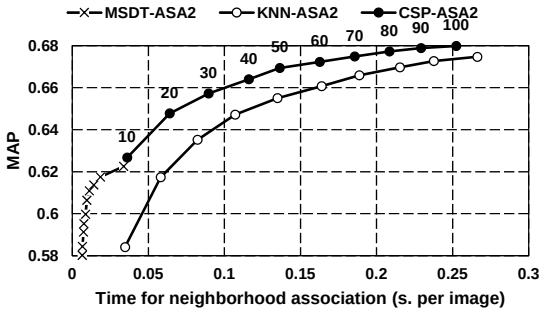
A. Setting

For a larger-scale evaluation, the common practice [12], [15], [19] is to employ a large database as distractors and to include it in a smaller database containing the ground truth. We follow the same scheme and use an unlabeled dataset known as Flickr 100K (F100K) [3] containing 100K images that are assumed not to contain the buildings in Oxford Buildings (OB) [3]. Although the assumption has not been validated, the configuration using F100K as distractors has been widely adopted in previous research [3], [9], [12]–[15], [19]. We put OB and F100K together for testing and use the visual

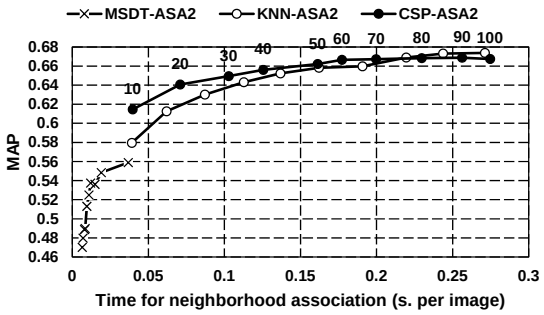


(a) Flickr Logos 32. (b) Holiday. (c) Oxford Buildings.

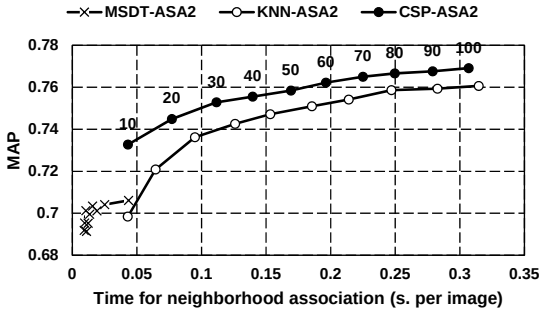
Fig. 12. Relationship between MAP and k used in k-NN.



(a) Flickr Logos 32.



(b) Holiday.



(c) Oxford Buildings.

Fig. 13. Relationship between MAP and time for neighborhood association. Numbers above CSP-ASA2 curves are numbers of levels L.

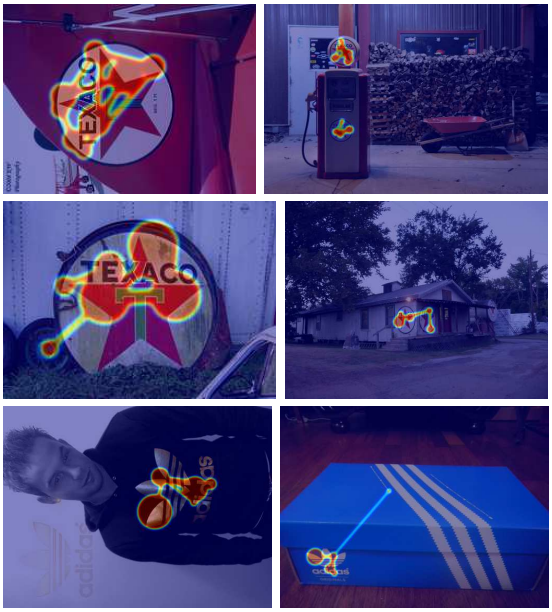


Fig. 14. Responses obtained using CSP-ASA2 with $L = 100$. kNN-ASA2 with $k = 100$ detected none of these true responses.

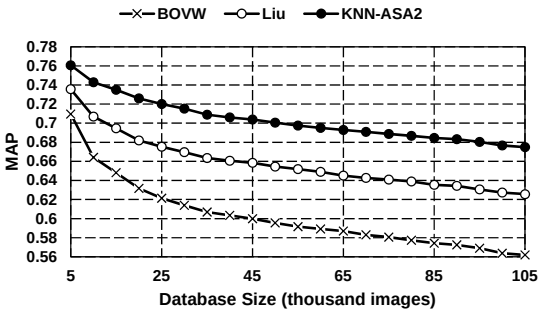


Fig. 15. MAP on OB dataset with various sizes of images as distractors from the ground truth. $k = 100$ in all cases.

vocabulary with 1M clusters built from OB, which is the same as that used in Section VI. We also test Liu's approach [10] in addition to BOVW for comparison because it is the most closely related to our proposal. k is set at 100 for both Liu's approach and our approach. 21 different sizes ranging from 5K to 105K are tested, in which 5K is the size of OB without distractors. The result is shown in Fig. 15.

B. Comparison

The performance degrades gradually as we increase the number of distractors, but it is obvious that the MAP of kNN-ASA2 degrades more smoothly than the others. When 100K images are included, we obtain a 11% MAP improvement of about 11% compared with BOVW and of 5% compared with Liu's approach [10]. We have discussed the low discriminative power of BOVW as regards confusing local features in Section VI-D1. As shown in Fig. 11b, kNN-ASA2 is

TABLE VII
MAP COMPARISON WITH STATE-OF-ART METHODOLOGIES. (%)

Approach	OB	OB+F100K
Our Approach	76.9	67.5
Avrithis and Tolias [25]	78.9	73.0
Shen et al. [23]	75.2	72.9
Qin et al. [42]	73.9	67.8
Avrithis [41]	71.6	65.7
Zhang et al. [9]	71.3	60.4

TABLE VIII
SCALABILITY COMPARISON ON OXFORD BUILDINGS.

Approach	Time[1]	Memory	#Distinct[2]	#Tuple
kNN-ASA2	321	14G	845M	1G
Liu et al. [10]	329	11G	714M	1G
SCK [6]	318	11G	697M	1G
MSDT [5]	62	1G	83M	85M
BOVW	34	231M	14M	18M

[1] Unit: millisecond per query.
[2] #Distinct indicates the total number of entries that must be inserted in the inverted index, and it is never more than #Tuple.

not only more robust but also achieves higher discriminative power than Liu's approach. That means kNN-ASA2 leads to fewer false responses when we employ a large database as distractors. This explains why kNN-ASA2 derived a smoother MAP degradation than the others in Fig. 15.

C. Comparison with State-of-Art Methodologies

For a more comprehensive evaluation, we compare the experimental results obtained with our approach with those reported in previous publications [9], [23], [25], [41], [42]. In all approaches, a specific vocabulary with 1M visual words is learned on all the images of OB and used for indexing and retrieval on the OB and OB+100K datasets. Some approaches [23], [25], [41] used a modified version [43] of the Hessian-affine region detector [30], where a gravity vector assumption is used to estimate the dominant orientation of features for descriptor extraction. This detector has a great advantage over general detectors if the dataset (e.g. OB and OB+F100K) includes no rotated images. Some authors [23], [25] also incorporated the orientation prior into their spatial models for higher MAP. In our approach, the orientation prior is not considered because it does not hold in general scenarios, e.g. in a logo search task.

Table VII presents the MAP of these approaches on OB and OB+F100K. Our approach achieves state-of-the-art MAP on OB, in fact it outperforms all approaches except for Avrithis and Tolias's approach [25]. In the experiment using OB+F100K, our approach is outperformed by Avrithis and Tolias's approach [25] and Shen's approach [23]. In both studies, the authors take the orientation prior into consideration by using the gravity vector assumption and switching off rotation for spatial matching. As a result, these approaches impose an additional constraint on matching such that matched features are disregarded if they differ significantly from each other in terms of orientation. This strategy greatly improves the performance especially when there are a very large number of distractor images. This explains why our approach outperforms Shen's approach [23] on the OB dataset but is inferior to the same approach on OB+F100K. Apart from that, our approach is very competitive and performs reasonably well even without using data-dependent prior knowledge. Note that Avrithis and Tolias's [25] and Shen's methods [23] are robust to uniform transformations but sensitive to anisotropic ones, which is in common with Liu's method [10].

The main limitation of spatial coherence models based on the higher-order structure of local features is their high memory usage, as shown in Table VIII. This problem is common to MSDT [5], SCK [6], Liu's approach [10] and our approach. In our experiments, the times were measured on a 2.93GHz QuadCore processor (single-threaded). Given n as the number of interest points and k as the number of nearest neighbors, SCK, Liu's approach, and our approach require $O(nk/2)$ memory usage. This is around $k/2$ times larger than the $O(n)$ memory usage of BOVW. Selecting $k = 100$ as we did in our experiments means that the spatial coherence models consume 50 times more memory than BOVW. This is a crucial issue if we consider a real application. We shall deal with this issue in our future research.

VIII. CONCLUSION

We have proposed a feature representation approach based on a second-order configuration of local features by extending the conclusion of affine shape adaptation. Image matching based on this approach is highly discriminative and more robust to 6DOF affine transformations. In a test using an FL32 dataset [39], we searched for images containing the same logo in the query with a MAP of 68.0%. This is the highest value yet reported for the retrieval protocols of this dataset, and is more than 8% higher than the second highest reported MAP [40]. The approach proposed for spatial neighborhood association is based on a Centrality-Sensitive Pyramid model. It is more robust as regards errors in interest point detection and description and achieves a higher speed than traditional solutions. Testing using datasets ranging from 1.5K to 105K in size demonstrated the reliability of our approach in the large-scale retrieval of various types of objects. Spatial coherence models such as SCK [6], Liu's approach [10], and our approach require a much larger memory than standard BOVW, which is a crucial issue for real applications. Possible solutions include feature selection and database augmentation. Boosting can be adapted for feature selection if we formulate the retrieval into a classification problem. Database augmentation can also be adapted because useless features likely exist in one image, while useful features are likely to be found in multiple images of the same object. Because we assume rigid objects, it is difficult to extend our method to video tasks such as action recognition where the target is deformable and prone to self-occlusion. However, it will be interesting to extend our method to the task of searching particular objects from videos, usually known as instance search [44], because a video is basically a series of images. These topics constitute our future direction.

REFERENCES

[1] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *ICCV*, 2003, pp. 1470–1477.

[2] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[3] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *CVPR*, 2007.

[4] G. Carneiro and A. D. Jepson, "Flexible spatial configuration of local image features," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 12, pp. 2089–2104, 2007.

[5] Y. Kalantidis, L. G. Pueyo, M. Trevisiol, R. van Zwol, and Y. S. Avrithis, "Scalable triangulation-based logo recognition," in *ICMR*, 2011, p. 20.

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

[7] J. Gao, Y. Hu, J. Liu, and R. Yang, "Unsupervised learning of high-order structural semantics from images," in *ICCV*, 2009, pp. 2122–2129.

[8] Z. Wu, Q. Ke, M. Isard, and J. Sun, "Bundling features for large scale partial-duplicate web image search," in *CVPR*, 2009, pp. 25–32.

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

[10] Z. Liu, H. Li, W. Zhou, and Q. Tian, "Embedding spatial context information into inverted file for large-scale image retrieval," in *ACM Multimedia*, 2012, pp. 199–208.

[11] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman, "Total recall: Automatic query expansion with a generative feature model for object retrieval," in *ICCV*, 2007, pp. 1–8.

[12] O. Chum, A. Mikulík, M. Perdoch, and J. Matas, "Total recall ii: Query expansion revisited," in *CVPR*, 2011, pp. 889–896.

[13] H. Jegou, M. Douze, and C. Schmid, "Hamming embedding and weak geometric consistency for large scale image search," in *ECCV (1)*, 2008, pp. 304–317.

[14] ——, "Improving bag-of-features for large scale image search," *International Journal of Computer Vision*, vol. 87, no. 3, pp. 316–336, 2010.

[15] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Lost in quantization: Improving particular object retrieval in large scale image databases," in *CVPR*, 2008.

[16] H. Jegou, M. Douze, and C. Schmid, "On the burstiness of visual elements," in *CVPR*, 2009, pp. 1169–1176.

[17] C.-Z. Zhu, H. Jegou, and S. Satoh, "Query-adaptive asymmetrical dissimilarities for visual object retrieval," in *The IEEE International Conference on Computer Vision*, 2013.

[18] P. Turcot and D. Lowe, "Better matching with fewer features: The selection of useful features in large database recognition problems," in *2009 IEEE 12th International Conference on Computer Vision Workshops*, 2009, pp. 2109–2116.

[19] R. Arandjelovic and A. Zisserman, "Three things everyone should know to improve object retrieval," in *CVPR*, 2012, pp. 2911–2918.

[20] Z. Ma, Y. Yang, F. Nie, and N. Sebe, "Thinking of images as what they are: Compound matrix regression for image classification," in *IJCAI*, 2013.

[21] S. Poullot, O. Buisson, and M. Crucianu, "Scaling content-based video copy detection to very large databases," *Multimedia Tools Appl.*, vol. 47, no. 2, pp. 279–306, 2010.

[22] W. Zhang, L. Pang, and C.-W. Ngo, "Snap-and-ask: answering multimodal question by naming visual instance," in *ACM Multimedia*, 2012, pp. 609–618.

[23] X. Shen, Z. Lin, J. Brandt, S. Avidan, and Y. Wu, "Object retrieval and localization with spatially-constrained similarity measure and k-NN re-ranking," in *CVPR*, 2012, pp. 3013–3020.

[24] G. Tolias, Y. Kalantidis, Y. S. Avrithis, and S. D. Kollias, "Towards large-scale geometry indexing by feature selection," *Computer Vision and Image Understanding*, vol. 120, pp. 31–45, 2014.

[25] Y. S. Avrithis and G. Tolias, "Hough pyramid matching: Speeded-up geometry re-ranking for large scale image retrieval," *International Journal of Computer Vision*, vol. 107, no. 1, pp. 1–19, 2014.

[26] K. Lebeda, J. Matas, and O. Chum, "Fixing the locally optimized RANSAC," in *BMVC*, 2012, pp. 1–11.

[27] D. Liu, G. Hua, P. A. Viola, and T. Chen, "Integrated feature selection and higher-order spatial feature extraction for object categorization," in *CVPR*, 2008.

[28] T. Lindeberg, "Non-uniform smoothing," in *Scale-Space Theory in Computer Vision*, ser. The Springer International Series in Engineering and Computer Science. Springer US, 1994, vol. 256, pp. 383–394.

[29] ——, "Scale-space," in *Wiley Encyclopedia of Computer Science and Engineering*. John Wiley & Sons, Inc., 2007.

[30] K. Mikolajczyk and C. Schmid, "Scale & affine invariant interest point detectors," *International Journal of Computer Vision*, vol. 60, no. 1, pp. 63–86, 2004.

[31] O. Boiman, E. Shechtman, and M. Irani, "In defense of nearest-neighbor based image classification," in *CVPR*, 2008.

[32] C. J. van Rijsbergen, *Information Retrieval*. Butterworth, 1979.

[33] M. d. Berg, O. Cheong, M. v. Kreveld, and M. Overmars, *Computational Geometry: Algorithms and Applications*, 3rd ed. Santa Clara, CA, USA: Springer-Verlag TELOS, 2008.

[34] D. W. Matula and R. 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.

[35] J. Cardinal, S. Collette, and S. Langerman, "Empty region graphs," *Comput. Geom.*, vol. 42, no. 3, pp. 183–195, 2009.

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

[37] ——, "Locally-scaled spectral clustering using empty region graphs," in *KDD*, 2012, pp. 1330–1338.

[38] P. Bose, J. Cardinal, S. Collette, E. D. Demaine, B. Palop, P. Taslakian, and N. Zeh, "Relaxed Gabriel graphs," in *CCCG*, 2009, pp. 169–172.

[39] S. Romberg, L. G. Pueyo, R. Lienhart, and R. van Zwol, "Scalable logo recognition in real-world images," in *ICMR*, 2011, p. 25.

[40] S. Romberg and R. Lienhart, "Bundle min-hashing," *International Journal of Multimedia Information Retrieval*, pp. 1–17, 2013.

[41] Y. S. Avrithis, "Quantize and conquer: A dimensionality-recursive solution to clustering, vector quantization, and image retrieval," in *ICCV*, 2013, pp. 3024–3031.

[42] D. Qin, C. Wengert, and L. J. V. Gool, "Query adaptive similarity for large scale object retrieval," in *CVPR*, 2013, pp. 1610–1617.

[43] M. Perdoch, O. Chum, and J. Matas, "Efficient representation of local geometry for large scale object retrieval," in *CVPR*, 2009, pp. 9–16.

[44] P. Over *et al.*, "TRECVID 2013 – an overview of the goals, tasks, data, evaluation mechanisms and metrics," in *TRECVID*, 2013.

Xiaomeng Wu received a B.E.E. degree in energy and power engineering from the University of Shanghai for Science and Technology (China) in 2001, and M.S.E.E. and Ph.D. degrees in information science and technology both from the University of Tokyo (Japan) in 2004 and 2007, respectively. He was a research associate at the National Institute of Informatics (Japan) from 2007 to 2013. He joined NTT Communication Science Laboratories (Japan) as a research associate in 2013. His research interests include image processing, information retrieval, multimedia and pattern recognition. He has served on the program committees of IEEE ISIEA and CBMI. He is a member of IEEE and ACM SIGMM.

Kunio Kashino (S'89-M'95-SM'05) received a Ph.D. degree in electrical engineering from the University of Tokyo, Tokyo, Japan, in 1995.

In 1995, he joined NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation, Atsugi, Japan, where he is currently a Senior Research Scientist and Supervisor. He has been working on multimedia information retrieval and music recognition. His research interests include acoustic signal processing, Bayesian information integration, and sound source separation.

Dr. Kashino was awarded the IEEE Transactions on Multimedia Paper Award in 2004.