Robust Spatial Matching as Ensemble of Weak Geometric Relations

Xiaomeng Wu
wu.xiaomeng@lab.ntt.co.jp
Kunio Kashino
kashino.kunio@lab.ntt.co.jp

NTT Corporation
3-1, Morinosato Wakamiya Atsugi-shi
Kanagawa, Japan 243-0198

Abstract

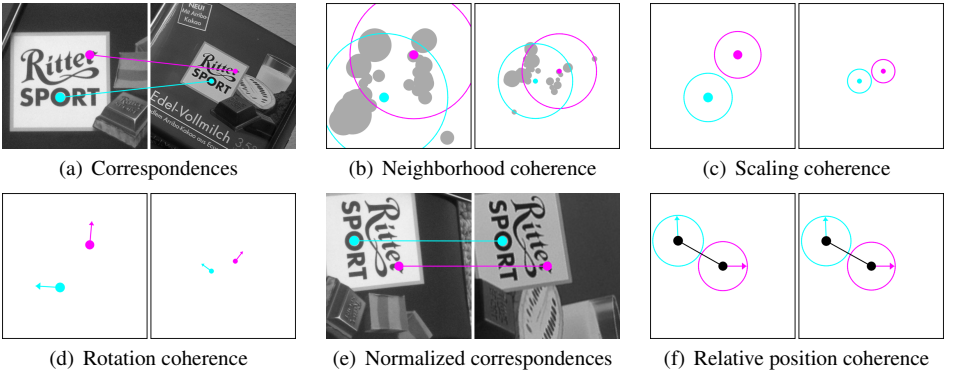
Existing spatial matching methods permit geometrically-stable image matching, but still involve a difficult trade-off between flexibility and discriminative power. To address this issue, we regard spatial matching as an ensemble of geometric relations on a set of feature correspondences. A geometric relation is defined as a set of pairs of correspondences, in which every correspondence is associated with every other correspondence if and only if the pair satisfy a given geometric constraint. We design a novel, unified collection of weak geometric relations that fall into four fundamental classes of geometric coherences in terms of both spatial contexts and between-image transformations. The spatial similarity reduces to the cardinality of the conjunction of all geometric relations. The flexibility of weak geometric relations makes our method robust as regards incorrect rejections of true correspondences, and the conjunctive ensemble provides a high discriminative power in terms of mismatches. Extensive experiments are conducted on five datasets. Besides significant performance gain, our method yields much better scalability than existing methods, and so can be easily integrated into any image retrieval process.

1 Introduction

Local feature-based image encoding [22, 23] has been shown to be successful in particular object retrieval. However, the direct matching of local features (hereafter features) [13, 14] leads to massive mismatches because they do not offer sufficient discriminative power. Spatial matching methods including RANSAC [1, 17, 18], Hough transform [9, 10] and spatial context methods [12, 24] were used to address this issue. In these methods, true correspondences are identified by imposing a constraint on one or two classes of geometric coherences, e.g. in terms of spatial contexts or between-image transformations. These methods are potentially less discriminative due to the limited number of coherence classes [10, 29], while forcibly enhancing the strength of constraints leads to the incorrect rejection of true correspondences [9]. Spatial matching still faces a difficult trade-off between flexibility and discriminative power. This constitutes the main problem we handle in this paper.

We aim at robust and fast spatial matching for the retrieval of near-rigid objects. We characterize spatial matching as an ensemble of geometric relations on the set of feature correspondences. A *correspondence* (Fig. 1(a)) is a pair of features detected from two images and located in immediate proximity to each other in a descriptor space. A *geometric relation*

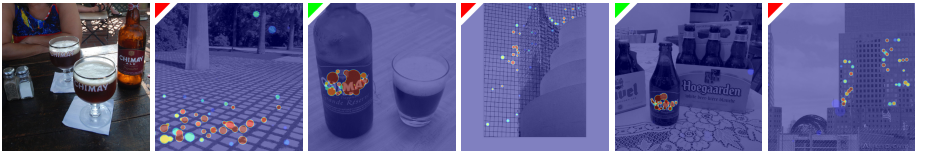
Pages 25.1-25.12
DOI: https://dx.doi.org/10.5244/C.29.25



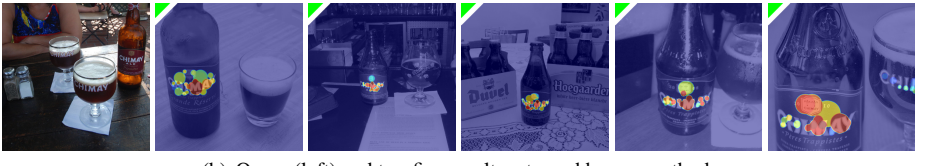
(a) Correspondences (b) Neighborhood coherence (c) Scaling coherence

(d) Rotation coherence (e) Normalized correspondences (f) Relative position coherence

Figure 1: Example of correspondence pair and fundamental classes of geometric coherences. (b) The smallest discs containing the k-NN of each correspondence are shown as open circles. (c)(d)(f) The scales and orientations of and the distances between the correspondences are shown as open circles, arrows and heavy black lines, respectively. (e) Correspondences are normalized in terms of the scale, orientation and coordinates of the magenta correspondence.



(a) Query (left) and top-five results returned by Hough pyramid matching [4].



(b) Query (left) and top-five results returned by our method.

Figure 2: Comparison of Hough pyramid matching [4] and our method. The green and red colors of the upper-left corners of the images indicate positive and negative results, respectively. Correspondences identified by the methods are highlighted in colors.

is a set of pairs of correspondences, in which every correspondence is associated with every other correspondence if and only if the pair satisfy a given geometric constraint. We design a novel, unified collection of multiple weak geometric relations. The relations fall into four fundamental classes of geometric coherences (Figures 1(b)-1(d) and 1(f)), which take both spatial contexts and between-image transformations into consideration. By a *weak geometric relation*, we mean a sufficiently flexible constraint which, nevertheless, may offer only a limited discriminative power. Our goal is to define such relations and to integrate them into a single strong constraint that is well-correlated with the true similarity (Fig. 2).

It is important to note that our method is not based on Hough transform. In contrast to Hough transform-based methods [4, 21] that target at single correspondences in a Hough space, our method directly identifies a set of pairs of correspondences on the basis of carefully designed geometric conditions. Since it does not rely on voting, our method spontaneously avoids the common issue of quantization errors in a Hough transform.

2 Related Research

Spatial matching methods [11, 16, 27, 28, 29, 30] can be categorized as prior or posterior: the former category, corresponding to spatial context methods, improves the discriminative power by embedding geometric information in indexing before matching; the latter rejects mismatches online. As an example of spatial context methods, Liu et al. [12] explored the co-occurrence and relative positions of nearby features, and embedded this information in an inverted index for fast spatial matching. Wu and Kashino [26] extended this method to handle anisotropic transformations. Tolias et al.'s method [24] serves as an alternative to Liu et al.'s method [12], in which each feature is described by a spatial histogram of the relative positions of all other features. Spatial context methods are limited to a reduced accuracy due to quantization of geometric information and has high index space requirements. Posterior matching is the factual solution of choice, where RANSAC and Hough transform dominate.

Exploiting the local shapes of features (e.g. scale, orientation, coordinates) to extrapolate between-image transformations, it is either possible to construct RANSAC hypotheses by single correspondences, or to see correspondences as votes in a transformation space. RANSAC [18] repeatedly computes an affine transformation, called a hypothesis, from each correspondence. All hypotheses are verified by counting the inlier correspondences that inversely fit the transformation. Perdoch et al. [17] proposed approximating RANSAC by vector-quantizing the shapes of features for less memory usage and less online complexity. Arandjelovic and Zisserman [1] used epipolar constraints for RANSAC-based spatial matching. However, RANSAC is known to perform poorly when the percentage of inliers falls much below 50%, e.g. when it comes to the retrieval of small objects. Meanwhile, Jegou et al. [10] used a weak geometric model realized with a 2D Hough transform whereby correspondences are determined as true correspondences if they agree in terms of scaling and, independently, in terms of rotation factor. Shen et al. [21] proposed uniformly sampling a fixed number of similarity transformations (hypotheses) from a transformation space. All hypotheses are verified in another 2D Hough space spanned by the normalized central coordinates of the common object. Avrithis and Tolias [4] followed the conventional practice of exploiting the shapes of features [10, 13]. The method explores a 4D Hough space of complete transformations including scaling, rotation and translation. The key contribution is an elegant pyramid model that distributes correspondences over a hierarchical partition of the transformation space and increases robustness as regards errors in feature detection. Despite exhausting efforts, Hough transform remains sensitive to noise generated during transformation estimation and quantization.

3 Ensemble of Weak Geometric Relations

3.1 Preliminaries

An image is represented by a set P of features. For each feature $p \in P$ we are given its visual word $u(p)$, position $\mathbf{t}(p) = [x(p)\ y(p)]^{\mathrm{T}}$, scale $\sigma(p)$ and orientation $\mathbf{R}(p)$. The geometries can be obtained from an affine covariant feature detector [14, 17] and $u(p)$ by vector quantization in a descriptor space [18, 22]. p can be mapped, from a unit circle heading a reference orientation, by a 3×3 transformation matrix $\mathbf{F}(p)$:

$$\mathbf{F}(p) = \begin{bmatrix} \mathbf{M}(p) & \mathbf{t}(p) \\ \mathbf{0}^{\mathrm{T}} & 1 \end{bmatrix} \tag{1}$$

where $\mathbf{M}(p) = \sigma(p)\mathbf{R}(p)$ is a linear transformation and homogeneous coordinates are to be used for the mapping. If $\sigma(p)$ is given by a real scalar, $\mathbf{F}(p)$ specifies a similarity transformation. $\mathbf{R}(p)$ is an orthogonal 2×2 matrix with $\det \mathbf{R}(p) = 1$, represented by an angle $\theta(p)$. Given two images P and Q, a correspondence $c \triangleq (p,q)$ is a pair of features $p \in P$ and $q \in Q$ with $u(p) = u(q)$. We assume $|C| \geq 2$ with $C = \{c\}$ and:

$$c = \big(u(c), \mathbf{t}(p), \sigma(p), \theta(p), \mathbf{t}(q), \sigma(q), \theta(q)\big). \tag{2}$$

3.2 Problem Formulation

Suppose that P and Q are related as regards a common near-rigid object and an unknown (geometric) transformation $\mathbf{F}$. It can be inferred that all parts of the object obey the same transformation. Therefore, given a correspondence set C constructed from P and Q, there is a subset $C_{\mathbf{F}} \subseteq C$ of correspondences that lie inside the object and show considerable similarity in terms of their local transformations. These local transformations must be close to $\mathbf{F}$. Spatial matching is then to identify such a subset, whose cardinality provides evidence for the belief that P and Q include the same object.

We focus on the Cartesian product $C^2 = C \times C$, i.e. the set of all ordered pairs (c_a, c_b) where $c_a, c_b \in C$. A constraint function $h : C^2 \to \{0,1\}$ is defined, which maps any arbitrary (c_a, c_b) to one if a given geometric constraint is satisfied, and zero otherwise. A geometric relation G is thus a subset of C^2 such that $\forall (c_a, c_b) \in G, h(c_a, c_b) = 1$. If h is sufficiently well-defined and if the geometries in Eq. 2 are accurately given, we have $G \approx C_{\mathbf{F}}^2$. Accordingly, the spatial similarity can be formulated by the cardinality of G instead of that of $C_{\mathbf{F}}$.

Instead of a single constraint h, we build a set $H = \{h\}$ of weak geometric constraints, resulting in a set $\mathcal{G} = \{G\}$ of geometric relations. Each $h \in H$ should be flexible as regards feature detection errors, but is allowed to offer a limited discriminative power. A conjunctive ensemble of such relations (Eq. 3) creates a single strong constraint that is expected to be highly discriminating in terms of mismatches. The spatial similarity thus becomes $|\hat{G}|$.

$$\hat{G} = \bigcap_{G \in \mathcal{G}} G = \left\{ (c_a, c_b) \in C^2 \,\middle|\, \left(\prod_{h \in H} h(c_a, c_b) \right) = 1 \right\} \tag{3}$$

3.3 Weak Geometric Relations

We focus on four fundamental classes of geometric coherence. The classes derive five weak geometric constraints as defined in Equations 4, 7, 8, 10 and 12, respectively.

3.3.1 Neighborhood Coherence

Since true correspondences lie inside the object (no larger than the image), correspondences with a large gap in an image space are more likely to be mismatches. This observation encourages us to employ a spatial neighborhood constraint. Given a feature p, let its k-nearest neighbors (k-NNs) be $\mathcal{N}_k(p)$. The constraint (our first constraint) can thus be described as:

$$h_{\mathcal{N}}(c_a, c_b) = \Big[\big(p_a \in \mathcal{N}_k(p_b)\big) \wedge \big(p_b \in \mathcal{N}_k(p_a)\big) \wedge \big(q_a \in \mathcal{N}_k(q_b)\big) \wedge \big(q_b \in \mathcal{N}_k(q_a)\big) \Big] \tag{4}$$

where the square brackets are Iverson brackets. An example of a 40-NN coherence is shown in Fig. 1(b). In addition to a fair discriminative power, the use of Eq. 4 offers a great advantage in efficiency. By disregarding pairs of non-adjacent correspondences, the complexity of

all subsequent processes can be reduced from $\mathcal{O}(|C^2|)$ to $\mathcal{O}(\min(|C|,k)|C|) \leq \mathcal{O}(k|C|)$. Our method thus operates in linear time in $|C|$ for a fixed k. As for the k-NN search, we use a randomized KD-tree [15], whose complexity is no more than $\mathcal{O}(k|C|\log|C|)$. These complexities do not contradict the discussion in Section 3.2 where we focused on the Cartesian product $C^2 = C \times C$. The computation of our method is dominated by $\mathcal{O}(k|C|)$ in the worst case because the k-NN search is much faster than geometric verifications.

Note that some spatial context methods, e.g. Liu et al.'s method [12] and Wu and Kashino's method [26], imposed the same constraint on pairs of features (rather than pairs of correspondences) before matching. Since in most cases features are more than 10 times larger than correspondences, these methods require much larger memory and search spaces than our method given the same k.

3.3.2 Scaling Coherence

Given $c = (p,q)$, a transformation from q to p is given by $\mathbf{F}(c) = \mathbf{F}(p)\mathbf{F}(q)^{-1}$. It consists of a linear transformation $\mathbf{M}(c) = \sigma(c)\mathbf{R}(c)$ and a translation $\mathbf{t}(c) = \mathbf{t}(p) - \mathbf{M}(c)\mathbf{t}(q)$. Scaling and rotation transformations are $\sigma(c) = \sigma(p)/\sigma(q)$ and $\mathbf{R}(c) = \mathbf{R}(p)\mathbf{R}(q)^{-1}$, respectively. True correspondences should show considerable similarity in terms of their local transformations $\mathbf{F}(c)$. This encourages us to employ a scaling and a rotation (Section 3.3.3) constraints.

The scaling constraint can be represented by $|\log(\sigma(c_a)) - \log(\sigma(c_b))| < \varepsilon_\sigma$ where $\varepsilon_\sigma \in \mathbb{R}^+$ is a threshold. To minimize the sensitivity to parameters, we approximate this constraint by imposing two weaker constraints on scale inequalities. In particular, Equations 5 and 6 define the two constraints in terms of an intra-image h'_σ and a between-image h''_σ scaling.

$$h'_\sigma(c_a,c_b) = \Big[\big(\sigma(p_a) > \sigma(p_b)\big) = \big(\sigma(q_a) > \sigma(q_b)\big)\Big] \tag{5}$$

$$h''_\sigma(c_a,c_b) = \Big[\big(\sigma(p_a) > \sigma(q_a)\big) = \big(\sigma(p_b) > \sigma(q_b)\big)\Big] \tag{6}$$

The overall scaling constraint (our second constraint) is given by:

$$h_\sigma(c_a,c_b) = h'_\sigma(c_a,c_b) \vee h''_\sigma(c_a,c_b). \tag{7}$$

An example of scaling coherence is shown in Fig. 1(c). We can find two minor yet similar intra-image enlargements (with a scaling factor of 1.04) from magenta to cyan correspondences. Two similar between-image enlargements from right to left can also be observed.

3.3.3 Rotation Coherence

Similar to the scaling, a rotation coherence (our third constraint) can be represented by:

$$h_\theta(c_a,c_b) = \Big[\big|\theta(c_a) - \theta(c_b)\big| < \varepsilon_\theta\Big] \tag{8}$$

where $\theta(c) = \theta(p) - \theta(q)$. An example of rotation coherence is shown in Fig. 1(d). Both magenta and cyan features are rotated, from right to left, by an anticlockwise angle of 32.7°.

3.3.4 Relative Position Coherence

If a given c_a is a true correspondence, its local transformation $\mathbf{F}(c_a)$ should be identical to the transformation $\mathbf{F}$ between P and Q. Consequently, P and Q should have the same appearance

if we regard c_a as a reference and normalize the images in terms of $\mathbf{F}(p_a)$ and $\mathbf{F}(q_a)$. Also, the spatial layout of c_a and any other true correspondence c_b should be consistent across P and Q after normalization. This relative position coherence is perfectly reflected in Figures 1(e) and 1(f) where the magenta correspondence serves as the reference.

Given p_a and p_b, let Eq. 9 define the relative position vector heading from p_a to p_b.

$$\mathbf{v}(p_b|p_a) = \mathbf{M}(p_a)^{-1}\big(\mathbf{t}(p_b) - \mathbf{t}(p_a)\big) \tag{9}$$

The relative position coherence (our fourth constraint) can thus be represented by:

$$h_{\mathbf{v}}(c_a,c_b) = \left[\max\Big(\big\|\mathbf{v}(p_b|p_a) - \mathbf{v}(q_b|q_a)\big\|_2, \big\|\mathbf{v}(p_a|p_b) - \mathbf{v}(q_a|q_b)\big\|_2\Big) < \varepsilon_{\mathbf{v}}\right]. \tag{10}$$

The reason of using maximum pooling instead of sum pooling is to effectively reject mismatches that occasionally satisfy either of the asymmetric constraints in Eq. 10. Equation 10 serves as the first constraint of the relative position coherence used in our method.

In addition, we project the relative position vector onto a polar space and impose another constraint on radius and polar angle inequalities:

$$h'_{\mathbf{v}}(c_b|c_a) = \left[\Big(\big(\rho(p_b|p_a) > 1\big) = \big(\rho(q_b|q_a) > 1\big)\Big) \wedge \Big(\big|\theta(p_b|p_a) - \theta(q_b|q_a)\big| < \varepsilon_\theta\Big)\right] \tag{11}$$

where ρ and θ are the radius and polar angle of $\mathbf{v}$, and ε_θ is the same as in Eq. 8. $\rho(p_b|p_a) > 1$ equals $\|\mathbf{t}(p_b) - \mathbf{t}(p_a)\|_2 > \sigma(p_a)$. Equation 11 is an asymmetric constraint. Combining Eq. 11 and its counterpart gives our fifth (symmetric) geometric constraint:

$$h'_{\mathbf{v}}(c_a,c_b) = h'_{\mathbf{v}}(c_b|c_a) \vee h'_{\mathbf{v}}(c_a|c_b). \tag{12}$$

Equation 12 serves as the second constraint of the relative position coherence.

Note that no mention has yet been made of the between-image translation. In this study, we do not directly impose any constraint on the translation coherence because it has been well incorporated in Eq. 10 (see the supplementary material for more detail).

3.4 Discussion

Our geometric constraint collection now becomes $H = \{h_{\mathcal{N}}, h_\sigma, h_\theta, h_{\mathbf{v}}, h'_{\mathbf{v}}\}$. Given a correspondence set C, our method finds the k-NNs of each $c \in C$ in the image space. Each pair of neighboring correspondences is then verified via the other constraints and assigned an integer in $\{0,1\}$ according to whether or not the constraints hold. The spatial similarity is computed on the basis of Eq. 3, and then combined with a non-spatial similarity:

$$S(P,Q) = \begin{cases} |\hat{G}| + 1 & \text{if } |\hat{G}| \neq 0 \\ S'(P,Q) & \text{else.} \end{cases} \tag{13}$$

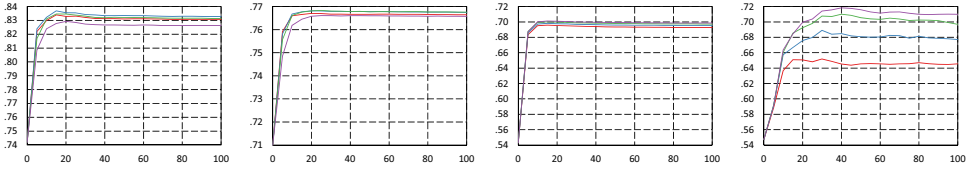
where $S(P,Q)$ is the overall similarity and $S'(P,Q) \in [0,1]$ the non-spatial similarity. We use the cosine similarity between TF-IDF histograms [22] as $S'(P,Q)$, but any local feature-based similarity [3, 23] can be used here. Equation 13 is the equivalent of first ranking the results according to $|\hat{G}|$ and then ranking those with zero similarities via $S'(P,Q)$.

The four classes of geometric coherences are fundamental in the sense that most spatial matching methods are based on one or two of these classes. RANSAC [17, 18] treats a correspondence c_b as an inlier to a geometric model $\mathbf{M}(c_a)$ if (c_a,c_b) satisfies a relative position

Table 1: Dataset comparison [1].

Dataset	Category	#Q	#I	#VW	Detector	Descriptor
OB [18]	Building	55	5.1K	1M	Perdoch et al. [17]	R-SIFT [1]
Paris [19]	Building	55	6.4K	1M	Perdoch et al. [17]	R-SIFT [1]
FL32 [20]	Logo	960	4.3K	1M	Mikolajczyk and Schmid [14]	R-SIFT [1]
Holiday [8]	Scenery	500	1.5K	200K	Mikolajczyk and Schmid [14]	SIFT [13]
F100K [18]	Distractor	n/a	100K	1M	Perdoch et al. [17]	R-SIFT [1]

[1] #Q, #I and #VW are the numbers of queries, images and visual words, respectively. R-SIFT stands for Root SIFT.



(a) OB [18] (b) Paris [19] (c) FL32 [20] (d) Holiday [8]

Figure 3: Relationship between MAP (y-axis) and k (x-axis) used in k-NN. The curves shown in red, blue, green and purple are obtained with $\varepsilon_\theta \in \{\pi, \pi/2, \pi/4, \pi/8\}$, respectively. $\varepsilon_\mathbf{v} = 5$.

constraint; Jegou et al.'s method [10] is a disjunction of scaling and rotation constraints; Liu et al.'s method [12] and Wu and Kashino's method [26] are a conjunction of Equations 4 and 12. More detail on the theoretical relation between current spatial matching methods and our method is given in the supplementary material.

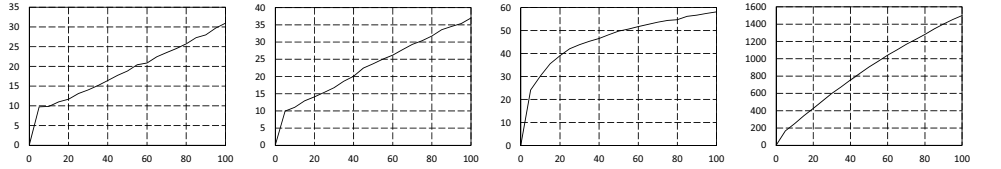
4 Experiments

4.1 Dataset

We tested our method on five datasets: Oxford Buildings (OB) [18], Paris [19], Flickr Logos 32 (FL32) [20], Holiday [8] and Flickr 100K (F100K) [18], which are compared in Table 1. For OB, Paris and F100K, we conformed to a widely-used configuration [4, 21] that assumed the datasets include no rotated images. For such datasets, we switched off rotation for feature detection and spatial matching. We used the feature set (SIFT [13]) and the visual vocabulary officially provided by Jegou et al. [8] for the Holiday dataset. For the other datasets, a visual vocabulary was built for each dataset via approximate k-means [18]. For instance, the vocabulary of the Paris dataset was trained on Paris itself. We measured the accuracy via mean average precision (MAP) [25]. All methods were implemented in single threads via C++ on a 3GHz CPU. We measured the memory use in terms of peak resident set size (PRSS). We excluded the time for feature detection and quantization from the evaluation since it is independent of the database size.

4.2 Parametric Analysis

We explored the dependence of the performance on the three parameters used in our method. They are the k used in k-NN (Eq. 4) and the two thresholds $\varepsilon_\theta \in (0, \pi)$ and $\varepsilon_\mathbf{v} \in \mathbb{R}^+$ used in Equations 8 and 10, respectively. Figures 3 and 4 show the relationship between the retrieval performance and $k \in \{10, 20, \cdots, 100\}$ with $\varepsilon_\theta \in \{\pi, \pi/2, \pi/4, \pi/8\}$. We can see that the MAP



(a) OB [18] (b) Paris [19] (c) FL32 [20] (d) Holiday [8]

Figure 4: Relationship between ε_θ-averaged search time (y-axis: msec per query and per 1K images) and k (x-axis) used in k-NN. $\varepsilon_\mathbf{v} = 5$.

Table 2: Performance comparison [1].

Methods	Oxford Buildings [18]			Paris [19]		
	MAP	PRSS	Time	MAP	PRSS	Time
BOVW [22]	.742	36M	.1	.710	32M	.1
Yang and Newsam [29]	.774	15G	72.7	.733	13G	59.6
Liu et al. [12]	.775	15G	45.8	.731	13G	44.7
Wu and Kashino [26]	.784	8G	69.4	.735	9G	79.6
HPM [4]	.794	70M	60.2	.729	**66M**	67.3
Our Method	**.827**	**69M**	**19.0**	**.766**	**66M**	**20.2**

Methods	Flickr Logos 32 [20]			Holiday [8]		
	MAP	PRSS	Time	MAP	PRSS	Time
BOVW [22]	.543	36M	.2	.547	35M	.9
Yang and Newsam [29]	.634	10G	58.2	.630	13G	**288.6**
Liu et al. [12]	.653	11G	59.4	.662	13G	269.9
Wu and Kashino [26]	.675	7G	93.7	.674	8G	438.9
HPM [4]	.614	**90M**	91.6			
Our Method	**.700**	**90M**	**42.2**	**.714**	**203M**	745.1

[1] All PRSSs are in increments of bytes per 1K images. All times are in increments of msec per query and per 1K images. The best performance among spatial matching methods is highlighted in bold.

is highly insensitive to ε_θ except for Holiday. As discussed in Section 3.3.1, the worst-case complexity of our method is $\mathcal{O}(k|C|)$ and so it is linear in terms of k for a fixed $|C|$. This is well reflected in Fig. 4. Searching Holiday was much slower than searching the other three datasets because the smaller visual vocabulary used for Holiday (Table 1) led to many more tentative correspondences being required for constraint checking. This also explains the exception of the ε_θ-sensitivity of our method for this dataset. We also compared MAPs obtained with various $\varepsilon_\mathbf{v} \in \{5, 10, 15, 20\}$, and the best MAP was achieved with $\varepsilon_\mathbf{v} = 5$ for all datasets. Instead of performing dataset-dependent tuning, we chose $\{k, \varepsilon_\theta, \varepsilon_\mathbf{v}\} = \{40, \pi/8, 5\}$ for all subsequent evaluations and for all datasets.

4.3 Evaluation and Comparison

We compared our method with the bag-of-visual-words (BOVW) method [22], three prior (spatial context) methods [12, 26, 29] and a posterior method called Hough pyramid matching (HPM) [4]. Other methods such as RANSAC [18] and Jegou et al.'s method [10] were not tested because they were reported to underperform HPM [4]. Table 2 compares the per-

formance obtained with various methods. Note that the results shown here were obtained with our own re-implementations for all competing methods. The highest MAPs were obtained with $k = 100$ for the k-NN used in all prior methods. For HPM, the best performance stabilized at five levels. The results obtained with the methods compared in Table 2 are even higher than those, e.g. .789 MAP and 210 msec for HPM (OB), reported in the literature [4, 26]. This demonstrates the propriety of our implementation.

Our method outperformed all the other methods in terms of accuracy. HPM obtained the second highest MAPs for OB and Paris, but could not match the others for FL32. This dataset includes rotated images, and so a full similarity transformation has to be considered [1]. The quantization led to 65,536 bins, making the Hough transform used in HPM very sensitive to feature detection errors. Even if a reasonable balance between flexibility and accuracy can be expected at the finest level of HPM, it is not guaranteed at coarse levels where the constraints are much less discriminating in terms of mismatches. Another reason lies in the small scale of the object (only 5% of the image) in FL32 . An example of a query and the top-five results returned by HPM and our method are shown in Fig. 2 (see the supplementary material for more examples).

In Table 2, posterior methods showed much less memory use than prior methods. Posterior methods operate in linear space as regards the number of features $|P|$, while prior methods in linear space as regards $k|P|$ with $k = 100$ being the parameter of k-NN. For HPM and our method, it is possible to process 1M images (up to 90GB) in a single thread via a CPU with 128GB memory. The large PRSS consumed by our method on Holiday is again because of the small visual vocabulary and in consequence the large number of tentative correspondences. This also explains the longer search time (linear in terms of $|C|$) of our method compared with prior methods for Holiday. In most cases, posterior methods are even faster than prior methods, which serves as a counter-example of the hypothesis behind prior methods [12, 26] (Section 1). The time consumption of prior methods derives from the large search space $k|P|$ composed of massive redundant features.

In our experiment, HPM suffered from long processing time due to recursive verifications of a one-one constraint (see Algorithm 2 in Avrithis and Tolias's paper [4] for more detail). The issue becomes significant at coarse levels when the Hough space is divided into larger bins (more verifications per bin). It is true that our method is in linear time not only in the number of correspondences $|C|$ but also in the number of neighbors k. However, it could achieve high MAPs with only a small $k = 40$ (Section 4.2). Therefore, HPM appeared to be slower than our method.

We included the F100K distractor dataset in OB for a larger scale examination. As shown in Fig. 5, the MAPs degrade gradually as we increase the number of distractors, but it is clear that the degradation with our method is much smoother than with the others. When all the distractors were included, we obtained a MAP improvement of 19% over BOVW and of 6% over HPM. Table 3 presents the reported MAPs of spatial matching methods on the OB, Paris and OB+F100K datasets, where Holiday is not taken into account because its uses in related works lack coherence. Note that since various detector-descriptor combinations were used in the related works [2], Table 3 is only a reference for readers who may be interested in the positioning of our method in the literature. Our method outperforms all methods on all datasets. Our search time per query and per 1K images was 19.0 msec for OB. The corre-

[1] Note that Avrithis and Tolias [4] and Shen et al. [21] did not assume a full similarity transformation when using Perdoch et al.'s feature detector [14].

[2] All the methods used Hessian affine feature detector [14] except that Perdoch et al. [17] and Shen et al. [21] used a modified one [17]; all the methods used SIFT [13] except that Arandjelovic and Zisserman [2] used R-SIFT [2].

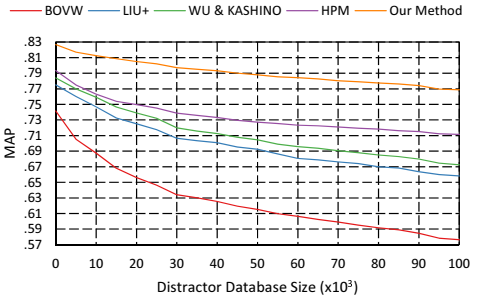


Figure 5: MAPs on OB+F100K.

Methods	OB	Paris	OB+F100K
Our Method	**.827**	**.766**	**.769**
Perdoch et al. [17]	.789	n/a	.726
Shen et al. [21]	.752	.741	.729
Arandjelovic [1]	.720	n/a	.642
Zhang et al. [30]	.713	n/a	.604
Cao et al. [5]	.661	.632	n/a

Table 3: Reported MAPs.

sponding time reported by Perdoch et al. [17] was 238 msec on 4 cores, and that reported by Shen et al. [21] was 17.6 msec. This reveals the high competitiveness of our scalability.

5 Conclusion

We have characterized spatial matching as identifying a subset, called a geometric relation, of the Cartesian product of a correspondence set. This relation is modeled as a conjunctive ensemble of multiple weak geometric relations, taking both spatial contexts and between-image transformations into consideration. Our method achieves a better trade-off between flexibility and discriminative power. Testing using five datasets ranging from 1.5K to 105K in size demonstrated the great superiority of our method with respect to the state of the art. Our method can be integrated in a retrieval system with other components such as query expansion [6, 7] and query adaptation [9, 31] to provide better object and image retrieval. Note that our method can easily estimate the underlying geometric transformation between images by identifying the most frequent correspondence in the conjunctive ensemble $\hat{G}$ (Eq. 3). The estimate is useful in query expansion, which has been shown to significantly improve the results. We recognize this as our future subject.

References

[1] Relja Arandjelovic and Andrew Zisserman. Efficient image retrieval for 3D structures. In *BMVC*, pages 1–11, 2010.

[2] Relja Arandjelovic and Andrew Zisserman. Three things everyone should know to improve object retrieval. In *CVPR*, pages 2911–2918, 2012.

[3] Relja Arandjelovic and Andrew Zisserman. All about VLAD. In *CVPR*, pages 1578–1585, 2013.

[4] Yannis S. Avrithis and Giorgos Tolias. Hough pyramid matching: Speeded-up geometry re-ranking for large scale image retrieval. *International Journal of Computer Vision*, 107(1):1–19, 2014.

[5] Yang Cao, Changhu Wang, Zhiwei Li, Liqing Zhang, and Lei Zhang. Spatial-bag-of-features. In *CVPR*, pages 3352–3359, 2010.

[6] Ondrej Chum, James Philbin, Josef Sivic, Michael Isard, and Andrew Zisserman. Total recall: Automatic query expansion with a generative feature model for object retrieval. In *ICCV*, pages 1–8, 2007.

[7] Ondrej Chum, Andrej Mikulík, Michal Perdoch, and Jiri Matas. Total recall II: Query expansion revisited. In *CVPR*, pages 889–896, 2011.

[8] Herve Jegou, Matthijs Douze, and Cordelia Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *ECCV*, pages 304–317, 2008.

[9] Herve Jegou, Matthijs Douze, and Cordelia Schmid. On the burstiness of visual elements. In *CVPR*, pages 1169–1176, 2009.

[10] Herve Jegou, Matthijs Douze, and Cordelia Schmid. Improving bag-of-features for large scale image search. *International Journal of Computer Vision*, 87(3):316–336, 2010.

[11] Yuning Jiang, Jingjing Meng, and Junsong Yuan. Randomized visual phrases for object search. In *CVPR*, pages 3100–3107, 2012.

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

[13] David G. Lowe. Distinctive image features from scale invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[14] Krystian Mikolajczyk and Cordelia Schmid. Scale & affine invariant interest point detectors. *International Journal of Computer Vision*, 60(1):63–86, 2004.

[15] Marius Muja and David G. Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. In *VISAPP*, pages 331–340, 2009.

[16] Shanmin Pang, Jianru Xue, Nanning Zheng, and Qi Tian. Locality preserving verification for image search. In *ACM Multimedia*, pages 529–532, 2013.

[17] Michal Perdoch, Ondrej Chum, and Jiri Matas. Efficient representation of local geometry for large scale object retrieval. In *CVPR*, pages 9–16, 2009.

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

[19] 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.

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

[21] Xiaohui Shen, Zhe Lin, Jonathan Brandt, and Ying Wu. Spatially-constrained similarity measure for large-scale object retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(6): 1229–1241, 2014.

[22] Josef Sivic and Andrew Zisserman. Video Google: A text retrieval approach to object matching in videos. In *ICCV*, pages 1470–1477, 2003.

[23] Giorgos Tolias, Yannis S. Avrithis, and Hervé Jégou. To aggregate or not to aggregate: Selective match kernels for image search. In *ICCV*, pages 1401–1408, 2013.

[24] Giorgos Tolias, Yannis Kalantidis, Yannis S. Avrithis, and Stefanos D. Kollias. Towards large-scale geometry indexing by feature selection. *Computer Vision and Image Understanding*, 120:31–45, 2014.

[25] Andrew Turpin and Falk Scholer. User performance versus precision measures for simple search tasks. In *SIGIR*, pages 11–18, 2006.

[26] Xiaomeng Wu and Kunio Kashino. Image retrieval based on anisotropic scaling and shearing invariant geometric coherence. In *ICPR*, pages 3951–3956, 2014.

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

[28] Hongtao Xie, Ke Gao, Yongdong Zhang, Jintao Li, and Yizhi Liu. Pairwise weak geometric consistency for large scale image search. In *ICMR*, page 42, 2011.

[29] Yi Yang and Shawn Newsam. Spatial pyramid co-occurrence for image classification. In *ICCV*, pages 1465–1472, 2011.

[30] Yimeng Zhang, Zhaoyin Jia, and Tsuhan Chen. Image retrieval with geometry-preserving visual phrases. In *CVPR*, pages 809–816, 2011.

[31] Cai-Zhi Zhu, Herve Jegou, and Shin'ichi Satoh. Query-adaptive asymmetrical dissimilarities for visual object retrieval. In *ICCV*, pages 1705–1712, 2013.