Exploiting descriptor distances for precise image search

Hervé Jégou, Matthijs Douze, Cordelia Schmid

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Exploiting descriptor distances for precise image search

Herveé Jégou — Matthijs Douze — Cordelia Schmid

N° 7656

June 2011

Thème COG

apport de recherche

ISSN 0249-6399 ISRN INRIA/RR--7656--FR+ENG

INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE

INRIA

centre de recherche
RENNES - BRETAGNE ATLANTIQUE

Exploiting descriptor distances for precise image search

Herveé Jégou* , Matthijs Douze , Cordelia Schmid

Thème COG — Systèmes cognitifs
Équipe-Projet Texmex

Rapport de recherche n° 7656 — June 2011 — 18 pages

Abstract: This report addresses precise image search based on local descriptors. Our approach extends a k-NN voting scheme in several ways. First, we introduce a query-adaptive criterion that is shown effective to weight the descriptor matches. Second, we exploit the distances between SIFT descriptors and the reciprocal neighbors to further refine the similarity measure between descriptors.

Each of these two complementary strategies leads to a significant improvement over the usual voting baseline, and significantly outperforms bag-of-features, at the cost of a very high computational and memory complexity due to the exact computation of distances and reciprocal nearest neighbors. In order to make our method tractable, we exploit an approximate search method which, in addition to returning nearest neighbors with high probability, provides precise distance estimates without accessing the full raw vectors, which is critical to avoid memory issues.

Experimental results show that our method outperforms the state of the art on four challenging datasets. Although our method is not as efficient as bag-of-features, we show that it can handle a database of up to 1 million images with reasonable query times.

Key-words: image search, image retrieval, nearest neighbor, reciprocal nearest neighbors

* Corresponding author. H. Jégou is with INRIA Rennes-Bretagne Atlantique. M. Douze and Cordelia Schmid are with INRIA Grenoble Rhône-Alpes.

Centre de recherche INRIA Rennes – Bretagne Atlantique
IRISA, Campus universitaire de Beaulieu, 35042 Rennes Cedex
Téléphone : +33 2 99 84 71 00 — Télécopie : +33 2 99 84 71 71

Amélioration de la précision en recherche d'image par l'utilisation des distances entre descriptors locaux

Résumé : Ce rapport considère le problème de la recherche d'image à partir de descripteurs locaux. Notre approche étend un système de vote par k plus proches voisins de plusieurs manières. Tout d'abord, nous introduisons un critère adaptatif, dérivé des distances associés aux plus proches voisins, afin de pondérer la qualité des appariemments. Nous exploitons ensuite les plus proches voisins réciproques et les distances associées pour améliorer la similarité entre chaque descripteur SIFT requête et ses voisins.

Chacune de ces deux méthodes apporte un gain significatif par rapport à la référence du système de vote initial, et fournit des résultats supérieurs à une approche par sac-de-mots. Cependant, elle nécessite le calcul coûteux, en mémoire et en CPU, des distances et du graphe des plus proches voisins associé à la base. Afin de rendre notre méthode utilisable à une plus grande échelle, nous utilisons une méthode de recherche approximative récente qui estime les distances entre les vecteurs requêtes et ceux de la base, sans avoir à stocker en mémoire la représentation pleine de descripteurs.

Nos expériences montrent que cette méthode approchée dépasse largement l'état de l'art sur 4 jeux de données couramment utilisés en recherche d'image. Bien qu'elle ne soit pas aussi efficace qu'une approche par sac-de-mots, nous montrons qu'elle reste utilisable pour une base comprenant jusqu'à 1 million d'images.

Mots-clés : recherche d'image, plus proches voisins, graphe de plus proche voisins

1 Introduction

Content based image search is a very active field. In recent years, the size of databases that could be handled increased dramatically. In this context, the bag of visual words (BOVW) framework [18, 26] and its recent extensions were shown to provide a good precision with high efficiency. In particular, precision was improved by adding binary codes [6] or by using soft-assignment in combination with large visual vocabularies [22]. Another way to improve matching within a BOVW framework is to learn descriptor projections adapted to the data [23, 28]. Probabilistic connections between visual words are advantageously learned to improve the image similarity in the context of large visual vocabularies [16].

The goal of this paper is to improve the precision by extending descriptor matching methods [13, 24]. High precision is required in applications such as grasping a limited set of objects [12], robot localization [25] or for re-ranking the results output by a large scale indexing system. Improving precision is especially necessary in the case of clutter and cropping, where the standard BOVW does no longer correctly measure the distance between images due to the small number of inliers and likewise the high rate of incorrect matches. In particular, recognizing locations or particular object instances on cropped images remains a challenging task. Although BOVW coupled with geometrical verification in a re-ranking stage partially addresses this issue, for large datasets a large proportion of images has to be verified, which severely impacts the efficiency and requires additional geometrical information[1]. In order to limit the number of images to be spatially verified, it is worth having a stronger matching system for the first stage. An image search system seeking high precision includes, in general, a geometric post-verification scheme [13, 21, 6], which can be combined with query expansion [2], if multiple relevant images are expected.

In this paper, we exploit the information provided by the distances between local descriptors. Except a few systems that perform feature selection [12, 27], the descriptor matches are usually assumed equal. Considering a k-NN voting framework, we first investigate how to use the distances to nearest descriptors to improve the voting quality, and propose a query-adaptive criterion extracted for the k-NN list. It is obtained by comparing the distance of the $k-1$ first neighbors with that of the k-th nearest neighbor. The informativeness of this criterion is validated by a mutual information analysis, which shows that the proposed quantity conveys more information about the correct image than ranks or absolute distances. This criterion is further combined with re-weighting techniques originally introduced in the BOVW framework [5] to improve precision.

As a second contribution, we exploit the distances associated with the reciprocal nearest neighbors (RNN). These RNN are the k-NN of the database descriptors within the database itself. These distances are exploited to further improve the quality of the score associated with each query. Indeed, the k-NN neighborhood is asymmetric: x being a k-NN of y does not necessarily mean that y is a k-NN of x. If a database vector y is a k-NN of the query x, then it is more likely to be a reliable match if x is also a k-NN of y if x would be added to the database.

A related work is the contextual dissimilarity measure [9], which proposes to regularize the distance between BOVW vectors by improving the reciprocality

[1] This representation may be compressed to a few dozen bits, as shown in [19].

of the neighborhood of BOVW vectors. Similarly, reciprocal nearest neighbors were exploited in [3], again on BOVW vectors.

In contrast to these works, our method exploits the reciprocal nearest neighbors of SIFT descriptors to improve the reliability measure of each individual match. This leads to a significant improvement when this confidence measure is used to adjust the voting weights. Both aforementionned methods assume that exact distances to nearest neighbors are available, as well as pre-computed nearest neighbors of each database vector. Moreover, they make use of raw descriptors, and require a lot of memory. These methods are therefore limited to relatively small datasets (a few thousands of images).

For these reasons, we propose an approximation of our method to trade precision against efficiency. It consists in first using the approximate search technique of [10], which, in addition to returning reliable nearest neighbors, produces distance estimates with sufficient precision. Second, we estimate based on an external dataset the typical distances associated with the RNN, which is the only information required by our scheme.

To the best of our knowledge, no other technique could handle nearest-neighbor searches of local descriptors with the required memory compactness, precision and search speed, and in addition providing exploitable distance estimates: LSH [1] and kNN trees [17] both require access to the raw descriptors of the whole dataset in a costly re-ranking stage.

The paper is organized as follows. Section 2 introduces the datasets and the image description scheme used in the evaluation. Section 3 presents the voting framework, introduces our query-adaptive criterion and our RNN method. Section 4 shows how to obtain a scalable approximation of our method, which is evaluated in Section 5 in terms of precision and efficiency and compared with the state of the art.

2 Datasets and image description

This section briefly presents the datasets and descriptors used for evaluating our approach. We use the descriptors/software available on-line, such that our results are directly comparable to those reported in the literature.

2.1 Oxford and Paris

The Oxford dataset [21] consists of 5062 images of building and 55 query images corresponding to 11 distinct buildings in Oxford. All queries are defined by a rectangle delimiting the building and are in "upright" orientation. The search quality is measured by the mean average precision (mAP) computed over the 55 queries, as defined in [21]. Images are annotated as either relevant, not relevant, or *junk*, which indicates that it is unclear whether a user would consider the image as relevant or not. These *junk* images are removed from the ranking before computing the mAP. The Paris dataset [22] consists of 6412 images collected from Flickr by searching for particular Paris landmarks. The definition of positives, their statistics, and the evaluation protocol is the same as for Oxford.

We use the Paris dataset to learn the parameters used for the Oxford dataset, and vice versa. Although learning the parameters on the dataset itself is reported to significantly improves the results [9, 18, 22], this does not reflect the

performance of the system on a large scale because it is prone to over-fitting. We have used the image descriptors available on-line[2]. They were obtained by extracting interest regions with the Hessian-affine detector [15] and describing them with SIFT descriptors [13].

2.2 Holidays and Flickr

The Holidays dataset [6] is composed of 1491 high resolution personal photos of different locations and objects, 500 of them being used as queries. The search quality is measured by mAP, with the query removed from the ranked list. For determining the parameters, we have used an independent dataset provided together with Holidays, denoted by Flickr60K. Large scale evaluation is performed by adding a set of 1 million images collected from Flickr referred to by Flickr1M and also used in [6] for large scale evaluation.

For all three datasets we have used the pre-computed features available on-line[3]. These features have been extracted with the Hessian-affine detector [15] and described by SIFT [13]. Note that the features are rotation invariant.

2.3 University of Kentucky dataset (UKB)

The UKB object recognition benchmark consists of a dataset of 10200 images representing 2550 objects or scenes (4 images per object). All images are used as queries, and the standard performance measure reported in the literature is the average number of images that are correctly ranked in the first four positions. As for the Holidays dataset we used the Hessian-affine detector to extract features and characterize them by rotation invariant SIFT descriptors. We used the same descriptors as in [9], which gives better results than using those of [6] (only the cornerness threshold is different, because the images are smaller).

3 Beyond majority vote

In this section, we consider a voting approach, where an indexing system returns a set of k-NN hypotheses for each query descriptor. We analyze how to better exploit this k-NN short-list to improve search precision, i.e., how to go beyond a simple majority vote.

For the purpose of analysis, we consider in this section that the k-NN search is exact: the true Euclidean nearest neighbors of the query descriptors are used. The way we leverage the use of exact nearest neighbors is presented in Section 4, where we introduce our approximate strategy and evaluate the trade-off between precision, memory and efficiency.

3.1 Voting criteria

We consider a set of m database images, each described by n_j descriptors ($j = 1 \dots m$). The set of vectors for the whole database is denoted by $\mathcal{Y} = \{y_1, ..., y_i, ...y_n\}$, where $n = \sum_j n_j$. The image associated with descriptor y_i is

[2] `http://www.robots.ox.ac.uk/~vgg/data/oxbuildings` and `http://www.robots.ox.ac.uk/~vgg/data/parisbuildings`

[3] `http://lear.inrialpes.fr/~jegou/data.php`

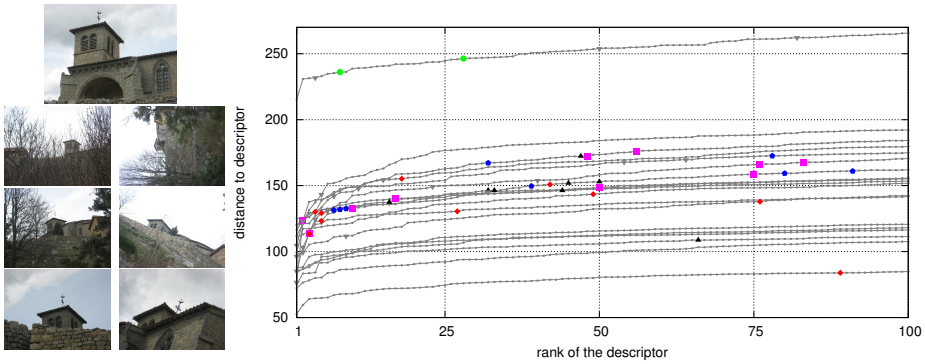


Figure 1: Illustrating descriptor distances. *Left*: a query (top-left) and corresponding images in the Holidays dataset (below). *Right*: distances of the 100-NN for 20 query descriptors (one curve per query). The true matches associated with one of the 6 database images are represented by bigger points (one color and point type per image). A filtering or weighting method based on absolute distances is unreliable, as these significantly vary across query descriptors. The rank information seems more informative, but useful for the first ranks only. The NN distance curves are remarkably similar in shape: a vertical translation approximately aligns them. *Best viewed in color.*

denoted by $\mathrm{im}(y_i)$. Let x be a descriptor of the query image. The k-nearest neighbors $\mathcal{N}_k(x)$ in $\mathcal{Y}$ of x are obtained by

$$\mathcal{N}_k(x) = k\text{-}\arg\min_{y_i \in \mathcal{Y}} d(x, y_i), \tag{1}$$

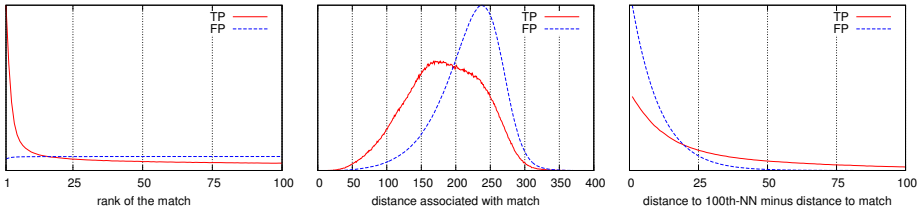
where $d(.,.)$ here refers to the Euclidean distance, without loss of generality. The r-th NN of x is denoted by $N_r(x)$, thus $\mathcal{N}_k(x) = \{N_1(x), \ldots, N_k(x)\}$.

Figure 1 shows, for 20 descriptors extracted from the same query image, the distances $d(x, N_r(x))$, r=1..100, associated with the 100-NN as a function of the rank r. We indicate the descriptors which belong to the corresponding images in the dataset, i.e., which may be correct and should be taken into account. On this small scale example we can make several observations, which are shown in the following to generalize to larger sets based on a quantitative evaluation:

- The absolute distance information is not reliable to distinguish query descriptors which obtain best matches.
- There is a concentration of true matches in the first ranks, but keeping only these first matches, for instance keeping the 10-NN only, ignores many matches associated with the more difficult images.

Once the nearest neighbors are retrieved, several voting strategies are possible to exploit the information provided by the ranked list of neighbors and their distances to query descriptor. In the following, we distinguish several types of voting schemes:

- C_{m}: majority vote, i.e., count the total number of votes for the image, independently of rank/distance information ;



(a) (b) (c)

Figure 2: Empirical distribution function of true (TP) and false (FP) matches for three criteria: the rank C_{r}, the Euclidean distance between descriptors C_{d}, and the proposed query-dependent distance criterion C_{a}.

- C_{r}: exploiting the rank information ;
- C_{d}: exploiting the absolute distance information.

More complex treatments exploiting the distance information of the ordered short-list of k-NN have also been proposed. For instance the distance ratio criterion [13] is used to discard unreliable votes [12, 23]. We observed that it does not significantly improve on Holidays.

Proposed criterion: We introduce a query adaptive criterion C_{a} derived from the distances to neighbors. For a fixed reference rank $r^* \leq k$, a query descriptor x and a descriptor y of the database, it is defined by

$$\delta^{r^*}(x, y) = \max\left(d\left(x, N_{r^*}(x)\right) - d\left(x, y\right), 0\right), \tag{2}$$

where $d\left(x, N_{r^*}(x)\right)$ is the distance between the query descriptor x and its r^*-th NN match.

The value $\delta^{r^*}(x, y)$ is more comparable across queries than the absolute distance $d(x, y)$. It exploits the regularity of the NN distance distribution shown in Figure 1: The criterion amounts to aligning the curves by a translation based on the distance to the reference r^*-th NN. As a result, our query adaptive criterion is better at distinguishing reliable matches from unreliable ones. We typically choose $r^* = k$, in which case the criterion is denoted δ^k.

3.2 Mutual information analysis

In this section we quantitatively measure the amount of information contained in the different voting criteria. To obtain an accurate measure of precision, all statistics presented in this subsection are computed from the 100-NN list of each descriptor of the Holidays dataset. This corresponds to the 1.45 million query descriptors extracted from the 500 query images, exhaustively computed by a linear scan on Holidays, or equivalently to a total of 145 million nearest neighbors.

We will consider descriptor matches as correct iff the images from which they come correspond. Figure 2 shows the empirical probability distribution functions of correct and incorrect retrieved descriptors as a function of a) the rank, b) absolute distance, and c) our query-adaptive criterion. These distributions give an intuitive idea of the properties of these criteria. For instance, the

	criterion		
k	$I(G, C_\mathrm{r})$	$I(G, C_\mathrm{d})$	$I(G, C_\mathrm{a})$
5	0.0210	0.0590	0.1433
10	0.0259	0.0362	0.1033
20	0.0225	0.0220	0.0652
30	0.0189	0.0163	0.0485
50	0.0144	0.0112	0.0329
100	0.0092	0.0066	0.0191

Table 1: Mutual information between the correctness G of a match ranked in k nearest neighbors and the rank C_r, the absolute distance C_d and the proposed criterion C_a. Note that this information is given per match (averaged over k nearest neighbors). It has been measured using the entire Holidays dataset.

rank is reliable for the first ranks, but for $k > 5$ the distribution of correct and incorrect matches is not distinguishable. Absolute distances give a slight prior on whether the match is correct or not. The distributions associated with our query-adaptive criterion significantly differ for true and false matches, suggesting a more reliable selection rule.

To quantitatively measure the informativeness of each criterion, we use mutual information [14], which measures the amount of information shared by two random variables, either discrete (for the rank) or continuous (for absolute distances or our criterion). Let's consider the random variable of quantity $C_\mathrm{x}(X, Y)$ (C_x=C_a, C_r or C_d) computed between query descriptor X and database descriptor Y. Denote by $G = (\mathrm{im}(X), \mathrm{im}(Y)) \in \{0, 1\}$ the random variable which indicates whether (X, Y) is a correct match or not. The conditional mutual information between the correctness of the match and the criterion C_x is computed as

$$I\left(G, C_\mathrm{x}\right) = \sum_{G} \sum_{C_\mathrm{x}} P(G, C_\mathrm{x}) \log_2 \left(\frac{P(G, C_\mathrm{x})}{P(G) P(C_\mathrm{x})} \right). \tag{3}$$

Here, we are only interested in extracting some measures from the k-NN shortlist, i.e., Y is assumed to be k-nearest neighbor of X. Therefore all probabilites are conditional probabilities knowing $Y \in \mathcal{N}_k(X)$. The values of the C_x=C_r and C_a are quantized for the estimation.

Table 1 reports the mutual information measured for C_r, C_d, and C_a. This mutual information is a decreasing function of k, as this measure is *averaged* over all k nearest neighbors: the additional neighbors are less reliable when k increases.

The informativeness of absolute distances is greater than that provided by the ranks for $k < 20$. This is not surprising, because what is measured is the conditional mutual information knowing that the descriptors are ranked in top k positions: the absolute distance provides a more complementary information. Overall, the query-adaptive criterion convey significantly more information about the correctness of the match than the other two.

3.3 Image similarity

Weighting method. We consider the following simple voting strategies. For the majority vote, the score of image b for query q is obtained as

$$s_{\mathrm{m}}(q,b) = \sum_{x \in \mathcal{D}(q)} \sum_{y \in \mathcal{N}_k(x) \cap \mathcal{D}(b)} 1, \tag{4}$$

where $\mathcal{D}(a)$ denotes the set of descriptors extracted from image a. Similarly, we consider a simple linear decreasing rank weighting method:

$$s_{\mathrm{r}}(q,b) = \sum_{x \in \mathcal{D}(q)} \sum_{y \in \mathcal{N}_k(x) \cap \mathcal{D}(b)} k - \mathrm{rank}(y), \tag{5}$$

where $\mathrm{rank}(y)$ refers to the rank of y in $\mathcal{N}_k(x)$. Our query-adaptive criterion is simply summed over matches of images b:

$$s_{\mathrm{a}}(q,b) = \sum_{x \in \mathcal{D}(q)} \sum_{y \in \mathcal{N}_k(x) \cap \mathcal{D}(b)} \delta^k(x,y). \tag{6}$$

It is possible to introduce a soft voting scheme exploiting the probabilities represented in Figure 2. However, we do not adopt such weights because they depend on the dataset and require ground-truth annotations.

Image score normalization: The score obtained by the voting scheme is not regularized. The images with many descriptors are, therefore, favored compared to those with fewer descriptors. Different normalization schemes have been evaluated [22], some of which are applicable in a BOVW framework only, such as the L2 normalization of the BOVW vector. In our voting framework, we evaluate:

- N0: normalization by the number of descriptors,
- SRN: the score is divided by the square root of the number of descriptors, i.e., the normalized similarity $s^*_{\mathrm{X}}(a,b)$ is obtained as

$$s^*_{\mathrm{X}}(a,b) = s_{\mathrm{X}}(a,b) \frac{1}{\sqrt{n}_a \sqrt{n}_b}, \tag{7}$$

  where n_a and n_b correspond to the number of descriptors in image a and b, respectively.

SRN has the desirable property that $s^*_{\mathrm{m}}(a,a) = 1$ if each descriptor of a is matched with itself only, in which case $s_{\mathrm{m}}(a,a) = n_a$. It is, equivalent in our voting framework, to the L2 normalization combined with the cosine similarity in a BOVW framework. However to our knowledge this has not been proposed in the k-NN framework, where the most used metrics are either the number or the rate of inliers.

Figure 3 shows the mAP results measured on Holidays for different combinations of scoring and normalization methods. The proposed criterion C_{a} is significantly better than the majority vote or ranked-based weighting for all tested k. This confirms the findings of the mutual information analysis. The SRN normalization is also consistently better for all weighting schemes. Based

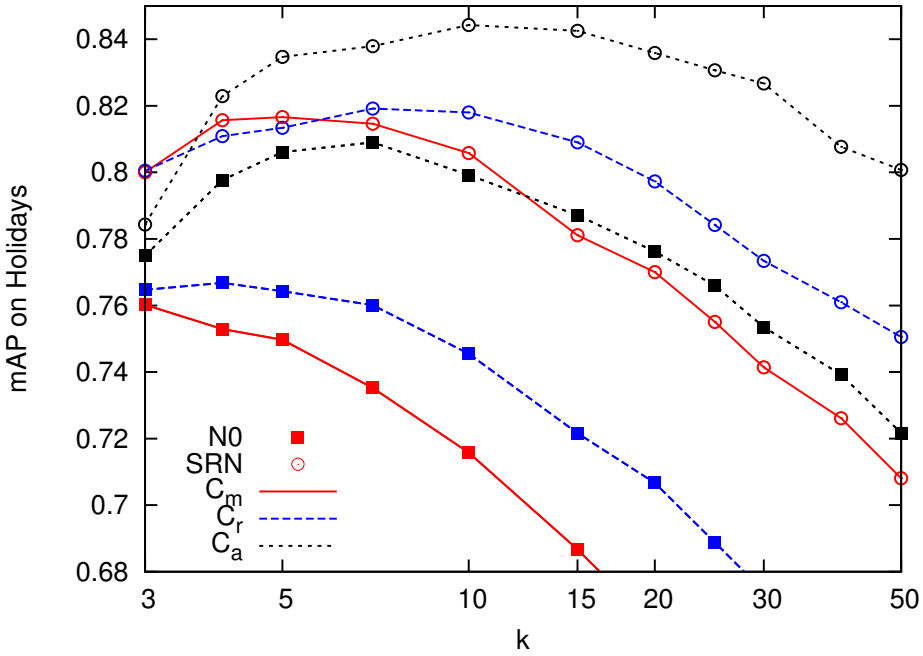


Figure 3: Impact of the scoring and normalization methods on the search accuracy. The line style and color represents the criterion used (C_{m}, C_{r} or C_{a}). The point type represents the kind of normalization (circle=N0, square=SRN). Results are presented for Holidays.

method	k =	10	20	30	40	50
C_a+SRN		84.4	83.6	82.7	80.8	80.1
+burst		84.4	83.9	83.1	82.3	81.6
+burst+reci		**85.2**	85.0	**85.2**	84.8	84.6
+burst+reci+WGC		86.3	86.7	86.7	**86.8**	86.6

Table 2: Impact of the "burstiness" score update [5], of our reciprocal nearest neighbor rule (C_a+SRN) and of a simple geometric check (WGC) [6]. Results are measured by mAP for Holidays.

on the exact Euclidean NNs, the mAP outperforms the state of the art by 7% of mAP: To our knowledge the best result without geometrical information is the mAP score of 77.5% reported in [6]. In the next sections, we will show that our approach still works when considering an approximate strategy using quantized descriptors.

Burstiness: As observed in [5], the visual elements are "bursty" by nature, which means that query descriptors may receive abnormally many votes from the same image. This phenomenon is observable in Figure 1: for the same query descriptor, several descriptors from the same image (same color and point type) are returned in the k-NN list. A simple strategy consists in using multiple match removal [5, 11, 16] to avoid counting multiple times matches obtained for the same query image and the same query descriptor: only the best one is considered. A more advanced strategy is the "intra-burstiness" method proposed in [5]. It is related to multiple match removal, since the scores associated with multiple matches are down-weighted. As shown in Table 2, this strategy slightly improves the results on Holidays for the best SRN +C_a variant, especially for large values of k where multiple matches are expected.

Term-Frequency and Inverse Document Frequency: In a BOVW framework, the inverse document frequency is a popular way of down-weighting the contribution of visual elements that appear more frequently. This strategy is not applicable and unnecessary in our voting framework because for all query descriptors the same number k of database descriptors are retrieved, thereby limiting the effect of unbalanced inverted lists in BOVW.

3.4 Reciprocal nearest neighbors

Unlike descriptor matching based on visual words, k-nearest neighbors are by nature asymmetric: y is a k-NN of x does not imply that x is a k-NN of y. In a BOVW framework, the paper [9] links this property with the observation that some images are very often returned while being irrelevant, and modifies the BOVW comparison metric to make the neighborhood more symmetric, showing that this improves the results.

To our knowledge the problem of neighborhood asymmetry has never been tackled for SIFT features in the context of image search, probably because it raises an important complexity issue: the full graph of k-nearest neighbors of the database must be pre-computed, which is feasible for small datasets only. Nevertheless, when aiming at very precise matching for small datasets, it is interesting to evaluate a reciprocal rule (denoted by *reci*) symmetrizing the

descriptor comparison. For this purpose, we consider the NNs associated with *the database image*. The voting strategy is modified as follows:

- for each $y \in \mathcal{Y}$ the distance to the k^*-NN is computed off-line
- the score is made symmetric by using the weight $\delta^k(x,y)+\delta^k(y,x)$ instead of $\delta^k(x,y)$ (note that $\delta^k(.,.)$ is not symmetric)
- in case the quantity is negative[4], the match is ignored.

As shown in Table 2, the results are further improved on Holidays, and the sensitivity to the parameter k is remarkably reduced. Using larger values of k is useful when geometric constraints are used to improve precision [13, 21], in which case more matches are required. Using a simple implementation of the efficient weak geometric consistency (WGC) check proposed in [6], we obtain a mAP of 86.8% which significantly outperforms the state-of-the-art mAP=84.8% reported in [5] by using a full geometrical verification [13]. The mAP score of 85.2% obtained without any geometrical information significantly outperforms the best score of 77.5% reported in [6]. However, this state of the art performance is obtained using exact matching. The next section addresses both the underlying CPU and memory issue by making use of approximate strategies.

4 Scaling the approach

Although the main objective of this paper is to improve the precision of image search without focusing on larger datasets such as those considered in [20, 8], the approach remains too costly to be applied even on a few thousands images. In this section, we adapt our method to improve its memory and CPU efficiency.

4.1 Approximate nearest neighbors

To improve the efficiency of our scheme, we use an approximate nearest neighbor (ANN) search algorithm, trading accuracy against efficiency. One of the most popular method is the FLANN algorithm [17]. However this package requires to access the full vectors for post-verification, which limits the number of descriptors and therefore of images that can considered without accessing disk storage. Instead, we use a recent extension [10] of the IVFADC method [7] to index local descriptors. It is based on inverted lists and quantization codes associated with the indexed vectors. The key parameters of this approach are:

- The total number of inverted lists. It is fixed to 20000 in our case;
- The number M of inverted lists visited per query descriptor. This parameter has the most impact on efficiency. We use M=1, 10 and 100.
- The number of bytes per descriptor in the inverted file system, which is fixed to 8 in our case (+4 bytes for the image identifier).

[4]This may happen if the query descriptor is not a k-NN of the database descriptor if being inserted in the database vector set: in this case $\delta^k(y,x) < 0$ is negative and the sum $\delta^k(x,y) + \delta^k(y,x)$ may be negative as well.

b	M	k=10	k=20	k=50	k=100
12	1	76.4	75.6	73.0	70.2
12	10	79.0	79.2	77.0	74.9
20	5	82.1	80.2	77.1	74.5
28	10	83.0	81.8	79.2	76.3
44	100	84.3	83.6	81.3	79.0
Exact	-	84.4	83.9	83.1	82.3

Table 3: Impact of the approximate nearest neighbor approach on search quality. Results are presents for Holidays. Values from Table 2 with the same setting (C_{a}+NRS+burst), but using exact search are duplicated for reference.

Although the approach of [7] returns mostly correct neighbors, we observe that the distance estimates obtained from the codes are not very accurate. The extension of [10] leverages this issue by using a vector re-ranking stage that improves the distance estimation. In contrast with most ANN methods, the re-ranking is done using quantized codes instead of using raw SIFT descriptors. The idea consists in encoding the error vector using quantization codes, which introduces an additional memory cost of 0, 8, 16 or 32 bytes, per descriptor, depending of the desired precision. This parameter has an important impact on memory, but little impact on efficiency, as the re-ranking is performed on a short-list of $10k$ neighbors only.

Overall, each descriptor is represented by $b = 12, 20, 28$ or 44 bytes, depending on the refinement parameters.

Table 3 shows the impact of the approximate nearest neighbor search technique. We consider different trade-offs in terms of memory (the b parameter) and efficiency (essentially related to parameter M, although parameter b has a slight impact as well). Timings measured on a large dataset of one million images will be given in Section 5. One can observe that for the best k, the method converges gracefully to the results obtained with exact search.

4.2 Reciprocal nearest neighbor

The main obstacle for using the reciprocal method of Subsection 3.4 on a large scale is the difficulty to obtain the full graph of k-NN. Although some methods (e.g., [4]) provide approximate strategies to break the complexity, this problem remains intractable when considering more than 100 million descriptors.

However, the key quantity exploited by our RNN method is not the full k-NN graph, but the typical distance of each descriptor to its k-th nearest neighbor. To estimate this quantity, our strategy considers an independent dataset of limited size (1 to 10 million descriptors) to estimate a typical k^* nearest neighbor distance associated with each database descriptor. The quantity k^* is not necessarily equal to k because the indexed database and the external dataset may have different sizes. It is set to k^*=20 for all external datasets (Flickr1M, Paris and Oxford).

This external dataset is fixed, and the typical distances associated with a database image are computed when adding it to the index. Storing this recip-

dataset →	UKB	Oxford	Paris	Holidays
SRN +C_{m}	3.46	63.0	64.8	79.2
SRN +C_{a}	3.55	71.3	69.9	84.3
SRN +C_{a}+burst	3.60	73.1	71.0	84.3
SRN +C_{a}+burst+reci	3.59	75.9	72.8	84.0

Table 4: Retrieval results for the four considered evaluation datasets. We set $k = 10$ for all results of Holidays and $k = 100$ for Oxford, Paris and UKB.

rocal distance information takes 2 bytes per descriptor[5]. The Flickr60K dataset is used to compute the typical k-NN distance associated with the database descriptors of Holidays. Paris is used for Oxford and Oxford for Paris.

5 Experiments with approximate neighbors

This section is dedicated to the performance analysis of the more scalable approximate version of our approach presented in Section 4. Unless specified, our results are obtained using $M = 100$ and $b = 44$ bytes per descriptor (+2 bytes if the reciprocal NN rule is used), without using any geometrical information. Our SRN normalization is used in all experiments, as it is consistently better results than N0. Hereafter, we discuss the respective merits of our methods and compare it with the state of the art.

The proposed criterion $C_{\mathbf{a}}$ significantly and consistently improves the results over the majority vote C_{m} on all datasets, see Table 4. The gain in mAP is +8.3% for Oxford, +5.1% for Paris, +5.1% for Holidays. Using the weighting strategy for multiple matches adapted from [5] further improves the results to +10.1%, +6.2% on Oxford and Paris. On Holidays there is no apparent gain for k=10, but this method reduces the sensitivity to k, see Table 3.

The reciprocal k-NN distance brings a significant improvement on both Oxford and Paris. In contrast to Table 2, the results are not improved on Holidays. Our explanation is that the Flickr60K database used to find the reference distances associated with reciprocal neighbors is very different from Holidays and UKB, while the dataset Oxford and Paris are more similar (both includes photos of building downloaded from Flickr).

Impact of k: The main parameter of our approach is the number k of nearest neighbors. The sensitivity of the search quality to this parameter can be judged from Figure 4 and Table 3 for Oxford and for Holidays, respectively. One can observe that the best variant SRN +C_{m}+burst+reci is not very sensitive to the parameter k, which depends on the number m of images in the databases and of the number m_{c} of correct images. An empirical value like $k = m_{\mathrm{c}} \log m$ is a satisfactory choice on the 4 datasets considered.

The comparison with the state of the art is performed by considering the most comparable setups, in particular without considering spatial verification. Our approach clearly outperforms the state of the art in search quality. The

[5]It could probably fit in 1 byte with sufficient precision, but since its memory usage is small compared to the descriptor representation, we store the exact distance.

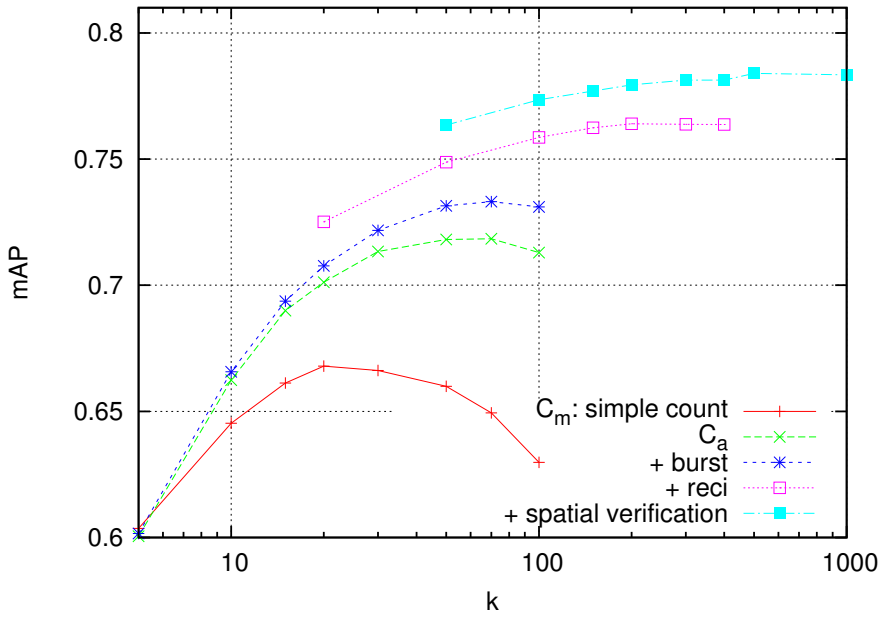


Figure 4: Precision of the proposed approaches on Oxford, depending on the number of neighbors per query point. The results are all provided for SRN normalization and using the approximate ANN method of [10], using 44 bytes per descriptor (raw descriptors are not used during search).

dataset →	Oxford	Paris	UKB	Holidays
Mikulík [16], geometry	*74.2*	*74.9*		*74.9*
Philbin [23], no geometry	66.2	67.8		
Philbin [23], geometry	*70.7*	*68.9*		
Philbin [23], +raw SIFT	*75.5*	*67.2*		
Jégou [9]			**3.68**	
Jégou [6], no geometry	56.1		3.42	77.5
Jégou [5], geometry	*68.5*		*3.64*	*84.8*
ours, no geometry	**76.4**	**72.8**	3.60	**84.4**

Table 5: Comparison between the state of the art and our method used with approximate search (without using raw descriptors). **Bold:** best results reported without geometry. *Italics:* results reported with spatial verification (without query expansion). These results are obtained using descriptors used in the literature. When using more descriptors with our method, we get further improved results (88.8% on Holidays without geometry, 91.5% with spatial verification).

concurrent approaches occasionally reports some better results on a particular dataset, but overall our results are more consistently high are all datasets. Interestingly, without using any geometry, our results on Oxford and Paris (76.4% and 72.8%, respectively) are better than those reported in [23] with full raw descriptors (linear scan), distance ratio testing, and spatial verification performed on *all* images (+0.9% on Oxford and +5.6% on Paris). Finally, we mention that by increasing the number of descriptors, our approach improves further: we obtain on Holidays a mAP of **91.5**% with geometry (**88.8**% without spatial verification) by using 16M instead of 4.5M descriptors.

Timings. We have measured the average time per query when querying the entire Flickr1M dataset merged with Holidays, i.e., 1,001,491 images in total. For this experiment, we allocate 28 bytes per descriptor, so that the structure fits on a server with 64 GB of memory. The corresponding mAP is 70.5%, against mAP=62% in [6] [4] (See [6]-Fig 15, which requires at at least 25GB of memory) and mAP=34% with a BOF representation. The search takes 12s using 20K inverted lists. This timing is obtained for k=20 on an 8-core machine. It does not include descriptor extraction.

Remarks:

1. For reference, exact search on Holidays only (1491 images) takes 710 seconds on the same machine with an optimized linear scan. Our method is therefore four orders of magnitude faster than this linear scan baseline.

2. Due to the relatively high amount of memory used to achieve a significantly improved precision, our approach is more intended to be used on smaller scales than those typically considered in the BOVW framework. However, it easily scales up to 100K images per machine, in which case a query takes about 1s. The evaluation we performed on 1 million images requires a powerful server (64GB of RAM).

6 Conclusion

This paper shows that exploiting distances between local descriptors significantly improves the accuracy of image search. Firstly, we introduce a distance criterion that provides additional information about correct matches. Secondly, we show that reciprocal nearest neighbors efficiently favor the correct matches and further improve the results.

Although our approach is not as scalable as some BOVW methods of the literature, it significantly improves precision, outperforming methods that use geometrical information extensively. An interesting open question is how to better approximate the reciprocal nearest neighbor method proposed in this paper.

References

[1] A. Andoni and P. Indyk. Near-optimal hashing algorithms for near neighbor problem in high dimensions. In *Proceedings of the Symposium on the Foundations of Computer Science*, pages 459–468, 2006.

[2] 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*, October 2007.

[3] Q. Danfeng, S. Gammeter, L. Bossard, T. Quack, and L. V. Gool. Hello neighbor: Accurate object retrieval with k-reciprocal nearest neighbors. In *CVPR*, 2011.

[4] W. Dong, M. Charikar, and K. Li. Efficient k-nearest neighbor graph construction for generic similarity measures. In *WWW*, 2011.

[5] H. Jégou, M. Douze, and C. Schmid. On the burstiness of visual elements. In *CVPR*, June 2009.

[6] H. Jégou, M. Douze, and C. Schmid. Improving bag-of-features for large scale image search. *IJCV*, 87(3):316–336, February 2010.

[7] H. Jégou, M. Douze, and C. Schmid. Product quantization for nearest neighbor search. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 33(1):117–128, January 2011.

[8] H. Jégou, M. Douze, C. Schmid, and P. Pérez. Aggregating local descriptors into a compact image representation. In *CVPR*, June 2010.

[9] H. Jégou, C. Schmid, H. Harzallah, and J. Verbeek. Accurate image search using the contextual dissimilarity measure. *PAMI*, 32(1):2–11, January 2010.

[10] H. Jégou, R. Tavenard, M. Douze, and L. Amsaleg. Searching in one billion vectors: re-rank with source coding. In *ICASSP*, Prague Czech Republic, 2011.

[11] A. Joly and O. Buisson. Logo retrieval with a contrario visual query expansion. In *ACM Multimedia*, pages 581–584, 2010.

[12] H. Kang, M. Hebert, and T. Kanade. Image matching with distinctive visual vocabulary. In *WACV'2011*, January 2011.

[13] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[14] D. MacKay. *Information Theory Inference and Learning Algorithms*, chapter 2. Cambridge University Press, 2006.

[15] K. Mikolajczyk and C. Schmid. Scale and affine invariant interest point detectors. *IJCV*, 60(1):63–86, 2004.

[16] A. Mikulik, M. Perdoch, O. Chum, and J. Matas. Learning a fine vocabulary. In *ECCV*, 2010.

[17] M. Muja and D. G. Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. In *VISAPP*, February 2009.

[18] D. Nistér and H. Stewénius. Scalable recognition with a vocabulary tree. In *CVPR*, pages 2161–2168, June 2006.

[19] M. Perdoch, O. Chum, and J. Matas. Efficient representation of local geometry for large scale object retrieval. In *CVPR*, June 2009.

[20] F. Perronnin, Y. Liu, J. Sanchez, and H. Poirier. Large-scale image retrieval with compressed Fisher vectors. In *CVPR*, June 2010.

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

[22] 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*, June 2008.

[23] J. Philbin, M. Isard, J. Sivic, and A. Zisserman. Descriptor learning for efficient retrieval. In *ECCV*, 2010.

[24] C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. *PAMI*, 19(5):530–534, 1997.

[25] S. Se, D. Lowe, and J. Little. Vision-based mobile robot localization and mapping using scale-invariant features. In *ICRA*, pages 2051–2058, 2001.

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

[27] P. Turcot and D. G. Lowe. Better matching with fewer features: The selection of useful features in large database recognition problems. In *ICCV Workshop WS-LAVD*, October 2009.

[28] S. Winder and M. Brown. Learning local image descriptors. In *CVPR*, June 2007.

Centre de recherche INRIA Rennes – Bretagne Atlantique
IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)

Centre de recherche INRIA Bordeaux – Sud Ouest : Domaine Universitaire - 351, cours de la Libération - 33405 Talence Cedex
Centre de recherche INRIA Grenoble – Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier
Centre de recherche INRIA Lille – Nord Europe : Parc Scientifique de la Haute Borne - 40, avenue Halley - 59650 Villeneuve d'Ascq
Centre de recherche INRIA Nancy – Grand Est : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex
Centre de recherche INRIA Paris – Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex
Centre de recherche INRIA Saclay – Île-de-France : Parc Orsay Université - ZAC des Vignes : 4, rue Jacques Monod - 91893 Orsay Cedex
Centre de recherche INRIA Sophia Antipolis – Méditerranée : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399