Local visual query expansion: Exploiting an image collection to refine local descriptors

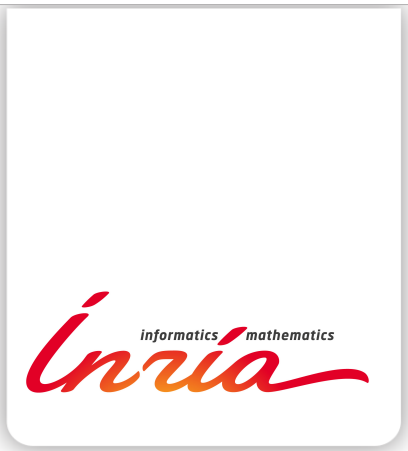
Giorgos Tolias, Hervé Jégou

Local visual query expansion: Exploiting an image collection to refine local descriptors

Giorgos Tolias, Hervé Jégou

RESEARCH REPORT
N° 8325
July 2013
Project-Team Texmex

ISSN 0249-6399 ISRN INRIA/RR--8325--FR+ENG



Local visual query expansion: Exploiting an image collection to refine local descriptors

Giorgos Tolias*, Hervé Jégou†

Project-Team Texmex

Research Report n° 8325 — July 2013 — 21 pages

Abstract: This paper proposes a query expansion technique for image search that is faster and more precise than the existing ones. An enriched representation of the query is obtained by exploiting the binary representation offered by the Hamming Embedding image matching approach: The initial local descriptors are refined by aggregating those of the database, while new descriptors are produced from the images that are deemed relevant.
This approach has two computational advantages over other query expansion techniques. First, the size of the enriched representation is comparable to that of the initial query. Second, the technique is effective even without using any geometry, in which case searching a database comprising 105k images typically takes 79 ms on a desktop machine. Overall, our technique significantly outperforms the visual query expansion state of the art on popular benchmarks. It is also the first query expansion technique shown effective on the UKB benchmark, which has few relevant images per query.

Key-words: image retrieval, query expansion, Hamming embedding

* National Technical University of Athens. This work was done in the context of a doctoral internship of Giorgos Tolias at INRIA Rennes, and supported by the ANR project FIRE-ID.
† INRIA Rennes

RESEARCH CENTRE
RENNES – BRETAGNE ATLANTIQUE

Campus universitaire de Beaulieu
35042 Rennes Cedex

Expansion de requête avec ou sans géométrie: raffinement des descripteurs locaux avec aggrégation de descripteurs

Résumé : Cet article propose une technique d'expansion de requête visuelle qui est plus rapide et précise que les approches existantes. Une représentation plus riche est obtenue en exploitant la représentation binaire offerte par la technique "Hamming Embedding" sous-jacente. Les descripteurs initiaux sont débruités en aggrégant ceux de la collection d'images, tandis que de nouveaux descripteurs sont synthétisés des images supposées pertinentes.
L'approche combine deux avantages in´dits en termes de coût de calcul. Tout d'abord, la requête enrichie a un coût comparable à celle de la requête initiale. Ensuite, la méthode est effective même lorsqu'aucune information géométrique n'est utilisée. Dans ce cas une recherche dans une base comprenant 105k images prend typiquement 79 ms sur une machine de bureau. L'approche dépasse plusieurs techniques de l'état de l'art en expansion de requête visuelle. Contrairement aux techniques concurrentes, elle présente un intérêt même lorsque il y a peu d'images correspondantes dans la collection pour une requête donnée.

Mots-clés : recherche d'image, expansion de requête, Hamming Embedding

1 Introduction

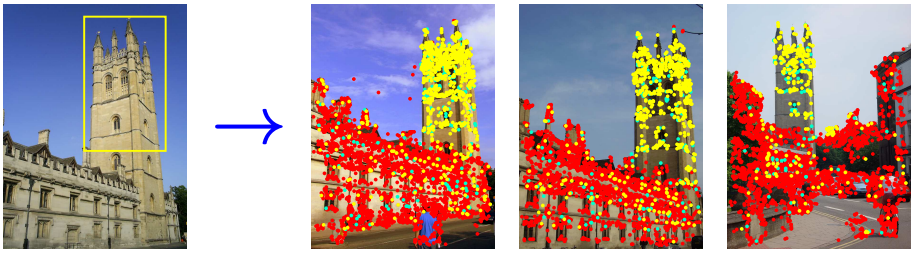


Figure 1: Query image (left) and the features selected (yellow+cyan) from the retrieved images to refine the original query. The features in red are discarded. Cyan features correspond to visual words that appear in the query image, and yellow ones to visual words that were not present in it. The selection of the depicted images and features has not involved any geometrical information.

This paper considers the problem of image and object retrieval in image databases comprising up to millions of images. The goal is to retrieve the images describing the same visual object(s) or scene as the query. In many applications, the query image is submitted by a user and must be processed in interactive time.

Most of the state-of-the-art approaches derive from the seminal Video-Google technique [1]. It describes an image by a bag-of-visual-words (BOVW) representation, in the spirit of the bag-of-words frequency histograms used in text information retrieval. This approach benefits from both the powerful local descriptors [2, 3] such as the SIFT, and from indexing techniques inherited from text information retrieval such as inverted files [4, 5]. Because it exploits the sparsity of the representation, BOVW is especially effective for large visual vocabularies [6, 7].

This analogy with text representation is a long-lasting source of inspiration in visual matching systems, and many image search techniques based on BOVW have their counterparts in text retrieval. For instance, some statistical phenomenons such as burstiness or co-occurrences appear both in texts [8, 9] and images [10, 11, 12] and are addressed in similar ways.

One of the most successful techniques in information retrieval is the query expansion (QE) principle [13], which is a kind of automatic relevance feedback. The general idea is to exploit the reliable results returned by an initial query to produce an enriched representation, which is re-submitted in turn to the search engine. If the initial set of results is large and accurate enough, the new query retrieves some additional relevant elements that were not present in the first set of results, which dramatically increases the recall.

Query expansion has been introduced to the visual domain by Chum et al. [14], who proposed a technique implementing the QE principle and specifically adapted to visual search. It proceeds as follows. First, the initial set of results is processed using a spatial verification method that filters out the images that are not geometrically consistent with the query. Second, the authors investigated several methods to build a new query from the images deemed relevant. The average query expansion (AQE) is of particular interest and usually consid-

ered as a baseline, as it is the most efficient variant [14] and provides excellent results. It is conceptually simple: A new *term-frequency inverse document frequency* (TFIDF) vector is obtained as the average of the results assumed correct and spatially back-projected to the original image.

Several extensions have been proposed to improve this initial QE scheme [15, 16, 17], for instance the use of a discriminative linear classifier [17] to define the new query instead of the average in AQE. Although these variants have improved the accuracy, they suffer from two inherent drawbacks which severely affect the overall complexity and quality of the search:

- First, they all require a costly geometrical verification step, which provides the automatic annotation of the relevant set and is typically performed on hundreds of images.
- Second, the augmented query representation contains significantly more non-zero components than the original one, which severely slows down the search. It is reported [17] that typically ten times more components are non-zeros. Since querying the inverted file has linear complexity in the number of features contained in the query vector, the second query is therefore one order of magnitude slower than the first.

Other kinds of expansion have been proposed for fixed image collections [17, 18]. They rely on the off-line pairwise matching of all images pairs and aim at identifying the features coming from the same object using spatial verification, which is rather costly as the complexity is quadratic in the number of images. They also assume that the image collection is fixed: The selection depends on a given set of images. Another related method [19] constructs a graph that links related images, and uses k-reciprocal nearest neighbors at query time to define a new similarity function that re-orders the images. Again, the cost of constructing and storing the graph in memory may be impracticable for large datasets.

In another line of research, several techniques address the loss in quantization underpinning BOVW, such as the use of multiple assignment [20] or soft quantization [21]. In a complementary manner, the Hamming Embedding (HE) technique [22] dramatically improves the matching quality by refining the descriptors with binary signatures. HE is not compatible with existing QE techniques because these assume a vector representation of the images. A noticeable exception is the transitive QE, which does not explicitly exploit the underlying image representation. However, this variant is not satisfactory with respect to query time and performance.

This paper, for the first time, proposes a novel way to exploit the QE principle in a system that individually matches the local descriptors, namely the HE technique. The new query expansion technique is both efficient and precise, thanks to the following two contributions:

- First, we modify the selection rule for the set of relevant images so that it does not involve any spatial verification. The images deemed relevant provide additional descriptors that are employed to improve the original query representation. Unlike other QE methods, it is done *on a per-descriptor basis* and not on the global BOVW vector. Figure 1 depicts an example of images and features that are selected by our method to refine the original query.

- The second key property of our method is that the set of local features is aggregated to produce new binary vectors defining the new query image representation. This step drastically reduces the number of individual features to be considered when submitting the enriched query.

To our knowledge, it is the first time that a visual QE is successful without any geometrical information: The only visual QE technique [14] that we are aware of performs poorly compared with other variants such as AQE. In contrast, our technique used without geometry reaches or outperforms the state of the art. Interestingly, it is shown effective even when a query has few corresponding images in the database, as shown by our results on the UKB image recognition benchmark [6]. Finally, incorporating geometrical information in the pipeline further improves the accuracy. As a result, we report a large improvement compared to the state of the art.

The paper is organized as follows. Section 2 introduces our core image system and Section 3 a post-processing technique for SIFT descriptors that is shown useful to improve the efficiency of the search. Section 4 introduces our Hamming Query Expansion (HQE) method and Section 5 describes our key aggregation strategy of local features. Section 6 describes how to exploit geometrical information with HQE. The experimental results presented in Section 7 demonstrate the superiority of our approach over concurrent visual QE approaches, with respect to both complexity and search quality, on the Oxford5k, Oxford105k and Paris benchmarks.

2 The core image system

This section describes the image search system based on Hamming Embedding upon which our query expansion techniques are built. This baseline method follows the guideline of the existing HE technique [22], which proceeds as follows. An image is represented by a set $\mathcal{P}$ of local SIFT descriptors [3] extracted with the Hessian-Affine detector [23].

BOVW and Hamming Embedding. The descriptors are quantized using a flat k-means quantizer, where k determines the visual vocabulary size. A descriptor $p \in \mathcal{P}$ is then represented by a quantization index, called a visual word $v(p)$. Computing and normalizing the histogram of visual words produces the BOVW representation. It can also be seen as a voting system in which all descriptors assigned to a specific visual word are considered as matching with a weight related to the inverse document frequency [1, 22].

In order to refine the quality of the matches and to provide more reliable weights to the votes, the HE technique [22] further refines each descriptor p by a binary signature $b(p)$, providing a better localization of the descriptor by subdividing the quantization cell $v(p)$. HE compares two local descriptors q and p that are assigned to the same visual word $v(p) = v(q)$ by computing the Hamming distance $h(q, p) = \|b(q) - b(p)\|_1$ between their binary signatures. If the Hamming distance is above a predefined threshold h_t, the descriptors are considered as non matching and zero similarity is attached. A significant benefit [22, 10] in accuracy is obtained by weighting the vote as a decreasing function of the Hamming distance. In this paper, we adopt the Gaussian function used in [10] with σ equal to one fourth of the binary signature size.

$\sqrt{}$	$-\mu$	mAP/BOVW	mAP/HE	IF
		47.7 ±*0.8*	67.1 ±*0.6*	1.200 ±*0.003*
x		47.7 ±*0.5*	69.5 ±*0.8*	1.290 ±*0.003*
x	x	48.1 ±*0.7*	69.6 ±*0.8*	1.238 ±*0.003*

Table 1: Evaluation with respect to mAP (performance) and IF (efficiency) of several post-processing procedures for SIFT: RootSIFT [17, 25] (denoted by $\sqrt{}$) and shifting (denoted by -μ). We have performed 10 measurements on Oxford5k with distinct vocabularies (k=16k) to report standard deviations.

The burstiness phenomenon in images was first revealed and tackled by Jegou *et al.* [10]. It takes into account descriptors that individually trigger multiple matches between specific pairs of images, which is often the case because of repetitive structures, or features which are abnormally common across all database images. Several normalizations have been proposed, from which we adopt the one that down-weights a given match score by the square root of the number of matches associated with the corresponding query descriptor [10]. This strategy is similar to the successful component-wise power-law normalization later proposed for BOVW or Fisher Kernels [24], but here applied to a voting technique.

Multiple assignment (MA). BOVW and HE handles descriptors assigned to the same visual word. However quantization losses are introduced when truly matching descriptors are assigned to different visual words. This has been addressed by assigning multiple visual words to each descriptor [20, 21]. We apply MA on the query side only in order to keep memory requirements unchanged [22]. In the rest of the paper, the initial method that assigns a descriptor a single visual word is denoted by SA (single assignment) to distinguish it with MA.

3 Root-SIFT and Shift-SIFT

It was recently shown [17, 25] that square rooting the components of SIFT descriptors improves the search performance. This is done either by L_1-normalizing the SIFT descriptor [17] prior to the square-rooting operation or, equivalently, by [25] square-rooting the components and normalizing the resulting vector in turn with respect to L_2. This operation amounts to computing the Hellinger distance instead of the Euclidean one. The impact of this scheme is evaluated in Table 1 on the Oxford5k building benchmark [7] for both BOVW and HE, without the burstiness processing. Following the standard evaluation protocol, we measure the mean average precision (mAP). In order to cope with the variability of the results due to the sensitivity of k-means to the initial random seeds, we average the results over 10 runs with different vocabularies and report the standard deviation. The improvement provided by square-rooting the components is statistically significant when used with HE.

However, as a side-effect of this processing, we observe that Root-SIFT introduces an unexpected complexity overhead, resulting from less balanced inverted

lists. The undesirable impact of uneven inverted lists was first noticed by Nister et al. [6] and is commonly measured by the imbalance factor (IF) [22], which is a multiplicative factor reflecting the deviation from perfectly balanced lists. For instance, IF=2 means that, on average, two times more individual descriptor comparisons are performed compared to the case where the lists have equal lengths. Table 1 shows that this negative effect, which was not reported for this RootSIFT variant [17, 25], is statistically significant.

Shift-SIFT. In order to reduce this undesirable effect, we introduce another processing method for SIFT descriptors referred to as shift-SIFT. It is inspired by the approach proposed for BOVW vectors [26], which aims at handling "negative evidences" by centering the descriptors and L_2 normalizing them in turn. It gives more importance in the comparison to the components which are close to 0, and improves the performance in the case of BOVW vectors.

Table 1 shows the interest of this shifting strategy applied to SIFT descriptors. We have use SA and no burstiness normalization in this experiment (conclusions are similar in other setups). The gain in accuracy is significant only with BOVW. Yet, this approach provides more balanced lists and therefore reduces the search complexity by about 4% at no cost, as reflected by the IF measure. Our interpretation is that, with this processing, the feature distribution is more uniform in the feature space, which is now the full L_2 ball and not only its positive quadrant. We use this shifting strategy combined with L_2 Root-SIFT [25] in all our experiments.

4 HE with query expansion

This section defines a query expansion technique based on HE and not involving any geometrical information. We revisit the different stages involved in the QE principle. We first describe how reliable images are selected from the initial result set. Then we detail the way an enriched query is produced from the images deemed relevant. The key subsequent aggregation step and the use of geometry will be introduced later in Sections 5 and 6, respectively.

4.1 Selection of reliable images

As in all query expansion methods [14, 15, 16, 17], the core image search system processes an initial query. The resulting set is analyzed to identify a subset of reliable images that are likely to depict the query object, and therefore to provide additional features that will be subsequently exploited in the augmentation stage.

In the following, we will denote the local features of the query image by $\mathcal{Q}$, and those of a given database image by $\mathcal{P}$, respectively. As a criterion to determine the relevant images, we count the number $C(\mathcal{Q}, \mathcal{P})$ of "strict" feature correspondences between the query and images in the short-list. It is given by

$$C(\mathcal{Q}, \mathcal{P}) = |\{(q, p) \in \mathcal{Q} \times \mathcal{P} : h(q, p) \leq h_{\mathrm{t}}^{\star}\}| \,, \tag{1}$$

where the threshold $h_{\mathrm{t}}^{\star}$ is lower than the Hamming embedding threshold h_{t} used for initial ranking. Such a lower threshold allows for a higher true positive to

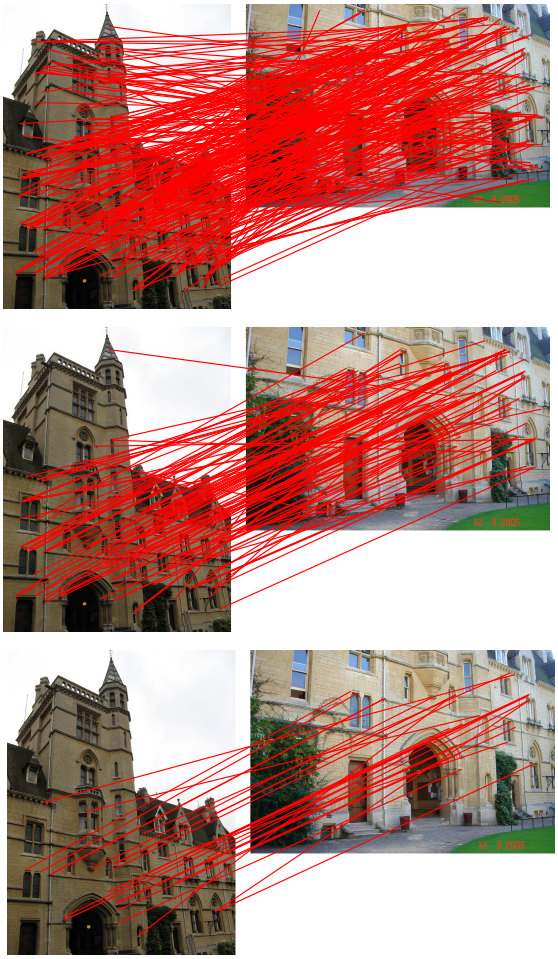


Figure 2: Matching features using BOVW (top), HE with $h_t = 24$ (middle) and HE with $h_t = 16$ (bottom).

false positive ratio of matches [22]. It provides a strict way to count correspondences in a manner that resembles the number of RANSAC inliers commonly used to verify the images [7]. It is less precise than RANSAC, yet it has the advantage of not using any geometry. It is therefore much faster.

Figure 2 illustrates, for a pair of images, the matching features obtained using BOVW and HE. We consider two different thresholds for HE to show the impact of the strict threshold $h_t^\star = 16$ on selected features. Observe that HE matching filters out many false matches compared to BOVW. With a lower threshold value, the filtering is not far in quality from that of a spatial matching method.

An image is declared reliable if at least c_t correspondences are satisfied,

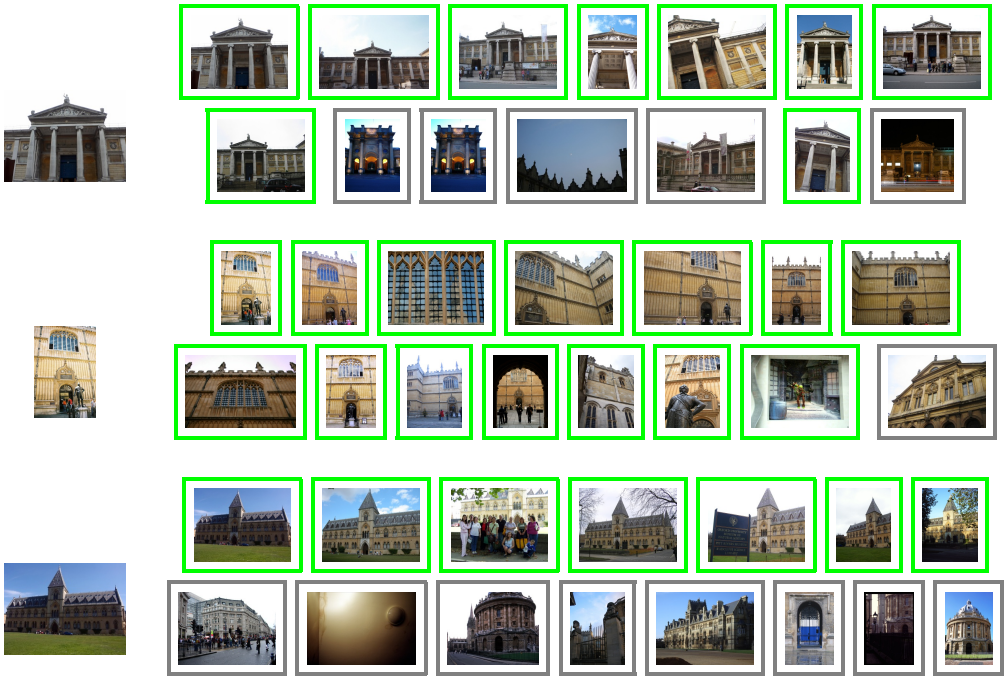


Figure 3: Examples of query images (left) and the corresponding top ranked lists by the baseline retrieval system. Images (not) selected as reliable are marked with (gray) green border.

which formally leads to define the set of *reliable images* as

$$\mathcal{L}_\mathcal{Q} = \{\mathcal{P} : C(\mathcal{Q}, \mathcal{P}) \geq c_\mathrm{t}\}. \tag{2}$$

In practice, only the images short-listed in the initial search are considered as candidates for the set of reliable images. In our experiments, we count the number of correspondences with Equation 1 only for the top 100 images. Figure 3 shows examples of queries and the corresponding reliable images. Although some negative images are selected and some positive ones are not, the result is not far from what spatial verification would produce. This suggests that selecting reliable images with HE and a low threshold is sufficient for the purpose of QE, as proposed in this section.

4.2 Feature set expansion

First, let us recall that a feature descriptor is described by both a visual word and a binary signature. Our augmentation strategy, *i.e.*, how we introduce new local features in the representation, is partly based on the selection of visual words that are not present in the original query.

Since a large proportion of the reliable images depicts the same object, the visual words frequently occurring in the images of the reliable set $\mathcal{L}_\mathcal{Q}$ are likely to depict the query object rather than the background. Our selection strategy is simple and consists in selecting the most frequent visual words occurring in $\mathcal{L}_\mathcal{Q}$. More precisely, we sort the visual words contained in the images of $\mathcal{L}_\mathcal{Q}$

Figure 4: Sample reliable images and features assigned to reliable visual words, when geometry is not used. *Left:* Query image. *Top:* Features assigned to reliable visual words that appear in the query image. *Bottom:* Features in the set of augmented visual words. Note: we only show a subsample of the actual reliable visual words. Each color represents a distinct visual word.

by the number of reliable images in which they appear. The top ranked words are selected and define the set of *reliable visual words* $\mathcal{V}$, which may include both visual words that are present or absent in the query image. The latter are referred to as the *augmented visual words*. Their count is controlled by a parameter α to ensure that the number of reliable visual words in the new query is proportional to that of the original query, as

$$|\mathcal{V} \setminus \mathcal{V}_\mathcal{Q}| = \alpha \cdot |\mathcal{V}_\mathcal{Q}|, \tag{3}$$

where $\mathcal{V}_\mathcal{Q}$ is the set of visual words occurring in the query. The parameter α typically takes values in the range $(0, 2]$.

The initial query set is enriched with the features of the reliable images assigned to the reliable visual words. Let define as

$$\mathcal{G} = \{p \in \mathcal{P} : \mathcal{P} \in \mathcal{L}_\mathcal{Q} \wedge v(p) \in \mathcal{V}\} \tag{4}$$

the union of all features of reliable images assigned to some reliable words. It defines the set of database features used to augment the initial query. In other terms, this set is merged with the initial query feature set to construct the augmented query as

$$\mathcal{Q}_E = \mathcal{Q} \cup \mathcal{G}. \tag{5}$$

Figure 4 depicts some features from reliable images assigned to reliable visual words. Observe that, even without any spatial information, selected visual words are detected on the foreground object. Moreover, each visual word corresponds not only to similar image patches, but often to the exact same patch of the object, as if spatial matching was used. This appears to be the case for either visual words which appear (top) or miss (bottom) in the query.

A simple way to construct an enriched query is to use the expanded set of features as the new image representation. However, similar to existing QE

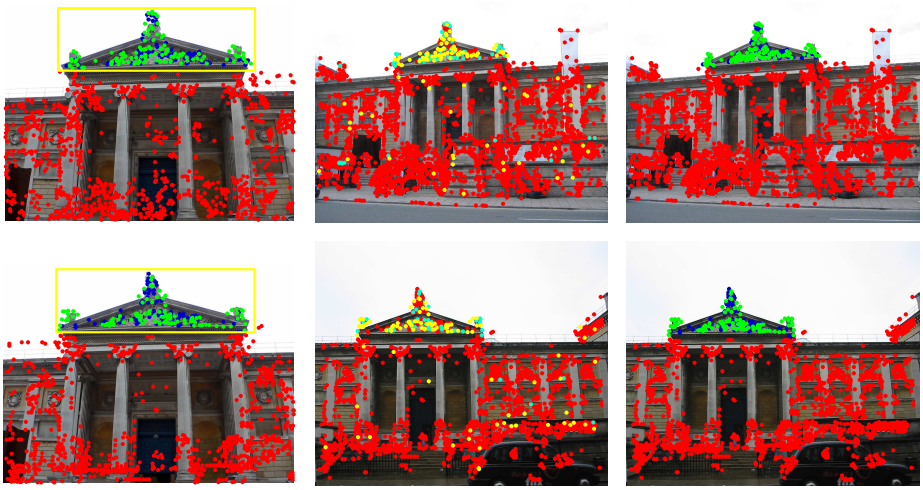


Figure 5: Features selected for the expanded set of a particular query image (left) without (middle) and with spatial matching (right). With spatial matching: features back-projected out of the bounding box are rejected (red), while the rest (blue and green) are kept. Those assigned to reliable visual words are shown in green. Without spatial matching: features assigned to reliable visual words are shown in cyan or yellow, with yellow being the ones assigned specifically to augmentation visual words. Rejected are shown in red. *Best viewed in color.*

strategies, such an approach leads to a high complexity because the number of features explodes. We observe that it is typically multiplied by a factor ranging from 10 to 20 for typical values of α, as analyzed in the experimental section 7. This drawback is shared by other effective techniques on query expansion [17], for which this problem leads to produce a BOVW vector having 10 times more non-zero components than the initial one. In the next section, we address this issue by proposing an aggregation strategy that drastically reduces the number of features.

5 QE with feature aggregation

The average query expansion technique [14] averages BOVW vectors to produce the new query. In this section, we explain how local descriptors are *individually* refined or created from binary signatures of the set of reliable features. At this stage, the augmented set contains multiple instances representing the same visual patches, either in the initial query or not. Descriptors associated to the same patch are expected to have similar binary signatures. The strategy presented below implicitly exploits this underlying property to produce the new set of query descriptors which is less redundant.

First, note that the selection strategies for images and features presented in the previous subsections introduce a few false positives in the augmented feature set. This is the cost to pay for not performing the selection with a stringent spatial matching technique: Our inliers are not selected as reliably as in

other query expansion methods. The aggregation operation proposed hereafter comes as a complement on our selection method, as it is robust enough to false positives. In contrast, averaging over normalized TFIDF vectors of similar images [14], as done in AQE, is sensitive to background and noisy features.

Our aggregation scheme is inspired by methods [27, 28] such as the VLAD technique, which aggregates the descriptors per visual word to produce a vector representation of an image. In our method, we aggregate the features of $\mathcal{Q}_E$ that are assigned to the same visual word. Therefore, our technique produces exactly one binary signature per visual word occurring in $\mathcal{Q}_E$. Our motivation is that the true matching patches are likely to overrule the false positives. This actually happens in practice because the true correspondences are more numerous and are associated with more consistent binary signatures.

For each visual word v appearing in $\mathcal{Q}_E$, a new binary signature $b(v)$ is obtained by computing the median values over all the bit vectors occurring in $\mathcal{Q}_E$ and assigned to v. If the numbers of 0 and 1 are equal for a particular bit component, the tie is arbitrarily resolved by assigning either 0 or 1 with equal probabilities. This new set of descriptors comprises exactly one binary signature per visual word and serves as the new query, which is then submitted to the system.

We will refer to this method as Hamming Query Expansion (HQE) in the remainder of this paper.

6 Geometrical information

This section proposes a variant of our method to further eliminate some incorrect matches by including some spatial information in the loop. For this reason and as shown later in the experimental section, it is not as fast as the HQE strategy proposed in Sections 4 and 5. However, this approach further improves the performance and is therefore interesting in some situations where one would trade an interactive time against any improvement in accuracy.

It proceeds as follows. The matches are collected with the regular HE technique, *i.e.*, they are returned by the first query. Instead of calculating the number of correspondences with Equation 1, we rely on the number $C_{\mathrm{g}}(\mathcal{Q}, \mathcal{P})$ of inliers found with a spatial matching technique. For this purpose, we have used the spatial verification procedure proposed by Philbin *et al.* [7]. Similar to other QE techniques, this procedure is applied on the top ranked images only. An image is declared reliable if the number of inliers is above a pre-defined threshold. The estimation of the affine transformation is then further exploited to filter the expanded feature set. As first suggested by Chum *et al.* [14], the matching features associated with the reliable images are projected back to the query image plane. Those falling out of the query image borders are filtered out.

The remaining steps of this variant then become similar to the HQE method of Sections 4 and 5. The only difference is that the input set of reliable features is different. Therefore, we first select the reliable visual words and perform the feature set expansion. The aggregation is similarly applied to produce one binary vector per visual word. Note that, the reliable images, as detected by spatial matching, are ranked in top positions.

Figure 5 depicts the descriptors selected for the HQE expanded set with

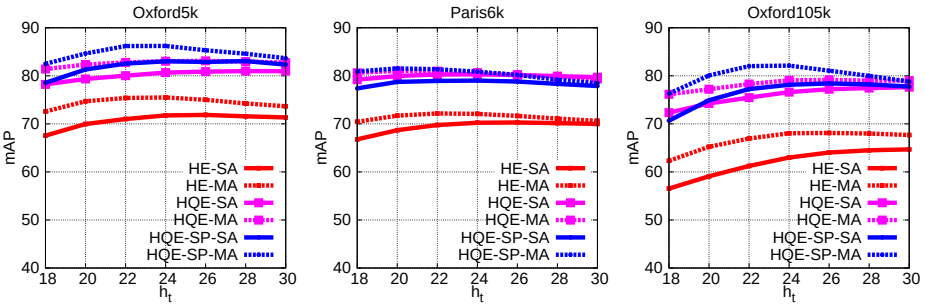


Figure 6: Impact of h_t on the performance of HE and HQE.

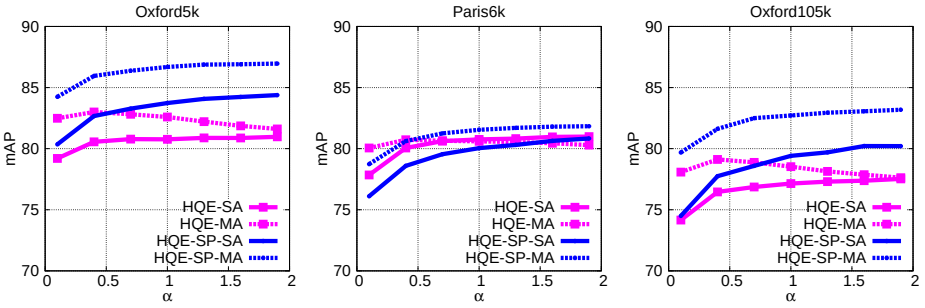


Figure 7: Impact of α. Performance of HQE when varying the number of new visual words in the expanded query.

and without geometry. Notice that even without geometry, most of the selected features are localized on the target object. The geometry effectively filters out the remaining features that do not lie on the query object.

7 Experiments

This experimental section first introduces the datasets, gives details about the experimental setup, and evaluates the impact of the parameters and variants. Our technique is then compared with the state of the art on visual QE before a discussion on the complexity.

7.1 Datasets and experimental setup

Datasets and measures. Query expansion techniques are only effective if the dataset consists of several relevant images for a given query. We evaluate the proposed method on two publicly available datasets of this kind, namely Oxford5k Buildings [7] and Paris [21], but also on a dataset where queries have only few corresponding images, that is UKB [6]. Following the standard evaluation protocols, we report mean Average Precision (mAP) for the two first and use the score definition associated with UKB: the average number of correct images ranked in first 4 positions (from 0 to 4). As for other QE works, the large scale experiments are carried out on the Oxford105k dataset, which augments Oxford5k with 100k additional distractor images.

Features and experimental setup. For Oxford5k and Paris, we used the modified Hessian-Affine detector proposed by Perdoch *et al.* [29] to detect local features. The extracted SIFT descriptors have been subsequently post-processed by using the L_2 Root-SIFT and shift-SIFT procedure, as described in Section 2. For UKB, we have used the same features provided by the authors of the papers [20, 19]. We follow the more realistic, less biased approach, of learning the vocabulary on an independent dataset. That is, when we use Oxford5k for evaluation, the vocabulary is learned with features of Paris and *vice versa*. Similarly, learning the medians of Hamming Embedding is carried out on the independent dataset.

Unless otherwise stated, we use a visual vocabulary comprising $k = 65,536$ visual words, binary signatures of 64 dimensions, and apply HE with weights and burstiness normalization. In all our experiments, the reliable images for our approach, either without or with spatial matching, are selected among top 100 ones returned by the baseline system. When using MA, it is applied on the query side using the 3 nearest visual words to limit the computational overhead of using more.

MA produces more correspondences than single assignment (SA), therefore the probability of finding a false positive match is increased even with spatial matching and the matching parameters should be stricter [22]. We set the minimum number of correspondences to $c_t = 4$ with SA and to $c_t = 5$ with MA.

Two factors introduce some randomness in the measure with our approach: The random projection matrix (in HE) and the random decision used to resolve ties when aggregating binary signatures. Therfore, each experiment is performed 5 times using distinct and independent parameters. We report the average performance and standard deviation to assess the significance of our measurements.

7.2 Impact of the parameters

Thresholds. The strict threshold $h_t^\star$ is constant and set to 16 in all our experiments. Figure 6 shows the impact of the parameter h_t. The performance is not very sensitive around the optimal values attained at $h_t = 22$ or $h_t = 24$, depending on the setup. Note already that HQE gives a significant improvement compared to the HE baseline. In the rest of our experiments, we set $h_t = 24$ in all cases, similar to most works based on HE. We have fixed $\alpha = 0.5$ for this preliminary experiment, which implies that the size of the new query is at most 1.5 times larger than the initial one. In practice, it is much smaller thanks to the descriptors aggregation. See, for instance, Table 3 to compare the average number of descriptors used in the original and augmented queries.

The parameter α (see Section 4) controls the size of the augmented query. Figure 7 presents its impact on the performance. HQE without spatial matching rapidly attains its maximum in performance and then decreases. This suggests that not too many visual words should be selected because the additional ones will introduce many outliers compared to inliers. In contrast, spatial matching filters out most of the outliers: Using more descriptors is better because the added ones are mostly inliers. As a compromise between performance and complexity, we set $\alpha = 0.5$ and 1.0 without and with spatial matching, respectively.

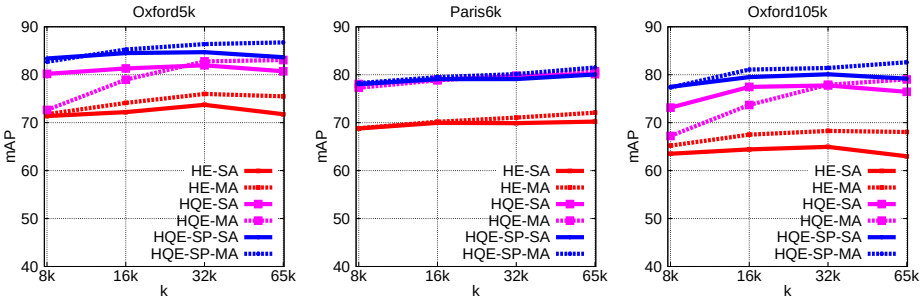


Figure 8: Impact of the vocabulary size on the performance of HE, HQE and HQE with spatial matching.

W	burst	MA	HQE	SP	Oxford5k	Paris6k	Oxford105k
					66.9	65.7	55.5
×					70.4	68.4	59.6
×	×				71.7	70.2	62.9
×	×	×			75.4	72.0	68.0
×	×	×	×		83.0	80.6	79.0
×	×	×	×	×	86.8	81.5	82.6
BOVW					53.3	54.8	44.2

Table 2: Mean average precision for separate components comprising the proposed method. Initial method is the original Hamming Embedding without weights. W denotes the use of weights.

The vocabulary size k is critical in most methods based on pure BOVW. Figure 8 that it is not the case with HE and our techniques, which achieves excellent performance for all the sizes. This confirms prior observations reported in the literature [22, 30].

Weights, burstiness and HQE. Table 2 summarizes the respective contributions of the different elements of our search engine. First note the large gain already achieved by weighting the Hamming distances in HE, using MA and applying the burstiness procedure [10]. Note also the even larger improvement obtained by using our HQE technique, either with or without spatial matching.

Aggregation. Table 3 reveals the double benefit of the local feature aggregation method proposed in Section 5 with respect to performance and query efficiency. Merging the binary signatures reduces the expanded query size and has a positive impact on complexity, as quantitatively measured in Table 3: The aggregation step reduces by about one order of magnitude the size of the enriched query, which becomes comparable to that of the initial query.

In addition, Table 3 also shows that this step significantly and consistently improves the performance. To demonstrate this, we have compared HQE (*i.e.*, with aggregation) to a method which issues the expanded query defined by Equation 5, *i.e.*, prior to aggregation. As already discussed in Section 5, our interpretation is that aggregating binary signatures filters out noisy features and removes the redundant features at the same time.

Dataset	Method	mAP		$\|Q\|$	
		SA	MA	SA	MA
Oxford5k	HE	71.7	75.4	1,362	4,089
	HQE/b.a.	79.0	82.0	11,937	27,345
	HQE	80.7	**83.0**	1,810	**5,030**
Paris6k	HE	70.2	72.0	1,460	4,382
	HQE/b.a.	76.6	77.3	35,982	66,665
	HQE	80.2	**80.6**	1,843	**5,045**
Oxford105k	HE	62.9	68.0	1,362	4,088
	HQE/b.a.	73.5	76.5	12,176	28,699
	HQE	75.6	**79.0**	1,810	**5,030**

Table 3: Performance and average query size $|Q|$ for the baseline HE, HQE and the use of the same expanded query before aggregation (HQE/b.a.). Note that the aggregation procedure is a key step: not only it significantly reduces the complexity (number of features), but it also improves the performance.

Detailed performance on Oxford5k. Table 4 presents some detailed performances and statistics we have collected on Oxford5k for HE and HQE. Our selection strategy for reliable images, even without spatial matching, does not suffer from the variability of the number of true similar images in the database, with an exception on *Cornmarket*, where HQE without spatial matching selects a few false positives as reliable images. Also observe that HQE notably outperforms HQE-SP for the *Bodleian* queries. It is because HQE-SP is stricter and does not select enough reliable images. This suggests that a weaker spatial matching model [31] could offer a good compromise to select these images.

More features. All our experiments are conducted with features extracted using the default threshold for the Hessian-Affine detector [29] to allow for a direct comparison with the literature. Using a lower threshold for the "cornerness" value produces a larger set of features. It might be useful for image matching but might also add noisy features and therefore arbitrary matches.

Table 5 investigates the impact of cornerness on both our methods and existing BOVW and HE baselines. With the default threshold, the software produces a total number of 12.53M features on Oxford5k. By using two smaller thresholds, we produced two other sets of features comprising 21.92M and 27.59M features, respectively. Table 5 shows that BOVW's performance increases with the medium-sized set, but its performance drops with the larger one. In contrast, HE benefits from having more features. The performance of the two larger sets is comparable, which suggests that HE better handle noisy matches in a better way and can use more features. As a consequence, HQE performs in a similar way. Interestingly, the performance increases up to **mAP=89.4** for HQE with geometry and MA, which is a large improvement over the state of the art.

7.3 Comparison with the state of the art

Oxford5k, Paris6k and Oxford105k. Table 6 compares the QE proposed

Building	GT	HE	HQE		HQE-SP	
		mAP	$\lvert\mathcal{L}_\mathcal{Q}\rvert$	mAP	$\lvert\mathcal{L}_\mathcal{Q}\rvert$	mAP
All Souls	183	78.2	47.0	94.6	44.8	97.3
Ashmolean	56	63.7	10.9	76.1	9.9	80.8
Balliol	30	72.7	15.8	81.0	8.0	82.1
Bodleian	54	66.4	33.4	94.5	19.8	86.9
Christ Church	211	74.9	39.5	75.7	45.1	90.7
Cornmarket	22	69.5	9.6	64.9	6.4	71.4
Hertford	55	87.7	41.5	95.0	43.5	98.3
Keble	18	93.0	9.5	96.5	7.6	99.5
Magdalen	157	29.9	15.6	36.5	8.8	48.6
Pitt Rivers	16	100.0	9.7	99.7	7.0	100.0
Radcliffe	569	93.9	97.1	98.5	96.0	98.7

Table 4: Oxford5k dataset: Summary of the number of ground truth images, the number of reliable images and the performance for HE, HQE and HQE with spatial matching. We report the average value of $|\mathcal{L}_\mathcal{Q}|$ per building, which is the number of reliable images in the short-list of 100 top-ranked ones.

# features		12.53M	21.92M	27.59M
method	MA	mAP		
BOVW		54.9	58.7	55.2
HE		74.2	78.6	78.3
HQE		81.0	84.8	84.4
HQE-SP		85.3	88.1	88.5
HQE-SP	×	88.0	**89.4**	89.3

Table 5: More features: Performance comparison on Oxford5k using lower detector threshold values, *i.e.*, larger sets of local features. Binary signatures of 128 bits are used.

method with previously published results on the same datasets. For a fair comparison, we have included the scores of other QE methods that have used the same local feature detector [29] as input and learned the vocabulary on an independent dataset. In this table, we also include the scores for our method when using 128-bit signatures for HE, which are better at the cost of higher memory usage and a slightly larger complexity.

Interestingly, even without spatial matching, our method outperforms all methods in Oxford105k and Paris6k dataset. HQE-SP outperforms them in all three datasets. All of the compared methods rely on spatial matching to verify similar images and expand the initial query. Moreover, the work of Mikulik *et al.* [32] requires a costly off-line phase and assigns the descriptors with a very large vocabulary of 16M, thereby impacting the overall efficiency.

To our knowledge, the performance of our method is the best reported to date on Oxford5k, Paris6k and Oxford105k, when learning the vocabulary on an independent dataset (89.1 was reported [17] by learning it on the Oxford5k comprising the relevant images). In addition, all these techniques are likely to

Method	SP	MA	Oxford5k	Paris6k	Oxford105k
Perdoch [29]	×		78.4	N/A	72.8
Perdoch [29]	×	×	82.2	N/A	77.2
Mikulik [32]	×	×	**84.9**	**82.4**	**79.5**
Chum [16]	×		82.7	80.5	76.7
Arandjelovic [17]	×		80.9	76.5	72.2
HQE			80.7±*0.9*	80.2±*0.2*	76.6±*1.1*
HQE-SP	×		83.7±*0.7*	80.0±*0.2*	79.4±*0.6*
HQE		×	83.0±*0.9*	80.6±*0.2*	79.0±*1.0*
HQE-SP	×	×	86.8±*0.3*	81.5±*0.3*	82.6±*0.4*
HQE 128bits			81.0±*0.5*	81.5±*0.2*	76.9±*0.6*
HQE-SP 128bits	×		85.3±*0.4*	81.3±*0.3*	80.8±*0.5*
HQE 128bits		×	83.8±*0.3*	82.8±*0.1*	80.4±*0.5*
HQE-SP 128bits	×	×	**88.0**±*0.3*	**82.8**±*0.2*	**84.0**±*0.2*

Table 6: Performance comparison with state-of-the-art methods on Oxford5k, Paris6k and Oxford105k. The standard deviation is obtained from 5 measurements.

Jégou [10]	Jégou [20]	Qin [19]	HE-MA	HQE-MA
3.64	**3.68**	**3.67**	3.59	**3.67**

Table 7: UKB: comparison with state-of-the-art methods.

be complementary, as they consider orthogonal aspects to improve the performance.

UKB [6] is a dataset with few corresponding images per query (4, including the query image). QE techniques are therefore not expected to perform well, and accordingly we are not aware of any competitive result reported with a QE method on this dataset. For this set only, we reduce the short-list of images selected in the short-list to reflect the expected result set. Table 7 shows that HQE improves the performance significantly compared to the HE baseline and is therefore effective even with few relevant images. It performs similar to other state-of-the-art techniques that perform well on this dataset. Note that these best techniques all require to cross-match (off-line) the whole image collection with itself, which may be infeasible on a large scale (quadratic complexity).

7.4 Complexity: timings and query size

First, note that the initial query includes a binary signature per local feature and several features can be assigned to the same visual word for a given image, especially with MA. In addition, the expanded query set, as defined before aggregation, is much larger as several images contribute to it with their reliable features, as previously shown in Table 3. Thanks to HQE, only one binary signature per visual word is kept. This favorably impacts the complexity of the enriched query in terms of the number of signatures. On average, the total number of features increases only by a small factor after aggregation, to be

Method	HE	HQE	HQE-SP
SA	30 ms	79 ms	731 ms
MA	76 ms	204 ms	955 ms

Table 8: Average query times for the HE baseline and our technique with and without spatial matching, measured on Oxford105k when using a single processor core. These timings do not include the description part (extracting and quantizing the SIFT descriptors), which does not depend on database size.

compared with queries which are one order of magnitude larger for other QE techniques.

Table 8 reports the average search times when querying Oxford105k. They have been measured on a single core desktop machine (3.2 Ghz). The spatial matching has been estimated by an external software and is included in the query time, unlike the SIFT extraction and quantization times. As expected, the search times are competitive for HQE without geometry, even when MA is used. As a reference, best time reported for QE with spatial matching [29] is 509ms on Oxford105k on a 4×3.0Ghz machine.

8 Conclusion

This paper makes several contributions related to visual query expansion. First, we introduce a QE method which is effective without using any geometry. While the general belief is that spatial verification is required to select the relevant images used to build the augmented query, exploiting the Hamming Embedding technique with a stringent selection rule and an aggregation strategy, we already achieve state-of-the-art performance. This method has a low complexity. We then show that combining our Hamming query expansion with geometry further improves the results and significantly outperform the state of the art.

In future work, we will investigate how to incorporate weak spatial matching models [31, 22] in our query expansion method, in order to find a compromise between a costly spatial verification or not using geometry at all.

Acknowledgments

This project was done in the context of the Project Fire-ID, supported by the Agence Nationale de la Recherche (ANR-12-CORD-0016).

References

[1] Sivic, J., Zisserman, A.: Video Google: A text retrieval approach to object matching in videos. In: International Conference on Computer Vision. (2003)

[2] Schmid, C., Mohr, R.: Local grayvalue invariants for image retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence **19** (1997) 530–534

[3] Lowe, D.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60** (2004) 91–110

[4] Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information Processing & Management **24** (1988) 513–523

[5] Zobel, J., Moffat, A., Ramamohanarao, K.: Inverted files versus signature files for text indexing. ACM Trans. Database Systems **23** (1998) 453–490

[6] Nistér, D., Stewénius, H.: Scalable recognition with a vocabulary tree. In: IEEE Conference on Computer Vision and Pattern Recognition. (2006) 2161–2168

[7] Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: IEEE Conference on Computer Vision and Pattern Recognition. (2007)

[8] Katz, S.M.: Distribution of content words and phrases in text and language modeling. Natural Language Engineering **2** (1996) 15–59

[9] Lund, K., Burgess, C.: Producing high-dimensional semantic spaces from lexical co-occurrence. Behavior Research Methods, Instruments & Computers **28** (1996) 203–208

[10] Jégou, H., Douze, M., Schmid, C.: On the burstiness of visual elements. In: IEEE Conference on Computer Vision and Pattern Recognition. (2009)

[11] Chum, O., Matas, J.: Unsupervised discovery of co-occurrences in sparse high dimensional data. In: IEEE Conference on Computer Vision and Pattern Recognition. (2010)

[12] Cinbis, R.G., Verbeek, J., Schmid, C.: Image categorization using fisher kernels of non-iid image models. In: IEEE Conference on Computer Vision and Pattern Recognition. (2012)

[13] Manning, C.D., Raghavan, P., Schütze, H.: Relevance feedback & query expansion. In: Introduction to Information Retrieval. Cambridge University Press (2008)

[14] Chum, O., Philbin, J., Sivic, J., Isard, M., Zisserman, A.: Total recall: Automatic query expansion with a generative feature model for object retrieval. In: International Conference on Computer Vision. (2007)

[15] Joly, A., Buisson, O.: Logo retrieval with a contrario visual query expansion. In: ACM International conference on Multimedia. (2009)

[16] Chum, O., Mikulik, A., Perdoch, M., Matas, J.: Total recall II: Query expansion revisited. In: IEEE Conference on Computer Vision and Pattern Recognition. (2011)

[17] Arandjelovic, R., Zisserman, A.: Three things everyone should know to improve object retrieval. In: IEEE Conference on Computer Vision and Pattern Recognition. (2012)

[18] Turcot, P., Lowe, D.: Better matching with fewer features: the selection of useful features in large database recognition problems. In: International Conference on Computer Vision Workshop. (2009)

[19] Danfeng, Q., Gammeter, S., Bossard, L., Quack, T., Gool, L.V.: Hello neighbor: Accurate object retrieval with k-reciprocal nearest neighbors. In: IEEE Conference on Computer Vision and Pattern Recognition. (2011)

[20] Jégou, H., Schmid, C., Harzallah, H., Verbeek, J.: Accurate image search using the contextual dissimilarity measure. IEEE Transactions on Pattern Analysis and Machine Intelligence **32** (2010) 2–11

[21] Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Lost in quantization: Improving particular object retrieval in large scale image databases. In: IEEE Conference on Computer Vision and Pattern Recognition. (2008)

[22] Jégou, H., Douze, M., Schmid, C.: Improving bag-of-features for large scale image search. International Journal of Computer Vision **87** (2010) 316–336

[23] Mikolajczyk, K., schmid, C.: Scale and affine invariant interest point detectors. International Journal of Computer Vision **60** (2004) 63–86

[24] Perronnin, F., J.Sánchez, Mensink, T.: Improving the Fisher kernel for large-scale image classification. In: European Conference on Computer Vision. (2010)

[25] Jain, M., Benmokhtar, R., Gros, P., Jégou, H.: Hamming embedding similarity-based image classification. In: ICMR. (2012)

[26] Jégou, H., Chum, O.: Negative evidences and co-occurences in image retrieval: The benefit of PCA and whitening. In: European Conference on Computer Vision. (2012)

[27] Perronnin, F., Dance, C.R.: Fisher kernels on visual vocabularies for image categorization. In: IEEE Conference on Computer Vision and Pattern Recognition. (2007)

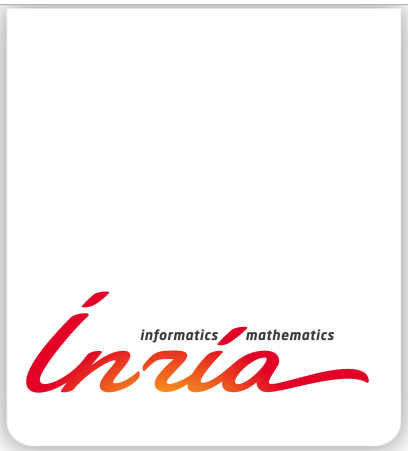
[28] Jégou, H., Perronnin, F., Douze, M., Sánchez, J., Pérez, P., Schmid, C.: Aggregating local descriptors into compact codes. In: IEEE Transactions on Pattern Analysis and Machine Intelligence. (2012)

[29] Perdoch, M., Chum, O., Matas, J.: Efficient representation of local geometry for large scale object retrieval. In: IEEE Conference on Computer Vision and Pattern Recognition. (2009)

[30] Sattler, T., Weyand, T., Leibe, B., Kobbelt, L.: Image retrieval for image-based localization revisited. In: IEEE Conference on Computer Vision and Pattern Recognition. (2012)

[31] Tolias, G., Avrithis, Y.: Speeded-up, relaxed spatial matching. In: International Conference on Computer Vision. (2011)

[32] Mikulík, A., Perdoch, M., Chum, O., Matas, J.: Learning a fine vocabulary. In: European Conference on Computer Vision. (2010)



RESEARCH CENTRE
RENNES – BRETAGNE ATLANTIQUE

Campus universitaire de Beaulieu
35042 Rennes Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399