# Query-adaptive asymmetrical dissimilarities for visual object retrieval

Cai-Zhi Zhu, Hervé Jégou, Shin'Ichi Satoh

# Query-adaptive asymmetrical dissimilarities for visual object retrieval

Cai-Zhi Zhu
NII, Tokyo
cai-zhizhu@nii.ac.jp

Hervé Jégou
INRIA, Rennes
herve.jegou@inria.fr

Shin'ichi Satoh
NII, Tokyo
satoh@nii.ac.jp

## Abstract

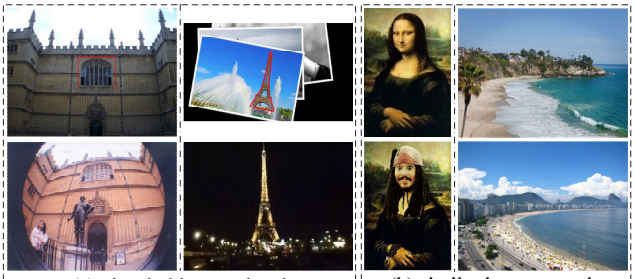
*Visual object retrieval aims at retrieving, from a collection of images, all those in which a given query object appears. It is inherently asymmetric: the query object is mostly included in the database image, while the converse is not necessarily true. However, existing approaches mostly compare the images with symmetrical measures, without considering the different roles of query and database.*

*This paper first measure the extent of asymmetry on large-scale public datasets reflecting this task. Considering the standard bag-of-words representation, we then propose new asymmetrical dissimilarities accounting for the different inlier ratios associated with query and database images. These asymmetrical measures depend on the query, yet they are compatible with an inverted file structure, without noticeably impacting search efficiency. Our experiments show the benefit of our approach, and show that the visual object retrieval task is better treated asymmetrically, in the spirit of state-of-the-art text retrieval.*

## 1. Introduction

The purpose of visual object retrieval is to search a specific object in large-scale image/video datasets. In contrast, similar image search or near duplicate detection aims at retrieving globally similar images. This difference is illustrated in Figure 1, where it appears that the two tasks mostly differ by how the query is defined. In object retrieval, a bounding box or a shape delimits the query entity, such as a person, place, or other object. In contrast, similar image search assumes that the query is the full image.

This task is the visual counterpart of searching by query terms in textual information retrieval, where a few words or a short descriptions are compared with large textual documents. Early in the 60's, the SMART system designed by Salton [20], considered text retrieval as an asymmetrical scenario. Similarly, state-of-the-art textual engines rely on asymmetrical measures, for instance by using different term weighting schemes for the query and database elements, such as in the Okapi [18, 19] method. For a recent



(a) visual object retrieval (b) similar image search

Figure 1. Differences between object retrieval and similar image search. In (a) object retrieval, the query is delimited by a bounding box or a shape, while in (b) similar image search, the query and database objects are of the same kind. This paper shows the importance of designing an asymmetrical dissimilarity measure for object retrieval, in order to better take into account the different inlier ratios between query objects and database images.

overview of these schemes, the reader may refer to a recent book by Manning *et al.* [11]. In this paper, we only consider unsupervised object retrieval, where no annotation is provided. Our goal is therefore not to determine the class of an image, but rather to find images containing visually similar objects, as in the Instance search task of the Trecvid evaluation campaign [14].

**Related work.** The bag-of-words (BoW) framework [21] is the long-lasting standard approach for large-scale image and visual object retrieval. Recent schemes [2, 17, 24] derived from this approach exhibit state-of-the-art performance on several benchmarks. This baseline method has been improved in several ways in recent years, in particular to compensate for quantization errors, *e.g.*, by using large vocabularies [13, 16], multiple- or soft-assignment [7, 17] and Hamming embedding [5]. Other techniques include improving the initial ranking list by exploiting spatial geometry [17, 21, 23] and query expansion [3].

All these approaches rely on a symmetrical metric to produce the initial ranking of the image collection, such as the $\ell_1$ [13] or Euclidean ($\ell_2$) [16] distances. Such choices are convenient: they correspond to some underlying vector space and allow the use of dedicated machine learning

techniques such as SVM or metric learning [4]. As a result, the body of literature on asymmetrical metrics is limited. An inspiring work is the framework proposed by Kulis *et al.* [8], who consider asymmetrical kernels for the problem of domain adaptation in image classification. However, this method requires annotation and is too general to address the unsupervised visual object retrieval problem. Note that, although Bregman divergences such as Kullback-Leibler [9] are asymmetrical by construction, they do not reflect the underlying assumptions underpinning visual object recognition and lead to poor comparison results.

Our paper specifically considers the visual object retrieval problem. We argue that symmetrical metrics are not optimal for judging the presence of query objects. This is because most of the area in the query image is useful: The bounding box or shape tightly delimits the relevant object. In contrast, the database images containing the object may also contain other objects or "stuff", *i.e.*, clutter. When the images are described by local descriptors, this leads to very different inlier ratios in the query and database images. This key aspect is not taken into account in existing schemes.

**Contributions.** First, we quantitatively analyze the different properties of the query and of the database images in visual object retrieval. We carried out our analysis on popular large-scale object retrieval datasets and the results show the extent to which this task is asymmetrical.

Focusing on the standard BoW method, we then propose new query-adaptive asymmetrical dissimilarities. They are specially designed to take into account the asymmetry of the comparison underpinning visual object retrieval. They are defined on-the-fly for each query in order to account for the expected inlier ratio. Yet they can be efficiently calculated by using an inverted file index.

The experiments are conducted on three large-scale datasets designed for visual object retrieval, namely Oxford105K and two datasets used in the instance search task of Trecvid. Our method improves the initial ranking in comparison with a symmetrical baseline that already achieves state-of-the-art performance for the initial ranking.

The rest of this paper is organized as follows. Section 2 introduces the datasets used through the paper to evaluate visual object retrieval, and illustrates the importance of asymmetry in this task. Section 3 describes our query-adaptive asymmetrical dissimilarities and how to calculate them with an inverted index. Our results on three large-scale datasets are reported in Section 4, along with a comparison with the state of the art. Section 5 concludes the paper.

## 2. Object retrieval: an asymmetrical scenario

This section shows that the asymmetry phenomenon is prevalent in visual object retrieval datasets. For this purpose, we first introduce three public benchmarks, which correspond to application scenarios where the query is an object instance. Then we describe the baseline system. Finally, we analyze the asymmetry of inliers in query and database images in visual object recognition tasks and discuss the limitations of the symmetrical BoW in this context.

### 2.1. Object retrieval benchmarks

**Oxford105K.** The Oxford buildings dataset (Oxford5K) [16] consists of 5062 high-resolution images crawled from Flickr. Another set comprising around 100,000 Flickr images is usually appended to form the Oxford105K dataset. A Region of Interest (ROI) is defined for each query image. It is a bounding box delimiting the building of interest. Following common practice, we consider two evaluation scenarios:

1. Oxford105K: The dataset is learned on Oxford5K [16].
2. Oxford105K$^*$: The vocabulary is independently trained on another dataset, namely the Paris building set [2, 17]. The performance is tested on the Oxford105K. This scenario corresponds to the case where the images are not known beforehand [2, 6, 17].

**TrecVid instance search: INS2011 and INS2012.** The TrecVid INstance Search (short: INS) datasets were released in the context of the evaluation campaign organized by NIST. BBC rushes and internet videos comprise the test data of the INS2011 and INS2012 datasets, respectively. The duration of the video clips in the test datasets is generally not longer than 1 minute (30 seconds on average).

The task description [14] is as follows. A large collection of video clips defines the dataset to be searched. Several query topics are defined. A query topic may refer to a person, an object or a place. Each query topic consists of several query images and corresponding masks delimiting the ROI. For each query topic, the system has to return the 1000 video clips that are most likely to contain a recognizable instance of the query topic. The INS task is rather challenging, as shown in Figure 1(a)-right: the objects are small and the database is represented by millions of frames. As a result, the quality of the initial ranking (before spatial verification and query expansion) is critical.

**Evaluation protocol.** The performance on each dataset is evaluated by the official score, *i.e.*, the mean average precision (mAP) on the Oxford105K [16] and the mean inferred average precision (infAP) for the TrecVid datasets.

Table 1 provides detailed information about the three benchmarks. All the images are described by SIFT descriptors [10] extracted with the Hessian-Affine detector [12]. For Oxford105K, we used the descriptors provided by Perďoch *et al.* [15] in order to allow for a direct comparison. The RootSIFT [2] post-processing is used on the Oxford105K and the INS2012 datasets, as it improves the re-

Table 1. Details of benchmark datasets.

| ↓ Dataset | #Images | #Videos | #SIFT points | #Queries |
|---|---|---|---|---|
| Oxford105K | 105,133 | N/A | 253,761,866 | 55 |
| INS2011 | 1,608,405 | 20,982 | 1,206,953,361 | 25 |
| INS2012 | 2,228,356 | 76,751 | 3,055,162,839 | 21 |

trieval performance at no cost. On TrecVid INS, the video clips are sampled with the rate of 3 frames per second.

### 2.2. The baseline BoW system and its limitations.

We briefly introduce the BoW system [21], which serves as our baseline, and discuss its limitation in the context of visual object retrieval.

Let us consider a database that consists of $N$ images. First, we extract SIFT features from each image. A large visual vocabulary comprising $k$=1 million visual words is trained by an efficient approximate k-means algorithm (AKM) [16]. After vector quantization with the visual vocabulary, each image is described by a $k$-dimensional histogram vector $\mathbf{T}_j \in \mathbb{R}^k$, $j = 1 \dots N$. Similarly, a query image $i$ is described by a histogram vector $\mathbf{Q}_i \in \mathbb{R}^k$ computed from the descriptors appearing in the ROI. The vectors $\mathbf{Q}_i$ and $\mathbf{T}_j$ correspond to BoW histograms, optionally weighted by *idf* terms.

In the standard scoring method, each vector is first $\ell_p$-normalized, with $p = 1$ or $p = 2$, and then the $\ell_p$-distance is computed between the query and all databases vectors to order the database images. In our notation, the distance is therefore computed as

$$\ell_p(\mathbf{Q}_i, \mathbf{T}_j) = \left\| \frac{\mathbf{Q}_i}{\|\mathbf{Q}_i\|_p} - \frac{\mathbf{T}_j}{\|\mathbf{T}_j\|_p} \right\|_p . \quad (1)$$

**A typical failure for visual object retrieval.** The toy example in Figure 2 illustrates the drawback of using Equation 1 as a scoring method in an asymmetrical object retrieval scenario. In the first row, the object region in the query image is delimited by an ellipse. The dataset consists of two test images. Let us assume that the object is described by two robust and repeatable visual words: a five-pointed star and a circle. In this case, Image 2 is the correct answer and contains *all* the features of the query object. But it also contains background corresponding to other visual content. The second row in Figure 2 shows that the standard scoring method produces the wrong result in this case. For the sake of exposition, let us assume that *idf* has no impact and consider the $\ell_1$ distance[1]. Such failures are frequent for small query objects like those of the Trecvid INS task, because the distance favors the selection of images described by a small number of features.

[1]The same conclusion holds for $\ell_2$ in this example.

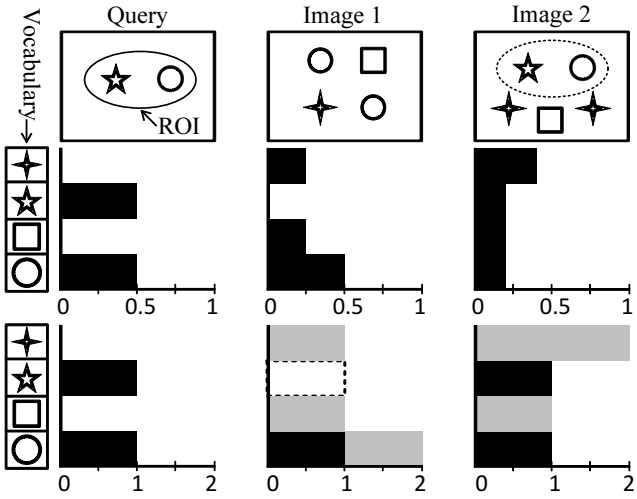



Figure 2. A toy example comparing the standard scoring method in the second row ($\ell_1(\mathbf{Q}, \mathbf{T}_1) = 1, \ell_1(\mathbf{Q}, \mathbf{T}_2) = 1.2$) with our asymmetrical dissimilarity in the third row ($\delta_1(\mathbf{Q}, \mathbf{T}_1, \infty) = 1, \delta_1(\mathbf{Q}, \mathbf{T}_2, \infty) = 0$).

### 2.3. Statistical analysis of asymmetry

In order to evaluate the extent of asymmetry in visual object retrieval, we consider the voting interpretation of the BoW framework [21, 5]. More specifically, a pair of features respectively from a query and test image is regarded as a match if these features are quantized to the same visual word. Each feature is allowed to be matched once at most. Our objective is to separate these features into three cases.

1. Inliers (Inl): Features belong to a matching pair (note that they may or may not correspond to a true match between the query and database images);
2. Query outliers (Qout): The query features (in the ROI) that do not correspond to any feature in the database image;
3. Database outliers (Dout): The features of the database that do not have any matching feature in the query ROI.

We estimate these quantities on the basis of the values of $\mathbf{Q}_i^l$ and $\mathbf{T}_j^l$, *i.e.*, the $l$-$th$ ($l = 1 \dots k$) component in the given query and database vectors. This is done by separately collecting the votes, as illustrated in Table 2. First, the maximum possible number of matching pairs $\min(\mathbf{Q}_i^l, \mathbf{T}_j^l)$ is an estimation of the number of inliers for this particular component $l$. The unmatched features are then counted as the outliers either of the query (if $\mathbf{Q}_i^l > \mathbf{T}_j^l$) or of the database (if $\mathbf{Q}_i^l < \mathbf{T}_j^l$ ) images. In summary, we separate the components $\mathbf{Q}_i^l$ and $\mathbf{T}_j^l$ according to the following equations:

$$\mathbf{Q}_i^l = \max(\mathbf{Q}_i^l - \mathbf{T}_j^l, 0) + \min(\mathbf{Q}_i^l, \mathbf{T}_j^l), \quad (2)$$

$$\mathbf{T}_j^l = \max(\mathbf{T}_j^l - \mathbf{Q}_i^l, 0) + \min(\mathbf{Q}_i^l, \mathbf{T}_j^l). \quad (3)$$

Table 2. Protocol to collect matching statistics from the histogram values $\mathbf{Q}_i^l$ and $\mathbf{T}_j^l$: the bottom-right cell collects the inliers, while the top-right and bottom-left cells respectively correspond to the outliers of the query and database images.

| $\downarrow \mathbf{Q}_i^l \quad \mathbf{T}_j^l \rightarrow$ | $=0$ | $>0$ |
|---|---|---|
| $=0$ | N/A | $\max(\mathbf{T}_j^l - \mathbf{Q}_i^l, 0)$ |
| $>0$ | $\max(\mathbf{Q}_i^l - \mathbf{T}_j^l, 0)$ | $\min(\mathbf{Q}_i^l, \mathbf{T}_j^l)$ |

| | |
|---|---|
| N/A | Dout $\approx 3686$ |
| Qout $\approx 367$ | Inl $\approx 15.8$ |

Figure 3. Query example ("bodleian") and its average number of inliers/outliers when matching the query ROI with the corresponding relevant images of Oxford105K.

For each relevant (query, database) pair, the quantities Inl, Dout, and Qout are estimated by summing the individual contributions of all the components $l = 1 \ldots k$. Figure 3 shows the estimation of these quantities for a particular query image contained in the Oxford105K benchmark.

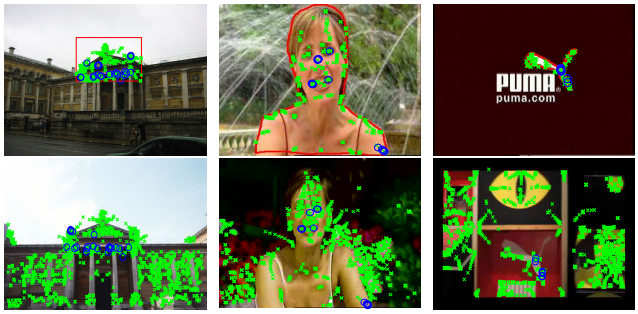
**Results of the analysis.** Table 3 reports the estimated inliers and outliers on the three datasets considered in this paper. These quantities are averaged over all query-database pairs that are relevant in terms of the ground-truth. Note that a joint average scheme [1, 24] is used for the TrecVid INS datasets: multiple images in each video clip and query topic are jointly quantized to form a single BoW vector by average pooling. This scheme was shown to be effective for image retrieval [24], and we have used it in all our experiments on the TrecVid benchmarks.

By defining the inlier ratio as the number of matched feature points divided by the total number of features, we calculate the inlier ratio associated with the query and database sides as Inl/(Inl + Qout) and Inl/(Inl + Dout), respectively. We define the outlier ratio in a similar way.

As is to be expected for visual object recognition of small objects, Table 3 clearly shows that Qout $<<$ Dout, meaning that the inlier ratio is much higher in the queries than in the corresponding database images; *i.e.*, the features points of the query ROI are significantly more likely to be present in a database image than the inverse. This is of course expected since additional visual content or clutter exists in database images. Figure 4 evidences this asymmetry in the inlier/outlier ratio by showing typical examples extracted from each of our evaluation datasets. Note also that some feature points labeled as matched do not strictly match each other. This is because a voting scheme based on BoW vectors, rather than a precise nearest neighbor search or stereo matching, implicitly builds loose correspondences.

Table 3. Estimation of the average number of Inl, Dout and Qout on the three datasets.

| $\downarrow$ Dataset | Dout | Qout | Inl |
|---|---|---|---|
| Oxford105K | 3 620.6 | 1 807.4 | 46.2 |
| INS2011 | 779.1 | 190.3 | 12.7 |
| INS2012 | 1 539.3 | 473.5 | 9.1 |




Oxford105K INS2011 INS2012

Figure 4. Examples visualizing the asymmetrical inlier/outlier ratio on the query and database side on each benchmark. Query regions are in red. Feature points labeled as inliers and outliers are marked with blue circles and green crosses, respectively.

In Table 3, note the average inlier ratio in the queries is very low on each dataset, especially the INS2012 dataset ($< 2\%$). This confirms the difficulty of object retrieval, and indicates that existing metrics are not likely to return images containing a small object surrounded by cluttered background.

## 3. Asymmetrical dissimilarity

The objective of the object retrieval task is to determine the existence of the query object, and it is inherently asymmetric: A appearing in B does not necessarily means that B also appears in A (see Figures 2 and 4). This is reflected in the asymmetry of the inlier ratio on the benchmarks. In the standard scoring framework, distance $\ell_p$ in Equation 1 is symmetrical, since $\ell_p(\mathbf{Q}_i, \mathbf{T}_j) = \ell_p(\mathbf{T}_j, \mathbf{Q}_i)$. For this reason, we deem that the standard BoW scoring method is better adapted to the symmetrical similarity image search problem (without ROI), but is not optimal for visual object retrieval. In short, we argue that a symmetrical metric is designed for measuring a symmetrical similarity problem, while the asymmetry of visual object recognition requires an asymmetrical dissimilarity.

This section describes asymmetrical dissimilarities that are specifically adapted to this task. Their design is motivated by the following observations:

- The normalization severely penalizes the database images in which the query object is small and corresponds to a small number of features (see Figure 2).

Table 4. Performance obtained with different parameter $w$ in Equation 4. Here $\ell_1(\mathbf{Q},\mathbf{T})$ is the baseline.

| Configurations | Oxford105K | Oxford105K* | INS2011 | INS2012 |
|---|---|---|---|---|
| $\delta_1(\mathbf{Q},\mathbf{T},0)$ | 0.3 | 0.3 | 0.02 | 0 |
| $\delta_1(\mathbf{Q},\mathbf{T},1)$ | 2.79 | 2.78 | 0.02 | 0 |
| $\delta_1(\mathbf{Q},\mathbf{T},\infty)$ | 65.29 | 38.85 | 44.88 | 19.51 |
| $\delta_1(\mathbf{Q},\mathbf{T},w_{\text{opt}})$ | 75.38 | 55.81 | 47.38 | 20.88 |
| $\ell_1(\mathbf{Q},\mathbf{T})$ | 73.88 | 54.47 | 45.16 | 19.83 |

- Ideally, the scoring should not depend too much on the amount of clutter in the database image; *i.e.*, Dout should not be penalized too much.
- In contrast, a feature appearing in the object has a higher probability of appearing in a relevant image; *i.e.*, Qout should receive a larger penalty.

After introducing our asymmetrical dissimilarities, we show how the computation is sped-up with an inverted file.

### 3.1. Asymmetrical penalties

We define our asymmetrical dissimilarity as follows:

$$\delta_p(\mathbf{Q}_i,\mathbf{T}_j,w) = \left\|\mathbf{d}(\mathbf{Q}_i,\mathbf{T}_j,w)\right\|_p, \tag{4}$$

where the $l\text{-}th$ component of the vector $\mathbf{d}(\mathbf{Q}_i,\mathbf{T}_j,w)$ is given by

$$d^l(\mathbf{Q}_i^l,\mathbf{T}_j^l,w) = w\times\max(\mathbf{Q}_i^l-\mathbf{T}_j^l,0)+\max(\mathbf{T}_j^l-\mathbf{Q}_i^l,0). \tag{5}$$

The parameter $w$ is a weight that takes into account the asymmetry of the problem. Equation 5 can be rewritten as

$$d^l(\mathbf{Q}_i^l,\mathbf{T}_j^l,w) = \begin{cases} w\left(\mathbf{Q}_i^l-\mathbf{T}_j^l\right) & \text{if } \mathbf{Q}_i^l > \mathbf{T}_j^l \\ \mathbf{T}_j^l-\mathbf{Q}_i^l & \text{if } \mathbf{Q}_i^l < \mathbf{T}_j^l \end{cases}. \tag{6}$$

Since we rely on relative values to establish the ranking list, Equation 5 only requires one weighting parameter $w$. It should be optimally related to the expected ratios between Qout and Dout (see Section 2). As one can deduce from Table 2, the values $w$ and 1 are penalties associated with the query and database (estimated) outliers, respectively. We intentionally give a larger weight, *i.e.*, $w > 1$, to the query outliers. This means that we severely penalize features that are detected in the query object regions having no corresponding features in the database image. In contrast, the database outliers receive a comparatively smaller penalty. This limits the impact, on the ranking, of the background appearing in the database images.

**Discussion.** We consider three particular choices for the parameter $w$, as shown in Table 4:

- The case $w = 0$ amounts to penalizing the database images based on Dout, *i.e.*, the estimated amount of background. Intuitively, this choice is not desirable because database images are expected to include clutter.
- The case $w = 1$ corresponds to a symmetrical case. It amounts to using the regular $\ell_p$ distance between the *unnormalized* histograms.
- The case $w \to \infty$, *i.e.*, using an arbitrarily large value, corresponds to the ideal case without considering the background in database images. It amounts to counting the number of Qout.

Compared with the baseline $\ell_1$, none of these choices is satisfactory, because none is adapted to the specific query and database. Instead, the next subsection introduces a query-dependent method that automatically adapts the weight $w$ to a given query and database.

### 3.2. Query-adaptive dissimilarity

The weight $w$ reflects the different inlier ratios between the query and database images. A naive strategy would be to fix it, as in the three particular cases mentioned before, thus we get $\delta_1(\mathbf{Q},\mathbf{T},w_{\text{opt}})$ in Table 4. A fixed optimal $w_{\text{opt}}$ yields better results than the baseline $\ell_1$. Yet the parameter $w_{\text{opt}}$ highly depends on the dataset, for instance, $w_{\text{opt}}$ is 700, 300, 1500 and 700 for the Oxford105K, Oxford105K*, INS2011 and INS2012, respectively.

In other terms, such a strategy implicitly assumes that the inlier ratio is constant across query and database images, which is not true in practice. We partially address this problem by automatically selecting $w$ on-the-fly, at query time. Substituting Equation 2, 3 into Equation 5 and then into Equation 4, we get:

$$\begin{aligned}\delta_p(\mathbf{Q}_i,\mathbf{T}_j,w) =& w\left\|\mathbf{Q}_i-\min(\mathbf{Q}_i,\mathbf{T}_j)\right\|_p \\ &+\left\|\mathbf{T}_j-\min(\mathbf{Q}_i,\mathbf{T}_j)\right\|_p.\end{aligned} \tag{7}$$

Recall that $\mathbf{Q}_i,\mathbf{T}_j$ are weighted by *idf* terms. Let us first consider the $\delta_1$ asymmetrical dissimilarity. Note also the vectors involved in Equation 7 are all positives. After dropping the constant term $w\|\mathbf{Q}_i\|_1$, which has not impact on the relative ranking of the images, and setting $\bar{w} = w+1$, we re-define an equivalent dissimilarity measure as

$$\delta_1(\mathbf{Q}_i,\mathbf{T}_j,\bar{w}) = \left\|\mathbf{T}_j\right\|_1 - \bar{w}\left\|\min(\mathbf{Q}_i,\mathbf{T}_j)\right\|_1. \tag{8}$$

The two terms on the right side of Equation 8 are intuitively understood as follows. Test images that are uncluttered (i.e. $\|\mathbf{T}_j\|_1$ is small) and have many matches with the query ($\|\min(\mathbf{Q}_i,\mathbf{T}_j)\|_1$ is large) will be regarded as similar to the query region. The quantity $\bar{w}$ balances the impact of clutter and positive matches in the scoring. In our method, instead of directly setting $\bar{w}$ to a fixed value, we set a parameter $\alpha_1$ related to $\bar{w}$ by the following equation:

$$\bar{w} = \alpha_1 \frac{\sum_{j=1}^{N}\left\|\mathbf{T}_j\right\|_1}{\sum_{j=1}^{N}\left\|\min(\mathbf{Q}_i,\mathbf{T}_j)\right\|_1}. \tag{9}$$

The benefit of this expression is that it automatically adapts the dissimilarity function to 1) the database and to 2) the particular query $\mathbf{Q}_i$ with the denominator. Overall, our method only requires the parameter $\alpha_1$ (whose impact is thoroughly analyzed in Section 4). Similarly, the dissimilarity $\delta_2$ becomes

$$\begin{aligned}
&\delta_2(\mathbf{Q}_i, \mathbf{T}_j, w) \\
&= w \left\| \mathbf{Q}_i - \min(\mathbf{Q}_i, \mathbf{T}_j) \right\|_2 + \left\| \mathbf{T}_j - \min(\mathbf{Q}_i, \mathbf{T}_j) \right\|_2 \\
&= w \left( \|\mathbf{Q}_i\|_2^2 - 2\mathbf{Q}_i \cdot \min(\mathbf{Q}_i, \mathbf{T}_j) + \|\min(\mathbf{Q}_i, \mathbf{T}_j)\|_2^2 \right)^{\frac{1}{2}} \\
&\quad + \left( \|\mathbf{T}_j\|_2^2 - 2\mathbf{T}_j \cdot \min(\mathbf{Q}_i, \mathbf{T}_j) + \|\min(\mathbf{Q}_i, \mathbf{T}_j)\|_2^2 \right)^{\frac{1}{2}}
\end{aligned} \tag{10}$$

For the same reason as in the $\delta_1$ case, we set a parameter $\alpha_2$ in Equation 11 instead of directly setting $w$:

$$w = \alpha_2 \frac{\sum_{j=1}^{N} \sqrt{\|\mathbf{T}_j\|_2^2 - 2\mathbf{T}_j \min(\mathbf{Q}_i, \mathbf{T}_j) + \|\min(\mathbf{Q}_i, \mathbf{T}_j)\|_2^2}}{\sum_{j=1}^{N} \sqrt{\|\mathbf{Q}_i\|_2^2 - 2\mathbf{Q}_i \min(\mathbf{Q}_i, \mathbf{T}_j) + \|\min(\mathbf{Q}_i, \mathbf{T}_j)\|_2^2}}. \tag{11}$$

### 3.3. Speeding-up retrieval with an inverted index

The direct calculation of the dissimilarities with Equations 4 or 5 requires one to access all the vector components. It is therefore inefficient when $k$ is large, as in the case of our million-sized vocabulary. However, it appears that the proposed dissimilarities (Equations 8 and 10) can be decomposed such that the calculation only involves 1) the shared nonzero components of two vectors to be compared, as in the case of BoW [21], along with 2) terms that separately depend on the database and query (computed offline and on-the-fly, respectively). It is therefore efficiently implemented based on an inverted index. The symmetrical distances and our asymmetrical dissimilarities have comparable complexities. The amount of visited memory is identical. The quantities $\|\mathbf{T}_j\|_p$ are pre-computed during the offline indexing stage. The computation burden of the query-specific terms of $\delta_p$ is comparable to that of the $\ell_p$ distance. For instance, the term $\|\min(\mathbf{Q}_i, \mathbf{T}_j)\|_1$ in Equation 8 is also calculated in the case of the $\ell_1$ distance. In practice, it takes less than 0.1 second to search an object in the Oxford105K dataset with the inverted file structure.

## 4. Experiments and analysis

This section describes our experiments on three large-scale datasets designed for object retrieval. In order to compare our asymmetrical dissimilarities with a competitive baseline, we first optimized the choices involved in the baseline system for each dataset. As we will see, our baseline outperformed the state of the art by itself on some benchmarks. After analyzing the impact of the additional parameter involved in our approach, we provide a comparison with the best baseline and the state of the art.

Table 5. Performance of the baseline for different configurations.

| Configurations | Oxford105K | Oxford105K* | INS2011 | INS2012 |
|---|---|---|---|---|
| *The selected* | *73.88* | *54.47* | *45.16* | *21.71* |
| Different AS | 70.93 | 48.95 | 44.89 | 21.14 |
| Different DS | 70.03 | 51.07 | 45.13 | 19.83 |
| With re-ranking | 76.59 | 71.82 | 30.76 | 14.15 |

Note: *the selected* configuration (top) is: using soft assignment and hard assignment on the Oxford and TrecVid datasets, respectively; utilizing $\ell_1$ metric on all datasets except the INS2012. *Legend for alternative choices:* AS: alternative assignment scheme (swap hard and soft with selected); DS: choice of the distance (swap $\ell_1$ with $\ell_2$).

In the experiment, we used a BoW baseline system without any re-ranking step, such as spatial re-ranking [17, 21] and query expansion [3], because we focus on improving the initial ranked accuracy, which is critical especially for difficult datasets. Most re-ranking algorithms, such as spatial verification [17, 21] or query expansion [3], require the short-list to be of sufficient quality to produce good results. Moreover, they are mostly complementary to our method.

**Configuration of the baseline system.** Table 5 evaluates the different options considered for the baseline system.

*Hard or soft assignment.* As previously reported in the literature [17], soft assignment improves the results on the Oxford105K dataset. But unexpectedly, it reduces the performance on the INS TrecVid datasets. Our interpretation is that the joint average pooling compensates the loss in quantization, at least to some extent, thus making the soft assignment unnecessary or even undesirable.

$\ell_1$ vs $\ell_2$. As shown in the literature [7, 13, 22], the best norm for BoW depends on the dataset. The $\ell_2$ metric is better on the INS2012 dataset, whereas the $\ell_1$ distance wins on the others. In our experiments, we used the best configuration for each dataset and kept this choice consistent with our asymmetrical dissimilarities.

*Spatial re-ranking* improves the performance only on Oxford. As mentioned above, we will not consider any re-ranking scheme like this in the remainder of this section, since we focus on improving the initial ranking list.

**Impact of the parameter $\alpha_p$ and relative improvement.** *The $\ell_1$/$\delta_1$ case.* Figure 5 shows the impact of the parameter $\alpha_1$ associated with the $\delta_1$ dissimilarity (see Equation 9). We include the performance of the baseline system (dash lines) provided by the $\ell_1$ distance for a fair comparison.

Our dissimilarity consistently outperforms the symmetrical baseline: The improvement is of +5.77%, +12.08%, +7.40% and +8.88% on the Oxford105K, Oxford105K*,

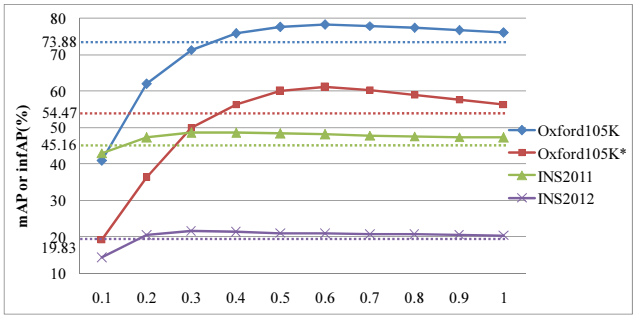



Figure 5. Impact of the parameter $\alpha_1$ (horizontal axis) in Equation 9: performance (vertical axis) of the $\delta_1$ asymmetrical dissimilarity.

INS2011, and INS2012, respectively. As expected, the performance monotonically increases with $\alpha_1$ until it attains a peak $\alpha_1^*$. Then it monotonically decreases. This shows the importance of balancing the clutter and matching terms in Equation 8. Interestingly the performance is remarkably stable around the peak: setting $\alpha_1 = 0.5$ leads to close-to-optimal results on all benchmarks, and which is consistently better than $\delta_1(\mathbf{Q}, \mathbf{T}, w_{opt})$ in Table 4.

*The $\ell_2$/$\delta_2$ case.* For the $\delta_2$ asymmetrical dissimilarity, we draw the same conclusions as above. However, as in the symmetrical case, the $\delta_2$ dissimilarity only slightly outperforms the corresponding $\delta_1$ on the INS2012 dataset (+1.30%) and gives worse results on other benchmarks. This dissimilarity systematically achieves its best performance in the extreme case of $\alpha_2 \to \infty$, which amounts to totally ignoring the clutter term.

**Sample results.** Figure 6 compares the ranked lists returned by our $\delta_1$ dissimilarity with those associated with the $\ell_1$ distance. Our method is especially better at returning relevant images containing a significant amount of clutter. One key problem of the symmetrical distance is that the same samples containing the query ROI are not necessarily ranked before the others: in the first example, the image same as the query is ranked second, and in the last example, the most bottom-right sample returned by the $\delta_1$ dissimilarity does not appear before some of the negative samples.

**Comparison with the state of the art.** The best results we are aware of are reported in Table 6. The best results reported for the quality of the initial short-list are given by $\text{Best}_1$. They reflect the score of the initial ranking and therefore correspond to the same setup as the one used in our technique. Note first that our baseline system (Best $\ell_p$) already outperforms this state of the art ($\text{Best}_1$) for producing the initial short-list.

Second, our asymmetrical method (Best $\delta_p$) is consistently better than its symmetrical counterpart for the best choice (Best $\ell_p$) of the baseline system. Recall that Best $\ell_p$ is optimally selected in Table 5. This shows the effectiveness of our asymmetrical dissimilarities. The improvement is very significant, except in the case of INS2012 (comparable results). This might be related to the fact that we generally observe that the relative improvement of our method is better for $p = 1$ than for $p = 2$, and that $p = 2$ is the best choice for $\ell_p$ and $\delta_p$ on the INS2012 dataset (only).

Table 6. Comparison with the baseline (Best $\ell_p$) and the state of the art ($\text{Best}_1$). The scores of $\text{Best}_2$ are reported for reference but are not directly comparable, as they generally include multiple features, spatial verification or/and query expansion.

| ↓ Dataset | Best $\ell_p$ | Best $\delta_p$ | $\text{Best}_1$ | $\text{Best}_2$ |
|---|---|---|---|---|
| Oxford105K | 73.88 | 78.14 | 62.2 [1] | 89.1 [2] |
| Oxford105K* | 54.47 | 61.05 | 34.3 [17] | 77.2 [15] |
| INS2011 | 45.16 | 48.50 | – | 55.6 [24] |
| INS2012 | 21.71 | 21.87 | – | 27.0 [14] |

*Remark:* For the sake of completeness, the table also reports the best results ($\text{Best}_2$) achieved by using, additionally, multiple features, spatial verification or other re-ranking schemes such as query expansion. Those results are therefore not directly comparable to our technique, and these additional techniques are arguably complementary to our method. In addition, we underline that for INS2011 and INS2012 benchmarks, the scores $\text{Best}_2$ are obtained by using the interest points outside the ROI, *i.e.*, by exploiting the context around the object. This does not correspond to our visual object recognition scenario[2].

## 5. Conclusions

This paper specifically addressed the asymmetrical phenomenon arising in an visual object retrieval scenario. This led us to propose new dissimilarities measures, adapted to the bag-of-words representation, that explicitly take into account this aspect to improve the retrieval quality. Our measures get rid of the normalization factor to address the cases where a small object appears in an image populated with many features. In addition, it takes into account the different inliers ratios. A key feature is to automatically adapt, per query, a parameter that reflects the different inlier ratios in the query and database images. Our dissimilarities come at not cost, as they are implemented with a vanilla inverted index like those used for symmetrical distances.

Its effectiveness is demonstrated in comprehensive experiments carried out on large-scale benchmarks. To conclude, we believe that our method is fully compatible with the standard object retrieval architecture [2, 16], meaning that further refinements such as spatial re-ranking or query expansion can be seamlessly integrated with it.

[2]This is effective on INS2011/INS2012 because the objects are often occurring with the same background.

Figure 6. Comparison of ranked lists. Query objects are on the left side. On the right side, the top 10 returns are ranked from left to right: For each example, the upper and lower rows are returned by $\ell_1$ and $\delta_1$, and the accuracies from top to bottom are 39.63 *vs*. 67.46, 27.2 *vs*. 50.17, and 39.71 *vs*. 56.35. Positive (negative) samples are marked with green (red) bounding boxes.

## References


[1] R. Arandjelović and A. Zisserman. Multiple queries for large scale specific object retrieval. In *BMVC*, 2012. 4, 7

[2] R. Arandjelović and A. Zisserman. Three things everyone should know to improve object retrieval. In *CVPR*, 2012. 1, 2, 7

[3] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman. Total recall: Automatic query expansion with a generative feature model for object retrieval. In *ICCV*, 2007. 1, 6

[4] S. C. Hoi, R. Jin, J. Zhu, and M. R. Lyu. Semi-supervised SVM batch mode active learning for image retrieval. In *CVPR*, 2008. 2

[5] H. Jégou, M. Douze, and C. Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *ECCV*, 2008. 1, 3

[6] H. Jégou, M. Douze, and C. Schmid. On the burstiness of visual elements. In *CVPR*, 2009. 2

[7] H. Jégou, H. Harzallah, and C. Schmid. A contextual dissimilarity measure for accurate and efficient image search. In *CVPR*, 2007. 1, 6

[8] B. Kulis, K. Saenko, and T. Darrell. What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In *CVPR*, 2011. 2

[9] S. Kullback and R. A. Leibler. On information and sufficiency. *Ann. Math. Statist.*, 22(1):79–86, 1951. 2

[10] D. Lowe. Distinctive image features from scale-invariant key points. *IJCV*, 60:91–110, 2004. 2

[11] C. D. Manning, P. Raghavan, and H. Schütze. *An Introduction to Information Retrieval*. Cambridge University Press, 2009. 1

[12] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *IJCV*, 1:63–86, 2004. 2

[13] D. Nistér and H. Stewénius. Scalable recognition with a vocabulary tree. In *CVPR*, 2006. 1, 6

[14] O. Paul, G. Awad, J. Fiscus, G. Sanders, and B. Shaw. Trecvid 2012 - an introduction of the goals, tasks, data, evaluation mechanisms and metrics. In *TRECVID*, 2012. 1, 2, 7

[15] M. Perďoch, O. Chum, and J. Matas. Efficient representation of local geometry for large scale object retrieval. In *CVPR*, 2009. 2, 7

[16] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *CVPR*, 2007. 1, 2, 3, 7

[17] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *CVPR*, 2008. 1, 2, 6, 7

[18] S. E. Robertson and K. S. arck Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27(3), 1976. 1

[19] S. E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In *TREC*, 1994. 1

[20] G. Salton and M. J. McGill. *Introduction to modern information retrieval*. McGraw-Hill, 1983. 1

[21] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV*, 2003. 1, 3, 6

[22] P. Tirilly and V. Claveau. Distances and weighting schemes for bag of visual words image retrieval. In *ICMR*, 2010. 6

[23] W. Zhou, Y. Lu, H. Li, Y. Song, and Q. Tian. Spatial coding for large scale partial-duplicate web image search. In *ACM Multimedia*, 2010. 1

[24] C.-Z. Zhu and S. Satoh. Large vocabulary quantization for searching instances from videos. In *ICMR*, 2012. 1, 4, 7