Asymmetric Distances for Binary Embeddings

Albert Gordo, Florent Perronnin, Yunchao Gong, Svetlana Lazebnik

Abstract—In large-scale query-by-example retrieval, embedding image signatures in a binary space offers two benefits: data compression and search efficiency. While most embedding algorithms binarize both query and database signatures, it has been noted that this is not strictly a requirement. Indeed, asymmetric schemes which binarize the database signatures but not the query still enjoy the same two benefits but may provide superior accuracy. In this work, we propose two general asymmetric distances which are applicable to a wide variety of embedding techniques including Locality Sensitive Hashing (LSH), Locality Sensitive Binary Codes (LSBC), Spectral Hashing (SH), PCA Embedding (PCAE), PCA Embedding with random rotations (PCAE-RR), and PCA Embedding with iterative quantization (PCAE-ITQ). We experiment on four public benchmarks containing up to 1M images and show that the proposed asymmetric distances consistently lead to large improvements over the symmetric Hamming distance for all binary embedding techniques.

Index Terms—Large-scale retrieval, binary codes, asymmetric distances.

1 Introduction

Recently, the computer vision community has witnessed an explosion in the scale of the datasets it has had to handle. While standard image benchmarks such as PASCAL VOC [2] or CalTech 101 [3] used to contain only a few thousand images, resources such as ImageNet [4] (14 million images) and Tiny images [5] (80 million images) are now available. In parallel, more and more sophisticated image descriptors have been proposed including the GIST [6], the bag-of-visual-words (BOV) histogram [7], [8], the Fisher vector (FV) [9], [10] or the Vector of Locally Aggregated Descriptors (VLAD) [11]. Descriptors with thousands or tens of thousands of dimensions have become the norm rather than the exception. Consequently, handling these gigantic quantities of data has become a challenge on its own.

When dealing with large amounts of data, there are two considerations of paramount importance. The first one is the *computational cost*: the computation of the distance between two image signatures should rely on efficient operations. The second one is the *memory cost*: the memory footprint of the objects should be small enough so that all database image signatures fit in RAM. If this is not the case, *i.e*. if a significant portion of the database signatures has to be stored on disk, then the response time of a query collapses because the disk access is much slower than that of RAM access.

These considerations have directly motivated research in learning compact binary codes [12], [13], [14], [15], [16], [11], [17], [18], [19], [20], [21]. A desirable property of such coding schemes is that they should map similar data points (with respect to a given metric such as the Euclidean distance) to similar binary vectors (*i.e*. vectors with a small Hamming distance). Transforming high-dimensional real-valued image descriptors into compact binary codes directly addresses both memory and computational problems. First the compression enables to store a large number of codes in RAM. Second, the Hamming distance is extremely efficient to compute in hardware, which enables the exhaustive computation of millions of distances per second, even on a single CPU.

However, it has been noted that compressing the query signature is not mandatory [22], [23], [17]. Indeed, the additional cost of storing in memory a single non-binarized signature is negligible. Also, the distance between an original signature and a compressed signature can still be computed efficiently through look-up table operations. As the distance is computed between two different spaces, these algorithms are referred to as *asymmetric*. A major benefit of asymmetric algorithms is that they can achieve higher accuracy for a fixed compression rate because they take advantage of the more precise position information of the query. We note however that the asymmetric algorithms presented in [22], [23], [17] are tied to specific compression schemes. Dong *et al*. [22] presented an asymmetric algorithm for compression schemes based on random projections. Jégou *et al*. [23] proposed an asymmetric algorithm for compression schemes based on vector quantization. Brandt [17] subsequently used a similar idea.

In this work, we generalize these asymmetric schemes to a broader family of binary embeddings. We first provide an overview of several binary embed-

- *A. Gordo is with the Computer Vision Center, Universitat Autònoma de Barcelona, Spain. E-mail: agordo@cvc.uab.es.*
- *F. Perronnin is with Textual Visual Pattern Analysis, Xerox Research Centre Europe, France. E-mail: florent.perronnin@xrce.xerox.com.*
- *Y. Gong is with the computer science department, University of North Carolina at Chapel Hill, USA. E-mail: yunchao@cs.unc.edu.*
- *S. Lazebnik is with the comoputer science department, University of Illinois at Urbana-Champaign, USA. E-mail: lazebnik@cs.unc.edu.*
- *A preliminary version of this paper [1] appears in CVPR2011.*

ding algorithms (including Locality Sensitive Hashing (LSH) [12], [13], Locality-Sensitive Binary Codes (LSBC) [14], Spectral Hashing (SH) [15], PCA Embedding (PCAE) [1], [21], PCAE with random rotations (PCAE-RR), and PCAE with iterative quantization (PCAE-ITQ) [21]) showing that they can be decomposed into two steps: i) the signatures are first embedded in an intermediate real-valued space and ii) thresholding is performed in this space to obtain binary outputs. A key insight that our asymmetric distances will exploit is that the Euclidean distance is a natural metric in the intermediate real-valued space.

Building on the previous analysis we propose two asymmetric distances which can be broadly applied to binary embedding algorithms. The first one is an expectation-based technique inspired by [23]. The second one is a lower-bound-based technique which generalizes [22].

We show experimentally on four datasets of different nature that the proposed asymmetric distances consistently and significantly improve the retrieval accuracy of LSH, LSBC, SH, PCAE, PCAE-RR, and PCAE-ITQ over the symmetric Hamming distance. Although the lower-bound and expectation-based techniques are very different in nature, they are shown to yield very similar improvements.

The remainder of this article is organized as follows. In the next section, we provide an analysis of several binary embedding techniques. In section 3 we build on the previous analysis to propose two asymmetric distance computation algorithms for binary embeddings. In section 4 we provide experimental results. Finally in section 5 we discuss conclusions and directions for future work.

2 Theoretical Analysis of Binary Embeddings

We now provide a review of several successful binary embedding techniques: LSH, LSBC, SH, PCAE, PCAE-RR, and PCAE-ITQ.

Let us introduce a set of notations. Let x be an image signature in a space Ω and let h_k be a binary embedding function, *i.e.* $h_k : \Omega \rightarrow \{0,1\}$ (some authors prefer the convention $h_k : \Omega \rightarrow \{-1,+1\}$). A set $\mathcal{H} = \{h_k, k = 1 \dots K\}$ of K functions defines a multi-dimensional embedding function $h : \Omega \rightarrow \{0,1\}^K$ with $h(x) = [h_1(x), \dots, h_K(x)]'$ (and the apostrophe denotes the transpose).

We show that for LSH, LSBC, SH, PCAE, PCAE-RR, and PCAE-ITQ, the functions h_k can be decomposed as follows:

$$h_k(x) = q_k[g_k(x)], \tag{1}$$

where $g_k : \Omega \rightarrow \mathbb{R}$ is the real-valued embedding function, and $q_k : \mathbb{R} \rightarrow \{0,1\}$ is the binarization function. We denote $g : \Omega \rightarrow \mathbb{R}^K$ with $g(x) = [g_1(x), \dots, g_K(x)]'$. If we have two image signatures x and y, we also show that the Euclidean distance is the natural metric between $g(x)$ and $g(y)$ and that it approximates the original distance between x and y. Thus, we can write the squared Euclidean distance as $d(x,y) \approx d(g(x), g(y)) = \sum_k d(g_k(x), g_k(y))$.

In the rest of this section, we survey a number of binary embeddings, which can be classified into two types: those based on random projections (LSH and LSBC), and those based on learning the hashing functions (PCAE, PCAE-RR, PCAE-ITQ, or SH).

2.1 Hashing with Random Projections

2.1.1 *Locality Sensitive Hashing (LSH)*

In LSH, the functions h_k are called hash functions and are selected to approximate a similarity function sim in the original space $\Omega \in \mathbb{R}^D$. Valid hash functions h_k must satisfy the LSH property:

$$Pr\left[h_k(x) = h_k(y)\right] = sim(x,y). \tag{2}$$

Here we focus on the case where sim is the cosine similarity $sim(x,y) = 1 - \frac{\theta(x,y)}{\pi}$, for which a suitable hash function is [13]:

$$h_k(x) = \sigma\left(r'_k x\right), \tag{3}$$

with

$$\sigma(u) = \begin{cases} 0 & \text{if } u < 0, \\ 1 & \text{if } u \geq 0. \end{cases} \tag{4}$$

The vectors $r_k \in \mathbb{R}^D$ are drawn from a multi-dimensional Gaussian distribution p with zero mean and identity covariance matrix I_D. We therefore have $q_k(u) = \sigma(u)$ and $g_k(x) = r'_k x$. In such a case the natural distance between $g(x)$ and $g(y)$ in the intermediate space is the Euclidean distance as random Gaussian projections preserve the Euclidean distance in expectation. This property stems from the following equality:

$$\mathbb{E}_{r \sim p}\left[||r'x - r'y||^2\right] = ||x - y||^2. \tag{5}$$

We note that centering the data around the origin can impact LSH very positively, especially when dealing with non-negative data such as GIST vectors [6]. This is because, given a set of points $\{x_i, i = 1 \dots N\}$ with zero-mean and a random direction r_k, we have the guarantee that $\frac{1}{N}\sum_{i=1}^{N} r'_k x_i = 0$, *i.e.* the distribution of values $r'_k x_i$ is centered around the LSH binarization threshold. If the data is not centered around the origin, the mean of the $r'_k x_i$ values can be very different from zero. We have observed cases where the projections all had the same sign, leading to weakly discriminative embedding functions [1].

The previous analysis can be readily extended to the Kernelized LSH approach of Kulis and Grauman [16]

[1] An alternative would be to choose a per-dimension threshold equal to the median of the $r'_k x_i$ values. However, this would not guarantee anymore the convergence to the cosine.

as the Mercer kernel between two objects is just a dot product in another space using a non-linear mapping $\phi(x)$ of the input vectors. This enables one to extend LSH as well as the asymmetric distance computations beyond vectorial representations.

2.1.2 Locality-Sensitive Binary Codes (LSBC)

Consider a Mercer kernel $K : \mathbb{R}^D \times \mathbb{R}^D \rightarrow \mathbb{R}$ that satisfies the following properties for all points x and y:

1) It is translation invariant: $K(x, y) = K(x - y)$.
2) It is normalized: *i.e.*, $K(x-y) \leq 1$ and $K(0) = 1$.
3) $\forall \alpha \in \mathbb{R} : \alpha \geq 1$, $K(\alpha x - \alpha y) \leq K(x, y)$.

Two well known examples are the Gaussian kernel and the Laplacian kernel. Raginsky and Lazebnik [14] showed that for such kernels, $K(x, y)$ can be approximated by $1-2\text{Ha}(h(x), h(y))/n$ with $h_k(x) = q_k(g_k(x))$ where:

$$g_k(x) = \cos(r'_k x + b_k), \tag{6}$$

$$q_k(u) = \sigma(u - t_k). \tag{7}$$

b_k and t_k are random values drawn respectively from unif$[0, 2\pi]$ and unif$[-1, +1]$, and n is the number of bits. The vectors r_k are drawn from a distribution p_K which depends on the particular choice of the kernel K. For instance, if K is the Gaussian kernel with bandwidth γ (note that the bandwidth is the inverse of the variance), then p_K is a Gaussian distribution with mean zero and covariance matrix γI_D. As the size of the binary embedding space increases, $1 - 2\text{Ha}(h(x), h(y))/n$ is guaranteed to converge to a function closely related to $K(x, y)$.

We know from Rahimi and Recht [24] that $g(x)'g(y)$ is guaranteed to converge to $K(x, y)$ since $\mathbb{E}_{w_k \sim p_K} g_k(x)g_k(y) = K(x, y)$ and therefore that the embedding g preserves the dot-product in expectation. Since $K(x, x) = K(x - x) = K(0) = 1$, the norm $||g(x)||^2$ is also guaranteed to converge to 1, $\forall x$. In that case, $||g(x) - g(y)||^2 = ||g(x)||^2 + ||g(y)||^2 - 2g(x)'g(y) = 2(1 - g(x)'g(y))$. Therefore the Euclidean distance is equivalent to the dot-product and the Euclidean distance is preserved in the intermediate space in expectation.

2.2 Learning Hashing Functions

Hashing methods based on random projections such as LSH and LSBC have important properties, such as the guarantee to converge to the target kernel when the number of bits grows to infinity. However, a large number of bits may be necessary to obtain a sufficiently good approximation. When aiming at short codes, it may be more fruitful to learn the hashing functions rather than to resort to randomness.

We will focus on unsupervised code learning techniques, and especially on those based on PCA, since PCA seems to be a core component of the best-performing binary embedding methods: in Product Quantization [23] and Transform Coding [17], PCA is used as a preliminary step before binarizing the data. In [1] and [21], a direct PCA embedding is used. Spectral Hashing [15] can be understood as a way to assign more bits to the PCA dimensions with more energy.

In the following, let $\mathcal{S} = \{x_i, i = 1 \ldots N\}$, be a set of N signatures in $\Omega \in \mathbb{R}^D$ that are available for training purposes.

2.2.1 PCA Embedding (PCAE)

A very simple encoding technique is PCA embedding (PCAE) [21][1]. We can define PCAE as $h_k(x) = q_k(g_k(x))$, with

$$g_k(x) = w'_k(x - \mu), \tag{8}$$

$$q_k(u) = \sigma(u), \tag{9}$$

where μ is the mean of the signatures of $\mathcal{S}$, $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$, and where w_k is the eigenvector associated with the k-th largest eigenvalue of the covariance matrix of the signatures of $\mathcal{S}$. Despite its simplicity, PCAE can obtain very competitive results as will be shown in the experiments of Section 4.

According to [15], when producing binary codes, two desirable properties are: i) that the bits are pairwise uncorrelated, and ii) that the variance of each bit is maximized. As seen in [18], this leads to an NP hard problem that needs to be relaxed to be solved efficiently. The analysis of [18] also shows that projecting the data with PCA is an optimal solution of the relaxed version of this problem, conferring some theoretical soundness to PCAE.

In the case of PCAE, the Euclidean distance is the natural distance in the intermediate space, since PCA projections preserve, approximately, the Euclidean distance (the PCA directions are those that minimize the mean squared reconstruction error).

A possible drawback of using PCA projections for binarization purposes is that not all the dimensions contain the same energy after the projection. Since, after thresholding, all bits have the same weight, it can be important to balance the variance before quantizing the vectors.

One possible solution is to rotate the projected data. This rotation can be random, such as in PCAE-RR (Section 2.2.2), or learned, such as in PCAE-ITQ (Section 2.2.3). Another option is to assign more bits to the more relevant dimensions. In practice, Spectral Hashing (Section 2.2.4) can be seen as a way to achieve this goal, even though its theoretical foundations are different.

2.2.2 PCA Embedding + Random Rotations (PCAE-RR)

As noted in [23],[21], one simple way to balance the variances is to project the data with the PCA projections and then rotate the result with a random orthogonal matrix $R \in \mathbb{R}^{K \times K}$ (or, equivalently, if we put the column eigenvectors in a matrix $W \in \mathbb{R}^{D \times K}$, to project and rotate the data at the same time with a matrix $\tilde{W} = WR$). One way to generate this random orthogonal matrix is to first create a random matrix drawn from a $\mathcal{N}(0,1)$ distribution and perform a QR decomposition, as done in [25]. Another option is to perform an SVD decomposition of such matrix, as done in [21]. We follow the latter approach.

Therefore, the PCAE-RR embedding can be defined as $h_k(x) = q_k(g_k(x))$, with

$$g_k(x) = \tilde{w}_k'(x - \mu), \tag{10}$$

$$q_k(u) = \sigma(u), \tag{11}$$

where $\tilde{w}_k$ is the kth column of $\tilde{W}$. Note that rotating the data with an orthogonal matrix after the PCA projection is still an optimal solution of the formulation of [18]. Also, since orthogonal rotations preserve the Euclidean distance – already approximately preserved after the PCA rotation –, the natural distance in the intermediate space for PCAE-RR is also the Euclidean distance.

2.2.3 PCA Embedding + Iterative Quantization (PCAE-ITQ)

In [21], the idea of rotating the projections to balance the variances after PCA is taken a step further. The goal is to find the optimal orthogonal rotation R that minimizes the quantization loss in a training set:

$$\underset{R}{\operatorname{argmin}} \sum_{x \in \mathcal{S}} ||q(g(x)) - g(x)||^2, \tag{12}$$

with

$$g_k(x) = \tilde{w}_k'(x - \mu), \tag{13}$$

$$q_k(u) = 2\sigma(u) - 1, \tag{14}$$

and, as before, $\tilde{w}_k = (WR)_k$.

Intuitively, the goal is to map the points into the vertices of a binary hypercube. The closer the points are to the vertices, the smaller the quantization error will be. This optimization problem is related to the Orthogonal Procrustes problem [26], in which one tries to find an orthogonal rotation to align one set of points with another. The optimization can be solved iteratively, and involves computing an SVD decomposition of a $K \times K$ matrix at every iteration. A random orthogonal matrix is used as initial values of this matrix. Since we are usually interested in compact codes (*e.g.*, $K \leq 512$), obtaining the orthogonal rotation matrix R is quite fast. Note also that this optimization has to be computed only once, offline.

As in the case of PCAE-RR, we perform an orthogonal rotation after the PCA projection, and so the Euclidean distance is approximately preserved in the intermediate space.

2.2.4 Spectral Hashing (SH)

Given a similarity sim between objects in $\Omega \in \mathbb{R}^D$, and assuming that the distribution of objects in Ω may be described by a probability density function p, SH [15] attempts to minimize the following objective function with respect to h:

$$\int_{x,y} ||h(x) - h(y)||^2 sim(x,y) p(x) p(y) dx dy, \tag{15}$$

subject to the following constraints:

$$h(x) \in \{-1, 1\}^K, \tag{16}$$

$$\int_x h(x) p(x) dx = 0, \tag{17}$$

$$\int_x h(x) h(x)' p(x) dx = I_D. \tag{18}$$

As optimizing (15) under constraint (16) is a NP hard problem (even in the case where $K = 1$, *i.e.* of a single bit), Weiss *et al*. propose to optimize a *relaxed* version of their problem, *i.e.* to remove constraint (16), and then to binarize the real-valued output at 0. This is equivalent to minimizing:

$$\int_{x,y} ||g(x) - g(y)||^2 sim(x,y) p(x) p(y) dx dy, \tag{19}$$

with respect to g and then writing $h(x) = 2\sigma(g(x)) - 1$, *i.e.* $q(u) = 2\sigma(u) - 1$. The solutions to the relaxed problem are eigenfunctions of the weighted Laplace-Beltrami operators for which there exists a closed-form formula in certain cases, *e.g*. when sim is the Gaussian kernel and p is separable and uniform. To satisfy, at least approximately, the separability condition a PCA is first performed on the input vectors.

Minimizing the objective function (19) enforces the Euclidean distance between $g(x)$ and $g(y)$ to be (inversely) correlated with the Gaussian kernel for points whose kernel value is high (equivalently, whose Euclidean distance is low). This shows that in the case of SH the natural measure in the intermediate space is the Euclidean distance.

3 Asymmetric Distances

In the previous section we decomposed several binary embedding functions h_k into real-valued embedding functions g_k and quantization functions q_k. Let d denote the squared Euclidean distance [2]. We also showed that:

$$d(g(x), g(y)) = \sum_k d(g_k(x), g_k(y)). \tag{20}$$

is a natural distance in the intermediate space. We now propose two approximations of the quantity (20). In the following, x is assumed non-binarized, *i.e.* we

[2] We use the following abuse of notation for simplicity: d denotes both the distance between the vectors $g(x)$ and $g(y)$, *i.e.* $d : \mathbb{R}^K \times \mathbb{R}^K \rightarrow \mathbb{R}$, and the distance between the individual dimensions, *i.e.* $d : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$.

have access to the values $g_k(x)$ (and therefore also to $h_k(x)$), while y is binarized, *i.e.* we only have access to the values $h_k(y)$ (but not to $g_k(y)$).

3.1 Expectation-Based Asymmetric Distance

In [23] Jégou *et al*. proposed an asymmetric algorithm for compression schemes based on vector quantization. A codebook is learned through k-means clustering and a database vector is encoded by the index of its closest centroid in the codebook. The distance between the uncompressed query and a quantized database signature is simply computed as the Euclidean distance between the query and the corresponding centroid.

We now adapt this idea to binary embeddings. We note that in the case of LSH, LSBC, SH, or PCAE we have no notion of centroid. However, we note that the centroid of a given cell in k-means clustering can be interpreted as the expected value of the vectors assigned to this particular cell. Similarly, we propose an asymmetric expectation-based approximation d_E for binary embeddings.

Assuming that the samples in the original space Ω are drawn from a distribution p, we define:

$$d_E(x,y) = \sum_k d(g_k(x), \mathbb{E}_{u\sim p}[g_k(u)|h_k(u)=h_k(y)]). \quad (21)$$

$d(g_k(x), \mathbb{E}_{u\sim p}[g_k(u)|h_k(u)=h_k(y)])$ is the distance in the intermediate space between $g_k(x)$ and the expected value of the samples $g_k(u)$ such that $h_k(u) = h_k(y)$.

Since we generally do not have access to the distribution p, the expectation operator is approximated by a sample average. In practice, we randomly draw a set of signatures $\mathcal{S} = \{x_i, i = 1 \ldots N\}$ from Ω. For each dimension k of the embedding, we partition $\mathcal{S}$ into two subsets: $\mathcal{S}_k^0$ contains the signatures x_i such that $h_k(x_i) = 0$ and $\mathcal{S}_k^1$ those signatures x_i that satisfy $h_k(x_i) = 1$. We compute *offline* the following *query-independent* values (see Fig. 1):

$$\alpha_k^0 = \frac{1}{|\mathcal{S}_k^0|} \sum_{u \in \mathcal{S}_k^0} g_k(u), \quad (22)$$

$$\alpha_k^1 = \frac{1}{|\mathcal{S}_k^1|} \sum_{u \in \mathcal{S}_k^1} g_k(u). \quad (23)$$

Online, for a given query x, we first pre-compute and store in look-up tables the following *query-dependent* values:

$$\beta_k^0 = d(g_k(x), \alpha_k^0), \quad (24)$$
$$\beta_k^1 = d(g_k(x), \alpha_k^1). \quad (25)$$

By definition, we have:

$$d_E(x,y) = \sum_k \beta_k^{h_k(y)}. \quad (26)$$

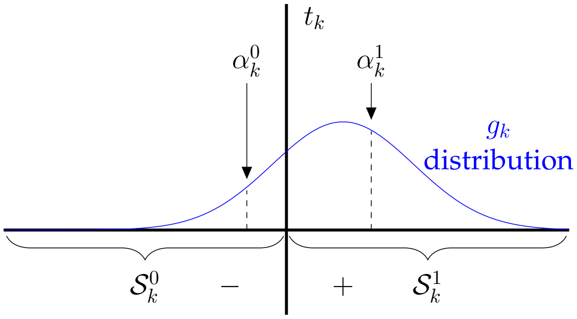


Figure 1. Expectation-based asymmetric distance

The cost of pre-computing the β values is negligible with respect to the cost of computing many $d_E(x,y)$'s for a large number of database signatures y. Note that the sum (26) can be efficiently computed, *e.g.* by grouping the dimensions in blocks of 8 bits and having one 256-dimensional look-up table per block (rather than one 2-dimensional look-up table per dimension). This reduces the number of summations as well as the number of accesses to the memory.

Instead of computing the distance between the query and the expected values of the dataset items as in Equation (21), one may find more intuitive to compute the expected distance between the query and the dataset items, *i.e.*,

$$d_{\hat{E}}(x,y) = \sum_k \mathbb{E}_{u\sim p}[d(g_k(x), g_k(u))|h_k(u)=h_k(y)]. \quad (27)$$

However, we did not notice any significant difference between this approach and d_E. Furthermore, we would like to note that computing $d_{\hat{E}}$ would require us to calculate and store more statistics than d_E. Particularly, we would need to compute the expected squared values, *i.e.*,

$$\hat{\alpha}_k^0 = \frac{1}{|\mathcal{S}_k^0|} \sum_{u \in \mathcal{S}_k^0} g_k^2(u),, \quad (28)$$

$$\hat{\alpha}_k^1 = \frac{1}{|\mathcal{S}_k^1|} \sum_{u \in \mathcal{S}_k^1} g_k^2(u)., \quad (29)$$

as well as α_k^0 and α_k^1.

3.2 Lower-Bound Based Asymmetric Distance

In [22], Dong *et al*. proposed an asymmetric algorithm for binary embeddings based on random projections. We now generalize this idea and show that a similar approach can be applied to a much wider range of binary embedding techniques. For the simplicity of the presentation, we assume that q_k has the form $q_k(u) = \sigma(u - t_k)$ where t_k is a threshold but this can be trivially generalized to other quantization functions.

The idea is to lower-bound the quantity (20) by bounding each of its terms. We note that t_k splits $\mathbb{R}$ into two half-lines and consider two cases:

- If $h_k(x) \neq h_k(y)$, *i.e.* $g_k(x)$ and $g_k(y)$ are on different sides of t_k, then a lower-bound between $g_k(x)$ and $g_k(y)$ is the distance between $g_k(x)$ and the threshold t_k, *i.e.* $d(g_k(x), g_k(y)) \geq d(g_k(x), t_k)$.
- If $h_k(x) = h_k(y)$, *i.e.* $g_k(x)$ and $g_k(y)$ are on the same half-line, then we have the following obvious lower-bound: $d(g_k(x), g_k(y)) \geq 0$ (actually, this bound is always true).

Merging the two cases in a single equation, we have the following lower-bound on $d(g(x), g(y))$:

$$d_{LB}(x,y) = \sum_k \bar{\delta}_{h_k(x),h_k(y)} d(g_k(x), t_k), \qquad (30)$$

where $\bar{\delta}_{i,j}$ is the negation of the Kronecker delta, *i.e.* $\bar{\delta}_{i,j} = 0$ if $i = j$ and 1 otherwise. We note that, in the case of LSH, equation (30) is equivalent to the asymmetric LSH distance proposed in [22].

In practice, for a given query signature x, we can pre-compute *online* the values (see Fig. 2):

$$\gamma_k^0 = \bar{\delta}_{h_k(x),0} d(g_k(x), t_k), \qquad (31)$$

$$\gamma_k^1 = \bar{\delta}_{h_k(x),1} d(g_k(x), t_k). \qquad (32)$$

We note that one of these two values is guaranteed to be 0 for each dimension k. We can subsequently pack these values by blocks of 8 dimensions for faster computation as was the case of the expectation-based approximation. By definition we have:

$$d_{LB}(x,y) = \sum_k \gamma_k^{h_k(y)}. \qquad (33)$$

A major difference between d_E and d_{LB} is that the former one makes use of the data distribution while the latter one does not. Despite its very crude nature, we will see that d_{LB} leads to excellent results on a variety of binary embedding algorithms.

3.3 Asymmetric Distances and Variance Preservation

As we will see in the experiments of Section 4, PCAE seems to benefit particularly from the asymmetric distances (see, *e.g.*, the results on CIFAR in Figures 3c and 4c). Remember from Section 2.2.1 that a disadvantage of PCAE is that each PCA dimension is assigned one bit, independently of the information it may carry. Especially, low-variance dimensions are given as much weight by the Hamming distance as high-variance dimensions. We now show that, in the case of PCAE, asymmetric distances in each dimension are proportional in expectation to the variance of the data, thus giving more weight to the first PCA dimensions. This may explain, at least partly, the significant improvements of asymetric distances for PCAE.

After PCA, the dimensions are uncorrelated and each dimension is supposed to have been generated by a Gaussian p_i (see top row of Figure 10) with zero mean and standard deviation σ_i. We assume the dimensions ordered such that $\sigma_i > \sigma_{i+1}$.

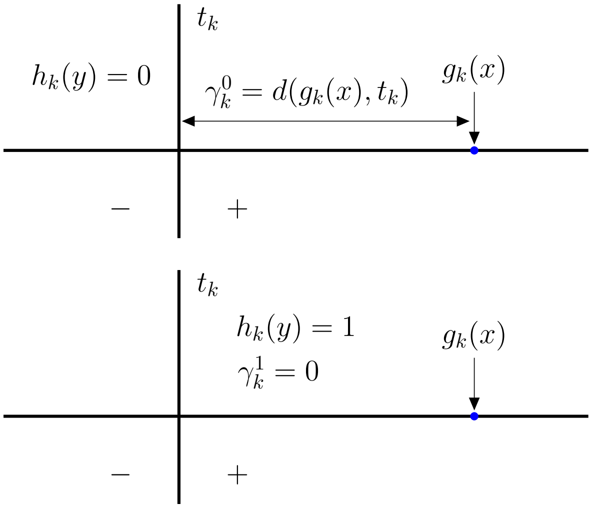


Figure 2. Lower-bound-based asymmetric distance. Top: case where $h_k(x) \neq h_k(y)$, and therefore $d(g_k(x), g_k(y)) \geq d(g_k(x), t_k)$. Bottom: case where $h_k(x) = h_k(y)$, and therefore $d(g_k(x), g_k(y)) \geq 0$.

a) d_E **case:** In a given dimension i, the expectation of the positive samples is the expectation of a half-normal distribution

$$E = \int_0^{\infty} x p_i(x) dx = \frac{\sigma_i}{\sqrt{2\pi}}$$

and the expectation of the negative samples is $-E = -\frac{\sigma_i}{\sqrt{2\pi}}$. These values correspond to the α's of equations (22) and (23). Consequently, the expectation of the distance to these values (i.e. the expectation of the β's of equations (24) and (25)) is:

$$\int_{-\infty}^{+\infty} (x - (\pm E))^2 p_i(x) dx = \sigma_i^2 (1 + \frac{1}{2\pi}). \qquad (34)$$

Therefore *in expectation* the contribution of a given dimension to the asymmetric distance is proportional to the variance in this dimension.

b) d_{LB} **case:** Let us consider the case where $h_k(x) = 1$ and $h_k(y) = 0$. The other relevant case can be treated analogously. The expectations of the γ's (*cf*. equations (31) and (32)) in dimension i are equal to

$$\int_{-\infty}^{0} 0\ p_i(x) dx + \int_0^{+\infty} x^2 p_i(x) dx = \frac{\sigma_i^2}{2}.$$

Again, *in expectation*, the contributions of the dimensions to the asymmetric distance decrease with the index i.

4 Experiments

We now show the benefits of the asymmetric distances proposed in section 3 on the binary embeddings we reviewed in section 2. We first describe in Section 4.1 the four public benchmarks we experiment on. We then provide in Section 4.2 implementation details for the different embedding algorithms. We finally report and discuss results in Section 4.3.

4.1 Datasets and Features

We run experiments on two category-level retrieval benchmarks, CIFAR and Caltech256, and on two instance-level retrieval benchmarks, the University of Kentucky Benchmark (UKB) and INRIA Holidays. This diverse selection of datasets allows us to experiment with different setups. On CIFAR, we evaluate both semantic retrieval and retrieval using Euclidean neighbors as groundtruth, and there are hundreds or even thousands of relevant items per query. On Caltech256, we evaluate the effect of different descriptors on the results: GIST, Bag of Words and Fisher Vector. UKB and Holidays are standard instance-level retrieval datasets. As opposed to CIFAR or Caltech256, the number of relevant items per query is much smaller, always 4 for UKB and from 2 to 10 for Holidays. We now describe these datasets as well as the associated standard experimental protocols in detail.

CIFAR. The CIFAR dataset [27] is a subset of the Tiny Images dataset [5]. We use the same version of CIFAR that was used in [21]. It contains 64,184 images of size 32×32 that have been manually grouped into 11 ground-truth classes. Images are described using a greyscale GIST descriptor [6] computed at 3 different scales (8, 8, 4) producing a 320-dimensional vector. 1,000 images are used as queries, 5,000 are used for unsupervised training purposes, and the remaining images are use as database images.

Similarly to [21], we report results on two different problems.

- Nearest neighbor retrieval: we discard the class labels. A nominal threshold of the average distance to the 50th nearest neighbor is used to determine whether a database image returned for a given query is considered a true positive. We note that the number of true positives varies widely from one query to another (from 0 to 2,353). We compute the Average Precision (AP) for each query (with a non-zero number of true positives) and report the mean over the 1,000 queries (Mean AP or MAP).
- Semantic retrieval: we use the class labels as groundtruth and report the precision at 1.

Caltech256. The Caltech256 dataset [28] contains approximately 30,000 images grouped in 257 classes. Through our experiments, we use only 256 classes and we discard the "clutter" class. As in CIFAR, we split the dataset in three different sets. We select 5 images per class (1,280 images in total) to serve as queries, and 5,000 random images to serve as unsupervised training data. The remaining images are used as the database. We describe the images and report precision at 1 using 3 different descriptors:

- GIST descriptor with 320 dimensions (same configuration as in CIFAR).
- Bag of Visual Words (BOV) with 1,024 vocabulary words using SIFT descriptors [29]. The low-level features are densely sampled and their dimensionality is reduced from 128 to 64 dimensions with PCA. Instead of k-means, we used a Gaussian Mixture Model to learn a probabilistic codebook and use soft assignment instead of hard assignment. The final histograms are L1-normalized and then square-rooted. As noted in [30], [31], this corresponds to an explicit embedding of the Bhattacharyya kernel when using the dot product as a similarity measure, and can yield very significant improvements at virtually zero cost.

  All the involved learning (PCA for the low-level features and vocabulary construction) is done using the 5,000 training images.
- Fisher Vector (FV) [9], which was recently shown to yield excellent results for object and scene retrieval [32], [11]. As is the case of BOV, low-level SIFT descriptors are densely sampled and their dimensionality is reduced from 128 to 64 dimensions. We use probabilistic codebooks (*i.e.* GMMs) with 64 Gaussians so that an image is described by a $64 \times 64 = 4,096$ dimensional FV. Again, the 5,000 training images are used to learn the low-level PCA and the vocabulary.

UKB. The University of Kentucky Benchmark (UKB) [33] contains 10,200 images of 2,550 objects (4 images per object). Each image is used in turn as query to search through the 10,200 images. The accuracy is measured in terms of the number of relevant images retrieved in the top 4, *i.e.* 4 × recall@4. We use the same low-level feature detection and description procedure as in [25] (Hessian-Affine extractor and SIFT descriptor) since the code is available online. As in Caltech256, we reduce the dimensionality of the SIFT features down to 64 dimensions and aggregate them into a single image-level descriptor of 4,096 dimensions using the FV framework. For learning purposes (*e.g.* to learn the PCA on the SIFT descriptors and the GMM for the FV), we use an additional set of 60,000 images (Flickr60K) made available by the authors of [25].

Holidays. The INRIA Holidays dataset [25] contains 1,491 images of 500 scenes and objects. The first image of each scene is used as query to search through the remaining 1,490 images. We measure the accuracy for each query using AP and report the MAP over the 500 queries. As was the case for UKB, images are described using 4,096 dimensional FVs, using the same low-level feature detection and description procedure, as well as the same Flickr learning set.

4.2 Implementation Details

To learn the parameters of the embedding functions and to compute offline the α values for d_E we need training data. For CIFAR and Caltech256, we will use the $5,000$ training samples, that are used neither as queries nor as database items. For UKB and Holidays we use the Flickr60K dataset [25].

For CIFAR and Caltech256, where no predefined partitions exist, we repeated the experiments 3 times using different queries and database partitions and averaged the results.

For LSH and LSBC, which perform binarization through random projections, as well as for PCAE-RR and PCAE-ITQ, which use random rotations, experiments are repeated 5 times with 5 different projection matrices and we report the average results.

As discussed in section 2.1.1, mean-centering the data can impact LSH very positively. Therefore, we have mean-centered the GIST and BOV descriptors for CIFAR and Caltech256, learning the means on their respective training sets. This centering has no significant impact for FVs since, by definition, they are already (approximately) zero-mean. We note that centering the data on the origin does not impact PCAE, PCAE-RR, PCAE-ITQ, and SH (which perform a PCA of the signatures) or LSBC (which is shift-invariant).

4.3 Results and Analysis

We report results on the four datasets in Figures 4 - 9 with the symmetric Hamming distance as well as with the proposed asymmetric distances d_E and d_{LB}. The following is a detailed discussion of our findings.

Asymmetric vs symmetric. The use of asymmetric distances consistently improves the results over the symmetric Hamming distance, independently of the dataset, of the descriptor used, and of the binary embedding technique. In general, the gain in accuracy is impressive both in terms of absolute and relative improvement. Here are just two examples: on the CIFAR dataset with semantic labels (Figure 4), when using PCAE, we can observe an improvement of 8% absolute and 22% relative at 128 bits. On Holidays, When using SH, we can observe an improvement of 8% absolute and 21% relative (Figure 9), also at 128 bits.

The GIST descriptor results on Caltech256 (Figure 5) are an exception to this rule. Indeed, in this setting the simpler Hamming distance can outperform the proposed asymmetric distances, see especially the PCAE results of Figure 5c . This seems to indicate that the Euclidean distance we are trying to approximate in the PCA projected space is suboptimal (at least on this dataset and with these features) and that the Hamming distance in the projected space approximates a better metric.

Expectation vs lower-bound. For almost all embedding techniques, d_E and d_{LB} yield very similar results, which is somewhat surprising given that the two approximations are very different in nature. The slight advantage of d_E over d_{LB} comes from the fact that the former approach uses information about the data distribution (through the pre-computed values α) while the latter does not.

We note, however, two exceptions. The first is LSBC, for which in most cases d_E performs significantly better than d_{LB} (see Figures 3b, 4b, 6b, 7b, 8b, 9b). We are still investigating this difference but we observed that the distributions of the values $g_k(x)$ in the intermediate real-valued space for LSBC are significantly different from those observed for the other embedding methods (typically U- or half-U-shaped for LSBC, as opposed to Gaussian-shaped for the others, particularly PCAE, see Figure 10). The second exception is on the Caltech256 with GIST descriptors on Figure 5, where d_{LB} usually and sometimes very clearly outperforms d_E. As we noted in the previous point, the Euclidean may not be the best measure for GIST descriptors.

Finally, preliminary experiments on fusing d_E and d_{LB} yielded only marginal improvements.

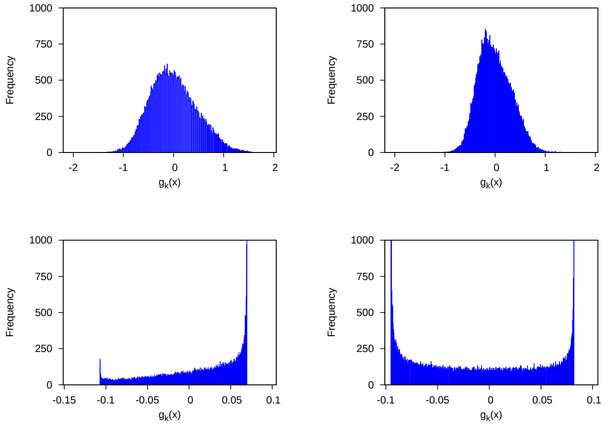
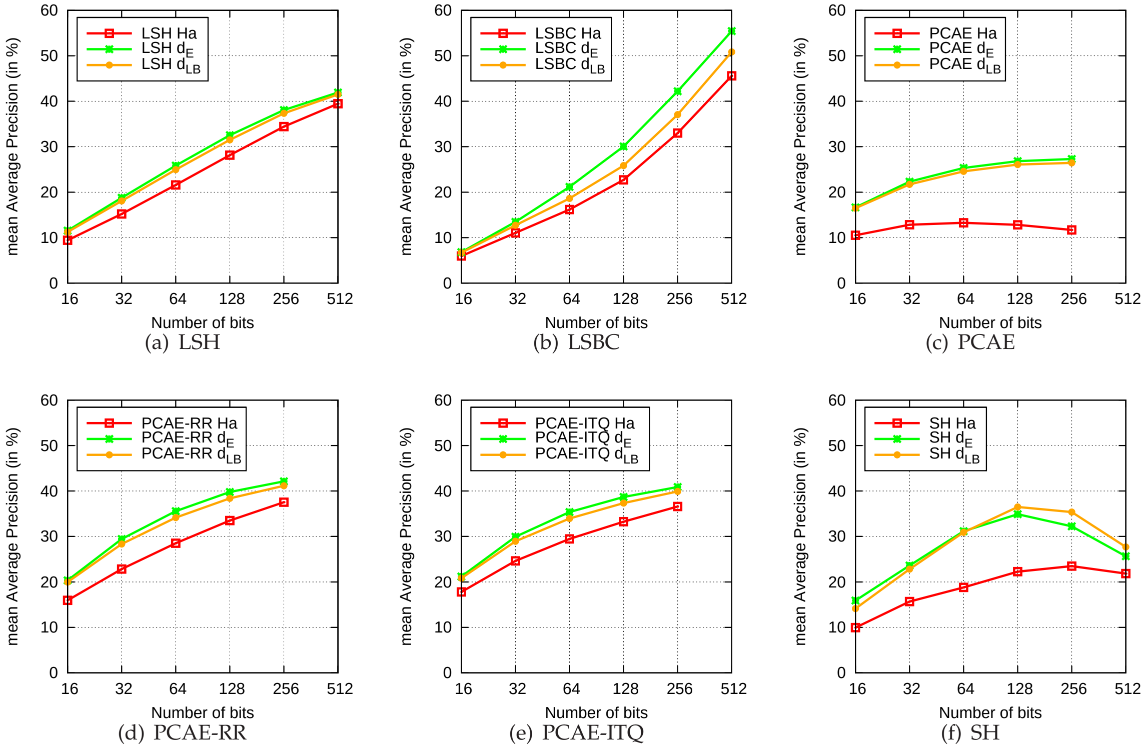


Figure 10. Histograms of the projected values of the 60,000 CIFAR images. Top: projected on the first two PCA dimensions. Bottom: projected with two random LSBC dimensions.

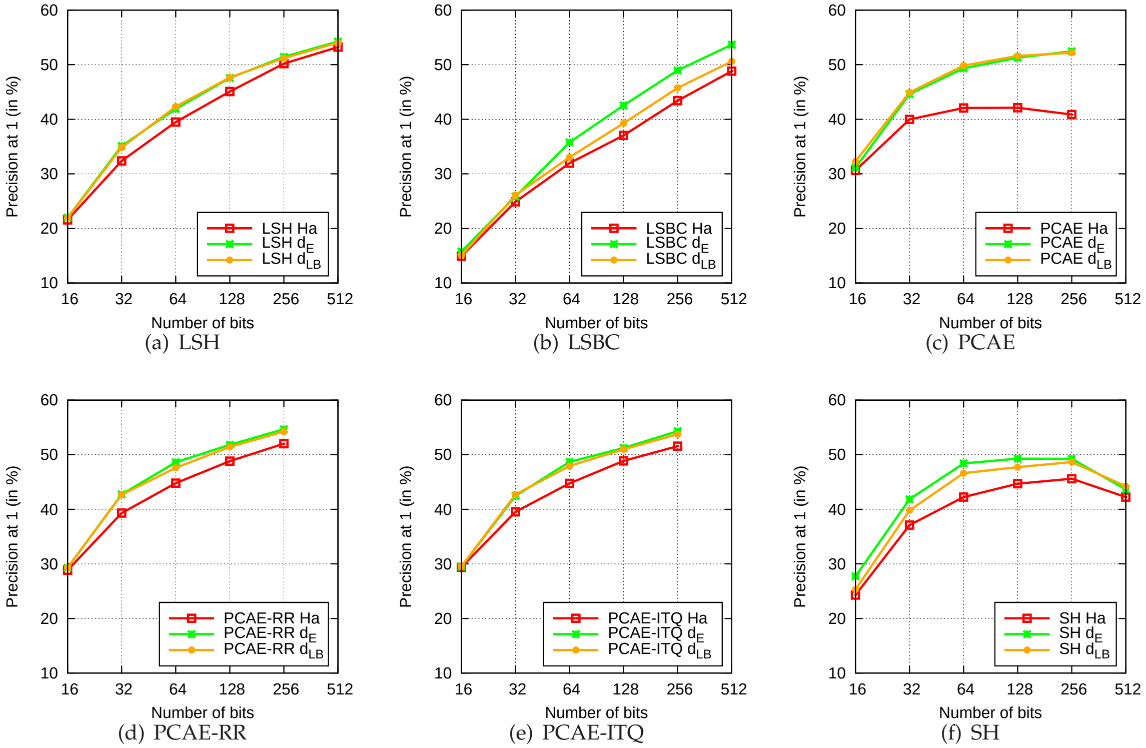
Influence of the retrieval problem. In Figures 3 and 4 we report results on the CIFAR dataset for two different problems: retrieval of Euclidean neighbors and semantic retrieval. We can observe how asymmetric distances provide similar improvements for both cases. We can also note how, in the Euclidean problem, PCAE with Hamming distance seems to perform poorly (Fig. 3c) compared to the results of PCAE on other datasets, and also how the asymmetric improvements seem larger in this case. We believe the difference stems not from the problem (semantic vs Euclidean retrieval) but because of the evaluation measure: this is the only experiment that combines,



(a) LSH (b) LSBC (c) PCAE

(d) PCAE-RR (e) PCAE-ITQ (f) SH

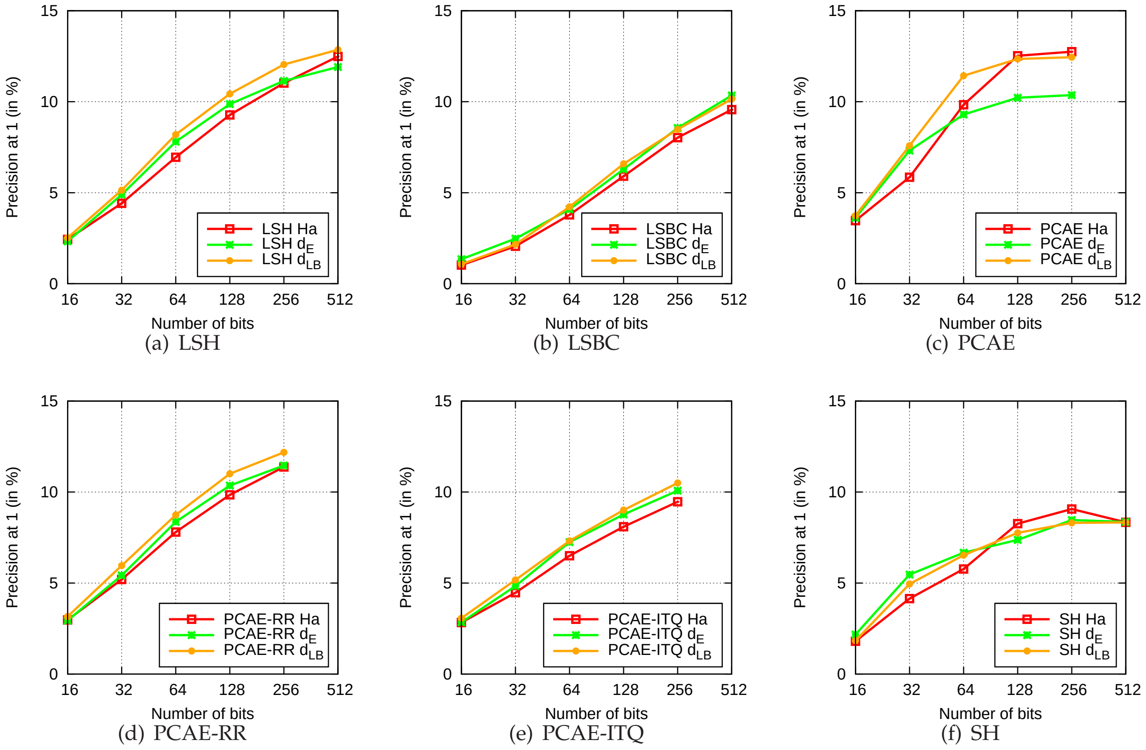
Figure 3. Influence of the asymmetric distances on the CIFAR dataset with Euclidean neighbors. The dimensionality of PCAE, PCAE-RR and PCAE-ITQ is limited by the dimensionality of the original GIST descriptor, 320 dimensions.



(a) LSH (b) LSBC (c) PCAE

(d) PCAE-RR (e) PCAE-ITQ (f) SH

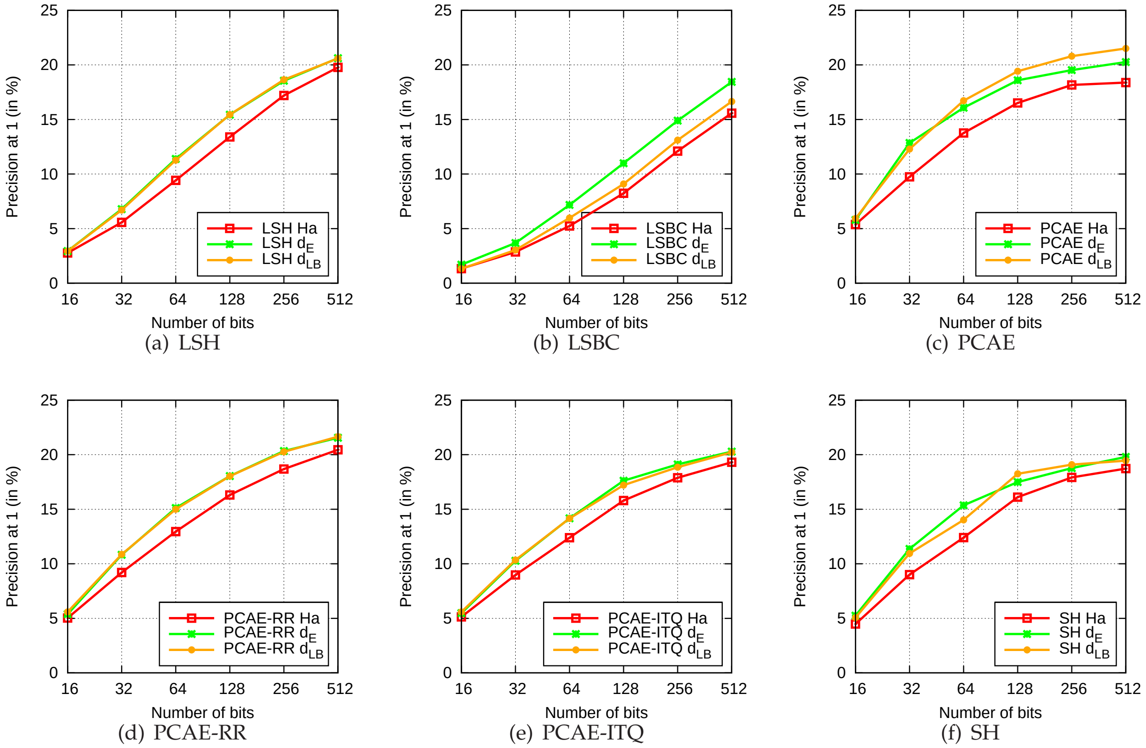
Figure 4. Influence of the asymmetric distances on the CIFAR dataset with semantic labels for 6 different encoding methods: LSH, LSBC, PCAE, PCAE-RR, PCAE-ITQ, and SH. The dimensionality of PCAE, PCAE-RR and PCAE-ITQ is limited by the dimensionality of the original GIST descriptor, 320 dimensions.



(a) LSH (b) LSBC (c) PCAE

(d) PCAE-RR (e) PCAE-ITQ (f) SH

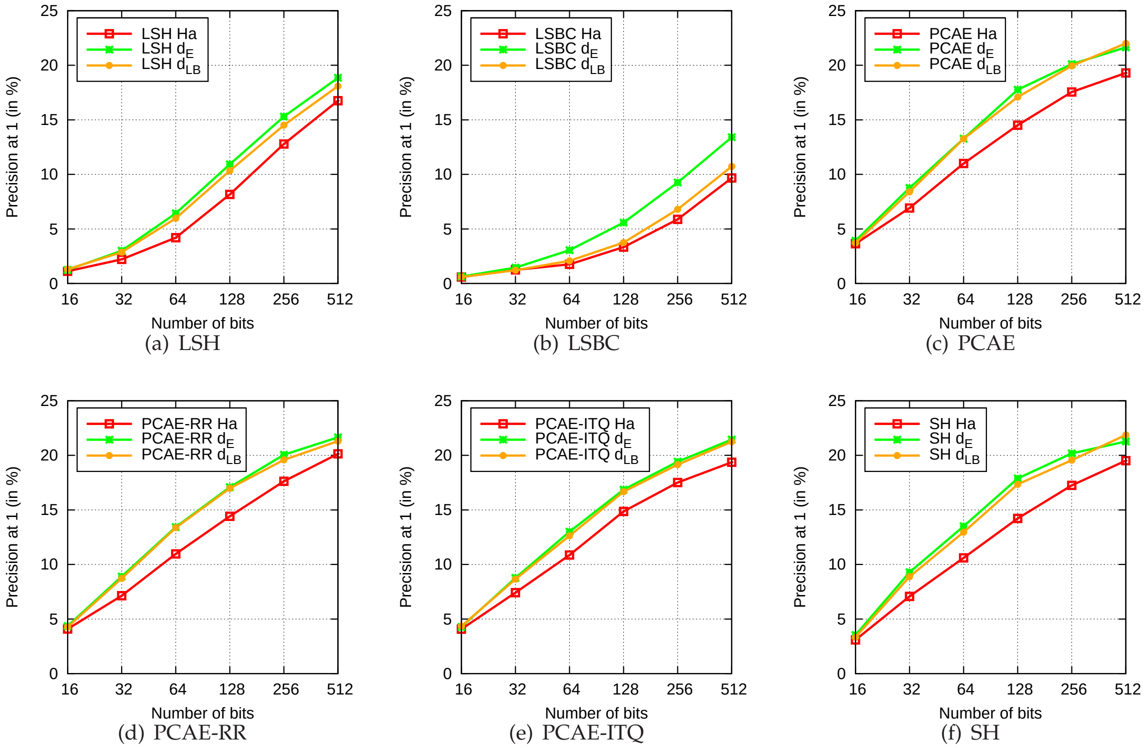
Figure 5. Influence of the asymmetric distances on the CALTECH256 dataset with GIST descriptors.



(a) LSH (b) LSBC (c) PCAE

(d) PCAE-RR (e) PCAE-ITQ (f) SH

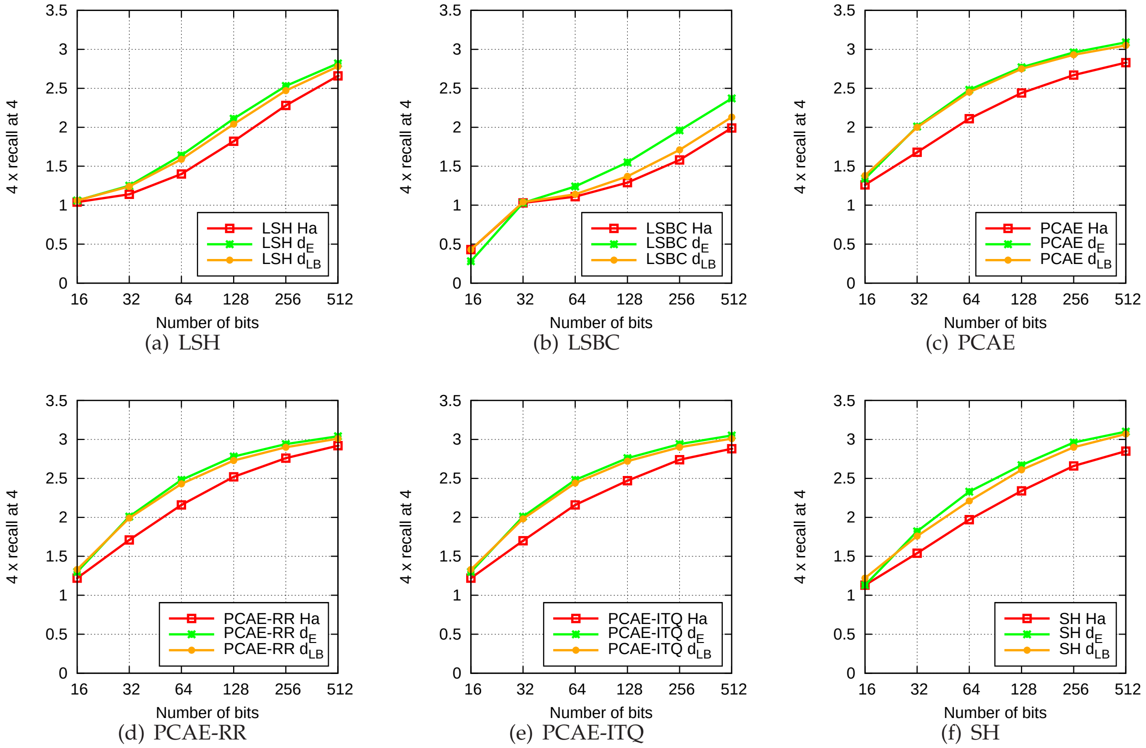
Figure 6. Influence of the asymmetric distances on the CALTECH256 dataset using BOV descriptors.



(a) LSH (b) LSBC (c) PCAE

(d) PCAE-RR (e) PCAE-ITQ (f) SH

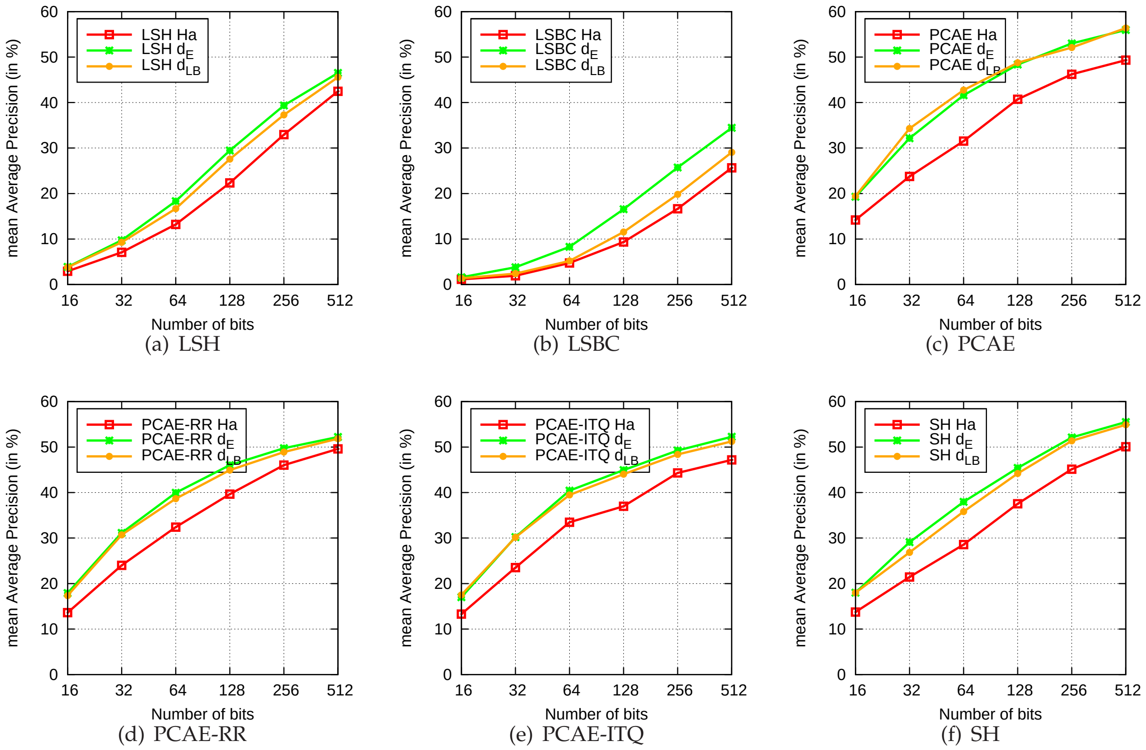
Figure 7. Influence of the asymmetric distances on the CALTECH256 dataset using FV descriptors.



(a) LSH (b) LSBC (c) PCAE

(d) PCAE-RR (e) PCAE-ITQ (f) SH

Figure 8. Influence of the asymmetric distances on the UKB dataset.



(a) LSH (b) LSBC (c) PCAE

(d) PCAE-RR (e) PCAE-ITQ (f) SH

Figure 9. Influence of the asymmetric distances on the Holidays dataset.

at the same time, a large number of relevant items per query and a global measure such as mAP. In such a case, balancing the data with PCAE-RR or PCAE-ITQ seems to yield a large benefit. To attest this, we experimented on Caltech256, which has many relevant items per query. Using BOV descriptors, we computed the mAP score instead of the precision at 1 reported in Figure 6c. In that case, PCAE results drastically dropped below those of PCAE-RR and PCAE-ITQ, supporting this idea.

Influence of the descriptor. In Figures 5 to 7 we can observe the influence of the descriptors on the Caltech256 dataset. In general, the improvements in BOV and FV are larger than the improvements obtained with GIST, showing that the improvement can be dependent on the feature type, particularly if the Euclidean distance was not a good measure in the original space.

We can also observe how, particularly in the PCA-based methods, FV has a slight edge over BOV when aiming at 256 bits or more. However, BOV can obtain better results than the FV when aiming at signatures of 128 bits or less. This is in line with the observations of [11], where they notice that, when producing small codes, it is usually better to start with a smaller image signature. In our case, the BOV has 1,024 dimensions and the FV has 4,096. When we can afford larger codes, the FV usually still outperforms BOV: the uncompressed BOV baseline is 22.11%, while the uncompressed FV baseline is 24.11%.

Influence of the embedding method. All methods benefit significantly from the asymmetric distances. This can be easily understood: since we are not binarizing the query, there is less loss of information on the query side.

PCAE seems to benefit particularly from the asymmetric distances (see, *e.g.*, the results on CIFAR in Figures 3c and 4c). This may be explained by the variance-preservation effect of the asymmetric distances (see Section 3.3). The variance problem of the other methods is not so severe: LSH and LSBC use random projections, and their variances are balanced in expectation. PCAE-RR, PCAE-ITQ, and SH all balance the variances, either explicitly as in PCAE-RR and PCAE-ITQ, or implicitly, as in SH, assigning more bits to the more important dimensions. Therefore, the impact of the asymmetric distances on these methods is not as pronounced, yet still sizeable. For example, on Caltech256 with FV (Figure 7), we show improvements on LSBC of about 4% absolute but almost 40% relative.

Asymmetric distances also seem to bridge the gap between the binary encoding methods, particularly between those based on PCA. Figure 11 compares the different encoding methods using Hamming distances and both d_E and d_{LB} on the CIFAR semantic problem with the same data we used for Figure 4. The results suggest that asymmetric distances can be used to compensate for the quality of the embedding method; the difference between the encoding methods is significant when using the Hamming distance (more than 10% absolute at 256 bits), but much less pronounced when using asymmetric distances (less than 5% absolute, again at 256 bits).

Qualitative results. Finally, Figure 12 shows qualitative results. We show the top 5 ranked images for four random queries of the Holidays dataset using PCAE with 128 bits, both for Hamming and for asymmetric distances. The false positives have been framed in red. We can observe how, in general, asymmetric distances obtain better and more consistent results than the Hamming distance. We believe that the asymmetric distances allow for better generalization. Figures 12b and 12c would be an example of this. However, this generalization capability can be pernicious when looking for almost-exact duplicates as in Figure 12d, where the Hamming distance retrieves better images.

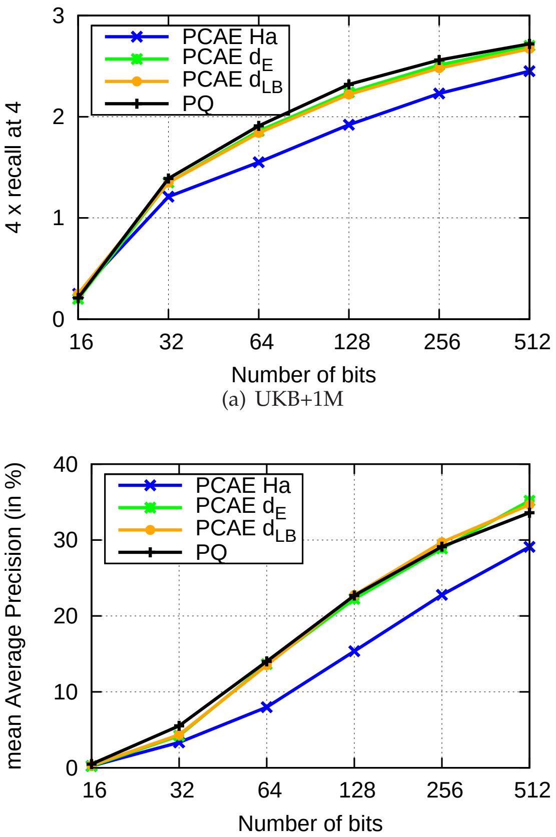
4.4 Large-Scale Experiments

We now show that the good results achieved with asymmetric distances scale to large datasets. For these experiments we merge Holidays and UKB with a set of 1M Flickr distractors made available by the authors of [25]. We refer to these combined datasets as Holidays+1M and UKB+1M. In both cases we use the original queries, 500 in Holidays and 10,200 in UKB. We experiment with PCAE, since it is the method that obtained, in general, the best results, and it is simpler than PCAE-RR and PCAE-ITQ.

Figure 13 shows the results on both datasets as a function of the number of bits. We compare them with Product quantization (PQ) [23] [11], since, to the best of our knowledge, these are the best results reported on Holidays+1M for very small operating points. For PQ, we employ the same pipeline as [11] which is composed of the following steps: i) PCA compression of the signatures, ii) random orthogonal rotation of the PCA projected signatures, iii) Product quantization and iv) comparison using PQ's asymmetric distances, referred to as ADC in [11].

To fix the number of dimensions D' in the PCA projection step of PQ, the authors minimized the mean square error of the projection and the quantization over a training set. We follow a different heuristic: we set D' to be equal to the number of output bits we are aiming at, and assign 8 dimensions to each subquantizer. We then fix the number of bits per subquantizer to 8, since this seems to be a standard choice that usually offers excellent results. Experimentally, we observed this heuristic to obtain comparable or better results than minimizing the mean square error, and in most cases was the best possible configuration. As was the case before, experiments are repeated 5 times with different projection matrices and the results are averaged.

We can observe how both asymmetric distances with PCAE perform comparably to PQ on both datasets although PCAE is simpler than PQ both from a conceptual and an engineering standpoint. Furthermore, as opposed to PQ, the lower-bound asymmetric distance does not require any training.



(a) UKB+1M

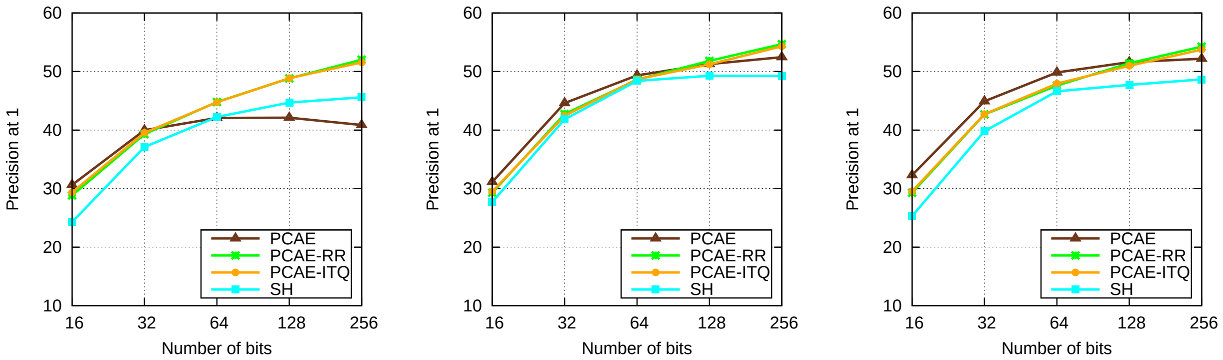
(b) Holidays+1M

Figure 13. Comparison of the proposed asymmetric distances and PQ [11] on UKB+1M (top) and Holidays+1M (bottom). MAP and 4 × recall at 4 as a function of the number of bits.

5 Conclusions and Future Work

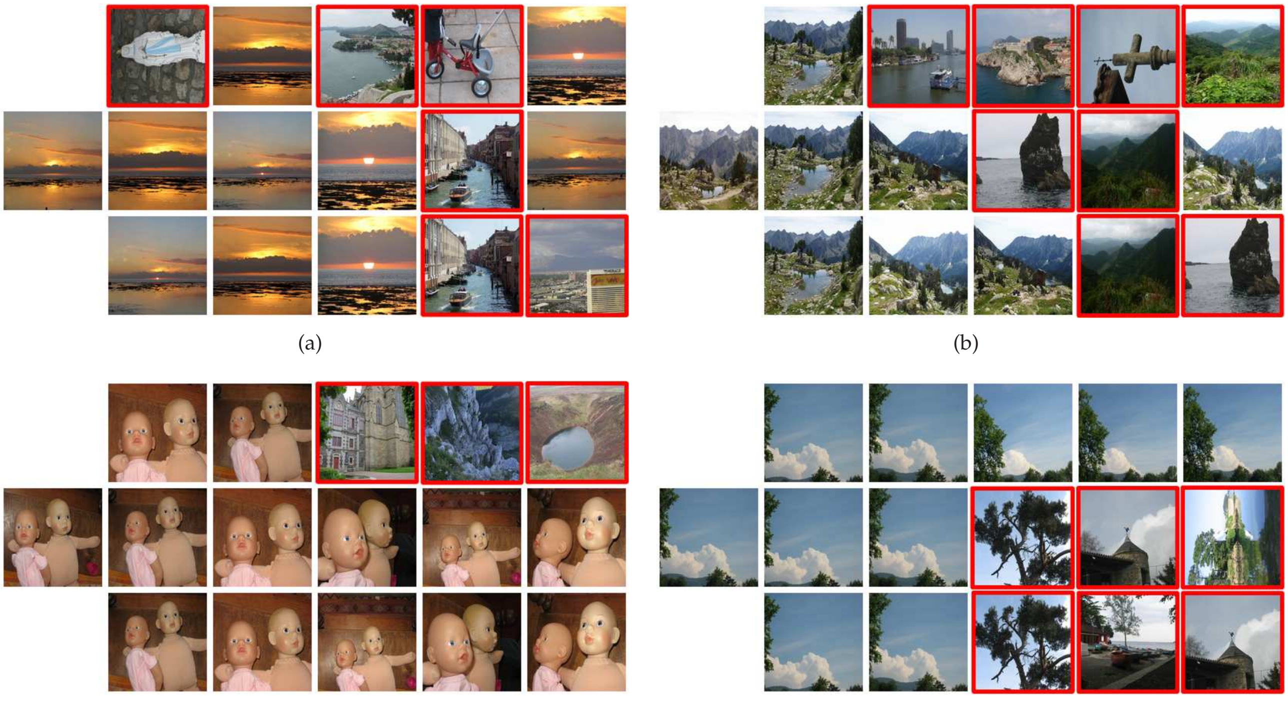
In this work, we proposed two asymmetric distances for binary embedding techniques, *i.e.* distances between binarized and non-binarized signatures. We showed their applicability to several embedding algorithms: LSH, LSBC, SH, PCAE, PCAE-RR, and PCAE-ITQ. We demonstrated on four datasets with up to 1M images that the proposed asymmetric distances consistently, and often very significantly, improve the retrieval accuracy over the symmetric Hamming distance. We also showed how this asymmetric distances can achieve results comparable to state-of-the-art methods such as PQ, while being conceptually much simpler. The lower-bound asymmetric distance can also be applied on datasets that have already been binarized, with no need to perform any reencoding or extra training.

In future work, we would be interested in investigating coding techniques which would be designed



(a) Hamming distance (b) Asymmetric d_E distance (c) Asymmetric d_{LB} distance

Figure 11. Comparison of Hamming and asymmetric distances on the CIFAR dataset with semantic labels. Same data as in Figure 4.



(a) (b) (c) (d)

Figure 12. Top five results of four random queries of Holidays using codes of 128 bits. First row: PCAE + Hamming. Middle row: PCAE + asymmetric d_E distance. Bottom row: PCAE + asymmetric d_{LB} distance.

with the asymmetric distances in mind. One possible example which was inspired to us by ITQ would be to learn a rotation matrix which does not minimize the quantization error between the signatures and the binary codes, but between the signatures and the *reconstructed* version of the binary codes using the expected values.

Also, in the asymmetric expectation-based approach, we currently learn the α coefficients (Equations (22) and (23)) which minimize a reconstruction error on the training data. It would be interesting to understand whether we could learn these coefficients with supervised data to optimize directly the retrieval objective function.

References

[1] A. Gordo and F. Perronnin, "Asymmetric distances for binary embeddings," in *CVPR*, 2011.

[2] M. Everingham, L. V. Gool, C. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *IJCV*, 2010.

[3] L. Fei-Fei, R. Fergus, and P. Perona, "One-shot learning of object categories," *IEEE PAMI*, vol. 28, no. 4, 2006.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009.

[5] A. Torralba, R. Fergus, and W. Freeman, "80 million tiny images: a large dataset for non-parametric object and scene recognition," *IEEE PAMI*, 2008.

[6] A. Oliva and A. Torralba, "Modeling the shape of the scene: a holistic representation of the spatial envelope," *IJCV*, 2001.

[7] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *ICCV*, 2003.

[8] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, “Visual categorization with bags of keypoints,” in *ECCV SLCV Workshop*, 2004.

[9] F. Perronnin and C. Dance, “Fisher kernels on visual vocabularies for image categorization,” in *CVPR*, 2007.

[10] F. Perronnin, J. Sánchez, and T. Mensink, “Improving the Fisher kernel for large-scale image classification,” in *ECCV*, 2010.

[11] H. Jégou, M. Douze, C. Schmid, and P. Pérez, “Aggregating local descriptors into a compact image representation,” in *CVPR*, 2010.

[12] P. Indyk and R. Motwani, “Approximate nearest neighbors: towards removing the curse of dimensionality,” in *STOC*, 1998.

[13] M. Charikar, “Similarity estimation techniques from rounding algorithms,” in *STOC*, 2002.

[14] M. Raginsky and S. Lazebnik, “Locality-sensitive binary codes from shift-invariant kernels,” in *NIPS*, 2009.

[15] Y. Weiss, A. Torralba, and R. Fergus, “Spectral hashing,” in *NIPS*, 2008.

[16] B. Kulis and K. Grauman, “Kernelized locality-sensitive hashing for scalable image search,” in *ICCV*, 2009.

[17] J. Brandt, “Transform coding for fast approximate nearest neighbor search in high dimensions,” in *CVPR*, 2010.

[18] J. Wang, S. Kumar, and S.-F. Chang, “Semi-supervised hashing for large scale search,” in *CVPR*, 2010.

[19] L. Torresani, M. Szummer, and A. Fitzgibbon, “Efficient object category recognition using classemes,” in *ECCV*, 2010.

[20] A. Bergamo, L. Torresani, and A. Fitzgibbon, “Picodes: Learning a compact code for novel-category recognition,” in *NIPS*, 2011.

[21] Y. Gong and S. Lazebnik, “Iterative quantization: A procrustean approach to learning binary codes,” in *CVPR*, 2011.

[22] W. Dong, M. Charikar, and K. Li, “Asymmetric distance estimation with sketches for similarity search in high-dimensional spaces,” in *SIGIR*, 2008.

[23] H. Jégou, M. Douze, and C. Schmid, “Product quantization for nearest neighbor search,” *IEEE TPAMI*, 2010.

[24] A. Rahimi and B. Recht, “Random features for large-scale kernel machines,” in *NIPS*, 2007.

[25] H. Jégou, M. Douze, and C. Schmid, “Hamming embedding and weak geometric consistency for large scale image search,” in *ECCV*, 2008. http://lear.inrialpes.fr/˜jegou/data.php.

[26] P. Schonemann, “A generalized solution of the orthogonal procrustres problem,” *Psychometrika*, vol. 31, 1966.

[27] A. Krizhevsky, “Learning multiple layers of features from tiny images,” tech. rep., University of Toronto, 2009.

[28] G. Griffin, A. Holub, and P. Perona, “Caltech-256 object category dataset,” tech. rep., California Institute of Technology, 2007.

[29] D. Lowe, “Distinctive image features from scale-invariant keypoints,” *IJCV*, 2004.

[30] F. Perronnin, J. Sánchez, and Y. Liu, “Large-scale image categorization with explicit data embedding,” in *CVPR*, 2010.

[31] A. Vedaldi and A. Zisserman, “Efficient additive kernels via explicit feature maps,” in *CVPR*, 2010.

[32] F. Perronnin, Y. Liu, J. Sánchez, and H. Poirier, “Large-scale image retrieval with compressed fisher vectors,” in *CVPR*, 2010.

[33] D. Nistér and H. Stewénius, “Scalable recognition with a vocabulary tree,” in *CVPR*, 2006.