

Fast Exact Search in Hamming Space with Multi-Index Hashing

Mohammad Norouzi, Ali Punjani, David J. Fleet,

Abstract—There is growing interest in representing image data and feature descriptors using compact binary codes for fast near neighbor search. Although binary codes are motivated by their use as direct indices (addresses) into a hash table, codes longer than 32 bits are not being used as such, as it was thought to be ineffective. We introduce a rigorous way to build multiple hash tables on binary code substrings that enables exact k -nearest neighbor search in Hamming space. The approach is storage efficient and straightforward to implement. Theoretical analysis shows that the algorithm exhibits sub-linear run-time behavior for uniformly distributed codes. Empirical results show dramatic speedups over a linear scan baseline for datasets of up to one billion codes of 64, 128, or 256 bits.

Index Terms—Binary codes, Hamming distance, nearest neighbor search, multi index hashing, large-scale image retrieval.

1 INTRODUCTION

There has been growing interest in representing image data and feature descriptors in terms of compact binary codes, often to facilitate fast near neighbor search and feature matching in vision applications (*e.g.*, [2], [7], [32], [33], [35], [19]). Binary codes are storage efficient and comparisons require just a small number of machine instructions. Millions of binary codes can be compared to a query in less than a second. But the most compelling reason for binary codes, and discrete codes in general, is their use as direct indices (addresses) into a hash table, yielding a dramatic increase in search speed compared to an exhaustive linear scan (*e.g.*, [38], [31], [25]).

Nevertheless, using binary codes as direct hash indices is not necessarily efficient. To find near neighbors one needs to examine all hash table entries (or *buckets*) within some Hamming ball around the query. The problem is that the number of such buckets grows near-exponentially with the search radius. Even with a small search radius, the number of buckets to examine is often larger than the number of items in the database, and hence slower than linear scan. Recent papers on binary codes mention the use of hash tables, but resort to linear scan when codes are longer than 32 bits (*e.g.*, [35], [31], [20], [25]). Not surprisingly, code lengths are often significantly longer than 32 bits in order to achieve satisfactory retrieval performance (*e.g.*, see Fig. 5).

This paper presents a new algorithm for exact k -nearest neighbor (k NN) search on binary codes that

is dramatically faster than exhaustive linear scan. This has been an open problem since the introduction of hashing techniques with binary codes. Our new multi-index hashing algorithm exhibits sub-linear search times, is storage efficient, and straightforward to implement. Empirically, on databases of up to 1B codes we find that multi-index hashing is hundreds of times faster than linear scan. Extrapolation suggests that the speedup gain grows quickly with database size beyond 1B codes.

1.1 Background: Problem and Related Work

Nearest neighbor (NN) search on binary codes is used for image search [29], [35], [38], matching local features [2], [7], [16], [33], image classification [6], object segmentation [19], and parameter estimation [32]. Sometimes the binary codes are generated directly as feature descriptors for images or image patches, such as BRIEF or FREAK [7], [6], [2], [36], and sometimes binary corpora are generated by discrete similarity-preserving mappings from high-dimensional data. Most such mappings are designed to preserve Euclidean distance (*e.g.*, [11], [20], [29], [33], [38]). Others focus on semantic similarity (*e.g.*, [25], [32], [31], [35], [26], [30], [21]). Our concern in this paper is not the algorithm used to generate the codes, but rather with fast search in Hamming space.¹

We address two related search problems in Hamming space. Given a dataset of binary codes, $\mathcal{H} \equiv \{\mathbf{h}_i\}_{i=1}^n$, the first problem is to find the k codes in \mathcal{H} that are closest in Hamming distance to a given query, *i.e.*, k NN search in Hamming distance. The 1-NN problem in Hamming space was called the *Best Match* problem by Minsky and Papert [23]. They observed that there are no obvious approaches significantly better than exhaustive search, and asked whether such approaches might exist.

1. There do exist several other promising approaches to fast approximate NN search on large real-valued image features (*e.g.*, [3], [17], [27], [24], [5]). Nevertheless, we restrict our attention in this paper to compact binary codes and exact search.

- M. Norouzi is with the Department of Computer Science, University of Toronto.
E-mail: norouzi@cs.toronto.edu
- A. Punjani is with the Department of Computer Science, University of Toronto.
E-mail: alipunjani@cs.toronto.edu
- D.J. Fleet is with the Department of Computer Science, University of Toronto.
E-mail: fleet@cs.toronto.edu

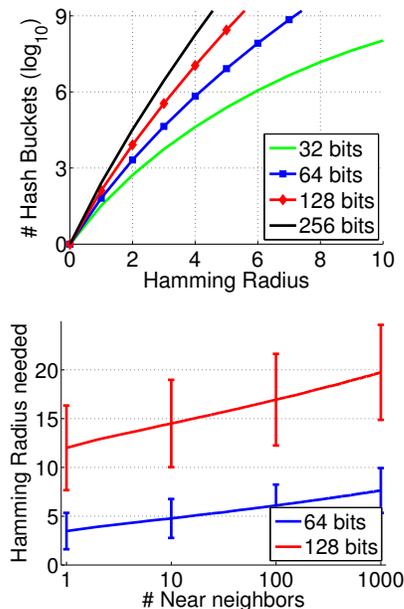


Fig. 1. (Top) Curves show the (\log_{10}) number of distinct hash table indices (buckets) within a Hamming ball of radius r , for different code lengths. With 64-bit codes there are about 1B buckets within a Hamming ball with a 7-bit radius. Hence with fewer than 1B database items, and a search radius of 7 or more, a hash table would be less efficient than linear scan. Using hash tables with 128-bit codes is prohibitive for radii larger than 6. (Bottom) This plot shows the expected search radius required for k -NN search as a function of k , based on a dataset of 1B SIFT descriptors. Binary codes with 64 and 128 bits were obtained by random projections (LSH) from the SIFT descriptors [18]. Standard deviation bars help show that large search radii are often required.

The second problem is to find all codes in a dataset \mathcal{H} that are within a fixed Hamming distance of a query, sometimes called the *Approximate Query* problem [13], or *Point Location in Equal Balls* (PLEB) [15]. A binary code is an r -neighbor of a query code, denoted \mathbf{g} , if it differs from \mathbf{g} in r bits or less. We define the r -neighbor search problem as: find all r -neighbors of a query \mathbf{g} from \mathcal{H} .

One way to tackle r -neighbor search is to use a hash table populated with the binary codes $\mathbf{h} \in \mathcal{H}$, and examine all hash buckets whose indices are within r bits of a query \mathbf{g} (e.g., [35]). For binary codes of q bits, the number of distinct hash buckets to examine is

$$L(q, r) = \sum_{z=0}^r \binom{q}{z}. \quad (1)$$

As shown in Fig. 1 (top), $L(q, r)$ grows very rapidly with r . Thus, this approach is only practical for small radii or short code lengths. Some vision applications restrict search to exact matches (i.e., $r = 0$) or a small search radius (e.g., [14], [37]), but in most cases of interest the desired search radius is larger than is currently feasible (e.g., see Fig. 1 (bottom)).

Our work is inspired in part by the multi-index hashing results of Greene, Parnas, and Yao [13]. Building on the classical Turan problem for hypergraphs, they construct a set of over-lapping binary substrings such that any two codes that differ by at most r bits are guaranteed to be identical in at least one of the constructed substrings. Accordingly, they propose an exact method for finding all r -neighbors of a query using multiple hash tables, one for each substring. At query time, candidate r -neighbors are found by using query substrings as indices into their corresponding hash tables. As explained below, while run-time efficient, the main drawback of their approach is the prohibitive storage required for the requisite number of hash tables. By comparison, the method we propose requires much less storage, and is only marginally slower in search performance.

While we focus on exact search, there also exist algorithms for finding *approximate* r -neighbors (ϵ -PLEB), or approximate nearest neighbors (ϵ -NN) in Hamming distance. One example is Hamming Locality Sensitive Hashing [15], [10], which aims to solve the (r, ϵ) -neighbors decision problem: determine whether there exists a binary code $\mathbf{h} \in \mathcal{H}$ such that $\|\mathbf{h} - \mathbf{g}\|_H \leq r$, or whether all codes in \mathcal{H} differ from \mathbf{g} in $(1 + \epsilon)r$ bits or more. Approximate methods are interesting, and the approach below could be made faster by allowing misses. Nonetheless, this paper will focus on the *exact* search problem.

This paper proposes a data-structure that applies to both k -NN and r -neighbor search in Hamming space. We prove that for uniformly distributed binary codes of q bits, and a search radius of r bits when r/q is small, our query time is sub-linear in the size of dataset. We also demonstrate impressive performance on real-world datasets. To our knowledge this is the first practical data-structure solving exact k -NN in Hamming distance.

Section 2 describes a multi-index hashing algorithm for r -neighbor search in Hamming space, followed by run-time and memory analysis in Section 3. Section 4 describes our algorithm for k -nearest neighbor search, and Section 5 reports results on empirical datasets.

2 MULTI-INDEX HASHING

Our approach is a form of multi-index hashing. Binary codes from the database are indexed m times into m different hash tables, based on m disjoint binary substrings. Given a query code, entries that fall *close* to the query in at least one such substring are considered *neighbor candidates*. Candidates are then checked for validity using the entire binary code, to remove any non- r -neighbors. To be practical for large-scale datasets, the substrings must be chosen so that the set of candidates is small, and storage requirements are reasonable. We also require that all true neighbors will be found.

The key idea here stems from the fact that, with n binary codes of q bits, the vast majority of the 2^q possible

buckets in a full hash table will be empty, since $2^q \gg n$. It seems expensive to examine all $L(q, r)$ buckets within r bits of a query, since most of them contain no items. Instead, we merge many buckets together (most of which are empty) by marginalizing over different dimensions of the Hamming space. We do this by creating hash tables on *substrings* of the binary codes. The distribution of the code substring comprising the first s bits is the outcome of marginalizing the distribution of binary codes over the last $q - s$ bits. As such, a given bucket of the substring hash table includes all codes with the same first s bits, but having any of the $2^{(q-s)}$ values for the remaining $q - s$ bits. Unfortunately these larger buckets are not restricted to the Hamming volume of interest around the query. Hence not all items in the merged buckets are r -neighbors of the query, so we then need to cull any candidate that is not a true r -neighbor.

2.1 Substring Search Radii

In more detail, each binary code \mathbf{h} , comprising q bits, is partitioned into m disjoint substrings, $\mathbf{h}^{(1)}, \dots, \mathbf{h}^{(m)}$, each of length $\lfloor q/m \rfloor$ or $\lceil q/m \rceil$ bits. For convenience in what follows, we assume that q is divisible² by m , and that the substrings comprise contiguous bits. The key idea rests on the following statement: When two binary codes \mathbf{h} and \mathbf{g} differ by r bits or less, then, in at least one of their m substrings they must differ by at most $\lfloor r/m \rfloor$ bits. This leads to the first proposition:

Proposition 1: If $\|\mathbf{h} - \mathbf{g}\|_H \leq r$, where $\|\mathbf{h} - \mathbf{g}\|_H$ denotes the Hamming distance between \mathbf{h} and \mathbf{g} , then

$$\exists 1 \leq z \leq m \quad \text{s.t.} \quad \|\mathbf{h}^{(z)} - \mathbf{g}^{(z)}\|_H \leq r', \quad (2)$$

where $r' = \lfloor r/m \rfloor$.

Proof of Proposition 1 follows straightforwardly from the Pigeonhole Principle. That is, suppose that the Hamming distance between each of the m substrings is strictly greater than r' . Then, $\|\mathbf{h} - \mathbf{g}\|_H \geq m(r' + 1)$. Clearly, $m(r' + 1) > r$, since $r = mr' + a$ for some a where $0 \leq a < m$, which contradicts the premise.

The significance of Proposition 1 derives from the fact that the substrings have only q/m bits, and that the required search radius in each substring is just $r' = \lfloor r/m \rfloor$. For example, if \mathbf{h} and \mathbf{g} differ by 3 bits or less, and $m = 4$, at least one of the 4 substrings must be identical. If they differ by at most 7 bits, then in at least one substring they differ by no more than 1 bit; *i.e.*, we can search a Hamming radius of 7 bits by searching a radius of 1 bit on each of 4 substrings. More generally, instead of examining $L(q, r)$ hash buckets, it suffices to examine $L(q/m, r')$ buckets in each of m substring hash tables.

While it suffices to examine all buckets within a radius of r' in all m hash tables, we next show that it is not always necessary. Rather, it is often possible to use a radius of just $r' - 1$ in some of the m substring hash

tables while still guaranteeing that all r -neighbors of \mathbf{g} will be found. In particular, with $r = mr' + a$, where $0 \leq a < m$, to find any item within a radius of r on q -bit codes, it suffices to search $a+1$ substring hash tables to a radius of r' , and the remaining $m - (a+1)$ substring hash tables up to a radius of $r' - 1$. Without loss of generality, since there is no order to the substring hash tables, we search the first $a + 1$ hash tables with radius r' , and all remaining hash tables with radius $r' - 1$.

Proposition 2: If $\|\mathbf{h} - \mathbf{g}\|_H \leq r = mr' + a$, then

$$\exists 1 \leq z \leq a + 1 \quad \text{s.t.} \quad \|\mathbf{h}^{(z)} - \mathbf{g}^{(z)}\|_H \leq r' \quad (3a)$$

OR

$$\exists a + 1 < z \leq m \quad \text{s.t.} \quad \|\mathbf{h}^{(z)} - \mathbf{g}^{(z)}\|_H \leq r' - 1. \quad (3b)$$

To prove Proposition 2, we show that when (3a) is false, (3b) must be true. If (3a) is false, then it must be that $a < m - 1$, since otherwise $a = m - 1$, in which case (3a) and Proposition 1 are equivalent. If (3a) is false, it also follows that \mathbf{h} and \mathbf{g} differ in each of their first $a + 1$ substrings by $r' + 1$ or more bits. Thus, the total number of bits that differ in the first $a + 1$ substrings is at least $(a+1)(r'+1)$. Because $\|\mathbf{h} - \mathbf{g}\|_H \leq r$, it also follows that the total number of bits that differ in the remaining $m - (a+1)$ substrings is at most $r - (a+1)(r'+1)$. Then, using Proposition 1, the maximum search radius required in each of the remaining $m - (a + 1)$ substring hash tables is

$$\begin{aligned} \left\lfloor \frac{r - (a+1)(r'+1)}{m - (a+1)} \right\rfloor &= \left\lfloor \frac{mr' + a - (a+1)r' - (a+1)}{m - (a+1)} \right\rfloor \\ &= \left\lfloor r' - \frac{1}{m - (a+1)} \right\rfloor \\ &= r' - 1, \end{aligned} \quad (4)$$

and hence Proposition 2 is true. Because of the near exponential growth in the number of buckets for large search radii, the smaller substring search radius required by Proposition 2 is significant.

A special case of Proposition 2 is when $r < m$, hence $r' = 0$ and $a = r$. In this case, it suffices to search $r + 1$ substring hash tables for a radius of $r' = 0$ (*i.e.*, exact matches), and the remaining $m - (r + 1)$ substring hash tables can be ignored. Clearly, if a code does not match exactly with a query in any of the selected $r + 1$ substrings, then the code must differ from the query in at least $r + 1$ bits.

2.2 Multi-Index Hashing for r -neighbor Search

In a pre-processing step, given a dataset of binary codes, one hash table is built for each of the m substrings, as outlined in Algorithm 1. At query time, given a query \mathbf{g} with substrings $\{\mathbf{g}^{(j)}\}_{j=1}^m$, we search the j^{th} substring hash table for entries that are within a Hamming distance of $\lfloor r/m \rfloor$ or $\lfloor r/m \rfloor - 1$ of $\mathbf{g}^{(j)}$, as prescribed by (3). By doing so we obtain a set of candidates from the j^{th} substring hash table, denoted $\mathcal{N}_j(\mathbf{g})$. According

2. When q is not divisible by m , we use substrings of different lengths with either $\lfloor q/m \rfloor$ or $\lceil q/m \rceil$ bits, *i.e.*, differing by at most 1 bit.

Algorithm 1: Building m substring hash tables.

```

Binary code dataset:  $\mathcal{H} = \{\mathbf{h}_i\}_{i=1}^n$ 
for  $j = 1$  to  $m$  do
  Initialize  $j^{\text{th}}$  hash table
  for  $i = 1$  to  $n$  do
    Insert  $\mathbf{h}_i^{(j)}$  into  $j^{\text{th}}$  hash table
  end for
end for

```

Algorithm 2: r -Neighbor Search for Query \mathbf{g} .

```

Query substrings:  $\{\mathbf{g}^{(j)}\}_{j=1}^m$ 
Substring radius:  $r' = \lfloor r/m \rfloor$ , and  $a = r - mr'$ 
for  $j = 1$  to  $a + 1$  do
  Lookup  $r'$ -neighbors of  $\mathbf{g}^{(j)}$  from  $j^{\text{th}}$  hash table
end for
for  $j = a + 2$  to  $m$  do
  Lookup  $(r'-1)$ -neighbors of  $\mathbf{g}^{(j)}$  from  $j^{\text{th}}$  hash table
end for
Remove all non  $r$ -neighbors from the candidate set.

```

to the propositions above, the union of the m sets, $\mathcal{N}(\mathbf{g}) = \bigcup_j \mathcal{N}_j(\mathbf{g})$, is necessarily a superset of the r -neighbors of \mathbf{g} . The last step of the algorithm computes the full Hamming distance between \mathbf{g} and each candidate in $\mathcal{N}(\mathbf{g})$, retaining only those codes that are true r -neighbors of \mathbf{g} . Algorithm 2 outlines the r -neighbor retrieval procedure for a query \mathbf{g} .

The search cost depends on the number of *lookups* (i.e., the number of buckets examined), and the number of candidates tested. Not surprisingly there is a natural trade-off between them. With a large number of lookups one can minimize the number of extraneous candidates. By merging many buckets to reduce the number of lookups, one obtains a large number of candidates to test. In the extreme case with $m = q$, substrings are 1 bit long, so we can expect the candidate set to include almost the entire database.

Note that the idea of building multiple hash tables is not novel in itself (e.g., see [13], [15]). However previous work relied heavily on exact matches in substrings. Relaxing this constraint is what leads to a more effective algorithm, especially in terms of the storage requirement.

3 PERFORMANCE ANALYSIS

We next develop an analytical model of search performance to help address two key questions: (1) How does search cost depend on substring length, and hence the number of substrings? (2) How do run-time and storage complexity depend on database size, code length, and search radius?

To help answer these questions we exploit a well-known bound on the sum of binomial coefficients [9]; i.e., for any $0 < \epsilon \leq \frac{1}{2}$ and $\eta \geq 1$.

$$\sum_{\kappa=0}^{\lfloor \epsilon \eta \rfloor} \binom{\eta}{\kappa} \leq 2^{H(\epsilon)\eta}, \quad (5)$$

where $H(\epsilon) \equiv -\epsilon \log_2 \epsilon - (1-\epsilon) \log_2 (1-\epsilon)$ is the entropy of a Bernoulli distribution with probability ϵ .

In what follows, n continues to denote the number of q -bit database codes, and r is the Hamming search radius. Let m denote the number of hash tables, and let s denote the substring length $s = q/m$. Hence, the maximum substring search radius becomes $r' = \lfloor r/m \rfloor = \lfloor sr/q \rfloor$. As above, for the sake of model simplicity, we assume q is divisible by m .

We begin by formulating an upper bound on the number of lookups. First, the number of lookups in Algorithm 2 is bounded above by the product of m , the number of substring hash tables, and the number of hash buckets within a radius of $\lfloor sr/q \rfloor$ on substrings of length s bits. Accordingly, using (5), if the search radius is less than half the code length, $r \leq q/2$, then the total number of lookups is given by

$$\text{lookups}(s) = m \sum_{z=0}^{\lfloor sr/q \rfloor} \binom{s}{z} \leq \frac{q}{s} 2^{H(r/q)s}. \quad (6)$$

Clearly, as we decrease the substring length s , thereby increasing the number of substrings m , exponentially fewer lookups are needed.

To analyze the expected number of candidates per bucket, we consider the case in which the n binary codes are uniformly distributed over the Hamming space. In this case, for a substring of s bits, for which a substring hash table has 2^s buckets, the expected number of items per bucket is $n/2^s$. The expected size of the candidate set therefore equals the number of lookups times $n/2^s$.

The total search cost per query is the cost for lookups plus the cost for candidate tests. While these costs will vary with the code length q and the way the hash tables are implemented, empirically we find that, to a reasonable approximation, the costs of a lookup and a candidate test are similar (when $q \leq 256$). Accordingly, we model the total search cost per query, for retrieving all r -neighbors, in units of the time required for a single lookup, as

$$\text{cost}(s) = \left(1 + \frac{n}{2^s}\right) \frac{q}{s} \sum_{k=0}^{\lfloor sr/q \rfloor} \binom{s}{k}, \quad (7)$$

$$\leq \left(1 + \frac{n}{2^s}\right) \frac{q}{s} 2^{H(r/q)s}. \quad (8)$$

In practice, database codes will generally not be uniformly distributed, nor are uniformly distributed codes ideal for multi-index hashing. Indeed, the cost of search with uniformly distributed codes is relatively high since the search radius increases as the density of codes decreases. Rather, the uniform distribution is primarily a mathematical convenience that facilitates the analysis of run-time, thereby providing some insight into the effectiveness of the approach and how one might choose an effective substring length.

3.1 Choosing an Effective Substring Length

As noted above in Sec. 2.2, finding a good substring length is central to the efficiency of multi-index hashing. When the substring length is too large or too small the approach will not be effective. In practice, an effective substring length for a given dataset can be determined by cross-validation. Nevertheless this can be expensive.

In the case of uniformly distributed codes, one can instead use the analytic cost model in (7) to find a near optimal substring length. As discussed below, we find that a substring length of $s = \log_2 n$ yields a near-optimal search cost. Further, with non-uniformly distributed codes in benchmark datasets, we find empirically that $s = \log_2 n$ is also a reasonable heuristic for choosing the substring length (e.g., see Table 4 below).

In more detail, to find a good substring length using the cost model above, assuming uniformly distributed binary codes, we first note that, dividing $cost(s)$ in (7) by q has no effect on the optimal s . Accordingly, one can view the optimal s as a function of two quantities, namely the number of items, n , and the search ratio r/q .

Figure 2 plots cost as a function of substring length s , for 240-bit codes, different database sizes n , and different search radii (expressed as a fraction of the code length q). Dashed curves depict $cost(s)$ in (7) while solid curves of the same color depict the upper bound in (8). The tightness of the bound is evident in the plots, as are the quantization effects of the upper range of the sum in (7). The small circles in Fig. 2 (top) depict cost when all quantization effects are included, and hence it is only shown at substring lengths that are integer divisors of the code length.

Fig. 2 (top) shows cost for search radii equal to 5%, 15% and 25% of the code length, with $n=10^9$ in all cases. One striking property of these curves is that the cost is persistently minimal in the vicinity of $s = \log_2 n$, indicated by the vertical line close to 30 bits. This behavior is consistent over a wide range of database sizes.

Fig. 2 (bottom) shows the dependence of cost on s for databases with $n = 10^6, 10^9$, and 10^{12} , all with $r/q = 0.25$ and $q = 128$ bits. In this case we have laterally displaced each curve by $-\log_2 n$; notice how this aligns the minima close to 0. These curves suggest that, over a wide range of conditions, cost is minimal for $s = \log_2 n$. For this choice of the substring length, the expected number of items per substring bucket, i.e., $n/2^s$, reduces to 1. As a consequence, the number of lookups is equal to the expected number of candidates. Interestingly, this choice of substring length is similar to that of Greene *et al.* [13]. A somewhat involved theoretical analysis based on Stirling's approximation, omitted here, also suggests that as n goes to infinity, the optimal substring length converges asymptotically to $\log_2 n$.

3.2 Run-Time Complexity

Choosing s in the vicinity of $\log_2 n$ also permits a simple characterization of retrieval run-time complexity, for

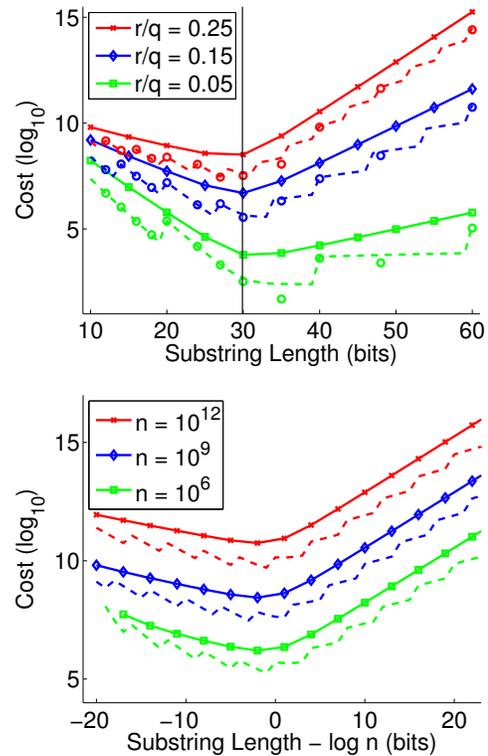


Fig. 2. Cost (7) and its upper bound (8) are shown as functions of substring length (using dashed and solid curves respectively). The code length in all cases is $q = 240$ bits. (Top) Cost for different search radii, all for a database with $n = 10^9$ codes. Circles depict a more accurate cost measure, only for substring lengths that are integer divisors of q , and with the more efficient indexing in Algorithm 3. (Bottom) Three database sizes, all with a search radius of $r = 0.25 q$. The minima are aligned when each curve is displaced horizontally by $-\log_2 n$.

uniformly distributed binary codes. When $s = \log_2 n$, the upper bound on the number of lookups (6) also becomes a bound on the number candidates. In particular, if we substitute $\log_2 n$ for s in (8), then we find the following upper bound on the cost, now as a function of database size, code length, and the search radius:

$$cost(s) \leq 2 \frac{q}{\log_2 n} n^{H(r/q)}. \quad (9)$$

Thus, for a uniform distribution over binary codes, if we choose m such that $s \approx \log_2 n$, the expected query time complexity is $O(q n^{H(r/q)} / \log_2 n)$. For a small ratio of r/q this is sub-linear in n . For example, if $r/q \leq .11$, then $H(.11) < .5$, and the run-time complexity becomes $O(b \sqrt{n} / \log_2 n)$. That is, the search time increases with the square root of the database size when the search radius is approximately 10% of the code length. For $r/q \leq .06$, this becomes $O(b \sqrt[3]{n} / \log_2 n)$. The time complexity with respect to q is not as important as that with respect to n since q is not expected to vary significantly in most applications.

3.3 Storage Complexity

The storage complexity of our multi-index hashing algorithm is asymptotically optimal when $\lfloor q/\log_2 n \rfloor \leq m \leq \lceil q/\log_2 n \rceil$. To store the full database of binary codes requires $O(nq)$ bits. For each of m hash tables, we also need to store n unique identifiers to the database items. This allows one to identify the retrieved items and fetch their full codes; this requires an additional $O(mn \log_2 n)$ bits. In sum, the storage required is $O(nq + mn \log_2 n)$. When $\lfloor q/\log_2 n \rfloor \leq m \leq \lceil q/\log_2 n \rceil$, as is suggested above, this storage cost reduces to $O(nq + n \log_2 n)$. Here, the $n \log_2 n$ term does not cancel as $m \geq 1$, but in most interesting cases $q > \log_2 n$.

While the storage cost for our multi-index hashing algorithm is linear in nq , the related multi-index hashing algorithm of Greene *et al.* [13] entails storage complexity that is super-linear in n . To find all r -neighbors, for a given search radius r , they construct $m = O(r2^{sr/q})$ substrings of length s bits per binary code. Their suggested substring length is also $s = \log_2 n$, so the number of substring hash tables becomes $m = O(rn^{r/q})$, each of which requires $O(n \log_2 n)$ in storage. As a consequence for large values of n , even with small r , this technique requires a prohibitive amount of memory to store the hash tables.

Our approach is more memory-efficient than that of [13] because we do not enforce exact equality in substring matching. In essence, instead of creating all of the hash tables off-line, and then having to store them, we flip bits of each substring at run-time and implicitly create some of the substring hash tables on-line. This increases run-time slightly, but greatly reduces storage costs.

4 k -NEAREST NEIGHBOR SEARCH

To use the above multi-index hashing in practice, one must specify a Hamming search radius r . For many tasks, the value of r is chosen such that queries will, on average, retrieve k near neighbors. Nevertheless, as expected, we find that for many hashing techniques and different sources of visual data, the distribution of binary codes is such that a single search radius for all queries will not produce similar numbers of neighbors.

Figure 3 depicts empirical distributions of search radii needed for 10-NN and 1000-NN on three sets of binary codes obtained from 1B SIFT descriptors [18], [22]. In all cases, for 64 and 128-bit codes, and for hash functions based on LSH [4] and MLH [25], there is a substantial variance in the search radius. This suggests that binary codes are not uniformly distributed over the Hamming space. As an example, for 1000-NN in 64-bit LSH codes, more than 10% of the queries require a search radius of 10 bits or larger, while for about 10% of the queries it can be 5 or smaller. Also evident from Fig. 3 is the growth in the required search radius as one moves from 64-bit codes to 128 bits, and from 10-NN to 1000-NN.

Algorithm 3: k NN Search with Query g .

```

Query substrings:  $\{g^{(i)}\}_{i=1}^m$ 
Initialize sets:  $N_d = \emptyset$ , for  $0 \leq d \leq q$ 
Initialize integers:  $r' = 0, a = 0, r = 0$ 
repeat
  Assert: Full radius of search is  $r = mr' + a$ .
  Lookup buckets in the  $(a+1)^{th}$  substring hash table
  that differ from  $g^{(a+1)}$  in exactly  $r'$  bits.
  For each candidate found, measure full Hamming
  distance, and add items with distance  $d$  to  $N_d$ .
   $a \leftarrow a + 1$ 
  if  $a \geq m$  then
     $a \leftarrow 0$ 
     $r' \leftarrow r' + 1$ 
  end if
   $r \leftarrow r + 1$ 
until  $\sum_{d=0}^{r-1} |N_d| \geq k$  (i.e.,  $k$   $r$ -neighbors are found)

```

A fixed radius for all queries would produce too many neighbors for some queries, and too few for others. It is therefore more natural for many tasks to fix the number of required neighbors, *i.e.*, k , and let the search radius depend on the query. Fortunately, our multi-index hashing algorithm is easily adapted to accommodate query-dependent search radii.

Given a query, one can progressively increase the Hamming search radius per substring, until a specified number of neighbors is found. For example, if one examines all r' -neighbors of a query's substrings, from which more than k candidates are found to be within a Hamming distance of $(r' + 1)m - 1$ bits (using the full codes for validation), then it is guaranteed that k -nearest neighbors have been found. Indeed, if all k NNs of a query g differ from g in r bits or less, then Propositions 1 and 2 above provide guarantees all such neighbors will be found if one searches the substring hash tables with the prescribed radii.

In our experiments, we follow this progressive increment of the search radius until we can find k NN in the guaranteed neighborhood of a query. This approach, outlined in Algorithm 3, is helpful because it uses a query-specific search radius depending on the distribution of codes in the neighborhood of the query.

5 EXPERIMENTS

Our implementation of multi-index hashing is available at [1]. Experiments are run on two different architectures. The first is a mid- to low-end 2.3Ghz dual quad-core AMD Opteron processor, with 2MB of L2 cache, and 128GB of RAM. The second is a high-end machine with a 2.9Ghz dual quad-core Intel Xeon processor, 20MB of L2 cache, and 128GB of RAM. The difference in the size of the L2 cache has a major impact on the run-time of linear scan, since the effectiveness of linear scan depends greatly on L2 cache lines. With roughly ten times the L2 cache, linear scan on the Intel platform is roughly

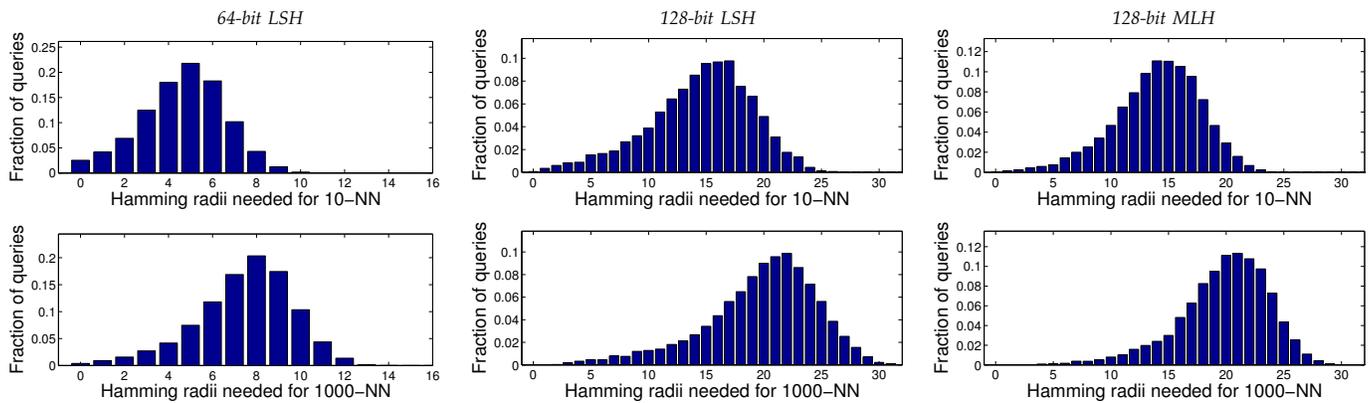


Fig. 3. Shown are histograms of the search radii that are required to find 10-NN and 1000-NN, for 64 and 128-bit code from LSH [4], and 128-bit codes from MLH [25], based on 1B SIFT descriptors [18]. Clearly shown are the relatively large search radii required for both the 10-NN and the 1000-NN tasks, as well as the increase in the radii required when using 128 bits versus 64 bits.

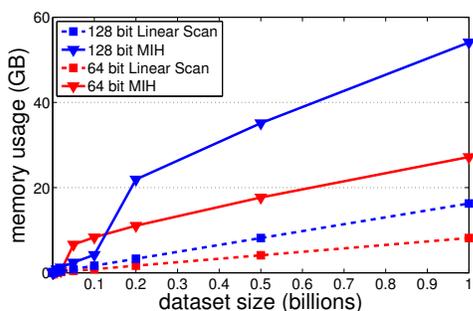


Fig. 4. Memory footprint of our implementation of Multi-Index Hashing as a function of database size. Note that the memory usage does not grow super-linearly with dataset size. The memory usage is independent of the number of nearest neighbors requested.

twice as fast as on the AMD machines. By comparison, multi-index hashing does not have a serial memory access pattern and so the cache size does not have such a pronounced effect. Actual run-times for multi-index hashing on the Intel and AMD platforms are within 20% of one another.

Both linear scan and multi-index hashing were implemented in C++ and compiled with identical compiler flags. To accommodate the large size of memory footprint required for 1B codes, we used the libhugetlbfs package and Linux Kernel 3.2.0 to allow the use of 2MB page sizes. Further details about the implementations are given in Section 6. Finally, despite the existence of multiple cores, all experiments are run on a single core to simplify run-time measurements.

The memory requirements for multi-index hashing are described in detail in Section 6. We currently require approximately 27 GB for multi-index hashing with 1B 64-bit codes, and approximately twice that for 128-bit codes. Figure 4 shows how the memory footprint depends on the database size for linear scan and multi-index

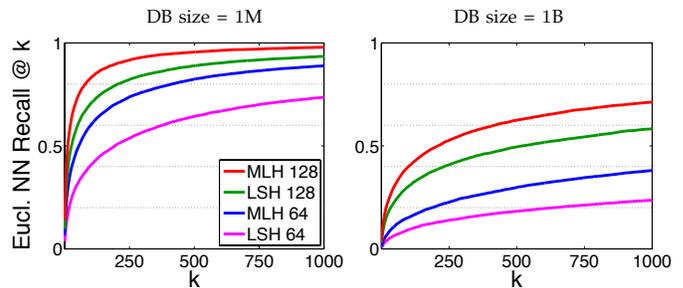


Fig. 5. Recall rates for BIGANN dataset [18] (1M and 1B subsets) obtained by k NN on 64- and 128-bit MLH and LSH codes.

hashing. As explained in the Sec. 3.3, and demonstrated in Figure 4 the memory requirements of multi-index hashing grow linearly in the database size, as does linear scan. While we use a single computer in our experiments, one could implement a distributed version of multi-index hashing on computers with much less memory by placing each substring hash table on a separate computer.

5.1 Datasets

We consider two well-known large-scale vision corpora: 80M Gist descriptors from 80 million tiny images [34] and 1B SIFT features from the BIGANN dataset [18]. SIFT vectors [22] are 128D descriptors of local image structure in the vicinity of feature points. Gist features [28] extracted from from 32×32 images capture global image structure in 384D vectors. These two feature types cover a spectrum of NN search problems in vision from feature to image indexing.

We use two similarity-preserving mappings to create datasets of binary codes, namely, binary angular Locality Sensitive Hashing (LSH) [8], and Minimal Loss Hashing (MLH) [25], [26]. LSH is considered a baseline random projection method, closely related to cosine similarity.

MLH is a state-of-the-art learning algorithm that, given a set of similarity labels, finds an optimal mapping by minimizing a loss function over pairs or triplets of binary codes.

Both the 80M Gist and 1B SIFT corpora comprise three disjoint sets, namely, a training set, a base set for populating the database, and a test query set. Using a random permutation, Gist descriptors are divided into a training set with 300K items, a base set of 79 million items, and a query set of size 10^4 . The SIFT corpus comes with 100M for training, 10^9 in the base set, and 10^4 test queries.

For LSH we subtract the mean, and pick a set of coefficients from the standard normal density for a linear projection, followed by quantization. For MLH the training set is used to optimize hash function parameters [26]. After learning is complete, we remove the training data and use the resulting hash function with the base set to create the database of binary codes. With two image corpora (SIFT and Gist), up to three code lengths (64, 128, and 256 bits), and two hashing methods (LSH and MLH), we obtain several datasets of binary codes with which to evaluate our multi-index hashing algorithm. Note that 256-bit codes are only used with LSH and SIFT vectors.

Figure 5 shows Euclidean NN recall rates for k NN search on binary codes generated from 1M and 1B SIFT descriptors. In particular, we plot the fraction of Euclidean 1st nearest neighbors found, by k NN in 64-bit and 128-bit LSH [8] and MLH [26] binary codes. As expected 128-bit codes are more accurate, and MLH outperforms LSH. Note that the multi-index hashing algorithm solves exact k NN search in Hamming distance; the approximation that reduces recall is due to the mapping from the original Euclidean space to the Hamming space. To preserve the Euclidean structure in the original SIFT descriptors, it seems useful to use longer codes, and exploit data-dependant hash functions such as MLH. Interestingly, as described below, the speedup factors of multi-index hashing on MLH codes are better than those for LSH.

Obviously, Hamming distance computed on q -bit binary codes is an integer between 0 and q . Thus, the nearest neighbors in Hamming distance can be divided into subsets of elements that have equal Hamming distance (at most $q+1$ subsets). Although Hamming distance does not provide a means to distinguish between equi-distant elements, often a re-ranking phase using Asymmetric Hamming distance [12] or other distance measures is helpful in practice. Nevertheless, this paper is solely concerned with the exact Hamming k NN problem up to a selection of equi-distant elements in the top k elements.

5.2 Results

Each experiment below involves 10^4 queries, for which we report the average run-time. Our implementation of the linear scan baseline searches 60 million 64-bit codes in just under one second on the AMD machine.

On the Intel machine it examines over 80 million 64-bit codes per second. This is remarkably fast compared to Euclidean NN search with 128D SIFT vectors. The speed of linear scan is in part due to memory caching, without which it would be much slower. Run-times for linear scan on other datasets, on both architectures, are given in Tables 1 and 2.

5.3 Multi-Index Hashing vs. Linear Scan

Tables 1 and 2 shows run-time per query for the linear scan baseline, along with speedup factors of multi-index hashing for different k NN problems and nine different datasets. Despite the remarkable speed of linear scan, the multi-index hashing implementation is hundreds of times faster. For example, the multi-index hashing method solves the exact 1000-NN for a dataset of 1B 64-bit codes in about 50 ms, well over 300 times faster than linear scan (see Table 1). Performance on 1-NN and 10-NN are even more impressive. With 128-bit MLH codes, multi-index hashing executes the 1NN search task over 1000 times faster than the linear scan baseline.

The run-time of linear scan does not depend on the number of neighbors, nor on the underlying distribution of binary codes. The run-time for multi-index hashing, however, depends on both factors. In particular, as the desired number of NNs increases, the Hamming radius of the search also increases (*e.g.*, see Figure 3). This implies longer run-times for multi-index hashing. Indeed, notice that going from 1-NN to 1000-NN on each row of the tables shows a decrease in the speedup factors.

The multi-index hashing run-time also depends on the distribution of binary codes. Indeed, one can see from Table 1 that MLH code databases yield faster run times than the LSH codes; *e.g.*, for 100-NN in 1B 128-bit codes the speedup for MLH is 353 \times *vs* 208 \times for LSH. Figure 3 depicts the histograms of search radii needed for 1000-NN with 1B 128-bit MLH and LSH codes. Interestingly, the mean of the search radii for MLH codes is 19.9 bits, while it is 19.8 for LSH. While the means are similar the variances are not; the standard deviations of the search radii for MLH and LSH are 4.0 and 5.0 respectively. The longer tail of the distribution of search radii for LSH plays an important role in the expected run-time. In fact, queries that require relatively large search radii tend to dominate the average query cost.

It is also interesting to look at the multi-index hashing run-times as a function of n , the number of binary codes in the database. To that end, Figure 6 and 7 depict run-times for linear scan and multi-index k NN search on the AMD machine. The left two figures in each show different vertical scales (since the behavior of multi-index k NN and linear scan are hard to see at the same scale). The right-most panels show the same data on log-log axes. First, it is clear from these plots that multi-index hashing is much faster than linear scan for a wide range of dataset sizes and k . Just as importantly, it is evident from the log-log plots that as we increase the database

dataset	# bits	mapping	speedup factors for k NN vs. linear scan				linear scan
			1-NN	10-NN	100-NN	1000-NN	
SIFT 1B	64	MLH	823	757	587	390	16.51s
		LSH	781	698	547	306	
	128	MLH	1048	675	353	147	42.64s
		LSH	747	426	208	91	
	256	LSH	220	111	58	27	62.31s
	Gist 79M	64	MLH	401	265	137	51
LSH			322	145	55	18	
128		MLH	124	50	26	13	3.37s
		LSH	85	33	18	9	

TABLE 1

Summary of results for nine datasets of binary codes on AMD Opteron Processor with 2MB L2 cache. The first four rows correspond to 1 billion binary codes, while the last four rows show the results for 79 million codes. Codes are 64, 128, or 256 bits long, obtained by LSH or MLH. The run-time of linear scan is reported along with the speedup factors for k NN with multi-index hashing.

dataset	# bits	mapping	speedup factors for k NN vs. linear scan				linear scan
			1-NN	10-NN	100-NN	1000-NN	
SIFT 1B	64	MLH	573	542	460	291	12.23s
		LSH	556	516	411	237	
	128	MLH	670	431	166	92	20.71s
		LSH	466	277	137	60	
	256	LSH	115	67	34	16	38.89s
	Gist 79M	64	MLH	286	242	136	53
LSH			256	142	55	18	
128		MLH	77	37	19	10	1.64s
		LSH	45	18	9	5	

TABLE 2

Summary of results for nine datasets of binary codes on Intel Xeon Processor with 20MB L2 cache. Note that the speedup factors reported in this table for multi-index hashing are smaller than in Table 1. This is due to the significant effect of cache size on the run-time of linear scan on the Intel architecture.

size, the speedup factors improve. The dashed lines on the log-log plots depict \sqrt{n} (up to a scalar constant). The similar slope of multi-index hashing curves with the square root curves show that multi-index hashing exhibits sub-linear query time, even for the empirical, non-uniform distributions of codes.

5.4 Direct lookups with a single hash table

An alternative to linear scan and multi-index hashing is to hash the entire codes into a single hash table, and then use direct hashing with each query. As suggested in the introduction and Figure 1, although this approach avoids the need for any candidate checking, it may require a prohibitive number of lookups. Nevertheless, for sufficiently small code lengths or search radii, it may be effective in practice.

Given the complexity associated with efficiently implementing collision detection in large hash tables, we do not directly experiment with the single hash table approach. Instead, we consider the empirical number of lookups one would need, as compared to the number of items in the database. If the number of lookups is vastly greater than the size of the dataset one can readily conclude that linear scan is likely to be as fast or faster than direct indexing into a single hash table.

Fortunately, the statistics of neighborhood sizes and required search radii for k NN tasks are available from the linear scan and multi-index hashing experiments reported above. For a given query, one can use the k^{th} nearest neighbor's Hamming distance to compute the number of lookups from a single hash table that are required to find all of the query's k nearest neighbors. Summed over the set of queries, this provides an indication of the expected run-time.

Figure 9 shows the total number of lookups required for 1-NN and 1000-NN tasks on 64- and 128-bit codes (from LSH on SIFT) using a single hash table. They are plotted as a function of the size of the dataset, from 10^4 to 10^9 items. For comparison, the plots also show the number of database items, and the number of lookups that were needed for multi-index hashing. Note that Figure 9 has logarithmic scales.

It is evident that with a single hash table the number of lookups is almost always several orders of magnitude larger than the number of items in the dataset. And not surprisingly, this is also several orders of magnitude more lookups than required for multi-index hashing. Although the relative speed of a lookup operation compared to a candidate check, as used in linear scan, depends on the implementation, there are a few important

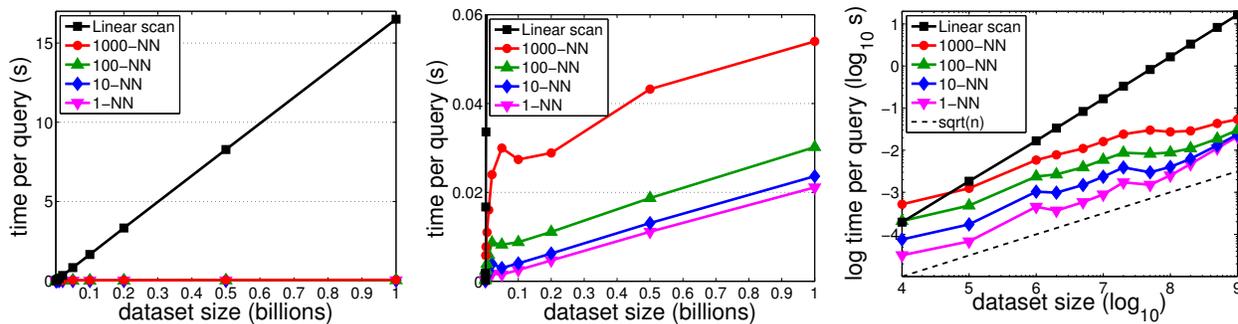


Fig. 6. Run-times per query for multi-index hashing with 1, 10, 100, and 1000 nearest neighbors, and a linear scan baseline on 1B 64-bit binary codes given by LSH from SIFT. Run on an AMD Optreron processor.

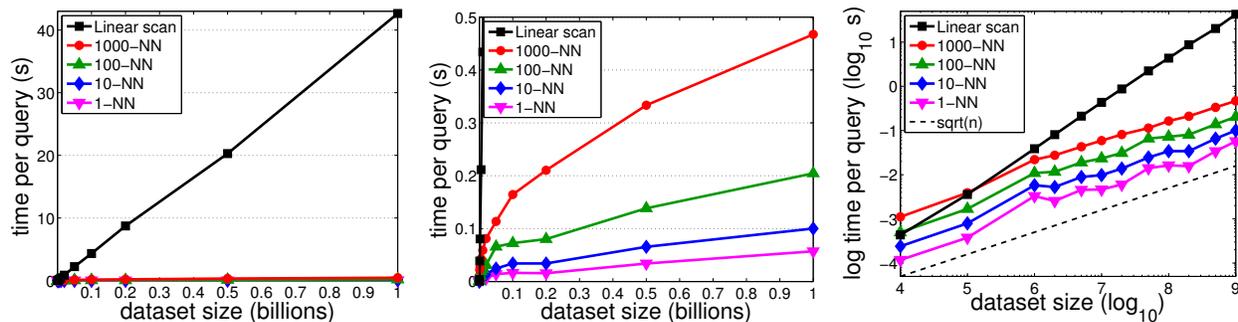


Fig. 7. Run-times per query for multi-index hashing with 1, 10, 100, and 1000 nearest neighbors, and a linear scan baseline on 1B 128-bit binary codes given by LSH from SIFT. Run on an AMD Optreron processor.

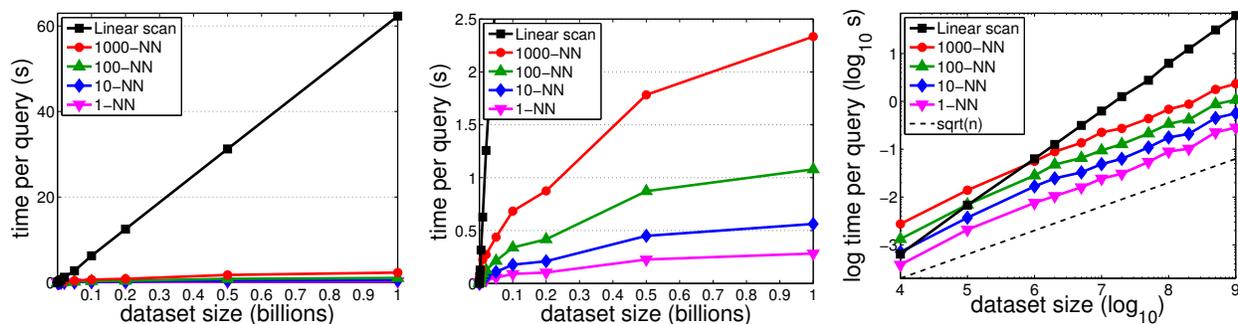


Fig. 8. Run-times per query for multi-index hashing with 1, 10, 100, and 1000 nearest neighbors, and a linear scan baseline on 1B 256-bit binary codes given by LSH from SIFT. Run on an AMD Optreron processor.

considerations. Linear scan has an exactly serial memory access pattern and so can make very efficient use of cache, whereas lookups in a hash table are inherently random. Furthermore, in any plausible implementation of a single hash table for 64 bit or longer codes, there will be some penalty for collision detection.

As illustrated in Figure 9, the only cases where a single hash table might potentially be more efficient than linear scan are with very small codes (64 bits or less), with a large dataset (1 billion items or more), and a small search distances (e.g., for 1-NN). In all other cases, linear scan requires orders of magnitude fewer operations. With any code length longer than 64 bits, a single hash table approach is completely infeasible to

run, requiring upwards of 15 orders of magnitude more operations than linear scan for 128-bit codes.

5.5 Substring Optimization

The substring hash tables used above have been formed by simply dividing the full codes into disjoint and consecutive sequences bits. For LSH and MLH, this is equivalent to randomly assigning bits to substrings.

It natural to ask whether further gains in efficiency are possible by optimizing the assignment of bits to substrings. In particular, by careful substring optimization one may be able to maximize the discriminability of the different substrings. In other words, while the

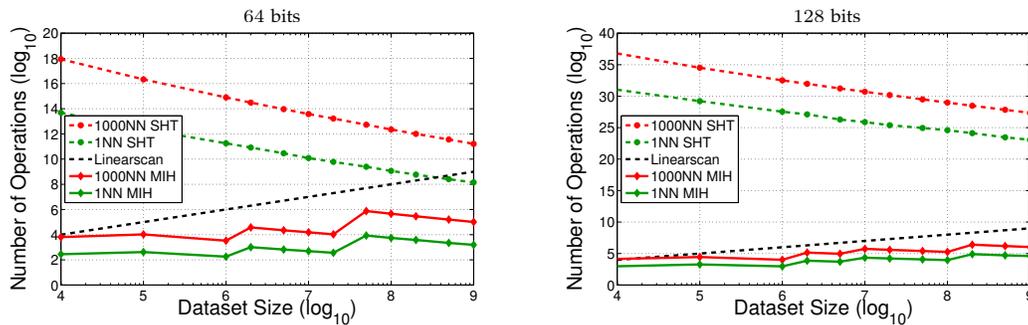


Fig. 9. The number of lookup operations required to solve exact nearest neighbor search in hamming space for LSH codes from SIFT features, using the simple single hash table (SHT) approach and multi-index hashing (MIH). Also shown is the number of candidate check operations required to search using linear scan. Note that the axes have a logarithmic scale. With small codes (64 bits), many items (1 billion) and small search distance (1 NN), it is conceivable that a single hash table might be faster than linear scan. In all other cases, a single hash table requires many orders of magnitude more operations than linear scan. Note also that MIH will never require more operations than a single hash table - in the limit of very large dataset sizes, MIH will use only one hash table and become equivalent.

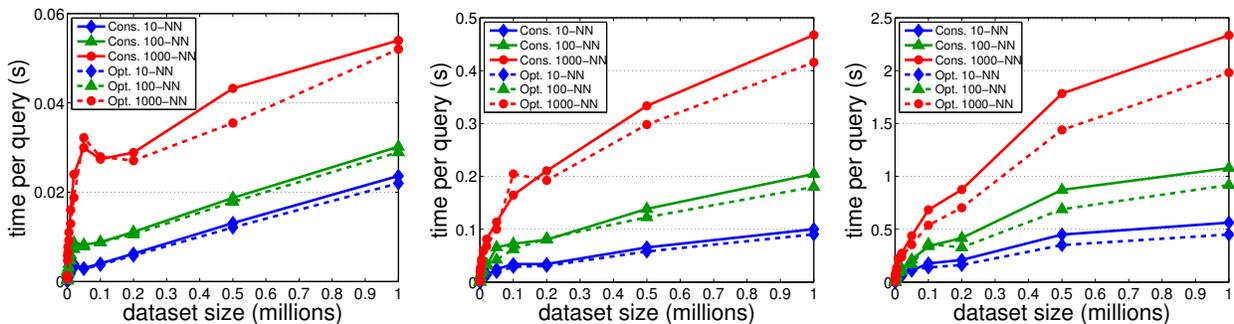


Fig. 10. Run-times for multi-index-hashing using codes from LSH on SIFT features with consecutive (solid) and optimized (dashed) substrings. From left to right: 64-bit, 128-bit, 256-bit codes, run on the AMD machine.

# bits	optimized speedup vs. linear scan (consecutive, % improvement)			
	1-NN	10-NN	100-NN	1000-NN
64	788 (781, 1%)	750 (698, 7%)	570 (547, 4%)	317 (306, 4%)
128	826 (747, 10%)	472 (426, 11%)	237 (208, 14%)	103 (91, 12%)
256	284 (220, 29%)	138 (111, 25%)	68 (58, 18%)	31 (27, 18%)

TABLE 3

Empirical run-time improvements from optimizing substrings vs. consecutive substrings, for 1 billion LSH codes from SIFT features (AMD machine). speedup factors vs. linear scan are shown with optimized and consecutive substrings, and the percent improvement. All experiments used 10M codes to compute the correlation between bits for substring optimization and all results are averaged over 10000 queries each.

radius of substring searches and hence the number of lookups is determined by the desired search radius on the full codes, and will remain fixed, by optimizing the assignment of bits to substrings one might be able to reduce the number of candidates one needs to validate.

To explore this idea we considered a simple method in which bits are assigned to substrings one at a time in a greedy fashion, based on correlations between bits. In particular, of those bits not yet assigned, the next substring is assigned the bit that minimizes the maximum correlation between that bit and all other bits already assigned to that substring. Initialization also occurs in a greedy manner: A random bit is assigned to the first

substring, after which the first bit to substring j is that which is maximally correlated with the first bit of substring $j-1$. This approach significantly decreases the correlation between bits within a single substring. This should make the distribution codes within substrings buckets more uniform, and thereby lower the number of candidates within a given search radius. Arguably an even better approach would be to maximize the entropy of the entries within each substring hash table, thereby making the distribution of substrings as uniform as possible. This entropic approach is, however, left to future work.

The results obtained with the correlation-based greedy

algorithm show that optimizing substrings can provide overall run-time reductions on the order of 20% against consecutive substrings for some cases. Table 3 displays the improvements achieved by optimizing substrings for different codes lengths and different values of k . Figure 10 shows the run-time performance of optimized substrings.

6 IMPLEMENTATION DETAILS

Our implementation of multi-index hashing is publicly available at [1]. Nevertheless, for the interested reader we describe some of the important details here.

As explained above, the algorithm hinges on hash tables built on disjoint s -bit substrings of the binary codes. We use direct address tables for the substring hash tables because the substrings are usually short ($s \leq 32$). Direct address tables explicitly allocate memory for 2^s buckets and store all data points associated with each substring in its corresponding bucket. There is a one-to-one mapping between buckets and substrings, so no time is spent on collision detection.

One could implement direct address tables with an array of 2^s pointers, some of which may be null (for empty buckets). On a 64-bit machine, pointers are 8 bytes long, so just storing an empty address table for $s = 32$ requires 32GB (as done in [27]). For greater efficiency here, we use sparse direct address tables by grouping buckets into subsets of 32 elements. For each bucket group, a 32-bit binary vector encodes whether each bucket in the group is empty or not. Then, a single pointer per group is used to point to a single resizable array that stores the data points associated with that bucket group. Data points within each array are ordered by their bucket index. To facilitate fast access, for each non-empty bucket we store the index of the beginning and the end of the corresponding segment of the array. Compared to the direct address tables in [27], for $s = 32$, and bucket groups of size 32, an empty address table requires only 1.5GB. Also note that accessing elements in any bucket of the sparse address table has a worst case run-time of $O(1)$.

Memory Requirements: We store one 64-bit pointer for each bucket group, and a 32-bit binary vector to encode whether buckets in a group are empty; this entails $2^{(s-5)} \cdot (8+4)$ bytes for an empty s -bit hash table ($s \geq 5$), or 1.5GB when $s = 32$. Bookkeeping for each resizable array entails 3 32-bit integers. In our experiments, most bucket groups have at least one non-empty bucket. Taking this into account, the total storage for an s -bit address table becomes $2^{(s-5)} \cdot 24$ bytes (3GB for $s = 32$).

For each non-empty bucket within a bucket group, we store a 32-bit integer to indicate the index of the beginning of the segment of the resizable array corresponding to that bucket. The number of non-empty buckets is at most $m \min(n, 2^s)$, where m is the number of hash tables, and n is the number of codes. Thus we need an extra $m \min(n, 2^s) \cdot 4$ bytes. For each data point per hash table

we store an ID to reference the full binary code; each ID is 4 bytes since $n \leq 2^{32}$ for our datasets. This entails $4mn$ bytes. Finally, storing the full binary codes themselves requires $nms/8$ bytes, since $q = ms$.

The total memory cost is $m2^{(s-5)}24 + m \min(n, 2^s)4 + 4mn + nms/8$ bytes. For $s = \log_2 n$, this cost is $O(nq)$. For 1B 64-bit codes, and $m = 2$ hash tables (32 bits each), the cost is 28GB. For 128-bit and 256-bit codes our implementation requires 57GB and 113GB. Note that the last two terms in the memory cost for storing IDs and codes are irreducible, but the first terms can be reduced in a more memory efficient implementation.

Duplicate Candidates: When retrieving candidates from the m substring hash tables, some codes will be found multiple times. To detect *duplicates*, and discard them, we allocate one bit-string with n bits. When a candidate is found we check the corresponding bit and discard the candidate if it is marked as a duplicate. Before each query we initialize the bit-string to zero. In practice this has negligible run-time. In theory clearing an n -bit vector requires $O(n)$, but in theory there are more efficient ways to store an n -bit vector without explicit initialization.

Hamming Distance: To compare a query and a candidate (for multi-index search or linear scan), we compute the Hamming distance on the full q -bit codes, with one XOR operation for every 64 bits, followed by a pop count to tally the ones. We used the built-in GCC function `__builtin_popcount` for this purpose.

Number of Substrings: The number of substring hash tables we use is determined with a hold-out validation set of database entries. From that set we estimate the running time of the algorithm for different choices of m near $q / \log_2 n$, and select the m that yields the minimum run-time. As shown in Table 4 this empirical value for m is usually the closest integer to $q / \log_2 n$.

Translation Lookaside Buffer and Huge Pages: Modern processors have an on-chip cache that holds a lookup table of memory addresses, for mapping virtual addresses to physical addresses for each running process. Typically, memory is split into 4KB pages, and a process that allocates memory is given pages by the operating system. The Translation Lookaside Buffer (TLB) keeps track of these pages. For processes that have large memory footprints (tens of GB), the number of pages quickly overtakes the size of the TLB (typically about 1500 entries). For processes using random memory access this means that almost every memory access produces a *TLB miss* - the requested address is in a page not cached in the TLB, hence the TLB entry must be fetched from slow RAM before the requested page can be accessed. This slows down memory access, and causes volatility in run-times for memory-access intensive processes.

To avoid this problem, we use the `libhugetlbfs` Linux library. This allows the operating system to allocate **Huge Pages** (2MB each) rather than 4KB pages. This reduces the number of pages; hence it reduces the frequency of TLB misses, improves memory access speed,

	n	10^4	10^5	10^6	2×10^6	5×10^6	10^7	2×10^7	5×10^7	10^8	2×10^8	5×10^8	10^9
$q = 64$	m	5	4	4	3	3	3	3	2	2	2	2	2
	$q/\log_2 n$	4.82	3.85	3.21	3.06	2.88	2.75	2.64	2.50	2.41	2.32	2.21	2.14
$q = 128$	m	10	8	8	6	6	5	5	5	5	4	4	4
	$q/\log_2 n$	9.63	7.71	6.42	6.12	5.75	5.50	5.28	5.00	4.82	4.64	4.43	4.28
$q = 256$	m	19	15	13	12	11	11	10	10	10	9	9	8
	$q/\log_2 n$	19.27	15.41	12.84	12.23	11.50	11.01	10.56	10.01	9.63	9.28	8.86	8.56

TABLE 4

Selected number of substrings used for the experiments, as determined by cross-validation, vs. the suggested number of substrings based on the heuristic $q/\log_2 n$.

and reduces run-time volatility. The increase in speed of multi-index hashing results reported here compared to those in [27] are attributed to the use of libhugetlbfs.

7 CONCLUSION

This paper describes a new algorithm for exact nearest neighbor search on large-scale datasets of binary codes. The algorithm is a form of multi-index hashing that has provably sub-linear run-time behavior for uniformly distributed codes. It is storage efficient and easy to implement. We show empirical performance on datasets of binary codes obtained from 1 billion SIFT, and 80 million Gist features. With these datasets we find that, for 64-bit and 128-bit codes, our new multi-index hashing implementation is often more than two orders of magnitude faster than a linear scan baseline.

While the basic algorithm is developed in this paper there are several interesting avenues for future research. For example our preliminary research shows that $\log_2 n$ is a good choice for the substring length, and it should be possible to formulate a sound mathematical basis for this choice. The assignment of bits to substrings was shown to be important above, however the algorithm used for this assignment is clearly suboptimal. It is also likely that different substring lengths might be useful for the different hash tables.

Our theoretical analysis proves sub-linear run-time behavior of the multi-index hashing algorithm on uniformly distributed codes, when search radius is small. Our experiments demonstrate sub-linear run-time behavior of the algorithm on real datasets, while the binary code in our experiments are clearly not uniformly distributed³. Bridging the gap between theoretical analysis and empirical findings for the proposed algorithm remains an open problem. In particular, we are interested in more realistic assumptions on the binary codes, which still allow for theoretical analysis of the algorithm.

While the current paper concerns exact nearest-neighbor tasks, it would also be interesting to consider approximate methods based on the same multi-index hashing framework. Indeed there are several ways that one could find approximate rather than the exact nearest

neighbors for a given query. For example, one could stop at a given radius of search, even though k items may not have been found. Alternatively, one might search until a fixed number of unique candidates have been found, even though all substring hash tables have not been inspected to the necessary radius. Such approximate algorithms have the potential for even greater efficiency, and would be the most natural methods to compare to most existing methods which are approximate, such as binary LSH. That said, such comparisons are more difficult than for exact methods since one must taken into account not only the storage and run-time costs, but also some measure of the cost of errors (usually in terms of recall and precision).

Finally, recent results have shown that for many datasets in which the binary codes are the result of some form of vector quantization, an asymmetric Hamming distance is attractive [12], [17]. In such methods, rather than converting the query into a binary code, one directly compares a real-valued query to the database of binary codes. The advantage is that the quantization noise entailed in converting the query to a binary string is avoided and one can more accurately using distances in the binary code space to approximate the desired distances in the feature space of the query. One simple way to do this is to use multi-index hashing and then only use an asymmetric distance when culling candidates. The potential for more interesting and effective methods is yet another promising avenue for future work.

ACKNOWLEDGMENTS

This research was financially supported in part by NSERC Canada, the GRAND Network Centre of Excellence, and the Canadian Institute for Advanced Research (CIFAR). The authors would also like to thank Mohamed Aly, Rob Fergus, Ryan Johnson, Abbas Mehrabian, and Pietro Perona for useful discussions about this work.

REFERENCES

- [1] <https://github.com/norouzi/mih/>.
- [2] A. Alahi, R. Ortiz, and P. Vandergheynst. Freak: Fast retina keypoint. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2012.
- [3] M. Aly, M. Munich, and P. Perona. Distributed kd-trees for retrieval from very large image collections. In *Proc. British Machine Vision Conference*, 2011.

3. In some of our experiments with 1 Billion binary codes, tens of thousands of codes fall into the same bucket of 32-bit substring hash tables. This is extremely unlikely with uniformly distributed codes.

- [4] A. Andoni and P. Indyk. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. *Communications of the ACM*, 51(1):117–122, 2008.
- [5] A. Babenko and V. Lempitsky. The inverted multi-index. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2012.
- [6] A. Bergamo, L. Torresani, and A. Fitzgibbon. Picodes: Learning a compact code for novel-category recognition. In *Proc. Advances in Neural Information Processing Systems*, volume 24, 2011.
- [7] M. Calonder, V. Lepetit, C. Strecha, and P. Fua. Brief: Binary robust independent elementary features. In *Proc. European Conference on Computer Vision*, page 778792, 2010.
- [8] M. Charikar. Similarity estimation techniques from rounding algorithms. In *ACM Symposium on Theory of Computing*. ACM, 2002.
- [9] J. Flum and M. Grohe. *Parameterized Complexity Theory*. Springer, 2006.
- [10] A. Gionis, P. Indyk, R. Motwani, et al. Similarity search in high dimensions via hashing. In *Proc. Int. Conf. Very Large Databases*, pages 518–529, 1999.
- [11] Y. Gong and S. Lazebnik. Iterative quantization: A procrustean approach to learning binary codes. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2011.
- [12] A. Gordo and F. Perronnin. Asymmetric distances for binary embeddings. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 729–736, 2011.
- [13] D. Greene, M. Parnas, and F. Yao. Multi-index hashing for information retrieval. In *IEEE Symposium on Foundations of Computer Science*, pages 722–731, 1994.
- [14] J. He, R. Radhakrishnan, S.-F. Chang, and C. Bauer. Compact hashing with joint optimization of search accuracy and time. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2011.
- [15] P. Indyk and R. Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. In *ACM Symposium on Theory of Computing*, pages 604–613, 1998.
- [16] H. Jégou, M. Douze, and C. Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *Proc. European Conference on Computer Vision*, volume I, pages 304–317, 2008.
- [17] H. Jégou, M. Douze, and C. Schmid. Product quantization for nearest neighbor search. *IEEE Trans. PAMI*, 33(1):117–128, 2011.
- [18] H. Jégou, R. Tavenard, M. Douze, and L. Amsaleg. Searching in one billion vectors: re-rank with source coding. In *IEEE Acoustics, Speech and Signal Processing*, pages 861–864. IEEE, 2011.
- [19] D. Kuettel, M. Guillaumin, and V. Ferrari. Segmentation propagation in imagenet. In *Proc. European Conference on Computer Vision*, 2012.
- [20] B. Kulis and T. Darrell. Learning to hash with binary reconstructive embeddings. In *Proc. Advances in Neural Information Processing Systems*, volume 22, 2009.
- [21] W. Liu, J. Wang, R. Ji, Y.-G. Jiang, and S.-F. Chang. Supervised hashing with kernels. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2012.
- [22] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. Journal of Computer Vision*, 60(2):91–110, 2004.
- [23] M. Minsky and S. Papert. *Perceptrons*. MIT Press, 1969.
- [24] M. Muja and D. Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. In *International Conference on Computer Vision Theory and Applications*, 2009.
- [25] M. Norouzi and D. J. Fleet. Minimal loss hashing for compact binary codes. In *Proc. International Conference on Machine Learning*, 2011.
- [26] M. Norouzi, D. J. Fleet, and R. Salakhutdinov. Hamming distance metric learning. In *Proc. Advances in Neural Information Processing Systems*, 2012.
- [27] M. Norouzi, A. Punjani, and D. Fleet. Fast search in hamming space with multi-index hashing. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2012.
- [28] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.
- [29] M. Raginsky and S. Lazebnik. Locality-sensitive binary codes from shift-invariant kernels. In *Proc. Advances in Neural Information Processing Systems*, volume 22, 2009.
- [30] M. Rastegari, A. Farhadi, and D. Forsyth. Attribute discovery via predictable discriminative binary codes. In *Proc. European Conference on Computer Vision*, 2012.
- [31] R. Salakhutdinov and G. Hinton. Semantic hashing. *International Journal of Approximate Reasoning*, 2009.
- [32] G. Shakhnarovich, P. Viola, and T. Darrell. Fast pose estimation

with parameter-sensitive hashing. In *Proc. IEEE International Conference on Computer Vision*, volume 2, 2003.

- [33] C. Strecha, A. Bronstein, M. Bronstein, and P. Fua. LDAHash: improved matching with smaller descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(1):66–78, 2012.
- [34] A. Torralba, R. Fergus, and W. Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE Trans. PAMI*, 30(11):1958–1970, 2008.
- [35] A. Torralba, R. Fergus, and Y. Weiss. Small codes and large image databases for recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2008.
- [36] T. Trzcinski, C. M. Christoudias, P. Fua, and V. Lepetit. Boosting binary keypoint descriptors. In *Proc. Advances in Neural Information Processing Systems*, 2012.
- [37] J. Wang, S. Kumar, and S. Chang. Sequential projection learning for hashing with compact codes. In *Proc. International Conference on Machine Learning*, 2010.
- [38] Y. Weiss, A. Torralba, and R. Fergus. Spectral hashing. In *Proc. Advances in Neural Information Processing Systems*, volume 21, 2008.



Mohammad Norouzi Sharif University of Technology in 2007, followed by the M.Sc. in Computer Science from Simon Fraser University in 2009. He is currently pursuing the PhD in Computer Science at the University of Toronto. His research lies at the intersection of computer vision and machine learning. His research interests include large scale image retrieval, large scale machine learning, and statistical image modeling.



Ali Punjani graduated with a BAsC in Aerospace Engineering from the University of Toronto in 2012 and is now a PhD student in Computer Science at U.C. Berkeley. His current research interests include large scale machine learning, biomedical applications of computer vision, and technological approaches to social change. He is a recipient of the NSERC Canada Graduate Scholarship.



David J Fleet received the PhD in Computer Science from the University of Toronto in 1991. He was on faculty at Queen's University in Kingston from 1991 to 1998, and then Area Manager and Research Scientist at the Palo Alto Research Center (PARC) from 1999 to 2003. In 2004 he joined the University of Toronto as Professor of Computer Science.

His research interests include computer vision, image processing, visual perception, and visual neuroscience. He has published research articles, book chapters and one book on various topics including the estimation of optical flow and stereoscopic disparity, probabilistic methods in motion analysis, modeling appearance in image sequences, motion perception and human stereopsis, hand tracking, human pose tracking, latent variable models, physics-based models for human motion analysis, and large-scale image retrieval.

He was awarded an Alfred P. Sloan Research Fellowship in 1996. He has won paper awards at ICCV 1999, CVPR 2001, UIST 2003, BMVC 2009. In 2010 he was awarded the Koenderink Prize for his work with Michael Black and Hedvig Sidenbladh on human pose tracking. He has served as Area Chair for numerous computer vision and machine learning conference. He was Program Co-chair for the CVPR 2003. He will be Program Co-Chair for ECCV 2014. He has been Associate Editor, and Associate Editor-in-Chief for IEEE TPAMI, and currently serves on the TPAMI Advisory Board.