Multi-Resolution and Multi-Bit Representation for Image Similarity Search

Tuncer C. Aysal and Daniel C. Heesch
Pixsta Research
Notting Hill, London, UK
{tuncer.aysal,daniel.heesch}@pixsta.com

Abstract

This paper explores the use of multi-bit quantisation of image features for similarity-based image retrieval. Our work builds on multi-resolution image similarity search algorithms which utilise one-bit representation of the largest magnitude wavelet coefficients. Given a query, images are ranked based on the number of quantised coefficients they have in common with the query. We explore the benefits of a finer-level quantisation (specifically with two bits) and one control parameter that can be chosen optimally based on the probability density of the wavelet coefficients. We show that this extension leads to significant performance improvements.

I. Introduction

The problem of search pervades almost every branch of artificial intelligence. The field of information retrieval has it at its core. Finding relevant information in large collections of data calls for object representations that are rich enough to capture salient properties and yet simple enough to support fast indexing and retrieval strategies. Common representation of multimedia objects utilises vectors in a high-dimensional metric space. Finding items similar to a query amounts to a nearest neighbour search in that space. Even when done approximately, nearest neighbour search scales poorly with the number of dimensions and is generally too costly for realtime image retrieval. The problem can be mitigated by reducing the dimensionality of the original space, for example through principal component analysis, probabilistic latent semantic indexing [1] or non-negative matrix decomposition [2].

A popular alternative is vector quantisation, which maps the vector space onto a finite set of labels or codevectors each representing one of the partition regions. By thus discretising the continuum, data can be stored in an inverted index, widely seen as the only viable indexing technique for large-scale search [3]: instead of mapping each object to the codewords it generates, each codeword is mapped to the objects generating it. Retrieval thus becomes a look-up and merge operation. First used for text retrieval, inverted indexes have become the *de facto* standard for large scale image matching [4] and similarity search [5].

Finding the closest code vector for a new data point involves a nearest neighbour search amongst all codevectors, which can be costly for large codebooks. A number of variants have therefore been proposed. One elegant technique suggested in [6] is to assign each point to the $m > 1$ closest codevectors or pivots, rather than just the closest, thus rendering the mapping function more expressive and allowing for smaller codebook sizes. Alternatively, the cost can be reduced by mapping each point not to the closest k-dimensional codevector, but the ordered set of the codevector's k quantised components. For an axis-parallel partitioning, this amounts to simple scalar quantisation along each of the k dimensions. For b bins along each dimension, there are thus bk codewords, each being identified by a unique combination of the component index and its quantisation level.

Amongst the earliest works that apply the idea of scalar quantisation to image retrieval is that of Jacobs *et al.* [7]. In their scheme, vector components are first sorted by magnitude and the largest components binarised. Retrieval involves extracting and binarising the largest magnitude coefficients of a query image and scoring images based on the number of matching codewords.

This paper extends the work of [7] by investigating the implications of multi-bit quantisation, in particular 2-bit quantisation with one parameter controlling the relative bin size. The intuition is that with finer quantisation, matches in the quantised domain should be closer in the original vector space. We formalise this intuition theoretically and demonstrate that whilst multi-bit quantisation does indeed improve with respect to two different performance measures for a broad range of parameter values, the relative benefit vanishes when only a small number of coefficients are retained. The results suggest that multi-bit quantisation can improve retrieval accuracy over one-bit qunantisation whilst still benefitting from the near-constant time complexity afforded by inverted index structures during retrieval.

The paper is structured as follows. Section II briefly presents background material. Section III motivates and describes the multi-bit extension, and suggests ways to estimate the optimal control parameter. Section IV presents simulation results comparing 1-bit with 2-bit quantisers in terms of difference in coefficient values and retrieval performance. Finally, we conclude in Section V.

II. Wavelets and Truncated Coefficient Quantisation

We briefly review the wavelet decomposition literature and its application to image similarity search in this section. We then detail the main components of the seminal work of Jacobs *et al.* [7] which our paper expands upon.

A. Wavelet Decomposition

Wavelet theory [8] has gained tremendous attention in the content-based image retrieval community. Do and Vetterli, for a given scale, model the wavelet coefficients with a generalised Gaussian density function as the feature extraction step and compute the Kullback-Leibler distance as a similarity measure for texture retrieval [9], [10]. Wang *et al.* characterises the color variations over the spatial extent of the image in a manner that provides semantically meaningful image comparisons [11]. The indexing algorithm applies a Daubechies' wavelet transform for each of the colour components. The wavelet coefficients in the lowest few frequency bands, and their variances, are stored as feature vectors. To speed up retrieval, a two-step procedure is used 1) first, a crude selection based on the variances, and then 2) a refinement of the search by performing a feature vector match between the selected images and the query [11]. Liapis and Tziritas extract the *texture features* using Discrete Wavelet Frames analysis, an over-complete decomposition in scale and orientation. Two-dimensional (2-D) or one-dimensional (1-D) histograms of the CIE Lab chromaticity coordinates are used as colour features. The 1-D histograms of the a, b coordinates were modeled by a generalised Gaussian distribution. The similarity measure defined on the feature distribution is based on the Bhattacharya distance [12]. Ma and Manjunath compare different wavelet transform-based features for content based texture search and retrieval [13]. Huang and Dai generate subband gradient vector and energy distributions from the subimages of the wavelet decomposition of the image in a two step similarity measure system [14]. Suematsu *et al.* propose a region-based image retrieval based on segmented texture features computed from wavelet coefficients [15].

B. Coefficient Truncation

Wavelets achieve very good image approximations with only a few coefficients, a property that makes them an important tool in lossy image compression [16]. Jacobs *et al.* were amongst the first to exploit this property for image retrieval. Following their work, we choose Haar wavelets as they are very fast and simple to compute. For a given query image I_q, we obtain a sequence of N coefficients $\{c_i : i = 1, 2, \ldots, N\}$. Rather than retaining all of them, it is desirable to "truncate" the sequence and keep only the M coefficients with largest magnitude:

$$t = \{c_j : j = 1, 2, \ldots, N; \Psi\{c_j\} \leq M\}$$

where Ψ denotes the order of the coefficient when their magnitudes are sorted in descending order. We thus obtain the truncated coefficient sequence $t = \{t_j : j = 1, 2, \ldots, M\}$. Truncation accelerates search times and reduces storage requirements. Moreover, we will show that it may also improve the discriminatory power of the signature.

C. Binary Quantisation

Quantisation brings similar benefits as truncation. Although the quantised coefficients retain little data about the precise magnitudes of major features in the images, the mere presence or absence of such features appears to provide enough information for image querying.

It is argued in [7] that quantising each significant coefficient to only two levels, *i.e.*,

$$q_j = \text{sign}\{t_j\}$$

works remarkably well. To our knowledge, the possibility of a multi-level representation of the truncated coefficient set $\{t_k : k = 1, 2, \ldots, M\}$ has not yet been investigated.

III. Multi–Level Quantisation

In this section, we first present an intuitive Lemma motivating the natural extension to multi-resolution representation for image similarity search. We subsequently extend the single-bit quantisation feature representation to multiple bits and focus on the simplest extension, *i.e.*, two bits. Also discussed is the problem of setting the *control parameter* of the proposed 2-bit quantiser.

A. Basic Idea

A finer quantisation has the effect of increasing the chance that two images matching each other on a given number of quantised components are also close in the non-quantised domain. Before we present a Lemma making this observation more formal, let us introduce some notation. Let $c(I) \in \mathbb{R}^N$, $t(I) \in \mathbb{R}^M$ and $q(I) \in \mathbb{R}^M$ (where $M \leq N$) denote, respectively, the representation vector (*i.e.*, Haar wavelet coefficients, Fourier coefficients, edge strength map, RGB values, etc.), truncated representation vector and quantised representation vector for the image I. Note that, without loss of generality, we take $t_k(I) = c_k(I)\mathbf{1}\{\lambda(c_k(I)) \leq M\}$, and $q_k(I) = \mathcal{Q}(t_k(I))$ where $\mathcal{Q}(\cdot)$ denotes the quantisation operator.

Lemma 1 *Let $\mathcal{S}_q(I, J)$ denote the set of indices of exact quantised value matches for images I and J. For a random query image I and a random image J in the database, the following statement holds:*

$$\Pr\{||c(I) - c(J)||_2 \geq \epsilon | |\mathcal{S}_q(I,J)| = m\} \\ \leq \epsilon^{-2}\Delta^2[m(1-4n^2) + 4Kn^2]$$

for any $\epsilon > 0$, where $|| \cdot ||_2$, $| \cdot |$ denotes the ℓ_2 norm, and cardinality, respectively. Moreover, $|c_k| \leq U = n\Delta$, $n \in \mathbb{N}^+$, for all $k = 1, 2, \ldots, N$.

Proof: See Appendix A. ■

Note that the above Lemma assumes that the representation elements are bounded and uniformly quantised. Thus, for a given number of perfect matches in the quantised domain, the probability of real non-truncated representation vector elements of two images being "close" to each increases with decreasing quantisation bin size and increasing number of matches. On the assumption that the non-quantised representation space provides a good approximation of image semantics, this lemma suggest that a finer quantisation will improve the "quality" of similarity results *quadratically* with the bin size and *linearly* with respect to the number of matches.

The simplest extension of the binary scheme is a 2-bit quantiser with four bins for each coefficient:

$$V_{-2} = (-\infty, -\theta], V_{-1} = (-\theta, 0], V_1 = (0, \theta], V_2 = (\theta, \infty)$$

with

$$t_j \in V_k \Rightarrow q_j = k.$$

Through its parameter θ, the 2-bit quantiser allows to differentiate between large and larger coefficients, and offers the possibility of adjusting the relative bin sizes for different image collections. Figure 1 depicts the traditional single bit quantiser along with the proposed two bit quantiser and various choices of the control parameter θ.

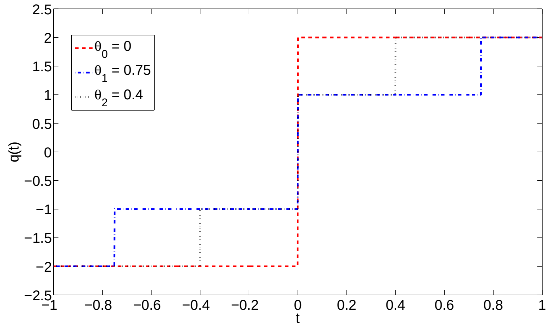

Fig. 1. 2-bit quantisers for control parameters (red) $\theta = \theta_0 = 0$, (blue) $\theta = \theta_1 = 0.75$ and (purple) $\theta = \theta_2 = 0.40$. Of note is that $\theta = 0$ reduces the 2-bit quantiser to a 1-bit one.

B. Indexing and Retrieval

The advantage of quantisation in the context of image retrieval and image matching is the ease with which images can be indexed and retrieved. Following truncation and 2-bit quantisation, images can now be represented as a finite set of labels, each being identified by the coefficient index and the quantisation level. There are thus a total of $4N$ labels where N is the number of non-truncated coefficients. Because only a fraction M of coefficients are retained, and each coefficient is quantised uniquely, each image gives rise to exactly M labels. We store for each label (combination of coefficient index and quantisation level) the list of images in which that particular coefficient has been quantised to that particular level. Let C be the set of images in the database and $\mathcal{L}$ the lexicon (set of all labels). The inverted index data structure is a mapping $\Phi : \mathcal{L} \mapsto \mathcal{P}(C)$ where $\mathcal{P}$ denotes the power set, with $\Phi(\lambda) = S$ if $\lambda \in Q(I_i)$ for all $I_i \in S$ (here $Q(\cdot)$ denotes any signature generator). Given a query signature, the inverted index allows fast subset selection as it circumvents the need to visit any non-matching database items. In particular, given a query image I_q with codewords $Q(I_q)$, we retrieve the union M over label-wise matches,

$$M(I_q) = \bigcup_{\lambda \in Q(I_q)} \Phi(\lambda).$$

The subset can then be ranked using a variety of scoring functions. We use the number of label matches as in [7].

C. Insights into Parameter Selection

The control parameter θ determines the relative widths of the four bins, for $\theta = 0$ and large θ, the 2-bit quantiser reduces to a 1-bit quantiser. In the context of compression and communication, quantiser performance is often measured in terms of the average rth-power distortion,

$\delta_r(Q_k|f(x)) = \mathbb{E}[||X - Q_k(X)||^r], r \in [1, \infty]$, where $f(x)$ denotes the probability density of the random variable X. For the purpose of retrieval, it seems preferable to choose any of the standard information retrieval measures, such as average precision or precision amongst the top k matches.

If class labels are absent, a surrogate measure may be obtained by measuring the average distance between the real-valued coefficients of the query image and those of the top J results (retrieved using the result of the 2-bit quantiser),

$$\mathcal{E}(\theta) = \frac{1}{J \times N} \sum_{i=1}^{N} \sum_{j=1}^{J} ||c(I_i) - c(I^j)||_2^2. \qquad (1)$$

We may approximate this empirical measure by

$$\mathbb{E}\{\mathcal{E}(\theta)\} = \int_{-\infty}^{\infty} \frac{1}{J} \sum_{j=1}^{J} ||\mathbf{x} - c(I_j)||_2^2 f(\mathbf{x}) d\mathbf{x}$$

where the expectation is taken over the query vector with probability density function $f(\mathbf{x})$. Assuming that the database is sufficiently large for the top J matches to have the same truncated vector quantisation as the query image, i.e. $q(c(I_j)) = q(c(I_q))$ for all $j = 1, \ldots, J$, and assuming further that all coefficients are i.i.d. according to some pdf, the following Lemma gives the expectation of interest.

Lemma 2 *Let* $V_{-2} = (-\infty, -\theta], V_{-1} = (-\theta, 0], V_1 = (0, \theta], V_2 = (\theta, \infty)$. *Then, the normalised expected error is given by*

$$\frac{1}{K}\mathbb{E}\{\mathcal{E}(\theta)|q(c(I_i)) = q(c(I_i^j)), \forall i, j\}$$
$$= \sum_{m \in \mathcal{M}} 2(\alpha_{m,0}(\theta)\alpha_{m,2}(\theta) - \alpha_{m,1}^2(\theta)) \qquad (2)$$

where

$$\alpha_{m,r}(\theta) = \int_{c_k(I_i) \in V_m} c_k^r(I_i) f(c_k(I_i)) dc_k(I_i)$$

with $r \in \{0, 1, 2\}$.

Proof: See Appendix B. ■

In essence, the above result attempts to provide an answer to the following question: Suppose you have two random vectors generated from i.i.d. processes. Given that they match perfectly after they are quantised to two bits with a given control parameter, what is the squared difference (on average) between the unquantised versions of these two random vectors?

TABLE I. Terms of interest along with their first and second derivatives

Parameter	First Derivative	Second Derivative
$\alpha_{-2,2}(\theta)$	$-\theta^2 f(-\theta)$	$-(2\theta f(-\theta) - \theta^2 f'(-\theta))$
$\alpha_{-1,2}(\theta)$	$\theta^2 f(-\theta)$	$2\theta f(-\theta) - \theta^2 f'(-\theta)$
$\alpha_{1,2}(\theta)$	$\theta^2 f(\theta)$	$2\theta f(\theta) + f'(\theta)\theta^2$
$\alpha_{2,2}(\theta)$	$-\theta^2 f(\theta)$	$-(2\theta f(\theta) + f'(\theta)\theta^2)$
$\alpha_{-2,1}(\theta)$	$\theta f(-\theta)$	$f(-\theta) - f'(-\theta)\theta$
$\alpha_{-1,1}(\theta)$	$-\theta f(-\theta)$	$-(f(-\theta) - f'(-\theta)\theta)$
$\alpha_{1,1}(\theta)$	$\theta f(\theta)$	$f(\theta) + f'(\theta)\theta$
$\alpha_{2,1}(\theta)$	$-\theta f(\theta)$	$-(f(\theta) + f'(\theta)\theta)$
$\alpha_{-2,0}(\theta)$	$-f(-\theta)$	$f'(-\theta)$
$\alpha_{-1,0}(\theta)$	$f(-\theta)$	$-f'(-\theta)$
$\alpha_{1,0}(\theta)$	$f(\theta)$	$f'(\theta)$
$\alpha_{2,0}(\theta)$	$-f(\theta)$	$-f'(\theta)$

For the given application at hand, *i.e.*, image similarity search, we would like to minimize the expected error. The optimal point, in the general case, can be found through numerical search algorithms guaranteeing the convergence to the global optimum value (if there is only a single minima) or to a local minimum in case there are several. Given the expected error expression, we are now in a position of calculating (at least numerically) the optimal control parameter as

$$\theta^\star = \arg\min_{\theta} \mathbb{E}\{\mathcal{E}(\theta)\}$$

under the rather restricted conditions we have considered. The optimal control parameter satisfies the following condition $\partial\mathbb{E}\{\mathcal{E}(\theta^\star)\}/\partial\theta = 0$ implying that

$$\sum_{m \in \mathcal{M}} (\alpha'_{m,0}(\theta^\star)\alpha_{m,2}(\theta^\star) + \alpha_{m,0}(\theta^\star)\alpha'_{m,2}(\theta^\star)$$
$$- 2\alpha_{m,1}(\theta^\star)\alpha'_{m,1}(\theta^\star)) = 0$$

where we define $\alpha'_{m,r}(\theta) = \partial\alpha_{m,r}(\theta)/\partial\theta$. Now, given an initial point θ_0, we can utilise the Newton-Raphson method to find a local (if the curve is (quasi)-convex, the global) minimum:

$$\theta_{t+1} = \theta_t - \left(\sum_{m \in \mathcal{M}} g_m(\theta_t)\right)^{-1} \sum_{m \in \mathcal{M}} h_m(\theta_t)$$

where

$$h_m(\theta_t) = \alpha'_{m,0}(\theta_t)\alpha_{m,2}(\theta_t) + \alpha_{m,0}(\theta_t)\alpha'_{m,2}(\theta_t)$$
$$- 2\alpha_{m,1}(\theta_t)\alpha'_{m,1}(\theta_t)$$

and

$$g_m(\theta_t) = \alpha''_{m,0}(\theta_t)\alpha_{m,2}(\theta_t) + \alpha'_{m,0}(\theta_t)\alpha'_{m,2}(\theta_t)$$
$$+ \alpha'_{m,0}(\theta_t)\alpha'_{m,2}(\theta_t) + \alpha_{m,0}(\theta_t)\alpha''_{m,2}(\theta_t)$$
$$- 2\alpha'_{m,1}(\theta_t)\alpha'_{m,1}(\theta_t) - 2\alpha_{m,1}(\theta_t)\alpha''_{m,1}(\theta_t)$$

with $g_m(x) = \partial h_m(x)/\partial x$ and $\alpha''_{m,r}(\theta) = \partial\alpha'_{m,r}(\theta)/\partial\theta$. The first and second derivatives of each term are given in Table I and derived in Appendix C. Newton iterations stop when $|\theta_{t+1} - \theta_t| \leq \epsilon$ for some small $\epsilon > 0$.

Moreover, although derived under strong assumptions, we show in later sections how good an approximation this provides of the empirical error.

Fig. 2. Example images from two categories ("bags" and "shoes") of the apparel collection.

IV. Experiments

This section 1) details the datasets utilised during simulations, specifically the MNIST and commercial apparel collection, 2) investigates the performance of the truncation level applied at the Haar transformation stage, and finally, 3) presents simulations to validate the insights developed in the previous sections by showing that the derived theoretical expression and the experimental optimal control parameter coincide almost fully, even under conditions violated by Lemma 2.

A. Datasets

The MNIST digits database, found at `http://yann.lecun.com/exdb/mnist`, is a set of 70,000 28×28 quasi-binary images of handwritten digits. It is commonly used for evaluating multi-class classifiers with state-of-the-art performance around 99.5% [17]. We scale the images to 32×32 before computing the Haar wavelet decomposition, leading to a vector of size 1024. The apparel collection contains 9,800 colour images from 25 different online retailers and 10 different categories: 'accessories', 'bags', 'dresses', 'jackets', 'jewellery', 'lingerie', 'shoes', 'skirts', 'tops' and 'trousers'. Examples for two classes are shown in Fig. 2. Images were cropped to the largest bounding square and resized to 32×32. Only the luminance data were used for wavelet decomposition.

B. Effect of Truncation

The first set of simulations seek to quantify the advantage of 2-bit over 1-bit quantisation, using two different measures: precision at 20 or PR20 (the number of relevant images retrieved amongst the top 20) and $\mathcal{E}(\theta)$ (the average difference between the query's non-quantised vector representation and the query's 20 closest matches). Note that the first measure directly quantifies the extent to which the system returns images of the same class as the query, while the second measure is independent of the images' class memberships. As well as varying θ between zero and the maximum coefficient value, we change the fraction of coefficients that are retained (truncation level) from 10% to 100%. The results are shown in Figures 3 and 4. Note that for the coefficient histograms, we display the distribution of the coefficients' *absolute* values; roughly 50% of the original values are negative. We make the following observations:

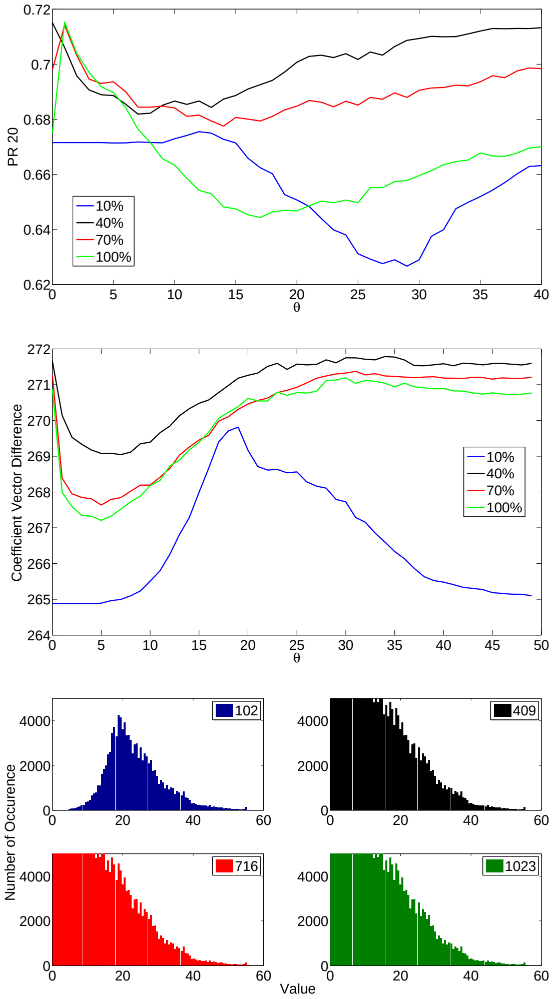

Fig. 3. MNIST: Effect of varying θ on PR20 (top) and $\mathcal{E}$ defined in Equation (1) (middle) for four different truncation levels. Bottom plot shows the coefficient histograms for each of the four truncation levels.

Observation 1. Performance for the top three truncation levels varies in very similar ways for both measures. Retrieval performance (PR20) peaks for smaller values of θ, and subsequently drops below the 1-bit quantiser baseline ($\theta = 0$) before recovering again for larger values. The percentage gain over 1-bit quantisation is 6% (MNIST) and 11% (Apparel) when all coefficients are retained (green curve) and somewhat smaller for the next two truncation levels. The average difference in coefficient values between query and its top matches, $\mathcal{E}(\theta)$, exhibits a steep drop

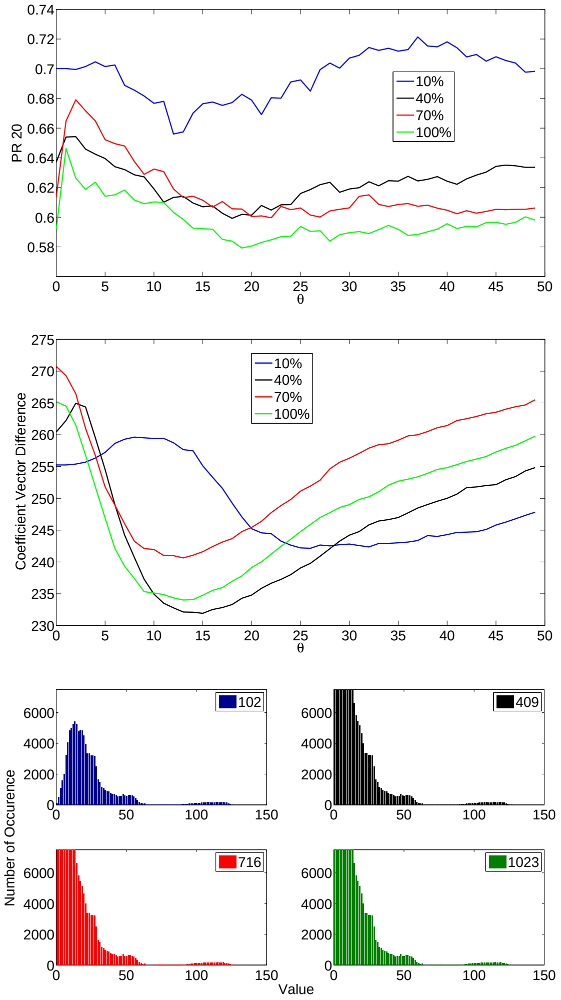

Fig. 4. Apparel: Effect of varying θ on PR20 (top) and $\mathcal{E}$ defined in Equation (1) (middle) for four different truncation levels. Bottom plot shows the coefficient histograms for each of the four truncation levels.

for $\theta \in [3, 7]$ (MNIST) and $\theta \in [10, 15]$ (Apparel) before climbing up to the 1-bit baseline. The histograms reveal that for these three truncation levels most of the coefficients are near zero with a shallow, albeit long tail. The long tail becomes relatively more pronounced as we retain progressively fewer coefficients, and the minima of $\mathcal{E}$ appear to mirror this right-shift of the probability mass.

Observation 2. Performance for extreme truncation levels (10%) varies with θ in quite a different manner. Although retrieval performance attains a small peak for non-zero values of θ for both datasets, it is much less marked than for the other three levels. Of particular interest is the qualitatively different behaviour for the $\mathcal{E}$ measure for the MNIST dataset with a minimum at $\theta = 0$ and a sharp rise for $\theta = 20$. At this level of truncation, 2-bit quantisation does not seem to add value.

Observation 3. Retaining only very few large magnitude coefficients seems beneficial for some collections and less for others: 10% truncation consistently outperforms all other truncation levels for all values of θ for the apparel collection, while it fares worst for the MNIST dataset.

Observation 4. Although image representation is diversified through multi-bit representation of wavelet coefficients, the standard inverse index structure does not incorporate the discrepancies between the features and only counts the exact matches to rank images (it is of importance to note that our algorithm outperformed the single-bit representation scheme despite this drawback). We believe that enriched inverted index structures based on soft assignment which are recently developed [18], could bring out more advantages of the technique proposed here.

C. Control Parameter Selection

To illustrate how a good value for θ could be chosen on the basis of Equation (1), we have fitted a two-component Gaussian density (it is shown that the mixture of Gaussians accurately represents both the modal and tail behavior of the wavelet coefficients [19]):

$$f(c;p,\mu_1,\sigma_1,\mu_2,\sigma_2) = pf(c;\mu_1,\sigma_1)+(1-p)f(c;\mu_2,\sigma_2)$$

where $f(\cdot,\mu,\sigma)$ denote the Gaussian density with mean μ and variance σ^2, to the coefficients of the digits dataset with no truncation. We utilise the EM (Expectation Maximization) to obtain the mixture model parameters [20].

As previously noted and shown in Fig. 5, the distribution of coefficients is roughly symmetric with respect to the origin and the Gaussian mixture model provide a very accurate representation both in the central and tail components. The best fitting Gaussian has $[p, \mu_1, \sigma_1, \mu_2, \sigma_2] = [0.7570, 0.0070, 0.5200, 0.1227, 17, 3660]$. We note that the shape of the theoretical $\mathbb{E}\{\mathcal{E}(\theta)\}$ is similar to the empirical curve displayed in Figure 3. Note, in particular, that the optimal value of θ, which is obtained through Newton-Raphson iterations, is very close to the empirical optimum.

V. Conclusions

We presented a multi-resolution and multi-bit image similarity search algorithm. Building on the work by Jacobs *et al.* [7], we show that it is possible to gain significant performance improvements, both in terms of representation vector difference and PR, by 1) simple extension to multi-bit (specifically two in order not to sacrifice the other advantages of quantisation, e.g. storage, computation speed, etc.) representation of wavelet coefficients and 2) carefully designing the proposed algorithm's

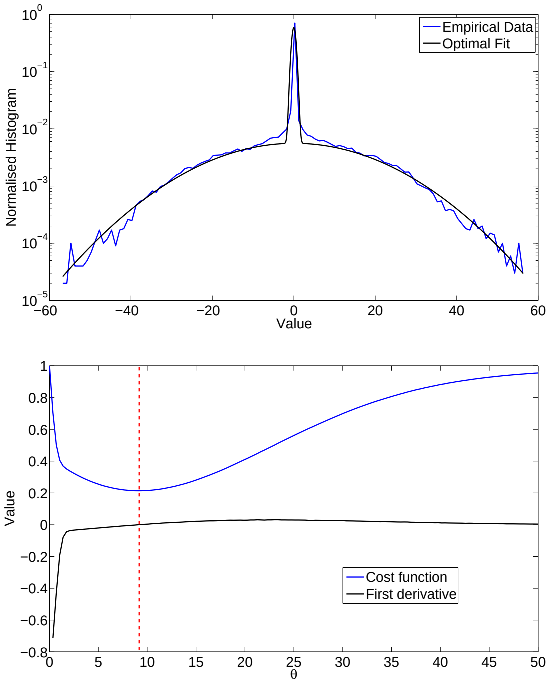

Fig. 5. (top:) The normalized histogram and the optimal two component Gaussian fit, and (below:) The theoretical cost function and its first derivative.

control parameter. Our current research efforts indicate that further improvements can be obtained utilising more sophisticated multi-resolution inverse index structures.

Appendix A - Proof of Lemma 1

The structure of the proof is as follows. First, we use a probabilistic bound to express the probability of interest in terms of an expectation. Then, we break the elements of the coefficient matrices into two disjoint sets: matching and non-matching elements. Finally, we upper-bound the terms appropriately for each set giving an upper-bound on the desired quantity. In the below, $|\cdot|$ denotes the absolute value operator for real-valued scalars and the cardinality for sets. Using the monotonicity of the function $f(x) = x^2$ for $x \geq 0$ and then applying Markov's inequality to the probability of interest yields

$$\begin{aligned}&\Pr\{||c(I)-c(J)||_2 \geq \epsilon||\mathcal{S}_q(I,J)| = m\}\\ &= \Pr\{||c(I)-c(J)||_2^2 \geq \epsilon^2||\mathcal{S}_q(I,J)| = m\}\\ &\leq \epsilon^{-2}\mathbb{E}\{||c(I)-c(J)||_2^2||\mathcal{S}_q(I,J)| = m\} \qquad (3)\end{aligned}$$

Now, using the definition of the ℓ_2 norm, equation (3) is decomposed into:

$$\begin{aligned}&\Pr\{||c(I)-c(J)||_2 \geq \epsilon||\mathcal{S}_q(I,J)| = m\}\\ &\leq \epsilon^{-2}\left(\mathbb{E}\left\{\sum_{k\in\mathcal{S}_q(I,J)}(c_k(I)-c_k(J))^2||\mathcal{S}_q(I,J)| = m\right\}\right.\\ &\quad\left.+\mathbb{E}\left\{\sum_{k\notin\mathcal{S}_q(I,J)}(c_k(I)-c_k(J))^2||\mathcal{S}_q(I,J)| = m\right\}\right).\end{aligned}$$

Now note that $|c_k(I) - c_k(J)| \leq \Delta$ for $k \in \mathcal{S}_q(I,J)$. Moreover, $|c_k(I) - c_k(J)| \leq 2U$ since $|c_{ij}| \leq U$ for all k. Noting that the summands are all bounded and positive (indicating that the expectation can be moved into the summations), and utilising the discussed bounds gives:

$$\begin{aligned}&\Pr\{||c(I)-c(J)||_2 \geq \epsilon||\mathcal{S}_q(I,J)| = m\}\\ &\leq \epsilon^{-2}\left(\sum_{k\in\mathcal{S}_q(I,J)}\mathbb{E}\left\{(c_k(I)-c_k(J))^2||\mathcal{S}_q(I,J)| = m\right\}\right.\\ &\quad\left.+\sum_{k\notin\mathcal{S}_q(I,J)}\mathbb{E}\left\{(c_k(I)-c_k(J))^2||\mathcal{S}_q(I,J)| = m\right\}\right)\\ &\leq \epsilon^{-2}\left(\sum_{k\in\mathcal{S}_q(I,J)}\Delta^2 + \sum_{k\notin\mathcal{S}_q(I,J)}4U^2\right)\\ &= \epsilon^{-2}\Delta^2[m + 4n^2(K-m)]\end{aligned}$$

where the last line follows from the fact that $|\overline{\mathcal{S}}_q(I,J)| = K - m$ where overline denotes the complement of its argument set. This concludes the proof of the claim.

Appendix B - Proof of Lemma 2

The proof assumes that the database is sufficiently large for the top J matches to a query to have all its coefficients quantised to the same value, i.e. $q(c(I_j)) = q(c(I_q))$ for all $j = 1, \ldots, J$ and that, without loss of generality, there is no truncation. We further assume that the coefficients are statistically independent between and within images.

Observe the following:

$$\mathbb{E}\{\mathcal{E}(\theta)\} = \frac{1}{NJ}\sum_{i=1}^{N}\sum_{j=1}^{J}\sum_{k=1}^{K}\mathbb{E}\left\{(c_k(I_i) - c_k(I_i^j))^2\right\}$$

which follows from the definition of the ℓ_2 norm and the fact that $[(c_k(I_i) - c_k(I_i^j))^2]$ is bounded and non-negative for all i, j, k. Given that $q(c(I_i)) = q(c(I_i^j))$ for all i, j, let us focus on each term in the summation above (to simplify the notation we utilise $x := c_k(I_i)$, $y := c_k(I_i^j)$, $q_k(c(I_i)) := w$ and $q_k(c(I_i)) := p$ to denote the corresponding random variables):

$$\begin{aligned}&\mathbb{E}\left\{(x-y)^2|w = p\right\}\\ &= \int_x\int_y (x-y)^2 f(x|w = p))f(y|w = p)dxdy\end{aligned}$$

where we used the fact that the coefficients for each image are independent and that $f(u,v|z) = f(u|z)f(v|z)$. Let us now focus on the pdf of interest:

$$f(x|w=p)) = \sum_{m\in\mathcal{M}} \alpha_{m,0} f(x|x\in V_m) = \sum_{m\in\mathcal{M}} \alpha_{m,0} g_m(x)$$

where the second equality follows since $w = m \Rightarrow x \in V_m$, and we define $g_m(x) = f(x)/\alpha_{m,0}$ if $x \in V_m$ and $g_m(x) = 0$ if $x \notin V_m$. Moreover we define $\mathcal{M} = \{-2,-1,1,2\}$ and

$$\alpha_{m,r}(\theta) = \Pr\{w = m\} = \int_{x\in V_m} x^r f(x)dx.$$

Substituting this information back into the above gives

$$\begin{aligned}
&\mathbb{E}\left\{(x-y)^2|w=p\right\} \\
&= \sum_{m\in\mathcal{M}} \alpha_{m,0}^2 \int_x\int_y (x-y)^2 g_m(x)g_m(y)dxdy \\
&= \sum_{m\in\mathcal{M}} \int_{x\in V_m}\int_{y\in V_m} (x-y)^2 f(x)f(y)dxdy \\
&:= \sum_{m\in\mathcal{M}} \Lambda_m(x,y;\theta)
\end{aligned}$$

Thus, the expectation of interest is given by

$$\mathbb{E}\{\mathcal{E}(\theta)|q(c(I_i) = q(c(I_i^j),\ \forall i,j\} = K\sum_{m\in\mathcal{M}} \Lambda_m(x,y;\theta)$$

since $f(x) = f(y)$ for all i,j,k. Let us consider the terms inside the summation:

$$\begin{aligned}
\Lambda_m(x,y;\theta) &= \int_{x\in V_m} x^2 f(x)dx \int_{y\in V_m} f(y)dy \\
&\quad + \int_{x\in V_m} f(x)dx \int_{y\in V_m} y^2 f(y)dy \\
&\quad - 2\int_{x\in V_m} xf(x)dx \int_{y\in V_m} yf(y)dy \\
&= 2\alpha_{m,0}(\theta)\alpha_{m,2}(\theta) - 2\alpha_{m,1}^2(\theta)
\end{aligned}$$

where the last line is due to the fact that x and y obey the same distribution and definitions given in the Lemma. Substituting this information into the summation concludes the proof.

Appendix C - Derivatives of Terms

We consider the derivatives of the terms of interest here. Consider first $\alpha_{-2,2}(\theta)$ as the derivation of the other terms follows similarly. Using Leibniz's rule [21], we obtain

$$\begin{aligned}
\alpha'_{-2,2}(\theta) &= \frac{\partial}{\partial\theta}\int_{-\infty}^{-\theta} c^2 f(c)dc \\
&= \int_{-\infty}^{-\theta} \frac{\partial c^2 f(c)}{\partial\theta}dc + \theta^2 f(-\theta)\frac{\partial}{\partial\theta}(-\theta) \\
&= -\theta^2 f(-\theta).
\end{aligned}$$

We also showed that $\alpha'_{-1,2}(\theta) = \theta^2 f(-\theta)$, $\alpha'_{1,2}(\theta) = \theta^2 f(\theta)$ and $\alpha'_{2,2}(\theta) = -\theta^2 f(\theta)$. The second derivatives are given by $\alpha''_{-2,2}(\theta) = -(2\theta f(-\theta) - f'(-\theta)\theta^2)$ where we used the product rule for derivatives.

References

[1] T. Hofmann, "Probabilistic latent semantic indexing," in *Proc Int'l SIG Information Retrieval (SIGIR)*, 1999.

[2] J. Tang and P. Lewis, "Non-negative matrix factorisation for object class discovery and image auto-annotation," in *Proc Int'l Conf Image and Video Retrieval*, 2008, pp. 105–112.

[3] C. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[4] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Proc Int'l Conf Computer Vision and Pattern Recognition*, 2007.

[5] P. Zezula, G. Amato, V. Dohnal, and M. Batko, *Similarity Search - The Metric Space Approach*, ser. Advances in Database Systems. Springer, 2006, vol. 32.

[6] J. Mamou, Y. Mass, M. Shmueli-Scheuer, and B. Sznajder, "A unified inverted index for effficient image and text retrieval," in *Proc Int'l SIG Information Retrieval (SIGIR)*, 2009.

[7] C. Jacobs, A. Finkelstein, and D. Salesin, "Fast multiresolution image querying," in *ACM SIGGRAPH*, Los Angeles, CA, USA, August 1995, pp. 277–286.

[8] S. G. Mallat, "A theory of multiresolution signal decomposition: The wavelet representation," *IEEE Trans Pattern Analysis and Machine Intelligence*, vol. 11, pp. 674–693, January 1989.

[9] M. N. Do and M. Vetterli, "Wavelet-based texture retrieval using generalized Gaussian density and Kullback-Leibler distance," *IEEE Trans Image Processing*, vol. 11, no. 2, pp. 146–158, February 2002.

[10] M. Do, S. Ayer, and M. Vetterli, "Invariant image retrieval using wavelet maxima moment," in *Proc Int'l Conf Visual Information Systems*, 1999, pp. 451–458.

[11] J. Z. Wang, G. Wiederhold, O. Firschein, and S. X. Wei, "Content-based image indexing and searching using Daubechies' wavelets," *Int'l Journal Digital Libraries*, vol. 1, pp. 311–328, 1997.

[12] S. Liapis and G. Tziritas, "Color and texture image retrieval using chromaticity histograms and wavelet frames," *IEEE Trans Multimedia*, vol. 6, no. 5, pp. 676–686, October 2004.

[13] W. Y. Ma and B. S. Manjunath, "A comparison of wavelet transform features for texture image annotation," in *IEEE Int'l Conf Image Processing*, 1995.

[14] P. W. Huang and S. K. Dai, "Image retrieval by texture similarity," *Pattern Recognition*, vol. 36, pp. 665–679, 2003.

[15] N. Suematsu, Y. Ishida, A. Hayashi, and T. Kanbara, "Region-based image retrieval using wavelet transform," in *Proc Int'l Workshop Database and Expert Systems*, 1999.

[16] R. DeVore, B. Jawerth, and B. Lucier, "Image compression through wavelet transfom coding," *IEEE Trans Information Theory*, vol. 38, no. 2, pp. 719–746, March 1992.

[17] M.-A. Ranzato, C. Poultney, S. Chopra, and Y. LeCun, "Efficient learning of sparse representations with an energy-based model," in *Proc Neural Information Processing Systems (NIPS)*, 2006.

[18] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Lost in quantization: Improving particular object retrieval in large scale image databases," in *Proc Int'l Conf Computer Vision and Pattern Recognition*, 2008.

[19] M. S. Crouse, R. D. Nowak, and R. G. Baraniuk, "Wavelet-based statistical signal processing using Hidden Markov Models," *IEEE Trans Signal Processing*, vol. 46, no. 4, April 1998.

[20] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal Royal Statistical Society, Series B*, vol. 39, no. 1, November 1977.

[21] O. Hijab, *Introduction to Calculus and Classical Analysis*. New York: Springer-Verlag, 1997.