Towards Compact Topical Descriptors

Ling-Yu Duan, Rongrong Ji, Jie Chen, Wen Gao
National Engineering Lab for Video Technology, Peking University, Beijing, China
{lingyu,rrji, cjie,wgao}@pku.edu.cn

Abstract

We introduce a Compact Topical Descriptor to learn a compact yet discriminative image signature from the reference image corpus. This descriptor is deployed over the well used bag-of-words image histogram, with two merits over the traditional topical features: First, we propose to directly control the topical sparsity to achieve the descriptor compactness. Second, we ensure the descriptor discriminability by minimizing the bag-of-words reconstruction errors during the topical histogram encoding. To this end, we have a generative viewpoint of the topical feature extraction, which is estimated as a sparse MAP estimation over the original bag-of-words. We learn such estimation by a bi-convex optimization, iterating between both hierarchical sparse coding from words to topical histograms and dictionary learning of the corresponding word-to-topic transform. Especially, supervised labels such as image ranking list can be also incorporated into our descriptor learning paradigm. We quantize our performance in both ImageNet10K and NUS-WIDE, with comparisons to bag-of-words, LDA, miniBoF, and Aggregated Local Descriptors. In practice, we also implement our descriptor for a low bit rate mobile visual search application, i.e. sending compact descriptors instead of the image to reduce the query delivery latency. Our descriptor has significantly outperformed the state-of-the-art compact descriptors by quantitative evaluations over 10 million reference images.

1. Introduction

Describing images is no doubt one of the fundamental challenges in computer vision. Besides its progressive efforts towards fully content understanding, the very recent works [1–4] have also focused on the descriptor compactness without serious loss of the descriptor discriminability, which has multidisciplinary benefits. For instance, in many mobile visual search systems like Google Goggles, it makes more sense to deliver a compact visual signature instead of the query image to reduce the query delivery latency, which is extremely worthwhile when facing a bandwidth-constraint wireless link. In image classification competitions like PASCAL VOC [7], the feature compactness is also desired to avoid the curse of dimensionality in classifier training. Motivations also come from large scale image/video classification endeavors [8,9], where a compact signature is very beneficial for scalable storing and fast accessing of millions of reference images or videos.

The descriptor compactness has been studied in the previous literature, for instance by reducing the local descriptor dimensions like PCA-SIFT [10], GLOH [11], SURF [12] and MSR descriptors [13], as well as by compressing the image-level signatures like miniBoF [15] and Aggregated Local Features [2]. With the ever growing mobile computing capability, recent works in mobile visual search [1,4,16] step forwarded to directly extract very compact descriptors at the mobile end to achieve a low bit rate query delivery. For instance, Chandrasekhar *et al.* [1] introduced a Compressed Histogram of Gradients (CHoG) descriptor, which adopts Huffman coding trees to compactly describe an interest point with approximately 60 bits. An alternative is to compress the bag-of-words histogram instead, as recently investigated in [4,16]. For instance, Chen *et al.* [16] proposed to encode position differences of non-zero bins in bag-of-words, which reportsing an $\sim 3KB$ code per image for a vocabulary with 1 million words. By using mobile contexts like GPS tags, Ji *et al.* [4] proposed a boosting based word selection to build location-adaptive vocabularies for mobile landmark search.

Inspiration. Extracting compact signatures given a large image corpus is *not alone*: In document retrieval [5,17–19], it is well admired that the distribution of document corpus plays a key role to design an optimal document signature. However, to the best of our knowledge, the distribution of image corpus is left unexploited in the existing works of compact image descriptors [1–3,10,13]. This contradiction inspires us to exploit the feasibility to achieve descriptor compactness from the statistics of image corpus. To that effect, previous works on supervised descriptor learning [20–23] cannot be scaled up well in this case since they

*This work was supported in part by the National Basic Research Program of China (2009CB320902), in part by grants from the National Science Foundation of China (60902057 and 61121002). Ling-Yu Duan is the corresponding author of this paper.

heavily depends on expensive image category labels. For instance, the work in [20,23] learn class-specific vocabularies from an initial vocabulary. Lazebnik *et al.* [21] built supervised visual codebook based on information loss minimization principle. Instead, in this paper, we propose to *unsupervised* learn compact descriptors from the image corpus distribution to guarantee our scalability (Section 2.2). The potential supervised labels, if any, can be treated as the optional additive (Section 2.4) to further boost our learning effectiveness.

The Approach. We propose a **C**ompact **T**opical **D**escriptor (**CTD**) paradigm to learn a data driven compact image signature. Similar to [24,25], we aim to learn a sparse abstraction from the well used bag-of-words image signature at the topical level. However, the sparsity of topical features can only be indirectly controlled by imposing a Dirichlet like sparse prior [27], or by adding an entropy like posterior regularization [18], which is due to the normalized mixture distribution constraints in traditional models like pLSA [26] or LDA [27].

Inspired by the works in sparse topical coding in document analysis [5], by avoiding the normalized distribution constraints in previous works [18,24–27], the direct control of the topical sparsity from a non-probabilistic perspective is feasible. Consequently, we can impose the topical sparsity into the topical feature extraction by adding a Lasso like ℓ_1-regularizer. In this sense, each visual word is maximally reconstructed by using a sparse linear combination of topics. And therefore, the image is described as a topical histogram built by assemble and truncate the topical reconstructions of individual words

In practice, we employ a log-Poisson loss with Laplace prior to model the discrete word distributions. We also learn a topic-to-word dictionary as the topical bases such that topics are still unigram distributions over the dictionary terms, similar to the traditional topic models [26,27]. Such relaxation enables us to deploy an efficient coordinate descent to learn both topic-to-word dictionary and topical features with a closed-form solution. Optionally, we can also incorporate supervised labels from the image corpus, such as ground truth ranking orders, to learn CTD in a supervised manner.

Benefits. Learning compact descriptors from the image corpus is of practical advantages. Our methodological explanations are two-fold:

(1) We seek an optimal tradeoff between the descriptor compactness and discriminability, from the perspective of image corpus distribution rather than from the single image statistics. In other words, we directly incorporate the inter-image similarity statistics to learn the topical features for individual images.

(2) Different from previous works in learning local descriptors [10,13], we prefer to extract compact image features *upon* the bag-of-words. Therefore, our descriptor is *data driven* and can be used to various local descriptors and codebooks [10–13].

Application. We further demonstrate the potentials of our topical descriptor in a *low bit rate* mobile visual search application, which is well advocated in the very recent works [1,3,4]. In this scenario, to ensure an efficient mobile query delivery, compact visual descriptors are directly extracted from the mobile end and sent through the wireless link. We have deployed our proposed descriptor on HTC DESIRE G7 smart phones, which is quantitatively tested on 10 million landmark images collected from five worldwide metropolitans. Our descriptor can maintain almost identical search accuracy comparing to a million scale bag-of-words histogram [36] with approximately 100-bit per image to achieve an astonishing 1:1000 compression rate, which significantly outperforms the state-of-the-art compact descriptors for mobile visual search [1,3].

Related Work. Besides the topical feature based image representation [24,25], our solution also relates to both sparse coding and nonnegative matrix factorization. The former learns a structured dictionary by sparse coding of image patches [28,29], typically with max or average pooling to incorporate spatial layouts [30]. The latter learns a document(image)-specific coding vector to reconstruct words extracted from a given document(image) [31]. Both schemes combine individual word-level sparse codes into an image signature, which therefore cannot well capture the inter-word dependency. Instead, we model such dependency by deriving topical features from words to topics, and then to images, hence can preserve such dependency at the topical level. Finally, while differing in methodology, works in compact global descriptors *e.g.* GIST [14] and compressed bag-of-words histograms *e.g.* miniBoF [15] Aggregated Local Features [2] also target at the image signature compactness, which will be quantitatively compared in Section 4 together with the topical features [24,25] like LDA .

2. Compact Topical Descriptor

Notations. We adopt a visual codebook $\mathbf{V} = \{V_1, ..., V_N\}$ to represent each given image I as a vector $\mathbf{w} = (w_1, ..., w_N)^T$, where each w_i ($i \in N$) represents the frequency of word i appears in I. ($w_i = 0$ denotes V_i does not appear). Our goal is to learn a topic-to-word dictionary $\mathbf{M} \in \mathbb{R}^{K \times N}$, where each row $\mathbf{M}_k$ is a topic basis, *i.e.* a unigram distribution over $\mathbf{V}$. For simplicity, we denote the nth column in $\mathbf{M}$ as $\mathbf{M}_{.n}$.

Main Idea. We aim to project an initial high-dimensional bag-of-words $\mathbf{w}$ into a sparse set of non-zero dimensions in $\mathbb{R}^K$ based on $\mathbf{M}$ as its abstraction. This basis projection is learnt from the image corpus by a hierarchical latent variable model (Figure 1 (a)), where $\mathbf{s}_n \in \mathbb{R}^K$ is the *word code* for word n, and $\mathbf{x}_i \in \mathbb{R}^K$ is the *image code* of ith

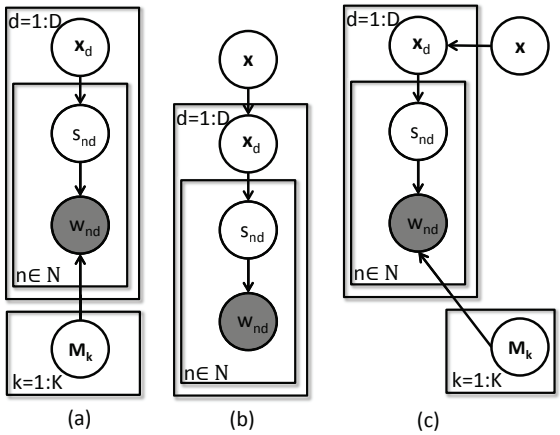

Figure 1. Difference explanation of our compact topical descriptor (a) versus other topical features like probabilistic Latent Semantic Analysis (pLSA) (b) and Latent Dirichlet Allocation (LDA) (c).

image as the targeted topical descriptor. In principle, this sparse latent representation is learnt for each $\mathbf{w}_i$ through a regularized loss minimization. We first revisit topic model from a probabilistic generative viewpoint proposed in [5,6]:

2.1. Generative Topical Description

For each image, we assume the word codes $\mathbf{S} = \{\mathbf{s}_n\}_{n=1}^N$ are conditionally independent given its image code $\mathbf{x}$, and also each observed word count w_n is independent given the corresponding word code $\mathbf{s}_n$, as in Figure 1 (a). Given a topic-to-word dictionary $\mathbf{M}$ initialized by sampling from a uniform distribution $\mathcal{P}$, this image is described generatively:

Phase 1 samples the image code $\mathbf{x}$ from a prior $p(\mathbf{x})$.

Phase 2 samples each word code $\mathbf{s}_n$ from a conditional distribution $p(\mathbf{s}_n|\mathbf{x})$ for each observed word V_n. Then, its word count w_n is again sampled from its Gaussian Distribution $Gaussian(w_n)$, which belongs to the expectation of a sufficient statistics of w_n and has its mean by $\mathbf{s}_n{}^T\mathbf{M}_{.n}$

$$\mathbb{E}_{p(w_n|\mathbf{s}_n,\mathbf{M})}\left[Gaussian(w_n)\right] \triangleq \mathbf{s}_n^T\mathbf{M}_{.n}. \tag{1}$$

In such a manner, the observed word count w_n is reconstructed using the linear combination of word code $\mathbf{s}_n$ and topic-to-word transform $\mathbf{M}_{.n}$ as $\mathbf{s}_n{}^T\mathbf{M}_{.n}$, which enjoys advantages such as non-negative, mean-valued describing power for Gaussian distribution.

Phase 3 aggregates individual word codes to output the image code, which depends on the choice of $p(\mathbf{x})$ and $p(\mathbf{s}_n|\mathbf{x})$, which is revisited in Section 2.4 for the case of supervised topical descriptor extraction.

2.2. Topical Sparsity by Sparse MAP Estimation

Based on the above probabilistic modeling, topics are derived by MAP estimation, during which we impose the sparsity constraint to achieve the descriptor compactness. For each image, the above generative process is interpreted as to infer the following joint distribution:

$$p(\mathbf{x}, \mathbf{S}, \mathbf{w}|\mathbf{M}) = p(\mathbf{x}) \prod_{n\in N, w_n\neq 0} p(\mathbf{s}_n|\mathbf{x}) p(w_n|\mathbf{s}_n, \mathbf{M}). \tag{2}$$

Similar to [5], a Poisson distribution is used to discretize the word observations in the right hand side of Equation 2:

$$p(w_n|\mathbf{s}_n, \mathbf{M}) \triangleq Poisson(w_n; \mathbf{s}_n^T\mathbf{M}_{.n}), \tag{3}$$

where $Poisson(x; y) = \frac{y^x e^{-y}}{x!}$ can be transferred into the KL-divergence of x and y after a log-Poisson operation.

To achieve the sparsity of both $\mathbf{x}$ and $\mathbf{S}$, following the setting of [32], we adopt a Laplace prior $p(\mathbf{x}) \propto exp(-\lambda\|\mathbf{x}\|_1)$, and $p(\mathbf{s}_n|\mathbf{x})$ is a supergaussian. We impose an ℓ_1-regularizer over $\mathbf{s}_n$ to directly control the word code sparsity[1]

$$p(\mathbf{s}_n|\mathbf{x}) \propto exp\left(-\gamma\|\mathbf{s}_n - \mathbf{x}\|_1 + \rho\|\mathbf{s}_n\|_1\right). \tag{4}$$

Learning from the Image Corpus. Subsequently, let $\mathbf{\Theta} = \{\mathbf{x}_i, \mathbf{S}_i, \mathbf{w}_i\}_{i=1}^I$ denote the codes for the entire image corpus containing $\{I_i\}_{i=1}^I$ images. We solve the problem of

$$\begin{aligned} &\min_{\mathbf{\Theta},\mathbf{M}} \sum_i \ell(\mathbf{s}_{in}, \mathbf{M}) + \lambda \sum_i \|\mathbf{x}_i\|_1 + \sum_{i, w_{in}\neq 0} (\gamma\|\mathbf{s}_{in} - \mathbf{x}_i\|_1 + \rho\|\mathbf{s}_{in}\|_1) \\ &s.t.: \ \forall i\ \mathbf{x}_i \geq 0;\ \forall n \in I_i,\ \mathbf{s}_{in} \geq 0;\ \forall k,\ \mathbf{M}_k \in \mathcal{P}. \end{aligned} \tag{5}$$

where ℓ is a log-Poisson loss function defined by

$$\ell(\mathbf{s}_{in}, \mathbf{M}) = -\log\ Poisson(w_{in}; \mathbf{s}_{in}^T\mathbf{M}_{.n}). \tag{6}$$

Minimizing Equation 6 is equivalent to minimizing an unnormalized KL-divergence between the observation w_{in} and the reconstruction $\mathbf{s}_n{}^T\mathbf{M}_{.n}$. We constraint both x_{in} and $\mathbf{s}_{in}$ as non-negative since w_{in} is non-negative, which can be interpreted as measuring the relative importance of topics.

2.3. Bi-Convex Optimization

The objective function of Equation 5 is bi-convex with a convex loss function ℓ (*e.g.*, log-loss of Gaussian distribution), with convex solution space. In other words, Equation 5 is convex over either Θ or $\mathbf{M}$ when fixing the other. This inspires us to solve Equation 5 using coordinate descent as in [31] that alternatively performs:

Sparse Topical Coding to optimize both image and word codes Θ when fixing $\mathbf{M}$. We solve the following minimization problem for each image respectively:

$$\min_{\mathbf{x},\mathbf{S}} \sum_{n\in N} \ell(\mathbf{s}_n, \mathbf{M}) + \lambda\|\mathbf{x}\|_1 + \sum_{n\in N} (\gamma\|\mathbf{s}_n - \mathbf{x}\|_1 + \rho\|\mathbf{s}_n\|_1), \tag{7}$$

with the constraint of $\mathbf{s}_n \geq 0$. Based on Equation 7, both image and word codes $\mathbf{x}$ and $\mathbf{S}$ are learnt through a coordinate descent with a closed-form solution, which alternates:

[1] We determine the hyper-parameters (λ, γ, ρ) via cross validation.

Learning Word Code $\mathbf{S}$: We first fix image code $\mathbf{x}$ to learn the word code $\mathbf{s}_n$ for each nth word, in which each topical basis dimension is independently estimated:

$$\hat{\mathbf{s}}_n = \arg\min_{\mathbf{s}_n} \ell(\mathbf{s}_n, \mathbf{M}) + \gamma\|\mathbf{s}_n - \mathbf{x}\|_2^2 + \rho \sum_{k=1}^{K} s_{nk}, \quad (8)$$

with the constraint of $\mathbf{s}_n \geq 0$. Based on Poisson distribution assumption, the corresponding solution is $s_{nk} = max(0, \hat{s}_{nk})$ by fixing s_{nj} for all the other $j \neq k$. Therefore, the following gradient is set as 0 to solve each s_{nk}

$$\nabla_{s_{nk}} \hat{\mathbf{s}}_n = \left(1 - \frac{w_n}{\mathbf{s}_n^T \mathbf{M}_{\cdot n}}\right) \mathbf{M}_{kn} + 2\gamma(s_{nk} - x_k) + \rho \quad (9)$$

Let $\mu = \sum_{j \neq k} s_{nj} \mathbf{M}_{jn}$ and $\tau = \mathbf{M}_{kn} + \rho - 2\gamma x_k$, solving Equation 9 is equal to solving $\hat{s}_{nk}$ for

$$2\gamma \mathbf{M}_{kn} \hat{s}_{nk}^2 + (2\gamma\mu + \mathbf{M}_{kn}\tau)\hat{s}_{nk} + \mu\tau - w_n x_{kn} = 0. \quad (10)$$

If $x_{kn} = 0$, we have $\hat{s}_{nk} = \theta = \frac{\rho}{2\gamma}$. Otherwise, we solve a quadratic equation to obtain $\hat{s}_{nk}$.

Learning Image Code $\mathbf{x}$: Then we fix $\mathbf{S}$ to learn each $\mathbf{x}$ for each image by solving

$$\min_{\mathbf{x}} \lambda\|\mathbf{x}\|_1 + \gamma\|\mathbf{s}_n - \mathbf{x}\|_2^2, \quad (11)$$

which is convex with the constraint of $\mathbf{x} \geq 0$. Since each dimension in $\mathbf{x}$ are independent, again we solve each x_k separately, such that

$$\forall k \;\; x_k = \max\left(0, \bar{S}_k - \frac{\lambda}{2\gamma N_{nonzero}}\right), \quad (12)$$

where $\bar{S}_k = \frac{1}{N_{nonzero}} \sum_{n \in N_{nonzero}} s_{nk}$. In other words, we use an ℓ_1-regularizer to have a truncated average over $N_{nonzero}$ individual word codes to obtain $\mathbf{x}$ for this image.

Learning Topic-to-Word Dictionary. Alternated, we fix both image and word codes $(\mathbf{x}, \mathbf{S})$ of individual images in the image corpus, then update the topic-to-word dictionary (transform) $\mathbf{M}$ by minimizing the log-Poisson loss. We solve this convex problem via the well-used projected gradient descent [33], where the projection to the simplex $\mathcal{P}$ can be performed with a linear algorithm as detailed in [33]. Algorithm 1 outlines the proposed bi-convex learning procedure.

2.4. Supervised Descriptor Learning

Optionally, we extend our topical descriptor learning into a *supervised* scenario. In this case, we also assume there are tags (side-information) available for a portion of the reference images[2]. Let $\mathbf{y} \in \mathbb{R}^L$ be the available tags with a set of L unique labels. In principle, we leverage $\mathbf{y}$ to supervise our dictionary learning by forcing co-labeled images as similar as possible.

[2] This is a commonsense scenario such as GPS or base station tags in mobile visual search systems [3,4].

Algorithm 1: The bi-convex optimization procedure of our compact topical descriptor (codes with light color correspond to supervised descriptor learning).

Input: Visual vocabulary $\mathbf{V} = \{V_1, ..., V_N\}$, image corpus $\mathbf{I} = \{I_i\}_{i=1}^I$,
Output: Image codes $\{\mathbf{w}_i\}_{i=1}^I$.
Initialize a topic-to-word transform $\mathbf{M}$ by sampling from the uniform distribution $\mathcal{P}$.
Hierarchical Sparse Coding by fixing $\mathbf{M}$:
for *each* $I_i \in \mathbf{I}$ **do**
  Learning word codes $\mathbf{S}_i = \{\mathbf{s}_{in}\}_{n=1}^N$ for I_i by solving Equation 10.
  Learning image code $\mathbf{x}_i$ by solving Equation 12 with truncated averaging.
end
Dictionary Learning by projective gradient descendent [33] based on fixing $\mathbf{\Theta}$ for $\mathbf{I}$.
Update supervised weighting vector $\mathbf{f}$ over image codes by solving Equation 13.

Modeling Ranking Loss. We formulate each image code $\mathbf{x}$ as input features for a retrieval system. The similarity is measured using an L2 distance. Note that others like Cosine distance can be also adopted. Given the training images with labels $\mathcal{I}' = \{(\mathbf{w}_i, y_i)\}_{i=1}^{I'}$, we learn a weighting vector $\mathbf{f} \in \mathbb{R}^K$ for $\mathbf{x}$ similar to the inverted document frequency [34]. Then, an optimal distance should have the following ranking loss over the labeled reference image set:

$$\begin{aligned} &\arg\min_{\mathbf{f}} Loss(\mathbf{f}, \mathbf{y}) = \\ &\arg\min_{\mathbf{f}} \sum_{i \in I'} \sum_{j \in I'} (y_i \oplus y_j) \mathbf{f}^T (\mathbf{x}_i - \mathbf{x}_j) \end{aligned} \quad (13)$$

We jointly learn a weighting vector $\mathbf{f}$ for the image corpus, learn a topic-to-word transform matrix (dictionary) $\mathbf{M}$, and learn the latent representations $\mathbf{\Theta} = \{\mathbf{x}_i, \mathbf{S}_i\}_{i=1}^{I'}$. To that effect, we rewrite the joint optimization as:

$$\begin{aligned} &\min_{\mathbf{\Theta}, \mathbf{M}, \mathbf{f}} f(\mathbf{\Theta}, \mathbf{M}) + CLoss(\mathbf{f}, \mathbf{y}) + \frac{1}{2}\|\mathbf{f}\|_2^2 \\ &s.t. : \; \forall i \; \mathbf{x}_i \geq 0, \; \forall n \in I_i \; \mathbf{s}_{in} \geq 0, \; \forall k, \; \mathbf{M}_k \in \mathcal{P}, \end{aligned} \quad (14)$$

where $f(\mathbf{\Theta}, \mathbf{M})$ refers to the loss in Equation 5, C is a positive constant.

Learning as Linear Regression. We solve Equation 14 by projective coordinate descent [33], with slight changes on solving for $\mathbf{x}$ by adding an additional step to learn $\mathbf{f}$. The learning of each $\mathbf{x}$ is again optimized by using Equation 12. The learning of $\mathbf{f}$ is achieved with the loss in Equation 13.

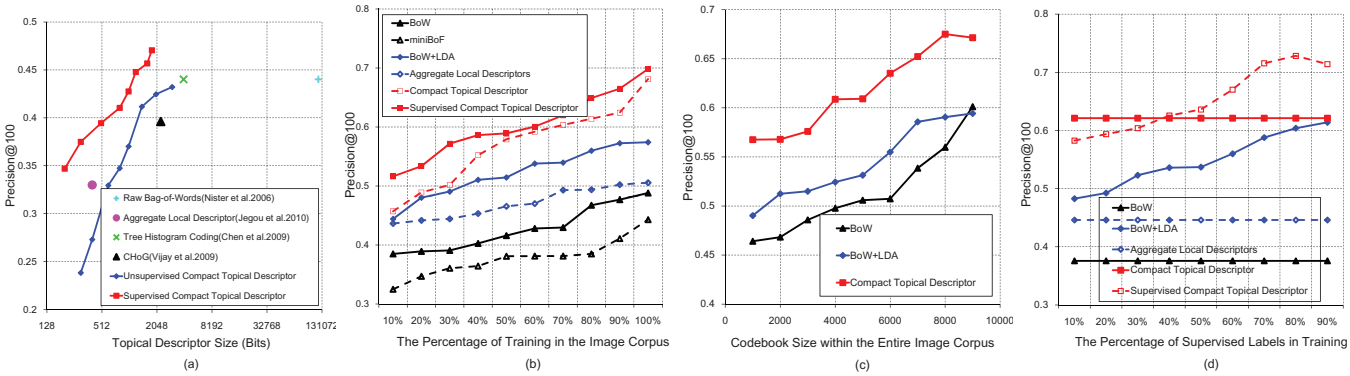

Figure 2. Detailed experimental analysis on the NUS-WIDE benchmark. (a). joint tuning of both γ and ρ; (b). the percentage of images to build CTD; (c). the tuning of codebook sizes; (d). the mAP improvements by adding supervised labels.

Then, for each image, we output its topical descriptor $\mathbf{S}$ by shifting the mean of each bin into

$$\bar{S}_k = \frac{1}{|N_{nonzero}|}\sum_{n=1}^{N_{nonzero}} s_{nk} + \frac{C}{2I|N_{nonzero}|\lambda}(f_{yk} - f_{\hat{y}k}) \quad (15)$$

where $\hat{y} = \arg\min_y(\triangle\ell(y,\hat{y}) + F(y,\mathbf{x}))$ is the prediction loss, such as their category label violation as in Equation 13, or ranking distortion as in [4]. The light color codes in Algorithm 1 outline our supervised topical descriptor learning.

3. Further Discussions

Our compact topical descriptor differs from the traditional topic models like pLSA or LDA [25,27]. Typically their methods assumes a Dirichlet prior to generate the topics of all the words, which are assumed to be from a given document-specific mixing proportion. And each word occurence in the document is supposed to be generated from a group of related topic. To this end, LDA based topic models do not have an explicit definition of the word code as in this paper, and therefore it is hard to directly control its code sparsity. Although the sparsity can be controlled by adjusting the Dirichlet prior indirectly, we will quantitative show in Section 4 that such controlling is less effective comparing with our proposed solution in this paper.

It is worth to mention that, methods such as Sparse Coding [28–30] and Non-negative Matrix Factorization (NMF) [31] also try to learn a hierarchical representation from the image (or document) corpus. by treating their learnt bases as "topics", aforementioned schemes are similar to the generative hierarchical topical generation schemes as in [5,6] and our methods. However, one fundamental difference is that our descriptor only encodes non-zero words. On the contrary, both sparse coding and NMF encode all words in the vocabulary, which are therefore less efficient in terms of coding speed and less compact in terms of descriptor size.

4. Experiments

In Section 4.1, we provide quantitative evaluations of our CTD feature over alternatives and state-of-the-art descriptors in both ImageNet10K and NUS-WIDE benchmarks, including BoW features [35,36], LDA [25] based topical feature, miniBoW [15], and Aggregate Local Descriptors [2]. In Section 4.2, we further demonstrate the real-world usage of our descriptor in a low bit rate mobile visual search system deployed on HTC DESIRE G7 smart phones, tested in 10 million reference images with comparisons to the state-of-the-art compact descriptors for mobile visual search, including CHoG [1,3], Tree Histogram Coding [16] and GIST [14].

4.1. Comparisons on Image Search Benchmarks

ImageNet10K. We first evaluate our CTD feature for image search tasks in ImageNet10K[3]. The original ImageNet contains more than 10,000,000 images in over 10,000 categories organized by the WordNet hierarchy. We choose the ImageNet10K, which contains 10,184 categories from the Fall 2009 release of ImageNet, including both internal and leaf nodes with more than 200 images each (a total of 9 million images). We use precision@100 for each query to evaluate the performance.

NUS-WIDE. The second dataset is NUS-WIDE, which contains around 270,000 web images associated with 81 ground truth concept tags[4]. Each image in NUS-WIDE contains multiple tags, which are treated as ground truth labeling to validate the image search performance. For evaluation, we consider 21 most frequent tags, such as "animal", "buildings", "person", *etc.*, each of which has abundant relevant images ranging from 5,000 to 30,000. We sample uniformly 100 images from each of the selected 21 tags to form a query set of 2,100 images with the rest as the training set. We use the average precision of these 100 images (average

[3] www.image-net.org/

[4] http://lms.comp.nus.edu.sg/research/NUS-WIDE.htm

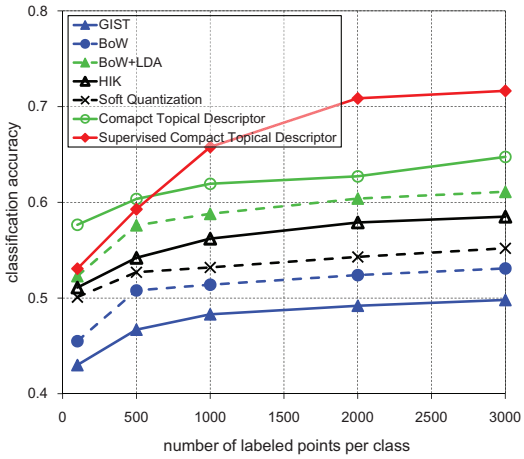

Figure 3. Performance figure on ImageNet10K dataset.

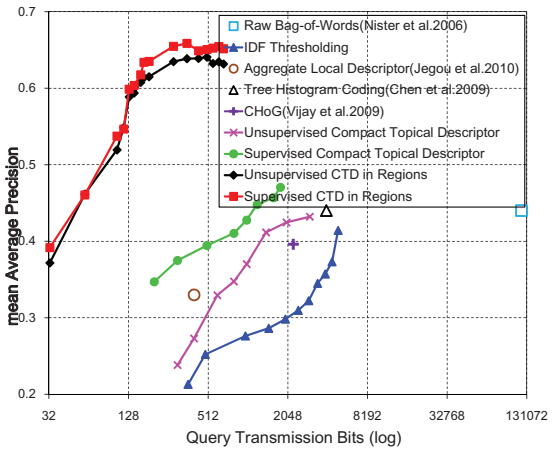

Figure 4. Compression rate vs. ranking distortion comparing to [1,2,16] using our ground truth query set.

classification accuracy) to evaluate the performance.

Baselines. (1) *GIST* represents each image by a single GIST [14] descriptor (a commonly accepted baseline descriptor for scene classification). (2) *BoW* represents each image by a 0.1 million dimensional BoW via hierarchical k-means clustering [36]. (3) *BoW+LDA* follows the setting of [25], where we adopt the LDA model [27] to compress the initial BoW into a topical feature. (4) *BoW Compression* compares our performance to three state-of-the-art works in compressed BoW histograms, including miniBoW [15] and Aggregate Local Descriptors [2]. All baselines adopt a kNN based search with L2 distance to find the top returning, running on a workstation with 2.53 GHz Intel Xeon CPU and 20GB RAM and costing less than 1 second for every query.

Performances. In both datasets, our CTD signature outperforms all state-of-the-art alternatives [2,14,15,25,36]. Note that the x-axis denotes the log-scale descriptor size per image, from which we can see that our CTD works quite well in the descriptor compactness. In addition, supervised labels helps a lot in *ImageNet10K*. However, the performance gain of supervised CTD is limited in *NUS-WIDE*, which is due to the fact that only a small portion of the dataset is manually labeled. As a result, the associated tags are very noisy.

Insights. Figure 2 provides the parameter tuning details as well as further insights of our supervised and unsupervised CTD in *NUS-WIDE* Subfigure (a) shows the joint tuning of both γ and ρ in hierarchical sparse coding in terms of precision@100. Subfigure (b) shows the percentage of images for CTD learning and testing, from which we can see our CTD tends to be stable with the increasing of the training scale. Subfigure (c) shows the tuning of codebook sizes, where the performance of our topical descriptor converges at 8,000 words, which means it already well captures the semantic distribution of the image corpus. Subfigure (d) shows the performance variations with how many percentages of supervised labels. As analyzed above, it has limited improvement by increasing the training scale due to the imprecise labels in *NUS-WIDE*.

4.2. Low Bit Rate Mobile Visual Search

Work Flow. We have applied our compact topical descriptor for the application of low bit rate mobile visual search. In this scenario, the descriptor is expected to be compact, discriminative, and meanwhile computationally efficient to reduce the overall query delivery latency. We take mobile location search for example, where each reference image contains a GPS tag and the delivered query contains both the compact visual signature and a coarse side information *e.g.* base station identity. The work flow of our descriptor for this application consists of four phases:

Phase 1 is a "region selection" operation in the mobile end. Its input can be mobile side information like base station tags that is directly available at the mobile end, which helps to locate the current query to one of the geographical regions in a given city[5]. *Phase 2* is to extract local features, quantize them into bag-of-words, and map into a topical descriptor which is binarized into an occurrence (hit/non-hit) histogram with Huffman coding to further reduce size. *Phase 3* is to transmit the encoded signature (with the base station identity) over the wireless link to a remote server. *Phase 4* decodes the topical descriptor in this server into the original bag-of-words, which is then combined with the region-specific **f** to search similar images.

Data Collection. We collect over 10 million geo-tagged photos from photo sharing websites of Flickr and Panoramio. We crawled photos from five worldwide metropolitans including *Beijing*, *New York City*, *Barcelona*, *Singapore* and *Florence*. This dataset is named as *10M*

[5]We do not use this location tag to post or pre-filter reference images, so as to make a fair comparison in Figure 4,

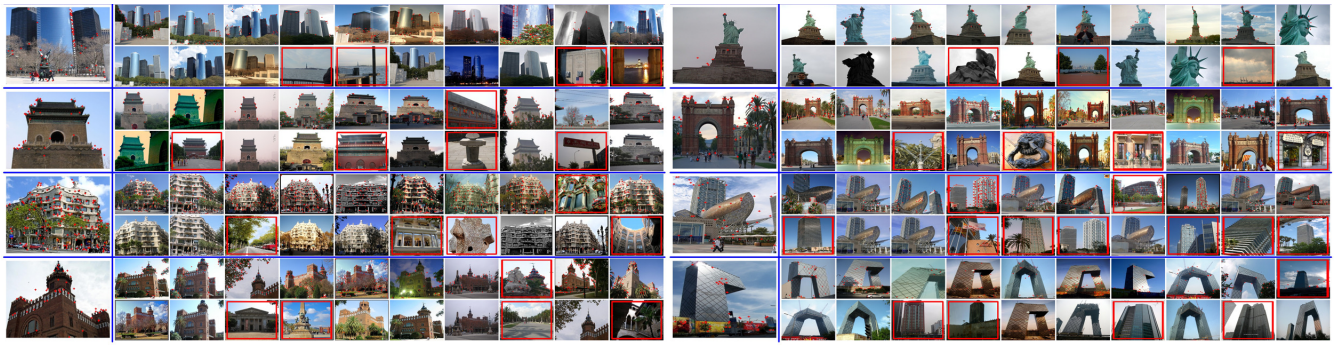

Figure 5. Search examples using CTD (first row) and CHoG (second row) in *10M Landmarks Photos* dataset, left is the query image. We visualize the non-zero local descriptor responses to the non-zero topical bins, showing that our topical descriptor only locates on discriminative landmark regions.

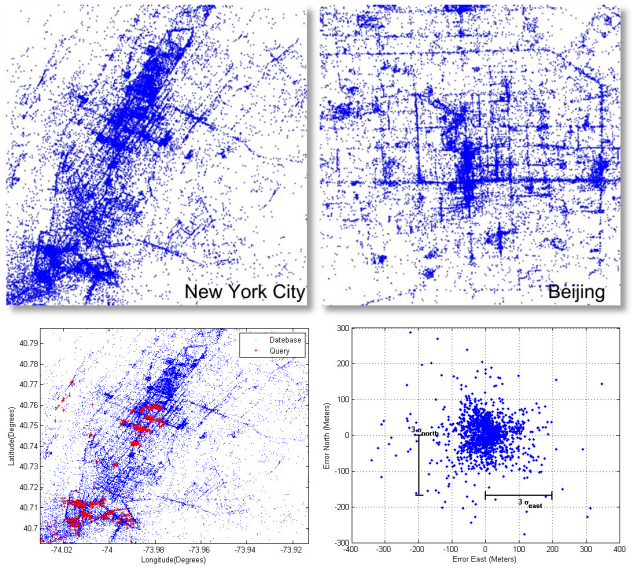

Figure 6. Clockwise: geographical distribution of reference images in New York and Beijing (top); geographical distortions of the returning images aligned with the ground truth locations: *red dot in the center* (down right); and the query (red) sampling in New York City (down left).

Landmark Photos. We run k-means clustering by using the GPS tags to partition photos of each city into multiple regions. For each city, we select the top 30 densest regions as well as 30 random regions. We then ask a group of volunteer to identify one or more dominant landmark views from each of these 60 regions. For a given dominant view, all their near-duplicate photos are manually labeled in its belonged region and nearby regions. Eventually we have 300 queries as well as their ground truth labels (corrected matches). Figure 6 further shows the geographical distribution of the collected reference images, the selected queries, and the search accuracy on geographical map.

Evaluation Protocol. We use *mAP* to evaluate the search performance, which is widely used in [1–4,35].

Table 1. Memory and time at HTC Desire G7.

Tree	Memory	Time
SIFT [37] SVT, H = 6, B = 10	59MB	1.361S
CHoG [1,3] SVT, H = 6, B = 10	24MB	1.624S
Tree Histogram Coding [16]	60MB	1.453s
Compact Topical Descriptor	68MB	1.974S

$$mAP = \frac{1}{K}\sum_{k=1}^{K}\frac{\sum_{r=1}^{N^k_{relevant}} P(r)}{N^k_{relevant}}. \quad (16)$$

where $k = 1$ to K is the total queries for evaluation. $N^k_{relevant}$ is the number of relevant documents to the kth query; r is the the rth relevant document; $P(r)$ is the precision at the cut-off rank of document r.

Rate Distortion Analysis. Figure 4 gives the rate distortion comparisons to [1–3,16], where the compression rate refers to the descriptor lengths and the search distortion refers to the *mAP* drops. Figure 4 shows our descriptor achieves the best tradeoff in the rate distortion: We report the highest compression rate (1:1000) with a very limited distortion (viewing Figure 4 horizontally), as well as the highest ranking mAP (significantly over [1–3,16]) with comparable compression rate (viewing Figure 4 vertically). In addition, without supervised descriptor learning, CTD still outperforms all state-of-the-art alternatives [1–3,16].

Efficiency Analysis. We deploy our low bit rate mobile visual search prototype on HTC Desire G7, which is equipped with an embedded camera with maximal 2592 × 1944 resolution, a Qualcomm MSM7201A processor at 528MHz, a 512M ROM + 576M RAM memory, 8G extended storage and an embedded GPS. Figure 7 shows a snapshot of our compact topical descriptor based mobile visual search system. Table 1 further shows the memory and time cost at the mobile end with comparisons to state-of-the-arts in [1,3,16]. The most time-consuming part is local feature extraction, which can be further accelerated by random sampling, instead of detecting interest points.

Case Study. We discovered that some empirical queries in our system happen to be taken at night. And some queries

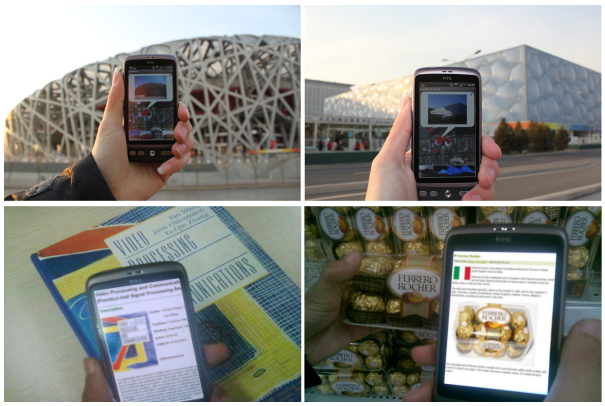

Figure 7. A snapshot of our low bit rate mobile visual search system deployed on HTC DESIRE G7 smart phone.

occur in different scales (from either nearby views or distant views). There is also a common problem that queries are with blurs (very common cases). We also selected some suboptimal queries with partial occlusions (objects or persons), as well as photos of partial landmark views. Figure 5 further shows the performances of our descriptor and the alternative of CHoG [3] in the abovel mobile query scenarios in New York, Beijing, Barcelona and Florence respectively. It is obvious that our descriptor performs much better, highly discriminative between foreground and background words by visualizing non-zero words after decoding, such that the selected topics only focus on the most discriminative foreground objects.

5. Conclusion

We introduce Compact Topical Descriptor (CTD), a non-probabilistic topical description learnt from the image corpus, which enables the direct control of topical sparsity to extract compact image signatures. Through relaxing the normalized constraints made in the traditional topical features, we perform a iterated bi-convex optimization including word-topic-image hierarchical sparse coding and topic-to-word dictionary learning. We further enable optional integration of supervised learning from side information such as partial category labels to improve its discriminability. Our proposed descriptor performs superior over the state-of-the-arts [2,14,15,25,36] on both ImageNet and NUS-WIDE. We also deploy our descriptor in a low bit rate mobile search system to provide location recognition in five metropolitans. We significantly outperform the state-of-the-art alternatives [1,3,16] by evaluating within 10 million reference landmark images.

References

[1] V. Chandrasekhar, G. Takacs, D. Chen, et. al. CHoG: Compressed histogram of gradients a low bit-rate feature descriptor. *CVPR*. 2009.

[2] H. Jegou, M. Douze, C. Schmid, P. Perez. Aggregating local descriptors into a compact image representation. *CVPR*. 2010.

[3] D. Chen, G. Baatz, K. Koser, *et al.* City-scale landmark identification on mobile devices. *CVPR*, 2011.

[4] R. Ji, L.-Y. Duan, J. Chen, H. Yao, J. Yuan, Y. Rui, and W. Gao. Location discriminative vocabulary coding for mobile landmark search. *IJCV*. 2011.

[5] Jun Zhu and Eric P. Xing. Sparse Topical Coding. *UAI*, 2011.

[6] Jun Zhu and Eric P. Xing. Conditional Topic Random Fields *ICML*, 2010.

[7] pascallin.ecs.soton.ac.uk/challenges/VOC/.

[8] www.image-net.org/.

[9] trecvid.nist.gov/.

[10] Y. Ke and R. Sukthankar. PCA-SIFT: A more distinctive rep. for local image descriptors. *CVPR*. 2004.

[11] K. Mikolajczyk and C. Schmid. Performance evaluation of local descriptors. *PAMI*, 2005.

[12] H. Bay, T. Tuytelaars, and L. Van Gool. SURF: Speeded up robust features. *ECCV*. 2006.

[13] G. Hua, M. Brown, and S. Winder. Discriminant embedding for local image descriptors. *ICCV*, 2007.

[14] Y. Weiss, A. Torralba, and R. Fergus. Spectral hashing. *NIPS*, 2008.

[15] H. Jgou, M. Douze and C. Schmid. Packing bag-of-features *ICCV*, 2009.

[16] D. Chen, S. Tsai, *et al.* Tree histogram coding for mobile image matching. *DCC*. 2009.

[17] S. Brody and M. Lapata. Bayesian word sense induction. *ACL*, 2009.

[18] M. Shashanka, B. Raj, and P. Smaragdis. Sparse overcomplete latent variable decomposition of counts data. *NIPS*, 2007.

[19] C. Wang and D. Blei. Decoupling sparsity and smoothness in the discrete hierarchical dirichlet process. *NIPS*, 2009.

[20] F. Perronnin, C. Dance, G. Csurka, and M. Bressan. Adapted vocabularies for generic visual category. *ECCV*, 2006.

[21] S. Lazebnik and M. Raginsky. Supervised learning of quantizer codebooks by information loss minimization. *PAMI*, 2009.

[22] J. Liu, Y. Yang, and M. Shah. Learning semantic visual vocabularies using diffusion distance. *CVPR*, 2009.

[23] J. Winn, A. Criminisi, and T. Minka. Object categorization by learned universal visual dictionary. *ICCV*, 2005.

[24] F.-F. Li and P. Perona. A Bayesian hierarchical model for learning natural scene categories. *CVPR*, 2005.

[25] A. Bosch, A. Zisserman, and X. Munoz. Scene classification using a hbrid generative/discriminative approach. *PAMI*, 2008.

[26] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 2001.

[27] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation, *JMLR*, 2003.

[28] S. Bengio, F. Pereira, Y. Singer, and D. Strelow. Group sparse coding. *NIPS*, 2009.

[29] R. Jenatton, J. Mairal, G. Obozinski, and F. Bach. Proximal methods for sparse hierarchical dictionary learning. *ICML*, 2010.

[30] J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. *CVPR*, 2009.

[31] D. Lee and H. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 1999.

[32] A. Hyvarinen. Sparse code shrinkage: Denoising of nongaussian data by maximum likelihood estimation. *Neural Computation*, 1999.

[33] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra. Efficient projections onto the L1-ball for learning in high dimensions. *ICML*, 2008.

[34] G. Salton *et al.* Term-weighting approaches in text retrieval. *Information Processing and Management*. 1988.

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

[36] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. *CVPR*, 2006.

[37] D. G. Lowe. Distinctive image features form scale-invariant keypoints. *IJCV*, 2004.