2-2014

Visual typo correction by collocative optimization: A case study on merchandize images

Xiao-Yong WEI

Zhen-Qun YANG

Chong-wah NGO
Singapore Management University, cwngo@smu.edu.sg

Wei ZHANG

Citation

1

Visual Typo Correction by Collocative Optimization: A Case Study on Merchandize Images

Xiao-Yong Wei, *Member, IEEE*, Zhen-Qun Yang, Chong-Wah Ngo, *Member, IEEE*, and Wei Zhang

***Abstract*— Near-duplicate retrieval (NDR) in merchandize images is of great importance to a lot of online applications on e-Commerce websites. In those applications where the requirement of response time is critical, however, the conventional techniques developed for a general purpose NDR are limited, because expensive post-processing like spatial verification or hashing is usually employed to compromise the quantization errors among the visual words used for the images. In this paper, we argue that most of the errors are introduced because of the quantization process where the visual words are considered individually, which has ignored the contextual relations among words. We propose a "spelling or phrase correction" like process for NDR, which extends the concept of collocations to visual domain for modeling the contextual relations. Binary quadratic programming is used to enforce the contextual consistency of words selected for an image, so that the errors (typos) are eliminated and the quality of the quantization process is improved. The experimental results show that the proposed method can improve the efficiency of NDR by reducing vocabulary size by 1000% times, and under the scenario of merchandize image NDR, the expensive local interest point feature used in conventional approaches can be replaced by color-moment feature, which reduces the time cost by 9202% while maintaining comparable performance to the state-of-the-art methods.**

***Index Terms*— Near-duplicate retrieval, visual word quantization, binary quadratic programming.**

I. Introduction

GIVEN the overwhelming number of online business websites nowadays, merchandize imagery, as a practical means of product promotion, is forming one of the most rapidly increasing set among web images. Many merchandize images, however, are indeed exact duplicates or near-duplicates (i.e., the images that share the same objects/scenes but appear differently due to the variations in capturing conditions, acquisition times, rendering conditions, or editing operations). For example, according to our study on a dataset including over one million merchandize images crawled from eBay.com and Taobao.com,[1] more than 12.9% of those images are near-duplicates, and together with exact duplicates, the proportion of the duplicate images is up to 29.7%. Therefore, searching near-duplicates in merchandize images is of great commercial value for the E-Commerce industry. It can help to identify pirated products, especially when they are sold at websites of different languages where the text-based retrieval may work inadequately; similarly, it can be used to detect unlicensed sellers, to monitor how a product is distributed among websites or countries through Internet, or even to find potential clients who are interested in a particular product. As shown in Fig. 1, near-duplicate images, once detected, can provide additional valuable clues such as price/comment comparison across websites or countries, and evaluation of the market potentials for products.

To conduct near-duplicate retrieval (NDR) in merchandize images, one can first think of employing general purpose NDR techniques, which have reached a level of sophistication. After a decade of intensive study, it is generally recognized that the use of the local features based on local interest points (LIPs) within a bag-of-word (BoW) framework [1]–[5] performs the best for NDR. LIPs are salient points of images whose features are invariant to the changes of scale, illumination, and viewpoint [6], [7]. With LIPs detected for each image, the idea of BoW is borrowed from the text retrieval, in the way that an image is considered a document with its LIPs as "words." On this basis, images can thus be indexed by LIPs so as to be retrieved with the advanced techniques as developed in text retrieval.

The LIPs+BoW framework is simple and straightforward, and due to its solid foundation in information retrieval, it has demonstrated promising performance in a wide range of applications such as copyright violation detection [8]–[10], object retrieval [4], scene recognition [11], and video summarization [12]. However, when facing large-scale dataset and online applications, the time efficiency of the framework is seriously below the expectations. First, the extraction of LIPs is inherently time-consuming, because to select candidate points, the process has to conduct a great amount of tests for every pixel of an image. Second, a point-to-point matching has usually been employed when determining the similarity between two images. The situation could be worse when

Manuscript received January 25, 2013; revised June 6, 2013, October 29, 2013, and November 15, 2013; accepted November 15, 2013. Date of publication December 3, 2013; date of current version December 17, 2013. This work was supported in part by the National Natural Science Foundation of China under Grants 61001148 and 61272256, and in part by the Research Grants Council of the Hong Kong Special Administrative Region, China, under Grant CityU 119610. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Bulent Sankur. *(Corresponding author: Z.-Q. Yang.)*

X.-Y. Wei and Z.-Q. Yang are with the College of Computer Science, Sichuan University, Chengdu 610065, China (e-mail: cswei@scu.edu.cn; jessicayang@scu.edu.cn).

C.-W. Ngo and W. Zhang are with the Department of Computer Science, City University of Hong Kong, Hong Kong (e-mail: cscwngo@cityu.edu.hk; wzhang34@student.cityu.edu.hk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2013.2293427

[1] eBay.com and Taobao.com are the most popular e-Commerce websites in US (or worldwide) and China respectively.

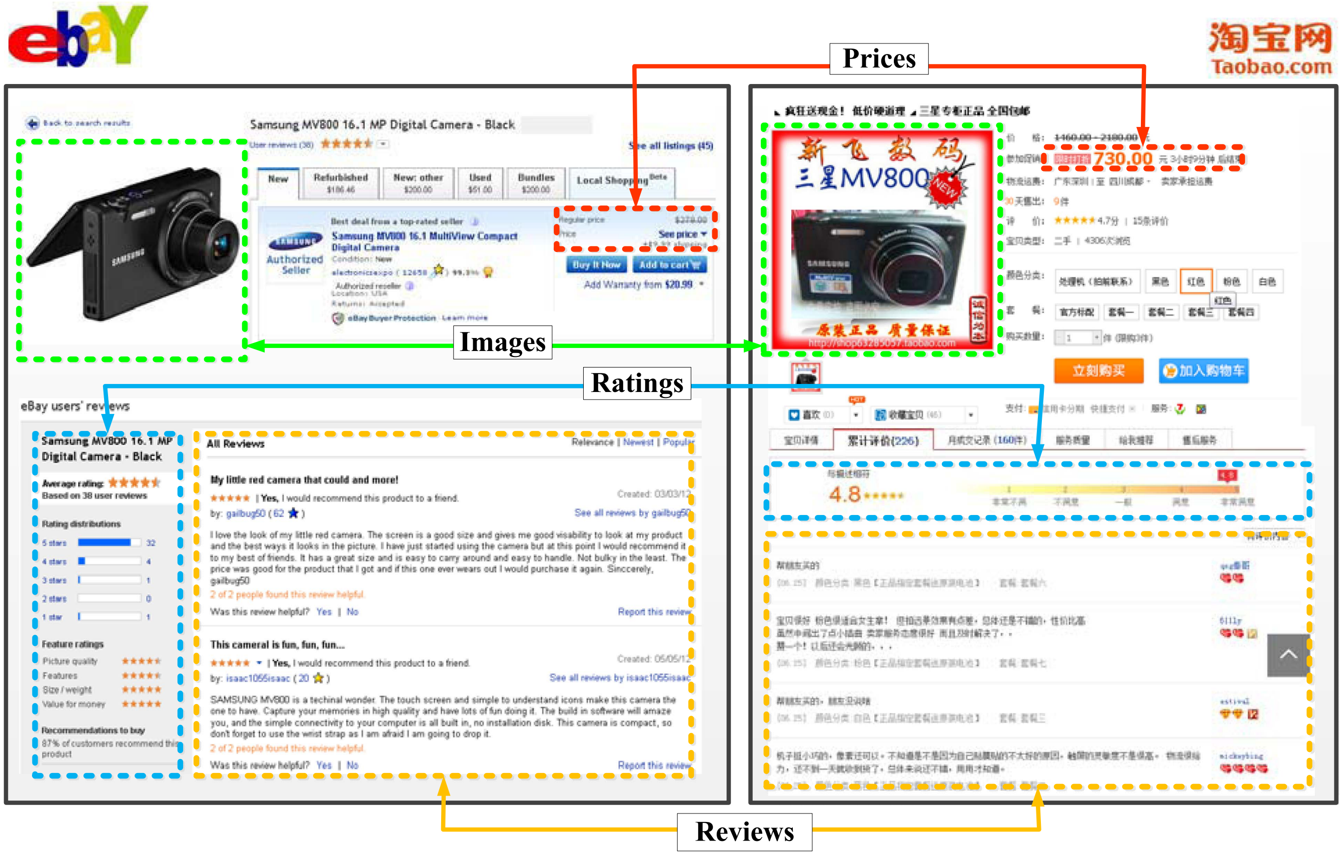
ebay
淘宝网
Taobao.com
Prices
Images
Ratings
Reviews
Samsung MV800 16.1 MP Digital Camera - Black
eBay users' reviews
All Reviews
My little red camera that could and more!
This cameral is fun, fun, fun...
730.00

Fig. 1. Valuable clues provided by cross-website (countries) near-duplicate detection.

some methods further integrate the checking of geometric consistency among LIPs into the process to improve the accuracy [13]–[16]. Therefore, the framework has seldom been studied in scenarios of Internet or mobile applications where the requirement for response time is critical and the target datasets are usually in web-scale.

In this paper, we conduct a domain-specific study of NDR on merchandize images, with the hope of serving the aforementioned online applications. We argue that the conventional LIPs+BoW framework does not exploit the potential representativeness of the LIPs, because LIPs are traditionally processed individually when mapping them into "words", disregarding the contextual information among LIPs and can easily lead to an inconsistent mapping. Therefore, we propose to integrate a "spelling or phrase correction" like process into the framework, which we name "visual typo correction". The basic idea is to extend the concept of *collocations* to the visual domain for modeling the contextual relations among visual words, on the basis of which we can enforce the mapped words for an image to be contextually consistent, so that the visual "typos" (i.e., the mis-mapped words) can be eliminated. The collocative optimization of visual words leads to the following advantages that are yet to be explored in the literature:

- *Robust Mapping*: Due to the variations in capturing conditions, acquisition times, rendering conditions, or editing operations, the appearances of a LIP can vary significantly in different near-duplicate images, which may introduce "typos/errors" into the mapped "words" when LIPs are considered individually. We propose a co-occurrence-based model to learn the visual word collocations and use it as a contextual information to correct *unreasonable* mapping. For example, a LIP from the surface of a car may be mapped to a "word" usually representing glass surfaces with reflections. This mapping can be corrected when the majority of other LIPs in the same image are mapped to car-related "words" (e.g., those representing wheels, lights or license plates);
- *Complexity Reduction*: To improve the representativeness of the LIPs within the LIPs+BoW framework, one has to increase the size of the vocabulary so as to cover as many "words" as possible. Several researches have shown that, to achieve a state-of-the-art performance, one has to build a vocabulary up to a scale of hundreds of thousands (e.g., 100k in [13] and 1M in [4]), which dramatically increases the time for feature quantization and thus is infeasible to online applications. Therefore, by introducing the contextual constraint in the proposed method, LIPs can work collaboratively to form patterns for feature representation, which relaxes the requirement for the large scale vocabulary. Our study will show that the proposed method can reduce the vocabulary size by 1,000% times while maintaining comparable performance (cf. Section V-C);
- *Feature Simplicity*: Existing works on NDR are mainly focusing on improving the accuracy (e.g., those

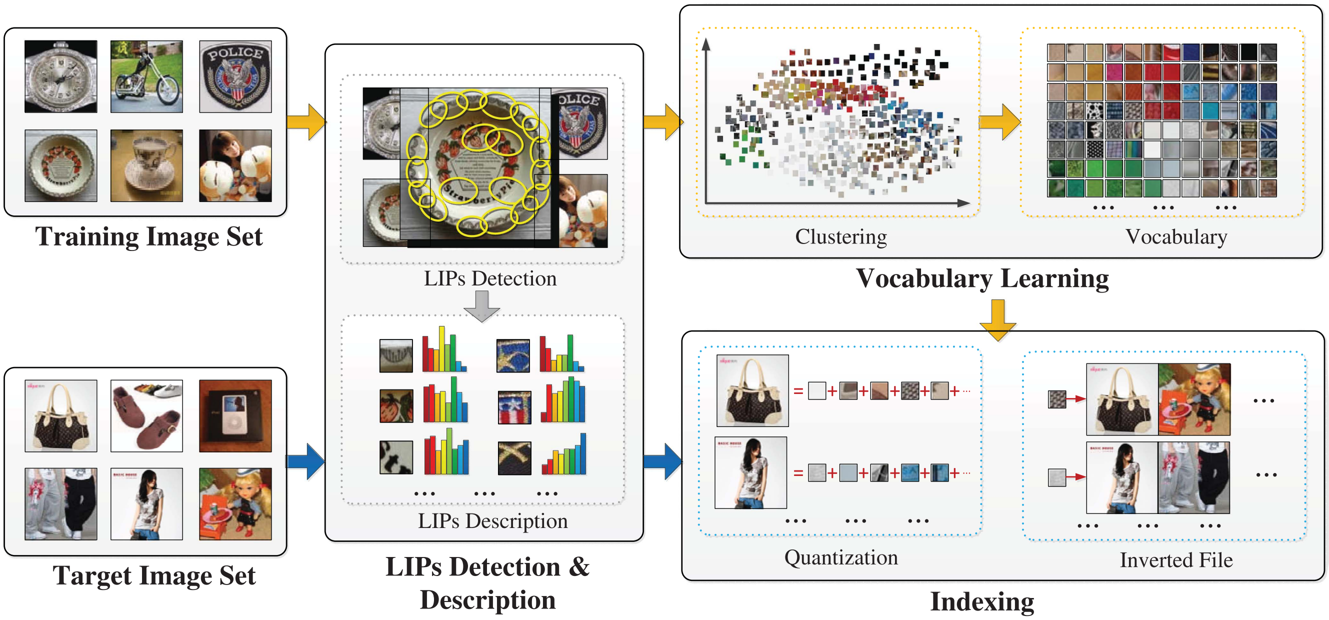

Fig. 2. LIPs+BoW framework for NDR.

by optimizing the detectors or descriptors for LIPs [17]–[19], by optimizing the vocabulary size [13], by introducing geometric constraints [13], [14]), while the efficiency has been simply addressed by employing standard techniques like inverted files [1] or hashing functions [20]. To the best of our knowledge, this is the first work studying the efficiency optimization for the domain-specific NDR on merchandize images. We will show that under certain scenarios, the expensive LIPs-based features can be replaced by simple features (e.g., color-moments) with comparable or even better performance, while reducing the time cost by 9,202% (cf. Section V-C).

The remainder of this paper is organized as follows. Section II reviews the LIPs+BoW framework and NDR techniques in the literature. Section III introduces the contextual consistency in a NDR sense, and Section IV describes our model of visual typo correction by collocative optimization. The experimental results will be presented in Section V. Finally, Section VI concludes this paper.

II. Related Work

A. A General Framework of NDR

A description of the LIPs+BoW based NDR framework is given in Fig. 2. The procedure starts by extracting LIPs for each image, where every pixel's stability to transformations (for example, the change of scale, illumination) is evaluated by investigating the patch(es) around the pixel, resulting in a set of pixels with good invariance properties as LIPs. A patch around each LIP is then extracted and described by the local features which are usually invariant to the geometric and photometric transformations [6], [7], [21]. LIPs of all images in the dataset are pooled and clustered into groups. A model feature vector representing the "word" of each cluster is calculated using the mean or medoid of member vectors within the cluster. By composing all "words" into a vocabulary, every LIP can be "quantized"[2] by assigning it to the closest word, so that one image is represented as a set of "words", resembling a traditional text document [1]–[5]. Finally, images are indexed into an inverted file and the rest of the process is almost the same as the conventional text retrieval.

The LIPs+BoW framework is simple and straightforward, and has demonstrated promising performance in many applications. However, when facing web-scale data, its performance has been challenged in both accuracy and time efficiency. On one hand, we should notice that decomposing an image into LIPs decreases the geometrical relations among pixels. Consequently, the precision of retrieval can be seriously affected, especially when facing a large amount of web images where the chance of encountering non-near-duplicate images having similar sets of visual words becomes non-negligible. On the other hand, as discussed in Section I, the searching of LIPs with a pixel-by-pixel manner is inherently inefficient, and the matching between the LIP sets of two images (usually over 1,000 LIPs for each) is even more costly.

B. Recent Trends of NDR

Efforts to improve the accuracy and efficiency are the main drive behind the recent trends in NDR. To improve the *accuracy*, the most intuitive idea is to add the geometrical constraint when comparing visual words (LIPs after quantization) between two images. Words in one image are usually related to their neighbors to construct (explicitly or implicitly) graph(s), hence image comparison using graphs resulting from LIPs of respective images yield more accurate similarity measures [13]–[16]. Alternatively, to mitigate some of the frequently occurring visual words from dominating the others,

[2] Note that the quantization process sometimes is called "encoding". We use the term "quantization" throughout this paper, because it is more often used in the literature of near-duplicate retrieval.

"burstiness" of visual words is often adopted to adjust the similarities between images by further considering the inter-image or intra-image frequency of the words [22]. Meanwhile, to improve the *efficiency*, a majority of researches is focusing on developing better hashing functions to map LIPs into a low-dimensional space so as to speed up the matching process (e.g., locality sensitive hashing (LSH) [17], [23], hamming embedding (HE) [13], [24], and min-Hash [25], [26]). State-of-the-art methods often integrate these two aspects by using hashing to retrieve the whole dataset and then employing the geometrical constraint to re-rank the top-ranked images. Nevertheless, the quest about which geometrical constraint should be exploited continues, since one has to set a balance between recall and precision. A similar dilemma also occurs in the hashing or encoding stage where one has to determine parameters of the hashing algorithms, the hashing function itself, the dimension of the target space, or the number/size of the bins.

Despite being accuracy or efficiency oriented, most of recent methods put their focus on *post*-quantization process. As discussed in Section I, however, due to the variations in the appearances of LIPs, the quantization usually introduces a certain amount of wrongly mapped words. Those "typos" will be propagated and further degrade performance of the processes afterwards. Therefore, in this paper, we propose a *pre*-quantization method which utilizes the contextual relations among visual words to correct the "typos" and thus improve the quality of LIPs-to-Words mapping. Specifically, instead of directly mapping each LIP to a single word, we select several candidates for every LIP and finally find an optimal set of words (one for each LIP) where the contextual consistency among words has been maximized. In this regard, the proposed method is also related to soft-weighting [27] or visual word ambiguity [28] which map each LIP to several words with each word assigned a weight proportional to the significance of the LIP. While soft assignment in soft-weighting or visual word ambiguity has been proven to be a practical way to compensate for the shortcomings of the single-word-mapping, though it will cause an increase in the number of words needed to represent the image as well as the complexity of the feature and thus affect the efficiency. Furthermore, it brings additional computational cost or may even cause ambiguity for the *post*-quantization processes. Therefore, soft assignment is only applied to the query side in most of the applications.

C. NDR in a Text Retrieval Point of View

While NDR has been studied over one decade resulting in a large number of methods, a significant amount of ideas employed in the LIPs+BoW framework have been obviously borrowed from text retrieval (TR) domain [29]. Therefore, when reviewing those ideas, it is better to refer to their analogues in TR that most readers are more familiar with, as summarized in Table I.

TABLE I

TECHNIQUES USED IN CONVENTIONAL TEXT RETRIEVAL AND THEIR ANALOGUES IN NEAR-DUPLICATE RETRIEVAL

Conventional Text Retrieval	Near-Duplicate Retrieval
Tokenization	LIPs Extraction [6] [7] [21]
Stop-List	Visual Word Stop-List [1]
Stemming	Vector Quantization [1-20]
Spelling/Phrase Correction	(No Analogues)
Inverted File	Visual Word Inverted File [1, 15]
Query Expansion	Contextual Synonyms [30]
Reranking	Spatial Reranking [13–16]

As introduced, the first step in TR is to parse documents into words (tokenization), while its analogue in NDR is to extract LIPs from images. Therefore, early efforts have been put into seeking effective types of "words" (i.e., LIPs). Representative ones include SIFT [6], SURF [7] and MSER [21]. Second, the quantization in NDR maps LIPs to visual words, which is analogous to stemming in TR. For each LIP, the mapped visual word is usually the one(s) with the largest similarity. A list of stop words is sometimes employed to filter out the non-informative visual words, which is directly adopted from TR [1]. In indexing, both NDR and TR use inverted file to index images (documents) [1], [15]. Alternatively, one can use "term frequency-inverse document frequency" (known as tf-idf) to represent images (documents) to vectors for similarity-based retrieval (i.e., Vector Space Model) [4]. In addition, query expansion, which expands query terms (i.e., the visual words of the query image) with their contextual synonyms (i.e., visual words that frequently co-occur within the same context), has also been practiced in NDR [30]. Finally, spatial consistency is measured between query images and the top-ranked ones in the initial retrieved list for re-ranking. There might be a lot of correspondences in TR for the idea of re-ranking. But given the fact that every graph edge is actually representing the spatial relation of a word-pair, the re-ranking scheme using edge-wise comparison in NDR is more like a bi-gram based method [13]–[16].

In Table I, one can observe that the spelling/phrase correction widely used in TR is missing its analogue in NDR. In TR, it is intuitive to remove typos/errors by using spelling/phrase correction to infer that a user should mean "a drop of water" even she has typed "a drap of wafer", indicated by the context of the sentence. One question may be raised immediately is that "Do visual words in NDR also include typos/errors?" In following sections, we will show that the answer is YES and the performance of NDR can be further improved when contextual consistency is considered in the LIPs+BoW framework.

III. CONTEXTUAL CONSISTENCY IN NDR

Compared with that of text retrieval, contextual consistency in NDR might be a less intuitive concept to the reader. In this section, we identify the "typos/errors" in visual words, on the basis which we then define the concept of contextual consistency for correcting visual spelling in NDR.

A. Typos/Errors in Visual Words

There are several chance events where "typos/errors" can occur in visual words. The inevitable variation in appearances is the first one to consider. Various conditions of capturing or various types of editing can make the same objects or scenes appear significantly different in near-duplicate images. Under these conditions, visual feature of an individual LIP may vary dramatically from the original. Serious deformation is the second important factor. This can happen during intentional editing (e.g., resizing, captioning, rotating), deformations of a product consisting of a soft material, and recapturing photos in an environment different than the query with significant changes in illumination, view-points and so on. Specular reflection is also responsible, given the overwhelming popularity of shiny surfaces in commercial products. The last one goes to the quantization step in NDR (cf. Fig. 2). Representing LIPs in the same cluster with its mean or medoid vector has reduced the computational cost, but inevitably will introduce ambiguity, which increases the probability of mapping the LIPs to the wrong visual words. When all those factors come together in NDR, the set of visual words for representing a certain image could be poor, and this can seriously degrade the subsequent steps.

B. What's the Contextual Consistency in NDR?

Nearly all the spelling/phrase correctors in TR are utilizing the contextual consistency (e.g., collocation of words) between words for eliminating typos/errors. In NDR, the concept of contextual consistency can be conveniently adopted since visual words work together to compose an image in almost the same way as text words do in a document. For example, in the conventional NDR approaches, LIPs from the surface of a car can easily be quantized as visual words representing surface of an airplane or other metal-like materials. However, if a certain amount of LIPs representing road, tires or other concept related to ground transportation are detected in the same image, one can be more confident that the mis-quantized LIPs are from the surface of a car. This is similar to the logic we use in text retrieval that “a drop” most likely comes with “water” instead of “wafer”. Using contextual consistency which considers visual words in an image collaboratively, therefore, has the advantage over conventional quantization process where visual words are mapped individually.

One may argue that in reranking methods based on *geometric consistency*, the *contextual consistency* has already been considered. Although this is partly true, these two concepts have fundamental differences: a) geometric consistency is employed to compare the relative positions among words, while contextual consistency is used to verify co-occurrence. In other words, geometric consistency prevents “drap of wafer” to be matched to phrases like “wafer of drap” or “wafer drap of”, but it does not take into consideration if “drap of wafer” is meaningful to the context. On the contrary, contextual consistency makes sure all visual words have semantic meanings.[3] consistent to the context, forcing “drap of wafer” to be mapped to “drop of water” during quantization; b) geometric consistency is used in the post-processing of the retrieval to eliminate the irrelevant results, while contextual consistency is used in the pre-processing to improve the quality of the inputs. These two are thus not mutually exclusive. It is worth mentioning that in contextual consistency checking discussed in this article, the order of the words is ignored, which means that “drap of wafer” can be mapped to “water of drop” or “water drop of”. Given the fact that how to define a linear order for visual words in a 2D image is still an open question, this relaxation is reasonable. Moreover, when both geometric consistency and contextual consistency are adopted in the same system, they can be complementary to each other.

IV. Utilizing Contextual Consistency for NDR

To utilize contextual consistency for NDR, we introduce two additional components into the vocabulary learning and quantization processes respectively. During vocabulary learning, we will first identify the visual word collocations for representing the contextual information. During quantization, instead of mapping each LIP to the visual word with the largest similarity (as in conventional approaches), we select for each LIP the top-k words as its candidates, and finally choose the best set of words that has the largest contextual consistency with the collocations learned in vocabulary learning. In this section, we will introduce these two components in detail.

A. Modeling Visual Word Collocations

In text domain, a collocation is defined as a sequence of words that co-occur more often than would be expected by chance [31]. A popular method for finding collocation candidates is to slide an n-word window along sentences to calculate the in-window co-occurrence of words so as to identify word sequences with high frequency. We extend this idea to image domain by defining the window as a circle of radius d pixels. Therefore, for any two words under investigation, it is easy to calculate the in-window statistics about their co-occurrence.[4] Finally we employ $\mathcal{X}^2$ test to select the word pairs which co-occur frequently but not by chance.

To model the contextual consistency in NDR, the most convenient way is to borrow existing techniques developed in natural language processing. We choose (point-wise) mutual information in this case, which is popularly employed to measure the collocation of words. More specifically, given two visual words w_1 and w_2, the mutual information is defined as

$$I(w_1, w_2) = \log \frac{p(w_1, w_2)}{p(w_1)p(w_2)} \tag{1}$$

where $p(w_1, w_2)$, $p(w_1)$ and $p(w_2)$ are the frequencies of the co-occurrence of w_1 and w_2, the occurrence of w_1, and the occurrence of w_2, respectively. Note that not like text word pairs which appear in sentence in order, visual words do not appear sequentially in an image. Therefore, in Eq. (1), we relax the order constraint so that $I(w_1, w_2) = I(w_2, w_1)$.

[3] The wording “semantic” means that a visual word shall carry some visual characteristics that is unique to the type of objects where it is extracted from. A visual word can thus be viewed as a symbol semantically meaningful for depicting its host objects, analog to using a text word (e.g., drop) to describe an object (e.g., water). Note that the visual words are neither associated to any semantic tags nor its meanings are learnt through any supervised learning methods.

[4] Similar ideas have been employed in [30] and [32] for learning the co-occurrence among visual words, but for different purposes (i.e., for detecting visual synonyms in [30] and for modeling visual phrases in [32]). More discussions can be found in the Sections IV-B and V-C.

It is worth mentioning that the collocation of visual words has also been studied by other authors (e.g., [33] and [34]) but mostly with the focus on how to identify different types of visual collocations, which is different from our purpose of enforcing contextual consistency for NDR. Therefore, in this paper, we employ $\mathcal{X}^2$ test and mutual information to model the collocations whose effectiveness has long been recognized in TR. While which technique for collocation extraction is the best remains an interesting topic for future work, it is beyond the topic of this paper.

B. Contextual Consistency Enforcement

During quantization, assuming that top-k visual words (with the largest similarities) have been selected as candidates for each LIP, the task of finding the best set of visual words (i.e., one word for each LIP) for representing an image then falls into a typical combinational optimization problem – to find an optimal combination (set) of visual words that maximizes the contextual consistency. To this end, we first define an energy function to measure the contextual consistency among words in a given set $\mathcal{W}$ as

$$E(\mathcal{W}) = \sum_{i=1}^{|\mathcal{W}|} \sum_{j=i+1}^{|\mathcal{W}|} I(w_i, w_j), \tag{2}$$

which is straightforwardly the accumulated mutual information of words in $\mathcal{W}$. In the rest of this section, we will see that searching an optimal set $\mathcal{W}^*$ which maximizes $E(\cdot)$ can be formulated as a binary quadratic programming (BQP) problem.

Assuming there are n LIPs in the current image and for each LIP, the top-k visual words (with the largest similarities) have been selected as candidates, the candidate pool then includes $n \times k$ words as

$$\underbrace{w_1, w_2, \ldots, w_k}_{\text{For the } 1^{st} \text{ LIP}}, \underbrace{w_{k+1}, \ldots, w_{2k}}_{\text{For the } 2^{nd} \text{ LIP}}, \ldots, \underbrace{w_{(n-1)k+1}, \ldots, w_{nk}}_{\text{For the } n^{th} \text{ LIP}}.$$

To each candidate word w_i, let us assign a binary variable $x_i \in \{0, 1\}$ to indicate its selection status, on the basis of which we can modify Eq. (2) to a BQP object function as

$$f(x_1, x_2, \ldots, x_{nk}) = \sum_{i=1}^{nk} \sum_{j=i+1}^{nk} x_i x_j \cdot I(w_i, w_j) \cdot N(w_i, w_j) \tag{3}$$

where $N(w_i, w_j)$ is a function which returns 1 if the corresponding LIPs of w_i and w_j fall into the same d-radius window, 0 otherwise. Note that the term $x_i x_j$ ensures $I(w_i, w_j)$ has its impact to the object function only if both w_i and w_j are selected (i.e., $x_i = x_j = 1$), while $N(w_i, w_j)$ enforces the selection of visual word for a LIP only need to be contextually consistent with those of its neighbors, which will significantly reduce the number of nonzero terms in Eq. (3) and thus simplify the problem. In addition, to make sure the selections of the visual words among the candidates of the same LIP are exclusive, we add n constraints for the BQP model as

$$\sum_{j=1}^{k} x_{(i-1)k+j} = 1, \quad i = 1, 2, \ldots, n. \tag{4}$$

Further composing all x_i's into a vector $\boldsymbol{x}$ and the pairwise neighboring and selection statuses (i.e., $N(w_i, w_j)$ and $I(w_i, w_j)$) into a matrix $\mathbf{C}$ (a nk-by-nk matrix) where

$$\mathbf{C}_{ij} = I(w_i, w_j) \cdot N(w_i, w_j), \quad i, j = 1, 2, \ldots, nk \tag{5}$$

the optimization problem can be summarized into a standard BQP form as

$$maximize \quad f(\boldsymbol{x}) = \boldsymbol{x}'\mathbf{C}\boldsymbol{x} \tag{6}$$

$$subject\ to \quad A\boldsymbol{x} = \mathbf{1} \tag{7}$$

$$x_i \in \{0, 1\}, \quad i = 1, 2, \ldots, nk \tag{8}$$

where $\mathbf{A}$ is a n-by-nk binary matrix for encapsulating the constraints in Eq. (4). Therefore, for each row (denoted as the i^{th} row) of $\mathbf{A}$, the $(i-1)k+1$ to $(i-1)k+k$ entities are all ones while the rest are all zeros. The BQP, even being a NP-hard problem, has been intensively studied, and can be efficiently solved with well-know discrete optimization techniques such as the branch and bound algorithm [35]. Moreover, both $\mathbf{C}$ and $\mathbf{A}$ are highly sparse, which will inherently reduce the time for solving of Eq. (6). Furthermore, to meet the need of online applications, we implement a coarse-to-fine branch and bound algorithm. We first perform clustering on all the LIPs and select only the medoids from the resulting clusters to conduct a coarse BQP. Then we fix the selections for those medoids so as to filter out a large number of branches in the search tree with which we conduct a fine BQP involving all the LIPs. This is based on the observation that similar LIPs will usually select similar words so that the selections of member LIPs in a cluster will generally follow that of the medoid. In our experiment, by simplifying the search tree, the computational cost has been reduced from 19 $\sim$ 116 to 12 $\sim$ 27 milliseconds.

One may argue that the BQP can be replaced by classical dynamic programming by arranging LIPs sequentially as *stages* and their candidate words as *states*. However, this is not exactly true since selections of visual words are highly interdependent on LIPs, it is difficult to enforce contextual consistency on an one-by-one basis. In terms of dynamic programming, it makes the optimization of the i-th state not only dependent on those of the previous states but also those of the succeeding ones. Therefore, arranging the states sequentially is tantamount to ignoring some of the interdependencies.

Another fact worth mentioning is that the BQP process depends on the learning of contextual consistency among visual words which were themselves obtained through learning. In principle, one could be concerned about the effect of word quantization errors to contextual consistency (i.e., Eq. (1)) as the quantization errors (i.e., some LIPs are mapped to wrong words) on the training dataset are inevitable. We argue this is not a critical issue for the proposed method, because the contextual consistency is learnt from the visual collocations which must be the words that co-occur "frequently". Since the quantization errors are basically "random" (in a statistical sense) and the mis-mapped words rarely have the chance to form collocations with others,

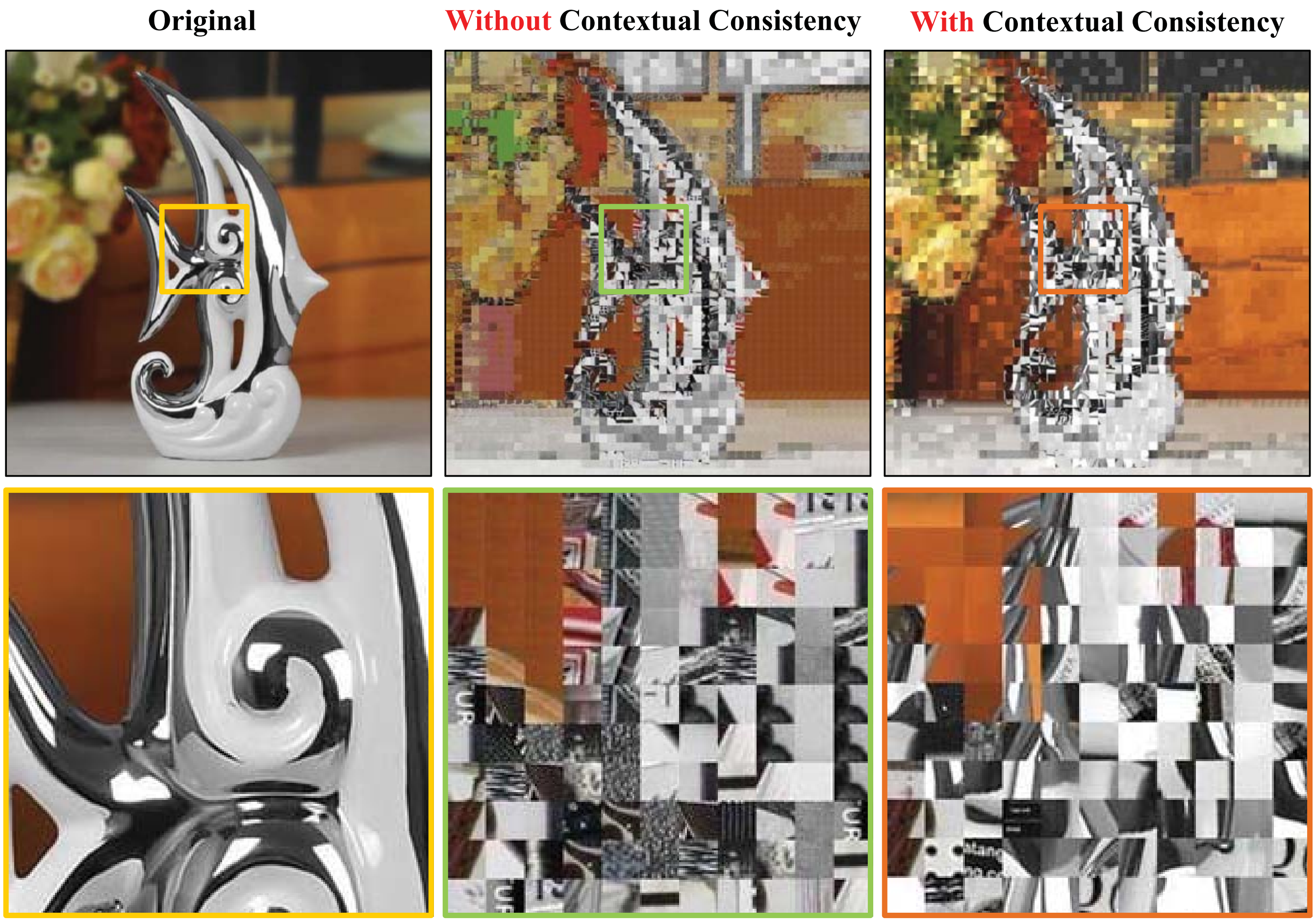

Fig. 3. Examples of reconstructing images with their visual words. The first row (from left to right): the original image, image reconstructions with and without contextual consistency. The second row: the enlarged details of images on the first rows.

these noisy cases can practically be ignored given a large enough training dataset.

Note that the use of spatial context in BQP is different from [30] in which the authors mine frequently appearing words as contextual synonyms for query expansion. While both approaches share similarity in learning co-occurrences statistics, BQP enforces the spatial consistency in visual word quantization and thus can be viewed as a pre-processing step for indexing. In contrast, [30] leverages the statistics during query time by expanding query words with contextual synonyms for boosting the recall of relevant images. Since the consistency is only considered in query time, mismatched words will still be indexed for online retrieval. In brief, BQP fundamentally corrects the mismatched words with spatial context, while [30] compensates the mismatches, rather than making correction, with contextually consistent words.

C. Visualization of Contextual Consistency

Up to this point, we have introduced and established the model of contextual consistency in NDR. However for the sake of clarity, we have borrowed examples from the text domain during description. In this section, we verify whether it is applicable to NDR in an "image" sense.

To visualize the effect of utilizing the contextual consistency on images, we modify the processing of quantization as follows: instead of using LIP detectors, images (with normalized resolutions) are decomposed into patches with a fixed size (20 × 20 pixels in our case, as shown in Fig. 3); patches are described using color-moment features (the first three moments, namely the mean, standard deviation, and skewness, are calculated per RGB channel [36]); feature vectors of all patches are clustered using K-Means so that the medoids are selected as visual words to construct the vocabulary; every time a new image comes, we decompose it into patches in the same way and assign the patches to the corresponding visual words using the BQP quantization approach. By this means, the resulting visual words can be attached back to the image so that the quantization process is visualized as a process of symbolic image reconstruction.

Fig. 3 shows an example of reconstructing image with its visual words (more examples can be found at our demo webpage[5]). It is easy to see that, by enforcing contextual consistency, the visual words after the BQP quantization are able to recover the original image more faithfully, by using more metal-like patches for the fish, more plant-like patches for the flower and leaves, and more wood-like patches for the drawers at background. From the enlarged details, we can see that the appearances of different parts of the metal fish

[5] http://vireo.cs.cityu.edu.hk/wxy/product.images/

vary largely due to the illumination and reflection. Therefore, simply selecting the visual words with the largest similarities to the corresponding patches will result in a diverse and random set of patches for reconstruction. For example, some of the highlights are represented by patches with white text and gray background, which are highly inconsistent to the patches of the other parts. By contrast, this disadvantage is eliminated when contextual consistency is enforced during quantization.

V. Experiment

A. Dataset Construction and Evaluation Metric

To evaluate the performance, we have crawled a dataset including one million (i.e., 1,106,280) merchandize images from eBay.com and Taobao.com, and divided them into three subsets: 1) **Training Set** which includes 1% (i.e., 10,000 images) of the entire dataset; 2) **Basic Set** which includes another 10% (i.e., 112,092 images) of the dataset; 3) **Distracter Set** which includes the rest part of the dataset (i.e., 984,188 images). There are no exact duplicates in the dataset and no overlaps among the three subsets. The ground-truth is generated by fully annotating the **Basic Set**, where 12,901 images are found to form 4,141 near duplicate groups. We use each of those images as a query. Note that the generation of the ground-truth is an interactive process, in which we manually check the retrieved list if images from the **Distracter Set** are top-ranked and move them to the **Basic Set** if they are near-duplicates. This process has been done for several rounds, so as to avoid the case (to the largest extent) that a near-duplicate image is included in the **Distracter Set**. All images are normalized to fit an 800 × 800 pixel box while preserving their aspect ratios. To the best of our knowledge, this is the dataset with the largest number of near duplicate images so far, which can be downloaded from our demo webpage[5] (will be open to public upon acceptance).

In the experiments, we use each of the near-duplicate images as a query. The retrieved images are ranked according to their similarities to the query. The search performance is then evaluated with mean average precision (MAP), where AP is defined as

$$AP = \frac{1}{\min(R, n)} \sum_{j=1}^{n} \frac{R_j}{j} I_j \tag{9}$$

where R is the number of duplicate images to a query, R_j is the number of duplicate images in the top-j retrieved images, and $I_j = 1$ if the image ranked at j^{th} position is a near-duplicate (ND) and 0 otherwise. We set n=100, while MAP is calculated as the mean AP of the 12,901 queries.

B. Collaborative vs. Individual Representations

We employ two baselines in the experiment which extract LIPs with

- $SIFT$ [6] based method, which is considered one of the most popular methods, where difference of Gaussians (DoG) is used for detecting the LIPs while SIFT feature is used for description;

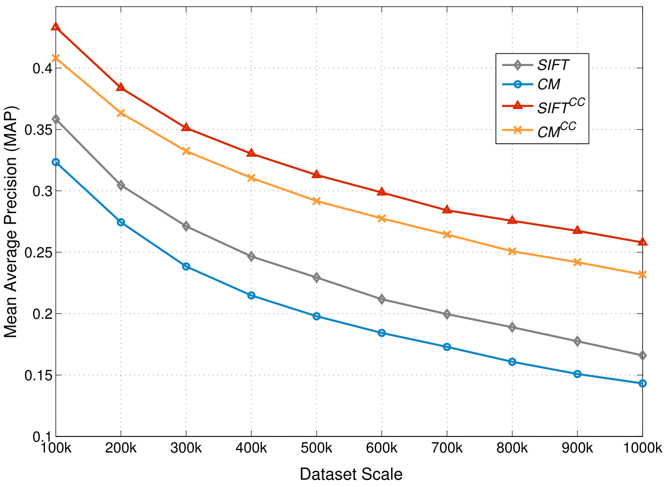

Fig. 4. Performance comparison over different dataset scales.

- Color-Moment (CM) based method (cf. Section IV-C), which extracts patches (of a fixed size of 20 × 20 pixels) for each image and uses color-moment feature for description. Given the fact that all images are normalized to fit an 800 × 800 pixel box during the experiments, the CM features are densely sampled, compared to the sparsely sampled SIFT features.

Both are implemented following the LIPs+BoW framework in Fig. 2. To investigate the effect of enforcing contextual consistency in NDR, we modify the two baselines by further imposing the contextual consistency constraint during quantization, resulting in two new methods $SIFT^{CC}$ and CM^{CC} where the superscripts have been added to distinguish from the baseline versions. The **Training Set** is used to learn the visual word collocations (i.e., co-occurrence patterns) for $SIFT^{CC}$ and CM^{CC}. In addition, to investigate the performance over different scales of datasets, our experiment starts by using the fully annotated **Basic Set** as a basis and randomly adding samples from the **Distracter Set** to generate larger sets. Due to a large number of experiments required for investigating the effects of the features and parameters to the performances, we temporarily fix the vocabulary at a small size 2k to speed up the process. The performances of using larger vocabulary sizes will be reported and discussed in Section V-C.

The results with MAP at the top-100 retrieved images are shown in Fig. 4. It is easy to see that by further considering contextual consistency, the performance of the two baselines have been improved significantly and the improvements are stable (ranging from 20.87% to 61.91%) over different dataset scales. Statistics on the **Basic Set** when dataset scale equals 100k show that $SIFT^{CC}$ (CM^{CC}) improves 6,164 (6,492) out of the 12,901 queries of from the baseline $SIFT$ (CM), while leaving 5,466 (4,802) queries unchanged and only a small portion of 1,271 (1,617) queries with slight drops. This has confirmed the advantage of considering contextual consistency in NDR, and is also consistent to our analysis in Section IV that enforcing contextual consistency in quantization can improve the quality of the LIPs-to-Visual-Words mapping and thus boosts the performance. To grasp an in-depth understanding of the method, we further conduct

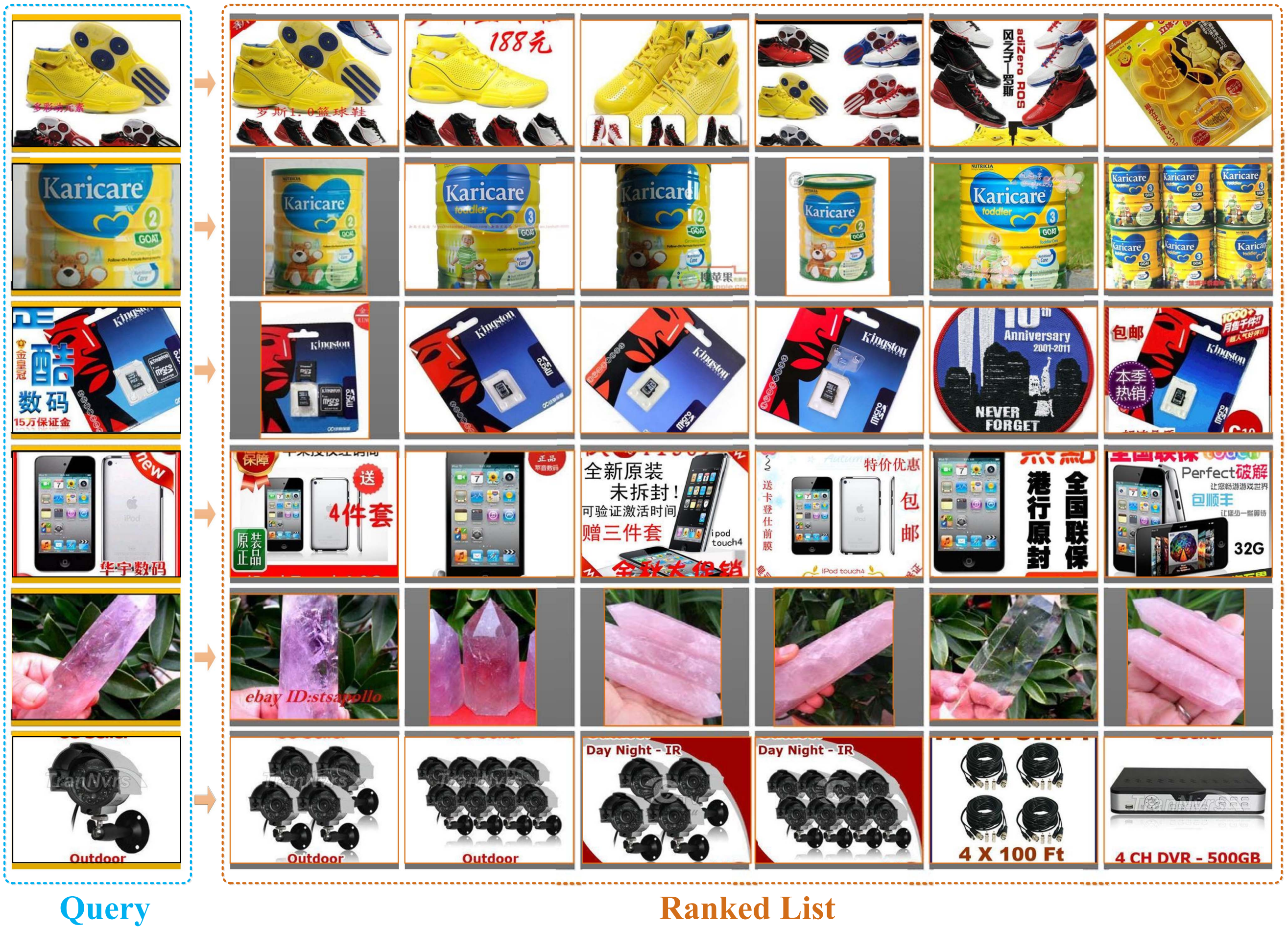

Fig. 5. Examples of the query images and the retrieved items by CM^{CC}.

several experiments to investigate its impacts of/on features, parameters, and the ranking list.

1) The Impacts of Features: In Fig. 4, we can see that, benefiting from the contextual consistency, the improvement of CM^{CC} from CM (47.18%±11.73) is much more significant than that of $SIFT^{CC}$ from $SIFT$ (38.20%±10.95). These results have indeed further revealed the advantage of the collaborative representation gained by considering contextual consistency. Color moment are usually extracted from patches of a grid-divided image so that features from those patches can be concatenated to represent color patterns in the image with respect to spatial relation given by the grid. However, the way of using color-moment in CM without considering contextual consistency has broken the spatial relations among patches because each patch is considered individually in the LIPs+BoW framework. This makes CM just a color histogram-like feature which is weak and lacks of representation power when compared to $SIFT$.

However, when contextual consistency is considered in CM^{CC}, spatial relations are taken into consideration, because the BQP quantization (cf. Section IV-B) will force the color-moment features of neighboring patches to be consistent with the co-occurrence patterns learned on the **Training Set**. Furthermore, the way of using visual word collocations to indirectly represent the spatial patterns can be more flexible than the conventional grid dependent method, because the fixed grid is not easy to be scale or rotation invariant. By contrast, the improvement of $SIFT^{CC}$ from $SIFT$ is not as significant as that of CM^{CC}, because $SIFT$ is inherently scale and rotation invariant, and with strong representativeness for local patterns. Therefore, the room for improving $SIFT$ is comparably limited.

To give an intuitive impression on the performance of CM^{CC} on the merchandize retrieval, in Fig. 5, we show several query images and the corresponding ranked lists returned by CM^{CC}. We can see that CM^{CC} is rotation invariant and is capable of tolerating small to moderate changes in scale. The rotation invariance is straightforwardly resulting from the definition of the color-moment feature, which is inherently a rotation invariant feature at patch-level. At image-level, since CM^{CC} models the inter-patch relationship with the co-occurrence among patches (instead of the strict spatial constraint), it is also invariant to the rotation. Furthermore, although CM^{CC} is not a scale invariant feature, it can still tolerate small changes in scale, thanks to the relatively simple textures in most of the merchandize images.

2) The Impact on Ranking: To investigate how the performance is improved by considering contextual consistency, we plot the change of APs on the **Basic Set** in Fig. 6. It is easy to see that the performance of a method is indeed determined

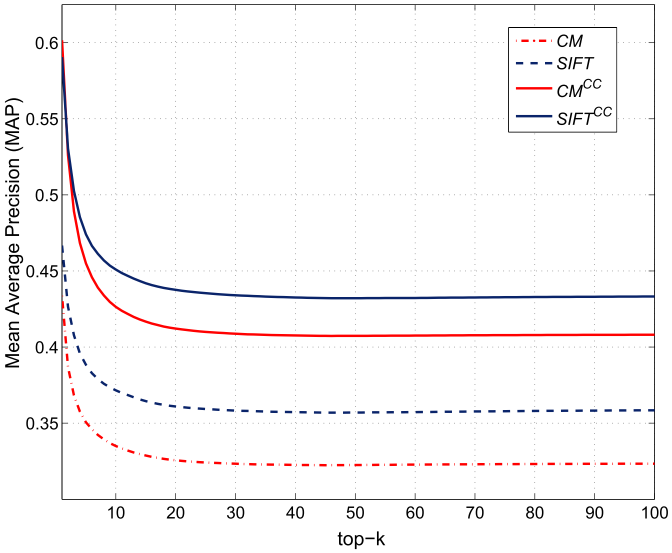

Fig. 6. Evaluation of MAP at top-k of the ranked lists.

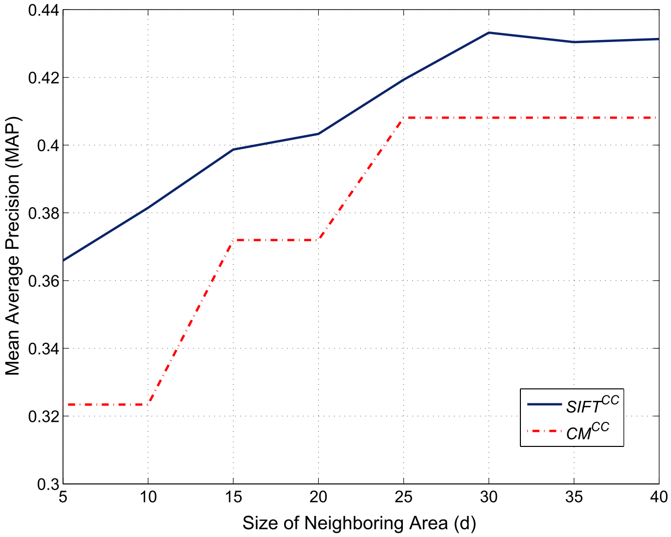

Fig. 7. Evaluation of MAP over the change of the size of the neighboring area (d, in pixels).

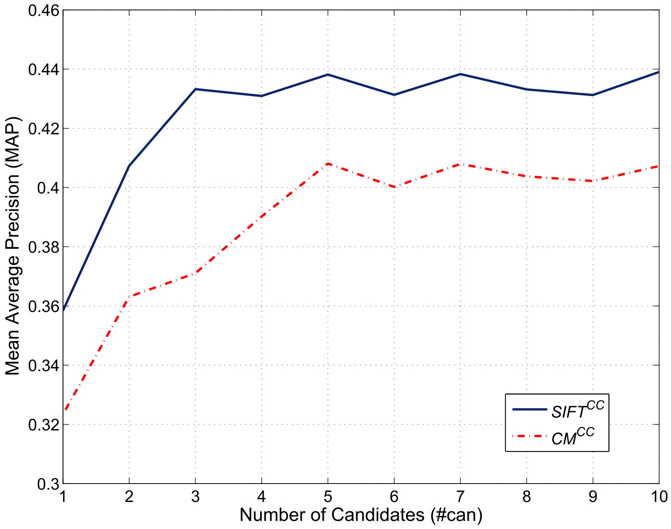

Fig. 8. Evaluation of MAP over the number of the candidates for each LIP.

by the number of NDs that it retrieved at the top-30 of the ranked list. This is an indication that most of the NDs distribute within the section of top-1 to 30 of the ranked lists. Furthermore, we can see that the APs of $SIFT^{CC}$ and CM^{CC} of this section drop more dramatically than those of $SIFT$ and CM, which is another indication that the improvements brought by considering contextual consistency has its impact mainly on the top-30 items of the ranked lists, where the NDs are boosted to the top positions. This is a highly desired characteristics, because users usually will or can access only the top-30 retrieved items in most of the applications (e.g., on the first page of the retrieved items in the e-Commerce websites, or on their smart phones where the screen size is too limited to display too many items).

3) The Impacts of the Parameters: In case that the vocabulary size is fixed, there are only two parameters in our approach, namely the radius d (in pixels) which defines the size of the neighboring area, and the number of visual word candidates for each LIP (denoted as #*can* hereafter[6]).

Fig. 7 shows the performances of $SIFT^{CC}$ and CM^{CC} over the changes of d on the **Basic Set**, where we can see that with the growth of d, the performance of $SIFT^{CC}$ (CM^{CC}) steps up continuously and is approaching stable at the point 30 (25).[7] This is due to the fact that, while the d is increasing, the overlaps among neighboring areas of LIPs are also enlarged, so that, at first, LIPs within the same neighboring areas are connected to form patterns which model the image more accurately. After that, while d increases further, the newly formed patterns are also connected, so that relations among LIPs not residing the same neighboring areas can be modeled indirectly, benefiting from the transitivity. Therefore, all the LIPs in an image can work collaboratively, resulting in a better representation. However, this will not bring further gain for the performance when d crosses certain threshold, since all LIPs has already been connected as a whole. Moreover, we can see that the performance of CM^{CC} converges at smaller d than $SIFT^{CC}$. This is because CM^{CC} (i.e., image blocks or patches) are tightly connected when we divide the image, while LIPs of $SIFT^{CC}$ (founded by LIP detector) are sparsely scattered.

Fig. 8 shows the performances of $SIFT^{CC}$ and CM^{CC} over the number of visual word candidates for each LIP (#*can*), where the performances of the two methods are approaching a stable plateau after crossing the thresholds, 3 and 5, respectively. Generally, the more the candidates are, the higher probability that the best visual word can be found after the BQP quantization. However, the optimal number of candidates is indeed dependent on the variance of

[6] Note the parameter has been denoted as k in Section IV-B for a better mathematical presentation. Here #*can* is used to make its meaning of visual word candidates more intuitive.

[7] Note that in $SIFT^{CC}$, two words have the neighborhood relationship only when the corresponding LIPs fall into a circle of radius d (pixels), while in CM^{CC}, the relationship is dependent on whether the *centers* of the two patches fall into the circle.

corresponding feature, because the larger the variance of the feature is, the more candidates need to be selected to make sure the best visual word is included. This in fact explains why $SIFT^{CC}$ converges at smaller number of candidates than that of CM^{CC}, because the $SIFT$ is claimed to be scale, rotation invariant, which we cannot guarantee with CM.

We have also conducted experiments for finding the best combination of the three parameters (i.e., the vocabulary size, d and #*can*) by using Grid Search. The experimental results show that CM^{CC} ($SIFT^{CC}$) converges when the vocabulary size, d and #*can* are set at $15K$, 25 and 5 ($25K$, 30 and 3), respectively. These values will be used in the subsequent sections as default settings for the parameters unless otherwise specified. Furthermore, the three parameters are observed independent of each other, because the change of the value for one parameter will not affect the conclusions made for the other parameters. This gives us advantage for the investigations.

C. Comparison With the State-of-the-Art

To investigate the accuracy and efficiency of the proposed methods, we compare $SIFT^{CC}$ and CM^{CC} to several state-of-the-art NDR methods in the literature:

- Hamming Embedding (HE) [24], which assigns hamming codes to LIPs as binary signatures for addressing the visual ambiguity introduced by soft-weighting (SOFT) [27] so as to improve the quality of the quantization and speed up the matching process at the same time;
- Weak Geometric Consistency (WGC) [13], which verifies the weak or partial geometric consistency between the query and target images for re-ranking;
- The enhanced WGC (EWGC) [14], which enhances the geometric consistency based reranking by further including translation information into WGC;
- Synonym-based Query Expansion (Syn-Expan) [30],[8] which defines the visual words that frequently co-occur in the same circular regions (cf. Section IV) as synonyms to each other and uses synonyms to conduct query expansion for improving the performance of LIPs+BoW framework;
- Delaunay Triangulation (DT) [15], the method reported with the highest performance in literature, which proposes to model the geometric relations among visual words in a more principled way by using Delaunay Triangulation, and thus makes the mapping between images more robust. The "burstiness" of visual words (BUR) [22] is also considered in this method.

According to our study on **Basic Set**, the performances of most methods are approaching optimal when the vocabulary is set to the range of $10k$ to $40k$ (no significant improvement when a larger vocabulary is used). Therefore, to be fair and consistent, for the experiments hereafter, we unify the vocabulary size to $20k$ for all the methods under investigation. Note the size is 10 times larger than the one we used in Section V-B (i.e., $2k$). Furthermore, to ease the discussion, we add a subscript for a method to indicate its vocabulary size (e.g., $SIFT^{CC}_{20k}$). All methods are implemented in C++ and all the experiments are conducted on a station with Intel Xeon(R) $2.67G$ Hz and $30G$ memory. The performances of all methods are optimized (with the condition that the vocabulary size equals to $20k$) with the configurations summarized in Table II. Note that to investigate the effects of BQP and Synonyms (i.e., different ways of using spatial context), we have not combined HE and other advanced techniques into Syn-Expan. This makes the performance of Syn-Expan lower than those reported in [30]. To fully investigate the nature of these methods, we compare them on three datasets, 1) **Basic Set**, the dataset with the largest number of fully annotated queries so far. It is employed for investigating the basic characteristics of each method; 2) **Kentucky** [37], a dataset similar to **Basic Set** which is composed of mostly product images and has been popularly employed in previous studies. It is employed for testing the generalizability of each method; and 3) **Oxford** [38], which is another standard benchmark which is composed of building images. It is employed for conducting the cross-domain experiments.

TABLE II
CONFIGURATIONS OF METHODS UNDER INVESTIGATION

Method	Configuration	Feature
$SIFT^{CC}_{20k}$	BoW+BQP	SIFT
$SIFT_{20k}$	BoW	SIFT
CM^{CC}_{20k}	BoW+BQP	Color-Moment
CM_{20k}	BoW	Color-Moment
HE	BoW+SOFT+HE	SIFT
WGC	BoW+SOFT+HE+WGC	SIFT
EWGC	BoW+SOFT+HE+EWGC	SIFT
Syn-Expan	BoW+Synonyms	SIFT
DT	BoW+SOFT+HE+DT+BUR	SIFT

*1) General Comparison on **Basic Set**:* The performance comparison of accuracy (MAP) and efficiency is shown in Table III. In terms of accuracy, $SIFT^{CC}_{20k}$ outperforms all conventional approaches, while, surprisingly, CM^{CC}_{20k} which employs much simpler feature than SIFT outperforms $SIFT^{20k}$, Syn-Expan and EWGC, and demonstrates comparable performance to WGC and DT. In addition, by comparing the performance of $SIFT_{20k}$ (MAP: 0.4080) to $SIFT^{CC}$ (MAP: 0.4332) (with $2k$ vocabulary) or CM_{20k} (MAP: 0.3894) to CM^{CC} (MAP: 0.4081), it is clear that with the help of BQP quantization, we can reduce the vocabulary size by 1,000% times without losing accuracy. Fig. 9 shows a more detailed comparison of MAP over top-k, where $SIFT^{CC}_{20k}$ and CM^{CC}_{20k}'s superiorities over other methods are more evidential at the range of top-10 retrieved images. Given the fact that there are on average 7.22 NDs for each query, this indicates that most of the NDs can be located at top-10. As mentioned in Section V-B, this is a highly desired property in merchandize image retrieval. Moreover, with simply color-comment feature, which was conventionally recognized as a much weaker feature than SIFT, CM^{CC}_{20k} earns a surprising performance without employing any advanced techniques (e.g., geometric constraint or hamming embedding). While this has again confirmed the advantage of the proposed BQP quantization, we should see that this is also due to the specificity of the merchandize

[8] We thank Drs. Zhang Lei and Cai Rui, the authors of [30], for sharing the binary code and helping with the experiments.

TABLE III

PERFORMANCE COMPARISON WITH THE STATE-OF-THE-ART ON **BASIC SET**. THE BEST RESULTS ARE BOLD

	$SIFT_{20k}^{CC}$	CM_{20k}^{CC}	$SIFT_{20k}$	CM_{20k}	WGC	EWGC	HE	Syn-Expan	DT
Mean Average Precision (MAP)	**0.5578**	0.5215	0.4080	0.3894	0.5404	0.5144	0.5352	0.4834	0.5406
Time Cost (milliscond/query)	231	**128**	246	134	1895	1604	705	367	3411

TABLE IV

PERFORMANCE COMPARISON WITH THE STATE-OF-THE-ART ON **KENTUCKY**. THE BEST RESULTS ARE BOLD

	$SIFT_{20k}^{CC}$	CM_{20k}^{CC}	$SIFT_{20k}$	CM_{20k}	WGC	EWGC	HE	Syn-Expan	DT
Mean Average Precision (MAP)	0.8491	**0.8780**	0.6238	0.6039	0.8645	0.8430	0.8608	0.6944	0.8586
Time Cost (milliscond/query)	35	**33**	38	34	150	163	87	42	264

TABLE V

PERFORMANCE COMPARISON WITH THE STATE-OF-THE-ART ON **OXFORD**. THE BEST RESULTS ARE BOLD

	$SIFT_{20k}^{CC}$	CM_{20k}^{CC}	$SIFT_{20k}$	CM_{20k}	WGC	EWGC	HE	Syn-Expan	DT
Mean Average Precision (MAP)	0.5242	0.3835	0.3861	0.2642	**0.5438**	0.5087	0.4975	0.4267	0.5349
Time Cost (milliscond/query)	52	**45**	54	57	330	341	186	82	420

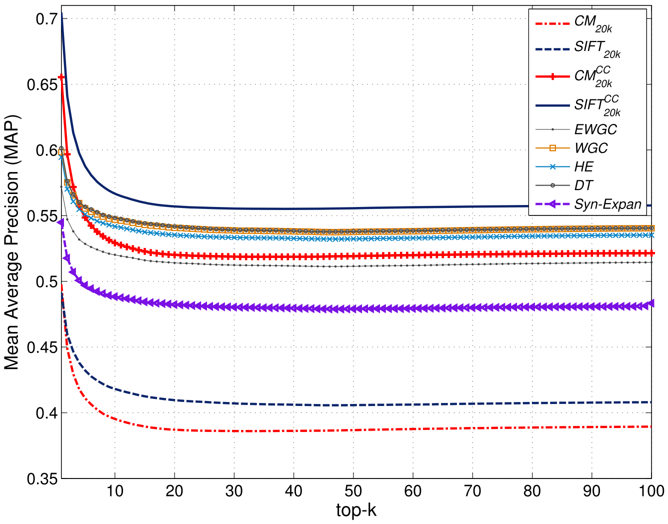

Fig. 9. Evaluation of MAP at the top-k of the ranked lists.

images. As color is an important feature for most of the products, the photographers usually try their best to maintain real color of the product when taking the pictures. However, most of the local feature (e.g., SIFT, SURF) has ignored the color, making their representations discount a lot. Therefore, color moment appears a better choice for such images.

In terms of time cost per query, as shown in Table III, $SIFT_{20k}^{CC}$ and CM_{20k}^{CC} are obviously more efficient than conventional approaches, while demonstrating comparable accuracy. It is worth mentioning that by including additional BQP quantization on $SIFT_{20k}$ and CM_{20k} respectively, the time cost for $SIFT_{20k}^{CC}$ and CM_{20k}^{CC} have not been increased but rather decreased. The reason is that with the BQP quantization, the errors/typos (i.e., the mis-assigned visual words) have been removed, resulting in a smaller and sparser set of visual words for representing each image, and thus saves the computational cost. For example, in CM^{CC}, the sparsity of the CM feature vectors has been improved from 15.07% to 11.24%, which can save about 1/3 of the time cost for calculating the similarities among feature vectors (the most expensive process in BoW framework since numerous times of multiplications are needed for computing the inner products). The saved cost is larger than the additional cost for BQP (only 27 milliseconds) and thus reduces the total time cost per query. Furthermore, considering both accuracy and efficiency, we can see that under certain scenario like merchandize image retrieval, CM_{20k}^{CC} can be used to replace some conventional methods based on complicated SIFT feature (e.g., to replace EWGC so as to obtain comparable performance while reducing the time cost by 9,202%). In terms of space complexity, one may have concern on the big matrix I that stores the collocative relations among words. Our experimental results show that the sparsity of the matrix is less than 10%, because only small amount of the word-pairs can pass the $\mathcal{X}^2$ test to be recognized as collocations. Therefore, at most 40M of the memory is needed for the matrix (in floating point numbers) with a 20k vocabulary.

2) Comparison of Generalizability on ***Kentucky****:* To test the generalizability of each method, we extend the performance comparison to **Kentucky** benchmark [37], which consists of 10,200 product images forming 2,550 ND groups (4 images for each group). We use each of the images as a query, resulting in 10,200 queries. The **Training Set** is used for training in $SIFT_{20k}^{CC}$, CM_{20k}^{CC}, and Syn-Expan. The results are shown in Table IV, where the performances of the methods under investigation are basically consistent with those on the **Basic Set**. This is not surprising, because the product images of **Kentucky** are naturally similar to the merchandize images of **Basic Set**. However, CM_{20k}^{CC} obtains the highest performance this time, even better than $SIFT_{20k}^{CC}$, WGC and DT in terms of MAP. The reason is that the scale variance on **Kentucky** is not as serious as that on **Basic Set**, because the same object usually occupies the similar portion of an image and the images of **Kentucky** are all with the same size. This helps the color-moment-based CM_{20k}^{CC} bypass the requirement for scale invariant (i.e., its biggest disadvantage compared with the SIFT-based features), and at the same time fully exhibit its strength in merchandize image retrieval (as we have learnt on **Basic Set**). Due to the same reason, the performances of spatial verification-based methods (e.g., WGC, EWGC, DT) are not as good as expected, indicated by the fact that there are no significant improvements observed from their MAPs when compared to that of HE. However, this observation is consistent with that reported in [13].

*3) Cross-Domain Experiments on **Oxford**:* To investigate whether the BQP works on datasets other than merchandize images, we conduct cross-domain experiments on **Oxford** [38], which is another standard benchmark consisting of 5,062 images of particular Oxford landmarks, with 55 query images and manually labeled ground truth. An additional set of 100,071 images (**Flickr 100k**) also provided by the authors of **Oxford** [38] is used for training in $SIFT_{20k}^{CC}$, CM_{20k}^{CC}, and Syn-Expan. The results are shown in Table V, where the performance of CM_{20k}^{CC} drops significantly due to fact that compared to **Basic Set** and **Kentucky**, the scale variance issue becomes much more serious on **Oxford**. However, by enforcing contextual consistency with BQP, $SIFT_{20k}^{CC}$ still outperforms some state-of-the-art methods such as HE, Syn-Expan, EWGC, while approaching WGC and DT, even without employing any advanced techniques like hamming embedding and soft-assignment.

VI. Conclusion

In this paper, we have studied the use of contextual consistency constraint for NDR to fulfill the critical requirement for response time when searching merchandize images. We have extended the concept of "collocation" in text domain to NDR for modeling the contextual relations among visual words, on the basis of which we use binary quadratic programming (BQP) to enforce the visual words selected for representing an image to be contextually consistent to each other so as to improve the quality of the quantization. The experimental results have validated the effectiveness and efficiency of the proposed method.

In terms of online application, one of the most favorable findings in this paper is that, for merchandize image retrieval, the conventional local features (e.g., SIFT) can be replaced by patch-wise feature based on color-moment, which can significantly improve the efficiency. While encouraging, we should see that the patch-wise color-moment feature is not scale invariant. It may cause problem when the scale change exceeds a certain degree (even this rarely happens for merchandize images). Further incorporating a pyramid based matching process into the method may address the problem. We will investigate this in our future work.

References

[1] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proc. 9th IEEE Int. Conf. Comput. Vis.*, vol. 2. Oct. 2003, pp. 1470–1477.

[2] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman, "Total recall: Automatic query expansion with a generative feature model for object retrieval," in *Proc. IEEE 11th ICCV*, Nov. 2007, pp. 1–8.

[3] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Nov. 2006, pp. 2161–2168.

[4] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.

[5] H. Jégou, H. Hedi, and C. Schmid, "A contextual dissimilarity measure for accurate and efficient image search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.

[6] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Jan. 2004.

[7] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Comput. Vis. Image Understand.*, vol. 110, no. 3, pp. 346–359, 2008.

[8] J. Yuan, L.-Y. Duan, Q. Tian, and C. Xu, "Fast and robust short video clip search using an index structure," in *Proc. ACM SIGMM Int. Workshop Multimedia Inf. Retr.*, 2004, pp. 61–68.

[9] E. Y. Chang, J. Z. Wang, C. Li, and G. Wiederhold, "RIME: A replicated image detector for the World-Wide Web," in *Proc. Symp. Voice, Video, Data Commun.*, 1998, pp. 58–67.

[10] A. Hampapur, K. Hyun, and R. M. Bolle, "Comparison of sequence matching techniques for video copy detection," in *Proc. Conf. Storage Retr. Media Databases*, Dec. 2002, pp. 194–201.

[11] R. Fergus, P. Perona, and A. Zisserman, "Object class recognition by unsupervised scale-invariant learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 2. Jun. 2003, pp. 264–271.

[12] F. Wang and C.-W. Ngo, "Rushes video summarization by object and event understanding," in *Proc. ACM Int. Workshop TRECVID Video Summarizat.*, 2007, pp. 25–29.

[13] H. Jégou, M. Douze, and C. Schmid, "Improving bag-of-features for large scale image search," *Int. J. Comput. Vis.*, vol. 87, no. 3, pp. 316–336, Feb. 2010.

[14] W.-L. Zhao, X. Wu, and C.-W. Ngo, "On the annotation of web videos by efficient near-duplicate search," *IEEE Trans. Multimedia*, vol. 12, no. 5, pp. 448–461, Aug. 2010.

[15] W. Zhang, L. Pang, and C.-W. Ngo, "Snap-and-ask: Answering multimodal question by naming visual instance," in *Proc. 20th ACM Int. Conf. Multimedia*, 2012, pp. 609–618.

[16] D.-Q. Zhang and S.-F. Chang, "Detecting image near-duplicate by stochastic attributed relational graph matching with learning," in *Proc. 12th ACM Int. Conf. Multimedia*, 2004, pp. 877–884.

[17] Y. Ke, R. Sukthankar, and L. Huston, "An efficient parts-based near-duplicate and sub-image retrieval system," in *Proc. 12th Annu. ACM Int. Conf. Multimedia*, 2004, pp. 869–876.

[18] Z. Wu, Q. Ke, M. Isard, and J. Sun, "Bundling features for large scale partial-duplicate web image search," in *Proc. IEEE Conf. CVPR*, Jun. 2009, pp. 25–32.

[19] W. Zhou, Y. Lu, H. Li, Y. Song, and Q. Tian, "Spatial coding for large scale partial-duplicate web image search," in *Proc. ACM Int. Conf. Multimedia*, 2010, pp. 511–520.

[20] A. Joly, O. Buisson, and C. Frelicot, "Content-based copy retrieval using distortion-based probabilistic similarity search," *IEEE Trans. Multimedia*, vol. 9, no. 2, pp. 293–306, Feb. 2007.

[21] M. S. Extremal, J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide baseline stereo from," in *Proc. British Mach. Vis. Conf.*, Sep. 2002, pp. 384–393.

[22] H. Jégou, M. Douze, and C. Schmid, "On the burstiness of visual elements," in *Proc. IEEE Conf. CVPR*, Jun. 2009, pp. 1169–1176.

[23] P. Jain, B. Kulis, and K. Grauman, "Fast image search for learned metrics," in *Proc. IEEE Conf. CVPR*, Jun. 2008, pp. 1–8.

[24] H. Jegou, M. Douze, and C. Schmid, "Hamming embedding and weak geometric consistency for large scale image search," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 304–317.

[25] O. Chum, J. Philbin, M. Isard, and A. Zisserman, "Scalable near identical image and shot detection," in *Proc. ACM Int. Conf. Image Video Retr.*, 2007, pp. 549–556.

[26] O. Chum, J. Philbin, and A. Zisserman, "Near duplicate image detection: Min-Hash and TF-IDF weighting," in *Proc. British Mach. Vis. Conf.*, 2008, pp. 812–815.

[27] Y.-G. Jiang, C.-W. Ngo, and J. Yang, "Towards optimal bag-of-features for object categorization and semantic video retrieval," in *Proc. 6th ACM Int. Conf. Image Video Retr.*, 2007, pp. 494–501.

[28] J. C. van Gemert, C. J. Veenman, A. W. M. Smeulders, and J.-M. Geusebroek, "Visual word ambiguity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 7, pp. 1271–1283, Jul. 2010.

[29] R. A. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*. Boston, MA, USA: Addison-Wesley, 1999.

[30] W. Tang, R. Cai, Z. Li, and L. Zhang, "Contextual synonym dictionary for visual object retrieval," in *Proc. ACM Int. Conf. Multimedia*, 2011, pp. 503–512.

[31] D. Pearce, "A comparative evaluation of collocation extraction techniques," in *Proc. Int. Conf. Lang. Resour. Eval.*, Jul. 2002, pp. 1–7.

[32] Y. Jiang, "Randomized visual phrases for object search," in *Proc. IEEE Conf. CVPR*, Nov. 2012, pp. 3100–3107.

[33] J. Yuan, Y. Wu, and M. Yang, "Discovery of collocation patterns: From visual words to visual phrases," in *Proc. IEEE Conf. CVPR*, Jun. 2007, pp. 1–8.

[34] Y. Huang, S. Shekhar, and H. Xiong, "Discovering colocation patterns from spatial data sets: A general approach," *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 12, pp. 1472–1485, Dec. 2004.

[35] G. L. Nemhauser and L. A. Wolsey, *Integer and Combinatorial Optimization*. New York, NY, USA: Wiley, 1988.

[36] M. A. Stricker and M. Orengo, "Similarity of color images," in *Proc. Symp. Electron. Imaging, Sci. Technol.*, 1995, pp. 381–392.

[37] D. Nistér and H. Stewénius, "Scalable recognition with a vocabulary tree," in *Proc. IEEE Conf. CVPR*, vol. 2. Nov. 2006, pp. 2161–2168.

[38] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Proc. IEEE Conf. CVPR*, Jun. 2007, pp. 1–8.

Xiao-Yong Wei (M'10) is an Associate Professor with the College of Computer Science, Chengdu, Sichuan University, China. His research interests include multimedia retrieval, data mining, and machine learning. He received the Ph.D. degree in computer science from the City University of Hong Kong in 2009. He is one of the founding members of the VIREO Multimedia Retrieval Group, City University of Hong Kong. He was a Senior Research Associate with the Department of Computer Science and Department of Chinese, Linguistics and Translation, City University of Hong Kong, in 2009 and 2010, respectively. He was a Manager with the Software Department, Para Telecom Ltd., China, from 2000 to 2003.

Zhen-Qun Yang is currently pursuing the Ph.D. degree with the College of Computer Science, Sichuan University, Chengdu, China. She was a Research Assistant with the Department of Computer Science, City University of Hong Kong, from 2007 to 2008, and a Senior Research Assistant with the Department of Chinese, Linguistics and Translation, City University of Hong Kong, in 2009. Her research interests include multimedia retrieval, image processing and pattern recognition, and neural networks.

Chong-Wah Ngo (M'02) received the B.Sc. and M.Sc. degrees in computer engineering from Nanyang Technological University, Singapore, and the Ph.D. degree in computer science from the Hong Kong University of Science and Technology, Hong Kong.

He was a Post-Doctoral Scholar with the Beckman Institute, University of Illinois at Urbana-Champaign, Champaign. He was a Visiting Researcher with Microsoft Research Asia. He is currently an Associate Professor with the Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong. His current research interests include large-scale multimedia information retrieval, video computing, and multimedia mining.

Dr. Ngo is currently an Associate Editor of the IEEE TRANSACTIONS ON MULTIMEDIA. He is the Conference Co-Chair of the ACM International Conference on Multimedia Retrieval 2015 and the Pacific Rim Conference on Multimedia 2014, and the Program Co-Chair of the ACM Multimedia Modeling Conference 2012 and the ACM International Conference on Multimedia Retrieval 2012, and the Area Chair of the ACM Multimedia 2012. He was the Chairman of the ACM (Hong Kong Chapter) from 2008 to 2009.

Wei Zhang received the B.Eng. degree from the School of Computer Software and the M.Eng. degree from the School of Computer Science and Technology, Tianjin University, Tianjin, China, in 2008 and 2011, respectively. He is currently pursuing the Ph.D. degree with the Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong.

He was a Former Member of the CV-TJU Laboratory, Tianjin University, from 2008 to 2011. He is currently with the VIREO Group, City University of Hong Kong. His research interests include large scale video retrieval and digital forensic analysis.