IEEE TRANSACTIONS ON
IMAGE PROCESSING

IEEE Xplore® Digital Library
http://ieeexplore.ieee.org/Xplore

Contextual Noise Reduction for Domain Adaptive Near-Duplicate Retrieval on Merchandize Images

Zhen-Qun Yang, Xiao-Yong Wei, *Member, IEEE*, Zhang Yi*, *Fellow, IEEE*, Gerald Friedland, *Senior Member, IEEE*

***Abstract*—In this paper, we have proposed a novel method which utilizes the contextual relationship among visual words for reducing the Quantization errors in near-duplicate image retrieval (NDR). Instead of following the track of conventional NDR techniques which usually search new solutions by borrowing ideas from the text domain, we propose to model the problem back to image domain, which results in a more natural way of solution search. The idea of the proposed method is to construct a context graph that encapsulates the contextual relationship within an image and treat the graph as a pseudo-image, so that classical image filters can be adopted to reduce the mis-mapped visual words which are contextually inconsistent with others. With these contextual noises reduced, the method provides purified inputs to the subsequent processes in NDR, and improves the overall accuracy. More importantly, the purification further increases the sparsity of the image feature vectors, which thus speeds up the conventional methods by 1662% times and makes NDR practical to online applications on merchandize images where the requirement of response time is critical. The way of considering contextual noise reduction in image domain also makes the problem open to all sophisticated filters. Our study shows the classic anisotropic diffusion filter can be employed to address the cross-domain issue, resulting in the superiority of the method to conventional ones in both effectiveness and efficiency.**

***Index Terms* — near-duplicate retrieval; contextual noise reduction; anisotropic diffusion.**

I. Introduction

THANKS to the prevalence of the e-Commerce nowadays, the merchandize imagery is becoming one of the rapidly growing genres of web images. However, according to the study on eBay.com and Taobao.com [1], over 42.6% of those images are indeed duplicates or near-duplicates. Therefore, near-duplicate retrieval (NDR), which helps to locate those duplicates, has been intensively studied in recent years, due to its potential to a wide range of commercial applications. For example, it can be used to build price comparison services for websites like PriceGrabber, Shopping.com and Shopzilla, where NDR is especially helpful for locating products selling on websites of different countries and described in different languages on which text retrieval works awkwardly. Under similar scenarios, NDR can also provide valuable clues for pirated product detection, unauthorized seller identification, market potential evaluation and so on.

A general purpose NDR framework is usually composed of local feature extraction and Bag-of-Words (BoW) indexing [4–8]. More specifically, local interest points (LIPs), which are the salient points whose features are invariant to the changes of scale, illumination, and viewpoint [9, 10], are extracted from an image and converted to "visual words" that mimic the words in the text documents, and thus the images can be indexed with BoW, and retrieved with the sophisticated techniques of traditional information retrieval. While the effectiveness of the LIPs+BoW framework has been recognized for general purpose NDR, however, it has seldom been studied on merchandize imagery domain.

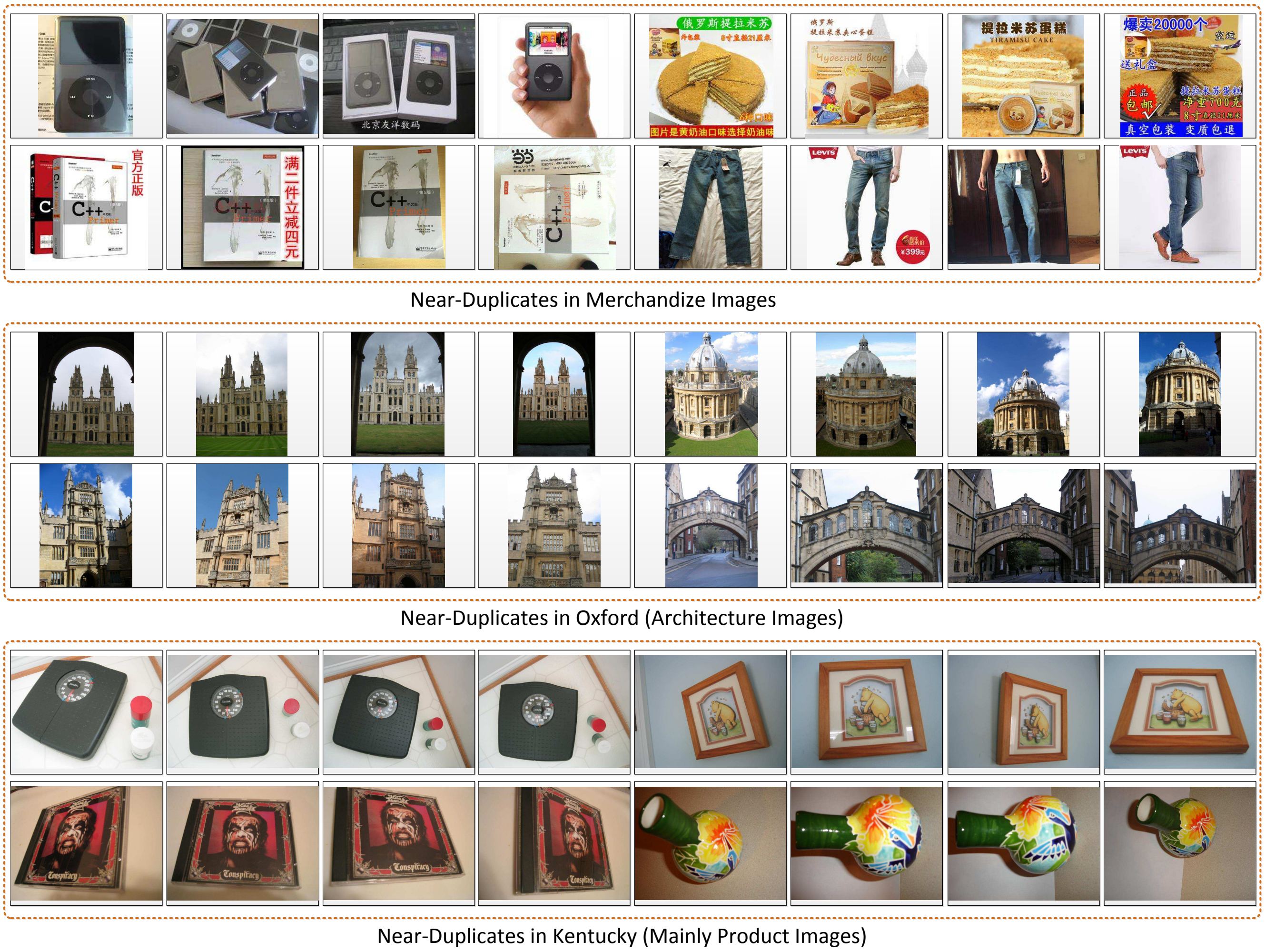
To adopt NDR in merchandize imagery domain, new challenges are resulting from the *scalability* and *diversity* of the datasets. First, the datasets are usually in web-scale which are overwhelmingly larger than the conventional ones used for the general purpose NDR. For instance, the merchandize imagery dataset used in [1] is in million-scale while the standard datastes (for general purpose NDR) such as Kentucky [3] and Oxford [2] are only in thousand-scale. The dramatically increased scale makes many (computationally) expensive NDR techniques infeasible, because the requirement of response time in the aforementioned online or mobile applications is critical. For example, geometric constraint verification [11–14], which is popularly employed to verify the geometrical consistency of the distributions of the visual words when comparing two images, is considered impractical in these applications, since the word-to-word comparison is required in the verification process which is with quadratic time complexity ($O(n^2)$). Second, the datasets are often composed of images from a diverse range of sub-domains (e.g., in Figure 1, the merchandize images include those of *electronics*, *food*, *books*, and *clothes*) rather than from a specific domain as in conventional datasets (e.g., in Figure 1, Oxford is composed of *architecture* images while Kentucky is mainly composed of *product* images). Furthermore, the datasets are with user-generated content (UGC) in which the image quality (even for the same product) varies significantly, because they are captured by various photographers from professionals to armatures, by devices from high resolution DCs to web cameras (thus to cause so-called sensor gap), under various conditions from photographic studios to randomly arranged outdoors, and

The work described in this paper was jointly supported by the grants from the National Natural Science Foundation of China (No. 61272256 and No. 61432012), the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344, and the National Science Foundation by ICSI under Award Number 1251276.

Zhen-Qun Yang, Xiao-Yong Wei and Zhang Yi are with the College of Computer Science, Sichuan University, Chengdu 610065, China (emails: {jessicayang, cswei, zhangyi}@scu.edu.cn). Xiao-Yong Wei and Gerald Friedland are with ICSI, Berkeley, California 94704, US (emails: {xiaoyong,fractor}@icsi.berkeley.edu).

* Corresponding author.

Digital Object Identifier: 10.1109/TIP.2017.2669842



Near-Duplicates in Merchandize Images

Near-Duplicates in Oxford (Architecture Images)

Near-Duplicates in Kentucky (Mainly Product Images)

Fig. 1: Near-Duplicate examples of merchandize imagery dataset [1], Oxford [2], and Kentucky [3].

modified with unpredictably heterogeneous styles. Compared to the conventional datasets which are usually built in a controlled environment (e.g., Kentucky images are collected in a laboratory space with the same DC and consistent illumination), therefore, the diversity of merchandize images leads to a much larger variance of the image features (for the same product) and thus makes the matching inherently much more difficult.

In terms of the LIPs+BoW framework, the challenges caused by the domain shift will affect the indexing in different ways. From a feature-level perspective, it results in "noises" among the visual words of an image, because the process of converting of a LIP to a visual word can easily be wrong when the feature of the LIP varies a lot from its original form (e.g., when affected by reflections, a LIP from a metal surface can be mistakenly converted to a visual word that representing glass-liked materials). From a domain-level perspective, due to the multi-domain nature of the merchandize datasets, LIPs from "unseen" sub-domains are inevitable and thus cannot be converted correctly (e.g., a LIP from a building will definitely be converted to a wrong visual word if there are no examples of *architecture* have been selected to train the codebook[1]).

To reduce the risk of the mis-mapping, a commonly adopted method is soft weighting [15, 16]. Rather than mapping a LIP to a single visual word, soft weighting maps a LIP to multiple words, in the hope that if one word is mis-mapped, the others could still represent the image correctly. However, since the multiple mapping scheme requires additional computational cost for mapping and comparison, it is impractical to large-scale datasets (i.e., the challenge of *scalability*). Furthermore, it increases the ambiguity of the feature because more words are used for representing a LIP. This becomes more serious when the words are from different domains (i.e., the challenge of *diversity*, a nature of the merchandize imagery datasets).

In this paper, we propose a novel quantization method for addressing both the *scalability* and *diversity* issues. Instead of being a post-process (to the LIPs+Bow framework) to "compensate" or a method using multiple words to "share" the risk of mis-mapping, the proposed method corrects the "errors" directly by verifying the contextual relationship among the mapped words and eliminating those contextually inconsistent

[1] In LIPs+BoW framework, a codebook (also called vocabulary in literature) is a dictionary of visual words and is the reference for the LIP-to-Word conversion.

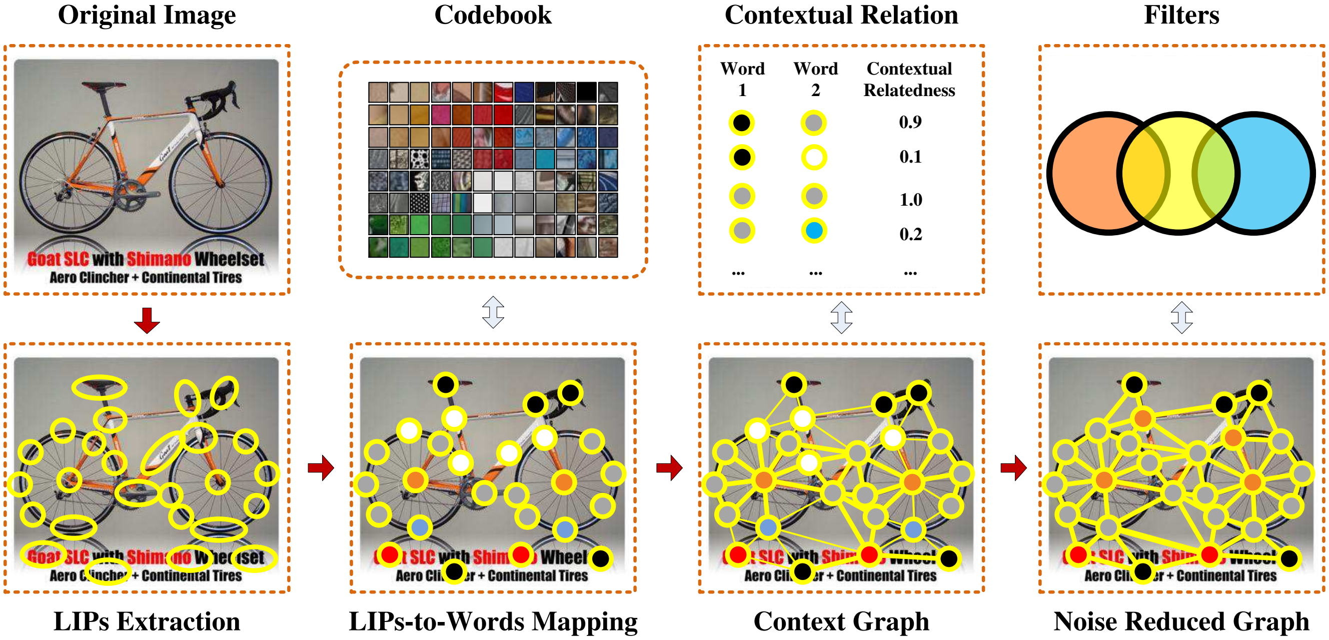


Fig. 2: Framework of the contextual noise reduction: 1) LIPs are extracted from the original image; 2) LIPs are mapped to visual words in the codebook; 3) Context Graph is constructed with the visual words as vertexes and edges are weighted by the contextual relatedness among words; 4) the visual words that are contextually inconsistent to others are removed by filters, and the final graph thus reaches a state of contextual harmony.

ones to others. More specifically, as shown in Figure 2, we first construct a context graph for each image with visual words as vertexes and their contextual relatedness as edge weights, and then treat the context graph as an "image" on which the visual words that are contextually inconsistent to others can be corrected by employing the sophisticated filters of image noise reduction. Compared to conventional methods, the advantages of the proposed method are as follows:

- Complexity Reduction: Existing methods for addressing the mis-mapping problem are usually with time complexity of $O(n^2)$ [11–14] or even NP-hard [1]. By simplifying the model into the "image" noise reduction on a context graph, the proposed method archives an approaching linear complexity (c.f. Section III-C), and is thus more practical for large-scale NDR.
- Domain Adaptability: For the multi-domain nature of the merchandize datasets, we address the issue by transforming it to an edge-preserving problem in image noise reduction and employing the classic Anisotropic Diffusion [17] as a solution. As demonstrated in Section IV, this gives the method a better domain adaptability.
- Extendibility: By transforming the contextual consistency optimization into an image noise reduction problem, in this paper, we propose not only a novel method for addressing the problem, but also a new way of seeking solutions, in the sense that image noise reduction has been intensively studied and there is a great amount of sophisticated techniques can be borrowed. For example, methods of adaptive noise reduction (e.g., [18]) can be used for improving the performance and neural networks can be employed for improving the efficiency (e.g., [19]).

The remainder of this paper is organized as follows. Section II reviews the LIPs+BoW framework and NDR techniques in the literature. Section III relates the problem of contextual noise reduction to the classical image noise reduction. In Section IV, we address the cross-domain problem by adapting the classic anisotropic diffusion. The experimental results will be presented in Section V. Finally, Section VI concludes this paper.

II. Related Work

A. *LIPs+BoW framework for NDR and Recent Trends*

LIPs+BoW might be the most popular and promising framework for NDR, which consists of four processes: LIPs Extraction, Codebook Learning, Quantization, and Indexing. Starting with LIPs Extraction, LIPs from images are pooled and clustered in Codebook Learning, where a model LIP is calculated for each cluster using mean or medoid of the member LIPs, and all model LIPs are then treated as visual words with which the codebook is established. The Codebook Learning is usually conducted only once as a training process. For every image, the Quantization process maps its LIPs to visual words in the codebook, so that the image can be indexed like a text document consisting of text words [4–8]. The most straightforward method for Quantization is to assign for each LIP a visual word that is with the largest similarity among the other words in the codebook. In Indexing, images are indexed into inverted files on which NDR can be carried out in the almost the same way as conventional text retrieval.

Due to its heritage from conventional text information retrieval, LIPs+BoW framework is straightforward and sophisticated, and has been successfully applied to a wide range of applications (e.g., [7, 20–24]. However, ever since the volumes of real world image datasets are growing up to web-scale, the framework has been challenged constantly, which makes the efforts to improve its efficiency and accuracy the main drive behind the recent trends in NDR. As aforementioned, the essential challenge to efficiency lies in the expensive word-to-word comparison process when matching two images. The main stream for addressing this issue is to build better hashing functions which map LIPs into a much lower dimensional space so as to speed up the matching process. Representative works include locality sensitive hashing (LSH) [25, 26], hamming embedding (HE) [11, 27], and min-Hash [28, 29]. Meanwhile, intensive studies have also been conducted to improve the accuracy by verifying geometrical consistency between two sets of visual words when matching two images. To this end, every image is converted into a graph in which its visual words are related to their neighbors to encapsulate their geometrical relationship, so that image comparison is transformed into a graph matching problem, which leads to higher accuracy [11–14]. Alternatively, to mitigate some of the frequently occurring visual words from dominating the others, "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 [30].

B. Addressing The Core Issue of The Framework

Despite the success of the methods mentioned above (either efficiency or accuracy oriented), they are not addressing the core issue of the LIPs+BoW framework directly, but rather providing compromised solutions. The core issue with the framework is that the LIPs or visual words of an image are always considered individually with their geometrical/contextual relationship ignored. When this happens in Quantization, it results in mis-mapped words that are inconsistent with others, which can seriously degenerate the quality of the subsequent processes and finally causes accuracy decline. It can also degrade the efficiency, in the way that when visual words are utilized individually, a NDR system has to adopt as many words as possible to enrich the representativeness of the codebook so that the natures of the images can be characterized explicitly. This often leads to large-scale codebooks consisting of hundreds of thousands words (e.g., $100k$ in [11] and $1M$ in [7]), which dramatically increases the time for Quantization. The efficiency and accuracy oriented methods mentioned above are all compromised solutions, because they are all carried out at post-Indexing stage, where the Quantization has been finished and the information of the geometrical/contextual relationship has already been lost.

To reduce the mis-mapped words as early as possible, pre-Indexing methods are proposed to reduce the mis-mapped words in the Quantization so that the quality of the mapped words can be purified at the earliest stage of the framework and the accuracy can be improved. These include soft assignment methods [15, 16]. Instead of directly mapping each LIP to a single word, soft assignment maps each LIP to several words with each word assigned with a weight that is proportional to its significance to the LIP (e.g., soft-weighting [15], visual word ambiguity [16]). With the multiple assignment, soft assignment increases the chance of a LIP to be mapped to a correct word. However, it also brings extra computational cost or may even cause ambiguity for the processes after Quantization. Therefore, soft assignment is only applied to the query side in most of the applications.

In [1], the authors propose a pre-Indexing method in which a word is considered mis-mapped if it is not contextually consistent to others and thus should be "corrected". The idea has been implemented as to solve a combinational optimization problem of mapping the LIPs of an image to the best set of visual words which maximizes the overall contextual consistency. Despite its success, however, the method in [1] is still costly, because, to solve the combinational optimization problem, binary quadratic programming (BQP) has been employed which is essentially NP-hard. Furthermore, BQP is a global optimization process which will sacrifice local consistency of the image for global harmony and causes words from "unseen" sub-domains being replaced by those from the "known" sub-domains forcibly and mistakenly. This limits its application to merchandize images due to the aforementioned multi-domain nature of the datasets. In this paper, we further explore the idea of utilizing contextual relationship for Quantization enforcement and propose a new method which is complexity reduced, domain adaptive, and logically simplified when compared to [1]. More insights about the difference between the proposed method and the method in [1] will be given in Section V-B3.

III. Contextual Noise Reduction from An Image point of view

In this section, we review the problem of contextual noise reduction from an image point of view for a more intuitive solution search, and propose a new framework that adopts the sophisticated image filters for noise reduction.

A. Constructing Context Graph

Before introducing our method, let us first represent the contextual relationship within an image into a context graph, then the idea will tell its own tale. A context graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{R})$ is composed of a set $\mathcal{V} = \{v_i\}$ of vertexes (where the vertex v_i is placed on the the i-th LIP of the image), a set $\mathcal{E} = \{e_{ij}\}$ of edges, and a set $\mathcal{R} = \{r_{ij}\}$ of weights for edges. $\mathcal{G}$ is not necessarily a complete graph, in the way that a vertex is only connected to its neighbors within a circle of radius d pixels. With a codebook $\mathcal{B} = \{w_n\}$ consisting of a set of visual words, we set $\varpi(v_i) = w_n$ when a visual word w_n is assigned to the i-th LIP during Quantization. An edge e_{ij} between two vertexes is then weighted with their contextual relatedness defined by (point-wise) mutual information of the corresponding visual words as

$$r_{ij} = MI(v_i, v_j) = \log\left(\frac{p(\varpi(v_i), \varpi(v_j))}{p(\varpi(v_i)) \cdot p(\varpi(v_j))}\right) \tag{1}$$

where $p(\varpi(v_i), \varpi(v_j))$, $p(\varpi(v_i))$ and $p(\varpi(v_j)$ are the frequencies of the co-occurrence of $\varpi(v_i)$ and $\varpi(v_j)$, the occurrence of $\varpi(v_i)$, and the occurrence of $\varpi(v_j)$, respectively.

The contextual relatedness among visual words is learnt in advance on a training dataset. We are interested in word pairs that are statistically meaningful, in the sense that the contextual relatedness is defined with value of non-zero only when the two words co-occur in the neighboring area more often than would be expected by chance [31]. $\mathcal{X}^2$ test is employed to verify this criteria, which is the most popular test in the text domain for the same purpose. Any word pair fails in the test will be set with a relatedness value of 0[2]. Figure 3 gives an example of the context graph. More examples are available at our demo page[3].

B. Adopting Image Noise Reduction to Context Graph

With the context graph, let us imagine the visual words represent the pixels of an image and the edges indicate the reversed gradients among pixels, it is easy to see that the visual words which are contextually inconsistent with its neighbors are similar to the noise pixels that are with abrupt contrast to the its neighbors. Therefore, techniques of image noise reduction can be adopted to fix those contextual noises.

Image noise reduction is often conducted with a convolution process which adjusts the value of a pixel by iteratively fusing it with those of its neighbors. The rationale behind is that, in case most of the surrounding pixels are with values at a normal level, through the iterative fusion, the value of a noise pixel can be set back to normal gradually. It is obvious that the idea can be easily adopted to reduce the contextual noises on the context graph by gradually adjusting the values of the mis-mapped words towards the normal level. However, regarding the fundamental differences between a context graph and a real image, we have to make following modifications to the reduction process:

1) Definition of Neighbors: Let us relax the constraints that the neighbors of a pixels have to be those at its adjacent rows or columns, and define the neighbors of a visual word as the ones located in its neighboring area as in Section III-A (i.e., within a circle of radius d pixels);
2) Definition of Pixel Value: Instead of being a single value (e.g., gray scale) for an image pixel, a "pixel" on the context graph is a visual word which is associated with a list of candidate words that consists of the top-k words which have the largest similarities with the corresponding LIP;
3) The Convolution: In image noise reduction, the convolution process can be considered as an evolutionary process that constantly adjusts the value of the target pixel towards the goal of reducing its gradients to its neighbors. In this regard, we define the convolution on the context graph as a process which adjusts the value of a node by selecting from its candidate list a visual word which maximizes the contextual consistency of the node to its neighbors.

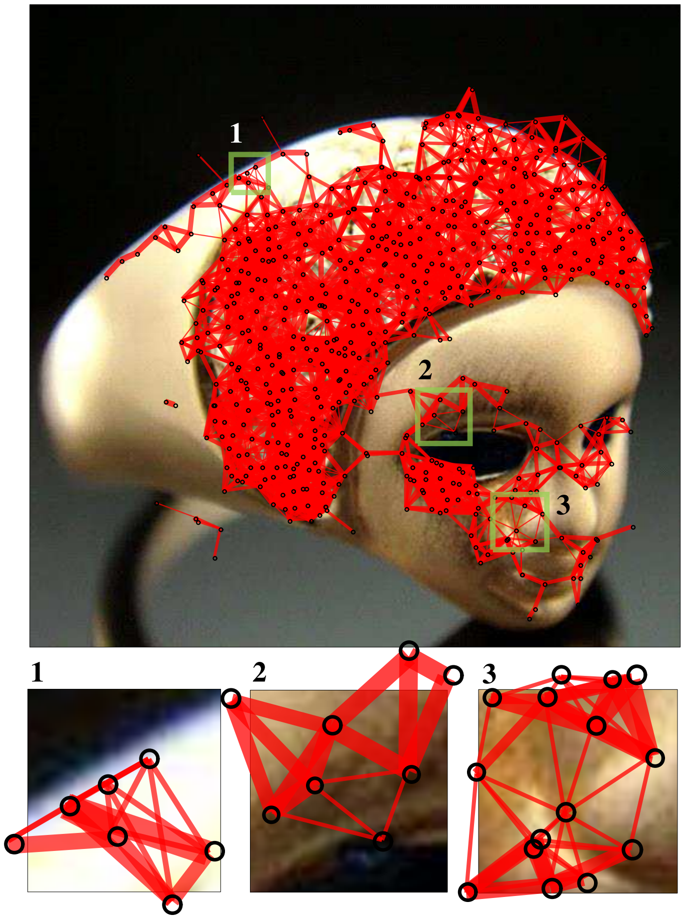


Fig. 3: Example of a context graph. Noise "pixels", which are contextually inconsistent with their neighbors, are indicated by the thin edges (e.g., those around the eyes and nose).

C. Contextual Noise Reduction

With the modifications, the contextual noise reduction can then be formulated as an iterative process. To facilitate the description, superscript will be hereafter used for iteration number. At each iteration (denoted as $\mathcal{I}^{(t)}$), we have a context graph $\mathcal{G}^{(t)}$ in which vertexes are assigned with visual words based on the result of the previous iteration ($\mathcal{I}^{(t-1)}$), and the graph weights are updated with the method introduced in Section III-A. The task of each iteration $\mathcal{I}^{(t)}$ is to select for each node a new visual word from its candidate list so as to maximize its contextual consistency to neighbors. That is,

$$\mathcal{I}^{(t)} : \quad \varpi^{(t)}(v_i) = \operatorname*{argmax}_{w \in \mathcal{C}(v_i)} \mathcal{H}\left(\varpi(v_i) = w \mid \mathcal{G}^{(t-1)}\right) \tag{2}$$

where $\mathcal{C}(\cdot)$ is a function for selecting the candidate list of visual words from the codebook $\mathcal{B}$ for a given vertex (i.e., the top-k most similar ones to the original LIP), and $\mathcal{H}\left(\varpi(v_i) = w \mid \mathcal{G}^{(t-1)}\right)$ is an object function to measure the neighborhood contextual consistency (harmoniousness) for the hypothesis of assigning a visual word w to the vertex v_i

[2] As the alternatives to learn the relationship, similar ideas can be found in [32, 33]

[3] http://www.cs.cityu.edu.hk/∼xiaoyong/product.images/context/

regarding the context graph $\mathcal{G}^{(t-1)}$ as

$$\mathcal{H}\left(\varpi(v_i)=w \mid \mathcal{G}^{(t-1)}\right)=\sum_{v_j\in\mathcal{N}(v_i)}\left\{\mathcal{F}(v_i,v_j)\cdot MI(v_i,v_j)\right\} \tag{3}$$

where $\mathcal{N}(v_i)$ is a function for selecting the set of neighbor vertexes of the target vertex v_i from the context graph, and $\mathcal{F}(v_i, v_j)$ servers as a filter function that determines the contribution or impact of a neighbor v_j to the target vertex v_i. With $\mathcal{F}(v_i, v_j)$ as an analogs to the noise filter in image domain, Eq.(3) can be considered a convolution defined on the context graph, which makes it convenient to adopt the existing image filter for contextual noise reduction. Let us simply take Gaussian filter as an example, which is one of the most popular image filters. $\mathcal{F}_g(v_i, v_j)$ can be defined for contextual noise reduction as

$$\mathcal{F}_G(v_i,v_j)=\frac{1}{2\pi\sigma^2}\exp\left(-\frac{\mathcal{D}(v_i,v_j)^2}{2\sigma^2}\right) \tag{4}$$

where the subscript G stands for Gaussian, σ is the standard deviation of the filter, and $\mathcal{D}(v_i, v_j)$ is a function to calculate the Euclidian distance between the two vertexes (based on their pixel coordinates). Once the word selection for the whole graph has been finished, the contextual relatedness (i.e., edge weights) are also recomputed accordingly. Through the context graph, the impact of the word selection of a node is indeed transferable to the other nodes in the graph, making the nodes "negotiate" to each other before they make decisions. This finally leads to a harmony state for the graph, in which the inter-word contextual consistency is ensured to a certain extent.

Note that the proposed reduction is compatible to almost any existing Quantization scheme. Therefore, as a starting point for the iteration, the initial results of word assignment for $\mathcal{I}^{(0)}$ can be obtained by conducting a conventional quantization (e.g., simply selecting the word with the largest similarity to the corresponding LIP). A formal definition of the process is given in Algorithm 1. It is easy to see the complexity of Algorithm 1 is $O(|\mathcal{C}(v_i)| \times m \times |\mathcal{V}|)$. According to our empirical study (cf. SectionV), the optimal settings for $\mathcal{C}(v_i)$ and m are $|\mathcal{C}(v_i)| \leq 10$ and $m \leq 5$, which makes $|\mathcal{C}(v_i)| \times m \ll |\mathcal{V}|$, and finally results in an approaching linear complexity of the algorithm (i.e., $O(c|\mathcal{V}|)$).

Algorithm 1 Iterative Contextual Noise Reduction

```
Input:
    The initial context graph, $\mathcal{G}^{(0)} = (\mathcal{V}, \mathcal{E}, \mathcal{R}^{(0)})$, where $\mathcal{V} = \{v_i\}$, $\mathcal{E} = \{e_{ij}\}$, and $\mathcal{R}^{(0)} = \{r_{ij}^{(0)}\}$;
    The codebook, $\mathcal{B} = \{w_n\}$ ;
    The contextual noise filter function, $\mathcal{F}(v_i, v_j)$;
    The maximum number of iterations, $m$;
Output:
    The resulting context graph, $\mathcal{G}^{(m)}$;
    The final word selection for all vertexes, $\{\varpi^{(m)}(v_i)\}$;
Conduct any conventional Quantization to obtain the initial visual selection, $\{\varpi^{(0)}(v_i)\}$;
Update edge weights $\{r_{ij}^{(0)}\}$ according to the word selection result, $\{r_{ij}^{(0)}\} = MI(v_i, v_j)$;
for $t = 1$ to $m$ do
    for $i = 1$ to $|\mathcal{V}|$ do
        Find the set of candidate visual words for current vertex $v_i$ from the codebook $\mathcal{B}$, $\mathcal{C}(v_i)$;
        For each candidate word $w$ in $\mathcal{C}(v_i)$, calculate the neighborhood contextual consistency using Eq.(3) by hypothetically assigning $w$ to $v_i$;
        Assign $v_i$ with the candidate word which maximizes the neighborhood contextual consistency, Eq.(2);
    end for
    Update the word selection $\{\varpi^{(t)}(v_i)\}$;
    Update edge weights $\{r_{ij}^{(t)}\} = MI(v_i, v_j)$;
    Update the context graph $\mathcal{G}^{(t)}$;
end for
return $\mathcal{G}^{(m)}$ and $\{\varpi^{(m)}(v_i)\}$.
```

IV. Domain Adaptive Contextual Noise Reduction

As shown in the previous section, by analyzing contextual noise reduction from an image point of view, it has greatly facilitated the solution search. We will see in this section that it can also provide a way for problem identification, in the sense that the edge preserving problem in image noise reduction is tightly related to domain-shift issue in contextual noise reduction.

A. Edge Preserving and Cross-Domain Issue

In the previous section, to determine the fusion weights, we adopt Gaussian filter for its popularity and simplicity. However, Gaussian filter is known to have the blurring effect problem that, while reducing the noise, Gaussian filter will blur the image and causes the loss of details. The most obvious impact of the blurring is that the edges of the image will dissolve into the adjacent regions. The reason causes the blurring of Gaussian filter is that it is isotropic, which means that, when fusing the neighbors into the target pixel, the contributions of the neighbors at different directions are considered equally without any "respects" to the edges. Therefore, the value of an edge pixel will be set to a weighted average of the pixels from the regions along the edge.

The reader might have the concern immediately that "*Will this affect the contextual noise reduction?*" Our empirical study shows the answer is YES. Ideally, we should expect the context graph to be as smooth as possible. In reality, however, the "edges" also exist on the graph, which is resulting from the domain shift from training to testing dataset. For example, in case that in the training dataset, there is only a limited number of images containing both *cars* and *trees*, the contextual relatedness between the visual words for the *cars* and those for the *trees* will thus tend to be low (cf. Eq.(1)). This will cause the *contextual* edges between these two types of visual words for a testing image which contains both *cars* and *trees*. In this case, using an isotropic filter like Gaussian will force the visual words along the contextual edges to be changed to those words that are with higher contextual relatedness to the words of both *cars* and *trees* (e.g., those of *buildings*, *sky*, *water* and so on). These mis-mapped edge words will significantly mislead the Indexing process and result in mis-matching during the retrieval. For instance, when visual words for *water* are used for indexing the images that do not contain *water*, there is a high chance for those images to be matched to images of *water*.

B. Anisotropic Diffusion

Fortunately, the problem of edge preserving has been intensively studied in image domain. The most popular method is called anisotropic diffusion which is proposed in 1990 by Perona and Mailik [17]. The idea is to construct an anisotropic filter which encourages the fusion between intra-region pixels and limits the fusion between inter-region pixels (i.e., those at two sides of an edge). By simply modifying the method proposed in [17], it is easy to construct an anisotropic filter for contextual noise reduction as

$$\mathcal{F}_A(v_i, v_j) = \frac{1}{\mathcal{D}(v_i, v_j)} \mathcal{P}(\|\nabla(v_i, v_j)\|) \tag{5}$$

where the subscript A stands for anisotropic diffusion, $\nabla(v_i, v_j)$ is the gradient from v_j to v_i which we can implement easily by reversing the contextual relatedness as

$$\nabla(v_i, v_j) = 1 - MI(v_i, v_j), \tag{6}$$

and $\mathcal{P}(\cdot)$ is the so-called conductance function[4] defined as

$$\mathcal{P}(\|\nabla(v_i, v_j)\|) = \exp\left[-(\frac{\|\nabla(v_i, v_j)\|}{\xi})^2\right] \tag{7}$$

where ξ is a constant which controls the fusion rate. In Eq.(8), the contribution of a neighbor is mainly determined based on its contextual gradient to the target vertex. The anisotropic diffusion is implemented with the non-increasing function $\mathcal{P}(\cdot)$ which encourages the fusion when the gradient is low (e.g., in the intra-region areas), and limits the fusion when the gradient is high so as to preserve the edges (where the inter-region gradient is high enough). This filter can be seamlessly integrated into the contextual noise reduction in Algorithm 1 by replacing the Gaussian filter.

One may doubt that, because pixels may have similarly large gradients on both edges and noises, how the scheme can reduce the noises and remain edges at the same time. To answer the question, we have to point out a fundamental difference between edge and noise pixels: *the edge pixels are comparably continuous and the noise pixels are usually isolated*. As illustrated in Figure 4, when the filter is on an edge pixel (e.g., C or D), the contributions of its neighbors on the same side of the edge will dominate those on the other side, and thus force the target pixel to be consistent with its intra-region neighbors. By contrast, when the filter is on a noise pixel (e.g., B or E), even the contributions of all the neighbors have been limited, their impacts on the target are even, so that the word selected is still the one which maximizes Eq. (3). There might be other anisotropic filters which perform better regarding the trade of between noise reduction and edge preserving. However, this is beyond the topic of this paper. The reader is referred to [34] for more details.

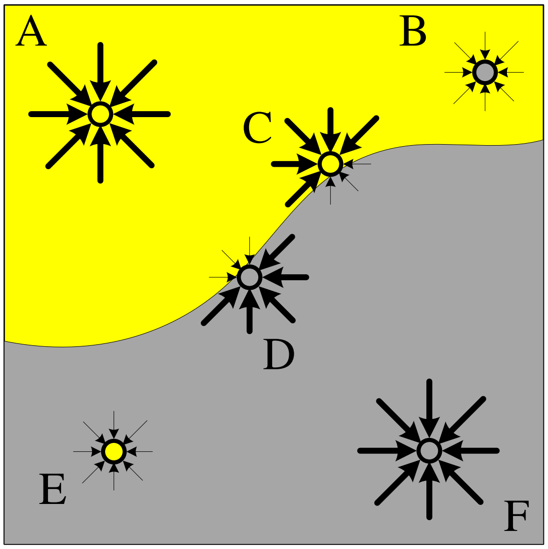


Fig. 4: Anisotropic contributions of the neighbor pixels to a target pixel at intra-region (A, B, E or F), and inter-region (C or D) areas.

C. Median Filtering

Beside anisotropic diffusion, median filtering is also widely used in image noise reduction due to its edge preserving property. To adopt median filtering into the framework, we can define a median filter as

$$\mathcal{F}_M(v_i, v_j) = \begin{cases} MI(v_i, v_j) & \text{if } MI(v_i, v_j) \text{ is the median,} \\ 0 & \text{otherwise,} \end{cases} \tag{8}$$

where M stands for the median filtering and the "median" means the median value among all contextual relatedness of v_i to its neighbours (i.e., $MI(v_i, v_j)$s). The filter is logically simple. However, it is computationally complex, because an additional sorting process has to be conducted for each candidate word to find the median.

V. Experiments

A. Dataset Construction and Evaluation Metric

To evaluate the performance, we use the dataset in [1], which includes one million (i.e., 1,106,280) merchandize images from eBay.com and Taobao.com. The dataset is open to public and is considered the one with the largest number of near duplicate images (12,901 images) so far. Due to its nature of user-generated-content, a wide range of modifications are included among the near-duplicates of the dataset, which contains the changes of illuminations, capturing devices, and scales, together with rotation, cropping, mirroring and text overlay. The dataset consists of three subsets:

[4] There are indeed two conductance functions that have proposed in [17]. The one presented in this paper has been randomly selected, because our empirical study shows there is no significant difference between the two functions when applied to NDR.

- **Training Set** which includes 1% (i.e., 10,000 images) of the entire dataset and will be used for learning the codebooks and inter-word contextual relationship;
- **Basic Set** which includes another 10% (i.e., 112,092 images) of the dataset and will be used as a basis for testing the performance of algorithms;
- **Distracter Set** which includes the rest part of the dataset (i.e., 984,188 images) and samples randomly selected from this set will be added to the **Basic Set** during the testing to generate larger datasets and investigate the robustness of the algorithms.

There are no exact duplicates in the dataset and no overlaps among the three subsets. The ground-truth of the dataset 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. In our experiments, all images are normalized to fit an 800 × 800 pixel box while preserving their aspect ratios.

There are five parameters in the proposed methods, namely the radius that defines the neighborhood d, the number of candidate visual words for each LIP ($\#can$), the maximum number of iterations for Algorithm 1 (m), the standard deviation for Gaussian filter (σ) and the fusion rate for Anisotropic filter (ξ). We set $d = 30$ which is an optimal setting found in [1]. We also set $\#can = 10$, because according to our empirical study, the proposed methods become less sensitive to this parameter when $\#can > 5$. Here we enlarge the number to increase the chance of a correct word is included in the candidate list. We set $\sigma = \frac{d}{3} = 10$ and $\xi = 0.2$ which are the popular setting in image domain. Unless mentioned otherwise, we set $m = 5$ in this paper. It has been confirmed as an optimal settings from the experiments in Section V-B2.

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), in which 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. Study of the Essential Characteristics

In this section, we study the essential characteristics of the proposed methods by utilizing them for contextual noise reduction on the most widely adopted Quantization scheme that selects for each LIP the visual word with the largest similarity. By using different methods for LIP extraction, we build two baselines as follows:

- $SIFT$ [9] based method, which is considered one of the most popular methods, where we employ Difference of Gaussians (DoG) for LIP detection and SIFT feature for description. SIFT is used as a representative of sparsely LIP sampling;
- Color-Moment (CM) based method [1], 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.

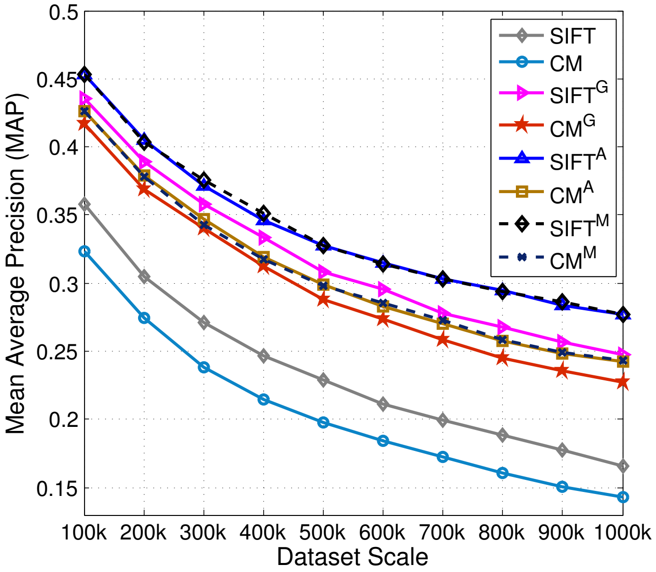


Fig. 5: Performance over different dataset scales. $SIFT^M$ (CM^M) is seriously overlapped with $SIFT^A$ (CM^A).

Gaussian filter $\mathcal{F}_G$, anisotropic filter $\mathcal{F}_A$ and median filter $\mathcal{F}_M$ are applied on the baselines respectively for contextual noise reduction, resulting in 6 new methods $SIFT^G$, CM^G, $SIFT^A$, CM^A, $SIFT^M$, and CM^M, in which superscript is used to indicate which filter has been used. We use the **Training Set** to learn the codebooks and the contextual relatedness among word pairs. All codebook are of size $2k$ (words) at this moment to speed up the computationally expensive processes of parameter tuning and feature comparison. The experiments using large codebooks will be reported and discussed in Section V-C.

As an overview for the performances of the 8 methods under investigation, the results with MAP at the top-100 retrieved images are shown in Figure 5, in which we vary the dataset scale to observe the scalability of different methods. The dataset scale is changed by first using the fully annotated **Basic Set** as a basis and randomly adding samples from the **Distracter Set** to generate larger sets. The results show that, overall, with contextual noise reduction, $SIFT^G$ and $SIFT^A$ ($SIFT^M$) outperform the baseline $SIFT$ by $36.57\% \pm 8.17$, $46.31\% \pm 12.69$ ($46.69\% \pm 12.68$), and CM^G and CM^A (CM^M) outperform CM by $46.28\% \pm 9.24$ and $52.03\% \pm 11.58$ ($52.02\% \pm 12.03$). In the following sections, we conduct more experiments to grasp an in-depth understanding of the proposed methods.

1) Does Contextual Noise Reduction Work?: In Figure 5, the advantage of using contextual noise reduction can be

TABLE I: Performance gains over the change of dataset scale.

Dataset Sacle	100k	200k	300k	400k	500k	600k	700k	800k	900k	1000k
$SIFT^G$ to $SIFT$	21.54%	27.84%	32.12%	35.16%	34.44%	39.49%	39.38%	41.64%	44.73%	49.34%
$SIFT^A$ to $SIFT$	26.45%	33.03%	36.98%	40.39%	42.81%	48.65%	51.70%	56.03%	59.89%	67.05%
$SIFT^M$ to $SIFT$	26.45%	32.40%	38.57%	42.42%	42.89%	48.42%	51.95%	55.40%	61.35%	67.05%
CM^G to CM	29.04%	34.51%	42.66%	45.30%	45.53%	48.45%	49.28%	52.49%	56.63%	58.87%
CM^A to CM	31.82%	38.30%	45.68%	48.56%	51.14%	53.77%	56.62%	60.20%	64.79%	69.41%
CM^M to CM	31.84%	37.94%	43.67%	47.77%	50.78%	54.80%	57.72%	60.63%	65.05%	69.97%

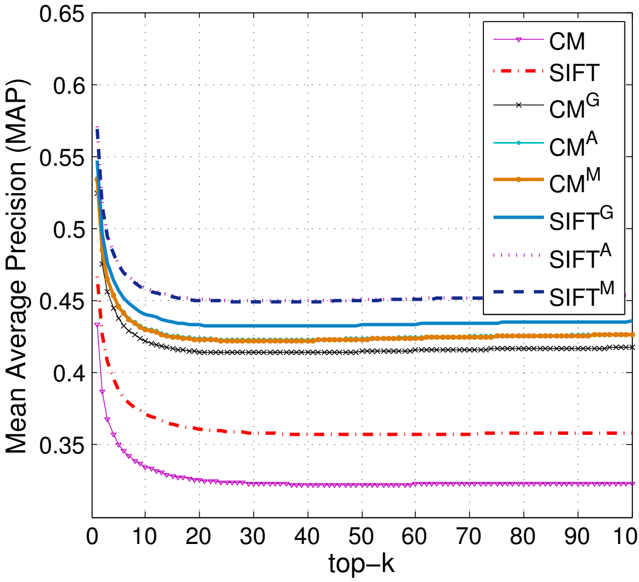


Fig. 6: Evaluation of MAP at the top-k of the ranked lists. $SIFT^M$ (CM^M) is seriously overlapped with $SIFT^A$ (CM^A).

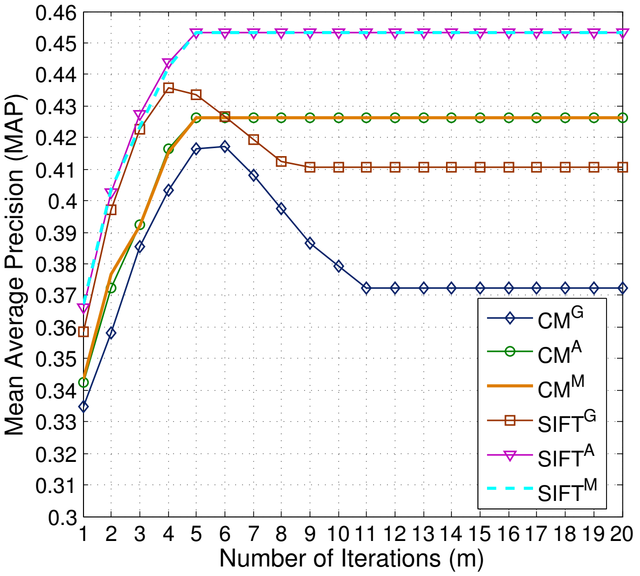


Fig. 7: Performance over the number of iterations. $SIFT^M$ (CM^M) is seriously overlapped with $SIFT^A$ (CM^A).

observed while $SIFT^G$, $SIFT^A$, $SIFT^M$, CM^G, CM^A and CM^M outperform the baselines. The rationale behind is that by further integrating the contextual relationship into the Quantization, visual words can work *collaborative* to represent the image features, which allows more patterns and more details to be contained in the resulting BoW feature vectors, and finally leads to higher accuracy in matching. This effect becomes more obvious when one compares CM^G and CM^A (CM^M) to CM, where the performance gain are $46.28\% \pm 9.24$ and $52.03\% \pm 11.58$ ($52.02\% \pm 12.03$) respectively, which are much more significant than those of $SIFT^G$ and $SIFT^A$ ($SIFT^M$) to $SIFT$. More importantly, we can also find in Figure 5 that when individual patches start to collaborate in CM^G and CM^A (CM^M), they can even outperform $SIFT$. This is an solid indication of the advantage of using the contextual noise reduction, because conventionally CM has been considered as a much weaker feature than $SIFT$.

Let us take a look at how this affects the ranking of NDR. In Figure 6, we evaluate the ranked lists of the 8 methods (on the **Basic Set**) by investigating how the AP changes from the top to the bottom of the lists. It can be seen that the majority of the improvement by utilizing contextual noise reduction is indeed coming from the section of the top-30 items in the lists. This is an indication that the near-duplicates at the bottom of the lists have been successfully boosted to the top position. In NDR, this is highly desired, because most users are only interested in the items ranked at the top positions, and in the applications built on smart phones or tablets, the screens are usually with too limited sizes to display too many items.

2) *Domain Adaptivity of Filters:* Comparing three filters (i.e., Gaussian, Anisotropic, and Median), Table I shows the results of performance gains of the methods over dataset scale changes. We can see that the Anisotropic and Median filters appear more reliable than Gaussian, in the senses that 1) in terms of performance gains, $SIFT^A$ (CM^A) is superior to $SIFT^G$ (CM^G) by $25.47\% \pm 8.00\%$ ($12.00\% \pm 3.53\%$) while $SIFT^M$ (CM^M) is superior to $SIFT^G$ (CM^G) by $25.51\% \pm 7.34\%$ ($20.92\% \pm 7.49\%$); and 2) divided by the scale point $500k$, the performance gain of $SIFT^A$ (CM^A) increases from $35.93\% \pm 6.44\%$ ($43.10\% \pm 7.93\%$) to $56.68\% \pm 7.21\%$ ($60.96\% \pm 6.30\%$) while that of $SIFT^M$ (CM^M) increases from $36.55\% \pm 7.04\%$ ($42.40\% \pm 7.62\%$) to $56.83\% \pm 7.44\%$ ($61.64\% \pm 6.00\%$). It is more significant than that of $SIFT^G$ (CM^G) which increases from $30.22\% \pm 5.63\%$ ($39.41\% \pm 7.32\%$) to $42.92\% \pm 4.20\%$ ($53.15\% \pm 4.55\%$). This is an indication for the advantage of using domain adaptive filters, because when more examples from the **Distracter Set** are added into the experiments, the domain-shift will be enlarged, and the chance of having words from "unseen" domains increases.

The comparison between the three filters can also been seen

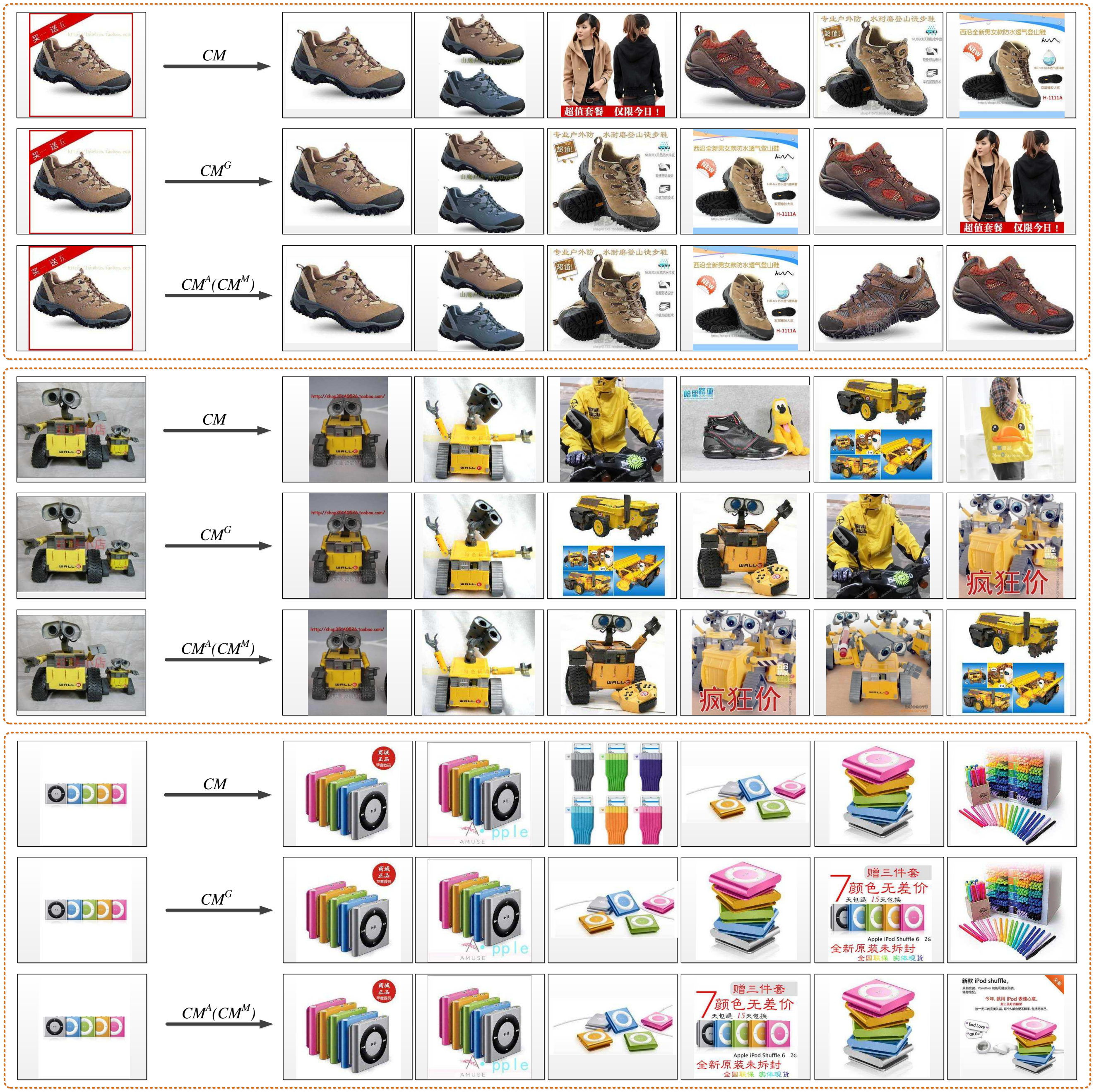


Fig. 8: Query images and corresponding ranked lists. The methods with domain adaptive filters (e.g., CM^A (CM^M)) are able to exclude items from irrelevant domains from the list.

in Figure 7, which has been obtained on the **Basic Set** during the parameter tuning for m the number of iteration. At the early stage of the iterations, both Gaussian and Anisotropic (Median) filters are able to boost the performance by using contextual information for noise reduction. However, the methods with Gaussian filter drops gradually after the 6-th and the 3-rd iteration in CM^G and $SIFT^G$ respectively. This is due to the isotropic nature of Gaussian filter, which blurs the contextual edges and causes unreasonable selection of visual words that are contextually related to the words at both sides of the edges but semantically unwarranted (cf. Section IV). Fortunately, the drop stops soon, since we have limited the number of candidates for each LIP, which prevents the Quantization from selecting more unreasonable words. For methods with Anisotropic (Median) filters, at the early stage of iterations, their performances are nearly the same as those of Gaussian filter, but continue to climb when Gaussian methods start to drop, and reaches a plateau afterwards. This further confirms the advantage of using domain adaptive filters. Nevertheless, we will see later in Sections V and VI, the difference between the two types of filters is in fact not apparent enough on the **Basic Set** where the domain shift from the **Training Set** is considered minor when compared to that of **Kentucky** (offline product images) or **Oxford** (architecture images).

To have an intuitive understanding, Figure 8 shows the query images and corresponding ranked lists returned by CM, CM^G and CM^A (CM^M, the results of CM^M for these queries are just the same as CM^A). We can see from the query image of "leather shoe" that, CM, which only considers color moments of the object, unreasonably ranks the clothes at the 3rd position, because the clothes are matched to the leather surface of the shoe and the red&white banner is matched to the banner and the red frame of the query image. The result is improved when contextual noise reduction has been considered in CM^G. However, the cloth image is still among top-6. Finally, when domain constraints have been considered in CM^A (CM^M), the cloth image has been successfully excluded. Similar observations can be found for the queries of "Wall-E" and the "iPod shuffle". Furthermore, for all the three queries, those non-ND images among the top-6 items retrieved by CM^A (CM^M) are all the products with similar colors and materials as the query images. This is a highly demand feature for product image retrieval, because, in case the ND images of the query cannot be found, users would be interested in the products with similar styles.

3) Global vs. Local Optimization: It is interesting to compare the methods proposed in this paper to the ones in [1] which are also utilizing the contextual relationship for optimizing the quantization (denoted as $SIFT^{CC}$ and CM^{CC} hereafter to be consistent with the ones used in the paper). In Table II, we summarize each method with its performance gains upon the $SIFT$ and CM, and corresponding time costs. The time cost is defined as the average time used for answering a query on the **Basic Set**. The two groups of methods obtain very similar performance in the comparison, in the sense that the average performance gain of the methods using Anisotropic or Median filter over the method in [1] is 4.52%, while the methods using Gaussian filter demonstrate comparable performance to the method in [1]. However, the methods proposed in this paper are with approaching linear complexity and are obviously more efficient than $SIFT^{CC}$ and CM^{CC} that are trying to solve a NP-hard problem (i.e., the binary quadratic programming), in the way that CM^G (CM^A) is faster than CM^{CC} by 36.90% (32.18%) and $SIFT^G$ ($SIFT^A$) is faster than $SIFT^{CC}$ by 13.02% (13.61%). Nonetheless, as the methods with Median filters, CM^M and $SIFT^M$ are exceptions of this observation, because the additional sorting process for searching medians is time consuming.

The essential difference of the two groups of methods lies in the scale of optimization, in the way that $SIFT^{CC}$ and CM^{CC} are global optimization which concerns more about the overall contextual consistency of the image, while proposed methods in this paper are all local-based optimization which focus on the local contextual consistency of the neighborhood and thus has the advantage of preserving the details of the image. However, through the iterative process, the optimization is able to extend the local consistency globally and gradually to the other parts of the image. By contrast, $SIFT^{CC}$ and CM^{CC} never has chance to extend its impacts to the details once the optimization is done. This is indicated by the results of significant tests, in which the superiority of the local-based optimization to the global optimization is incremental, but significantly stable. To further demonstrate this advantage, in Figure 9, we reconstruct the query image by attaching the mapped visual words to their original positions of the image. With global optimization, the signature on the baseball has been "merged" into the background, because the red threads around the ball are dominating patterns in the image, which leads the overall selection of visual words to cloth liked materials and the details of the signature have thus been sacrificed, resulting in the "blur" of the contextual boundary of the signature to the background. By contrast, with local-based optimization, the signature has been better represented by the words of strokes on the white background and the contextual details haven preserved successfully.

TABLE II: Comparison of mean average precision (MAP), time cost per query (milliseconds) between pre-Quantization and post-Quantization methods, and the results of significance test (at level 0.05, and X $\gg$ Y indicates that X is significantly better than Y).

Baseline	**Method**	**Group**	**MAP**	**Time Cost**
CM	CM^{CC}	pre-Quantization	0.4081	115
CM	CM^G	post-Quantization	0.4173	84
CM	CM^A	post-Quantization	0.4263	87
CM	CM^M	post-Quantization	0.4264	246
$SIFT$	$SIFT^{CC}$	pre-Quantization	0.4332	217
$SIFT$	$SIFT^G$	post-Quantization	0.4356	191
$SIFT$	$SIFT^A$	post-Quantization	0.4531	192
$SIFT$	$SIFT^M$	post-Quantization	0.4531	408
Results of Significance Tests	$SIFT^A, SIFT^M \gg SIFT^G, SIFT^{CC}$, $CM^A, CM^M \gg CM^G, CM^{CC}$			

TABLE III: Configurations of methods under investigation.

Method	Configuration	Feature
$SIFT^A_{20k}$	BoW+Anisotropic Filter	SIFT
$SIFT^G_{20k}$	BoW+Gaussian Filter	SIFT
$SIFT_{20k}$	BoW	SIFT
CM^A_{20k}	BoW+Anisotropic Filter	Color-Moment
CM^G_{20k}	BoW+Gaussian Filter	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

C. Comparison with the State-of-the-art

In this section, we conduct more comprehensive comparisons to investigate the accuracy and efficiency of the proposed methods. Five state-of-the-art NDR methods in the literature have been added in the experiments including

- Hamming Embedding (HE) [27], the method employs hamming codes to assign binary signatures to LIPs so as to decrease the visual ambiguity introduced by soft-weighting (SOFT) [15];
- Weak Geometric Consistency (WGC) [11], the method improves the accuracy of macthing by adding weak or partial geometric constraints;
- The enhanced WGC (EWGC) [12], the method improves WGC by further including translation information;

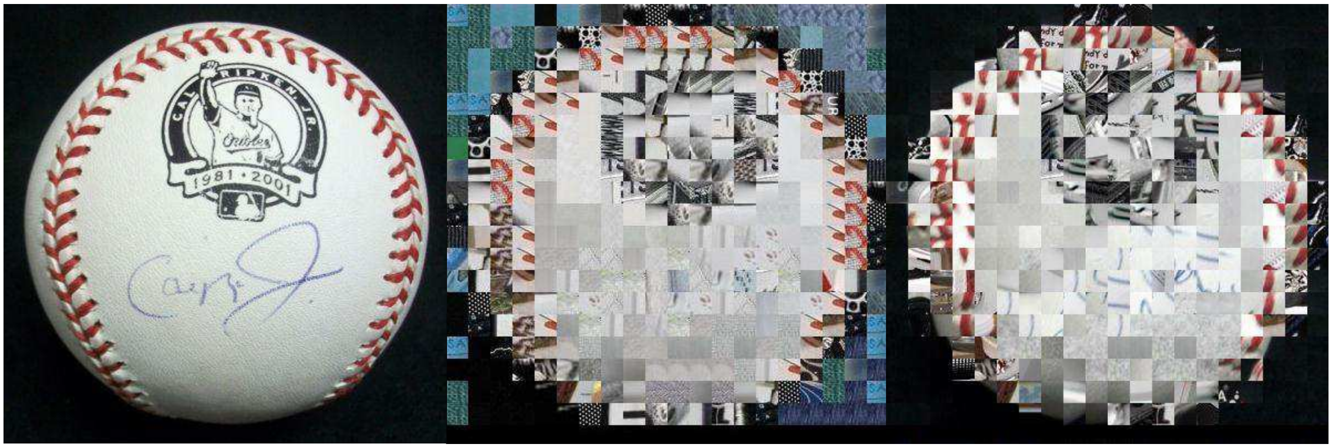


Fig. 9: Query image (left) and the results of reconstructing the query with the visual words selected by global optimization of CM^{CC} and local-based optimization of CM^{A} (shown on the middle and right respectively).

TABLE IV: Performance comparison with the state-of-the-art on **Basic Set** and the results of significance test (at level 0.05, and X $\gg$ Y indicates that X is significantly better than Y). The best results are bold.

	SIFT$^{A}_{20k}$	**SIFT**$^{G}_{20k}$	**SIFT**$_{20k}$	**CM**$^{A}_{20k}$	**CM**$^{G}_{20k}$	**CM**$_{20k}$	**WGC**	**EWGC**	**HE**	**Syn-Expan**	**DT**
Mean Average Precision (MAP)	**0.5617**	0.5438	0.4080	0.5263	0.5105	0.3894	0.5404	0.5144	0.5352	0.4834	0.5406
Time Cost (milliscond/query)	205	202	246	98	**96**	134	1895	1604	705	367	3411
With Anisotropic Filter							**WGC**^{A}	**EWGC**^{A}	**HE**^{A}	**Syn-Expan**^{A}	**DT**^{A}
Mean Average Precision (MAP)							0.6552	0.6193	0.6487	0.5929	**0.6567**
Time Cost (milliscond/query)							2045	1731	749	**518**	3582
Results of Significance Tests	$DT^{A}, WGC^{A} \gg HE^{A} \gg EWGC^{A} \gg$ Syn-Expan$^{A} \gg SIFT^{A}_{20k} \gg SIFT^{G}_{20k}, DT, WGC \gg HE,$ $CM^{A}_{20k} \gg EWGC, CM^{G}_{20k} \gg$ Syn-Expan $\gg SIFT_{20k} \gg CM_{20k}$										

- Synonym-based Query Expansion (Syn-Expan) [32][5], the method defines visual words that highly co-occur to each other as visual synonyms and uses them to conduct query expansion for performance improvement;
- Delaunay Triangulation (DT) [13], the method 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) [30] is also considered in this method. It has been reported with the highest performance in literature.

According to the study [1], the performances of most methods are approaching optimal on **Basic Set** when the vocabulary is set to the range of $10k$ to $40k$. Therefore, to be fair and consistent, for the experiments in this section, we unify the vocabulary size to $20k$ which is 10 times larger than the size we used in Section V-B (i.e., $2k$)[6]. To ease the discussion, we add a subscript for each method to indicate the vocabulary size (e.g., $SIFT^{A}_{20k}$). The performances of all methods are optimized (with the condition that the vocabulary size equals to $20k$) with the configurations summarized in Table III. Note that to compare 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 [32]. In addition, we exclude Median filter from this experiment, because its time complexity increase significant with the larger codebook size and its performance always appears similar to that of Anisotropic filter. 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.

To fully explore 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** [3], 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** [2], which is another standard benchmark which is composed of building images. It is employed for conducting the cross-domain experiments.

*1) General Comparison on **Basic Set**:* The comparison of the 11 methods is shown in Table IV. In terms of accuracy, for the methods using $SIFT$ feature, $SIFT^{A}_{20k}$ and $SIFT^{G}_{20k}$ outperform all conventional methods by $11.69\% \pm 12.18$, which has further confirmed the advantage of using contextual noise reduction. For CM^{A}_{20k} and CM^{G}_{20k}, even with comparably much weaker feature than SIFT, they can outperform $SIFT_{20k}$, EWGC and Syn-Expan by $21.21\% \pm 10.56$, and has demonstrated comparable performances to HE. Beside benefiting from the contextual noise reduction, CM-based methods, which are built based on the color information of the images, have taken the advantage of working on the merchandize imagery domain, because colors are of great importance to identify a product. However, popular LIP features like SIFT

[5] We thank Drs. Zhang Lei and Cai Rui, the authors of [32], for sharing the binary code and helping with the experiments.

[6] In fact, the size $20k$ is also a popular setting in literature (e.g., [11][13][30]).

TABLE V: Performance comparison with the state-of-the-art on **Kentucky** and the results of significance test (at level 0.05, and X $\gg$ Y indicates that X is significantly better than Y). The best results are bold.

	SIFT$_{20k}^{A}$	**SIFT**$_{20k}^{G}$	**SIFT**$_{20k}$	**CM**$_{20k}^{A}$	**CM**$_{20k}^{G}$	**CM**$_{20k}$	**WGC**	**EWGC**	**HE**	**Syn-Expan**	**DT**
Mean Average Precision (MAP)	0.8683	0.8213	0.6238	**0.8913**	0.8101	0.6039	0.8645	0.8430	0.8608	0.6944	0.8586
Time Cost (milliscond/query)	31	31	38	29	**28**	34	150	163	87	42	264
With Anisotropic Filter							**WGC**^{A}	**EWGC**^{A}	**HE**^{A}	**Syn-Expan**^{A}	**DT**^{A}
Mean Average Precision (MAP)							0.9263	0.9024	0.9187	0.8725	0.9125
Time Cost (milliscond/query)							174	183	105	62	279
Results of Significance Tests	$WGC^A \gg HE^A, DT^A \gg EWGC^A \gg CM_{20k}^A \gg$ Syn-Expan$^A, SIFT_{20k}^A, WGC, HE \gg DT \gg$ $\gg EWGC \gg SIFT_{20k}^G \gg CM_{20k}^G \gg$ Syn-Expan $\gg SIFT_{20k} \gg CM_{20k}$										

TABLE VI: Performance comparison with the state-of-the-art on **Oxford** and the results of significance test (at level 0.05, and X $\gg$ Y indicates that X is significantly better than Y). The best results are bold.

	SIFT$_{20k}^{A}$	**SIFT**$_{20k}^{G}$	**SIFT**$_{20k}$	**CM**$_{20k}^{A}$	**CM**$_{20k}^{G}$	**CM**$_{20k}$	**WGC**	**EWGC**	**HE**	**Syn-Expan**	**DT**
Mean Average Precision (MAP)	0.5417	0.4921	0.3861	0.3041	0.2713	0.2642	**0.5438**	0.5087	0.4975	0.4267	0.5349
Time Cost (milliscond/query)	42	41	54	**33**	33	57	330	341	186	82	420
With Anisotropic Filter							**WGC**^{A}	**EWGC**^{A}	**HE**^{A}	**Syn-Expan**^{A}	**DT**^{A}
Mean Average Precision (MAP)							0.6847	0.6223	0.5982	0.5424	0.6724
Time Cost (milliscond/query)							352	368	214	97	442
Results of Significance Tests	$WGC^A \gg DT^A \gg EWGC^A \gg HE^A \gg WGC,$ Syn-Expan$^A, SIFT_{20k}^A \gg DT \gg EWGC \gg HE,$ $SIFT_{20k}^G \gg$ Syn-Expan $\gg SIFT_{20k} \gg CM_{20k}^A \gg CM_{20k}^G \gg CM_{20k}$										

usually ignore this information, and their performances are thus less promised in this specific domain.

In terms of efficiency, $SIFT_{20k}^A$, $SIFT_{20k}^G$, CM_{20k}^A and CM_{20k}^G are apparently much faster than conventional methods, in the sense that their time costs are (on average) 1662% times less than those of WGC, EWGC, and DT. It is worth mentioning that with an additional contextual noise reduction process employed in the Quantization, the total time costs of $SIFT_{20k}^A$ and CM_{20k}^A are even reduced when compared to $SIFT_{20k}$ and CM_{20k}. The reason is that with when the contextual noises being reduced from the Quantization, the purity and sparsity of the BoW feature vectors have been improved, which has significantly saved the time for matching (the most expensive process in LIPs+BoW framework).

As discussed in Section II, the method proposed in this paper work in Quantization stage of the LIPs+BoW framework and thus are compatible to most of the conventional methods that are conducted in the post-Indexing stage. To investigate if this is true, we further combine the five state-of-the-art methods with Anisotropic filters, resulting in WGC^A, EWGC^A, HE^A, Syn-Expan^A and DT^A. Their results are shown in the second section of Table IV, in which the performances of five methods have been improved by 21.39% on average with slight time cost increased (less than 8% for WGC, EWGC, HE, and DT). This confirms our claim in Section II that, by using Anisotropic filter to improve the clarity of the Quantization directly, the quality of the proceeding processes will also be improved.

2) Comparison of Generalizability on ***Kentucky****:* To test the generalizability of each method, we extend the performance comparison to **Kentucky** benchmark [3], which consists of 10,200 product images forming 2,550 near-duplicate 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 $SIFT_{20k}^G$, $SIFT_{20k}^A$, CM_{20k}^G, CM_{20k}^A and Syn-Expan. The results are shown in Table V, where the methods proposed in this paper continue to show superiority to conventional methods. This is consistent with the results in the previous section, thanks to the similarity between **Kentucky** and **Basic Set**. Moveover, the CM-based method CM_{20k}^A surprisingly outperform all its competitors on **Kentucky**. This is due to the fact that scale change has not been considered in **Kentucky** when the images were captured, which has relaxed the requirement for scale invariant of a local feature, and thus reduced the biggest advantage of CM patches when compared to SIFT LIPs. The same observation has also been found in [1, 11]. Furthermore, comparing between the methods using different contextual noise filters, the advantage of the domain adaptive filter becomes more obvious than the experimental results in Section V-B, in the way that CM_{20k}^A ($SIFT_{20k}^A$) outperform CMG_{20k} ($SIFT_{20k}^G$) by 10.02% (5.71%).

In the second section of Table V, the results again confirm the advantage of integrating Anisotropic filter, which brings additional 10.57% performance gain (on average) for the five state-of-the-art methods. The improvement is significantly stable.

3) Cross-Domain Experiments on Oxford: In this section, we further enlarge the domain shift by moving from merchandize imagery dataset to architecture image dataset. **Oxford** [2] is employed for this purpose, 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** [2] is used for training in $SIFT_{20k}^G$, $SIFT_{20k}^A$, CM_{20k}^G, CM_{20k}^A and Syn-Expan. The results in Table VI show that $SIFT_{20k}^A$ can still outperform all the conventional methods except WGC while demonstrating comparable performances to WGC. The CM-based methods lose their advantage in architecture image domain when the scale invariance increases significantly. However, the advantage of using the domain adaptive filter is still apparent, indicated by the fact that CM_{20k}^A ($SIFT_{20k}^A$) outperform CM_{20k}^G ($SIFT_{20k}^G$) by 12.09% (10.08%).

In the second section of Table VI, the results confirm the advantage of integrating Anisotropic filter, which brings additional 24.26% performance gain for the five state-of-the-art methods. The improvement is significantly stable.

D. Issues and Possible Solutions

While the encouraging results of the proposed methods have been observed in the experiments, we have also found three issues which are worth further investigations. First, the Anisotropic filter may fail when a certain number of mis-mapped words is distributing closely, because in this case, these mis-mapped works will form a "region", to which the filter cannot distinguish from a normal region and thus will choose to "respect" it. This phenomenon is usually found on images including reflections of lights which make the regions around the reflections overexposed and further cause densely distributed mis-mapped words. Reflection removal algorithms developed in image processing can be used as a pre-processing to address this problem. Second, the optimal number of iterations is indeed image dependent, which is within a range of 2 to 17 according to our investigation on a randomly selected subset consisting of 30 images. An adaptive algorithm that determines the number of iterations on-the-fly for each image can be possibly developed based on the number of words, and the overall strength of the intra-image contextual relatedness. Third, for objects with "smooth" textures, the features on the boundaries are of great importance. However, the current framework of the proposed method will make the boundaries being "absorbed" by the inner regions of the objects or backgrounds during the quantization. This can be solved by specifically maintain a certain portion of boundary words in the codebook and give them priority to be assigned along boundaries. Alternatively, boundary descriptors, which have been developed in computer vision, can be employed to compensate for the shortages of the descriptors used in this paper.

VI. Conclusions

In this paper, we study the use of contextual relationship among visual words to reduce the mis-mapped words in Quantization. A context graph is constructed to encapsulate the contextual relationship within an image, and treated as a pseudo-image, on which we can apply the sophisticated image filters to reduce the contextual noises and thus improve the quality of the Quantization. The cross-domain issue has also been addressed in the method by integrating the classic Anisotropic diffusion filter and Median filter into the scheme. The experimental results have validated the effectiveness and efficiency of the proposed method.

Beside being a practical approach for near-duplicate retrieval, the method proposed in this paper has demonstrated that, by treating the context graph as a pseudo-image, it greatly facilities the solution search and problem identification, which leaves rooms for future explorations of adopting classical image noise reduction techniques to Quantization enhancement. We will conduct more studies following this direction in our future work. Moreover, it is interesting to investigate the performances of different filters on different near-duplicate modifications, in the sense that they might cause different types of contextual noises.

References

[1] X.-Y. Wei, Z.-Q. Yang, C.-W. Ngo, and W. Zhang, "Visual typo correction by collocative optimization - a case study on merchandize images," *IEEE Transactions on Image Processing*, vol. 23, no. 2, pp. 527–540, 2014.

[2] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[3] D. Nistér and H. Stewénius, "Scalable recognition with a vocabulary tree," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2006.

[4] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *IEEE International Conference on Computer Vision*, 2003.

[5] 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 *IEEE International Conference on Computer Vision*, 2007, pp. 1–8.

[6] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2006, pp. 2161–2168.

[7] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[8] H. Jégou, H. Hedi, and C. Schmid, "A contextual dissimilarity measure for accurate and efficient image search," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[9] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

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

[11] H. Jégou, M. Douze, and C. Schmid, "Improving bag-of-features for large scale image search," *International Journal of Computer Vision*, vol. 87, no. 3, pp. 316–336, 2010.

[12] W.-L. Zhao, X. Wu, and C.-W. Ngo, "On the annotation of web videos by efficient near-duplicate search," *IEEE Transactions on Multimedia*, vol. 12, no. 5, 2010.

[13] W. Zhang, L. Pang, and C.-W. Ngo, "Snap-and-ask: answering multimodal question by naming visual instance," in *ACM International Conference on Multimedia*, 2012, pp. 609–618.

[14] D.-Q. Zhang and S.-F. Chang, "Detecting image near-duplicate by stochastic attributed relational graph matching with learning," in *ACM International Conference on Multimedia*, 2004, pp. 877–884.

[15] Y.-G. Jiang, C.-W. Ngo, and J. Yang, "Towards optimal bag-of-features for object categorization and semantic video retrieval," in *ACM International Conference on Image and Video Retrieval*, 2007, pp. 494–501.

[16] J. C. van Gemert, C. J. Veenman, A. W. M. Smeulders, and J.-M. Geusebroek, "Visual word ambiguity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 7, pp. 1271–1283, 2010.

[17] P. Perona and J. Malik, "Scale-space and edge detection using anisotropic diffusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 7, pp. 629–639, 1990.

[18] B. Widrow, J. Glover, and et. al, "Adaptive noise cancelling: Principles and applications," *Proceedings of the IEEE*, vol. 63, no. 12, pp. 1692–1716, 1975.

[19] H. Burger, C. Schuler, and S. Harmeling, "Image denoising: Can plain neural networks compete with bm3d?" in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[20] J. Yuan, L.-Y. Duan, Q. Tian, and C. Xu, “Fast and robust short video clip search using an index structure,” in *ACM SIGMM International Workshop on Multimedia Information Retrieval*, 2004, pp. 61–68.

[21] E. Y. Chang, J. Z. Wang, C. Li, and G. Wiederhold, “Rime: A replicated image detector for the world-wide web,” in *SPIE Symposium of Voice, Video, and Data Communications*, 1998, pp. 58–67.

[22] A. H. A, K. ho Hyun B, and R. B. A, “Comparison of sequence matching techniques for video copy detection,” in *Conference on Storage and Retrieval for Media Databases*, 2002.

[23] R. Fergus, P. Perona, and A. Zisserman, “Object class recognition by unsupervised scale-invariant learning,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2003, pp. 264–271.

[24] F. Wang and C.-W. Ngo, “Rushes video summarization by object and event understanding,” in *TRECVID BBC Rushes Summarization Workshop at ACM Multimedia*, 2007, pp. 25–29.

[25] Y. Ke, R. Sukthankar, and L. Huston, “An efficient parts-based near-duplicate and sub-image retrieval system,” in *ACM International Conference on Multimedia*, 2004, pp. 869–876.

[26] P. Jain, B. Kulis, and K. Grauman, “Fast image search for learned metrics,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[27] H. Jegou, M. Douze, and C. Schmid, “Hamming embedding and weak geometric consistency for large scale image search,” in *European Conference on Computer Vision: Part I*, 2008, pp. 304–317.

[28] O. Chum, J. Philbin, M. Isard, and A. Zisserman, “Scalable near identical image and shot detection,” in *ACM International Conference on Image and Video Retrieval*, 2007.

[29] O. Chum, J. Philbin, and A. Zisserman, “Near duplicate image detection: min-hash and tf-idf weighting,” in *British Machine Vision Conference*, 2008.

[30] H. Jégou, M. Douze, and C. Schmid, “On the burstiness of visual elements,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1169–1176.

[31] D. Pearce, “A comparative evaluation of collocation extraction techniques,” in *International Conference on Language Resources and Evaluation*, 2002.

[32] W. Tang, R. Cai, Z. Li, and L. Zhang, “Contextual synonym dictionary for visual object retrieval,” in *ACM international conference on Multimedia*, 2011, pp. 503–512.

[33] Y. Jiang, “Randomized visual phrases for object search,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Washington, DC, USA, 2012, pp. 3100–3107.

[34] J. Weickert, *Anisotropic Diffusion In Image Processing*. Stuttgart, Germany: Teubner-Verlag, 1998.

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

Xiao-Yong Wei is a Professor with the College of Computer Science, Sichuan University, China, and a postdoctoral fellow with the International Computer Science Institute, University of California, Berkeley, USA. He received his Ph.D. degree in Computer Science from 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 in Dept. of Computer Science and Department of Chinese, Linguistics and Translation of City University of Hong Kong in 2009 and 2010, respectively. He had worked as a Manager of Software Department at Para Telecom Ltd., China, from 2000 to 2003. His research interests include multimedia retrieval, data mining, and machine learning.

Zhang Yi received the Ph.D. degree in mathematics from the Institute of Mathematics, The Chinese Academy of Science, Beijing, China, in 1994. Currently, he is a Professor at the Machine Intelligence Laboratory, College of Computer Science, Sichuan University, Chengdu, China. He is the co-author of the books: Convergence Analysis of Recurrent Neural Networks (Kluwer Academic Publishers, 2004), Neural Networks: Computational Models and Applications (Springer, 2007), and Subspace Learning of Neural Networks (CRC Press, 2010). He was an Associate Editor of IEEE Transactions on Neural Networks and Learning Systems (2009 2012). He is an Associate Editor of IEEE Transactions on Cybernetics (2014). He is a Fellow of IEEE. His current research interests include Neural Networks and Big Data.

Gerald Friedland is the director of the Audio and Multimedia group at the International Computer Science Institute (ICSI), a private, non-profit research lab affiliated with the University of California, Berkeley. He is also teaching in the EECS department as an adjunct professor. His research interests focus on multimedia content analysis, its applications, and social implications. Dr. Friedland has published more than 200 peer-reviewed articles in conferences, journals, and books. His co-authored text book on Multimedia Computing published in August 2014 with Cambridge University Press and his co-authored book on “Multimodal Location Estimation of Images and Videos” appeared with Springer in October 2014. He is the recipient of several research and industry recognitions, among them the European Academic Software Award and the Multimedia Entrepreneur Award by the German Federal Department of Economics. He also leads the team that won the ACM Multimedia Grand Challenge in 2009 and 2015, respectively. Dr. Friedland received his doctorate and master degree in computer science from Freie Universitaet Berlin in 2002 and 2006, respectively. He is a senior member of ACM and IEEE.