Bundling Features for Large Scale Partial-Duplicate Web Image Search

Zhong Wu*, Qifa Ke, Michael Isard, and Jian Sun

Microsoft Research

Abstract

In state-of-the-art image retrieval systems, an image is represented by a bag of visual words obtained by quantizing high-dimensional local image descriptors, and scalable schemes inspired by text retrieval are then applied for large scale image indexing and retrieval. Bag-of-words representations, however: 1) reduce the discriminative power of image features due to feature quantization; and 2) ignore geometric relationships among visual words. Exploiting such geometric constraints, by estimating a 2D affine transformation between a query image and each candidate image, has been shown to greatly improve retrieval precision but at high computational cost. In this paper we present a novel scheme where image features are bundled into local groups. Each group of bundled features becomes much more discriminative than a single feature, and within each group simple and robust geometric constraints can be efficiently enforced. Experiments in web image search, with a database of more than one million images, show that our scheme achieves a 49% improvement in average precision over the baseline bag-of-words approach. Retrieval performance is comparable to existing full geometric verification approaches while being much less computationally expensive. When combined with full geometric verification we achieve a 77% precision improvement over the baseline bag-of-words approach, and a 24% improvement over full geometric verification alone.

1. Introduction

Our goal, given a query image, is to locate its near- and partial-duplicate images in a large corpus of web images. There are many applications for such a system, for example detecting copyright violations or locating high-quality, or canonical, versions of a low-resolution or altered image.

Web image search differs from image-based object retrieval, where image variations can be due to 3D view-point change, lighting, object deformations, or even object-class variability. In our case, target images are obtained by editing the original 2D image through changes in scale, cropping, partial occlusions, etc. This is a less challenging task than full object retrieval, and so the bar is set higher for a system's performance, scalability and accuracy. Nevertheless, modifications to the original web images are often substantial and cannot be described by a single 2D transformation such as a homography. Figure 1 shows some examples of partial-duplicate web images. As the figure shows, users often take different portions from the original image and paste them to the target image with modifications, resulting in a partial-duplicate image that differs from the original not only in appearance, but also in 2D layout.

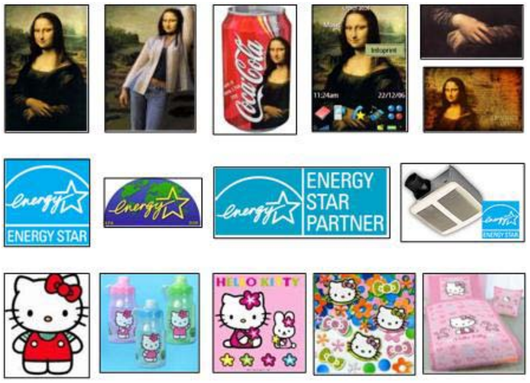

Figure 1. Examples of partial-duplicate web images.

State-of-the-art large scale image retrieval systems [1, 3, 11, 12] have relied on quantizing local SIFT descriptors [6] into visual words, and then applying scalable textual indexing and retrieval schemes [15]. The discriminative power of local descriptors, however, is limited due both to quantization and to the large number of images (e.g. greater than a million images). Geometric verification [3, 6, 12, 15] becomes an important post-processing step for getting a reasonable retrieval precision, especially for low-resolution images. But full geometric verification is computationally expensive. In practice therefore it is only applied to a subset of the top-ranked candidate images. For web image retrieval the number of near or partial duplicates could be large, and applying full geometric verification to only these top-ranked

*Zhong Wu is affiliated with Tsinghua University, Beijing, P.R.China, and this work was performed while Zhong Wu was a summer intern at Microsoft Research.

images may not be sufficient for good recall.

In this paper we propose a novel scheme to bundle SIFT features into local groups. These bundled features are repeatable and much more discriminative than an individual SIFT feature. Equally importantly, they provide a flexible representation that allows simple and robust geometric constraints to be efficiently enforced when querying the index. Experiments in web image search, with a database of more than a million web images, show that our simple scheme achieves a 49% improvement over the baseline bag-of-words approach. Retrieval performance is comparable to existing full geometric verification approaches while being much less computationally expensive. When combined with full geometric verification we achieve a 77% improvement in average precision over the baseline bag-of-words approach and a 24% improvement over full geometric verification alone.

1.1. Related work

State-of-the-art large scale image retrieval systems [1, 3, 11, 12] have been significantly advanced by two seminal works: 1) the introduction of local SIFT descriptors [6] for invariant image representation; and 2) the quantization of local descriptors into visual words for scalable image indexing and query [15]. By using an inverted-file index of visual words one not only avoids storing and comparing high-dimensional local descriptors, but also reduces the number of candidate images since only those images sharing common visual words with the query image need to be examined. While critical for scalability, quantization has two major issues. First, modifications to an image patch can lead to its corresponding descriptor being quantized into different visual words. Second, quantization reduces the discriminative power of local descriptors since different descriptors quantized to the same visual word are considered to match with each other. These two issues reduce the precision and recall in image retrieval, especially for low resolution images.

Soft-quantization [4, 13] has been proposed to improve recall by quantizing a descriptor to more than one visual word. Query expansion [2] is another successful technique to boost recall but it can fail on queries with poor initial recall. To improve precision, a visual word may be augmented with compact information from its original local descriptor, including a Hamming code [3], descriptor scale and angle [3], and the distance (in descriptor space) to its neighboring visual words [13]. While these methods are effective at improving the discriminative power at a reasonable cost in index file size, they still ignore the geometric relationship between feature points—an important characteristic for images that has no direct analog in traditional text retrieval. Exploiting such geometric relationships with full geometric verification (e.g. by estimating an affine transformation between the query image and a candidate image) has been shown to significantly improve the retrieval precision [3, 6, 12]. But full geometric verification is computationally expensive. Local spatial consistency from k (=15) spatial nearest neighbors, a weaker but computationally more feasible geometric constraint, is used in [15] to filter false visual-word matches. However, we have found spatial nearest neighbors to be sensitive to the image noise and resolution changes that are typical in web images. Other related works on higher-order features(c.f. [5, 14, 16, 17]) require data-dependent training to select co-occurrence features, or combinatorial feature pairing/grouping. Such training schemes do not scale for web images containing all kinds of object categories.

2. Bundled Features

In this section we first compare the discriminative power of two popular local image features: SIFT and MSER. We then motivate the use of grouped features to improve discrimination, and introduce the notion of a bundled feature which is a flexible representation that facilitates partial matches between two grouped features.

Notation. In the rest of this paper we adopt the following terminology:

- *SIFT feature*: a keypoint and the SIFT descriptor computed from the scale-invariant region centered at the keypoint [6];
- *MSER detection*: a maximally stable region [8];
- *MSER feature*: an MSER detection and the SIFT descriptor computed from the detected region.

2.1. Point feature: SIFT

The SIFT feature (keypoint and descriptor) is one of the most robust and distinctive point features [6]. The SIFT keypoint gains invariance to scale and rotation by exploiting scale-space extrema and the local dominant orientation. The SIFT descriptor assembles a 4x4 array of 8 gradient orientation histograms around the keypoint, making it robust to image variations induced by both photometric and geometric changes.

In large scale image search, however, we need to match a single SIFT feature to millions or billions of SIFT features computed from a corpus of web images. In this scenario the discriminative power of the quantized SIFT feature decreases rapidly, resulting in many false positive matches between individual features. Figure 2(a) shows an example. On the left column is a query SIFT feature, followed by its top five matches from a dataset with 100,000 images. The figure shows that the four best candidates are actually mismatches.

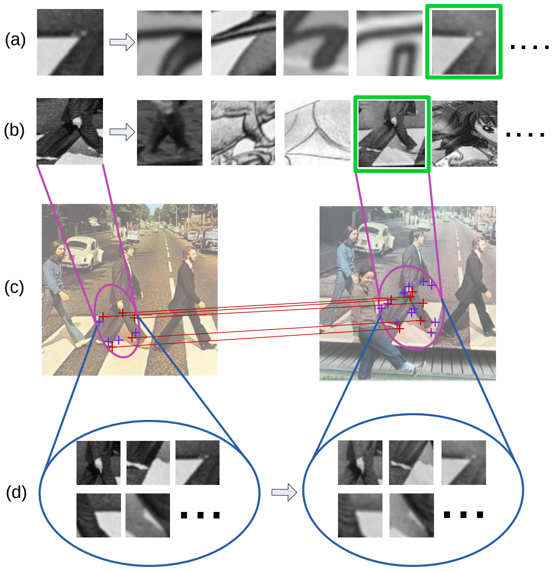

Figure 2. Matching bundled features: (a) a query SIFT feature and its top matches; (b) a query MSER feature and its top matches (green boxes in (a) and (b) indicate correct matches); (c) **partial** matching of two bundled features (some features are not matched); (d) matched SIFT features inside the MSER detection region.

2.2. Region feature: MSER

The Maximally Stable Extremal Region (MSER) [8] is another widely-used feature in large scale image-retrieval systems. Unlike the SIFT feature detector, MSER detects affine-covariant stable regions. Each detected elliptical region is normalized into a circular region from which a SIFT descriptor is computed.

Usually the MSER detector outputs a relatively small number of regions per image and their repeatability and distinctness are higher [9, 10] than that of the SIFT keypoint. However, false positive matches remain an issue for large image databases. Figure 2(b) shows a query MSER feature and several retrieved, mismatched features from the same database. The sources of false positives are twofold: 1) each MSER feature is still represented by a single SIFT descriptor no matter how large the region is; and 2) quantization further decreases the discriminative power of the feature.

2.3. Bundled features

A straightforward way to increase the discriminative power of a local feature is to increase its region size in the image (e.g. increasing its scale by a constant factor) and/or the dimensionality of the descriptor. A larger feature, however, is less repeatable and has a lower localization accuracy and it is more sensitive to occlusion and image variations caused by photometric and geometric changes. Figure 2(c) shows two MSER detections from two near-duplicate images that are of different sizes and corrupted by noise. Since the two regions are not well aligned, a larger descriptor with higher dimension will be more sensitive to overlap error [10]. However, if we closely inspect the two MSER detections in Figure 2(c) we can observe that some parts of these detected regions match very well. Figure 2(d) highlights the matched partial regions. If we can partially match two MSER regions and represent them with a more discriminative descriptor, we may improve both precision and recall. This motivates us to combine several point features within a region to form a *bundled feature*.

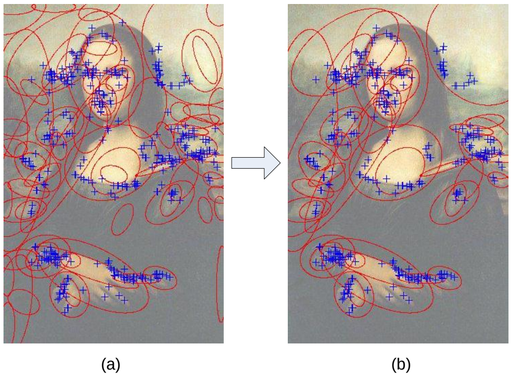

Figure 3. Examples of bundled features. (a) SIFT features and MSER detections; (b) bundled features from (a). The bundling MSER detections are indicated by ellipses, and the SIFT features in each bundled feature are shown in blue "+".

Denote $\mathbf{S} = \{s_j\}$ the SIFT features and $\mathbf{R} = \{r_i\}$ the MSER detections computed in an input image. We define the bundled feature $\mathbf{B} = \{\mathbf{b}_i\}$ to be:

$$\mathbf{b}_i = \{s_j | s_j \propto r_i, s_j \in S\}, \tag{1}$$

where $s_j \propto r_i$ means that the point feature s_j falls inside the region r_i. (In fact we enlarge the ellipse of the bundling MSER slightly when computing $s_j \propto r_i$. We considered enlargement factors between 1 and 2 and get the best performance using a factor of 1.5.) $\mathbf{b}_i$ is discarded if it is empty. A bundled feature is simply several SIFT features "bundled" by an MSER detection. Note that one SIFT feature may belong to multiple bundled features or may not belong to any bundled feature. Figure 3 shows several bundled features. We discard any MSER detection whose ellipse spans more than half the width or height of the image, since such large regions are generally not repeatable.

A bundled feature is more discriminative than a single SIFT feature as it consists of multiple SIFT features. Unlike a single large feature, a bundled feature provides a flexible representation that allows us to *partially* match two groups of SIFT features. Specifically, two matched bundled features are allowed to have large overlap error in their bundling MSER, and to have different number of SIFT features with only a subset of them matched. Thus the more discriminative bundled feature is also robust to occlusion and other image variations induced by photometric and geometric changes, making it possible to achieve both high precision and recall. The remaining challenge is how to efficiently perform partial matching of two bundled features in a large scale image search system.

3. Image Retrieval using Bundled Features

This section shows how to exploit two simple weak geometric constraints for efficient partial matching of bundled features in a large scale image-search system. First, SIFT features that are bundled in the query image should typically match with corresponding SIFT features in a target image that also reside in a common bundle. Second, the relative spatial configuration of the SIFT features within one bundled feature should remain approximately the same in query and target images.

3.1. Feature quantization

To build a large scale image indexing and retrieval system, we need to quantize local descriptors into visual words. For a given bundled feature, we quantize its constituent SIFT descriptors individually. We use hierarchical K-means [11] to obtain a vocabulary of one million visual words from a training set of 50 thousand images. Following [12] we use a k-d tree to organize these visual words for nearest neighbor search during quantization. To reduce quantization error, we use a soft quantization scheme [13], mapping a descriptor to its n-nearest visual words in the k-d tree.

3.2. Matching bundled features

Let $\mathbf{p} = \{p_i\}$ and $\mathbf{q} = \{q_j\}$ be two bundled features with quantized visual words $p_i, q_j \in W$, where W is our visual vocabulary. First, we sort $\{p_i\}$ and $\{q_j\}$ in a geometric order (as explained below). Next, we discard any $p_i \in \mathbf{p}$ that does not have a matching $q_i \in \mathbf{q}$. Then for each remaining visual word p_i in the bundled feature $\mathbf{p}$, we find its matched visual word $q^*(p_i)$ in the bundled feature $\mathbf{q}$ and denote the order of $q^*(p_i)$ in $\mathbf{q}$ by $O_q[p_i]$.

Now, we define a matching score $M(\mathbf{q}; \mathbf{p})$ between $\mathbf{p}$ and $\mathbf{q}$. The score $M(\mathbf{q}; \mathbf{p})$ consists of a membership term $M_m(\mathbf{q}; \mathbf{p})$ and a geometric term $M_g(\mathbf{q}; \mathbf{p})$:

$$M(\mathbf{q}; \mathbf{p}) = M_m(\mathbf{q}; \mathbf{p}) + \lambda M_g(\mathbf{q}; \mathbf{p}), \qquad (2)$$

where λ is a weighting parameter.

Membership term. We simply use the number of common visual words between two bundled features to define the membership term $M_m(\mathbf{q}; \mathbf{p})$:

$$M_m(\mathbf{q}; \mathbf{p}) = |\{p_i\}|. \qquad (3)$$

This term gives a higher score for matched bundles with more common visual words, enforcing a weak spatial consistency. This score is not normalized by the total number of matched and unmatched features in $\mathbf{p}$ and $\mathbf{q}$ so regions with many matching features score higher than regions with fewer matching features, even if the proportion of features that match is higher in the latter case.

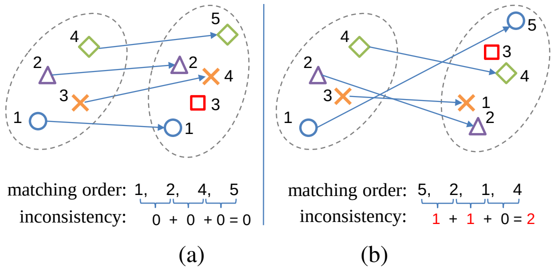
4
2
3
1
5
2
4
3
1
matching order: 1, 2, 4, 5
inconsistency: 0 + 0 + 0 = 0
4
2
3
1
5
3
4
1
2
matching order: 5, 2, 1, 4
inconsistency: 1 + 1 + 0 = 2
(a)
(b)

Figure 4. The ordering constraint for bundled features. (a) Correct matches preserve the relative ordering of the SIFT features inside the bundled feature; (b) Wrong matches result in inconsistent relative orders.

Geometric term. Our geometric term performs a weak geometric verification between two bundled features $\mathbf{p}$ and $\mathbf{q}$ using relative ordering:

$$M_g^D(\mathbf{q}; \mathbf{p}) = -\sum_i \delta(O_q[p_i] > O_q[p_{i+1}]), \qquad (4)$$

where D is some pre-defined geometric order, and $\delta(O_q[p_i] > O_q[p_{i+1}])$ is an indicator function that measures the consistency between the order $i < i+1$ (before matching) and the order $O_q[p_i] > O_q[p_{i+1}]$ (after matching). In other words, we penalize geometric inconsistency (as defined by our ordering) of the matching between two bundled features. We do not compare the absolute ordered rank of matched SIFT features. Instead, we only use the relative ordering relationship since it is more robust to inconsistencies resulting from partial matches of features between bundles.

So far we have not defined what the geometric order is, and in fact it may be application dependent. Since there is no significant rotation between duplicate images in our web image search scenario, we use the X- and Y-coordinates of $\{p_i\}$ and $\{q_j\}$ to define the geometric order:

$$M_g(\mathbf{q}; \mathbf{p}) = \min(M_g^X(\mathbf{q}; \mathbf{p}), M_g^Y(\mathbf{q}; \mathbf{p})), \qquad (5)$$

where $M_g^X(\mathbf{q}; \mathbf{p})$ is computed by sorting $\{p_i\}$ and $\{q_j\}$ according to their X-coordinates, and $M_g^Y(\mathbf{q}; \mathbf{p})$ by sorting on their Y-coordinates. The generalization to handle larger rotations is straightforward, e.g. by ordering features along the dominant orientation of the bundling MSER detection.

Figure 4 shows two matching pairs. In the correctly matching case (a), the geometric score is $M(\mathbf{q}; \mathbf{p}) = 4 - 0 = 4$. In the mismatched case (b), there are four matched features of which two are in reversed relative orders, leading to a lower score $M(\mathbf{q}; \mathbf{p}) = 4 - 2 = 2$ (where in this example $\lambda = 1$).

Note that the definition of matching score in Equation (2) is very general. We show particular implementation choices

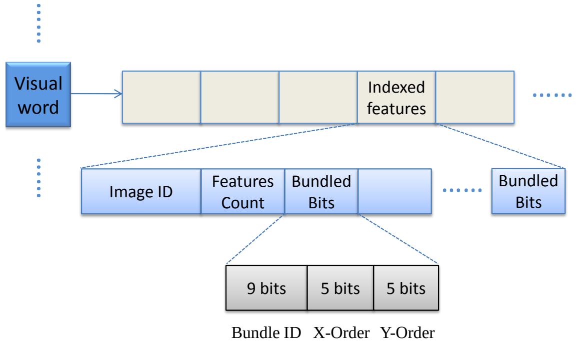

Figure 5. Inverted file structure. "Feature count" is the number of records of "bundled bits" in one image.

of $M_m(\mathbf{q};\mathbf{p})$ and $M_g(\mathbf{q};\mathbf{p})$ for our system, however more suitable forms may be defined for other image retrieval applications.

3.3. Indexing and retrieval

We use an inverted-file index [7] for large-scale indexing and retrieval. Fig. 5 shows the structure of our index. Each visual word has an entry in the index that contains the list of images in which the visual word appears. In addition to the image ID, for each occurrence of a visual word in a bundled feature we use 19 "bundled bits" to record the geometric information: 9 bits for the ID of the bundled feature within the image, 5 bits for X-order, and 5 bits for Y-order. This format supports at most 512 bundled features per image. If an image contains more than 512 bundles, the bundles containing the fewest features are discarded to remain within this limit. If a bundle contains more than 32 features, the ordinals denoting order are projected onto the range $[0, 31]$ to fit into 5 bits, so adjacent features may end up mapped to the same position in the order. If two bundled features have greater than 97% overlap in their constituent SIFT features, we only index one bundled feature. A traditional text index would contain the location of each word within the document in place of the bundled bits and thus this representation uses indexing space comparable to a text index containing an equivalent number of terms and documents.

Image retrieval is formulated as a voting problem. Each visual word in the query image votes on its matched images. The matched images are ranked by the sum of weighted votes. Suppose a query visual word and its matched visual word belong to the bundle feature $\mathbf{p}$ in the query image and the bundle feature $\mathbf{q}$ in the matched image respectively, we weight this vote using the matching score between two bundled features:

$$v = v_{\text{tfidf}} \cdot M(\mathbf{q};\mathbf{p}), \tag{6}$$

where v_{tfidf} is standard *tf-idf* weight [15] and v is the final weight. Thus features that occur as part of spatially-consistent groups across the two images score more highly, and we have augmented the bag-of-words model with a weak local geometric matching.

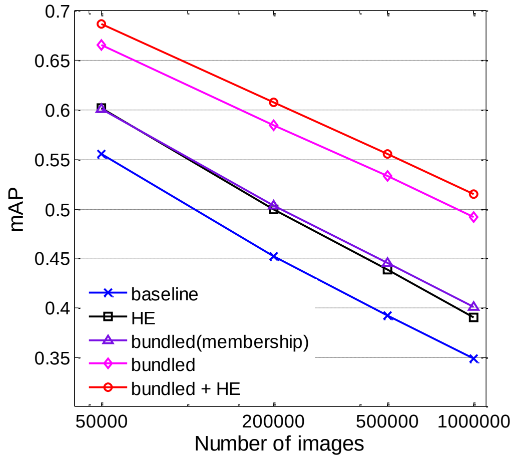

Figure 6. Comparison of different methods using mAP: "HE" is the baseline approach enhanced by hamming embedding; "bundled (membership)" is the approach using bundled features but with only the membership term in (2), i.e. without the geometric term. We also combine our bundled features with hamming embedding: the "bundled + HE" approach.

4. Experimental Results

We crawled one million images that are most frequently clicked in a popular commercial image-search engine to form our basic dataset. Then, we collected and manually labeled 780 partial-duplicate web images for our experiments[1]. These "ground truth" images form 19 groups and the images in each group are partial duplicates of each other. There are no exact (or very near-exact) duplicates in these images. Figures 1 and 11 show typical examples. We add the labeled images into the basic dataset to construct an evaluation dataset. To evaluate the performance with respect to the size of the dataset, we also build three smaller datasets (50K, 200K, and 500K) by sampling the basic dataset.

In the following evaluation, we select 150 representative images from the ground truth set as our queries. Following [3, 12] we use mean average precision (mAP) as our evaluation metric. For each query image we compute its precision-recall curve, from which we obtain its average precision and then take the mean value over all queries.

4.1. Evaluation

Baseline. We use a bag-of-features approach with soft assignment [13] as the "baseline" approach. We use a vocab-

[1] The basic dataset contains partial duplicates of our ground truth dataset. For evaluation purposes we identify and remove these partial-duplicate images from the basic dataset by querying the database using every image from the ground-truth dataset, and manually examining the returned images sharing any common visual words with the query images.

ulary of 1M visual words and the number of nearest neighbors in the soft assignment is set to 4. We experimented with different sizes (both larger and smaller) of visual word vocabulary, and found the 1M vocabulary to give the best overall performance.

Comparisons. We also enhance the baseline method with hamming embedding [3] by adding a 24-bit hamming code to filter out target features that have the same quantized visual word but have a large hamming distance from the query feature. We call this method "HE." Our bundled-feature based approach has three variants: 1) "bundled (membership)," in which we only use the membership term in (2); 2) "bundled," in which we use both the membership term and the geometric term; and 3) "bundled + HE," our "bundled" approach enhanced by the hamming embedding. In our implementation, λ in (2) is set to be 2 (see below for an experimental study of the effect of varying λ).

Figure 6 compares the above five approaches using mAP, leading to three major observations. First, the bundle membership term (see (2)) significantly improves the mAP, as can be seen by comparing the results for "bundled (membership)" to "baseline." On the 1M dataset, mAP is increased from 0.35 to 0.40, a 14% improvement. Second, the weak geometric term (relative ordering) plays a role as important as that of the membership term. The mAP reaches 0.49 (a 40% improvement) with both bundled-feature terms (membership + geometric), as shown by the curve labeled "bundled." Finally, the hamming embedding boosts both approaches. With the hamming embedding, we achieve the highest mAP 0.52 ("bundled + HE") on the 1M dataset, a 49% improvement over the baseline approach. Thus the bundled features are complementary to the hamming embedding.

Re-ranking. In the pipeline of an image retrieval system, re-ranking (full geometric verification) can substantially improve the retrieval performance as well as removing false positives by filtering out images that do not arise from valid 2D geometric transformations of the query image [6, 12]. We therefore evaluate the results of applying a full affine-transformation based re-ranking in our system. Re-ranking can be expensive since it requires additional disk IO and, for a distributed index, network communication on top of the computational cost of estimating affine transformations using RANSAC. Therefore we only re-rank a short list of the top 300 candidate images.

Figure 7 shows the re-ranking results. As can be expected, all approaches ("baseline," "bundled," and "bundled + HE") benefit from re-ranking. Note that even *without* the re-ranking, our approach ("bundled") achieves almost the same performance as "baseline + re-ranking" on the 1M dataset. When combined with re-ranking, our approach achieves mAP of 0.62, while the mAP of "baseline + reranking" is 0.50. We obtain a 77% improvement over "baseline," and a 24% improvement over "baseline + reranking." This is because re-ranking can only re-sort the short list, whereas our bundled-feature approach can fundamentally improve the ranking quality and bring more correctly matched images into the short list.

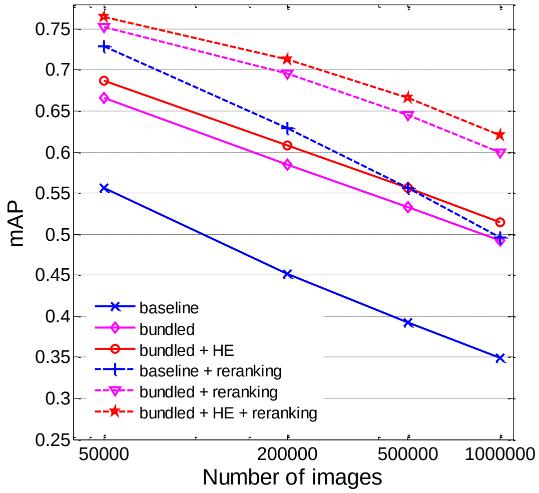

Figure 7. Comparison of different methods with and without reranking. In our system, re-ranking is performed on a short list of the top 300 images. The bundled features without re-ranking achieve performance comparable to the "baseline + re-ranking" approach.

Impact of λ. The λ value in (2) determines the weight of the geometric consistency term. We test the performance of our bundled features ("bundled" approach) using different λ values on the 1M dataset. Geometric consistency plays an important role in improving the mAP. Relying too much on geometric consistency at the expense of other signals, however, can reduce mAP. As Table 1 shows, $[1.5, 2.5]$ is the most effective range for the value of λ and in our reported results we use $\lambda = 2$.

λ	0	1.0	1.5	2.0	2.5	3.0
mAP	0.401	0.470	0.484	**0.492**	0.482	0.474

Table 1. Comparing the performance of bundled features on 1M dataset for different values of λ.

Runtime. We perform our experiments with a *single* CPU on a 3.0GHz Core Duo desktop with 16G memory. Table 2 shows the average query time for one image query. The feature extraction time is not included. As can be seen, the hamming embedding speeds up the system by filtering out some feature mismatches and thus reducing the number of features used for voting in the retrieval stage. Our bundled-feature approach achieves a significant improvement in the retrieval accuracy, while only introducing a modest time penalty (0.7 seconds). In comparison, full geometric reranking on the top 300 candidates introduces an additional

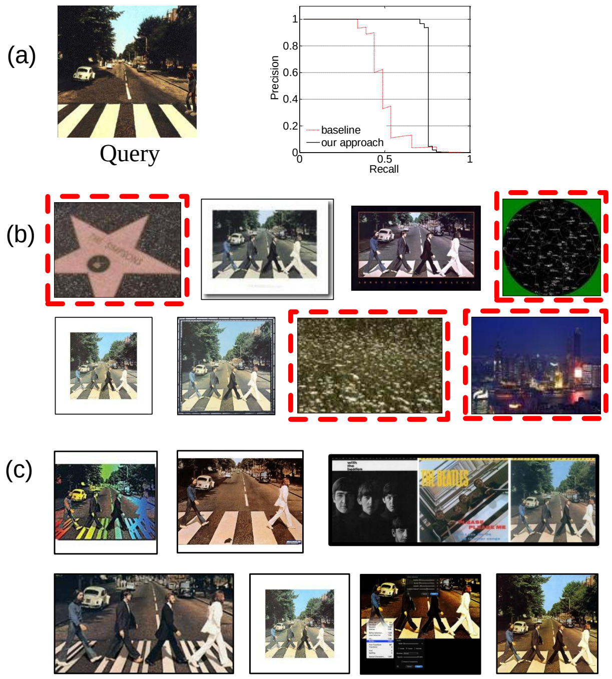

Figure 8. Sample results comparing our approach (without full geometric re-ranking) to the baseline approach. (a) Query image and a comparison of the Precision-recall curves. (b) The top images returned by the baseline approach (starting from the 13^{th} image). (c) The top images returned by our approach (starting from the 13^{th} image). The false positives are shown with red dashed bounding boxes. Note that the 3^{rd} image of (c) is a combination of three low resolution sub-images.

cost of 3.0 seconds, a much larger overhead but with mAP comparable to our bundled-feature approach.

	baseline	bundled features
without HE	1.7s	2.5s
with HE	1.2s	1.9s

Table 2. Average query time (not including feature extraction time).

4.2. Sample results

Figure 8 gives examples of our results on the 1M dataset. We show all results without full geometric re-ranking to emphasize the power of the bundled features. For this query, compared to baseline approach, our "bundled" approach improves the mAP from 0.51 to 0.74, a 45% improvement. Figure 8 (b) and (c) show the top images returned by the baseline approach and our approach, respectively. Because the top 12 images of both approaches are all correct (though they may be different), we show results starting from the 13^{th} returned image. The false positives are marked by red dashed bounding boxes. Although these false positives look irrelevant to the query image, they contain many local patches similar to those in the query image. These similar features are quantized into same visual words and contribute to false positives appearing in top images in the baseline approach.

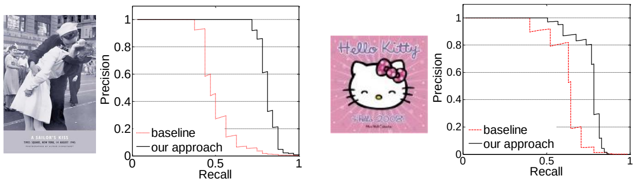

Figure 9. Two example queries and their precision-recall curves.

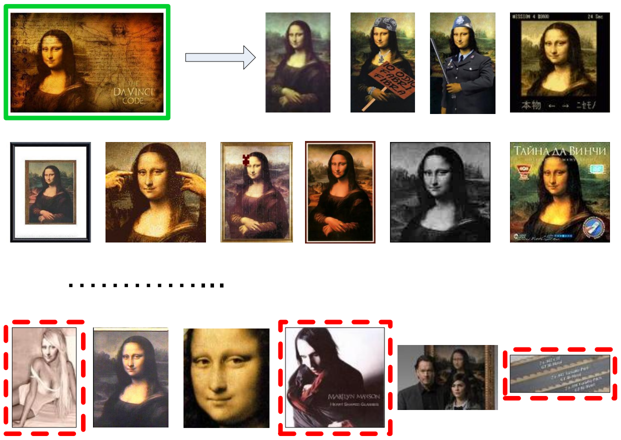

Figure 10. Top-ranked images returned from a "Da Vinci Code" image query. The query image is shown with a green bounding box and false positives with red dashed bounding boxes.

Figure 9 shows the precision-recall curves for another two example queries. The improvement on the "Hello Kitty" image is not as significant as for the other queries, since the baseline approach works relatively well on this image. In Figure 10 we show some false positives with low ranks. These false positives actually have similarities to the "Da Vinci Code" query image, in regions containing face, hair, shoulder, or text. Figure 11 shows more example results using our "bundled" approach without re-ranking. The retrieved images are diverse and contain large changes in scale and/or contrast, additions of text or framing, or significant editing (cropping and composition).

5. Conclusion

We have introduced bundled features for large scale partial duplicate web image search. Bundled features are a flexible representation with several desirable properties. First, they are more discriminative than individual SIFT features. Second, they allow us to enforce simple and robust geometric constraints at the bundle level. Finally, they allow us to *partially* match two groups of SIFT features, improving robustness to occlusion and image variations induced by photometric and geometric changes. Feature bundling and partial matching are a general and powerful framework. Our current implementation uses an MSER detec-

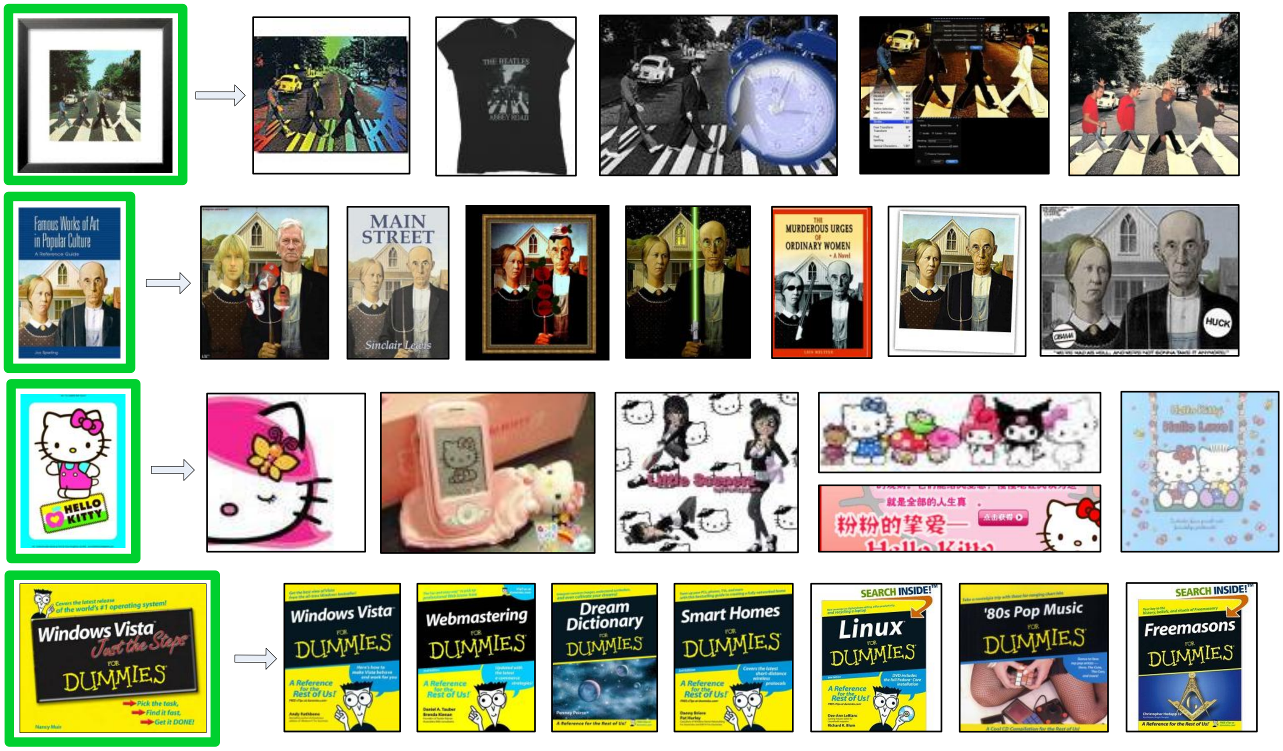

Figure 11. Example results. Queries are shown with green bounding boxes, and highly-ranked images (selected from those before the first false positive) from the query results are shown on the right.

tion to bundle SIFT features, but other bundling approaches could be applied. In the future we plan to investigate alternative bundling approaches as well as new bundle-level constraints for robust partial matching. As a flexible representation that is capable of partial matching, bundled features are also attractive for image-based object retrieval. Our recent experiments on reference data sets [11] showed a substantial improvement by using our bundled features. We plan to pursuit further along this direction.

Acknowledgment

We would like to thank Heung-Yeung Shum for fruitful discussions, and Dennis Fetterly for helping us crawl the web images.

References

[1] O. Chum, J. Philbin, M. Isard, and A. Zisserman. Scalable near identical image and shot detection. In *Proc. of the Int. Conf. on Image and Video Retrieval*, 2007.

[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 *ICCV*, 2007.

[3] H. Jegou, M. Douze, and C. Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *ECCV*, 2008.

[4] H. Jegou, H. Harzallah, and C. Schmid. A contextual dissimilarity measure for accurate and efficient image search. In *CVPR*, 2007.

[5] D. Liu, G. Hua, P. Viola, and T. Chen. Integrated feature selection and higher-order spatial feature extraction for object categorization. In *CVPR*, 2008.

[6] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 20:91–110, 2003.

[7] C. D. Manning, P. Raghavan, and H. Schtze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[8] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *BMVC*, 2002.

[9] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *PAMI*, 27(10):1615–1630, 2005.

[10] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *IJCV*, 65:43–72, 2005.

[11] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *CVPR'2006*.

[12] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *CVPR*, 2007.

[13] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *CVPR*, 2008.

[14] T. Quack, V. Ferrari, B. Leibe, and L. J. V. Gool. Efficient mining of frequent and distinctive feature configurations. In *ICCV*, 2007.

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

[16] P. Tirilly, V. Claveau, and P. Gros. Language modeling for bag-of-visual words image categorization. In *Proc. Int. conf. on Content-based image and video retrieval*, 2008.

[17] J. Yuan, Y. Wu, and M. Yang. Discovery of collocation patterns: from visual words to visual phrases. In *CVPR*, 2007.