GRID-BASED LOCAL FEATURE BUNDLING FOR EFFICIENT OBJECT SEARCH AND LOCALIZATION

Yuning Jiang, Jingjing Meng, Junsong Yuan

School of Electrical and Electronics Engineering,
Nanyang Technological University, Singapore, 639798

ABSTRACT

We propose a new grid-based image representation for discriminative visual object search, with the goal to efficiently locate the query object in a large image collection. After extracting local invariant features, we partition the image into non-overlapping rectangular grid cells. Each grid bundles the local features within it and is characterized by a histogram of visual words. Given both positive and negative queries, each grid is assigned a mutual information score to match and locate the query object. This new image representation offers two great benefits for efficient object search: 1) as the grid bundles local features, the spatial contextual information enhances the discriminative matching; and 2) it enables faster object localization by searching visual object in the grid-level image. To evaluate our approach, we perform experiments on a very challenging logo database *BelgaLogos* [1] of 10,000 images. The comparison with the state-of-the-art methods highlights the effectiveness of our approach in both accuracy and speed.

Index Terms— *grid feature, mutual information*

1. INTRODUCTION

Visual object search in large image collections is an important component for many applications, such as object recognition, image annotation and image understanding. Given a query object, our objective is to not only find out in the database all images that contain the object, but also locate the object in these images (see Figure 1). In this respect, visual object search can be viewed as two tasks: object matching and object localization.

Though much work [2] [1] [3] has focused on this area in recent years, visual object search, especially for small objects (e.g. logos), remains a challenging problem. On one hand, many challenges in object matching can be attributed to the visual differences between the target objects and the query, such as changes in scale, view point and color, and differences caused by partial occlusions. In light of the difficulties in object matching, we raise the need for a highly discriminative feature. [2] [4] opt for the Nearest-Neighbor (NN) classifier to avoid the quantization error caused by the bag-of-visual-words (*BOVW*) scheme and thus improve the matching accuracy. However, these NN-based algorithms are all under the Naive-Bayes assumption that each feature point is independent from the others. Without considering the spatial context, matching individual features can not provide satisfactory results. Besides, searching nearest neighbors for all query feature points is costly in both memory and time, hence prohibiting the application of NN classifiers to large datasets.

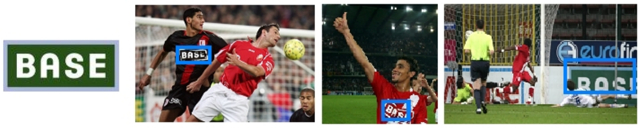
Fig. 1. *An example of visual object search.* Left: a query object, such as a logo, selected by the user. Right: resulting images from visual object search, where object locations are identified and marked by blue bounding boxes.

On the other hand, object localization is formulated as the problem of finding the subimage with maximum similarity to the query object [2] [5]. Although use of branch-and-bound algorithm can avoid linear search through all the subimages of an image, object localization is still a computationally expensive job for high resolution images (i.e., 800×800 and higher), especially when the target object is small.

To address the two tasks mentioned above, we propose a grid-based visual object search approach in this paper. We first partition each image into non-overlapping rectangular grids and bundle local features in each grid into a *grid feature*, which is described as a visual word histogram under the *BOVW* framework. Then given the positive and negative queries, each grid is assigned a mutual information score determined by its histogram intersections with both positive and negative sets. Finally the subimage with maximum mutual information, computed as the summation of the mutual information scores of all its grids, is retrieved using the branch-and-bound algorithm.

Our approach contributes to both tasks for visual object search: (1) For object matching, it improves the matching accuracy via the discriminative grid matching; On one hand, instead of matching individual local features, we match grids as a whole using the bundled features. By considering the spatial

This work is supported in part by the Nanyang Assistant Professorship (SUG M58040015) to Dr. Junsong Yuan.

context, it thus improves the matching quality. On the other hand, instead of matching the query object only, each grid matches both positive and negative queries to enable a more discriminative matching. (2) For object localization, branch-and-bound search at the grid level drastically reduces both the time and space complexity, as it is essentially performing search on down-sampled images.

2. ALGORITHM

2.1. Grid Feature

Given an image database $\mathcal{D} = \{\mathcal{I}_i\}$, we denote by $\{f_{i,j}\}$ all the high-dimensional local descriptors extracted from the image $\mathcal{I}_i$. Follow the *BOVW* scheme, each local descriptor f is quantized to a visual word using a vocabulary of K words, represented as $w = \{x, y, d\}$, where (x, y) is the location and $d \in \{1, \dots, K\}$ is the corresponding index of the visual word.

Then we partition each image $\mathcal{I}_i$ into $M_i \times N_i$ non-overlapping rectangular grid cells $\{R_{i,m,n}\}$, $m \in \{1, \dots, M_i\}$ and $n \in \{1, \dots, N_i\}$. A grid feature is then defined as:

$$G_{i,m,n} = \{w_{i,j} | w_{i,j} \propto R_{i,m,n}\}, \qquad (1)$$

where $w_{i,j} \propto R_{i,m,n}$ means the point feature $w_{i,j}$ falls inside the grid cell $R_{i,m,n}$. Empty grids are discarded. Furthermore, each grid feature $G_{i,m,n}$ is represented as a K-dimensional histogram of visual word occurrences $h_{i,m,n}$, and indexed by an inverted file to take advantage of its sparsity. Figure 2 illustrates how to construct and index the grid features.

A grid feature is more discriminative than an individual local feature, as it contains multiple features as the context [6] [7]. And thanks to the *BOVW* scheme, we need not to store and match all local features in a high-dimensional space. In practise, the inverted index results in a substantial speedup as only grids containing the words that also occur in the query need to be examined.

2.2. Mutual Information Score based on Histogram Intersection

After representing each grid as a sparse histogram $h_{i,m,n}$, our next step is to evaluate the similarity between each subimage I and the query object. Here we propose the mutual information score as the similarity measure based on histogram intersection. First let us introduce the *Normalization Histogram Intersection (NHI)*. For any two histogram h_1 and h_2, we have:

$$NHI(h_1, h_2) = \frac{|h_1 \cap h_2|}{|h_1 \cup h_2|} = \frac{\sum_k \min(h_1^k, h_2^k)}{\sum_k \max(h_1^k, h_2^k)}, \qquad (2)$$

where h_1^k and h_2^k are the k_{th} dimensions of the histograms.

Then given the positive query Q_+ and negative query Q_-, which are represented as the visual word histograms h_{Q_+} and h_{Q_-} respectively, we define the distance between a grid feature G and the query as:

$$D(G, Q) = 1 - NHI(h_G, h_Q) \in [0, 1]. \qquad (3)$$

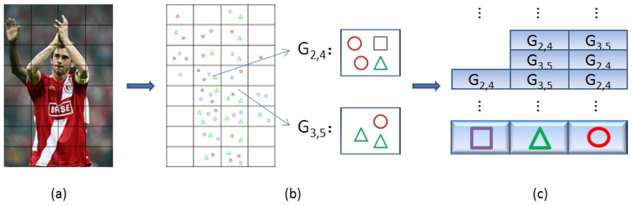

Fig. 2. *Construct and index the grid features.* (a) Partition the original image into grid cells. (b) Construct grid features using *BOVW* scheme. The rectangles, circles and triangles stand for different visual words. (c) Index grid features using an inverted file.

Assume that grids are independent from each other, the mutual information score of the subimage I can be calculated as the summation of the scores of all the grids it contains [2] [8]:

$$\begin{aligned} s(I) &= MI(Q_+, I) = \log \frac{p(I|Q_+)}{p(I)} \\ &= \log \frac{\prod_{G \in I} p(G|Q_+)}{\prod_{G \in I} p(G)} = \sum_{G \in I} \log \frac{p(G|Q_+)}{p(G)} \\ &= \sum_{G \in I} \log \frac{p(G|Q_+)}{p(G|Q_+)p(Q_+) + p(G|Q_-)p(Q_-)} \\ &= \sum_{G \in I} \log \frac{1}{p(Q_+) + \frac{p(G|Q_-)}{p(G|Q_+)} p(Q_-)} \\ &= \sum_{G \in I} s(G), \end{aligned} \qquad (4)$$

where $s(G)$ is the mutual information score of a grid feature G. To evaluate the conditional distributions $p(G|Q_-)$ and $p(G|Q_+)$, the Gaussian kernel based on histogram intersection is adopted:

$$\begin{aligned} \frac{p(G|Q_-)}{p(G|Q_+)} &= e^{-\frac{1}{2\sigma^2}(D(G,Q_-) - D(G,Q_+))} \\ &= e^{-\frac{1}{2\sigma^2}(NHI(h_G, h_{Q+}) - NHI(h_G, h_{Q-}))}. \end{aligned} \qquad (5)$$

Compared to the NN-based method [2] assigning each local feature a mutual information score, the grid-based approach relaxes the Naive-Bayes assumption, as we allow intra-grid dependency over feature points, while still enforcing inter-grid independency.

2.3. Branch-and-Bound Search

For an image $\mathcal{I}$, object localization is formulated as the problem of finding a rectangular region I^* of $\mathcal{I}$ that has the maximum mutual information score to the query:

$$I^* = \arg\max_{I \subseteq \mathcal{I}} MI(Q_+, I) = \arg\max_{I \subseteq \mathcal{I}} \sum_{G \in I} s(G). \qquad (6)$$

Since exhaustively locating the subimage is $O(M^2N^2)$ if the image $\mathcal{I}$ consists of $M \times N$ grids, we employ the branch-and-bound algorithm to avoid the exhaustive search. Now given

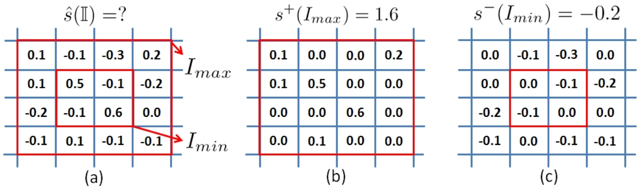

Fig. 3. *Illustration of the definition of upper bound function.* (a) is the original confidence map, in which each grid is assigned a mutual information score. (b) is the confidence map containing only positive grids, while (c) contains only negative ones. Hence the upper bound of $\mathbb{I}$ is $\hat{s}(\mathbb{I}) = s^+(I_{max}) + s^-(I_{min}) = 1.4$.

the mutual information score $s(I)$ as the quality function, in the following we explain how to derive the upper bound function $\hat{s}(\mathbb{I})$, where $\mathbb{I}$ is a collection of subimages in image $\mathcal{I}$.

Similar to the ESS algorithm [5], we assume that there exist two subimages I_{min} and I_{max} such that for any $I \in \mathbb{I}$, $I_{min} \subseteq I \subseteq I_{max}$. Then the upper bound function is defined as:

$$\hat{s}(\mathbb{I}) = s^+(I_{max}) + s^-(I_{min}), \tag{7}$$

where $s^+(I) = \sum_{G \in I} \max(s(G), 0)$ contains only positive grids, while $s^-(I) = \sum_{G \in I} \min(s(G), 0)$ contains only negative ones, as shown in Figure 3. Both $s^+(I)$ and $s^-(I)$ can be computed in $O(1)$ operations using the integral images. It is easy to see that $\hat{s}(\mathbb{I})$ meets the two conditions of an upper bound function, as proposed in [5]:

$$i) \quad \hat{s}(\mathbb{I}) \geq \max_{I \in \mathbb{I}} s(I), \tag{8}$$

$$ii) \quad \hat{s}(\mathbb{I}) = s(I), \text{ if } I \text{ is the only element in } \mathbb{I}. \tag{9}$$

Consider that our objective is to find the top-K subimages from the entire image database $\mathcal{D}$, the branch-and-bound algorithm is initialized using all images $\mathcal{I}_i \in \mathcal{D}$. The iteration stops after the top-K results are returned so that the images with low scores will never be processed.

Essentially, the grid-based subimage search down-samples the images to a lower resolution, which greatly decreases the total number of subimages. In our experiment it performs object search in a database of $10,000$ images within seconds. At the same time, memory usage is reduced because the integral images are constructed at the grid level as well. For example, an image of resolution 800×800 only costs $10K$ of RAM when the grid size is fixed at 16×16, while it costs more than $2.5M$ if using the original resolution of 800×800.

3. EXPERIMENTS

3.1. Experimental Setup

We evaluate our approach on a very challenging logo database *BelgaLogos* [1] of $10,000$ images covering various aspects of life and current affairs. As in [1], all images are re-sized with a maximum value of height and width equal to 800 pixels, preserving the original aspect ratio. And in total more than 24 millions scale and affine invariant interest points are extracted by the Harris-Affine detector and described by 128-dimensional SIFT descriptors [9]. Finally all the descriptors are clustered into a vocabulary of 1M visual words using the Hierarchical K-Means (HKM) method in [10].

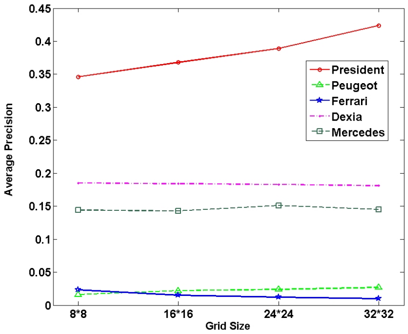

Fig. 4. *Performance of grid features at different scales, from* 8×8 *to* 32×32*.*

Since the images in BelgoLogos are of different aspect ratios, in practice we fix the grid size when dividing up the images. We test 4 different grid sizes (8×8, 16×16, 24×24 and 32×32) and compare their performance. To test the effectiveness of our object search algorithm, 5 external logos used in [2] are selected as the query objects. Meanwhile, we randomly pick out two images containing no logos from the database as the negative queries.

3.2. Result Evaluation

To make a fair comparison with previous work, we evaluate our approach using both Precision/Recall (P/R) scores and Average Precision (AP). Since the BelgaLogo database does not provide the location of each logo in its groundtruth images, we consider a retrieved image containing the query logo as a correct detection. In our experiments, we manually check each returned detection to ensure the bounding box touches the target object. For each query, the top 100 subimages are returned as the retrieval results.

First, we test how the grid size affects AP, as shown in Figure 4. We can see that as the grid size increases, AP of each logo changes in different ways. For instance, AP for the Presidential logo increases while AP for the Ferrari logo falls slightly. The reason is that in the database the President logos always appear in a larger size than that of Ferrari, and enlarging grid size may risk introducing noise for small logos, hence affecting the precision adversely. Because of space limitation we only discuss results of 24×24 grid size in later experiments.

Then we compare our approach with the NN-based mutual information algorithm (NN-MI) [2] and the baseline method [1]. Since the published NN-MI results were evaluated by P/R score, here we compare our precision with it given the same recall. To make a fair comparison, our initial retrieved results are re-ranked by the RANSAC algorithm as is done in the baseline method. The comparison results are showed in Table 1 and Table 2 respectively. It demonstrates that our approach has a significant improvement over NN-MI; and the re-ranked results are sightly better than that of the baseline method. Moreover, our algorithm can accurately

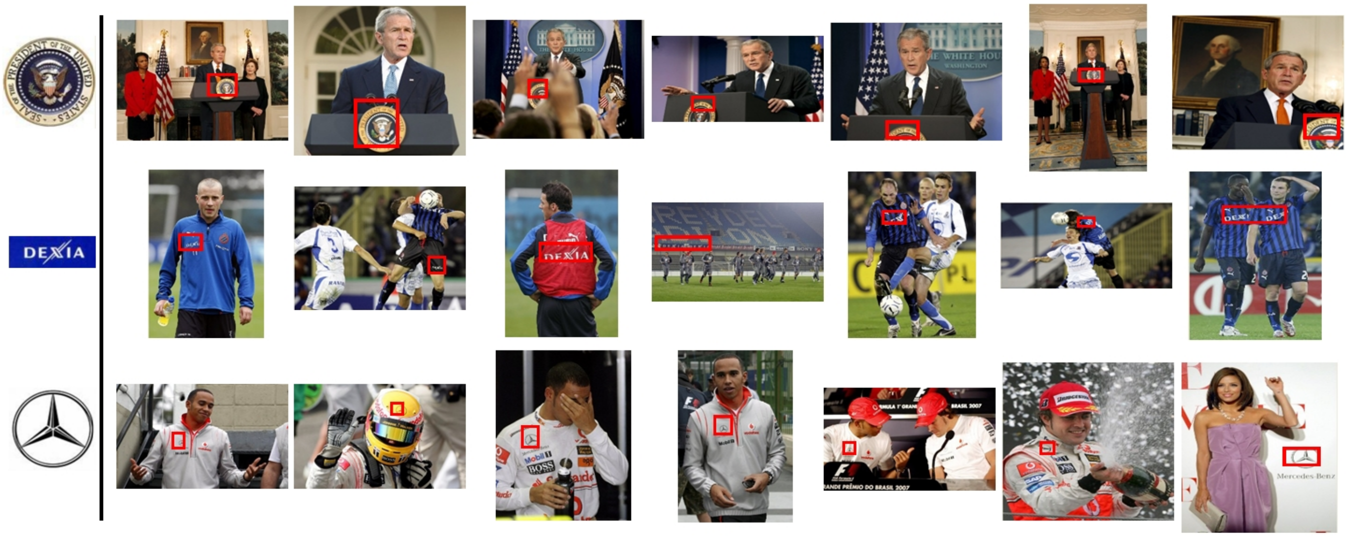

Fig. 5. *Examples of search results for 3 logos: President, Dexia and Mercedes.* The query is shown on the left, with selected top ranked retrieved images shown on the right.

		NN-MI[2]	Grid-based
	recall	precision	precision
Dexia	0.096	0.810	0.699
Ferrari	0.013	0.010	0.333
Mercedes	0.145	0.917	0.917
Peugeot	0.167	0.010	0.053
President	0.357	0.050	0.455
Average		0.359	0.491

Table 1. *Comparison with the NN-MI[2] using precision given the same recall.*

separate the object from cluttered background, as shown in Figure 5.

3.3. Running Time

As the time cost was not published in previous papers, here we just present the time cost of our approach and make the comparison between different grid scales. All algorithms are implemented by C++ and run on a single PC of 2.6G Intel CPU and 2G main memory. The running time shown in Table 3 is the average time cost for 5 logos, including query feature extraction, similarity measurement and subimage search. From Table 3 we can see that enlarging grid size significantly speeds up the subimage retrieval.

4. CONCLUSION

	Baseline[1]	Grid-based + RANSAC
Dexia	0.293	0.211
Ferrari	0.075	0.031
Mercedes	0.185	0.245
Peugeot	0.207	0.202
President	0.603	0.688
Average	0.273	0.276

Table 2. *Comparison with the Baseline[1] using AP.*

Grid size	8 × 8	16 × 16	24 × 24	32 × 32
Running time(s)	26.1	13.9	7.2	4.9

Table 3. *Time cost at different grid scales.*

In this paper, we introduce a grid feature to search visual object in a large image collection. By bundling the spatial nearest neighbors, grid feature is more discriminative than individual local features. Moreover, it significantly reduces both time and memory usage when combined with branch-and-bound subimage search scheme. Although we implement the grid feature using only quantized visual words, other regional features, e.g. color histogram, can also be bundled and contribute to the mutual information score. We believe that as a flexible image representation, the grid feature will be of great value to other image-related applications. Our experiments on the *BelgaLogos* logo dataset validate the effectiveness and efficiency of our grid-based method.

5. REFERENCES

[1] Alexis Joly and Olivier Buisson, "Logo retrieval with a contrario visual query expansion," in *Proc. ACM Multimedia*, 2009.

[2] Jingjing Meng, Junsong Yuan, Yuning Jiang, Nitya Narasimhan, Venu Vasudevan, and Ying Wu, "Interactive visual object search through mutual information maximization," in *Proc. ACM Multimedia*, 2010.

[3] Ondrej Chum, James Philbin, Josef Sivic, Michael Isard, and Andrew Zisserman, "Total recall: automatic query expansion with a generative feature model for object retrieval," in *Proc. IEEE Intl. Conf. on Computer Vision*, 2007.

[4] O. Boiman, E. Shechtman, and M. Irani, "In defense of nearest-neighbor based image classification," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2008.

[5] C. H. Lampert, M. B. Blaschko, and T. Hofmann, "Efficient subwindow search: a branch and bound framework for object localization," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2009.

[6] Junsong Yuan and Ying Wu, "Spatial random partition for common visual pattern discovery," in *Proc. IEEE Intl. Conf. on Computer Vision*, 2007.

[7] Zhong Wu, Qifa Ke, M. Isard, and Jian Sun, "Bundling features for large scale partial-duplicate web image search," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2009.

[8] Junsong Yuan, Zicheng Liu, and Ying Wu, "Discriminative subvolume search for efficient action detection," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2009.

[9] David Lowe, "Distinctive image features from scale-invariant keypoints," *Intl. Journal of Computer Vision*, 2004.

[10] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2007.