Visual Phraselet: Refining Spatial Constraints for Large Scale Image Search

Liang Zheng and Shengjin Wang, *Member, IEEE*

***Abstract*—The Bag-of-Words (BoW) model is prone to the deficiency of spatial constraints among visual words. The state of the art methods encode spatial information via visual phrases. However, these methods discard the spatial context among visual phrases instead. To address the problem, this letter introduces a novel visual concept, the *Visual Phraselet*, as a kind of similarity measurement between images. The visual phraselet refers to the spatial consistent group of visual phrases. In a simple yet effective manner, visual phraselet filters out false visual phrase matches, and is much more discriminative than both visual word and visual phrase. To boost the discovery of visual phraselets, we apply the soft quantization scheme. Our method is evaluated through extensive experiments on three benchmark datasets (Oxford 5 K, Paris 6 K and Flickr 1 M). We report significant improvements as large as 54.6% over the baseline approach, thus validating the concept of visual phraselet.**

***Index Terms*—Image search, spatial constraint, visual phrase, visual phraselet.**

I. Introduction

CONTENT Based Image Retrieval (CBIR) is one of the key techniques of manipulating the sheer amount of image/video collections. The variety of visual content of images poses great challenge for CBIR systems, especially in the setting of large-scale datasets. Given a query image or region, our goal is to locate the most relevant ones from a large corpus of images.

The Bag-of-visual Words (BoW) model appears to be the *de-facto* approach in the state-of-the-art CBIR systems [7], [8]. It functions by quantizing local feature descriptors (such as SIFT) into visual words, and then employing scalable text indexing and retrieval schemes.

There exists a plethora of work aiming at improving the search performance. The most significant advancement is the adoption of SIFT descriptors [14]. To speed up the assignment of local descriptors to visual words, tree-like data structures [2] are employed. Hamming embedding [4] uses binary codes to filter our false matches, and soft assignment [16] associates each descriptor with multiple visual words to improve the initial recall. Post-processing techniques, e.g., spatial verification by RANSAC [2] and query expansion [15] which re-issues the initial result as queries, have also been leveraged.

However, the BoW model represents an image as a bag of orderless visual words, without incorporating spatial constraints. This problem is the hot spot of current research and many literatures attempt to overcome this issue.

Typically, post-processing methods involve full geometric verification (RANSAC) [2]. It is computationally expensive and can only be applied to a subset of the top-ranked candidate images. Others incorporate spatial information into the retrieval framework. Spatial Pyramid Matching (SPM) [9] utilizes rigid spatial information encoded by coarse image quantization, but lacks invariance to image translation and rotation [10]. Spatial information is also considered by building contextual visual vocabulary [11], or by contextual weighting for visual words [7]. Furthermore, constructing feature of high orders [3], [12], [13] is also a challenging yet effective method. Bundled-features [12] are detected by grouping local features within MSER regions. Some [3], [13] employ supervised approaches to find the co-occurrences of visual words and form a visual phrase vocabulary. However, this method is sensitive to image noise and requires a data-dependent training process, which do not scale for large scale applications.

In this letter, we are interested in incorporating more spatial information. Current methods neglect the spatial information between visual phrases. This problem deteriorates the performance as described in Fig. 1. To tackle this, we propose a novel visual element, called *Visual Phraselet*, to describe spatially consistent sets of visual phrases. We extend the previous work [5] on mining geometry-preserving visual phrases (GVPs) and propose a simple yet effective approach to efficiently discover visual phraselets. We show that visual phraselet encodes more strict spatial information and achieves significant improvement over the baseline approach.

II. Visual Phraselet

In this letter, we define the *Visual Phraselet* as a group of visual phrases that manifest consistent spatial relationships.

Our work is built upon the geometry-preserving visual phrase (GVP) [5], as illustrated in Fig. 2. Briefly, this method quantizes an image into predefined bins, for example a 10 × 10 partition. Then, each visual word is represented by the word index and its coordinates in terms of the quantized bins. On the matching of two identical visual words j in the images I and I', their offset $(\Delta x_j, \Delta y_j)$ is calculated via subtracting the coordinates of word in images I from that in image I'. Then, a vote is generated in the offset space at $(\Delta x_j, \Delta y_j)$. In this manner, the spatial information of visual words are recorded

Manuscript received December 04, 2012; revised February 05, 2013; accepted February 16, 2013. Date of current version March 07, 2013. This work was supported by the National High Technology Research and Development Program of China (863 program) under Grant 2012AA011004 and the National Natural Science Foundation of China under Grant 61071135. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Chao Tian.

The authors are with the Department of Electronic Engineering, Tsinghua University, Beijing 100086, China (e-mail: zheng-l06@mails.tsinghua.edu.cn; wgsgj@tsinghua.edu.cn).

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

Digital Object Identifier 10.1109/LSP.2013.2249513

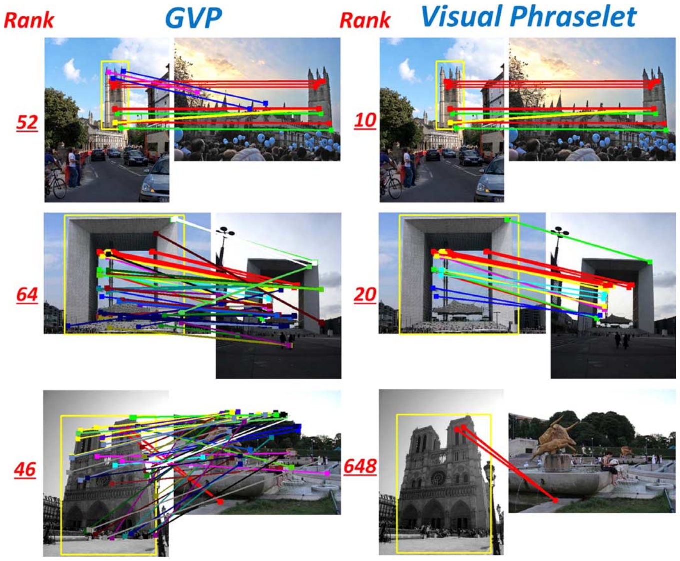

Fig. 1. Matched images are shown between query and candidate images. To the left are pairs obtained by visual phrases, while those by visual phraselets are aligned on the right. Also shown are their ranks.

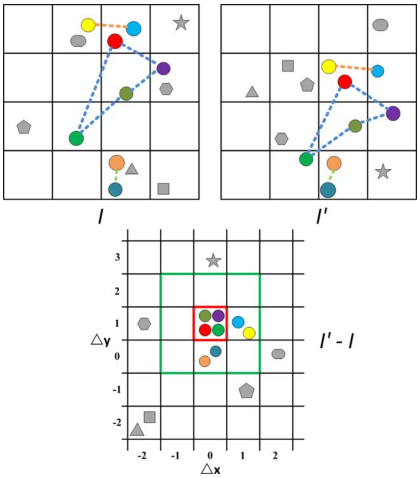

Fig. 2. Construction of GVP and Visual Phraselet. (Top): A pair of matched images. Each different symbol represents a distinct visual word. The colored circles are points located on the target object, while the gray symbols indicate distracter points in the background. Matching pairs of points are in the same color and of the same shape. (Bottom): Offset space produced from the two matching images. After subtracting the coordinate of points in the second image by points in the first image, votes are generated and fall into the corresponding bins in the offset space. Visual words of the same object generally locate in neighboring bins while distracter words scatter about. By considering only the visual phrases within the eight-neighborhood (green boxed) of the maximum response bin (red boxed), a visual phraselet is generated which filters out spatial inconsistent visual phrases.

in the offset space: by checking the values in the offset space, visual phrases of length k can be efficiently located. Since GVP of length-2 was shown to outperform other lengths in [5], in what follows, we always refer to GVP of length-2 if not specified. Same with the common weakness of visual phrase mining algorithms, GVP method again neglects the spatial information among visual phrases.

The basic idea of visual phraselet is rooted on the key inference: votes in the offset space tend to concentrate for relevant images, but distribute ”randomly” for irrelevant ones. To prove it, we assume a global geometrical transform between the query and database images [4], [2]. Assume images I and I' both contain the same object (thus are a relevant image pair). The location of the object undergoes a shift of (δ_x, δ_y) from I to I' in the 10×10 coordinate system. Let $q = \{w_i, (x_i, y_i)\}$ and $p = \{w'_i, (x'_i, y'_i)\}$ $(i = 1 \sim m)$ be the m matched features belonging to the query and target image, respectively, where inside the parentheses are the visual word indexes and locations. Ideally, all these matched points reside in the foreground object, and the background has no matches,

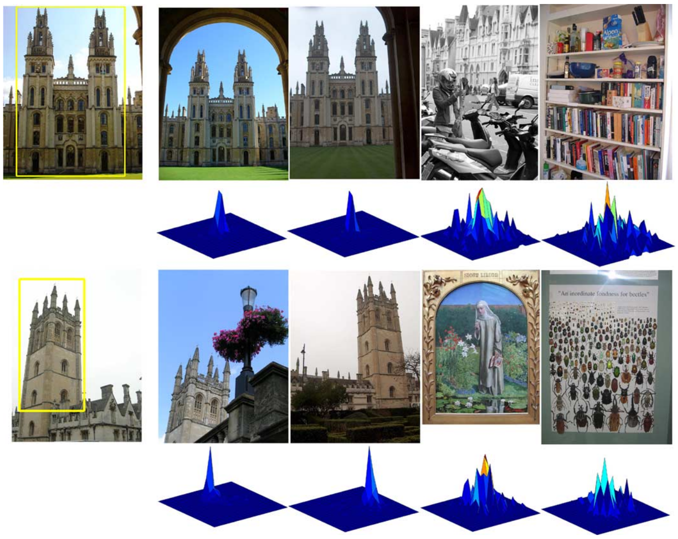

Fig. 3. Profile of offset space under different image pairs. The first and third rows are image pairs: (left) a query image; (middle) two positive retrieved images; (right) two irrelevant images. The second and fourth rows are the corresponding offset tables.

$$w_i = w'_i, \text{ for } i = 1, 2, \ldots, m \tag{1}$$

$$(x'_i, y'_i) = (x_i, y_i) + (\delta_x, \delta_y), \text{ for } i = 1, 2, \ldots, m. \tag{2}$$

Therefore, the offset table is expressed as:

$$S_{i,x,y} = \begin{cases} m, & x = \delta_x, y = \delta_y, \\ 0, & \text{others.} \end{cases} \tag{3}$$

Taking into account affine transformations, the situation may deviate from optimality: a few bins in the offset table take nonzero values, and take on a peak-like profile. In a sharp contrast, descriptor matching pairs in two irrelevant images are not consistent in geometry relationship. As a result, votes in the offset table are scattered (Fig. 3).

In other words, votes in the ”peak” region of the offset table correspond to spatial consistent visual phrases, while those far apart are inconsistent phrases. We identify the largest value in the offset table, preserve it as well as its 8 neighbors, and only employ these 9 bins to calculate matching scores between the query and target images. The length-2 GVPs within these bins are assembled as a visual phraselet (Fig. 2).

Visual phraselet alone imposes geometry constraints to the matching process, but due to errors introduced by illumination, detector drift, quantization, etc, we may lose some information of visual word itself. Therefore, our final scoring function is a weighted sum of visual word and visual phraselet matching:

$$\mathcal{K}(I, I') = (1 - w) \cdot \Theta(I, I') + w \cdot \hat{\Theta}(I, I') \tag{4}$$

where $\Theta(\cdot)$ and $\hat{\Theta}(\cdot)$ are similarity measures for visual word matching and visual phraselet matching, respectively, and $\mathcal{K}(\cdot)$ indicates the matching score between query image I and target image I'. The weight w is between 0 and 1. Empirically we set w to 0.2.

In this letter, the similarity functions are defined as follows.

$$\Theta(I, I') = \sum_{m} D_{i,m} \tag{5}$$

$$\hat{\Theta}(I, I') = \sum_{m \in neigh(\delta_x, \delta_y)} D_{i,m} \begin{pmatrix} S_{i,m} - 1 \\ k - 1 \end{pmatrix} \tag{6}$$

where $D_{i,m}$ is the IDF weight table as derived in [5], $neigh(\delta_x, \delta_y)$ denotes the 9 bins with highest responses, and k encodes the GVP length (equals 2).

III. Experiments

To evaluate the proposed approach on visual phraselet based representation, we conducted experiments on three benchmark datasets. For all the experiments, codebooks are generated on the Oxford 5 K dataset. Mean Average Precision (mAP) is used to measure image search accuracy.

A. Baseline

We adopt the image search procedure introduced in [2] as the baseline approach.

We first extract Hessian-affine local regions from which the SIFT descriptors are computed, using the publicly available software from [2]. Then, we employ the Approximate K-means (AKM) algorithm to construct visual word codebooks of various sizes, i.e., 50 K, 100 K, 250 K, 500 K, 750 K, and 1 M, respectively. Quantification is accelerated by the approximate nearest neighbors (ANN) indexing structure. Furthermore, inverted files are built to index each image and the associated attributes. In the searching step, scores for each image are obtained and compared based on standard TF-IDF weighted visual words.

B. Datasets

1) Oxford 5 K [2]: A total number of 5062 images were obtained from Flickr. By manual annotation based on the relevance to the queries, this dataset has been generated as a comprehensive ground truth for 11 distinct landmarks, each containing 5 queries. In total there are 55 query images.

2) Paris 6 K [6]: This dataset contains about 6400 high resolution images from Flickr by text queries of Paris landmarks. Again, Paris dataset is featured by 55 queries of 11 different landmarks.

3) Flickr 1 M [4]: The Flickr 1 M dataset are distractor images arbitrarily retrieved from Flickr. These images are added into the Oxford 5 K and Paris 6 K datasets to test the scalability of our approach.

C. Effect of Codebook Size

The proposed method is evaluated against different codebook sizes. We compare the visual phraselet method with the baseline and GVP based approach. To boost visual word matches and enhance visual phraselet discovery, we adopt the soft quantization, similar to the multiple descriptor assignment proposed in [4] and [16]. In our implementation, soft quantization is applied to the query images only, so that the memory usage of inverted files remains unchanged. One descriptor is assigned to its 3 nearest neighbors in the codebook, with weights to each center $\exp(-d^2/2\sigma^2)$, where d is the Euclidean distance between cluster centres and $\sigma^2 = 6250$.

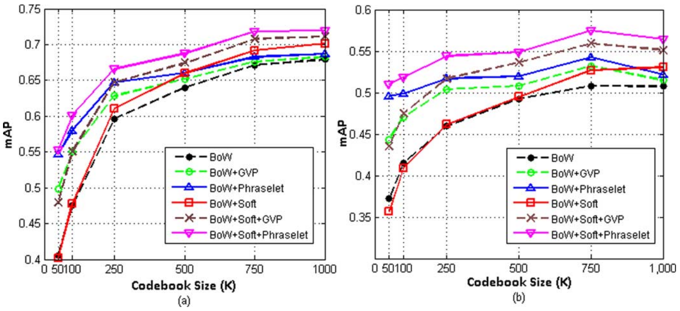

Fig. 4. mAP as a function of the codebook size. Results for (a) Oxford 5 K and (b) Paris 6 K datasets are presented.

The experimental results are demonstrated in Fig. 4. It is shown that, for different codebook sizes, the proposed method improves the search accuracy, over both the BOW and the GVP approaches. Moreover, the improvement is more significant on smaller codebooks. For small codebooks, the defined visual words are more ambiguous, and visual phraselet increases the discriminative power of visual word by filtering out false positive matches. Further, to boost visual word matching, soft quantization of visual words is employed. Large codebooks benefit from more matches, but small codebooks suffer from more noise, the case shown by adopting soft quantization alone. However, when visual phraselet representation is combined, the improvement of small codebook over the GVP method is more significant. It is because soft quantization not only boosts recall, but also bring about wrong GVP matches, deteriorating its performance. Our method rectifies this phenomenon, so the amplitude of improvement is larger for small codebooks.

D. Scalability and Efficiency

To test the scalability and efficiency of our method, we further combined the Oxford 5 K and Paris 6 K datasets with the Flickr 1 M dataset [4] as distractor images, and the results are summarized in Fig. 5, Tables I and II.

Fig. 5(a) and (c) indicate that our proposed method consistently outperforms the other models. The mAP of BOW and GVP can be improved by 24.7% and 7.8% on Oxford dataset and by 40.6% and 8.6% on Paris dataset populated with one million images. Moreover, from Fig. 5(b) and (d), it is shown that the gain in mAP is more significant if soft quantization is applied. The relative improvement arrives at 33.9% and 54.6% over the BOW method for the two datasets, respectively. This is because soft quantization leads to more invalid matches which are filtered out by visual phraselet. Further, we note that when the database gets scaled up, the mAP of our method drops much slower compared with other methods. In other words, more notable improvement is obtained on larger databases, thus confirming the scalability of the visual phraselet. Last but not least, visual phraselet-based method brings about more improvement on Paris 6 K + Flickr 1 M dataset than on Oxford 5 K + Flickr 1 M dataset. In fact, the codebook trained on Oxford 5 K is quite

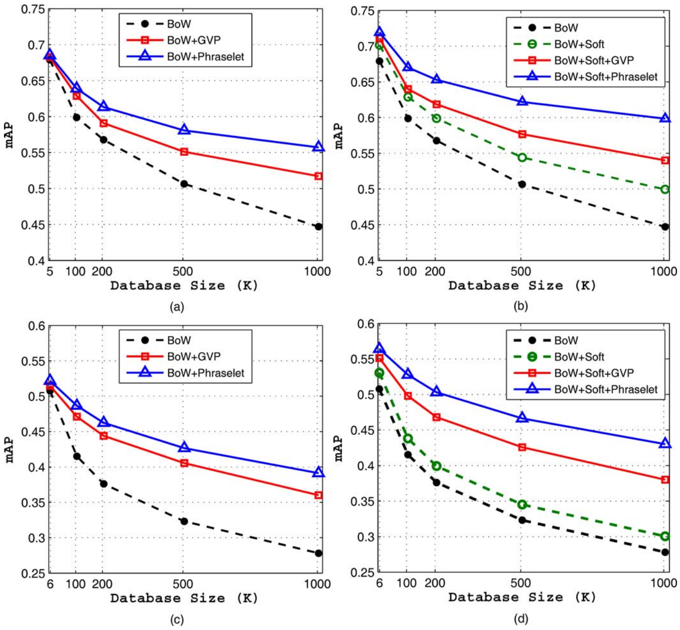

Fig. 5. mAP for Oxford 5 K with (a) hard quantization, (b) soft quantization and Paris 6 K with (c) hard quantization, (d) soft quantization, scaled with Flickr 1 M dataset. Three methods are employed, i.e., the BoW baseline, the GVP method and the proposed visual phraselet, respectively. (a) Oxford 5 K, hard quantization. (b) Oxford 5 K, soft quantization. (c) Paris 6 K, hard quantization. (d) Paris 6 K, soft quantization.

TABLE I
COMPARISON OF mAP FOR VARIOUS METHODS ON BENCHMARK DATASETS

	GVP	Phraselet	Oxford		Paris	
			$5K$	$5K+1M$	$6K$	$6K+1M$
Hard			0.679	0.447	0.508	0.278
Hard	×		0.683	0.517	0.515	0.360
Hard		×	0.685	0.557	0.522	0.391
Soft			0.702	0.500	0.531	0.301
Soft	×		0.711	0.540	0.551	0.380
Soft		×	**0.719**	**0.599**	**0.564**	**0.430**

TABLE II
EFFICIENCY RESULTS ON OXFORD DATASETS

Average Query Time (s)	BoW	BoW+GVP	BoW+Phraselet
Oxford 5K	0.026	0.183	0.186
Oxford 5K + Flickr 1M	0.677	2.595	2.780

discriminative on the same dataset, but is ambiguous for an irrelevant dataset. The proposed method functions by enhancing the discriminative power of visual representation, so more improvement could be observed on Paris dataset. This indicates that our method generalizes well to the case where the codebook is trained on irrelevant data and the improvement is much more considerable.

We run our experiment using Matlab 2010b on a 2.40-GHz CPU of a Sixteen-Core Intel Xeon server with 32 GB memory. The results for efficiency are shown in Table II. On both the Oxford 5 K and Oxford 5 K+Flickr 1 M datasets, the query time of phraselet-based system increases marginally over the GVP approach. The increased response time is mainly attributed to the filtering process. In fact, compared with the GVP method, phraselet achieves a noticeable gain in accuracy on all datasets, especially on the 1 M dataset, so the limited efficiency loss is worthwhile. Moreover, [5] reports a runtime of 0.248 s on Oxford 5 K+Flickr 1 M dataset, so the efficiency of visual phraselet based approach is also demonstrated.

IV. CONCLUSIONS

In this letter, a novel visual concept, the *Visual Phraselet*, is proposed for large scale image search. We posit a global geometric transformation between the query and database images. Spatially consistent visual phrases of length-2 are extracted from the maximum response bin and its eight neighbors in the offset space, and then are assembled into a visual phraselet. By filtering out outlier visual phrases in an unsupervised manner, more strict geometric constraint is encoded in the searching step. Experimental results on three annotated datasets demonstrate that phraselet-based method outperforms the other approaches. The improvement is more significant when soft assignment is applied. We also show that visual phraselet is suitable for tasks where the codebook is trained on irrelevant data. Furthermore, large scale experiments confirm the scalability and efficiency of our method and the mAP improvement can be as large as 54.6% on the 1 M dataset. Our future work involves incorporating scale and translation invariance into the visual phraselet framework.

REFERENCES

[1] H. Jégou, M. Douze, and C. Schmid, “On the burstiness of visual elements,” in *CVPR*, 2009.

[2] J. Philbin, O. Chum, M. Isard, and A. Zisserman, “Object retrieval with large vocabularies and fast spatial matching,” in *CVPR*, 2007.

[3] S. Zhang, Q. Tian, G. Hua, Q. Huang, and S. Li, “Descriptive visual words and visual phrases for image applications,” in *ACM Multimedia*, 2009.

[4] H. Jégou, M. Douze, and C. Schmid, “Hamming embedding and weak geometric consistency for large scale image search,” in *Proc. Eur. Conf. Comput. Vision*, 2008, pp. 304–317.

[5] Y. Zhang, Z. Jia, and T. Chen, “Image retrieval with geometry-preserving visual phrases,” in *CVPR*, 2011.

[6] M. Perd’och, O. Chum, and J. Matas, “Efficient representation of local geometry for large scale object retrieval,” in *CVPR*, 2009.

[7] X. Wang, M. Yang, T. Cour, S. Zhu, K. Yu, and T. X. Han, “Contextual weighting for vocabulary tree based image retrieval,” in *ICCV*, 2011.

[8] W. Zhou, Y. Lu, H. Li, Y. Song, and Q. Tian, “Spatial coding for large scale partial-duplicate web image search,” in *ACM Multimedia*, 2009.

[9] S. Lazebnik, C. Schmid, and J. Ponce, “Beyond bags of features: Spatial pyramid matching for recognizing natual scene categories,” in *CVPR*, 2006.

[10] Y. Cao, C. Wang, Z. Li, L. Zhang, and L. Zhang, “Spatial bag-of-features,” in *CVPR*, 2010.

[11] S. Zhang, Q. Huang, G. Hua, S. Jiang, W. Gao, and Q. Tian, “Building contextual visual vocabulary for large-scale image applications,” in *ACM Multimedia*, 2010.

[12] Z. Wu, Q. Ke, M. Isard, and J. Sun, “Bundling features for large scale partial-duplicate web image search,” in *CVPR*, 2009.

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

[14] D. G. Lowe, “Distinctive image features from scale invariant keypoints,” in *IJCV*, 2004.

[15] 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.

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