Effective Object-based Image Retrieval Using Higher-level Visual Representation

Ismail El sayad, Jean Martinet, Thierry Urruty, Samir Amir, and Chabane Djeraba
LIFL/CNRS-UMR 8022
University of Lille 1 & Telecom Lille 1
Lille, France
Email: {ismail.elsayad, jean.martinet, thierry.urruty, samir.amir, chabane.djeraba}@lifl.fr

***Abstract*—Having effective methods to access the desired images is essential nowadays with the availability of huge amount of digital images. The proposed approach is based on an analogy between image retrieval containing desired objects (object-based image retrieval) and text retrieval. We propose a higher-level visual representation,for object-based image retrieval beyond visual appearances. The proposed visual representation improves the traditional part-based bag-of-words image representation, in two aspects. First, the approach strengthens the discrimination power of visual words by constructing an mid level descriptor, visual phrase, from frequently co-occurring and non noisy visual word-set in the same local context. Second, to bridge the visual appearance difference or to achieve better intra-class invariance power, the approach clusters visual words and phrases into visual sentence, based on their class probability distribution.**

***Index Terms*—Object-based Image Retrieval; Feature extraction; Bag of visual words; Visual phrases.**

I. Introduction

With the increasing convenience of capture devices and wide availability of large capacity storage devices, the amount of digital images that an ordinary people can reach has become so vast that effective and efficient ways are being called for to locate the desired images in the sea of images. This paper investigates an important branch of content-based image retrieval: object-based image retrieval (OBIR). The goal of OBIR is to find images containing desired object by providing retrieval system an example image or some example images of the desired object. In typical image retrieval systems, it is always important to select an appropriate representation for images.

Indeed, the quality of the retrieval depends on the quality of the internal representation for the content of the image. Bag-of-visual-words [1], [2], [3] has drawn much attention between other approaches in the *part-base image represnetaion*. Analogous to document representation in terms of words in text domain, the bag-of-visual-words approach models an image as an unordered bag of visual words. These visual word does not possess any semantics, as it is only a quantized vector of sampled local regions. However, if neglecting the semantic factor, what really distinguishes textual word from visual word is the discrimination and invariance power. Hence, in order to achieve better image retrieval performance, the low discrimination and invariance issues of visual words must be tackled.

Firstly, the *low discrimination power* of visual words leads to low correlations between the image features and their semantics. In our work, we build a higher-level representation, namely *visual phrase*from groups of adjacent words using *association rules* extracted with the *Apriori* algorithm [4]. Having a higher-level representation, from mining the occurrence of groups of low-level features (visual words), enhances the image representation with more discriminative power since structural information is added.

Secondly, the images of the same semantic class can have arbitrarily different visual appearances and shapes. Such visual diversity of object causes one image semantics to be represented by different visual words and phrases. This leads to *low invariance* of visual words and phrases. In this circumstances, the visual words and phrases become too primitive to effectively model the image semantics, as their efficacy depends highly on the visual similarity and regularity of images of the same semantics. To tackle this issue, a higher-level visual content unit(*visual sentence*) is needed which is in in upper level comparing to words and phrases.

The remainder of the article is structured as follows: Section II, we describe the method for constructing visual words from images , mining visual phrases from visual words, and clustering both of them(visual words and phrases)to obtain the higher representation level which is the visual sentence. In Section III, we present an image similarity method based on visual words and visual phrases. We report the experimental results in Section IV, and we give a conclusion to this article in Section V.

II. Multilayer Image Representation

In this section, we describe different components of the chain of processes in constructing the visual words, visual phrases and visual sentences. Figure 1 presents the different process starting from detecting interest and edge points till the image description of the image by visual words , phrases and sentences.

A. Visual Word Construction

We use the fast Hessian detector [5] to extract interest points. In addition, the canny edge detector [6] is used to detect edge points. From both sets of interest and edge points, we use a clustering algorithm to group these points into different

978-1-4244-8611-3/10/$26.00 ©2010 IEEE

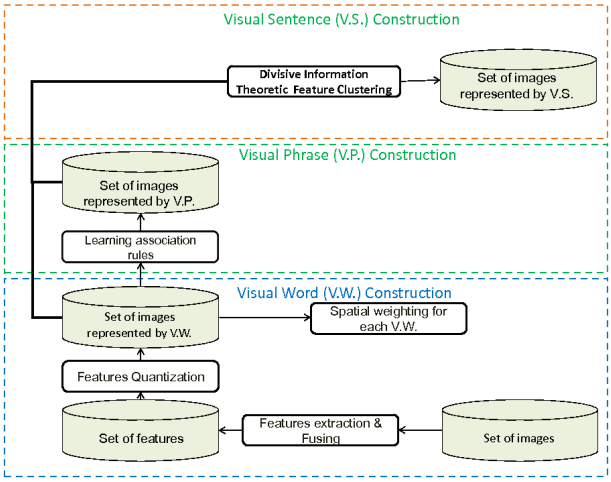

Fig. 1. Flow of information in the proposed image representation model.

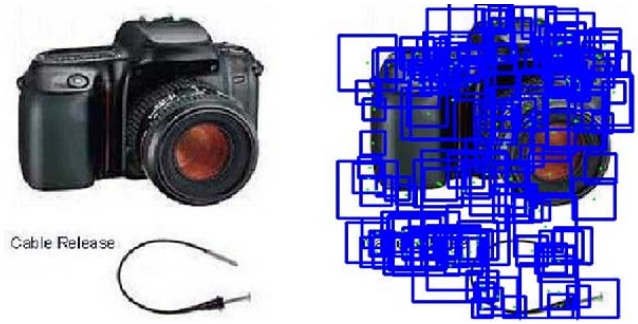

Fig. 2. Examples of images after Surf features extraction.

clusters in the 5-dimensional color-spatial feature space. The clustering result is necessary to extract the Edge context descriptor [7]and to estimate the spatial weighting scheme[7] for the visual words.

1) Extracting And Describing Local Features: In our approach, we use the SURF low-level feature descriptor which is 64 dimensional vector that describes the distribution of pixel's intensities within a scale-dependent neighborhood of each interest point detected by the Fast-Hessian. Figure 2 shows an example of an image after SURF features extraction. In addition to the *SURF* descriptor, we used another descriptor *(Edge context descriptor)* introduced by El sayad et al.[7]. This descriptor is inspired by the shape context descriptor proposed by Belongie et al. [8] with respect to the extracted information from edge point distribution. It describes the distribution of the edge points in the same Gaussian (by returning to the 5-dimensional color-spatial feature space). It is represented as a histogram of 6 bins for R (*magnitude* of the drawn vector from the interest point to the edge points) and 4 bins for θ (*orientation angle*).

Finally, the two descriptors are fused to form a feature vector composed of 88 dimensions (64 from SURF + 24 from the Edge context descriptor). Hence, the new feature vector describes information on the distribution of the intensity and the edge points of the image.

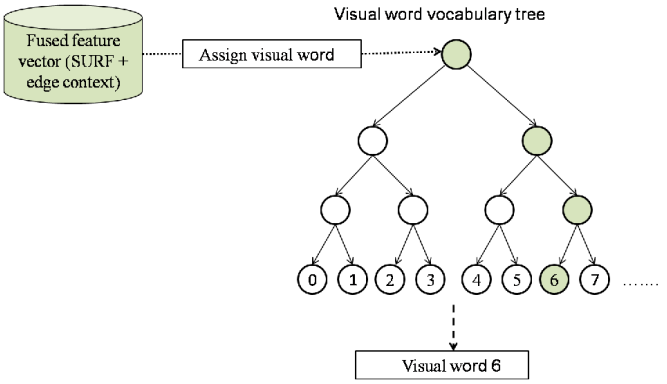

Fig. 3. Example of assigning a fused feature vector into a discrete visual word.

2) Quantizing the Local Features: Visual words are created by clustering the fused feature vectors (SURF + Edge context feature vector) in order to form a visual vocabulary. Quantization of the features into visual words is performed by using a vocabulary tree [9] in order to support large vocabulary size. The vocabulary tree is computed by repeated k-means clusterings that hierarchically partition the feature space.

This hierarchical approach overcomes two major problems of the traditional direct k-means clustering in cases where k is large. Firstly clustering is more efficient during visual word learning and secondly the mapping of visual features to discrete words is way faster than using a plain list of visual words. Finally,we map each feature vector of an image to its closest visual word. Therefore we query the vocabulary tree for each extracted feature, and the best matching visual word index is returned. Figure 4 shows an example of a fused feature vector assigned int a discrete visual word 6.

3) Filtering The Noisy Visual Words: In this section, we introduce another method to eliminate presumed useless visual words. This method aims at eliminating the most noisy words generated by the vocabulary building process, using *multilayer pLSA*. Lienhart et al.[10] proposed a multilayer multimodal probabilistic Latent Semantic Analysis (*mm-pLSA*). The proposed approach (*mm-pLSA*) has two modes: one mode for visual words and the other one for image tags. We used only the visual word mode. In the *multilayer pLSA* (*m-pLSA*), we have two different hidden topics.

- Top-level latent topics z_i^t.
- Visual latent topics z_j^v.

This generative model is expressed by the following probabilistic model:

$$P(I/w_l) = \sum_{i=1}^{P} \sum_{j=1}^{V} P(I)P(z_i^t/I)P(z_j^v/z_i^t)P(w_l/z_j^v) \quad (1)$$

where $P(I)$ denotes the probability of a an image I of the database to be picked, $P(z_i^t/I)$ the probability of a top-level topic z_i^t given the current image, $P(z_j^v/z_i^t)$ the probability of a

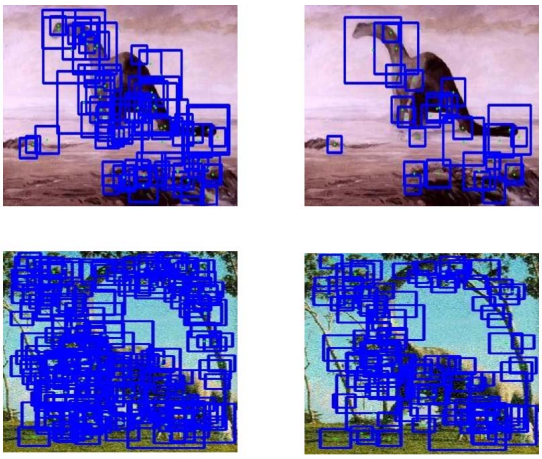

Fig. 4. Examples of images after filtering the noisy visual words using *m-pLSA*.

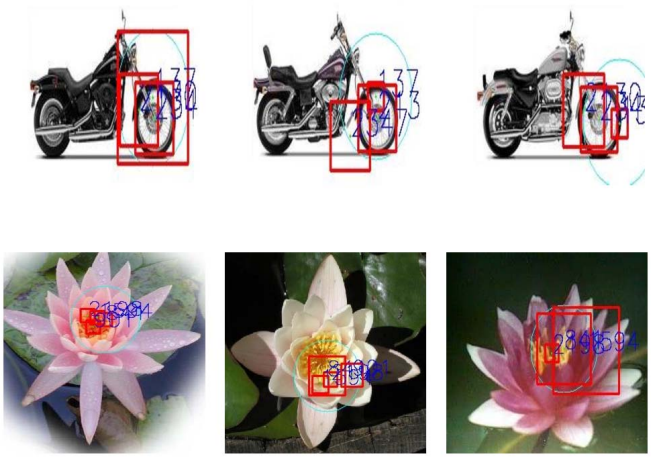

Fig. 5. Examples of visual phrases. The square resembles a local patch, which denotes one of the visual words, and the circle around the center of the patch denotes the local context .

visual latent topic z_j^v given a z_i^t and $P(w_l/z_j^v)$ the probability of a visual word w_l given a z_j^v.

We assigned one top-level latent topic per category of images, the total number of top-level latent topics (P) is the same as the total number of categories of the image dataset. The total number of visual concepts is V where $V < P$. We categorized visual concepts according to their joint probabilities with all top-level latent topics $P(z_j^v/z_i^t)$. All visual concepts whose joint probability to all top-level latent concepts are lower than a given threshold is categorized as irrelevant. After that, we eliminated all visual words whose probability $P(w_l/z_j^v)$ is low to a given threshold for every relevant visual concept, since they are not informative for any relevant visual concept. Therefore, we propose to keep only the most significant words for each relevant visual concept.

Figure 4 shows examples of images after eliminating ambiguous visual words. Experiments reported in Section 5 show that this technique improves the performance of the image retrieval. An important aspect of this model is that every image consists of one or more visual aspects, which in turn are combined to one or more higher-level aspects. This is very natural since images consist of multiple objects and belong to different categories.

B. Visual Phrase Construction

Before proceeding to the construction phase of visual phrases for the set of images, let us examine phrases in text. A phrase can be defined as a group of words functioning as a single unit in the syntax of a sentence and sharing a common meaning. For example, from the sentence "*James Gordon Brown is the Prime Minister of the United Kingdom and leader of the Labor Party*", we can extract a shorter phrase "*Prime Minister*". The meaning shared by these two words is the governmental career of James Gordon Brown.

Analogous to documents, which are particular arrangements of words in 1D space, images are particular arrangements of patches in 2D space. The inter-relationships among patches encode important information for our perception. Applying association rules, we used both the patches themselves and their inter-relationships to obtain a higher-level representation of the data known as visual phrase.

1) Association Rules: In the proposed approach, the visual phrase is constructed from group of non-noisy visual words that share strong association rules and are located within the same local context (see the green circles in Figure 5). Considering the set of all visual words (visual vocabulary) $W = \{w_1, w_2, ..., w_k\}$ which denotes the set of items, D is a database (set of images I), $T = \{t_1, t_2, ..., t_n\}$ is the set of all different sets of visual words located in the same context which denotes the set of transactions.

An association rule is a relation of an expression $X \Rightarrow Y$, where X and Y are sets of items(sets of one or more of frequent visual words that are within the same context). The properties that characterize association rules are:

- The rule $X \Rightarrow Y$ holds in the transaction set T with support s if s % of transaction in T contain X and Y.
- The rule $X \Rightarrow Y$ holds in the transaction set T with confidence c if c % of transactions in T that contain X also contain Y.

Given a set of images D, the problem of mining association rules is to discover all strong rules, which have a support and confidence greater than the pre-defined minimum support (*minsupport*) and minimum confidence (*minconfidence*).

C. Visual Sentence Construction

Studying the co-occurrence and spatial scatter information make the image representation more distinctive, the invariance power of visual words or phrases is still low. Returning to text documents, the synonymous words are usually clustered into one synonymy set to improve the document categorization performance [11]. Such approach inspires us to enhance the

invariance power of visual words and phrases by generating a higher-level visual representation, called visual sentence, by clustering the visual words and phrases based on their probability distributions to all relevant latent visual concepts .

The visual sentence is a semantic cluster of visual words and phrases, in which the member visual words and phrases might have different visual appearances but similar semantic inferences towards the latent visual concepts.

By defining visual sentence construction as a task of clustering based on visual concept probability distributions, the next step will be how to select an optimal distributional clustering framework. In this paper, we use an information-theoretic framework that was introduced by Dhillon et al. [12] that is similar to Information Bottleneck [13] to derive a global criterion that captures the optimality of distributional clustering. The main criterion is based on the generalized Jensen-Shannon divergence [14] among multiple probability distributions.

In order to find the best distributional clustering, i.e., the clustering that minimizes this objective function, Dhillon et al. introduced a new divisive algorithm for distributional clustering.They showed that their algorithm minimizes 'within-cluster divergence' and simultaneously maximizes 'between-cluster divergence'. This approach is markedly better than the agglomerative algorithms of Baker and McCallum [15] and the one introduced by Slonim and Tishby [16] .

Having $Z^v = \{z_1^v, z_2^v, ..., z_V^v\}$ as set relevant visual latent topics, $G = \{g_1, g_2, ..., g_M\}$ as a set of visual *glossary* (visual words and phrases), and $S = \{s_1, s_2, ..., s_N\}$ as a set of clusters (visual sentence). The joint distribution $P(G/Z^v)$ can be estimated from the training set as dicussed in section II-A3.

Dhillon et al. used an information-theoretic measure to judge the quality of the clusters. The information about Z^v captured by G can be measured by the mutual information $I(Z^v; G)$. The *best* clustering is the one that minimizes the decrease in mutual information, $I(Z^v; G) - I(Z^v; S)$, for a given number of clusters. The following theorem states that the change in mutual information can be expressed in terms of the generalized Jensen-Shannon divergence of each cluster.

$$I(Z^v;G) - I(Z^v;S) = \sum_{j=1}^{k} \pi(s_j) JS_{\pi'}(\{P(Z^v/s_j) : g_t \in s_j\}) \tag{2}$$

where $\pi(s_j) = \sum_{g_t \in s_j} \pi(g_t)$, $\pi(g_t) = P(g_t)$, $\pi_t^{'} = \pi_t / \pi(s_j)$ for $g_t \in s_j$, and JS denotes the generalized Jensen-Shannon divergence. Figure 6 shows examples of visual sentences. Each visual sentence has a semantic interruption, for example the fist image in the upper right of the Figure is a visual sentence that contains all the visual phrases and visual words that describe the windows of the airplane.

III. Image Representation, Indexing And Retrieval

Given the proposed image representation discussed in Section II, we describe here how images are represented, indexed and retrieved .

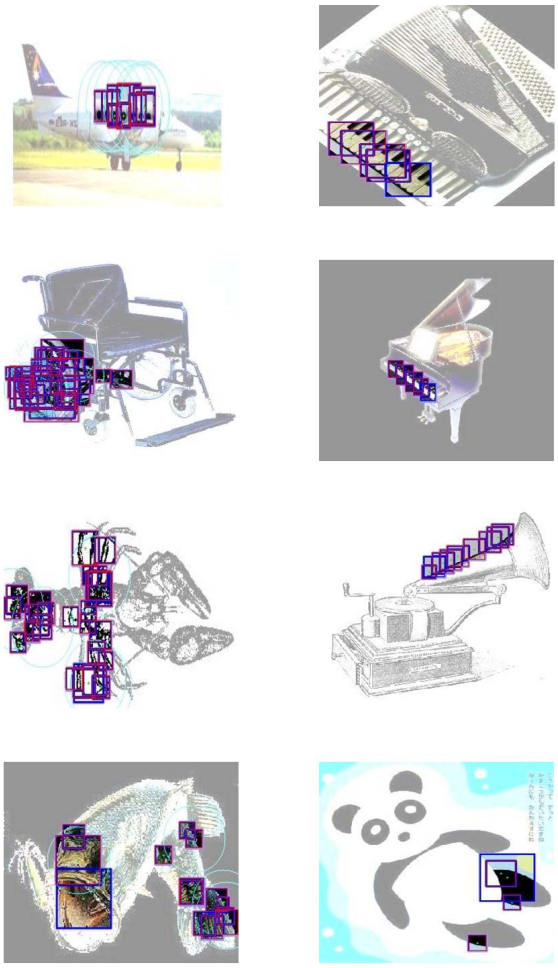

Fig. 6. The images on the right side are examples of visual sentences that are constructed from visual words and phrases while the images on the left sides are for visual sentences that are constructed from visual words only.

A. Image Representation

The traditional Vector Space Model [17] of Information Retrieval [17] is adapted to our representation, and used for similarity matching and retrieval of images. The following doublet represents each image in the model:

$$I = \begin{cases} \vec{W}_i \\ \vec{P}_i \\ \vec{S}_i \end{cases} \tag{3}$$

where $\vec{W}_i$,$\vec{P}_i$, and $\vec{S}_i$ are the vectors for the word, phrase and sentence representations of an Image respectively:

$$\vec{W}_i = (w_{1,i},\ ...,\ w_{n_w,i})\ ,\ \vec{P}_i = (p_{1,i},\ ...,\ p_{n_p,i})\ ,$$
$$\vec{S}_i = (s_{1,i},\ ...,\ p_{n_s,i}) \tag{4}$$

Note that the vectors for each level of representation lie in a separate space. In the above vectors, each component

represents the weight of the corresponding dimension. We used the ***spatial weight scheme*** that is introduced by El sayad et al. [7], for the words and the standard ***td.idf-weighting scheme*** for the phrases and the sentences. Thus, we map images into documents and we apply document retrieval techniques to image retrieval.

B. Image Indexing

In our approach, we use an inverted file [18] to index images. The inverted index consists of two components: one includes indexed visual words, visual phrases and visual sentences. The other includes vectors containing the information about the spatial weighting of the visual words and the occurrence of the visual phrases.

C. Similarity Measure And Retrieval

After represented the query image as a doublet of visual words, phrases and sentences, we consult the inverted index to find candidate images. All candidate images are ranked according to their similarities to the query image. We have designed a simple measure that allows evaluating the contribution of words and phrases. The similarity measure between a query I_q and a candidate Image I_c is estimated with:

$$sim(I_q, I_c) = (\alpha)RSV(\vec{W}_c, \vec{W}_q) + (\beta)RSV(\vec{S}_c, \vec{S}_q) + (\gamma)RSV(\vec{P}_c, \vec{P}_q) \quad (5)$$

The Retrieval Status Value (*RSV*) of the three vectors is estimated with the cosine distance. The non-negative parameter α is to be set according the experiment runs in order to evaluate the contribution between visual words, visual phrases and visual sentence.

IV. Experiments

This section describes the set of experiments we have performed to explore the performance of the proposed methodology. Firstly, we investigate the performance of the proposed approach and the average number of visual words on each class of images after filtering the noisy visual words. Secondly, we evaluate the effect of fusion Edge context descriptor with SURF

A. Dataset And Experimental Setup

The image dataset used for these experiments is the Caltech101 Dataset1 [19]. It contains 8707 images, which includes objects belonging to 101 classes. The number of images in each class varies from about 40 to about 800 with an average of 50 images. For the various experiments, we construct the test data set by selecting randomly 10 images from each class (1010 images). The query images are picked from this test data set during the experiment. The visual word vocabulary size (*K*)=3000 and the visual phrase vocabulary size is 960 and for the visual sentence vocabulary size, we investigate it for different values . We start our experiment with $\alpha = 1$,$\beta = 0$, and $\gamma = 0$ to check the performance of the visual words. Later we investigate the performance of the system with different values for α, β, and γ.

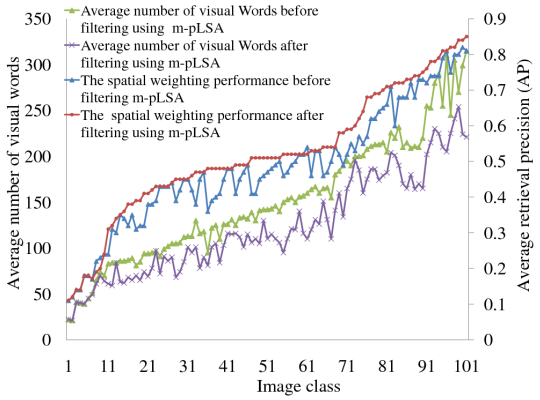

Fig. 7. Evaluation of the performance of the proposed approach and the average number of visual words on each class of images after filtering the noisy visual words.

B. Assessment of the Visual Glossary Performance

1) Evaluation Of The Performance Of the Proposed bag-of-visual-words and the Average Number of Visual Words After Filtering the Noisy Visual Words: In this section we show the influence of filtering noisy visual words based on the m-pLSA and we contribute on the relation between the average number of visual words in each class and the corresponding retrieval performance. Figure 7 plots the average retrieval precision (*AP*) for our spatial weighing approach before and after filtering. In addition, it plots the corresponding average number of visual words for each class. For a clearer presentation, we arrange the 101 classes from left to right in the figure with respect to the ascending order of their average precision of each class after filtering.

On one hand, it is obvious from the results displayed that the performance has slightly improved after filtering especially in the classes that have huge amounts of words comparing to others that have small amounts. On the other hand, there is a variation of retrieval performance among all 101 classes, and this variation is related to the average number of the visual words. Figure 7 shows a clear difference in the average number of visual words between the classes that highly perform and the classes that have poor performance.

The number of visual words on an image depends on the interest point detector, as we mentioned before that we use the Fast Hessian detector which is faster comparing to others. The computational time for detecting the interest points is reduced by using image convolutions based on integral images. Having these convolutions decreases the number of detected interest points and this contributes as a limitation for Fast Hessian in images with rare texture.

2) Evaluation of the perofmance of the visual glossary: We combine visual phrase and visual word representations by varying the parameter α and β used in the similarity matching approach. Figure 8 plots the *MAP* for different values of α and β over all 101 classes. When considering only visual phrases in the similarity matching $(\alpha = 0;, \beta = 1, and \gamma = 0)$, the *MAP* is slightly better than the scenario in which only visual

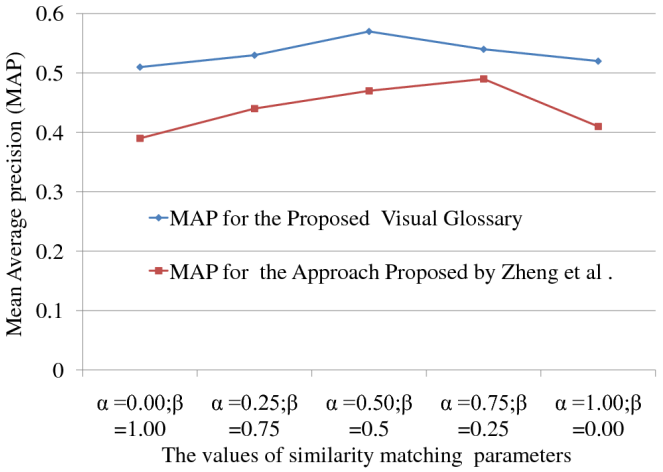

Fig. 8. Contribution of visual words and visual phrases in our approach and in Zheng et al. approach.

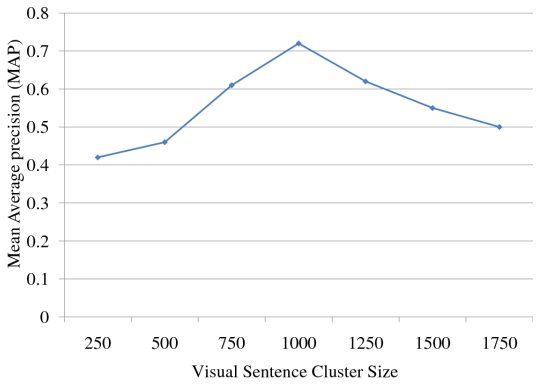

Fig. 9. visual sentence performance

words are used $(\alpha = 1;, \beta = 0, and \gamma = 0)$. However, the combination of both yields better results than using words or phrases separately.

C. Evaluation of the performance of the Higher-Level Visual Representation (Visual Sentence)

We evaluate the effectiveness of visual sentence by performing the distributional clustering. We set the size of visual sentence vocabulary to different values and we set the similarity parameters for the following values: $\alpha = 0, \beta = 0, and \gamma = 1$. Figure 9 displays the Mean Average Precision (MAP) of image retrievals based on different number of visual sentence. We observe that with proper cardinality, the visual sentence representation can deliver superior results over both visual words and visual phrases with a more compact representation. For example, the run with only 750 visual sentence can achieve a MAP of 0.62, which is superior to the maximum run for the visual glossary. This representation compactness does not only enable high computational efficiency but also alleviates the curse of dimensionality.

V. Conclusion

In order to retrieve images beyond their visual appearances, we proposed a higher level image feature, visual sentence , for object-based image retrieval. First, we exploit the spatial co-occurrence information of visual words to generate a more distinctive visual configuration, i.e. visual phrase. This improves the discrimination power of visual word representation with better interclass distance. Second, we proposed to group the visual words and phrases with similar semantic into a visual sentence. Rather than in a conceptual manner, the semantics of a visual phrase is probabilistically defined as its image class probability distraction.

In our future work, Several open issues remain. First, the generation of visual phrase is a time-consuming task. A more efficient algorithm is demanded. Second, the questions as how the number of classes changes the semantic inference distribution of visual lexicons and how this affects the visual sentence generation and final classification, have not been investigated.

References

[1] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *ICCV*. IEEE Computer Society, 2003, pp. 1470–1477.

[2] J. Willamowski, D. Arregui, G. Csurka, C. R. Dance, and L. Fan, "Categorizing nine visual classes using local appearance descriptors," in *In ICPR Workshop on Learning for Adaptable Visual Systems*, 2004. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.70.9926

[3] F. Jurie and B. Triggs, "Creating efficient codebooks for visual recognition," in *ICCV*, 2005, pp. 604–610.

[4] R. Agrawal, T. Imielinski, and A. N. Swami, "Mining association rules between sets of items in large databases," in *SIGMOD Conference*, P. Buneman and S. Jajodia, Eds. ACM Press, 1993, pp. 207–216.

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

[6] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 8, no. 6, pp. 679–698, November 1986. [Online]. Available: http://portal.acm.org/citation.cfm?id=11275

[7] I. Elsayad, J. Martinet, T. Urruty, and C. Djeraba, "A new spatialweighting scheme for bag-of-visual-words," in *IEEE International Workshop on Content-Based Multimedia Indexing (CBMI)*, 2010.

[8] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 4, pp. 509–522, 2002.

[9] D. Nistér and H. Stewénius, "Scalable recognition with a vocabulary tree," in *CVPR (2)*, 2006, pp. 2161–2168.

[10] R. Lienhart, S. Romberg, and E. Hörster, "Multilayer plsa for multimodal image retrieval," in *CIVR*, 2009.

[11] R. Bekkerman, R. El-Yaniv, N. Tishby, and Y. Winter, "Distributional word clusters vs. words for text categorization," *J. Mach. Learn. Res.*, vol. 3, pp. 1183–1208, 2003.

[12] I. S. Dhillon, S. Mallela, and R. Kumar, "A divisive information-theoretic feature clustering algorithm for text classification," *Journal of Machine Learning Research*, vol. 3, pp. 1265–1287, 2003.

[13] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[14] J. Lin, "Divergence measures based on the shannon entropy," *IEEE Transactions on Information Theory*, vol. 37, no. 1, pp. 145–, 1991.

[15] L. D. Baker and A. McCallum, "Distributional clustering of words for text classification," in *SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 24-28 1998, Melbourne, Australia*. ACM, 1998, pp. 96–103.

[16] N. Slonim and N. Tishby, "The power of word clusters for text classification," in *In 23rd European Colloquium on Information Retrieval Research*, 2001.

[17] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Commun. ACM*, vol. 18, no. 11, pp. 613–620, 1975.

[18] I. H. Witten, A. Moffat, and T. C. Bell, *Managing Gigabytes: Compressing and Indexing Documents and Images, Second Edition*. Morgan Kaufmann, 1999.

[19] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories," *Comput. Vis. Image Underst.*, vol. 106, no. 1, pp. 59–70, 2007.