Improving retrieval framework using information gain models

Huu Ton Le, Thierry Urruty, Syntyche Gbehounou, François Lecellier, Jean Martinet, Christine Fernandez-Maloigne

Signal, Image and Video Processing manuscript No.
(will be inserted by the editor)

Improving Retrieval Framework using Information Gain Models

Huu Ton Le · Thierry Urruty · Syntyche Gbèhounou · François Lecellier · Jean Martinet · Christine Fernandez-Maloigne

Received: date / Accepted: date

Abstract Content Based Image Retrieval (CBIR) systems are meant to retrieve the most similar images of a collection to a query image. One of the most well-known models and widely applied for this task is the Bag of Visual Words model (BoVW). In this paper, we introduce a study of different information gain models used for the construction of a visual vocabulary. In the proposed framework, information gain models are used as a discriminative information to index image features and select the ones that have the highest information gain values. We introduce some extensions to further improve the performance of the proposed framework: mixing different vocabularies, and extending the BoVW to Bag of Visual Phrases (BoVP). Exhaustive experiments show the interest of Information Gain models on our retrieval framework.

Keywords Image Retrieval, Vocabulary Construction, Bags of Visual Words, Saliency Map

Huu Ton Le
ICTLab Research Laboratory, University of Science and Technology of Hanoi, Vietnam
E-mail: le-huu.ton@usth.edu.vn

Thierry Urruty, Syntyche Gbèhounou, François Lecellier, and Christine Fernandez-Maloigne
XLIM, UMR CNRS 7252, University of Poitiers, France
E-mail: firstname.lastname@xlim.fr

Jean Martinet
Univ. Lille, CNRS, Centrale Lille, UMR 9189 - CRIStAL - Centre de Recherche en Informatique Signal et Automatique de Lille, IRCICA, F-59000 Lille, France
E-mail: jean.martinet@univ-lille1.fr

1 Introduction

In the last decade, along with the development of new technologies, it has become really easy to create and share pictures online. This phenomenon leads to an exponential growth of image databases implying a need for new methodologies to manage such very large number of images, not only effectively but also efficiently.

One of the most popular methods for CBIR is the Bag of Visual Words model [4,25]. In this model, images are represented by histograms of visual words and those histograms are used to index all images in the database. Much research was done to enhance the performance of BoVW model [12,19]. In [12], authors added some spatial information into BoVW model by dividing the image into a spatial pyramid (sub-regions), and then a BoVW framework is applied to calculate the histogram of local features in each sub-region. Those histograms are then combined together to form the image representation. The method proposed by Pedrosa et al. [19] finds the co-occurrences of local features in the images to make visual phrases which are claimed to be more discriminative than visual words. Histograms of visual phrases are then used by authors for indexing images.

In [27], we proposed a new method to construct incrementally a visual vocabulary based on Information Gain (IG) that combines a saliency map and an IG model, *tf-idf*. The saliency map of an image is generated with a visual attention model. These bio-inspired models [9,16] are based on psycho-visual features while others also make use of several low-level features [7]. In this paper, we extend our research [14] and evaluate the effects of different IG models using the iterative visual words selection algorithm. We present a detailed study of four different IG models: *tf-idf* [24], *tfc* [24], *bm25* [23] and *entropy* [15]. We exploit those models

to iteratively filter out candidate words with low IG value, in order to retain only the *best* words in the final vocabulary. Exhaustive experiments on well-known datasets with several descriptors are implemented to highlight the difference between models. Some extension methods that further improve the performance of the BoVW model are also introduced: mixing the vocabularies obtained by different IG models as well as extending the BoVW to Bag of Visual Phrases (BoVP) to see the effect of new vocabulary selection scheme on the performance of BoVP.

The remainder of this article is structured as follows: we provide a brief overview of related works in Section 2. We present our global framework and the information gain evaluation in Section 3 and detail the related experiments in Section 4. Further improvement mechanisms are introduced in Section 5, including a discussion on the findings of our study. Section 6 concludes and opens some perspectives.

2 State of the art

The Bag of Visual Words (BoVW) model proposed by Csurka et al. [4] was inspired by the Bag of Words model [8] of the Information Retrieval domain. The BoVW model is composed of three main steps: feature detection, feature extraction and vocabulary construction. The purpose of the BoVW model is to represent images by histograms of local features, i.e. visual words, which defines the visual vocabulary. The feature detection step selects a set of interest points which contain rich local information about the image and uses the local patches around these points to describe an image. Feature extraction converts each local patch around all interest points of the image into a numerical vector. Finally, the vocabulary construction defines a set of visual words as a base, i.e. the visual vocabulary. Each local feature in the image is then assigned to one visual word presented in the visual vocabulary. Finally, this frequency histogram of visual words defines the image signature, which is used for indexing all the images in a database.

Improving the BoVW model has been an active research topic recently, and lots of methodologies have been introduced to enhance the performance of this model, as detailed below. For example, Fisher Kernel [20] or Vector of Locally Aggregated Descriptors (VLAD) [11] are efficient models introduced to enhance the BoVW model. The first approach has been used by Perronnin and Dance [20] on visual vocabularies for image categorisation. They proposed to apply Fisher Kernels to visual vocabularies represented by means of a Gaussian Mixture Model (GMM). In comparison to the BoVW representation, fewer visual words are required by this more sophisticated representation. VLAD is introduced by Jégou et al. [11] and can be seen as a simplification of the Fisher kernel. Considering a codebook C $= c_1, \ldots, c_k$ of k visual words generated with *k-means*, each local descriptor x is associated to its nearest visual word $c_i = 1 - NN(x)$. The idea of VLAD is to accumulate, for each visual word c_i, the differences $x - c_i$ for all vectors x assigned to c_i.

Ren et al. [22] extend the BoVW model into Bag of Bags of Visual Words. First, images are represented as a connected graph by segmenting the images into partitions using the Normalized Cut methodology. Then the classical BoVW model is applied to each individual sub-graph and each sub-graph has its own histogram of visual words. The signature of the image is the concatenation of histograms of every sub-graph. By using several resolutions, which define the number of sub-graphs per image, they also use an approach named Irregular Pyramid Matching (IPM) for image representation instead of classical spatial pyramid matching [13]. Alqasrawi et al. [2] have also used the BoVW model and colour information with a spatial pyramid representation to obtain good performance for nature image classification. For a similar application, Yeganli et al. [30] have proposed an approach mixing several dictionaries learned from multiple image resolution patches.

More recent researches show the needs to improve the BoVW model. First, the Extended Bag of Features (EBoF), which has been introduced by Tsai et al. [26]. In comparison to classical BoVW, EBoF is found to be more robust to rotation, translation and scale variations thanks to the proposed circular-correlation-based algorithm. EBoF model divides an image into fan-shaped sub-images and the BoVW model is applied to each of them. A 2D Gaussian weighting scheme is then applied to remove the contribution of visual words that are located far from the center. The histograms of all sub-images are then combined together to build the image representation. Abolghasemi et al. [1] have proposed to construct a dictionary by using an incoherent K-SVD approach, which speeds up the learning stage and has shown promising results on medical images. Testing this interesting way to construct a dictionary is part of our perspectives.

We have selected 4 information gain models in this study (as detailed in the next section), that are part of the most widely-used models. The IG models serve for generating several vocabularies, that are used in a CBIR system for an image search task, with the central objective to study the effects of IG models on the quality of the vocabulary.

3 Evaluation of information gain models

Global Framework

The selection of image features to build the visual words vocabulary plays an important role in the BoVW model; a good vocabulary leads to a good image representation and thus to a good retrieval accuracy. However, it is difficult to define what a good vocabulary is. The original BoVW model employs a clustering algorithm such as *k-means* to group similar features in the same cluster and uses the centroids of each group as visual vocabulary. However, Parsons et al. [18] show that such clustering algorithms do not have great performance in a high dimensional feature space which is the case of many visual image descriptors. Thus, clustering algorithms may not be the best choice to construct a visual vocabulary.

In our previous works [27], we introduced a new methodology to construct the visual vocabulary that uses an information gain model based on *tf-idf* (term frequency - inverse document frequency [24]), combined to a saliency information. Instead of using a clustering algorithm, the proposed method randomly selects visual words among all features, and then iteratively filters out words according to an IG criterion:

1. In the first step, a subset of features are randomly selected from a large and heterogeneous set (typically the whole set of features from all images in a dataset) to form the initial vocabulary of visual words. All remaining features are assigned to their closest visual word.
2. The second step iteratively identifies visual words that have the highest IG values in the vocabulary: visual words with low IG value are then discarded, and "orphan" descriptors (that were previously assigned to discarded words) are re-assigned to the closest remaining words.

The initial vocabulary is purposely large, and the process iterates until the desired vocabulary size is reached. Figure 1 shows the steps of the proposed approach. The information gain formulation IG for this work combines two sources: the *tf-idf* weighting scheme [24] and Itti's saliency maps [9]:

$$IG_w = \frac{n_{wD}}{n_D} \log \frac{N}{n_w} + \frac{\sum Sal_{wD}}{n_{wD}}, \quad (1)$$

where IG_w is the IG value of the visual word w, n_{wD} is the frequency (i.e. the number of occurrences) of w in the dataset D, n_D is the total number of descriptors in the dataset, N is the number of images in the dataset, n_w is the number of images containing the word w, and $\sum Sal_{wD}$ is the accumulated saliency values from all the keypoints assigned to word w. Note that the traditional *tf* for text is defined document-wise, and the formulation $\frac{n_{wD}}{n_D}$ in Eq. 1 is defined collection-wise. During each iteration, an amount of words (defined by a fixed ratio of the current vocabulary size) with lowest IG_w value are discarded. This formulation means that the only words to be kept in the vocabulary have:

- either a high *tf-idf* value, i.e. large number of occurrences in the collection, only in a limited number of images,
- or a high saliency score, i.e. high saliency values for keypoints attached to the word),
- or both a high *tf-idf* value and a high saliency score.

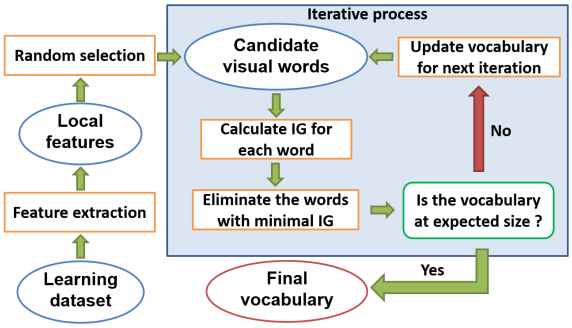

Fig. 1 Using Information Gain for Vocabulary Construction

Experimental results in [27] have shown a precision increase of around 7% (regardless of the descriptor) for the proposed framework compared to the original BoVW model and other state-of-the-art methods. It highlights the interest of using IG in the selection of the visual vocabulary. Since the initial set of image features is selected randomly, the stability of the proposed method may be questioned. Experiments described in [27] with University of Kentucky Benchmark [17] and PASCALVOC2012 [6] have proved little variations regardless of the descriptor and dataset.

Another interesting results is the fact that there may exist an optimal vocabulary size (number of visual words). However, this optimal size changes with respect to the selected dataset and the used descriptor. For example, with CMI on UKB, the optimal number is around 300 and with OpponentSift on Holidays, it is around 1000.

Information gain models

We propose to evaluate the effect of different IG models on the construction of visual vocabulary. Other than *tf-idf* models, three other IG models are taken into consideration. We implement the same framework as introduced in [27], however we use neither the saliency model nor the stabilization process in order to insure a fair comparison between the IG. We use the 4 following

IG formulations for our study [24]: *tf-idf*, *Okapi bm25*, *entropy*, and *tfc*. Other IG models exist [3], we based our study on the mostly used in the Information Retrieval field.

tf-idf stands for term frequency, inverse document frequency. It weights a term based on its term frequency (tf), the more frequently a term appears in a document, the more important it is for the document (notion of informativeness) and on its inverse document frequency (idf): the more frequently a term appears in the collection of documents, the less important it is for the collection (notion of discriminance). *Okapi bm25* is a function that measures the similarity between 2 documents based on the query term appearing in each document. *tfc* stands for term frequency component. tfc includes the differences in documents' length, it can be considered as the normalised version of *tf-idf*. *Entropy* weighting is calculated based on the distribution of a term in a single document as well as in the whole collection.

The IG models are implemented and used in our framework to estimate the IG value for each visual word w, in the same way as in Eq. 1:

$$tf\text{-}idf_w = \sum_{j=0}^{N-1} \frac{n_{wj}}{C_j}.log\frac{N}{n_w},$$

$$bm25_w = \sum_{j=0}^{N-1} \frac{n_{wj}(k_1+1)}{n_{wj}+k_1.(1-b+b.\frac{C_j}{avgdl})}.\log\frac{N-n_w+0.5}{n_w+0.5},$$

$$entropy_w = -\sum_{j=0}^{N-1} n_{wj}.\log(n_{wj}),$$

$$tfc_w = \sum_{j=0}^{N-1} \frac{\frac{n_{wj}}{C_j}.log\frac{N}{n_w}}{\sqrt{\sum_{k=0}^{D-1}(\frac{n_{kj}}{C_j}.log\frac{N}{n_k})^2}},$$

where N is the total number of images in the whole dataset, n_{wj} is the number of occurrences of visual word w in image j, C_j is the total number of visual words in image j and n_w is the number of images that contain w, D is the number of visual word that is used to represent image j, k_1 and b are 2 constant parameters. In our experiments, we have set $k_1 = 1.2$ and $b = 0.75$, as suggested in [23].

4 Experimental results

Three image datasets were considered:

University of Kentucky Benchmark which is proposed by Nistér and Stewénius [17]. This dataset is referred as UKB to simplify the reading. UKB contains of 10200 images divided into 2550 groups, each group consists of 4 images of the same object with different conditions (rotated, blurred...).

INRIA Holidays [10]: this dataset, referred as Holidays, is a collection of 1491 images, 500 of them are query images, and the remaining 991 images are the corresponding relevant images. The evaluation on Holidays is based on mean average precision score (mAP) [21].

PASCAL Visual Object Classes challenge 2012 [6] called PASCAL VOC2012, also referred as PASCAL. There are 17225 images in that database, 11530 are classified. Images in PASCAL VOC2012 are categorised into 20 object classes. 11530 classified will be the input for the retrieval test, the 100 nearest neighbour images to a query image are retrieved, object classes are the ground-truth.

Feature extraction: we use the feature extraction tool provided by Van de Sande et al. [28] to extract 4 widely-used descriptors: SIFT, Color Moment (CM), Color Moment Invariant (CMI) and OppSIFT. Note that we apply no keypoint filtering or selection in order to guarantee a common evaluation framework for all methods. Even though the results could be easily improved for some methods by using filtering and selection, we wish to ensure a fair comparison.

Vocabulary construction: the vocabularies are built on PASCAL VOC2012 and used for testing on PASCAL itself and other 2 datasets: UKB and Holidays. For each descriptor, we randomly select a set of random features from PASCAL database, and use those features as the initial visual vocabulary. For a fair evaluation, each time a random visual vocabulary is created, it will be used for all IG models: *tf-idf*, *tfc*, *bm25* and *entropy*. In each iteration, we discard 10% of features that have smallest IG values to obtain a new vocabulary and use that vocabulary as the input of the next iteration. Thus, we have the same initial input for all 4 IG models, however the final visual vocabularies used for retrieval are different.

Information gain BoVW vs Original BoVW

The global comparison between visual words selection using one of the four IG models and the classic *k-means* from the original BoVW model (denoted *Baseline*) is presented in Table 1. All experiments have been run using a vocabulary of 256 visual words; the results in bold highlight the experiments that obtain a better retrieval result than the classical BoVW method using *k-means*. The values between parenthesis indicate differences in percentage between both results.

It is noticeable that except *bm25*, using the vocabulary obtained by IG selection gives us better results than the classical *k-means* algorithm. The bottom row

Table 1 k-means vs Information Gain for BoVW

Descriptor	Database	Baseline	tf-idf	bm25	tfc	entropy
CM	UKB	2.68	**2.74(2.2%)**	2.58 (-3.7%)	**2.73(1.9%)**	2.68(0%)
	PASCAL	25.29	25.07(-0.9%)	**25.58(1.2%)**	25.1(-0.8%)	**25.44(0.6%)**
	Holidays	0.5	**0.516(3.2%)**	0.468(-6.4%)	**0.522(4.4%)**	**0.514(2.8%)**
SIFT	UKB	2.17	**2.19(0.9%)**	2.11(-2.8%)	**2.21(1.8%)**	**2.26(4.1%)**
	PASCAL	35.64	35.58(-0.2%)	34.54(-3.1%)	**35.69(0.2%)**	35.52(-0.3%)
	Holidays	0.316	**0.347(9.8%)**	**0.324(2.5%)**	**0.349(10.4%)**	**0.347(9.8%)**
CMI	UKB	2.96	**3.04(2.7%)**	**2.98(0.7%)**	**3.06(3.4%)**	**3.03(2.4%)**
	PASCAL	28.52	28.18(-1.2%)	27.8(-2.5%)	28.27 (-0.9%)	**29.18(2.3%)**
	Holidays	0.506	**0.508(0.4%)**	0.49(-3.2%)	**0.517(2.2%)**	**0.524(3.6%)**
OppSIFT	UKB	2.33	**2.39(2.6%)**	**2.34(0.4%)**	**2.41(3.4%)**	**2.4(3%)**
	PASCAL	35.04	34.8(-0.7%)	34.5(-1.5%)	34.62(-1.2%)	34.7(-1.0%)
	Holidays	0.483	**0.486(0.7%)**	0.47(-2.6%)	0.483(0%)	0.483(0%)
Avg diff			1.63%	-1.76%	1.98%	2.28%

Table 2 Mean differences with respect to the datasets

Database	tf-idf	bm25	tfc	entropy
UKB	2.1%	-1.35%	2.63%	2.38%
PASCAL	-0.75%	-1.5%	-0.68%	0.4%
Holidays	3.53%	-2.43%	4.23%	4.05%

of the table shows the average difference in score of each IG model against *k-means* algorithm. We can see that the results increase by 1.63 % with *tf-idf*, 1.98 % with *tfc* and 2.28 % with *entropy*.

Table 2 highlights another behavior: the IG model has different effects with respect to the datasets. This table presents the average difference score of each IG model with respect to each dataset. We can see that IG models have a great performance over UKB and Holidays databases, e.g. it increases the average results by 4.23% on Holidays and 2.63% on UKB with *tfc* model. However, using IG models does not seem to work very well on PASCAL, the score differences are very close to the classical BoVW. This comes from the fact that PASCAL dataset main objective is classification. Image categories given in PASCAL are more cluttered and therefore less adapted to our retrieval context.

5 Framework extensions

Previous experiments have demonstrated the importance of using IG to construct vocabularies.

This observation leads us to the importance of mixing the results to construct a visual vocabulary that accurately reflects the data. Thus, in this section, we introduce 2 extensions to improve the performance of our framework: (1) combination of several vocabularies to keep only the "best" words – in addition, we compare the results of *late* versus *early* combinations, i.e. *after* versus *during* the construction step, and (2) we illustrate how the proposed framework can be applied to the Bag of Visual Phrases model [5] in order to achieve even more discriminative representations.

Combination of vocabularies

We propose to mix the vocabularies obtained by iterative visual word selection based on all different IG models, either with a *late* or with an *early* combination strategy. Besides, for each strategy, two methods are compared to select the best visual words:

Round Robin: best words are selected alternatively from N_i lists of sorted words (N_i is the number of IG models) to form a unique sorted list with no duplicate. *Average rank*: an average IG-based rank across vocabularies is calculated for each word w using:

$$r(w) = \frac{1}{N_i} \sum_{i=1}^{N_i} rank_i(w), \tag{2}$$

where $rank_i(w)$ is the rank of w given by the IG model i, and N_i is the number of IG models. Note that we consider only words that belong to all vocabularies, the others disregarded. The final vocabulary is made up of the lowest-rank words.

Late combination: starting from a set of initial candidate visual words, we employ different IG models separately in our iterative process to obtain 4 different vocabularies containing 256 visual words. Then round robin or average rank method is applied to build the final vocabulary of the same size. Since the performance of *bm25* model is low (because the vocabulary size is small), we also run a test without including this IG model.

Table 3 presents all the experiment results of mixing vocabularies after the iterative step. The baseline is still the classical BoVW model: results in bold indicate retrieval results higher than the baseline. The column marked as rrAll shows the results of round robin method using all 4 IG models, column marked as rrBest shows the result of round robin method with the 3 best IG models, i.e. all but *bm25*. Similarly, column rankAll

Table 3 Late combination scores

Descriptor	Database	Baseline	rrAll	rrBest	rankAll	rankBest
CM	UKB	2.68	**2.73(1.7%)**	**2.72(1.5%)**	**2.72(1.5%)**	**2.72(1.5%)**
	PASCAL	25.29	25.21(-0.3%)	**25.36(0.3%)**	25.1(-0.8%)	25.18(-0.4%)
	Holidays	0.5	**0.518(3.5%)**	**0.529(5.7%)**	**0.51(2%)**	**0.521(4.2%)**
SIFT	UKB	2.17	**2.24(3.1%)**	**2.27(4.7%)**	**2.22(2.5%)**	**2.23(2.9%)**
	PASCAL	35.64	35.24(-1.1%)	35.56(-0.2%)	35.19(-1.3%)	35.45(-0.5%)
	Holidays	0.316	**0.34(7.6%)**	**0.345(9%)**	**0.334(5.8%)**	**0.344(8.8%)**
CMI	UKB	2.96	**3.06(3.2%)**	**3.07(3.7%)**	**3.06(3.4%)**	**3.06(3.4%)**
	PASCAL	28.52	**28.65(0.5%)**	**28.7(0.6%)**	**28.85 (1.2%)**	28.51(0%)
	Holidays	0.506	**0.519(2.6%)**	**0.522(3.2%)**	**0.52(2.7%)**	**0.516(2.0%)**
OppSIFT	UKB	2.33	**2.41(3.3%)**	**2.41(3.3%)**	**2.4(2.8%)**	**2.41(3.3%)**
	PASCAL	35.04	34.39(-1.9%)	34.87(-0.5%)	34.46(-1.7%)	34.61(-1.2%)
	Holidays	0.483	0.476(-1.4%)	0.479(-0.7%)	0.479(-0.7%)	**0.485(0.4%)**
Mean Difference			1.74%	2.56%	1.45%	2.04%

shows the result of average rank method with all 4 vocabularies and column rankBest with only 3 vocabularies. The bottom row shows the average difference for all descriptors and datasets.

As expected, eliminating the *bm25*-based vocabulary gives better average results. Mixing the vocabulary from *tf-idf*, *entropy* and *tfc* almost yields best retrieval performance regardless of the setup. rrBest is also the best mixing method as it increases 2.56% the retrieval results compared to the baseline (classical *k-means* BoVW model), in comparison with 1.74% for rrAll, 1.45 % for rankAll and 2.04% for rankBest.

Early combination: One problem with the late combination is that it requires 4 times as much computation time since each IG model requires an individual implementation. We overcome this problem by mixing all the IG inside the iterative step. We start from the initial set of 4096 visual words. At each loop of the iterative steps, we employ all IG models to create N_i sorted lists of words. Round robin or average rank method is then applied to create the global sorted list. We then discard the last 10% of visual words from this list to build the new candidate vocabulary, that serves as an input for the next iteration. This procedure is iterated until the final vocabulary size (256) is reached. This way, only one iterative step is needed instead of 4 as previously.

In all next experiments, *bm25* model is taken out of consideration due to poor performance. Table 4 presents the early combination results in comparison to BoVW model; Both the minimum ranking and the round robin methods return more positive results as the retrieval score increases up to 2.37% and 2.41% respectively.

The results presented in both tables show that mixing vocabularies may not always return better retrieval scores in comparison to a single IG model, but it is close to the best one. Note also that the construction complexity (offline process) of these improvements is multiplied by the number of IG model used. However, the retrieval (online process) is similar to the initial framework and to the BoVW model. Globally, we can conclude that mixing vocabularies is a right choice when there is a lack of knowledge on the nature of database as well as when several features are used.

Table 4 Early combination scores

Descriptor	Database	rrBest	rankBest
CM	UKB	**2.71 (1.1%)**	**2.71 (1.1%)**
	PASCAL	**25.41 (0.5%)**	**25.83(2.1%)**
	Holidays	**0.52(4%)**	**0.51(1.5%)**
SIFT	UKB	**2.26(4%)**	**2.25(3.8%)**
	PASCAL	35.57(-0.2%)	35.45(-0.5%)
	Holidays	**0.34(8.9%)**	**0.34(8.7%)**
CMI	UKB	**3.04(2.7%)**	**3.05(3.2%)**
	PASCAL	**29.1(2.0%)**	**29.2(2.4%)**
	Holidays	**0.53(4.0%)**	**0.53(3.8%)**
OppSIFT	UKB	**2.39(2.7%)**	**2.4(3.0%)**
	PASCAL	34.61(-1.2%)	34.73(-0.9%)
	Holidays	**0.49(0.4%)**	**0.48(0.4%)**)
Mean Diff		2.41%	2.37%

Extension to visual phrases

To enhance the performance of BoVW model, an interesting approach consists in using *phrases* (i.e. groups of words) to create more discriminative descriptors which are called visual phrases. Bag of Visual Phrases (BoVP) model has been proved to outperform the effectiveness of BoVW model [29], [19]. In [29], authors combine SIFT and MSER (maximally stable region) detector to build a new feature. In [19], a visual word is combined with its nearest neighbours to make visual phrases; the authors implement several experiments with different phrase lengths and show that the best result can be achieved with visual phrases combining 2 or 3 visual words. Note that most BoVP models use an indexing structure to speed up the retrieval as BoVP is more complex than BoVW model, even during the retrieval process.

Table 5 Result of Visual Phrases on UKB database

Desc	Baseline	tf-idf	tfc	entropy
CMI	3.29	3.36(2.1%)	3.37(2.5%)	3.32(1.2%)
CM	3.02	3.06(1.6%)	3.06(1.3%)	3.07(1.6%)

Our visual phrase algorithm is built upon BoVW model. First, all keypoints in the image are represented by their closest visual words in the visual word vocabulary, and then each keypoint in the image is linked to other keypoints in the spatial neighbourhood to make visual phrases. In our experiments, the neighbour windows size, set to 10% of the image size, is centered on the keypoint coordinates. Table 5 shows the result of our experiment with BoVP model on UKB database using the vocabularies obtained by different IG models. *Baseline* here denotes the BoVP model with k-means-based BoVW model. The numbers inside brackets indicate the difference in percentage of the retrieval score by replacing the vocabulary based *k-means* algorithm by the one obtained by information model.

We see that, although vocabularies are constructed based on the information contained in each visual word individually, they also enhance the performance of BoVP model where visual words are linked together. This result once more demonstrates the importance of using an IG model to construct vocabularies.

Discussion

Results obtained from our exhaustive experiments are mainly positive towards using IG models to construct a visual vocabulary. Except for the *bm25* model, all others models yield better retrieval scores than the classical BoVW using *k-means* clustering algorithm. However, one can observe a difference in the results with respect to the datasets.

Indeed, most of PASCAL scores are very close one to another. The maximum variation is -3.1% for *bm25* model using SIFT descriptor, but the variations generally range from -1% to +1%. This means that for this dataset, the vocabulary construction methodology has little impact in our study. This might be due to PASCAL dataset nature which is a classification benchmark (20 classes) where machine learning techniques are generally used to obtain a good performance. UKB and Holidays are retrieval datasets, with similar objects or scenes to find in the whole collection. Score differences are much wider, up to +10,4% for Holidays with *tfc* model using SIFT descriptor. The results with those datasets also show the importance of constructing a good visual vocabulary by selecting an appropriate IG model. The proposed approach proves a good performance with retrieval datasets using a very small vocabulary (256). Such a vocabulary size makes it difficult to perform well in classification, which explains the results with PASCAL.

Mixing visual vocabularies built with different IG models yields promising results. The obtained scores are mainly higher than classical BoVW showing its interest when no a priori knowledge on the used dataset or descriptors. Although the selection of visual vocabulary is based on the IG of each individual visual word, it improves the performance of BoVP where visual words are linked together to make visual phrases. This result demonstrates the discriminative power of IG models.

In a last experiment, we have used a saliency model as a extra weight to build the histogram of visual words or phrases. Results show that keypoints with higher saliency value plays a more important role in the retrieval process, and thus, saliency values can be use as a weight to enhance the performance of BoVW and BoVP. Note that using saliency map as a weight to build the image signature is just one way of taking advantage of visual attention models in CBIR tasks. The potential of visual attention model is still very open and will be part of future works.

6 Conclusion

In this paper, we have described an evaluation study of four information gain models for the construction of a visual vocabulary: *tf-idf*, *tfc*, *bm25* and *entropy*. Starting with an initial set of randomly-selected candidate visual words, these models are used to iteratively select words with highest IG scores, so as to generate the final vocabulary. The final vocabulary is the set of the best visual words in term of IG for one specific IG model.

To enhance the performance of our framework, we propose two extensions: one extension consists in combining several vocabularies obtained with different IG models in order to further improve the quality, and the other extension illustrates how the proposed framework can be applied to visual phrases. This demonstrates the interest of the proposed framework when no prior knowledge is available regarding datasets and descriptors. Experiments on different image datasets with several widely-used descriptors have demonstrated that except for *bm25*, all other IG-based vocabularies achieve better results than classical BoVW model.

Future works will focus on the combination of IG models and saliency models in the word selection process, and also for constructing or filtering visual phrases in order to retain only those composed of informative and salient words.

References

1. V. Abolghasemi, S. Ferdowsi, and S. Sanei. Fast and incoherent dictionary learning algorithms with application to fmri. *Signal, Image and Video Processing*, 9(1):147–158, 2015.
2. Y. Alqasrawi, D. Neagu, and P. I. Cowling. Fusing integrated visual vocabularies-based bag of visual words and weighted colour moments on spatial pyramid layout for natural scene image classification. *Signal, Image and Video Processing*, 7(4):759–775, 2013.
3. G. Amati and C. J. van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans. Inf. Syst.*, 20(4):357–389, 2002.
4. G. Csurka, C. Bray, C. Dance, and L. Fan. Visual categorization with bags of keypoints. *Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22, 2004.
5. I. Elsayad, J. Martinet, T. Urruty, and C. Djeraba. Toward a higher-level visual representation for content-based image retrieval. *Multimedia Tools Appl.*, 60(2):455–482, 2012.
6. M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/-VOC/voc2012/workshop/index.html, 2012.
7. K. Gao, S. Lin, Y. Zhang, S. Tang, and H. Ren. Attention model based sift keypoints filtration for image retrieval. In R. Y. Lee, editor, *ACIS-ICIS*, pages 191–196. IEEE Computer Society, 2008.
8. Z. Harris. Distributional structure. *Word*, 10(23):146–162, 1954.
9. L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(11):1254–1259, Nov. 1998.
10. H. Jégou, M. Douze, and C. Schmid. Hamming embedding and weak geometric consistency for large scale image search. In A. Z. David Forsyth, Philip Torr, editor, *European Conference on Computer Vision*, volume I of *LNCS*, pages 304–317. Springer, oct 2008.
11. H. Jégou, M. Douze, C. Schmid, and P. Pérez. Aggregating local descriptors into a compact image representation. In *23rd IEEE Conference on Computer Vision & Pattern Recognition (CVPR '10)*, pages 3304–3311, San Francisco, United States, 2010. IEEE Computer Society.
12. S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2*, CVPR '06, pages 2169–2178, Washington, DC, USA, 2006. IEEE Computer Society.
13. S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), 17-22 June 2006, New York, NY, USA*, pages 2169–2178, 2006.
14. H. T. Le, S. Gbèhounou, T. Urruty, F. Lecellier, and C. Fernandez-Maloigne. Information gain study for visual vocabulary construction. In A. G. Hauptmann, C. Ngo, X. Xue, Y. Jiang, C. Snoek, and N. Vasconcelos, editors, *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, Shanghai, China, June 23-26, 2015*, pages 503–506. ACM, 2015.
15. F. R. López, H. Jiménez-Salazar, and D. Pinto. A competitive term selection method for information retrieval. In *CICLing*, pages 468–475, 2007.
16. O. L. Meur, P. L. Callet, D. Barba, and D. Thoreau. A coherent computational approach to model the $bottom-up$ visual attention. *IEEE Transactions on Pattern Analysis and MLachine Intelligence (PAMI)*, 28(5):802–817, 2006.
17. D. Nistér and H. Stewénius. Scalable recognition with a vocabulary tree. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 2161–2168, June 2006.
18. L. Parsons, E. Haque, and H. Liu. Subspace clustering for high dimensional data: a review. In *ACM SIGKDD*, volume 6, pages 90–105. Explorations Newsletter, 2004.
19. G. Pedrosa and A. Traina. From bag-of-visual-words to bag-of-visual-phrases using n-grams. In *Graphics, Patterns and Images (SIBGRAPI), 2013 26th SIBGRAPI - Conference on*, pages 304–311, Aug 2013.
20. F. Perronnin and C. R. Dance. Fisher kernels on visual vocabularies for image categorization. In *2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007), 18-23 June 2007, Minneapolis, Minnesota, USA*. IEEE Computer Society, 2007.
21. J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2007.
22. Y. Ren, A. Bugeau, and J. Benois-Pineau. Bag-of-bags of words irregular graph pyramids vs spatial pyramid matching for image retrieval. In *Image Processing Theory, Tools and Applications (IPTA), 2014 4th International Conference on*, pages 1–6, Oct 2014.
23. S. E. Robertson, S. Walker, and M. Hancock-Beaulieu. Experimentation as a way of life: Okapi at TREC. *Inf. Process. Manage.*, 36(1):95–108, 2000.
24. G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.*, 24(5):513–523, Aug. 1988.
25. J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *Proceedings of the International Conference on Computer Vision*, pages 1470–1477, Oct. 2003.
26. C.-Y. Tsai, T.-C. Lin, C.-P. Wei, and Y.-C. Wang. Extended-bag-of-features for translation, rotation, and scale-invariant image retrieval. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 6874–6878, May 2014.
27. T. Urruty, S. Gbèhounou, H. T. Le, J. Martinet, and C. Fernandez-Maloigne. Iterative random visual word selection. In *International Conference on Multimedia Retrieval, ICMR '14, Glasgow, United Kingdom - April 01 - 04, 2014*, page 249, 2014.
28. K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1582–1596, 2010.
29. Z. Wu, Q. Ke, M. Isard, and J. Sun. Bundling features for large scale partial-duplicate web image search. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 25–32, June 2009.
30. F. Yeganli, M. Nazzal, and H. Özkaramanli. Image super-resolution via sparse representation over multiple learned dictionaries based on edge sharpness and gradient phase angle. *Signal, Image and Video Processing*, 9(Supplement-1):285–293, 2015.