Distances and Weighting Schemes for Bag of Visual Words Image Retrieval

Pierre Tirilly
CNRS
IRISA
Rennes, France
ptirilly@irisa.fr

Vincent Claveau
CNRS
IRISA
Rennes, France
vclaveau@irisa.fr

Patrick Gros
INRIA
Centre Rennes - Bretagne Atlantique
Rennes, France
pgros@inria.fr

ABSTRACT

Current text retrieval techniques allow to index and retrieve text documents very efficiently and with a good accuracy. Image retrieval, on the contrary, is still very coarse and does not yield satisfying results. Therefore, computer vision researchers try to benefit from text retrieval contributions to enhance their retrieval systems. In particular, Sivic and Zisserman, with their *video-google* framework [1], propose a description of images similar to standard text descriptors: images are described by elementary image parts, called *visual words*. Thus, they perform image retrieval using the standard Vector Space Model (VSM) of Information Retrieval (IR) and benefit from some classical IR techniques such as inverted files. Among available text retrieval techniques, automatically finding the most relevant words to describe a document has been intensively studied for texts, but not for images. In this paper, we propose to explore the use of term weighting techniques and classical distances from text retrieval in the case of images. These weights are standard VSM weights, weights derived from probabilistic models of IR or new weighting schemes that we propose. Our experiments, performed on several datasets, show that no weighting scheme can improve retrieval on every dataset, but also that choosing weights fitting the properties of the dataset can improve precision and MAP up to 10%. This study provides some interesting insights about the semantic and statistical differences between textual and visual words, and about the way visual word-based image retrieval systems can be optimized. It also shows some limits of the bag of visual words model, and the relation existing between Minkowski distances and local weighting. At last, this study questions some experimental habits commonly found in the literature (choice of L1 distance, TF*IDF weights and evaluation using one dataset only).

Categories and Subject Descriptors

H.3.1 [**INFORMATION STORAGE AND RETRIEVAL**]: Content Analysis and Indexing; H.3.3 [**INFORMATION STORAGE AND RETRIEVAL**]: Information Search and Retrieval—*Retrieval models*; I.4.10 [**IMAGE PROCESSING AND COMPUTER VISION**]: Image Representation

General Terms

EXPERIMENTATION

Keywords

Image retrieval, Bags of visual words, Weighting schemes, Text retrieval, TF*IDF, Divergence from randomness, Minkowski distance, Fractional distance

MIR'10, March 29–31, 2010, Philadelphia, Pennsylvania, USA.

1. INTRODUCTION

The quick growth of professional and personal image databases (such as `www.flickr.com`, that hosts more than 3 billion images) raises the problem of accessing their content efficiently with a good precision. Content-Based Image Retrieval (CBIR) systems typically work as any retrieval system: given an image collection, image descriptors (*e.g.* vectors) are computed; then, for any image query, its descriptor is computed and matched to the others (using a distance or a similarity measure, *e.g* euclidean distance L2) to find the images that are close to the query. A particular difficulty is to find a good descriptor of the image content, so that the system can retrieve semantically similar images. Such a descriptor can also be used to perform image classification or annotation. Former CBIR systems relied on global descriptors such as color histograms, but such descriptors cannot handle local information, and therefore cannot detect the objects contained in a picture. More recent approaches rely on *bags of visual words* or *bags of features*: images are described as sets of elementary patches called visual words, by analogy with text. By counting the occurrences of visual words, we can describe images as vectors where each dimension corresponds to the frequency of a given visual word. This representation of images is similar to the traditional *Vector Space Model* (VSM) for text retrieval [1], and can therefore benefit from previous text retrieval studies such as inverted files [1] or latent semantic analysis [2]. In particular, finding the most significant words to describe a document, and therefore to better match documents, has been intensively studied in text retrieval. It can be done by weighting the term frequencies in the VSM, or defining new matching measures, generally in a probabilistic framework.

Currently, such weights have not been extensively studied in the case of images, except TF*IDF and binary weights. In this paper, we propose to apply to images traditional text retrieval weighting schemes derived from the VSM or from probabilistic models. We also study the influence of the distance used to match documents on the retrieval performance. Our experiments are performed on various retrieval tasks and datasets, which is very rarely the case in the literature. This work provides a comparison between textual and visual words, and some interesting insights about the statistical distribution of visual words, their significance and the ways a CBIR system based on visual words can be optimized.

The paper is organized as follows: first we detail the bag of visual words model and its relation with text retrieval, then we give an overview of related studies from the literature. In Sect. 4 we present the standard distances of information retrieval. In Sect. 5 we detail the weighting schemes we consider in our study. Section 6 describes the experimental settings and results. Some specific conclusions are discussed in Sect. 7. Last, we conclude in Sect. 8.

2. BAG OF VISUAL WORDS AND RELATIONS WITH TEXT RETRIEVAL

The bag of visual words model has been initially proposed by Sivic *et al.* [1]. It represents images as sets of elementary image patches called visual words. The whole model is shown on Fig. 1. The first step to describe images using visual words is to build a set of visual words, called *visual vocabulary*; then each image to index can be described using the words of the vocabulary that occur in it.

2.1 Visual vocabulary

The visual vocabulary is computed on a set of images. First, interest regions (*i.e.* image regions containing interesting features such as corners) are detected on each image using an interest point detector [3]. Each interest region is then described as a multidimensional numerical vector called a local descriptor [4]. These descriptors are grouped using a clustering algorithm such as *k-means*. The resulting clusters are considered as visual words, *i.e.* all the local descriptors that fall in a given cluster are considered as instances of the same visual word. So, visual words represents elementary parts of objects, such as eyes or wheels. Using visual words is then quite similar to using a stemmer for text retrieval: local descriptors that describe the same object part are expected to correspond to the same visual word, making the overall system more robust, but also faster, by limiting the number of possible descriptors. It is important to notice that the vocabulary size has to be manually set, as a parameter of the clustering algorithm. This parameter, as well as the algorithms used to detect, describe and cluster the interest regions, are known to have an influence on the vocabulary obtained, and then on the retrieval performance, but the study of these parameters are beyond the scope of this paper.

2.2 Image description

Given a visual vocabulary, an image can then be described as a set of visual words, or as a vector of visual word frequencies. First, local descriptors are extracted from the image (detection, then description, of interest regions). Then, each local descriptor is associated with its visual word, *i.e.* the cluster it falls into.

2.3 Relation with text retrieval

This way of describing images is very similar to the descriptors used for text retrieval: visual words can be seen as index terms, and any retrieval model (VSM, boolean, probabilistic...) can be used to match images. The first paper about visual word-based retrieval emphasizes this similarity [1], and some studies apply text retrieval techniques to image retrieval [2, 1]. However, some inherent differences exist and must be taken into account when adapting text retrieval techniques to image retrieval (a more detailled presentation of these differences is available in [5]):

- the initial vocabulary does not only depend on the collection (words that occur in it), but also on the algorithms and parameters used (clustering algorithm, interest regions detector and descriptor, size of the vocabulary). Therefore many different visual vocabularies can be obtained for any collection;
- the queries are full images, they are therefore much longer than usual text queries;
- the meaning of the words: whereas one given textual word has generally one (or more) given meaning, an object in an image is described by several visual words. The *word independance assumption* commonly used in text retrieval is therefore much less suited to the case of images;
- word frequencies, word document frequencies and document length vary according to many parameters: the vocabulary used, the size of images, the size of the objects in the images, the number of objects in images...

3. RELATED WORK

Bag of visual words image retrieval has been intensively studied recently but there is little work about the efficiency of using traditionnal distance and weighting schemes in this context. Jiang *et al.* have studied the effect of TF*IDF weighting and binary weighting in the case of image retrieval [6] and classification [7]. To our knowledge, this is the only work about weighting schemes based on the statistical distribution of visual words in documents. The other weights proposed in the litterature rely on information from the clustering stage of the vocabulary building process [8, 9] or spatial information from a color-based image segmentation [10]. Distances $L1$ and $L2$ have been concisely compared by Nister and Stewenius [11] on one dataset. They conclude that $L1$ performs better but do not explain this result. Distance studies by Jegou *et al* include the definition of custom distances: one uses the fact that the vocabulary is based on clusters [12], the other on nearest-neighbor search properties [13]. These distances were tested on only one dataset each. Jiang *et al.* also proposed a distance based on the vocabulary clusters [8].

4. DISTANCE MEASURES

Here we present the distances that we study in our experiments. The most common matching measures used to compare documents in the VSM are the cosine similarity

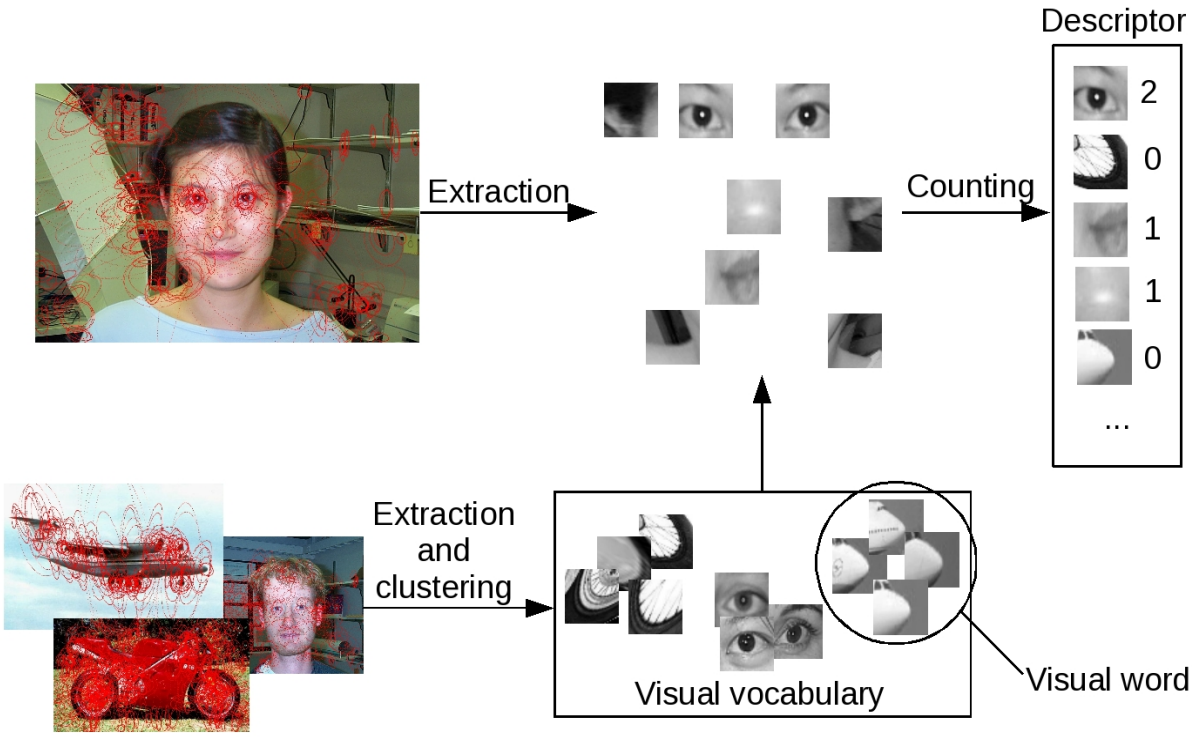

Figure 1: Construction of a visual vocabulary and description of an image as a bag of visual words

(Equation 2) and the distances $L1$ and $L2$ derived from the general Minkowski distance Lk (Equation 3). The first remark to do is that $L2$ and the cosine similarity are equivalent when the vectors are normalized: given a query q, they provide the same document rankings, since, for any document d:

$$d_{L2}(d,q)^2 = 2(1 - d_{\cos}(d,q)) \quad (1)$$

$$d_{\cos}(d,q) = \frac{\sum_i d_i.q_i}{\sqrt{\sum_i d_i^2}.\sqrt{\sum_i q_i^2}} \quad (2)$$

$$d_{Lp}(d,q) = (\sum_i |d_i - q_i|^p)^{\frac{1}{p}} \quad (3)$$

However, it is possible to compute distances using a Minkowski distance Lk with any real value of the distance parameter k. Such distances have useful properties when considering high dimensional vectors [14]:

- using low values of k limits the effect of the curse of dimensionality, which tends to make the distances between vectors equal when the number of dimensions increases;
- low values of k give less importance to large local distances[1] whereas high values of k increase their importance;
- low values of k make the distance robust to the presence of noise in the data. However, when the data is too noisy, all distances tend to be equivalent.

[1]distance computed on one given dimension

We also performed some experiments using information theoretic measures (Kullback-Liebler Divergence, Kullback-Liebler Distance, Jensen-Shannon divergence), but we do not report these results here as these similarity measures significantly worsen the retrieval performance.

5. WEIGHTING SCHEMES FROM TEXT RETRIEVAL

In this section, we present the weighting schemes that we use in our experiments. They are standard text weights, weights derived from probabilistic IR models or new weighting schemes that we propose. Table 1 shows the notations used hereafter.

5.1 Vector space model weighting schemes

Weights w_{ij} in the vector space model are usually divided into three parts l_{ij}, g_i and n_j, so that $w_{ij} = l_{ij}.g_i.n_j$:

- a local weight l_{ij}, that reflects word t_i importance in document j. Traditional local weights are $l1$, $l2$, $l3$, $l4$, $l6$ from Table 2. Some emphasizes high-frequency words ($l6$) whereas some limit the influence of high frequencies ($l2$, $l3$, $l4$). We added weights $l5$ and $l7$ that we derived from probabilistic matching measures (see Sect. 5.2);
- a global weight g_i: this weight reflects the importance of the word in the collection. The underlying idea is generally that, the less document a word occurs in, the more importance it has. It justifies weights $g1$ and $g3$ from Table 3. g_2 also follows this idea, although it is initially based on probabilistic considerations (see Sect. 5.2). We also propose two new weights,

t_i	i-th term (or word)
d_j	j-th document
N	Number of documents
n_i	Number of documents containing t_i
tf_{ij}	Number of occurrences of t_i in d_j

$\overline{\mathrm{tf}_i}$	Mean occurrences of t_i
CF_i	Number of occurrences of t_i in the collection
CF^*	Total number of word occurrences in the collection
l_j	Length of d_j (number of word occurrences)
avg_l	Average document length

Table 1: Notations used in this paper

g_4 and g_5. They are based on the assumption that words that are repeated in a picture are usually relevant (windows, eyes, wheels...);

- a normalization factor n_j: it aims at normalizing the weighted vectors according to their length, so that the distances between vectors are comparable. When using Lk Minkowski distances, it just consists in dividing the vectors by the appropriate norm.

5.2 Probabilistic models weights

Probabilistic retrieval has raised in the sixties. It relies on Robertson's *Probability Ranking Principle* which states that an optimal retrieval system should rank documents according to their probabilities of relevance to the query [15]. It provides probabilistic measures to match documents according to a given data model. These measures usually take the following form:

$$PM(d_j, q) = \sum_i q_i.w(d_{ij}) \quad (4)$$

This is an inner product of vectors q and $w(d_j)$. It is therefore possible to consider any Minkowski distance instead of the inner product. Of course, the vectors first need to be normalized with a well-suited norm. In our case, we will also consider that d_j and q need to be weighted (since we use full images as queries), hence we will compute the distance between $w(d_j)$ and $w(q)$.

5.2.1 BM25-based local and global weights

BM25 is often considered as the best weighting scheme for text retrieval. Equation (5) is a simplified version of BM25 matching score that assumes that no relevance information is available. We can extract two parts from it:

- a local weight part, which corresponds to local weight $l7$ in Table 2. This local weight is obtained by considering that word occurrences are distributed following two Poisson distributions (see [16] for details).
- a global weight part, which corresponds to global weight $g2$ from Table 3. It is supposed to provide optimal results, according to the *probability ranking principle* [17].

$$\mathrm{BM25}(d_j, q) = \sum_{t_i \in q} \frac{tf_{ij} * (k_1 + 1)}{K + tf_{ij}} \cdot tf_{iq} \cdot \log \frac{N - n_i}{n_i} \quad (5)$$

5.2.2 Divergence From Randomness-based weights

Divergence From Randomness (DFR) matching measures have been proposed by Amati and van Rijsbergen [18]. Their general matching measure provides two-parts weights, as shown in (6). It contains two information sources, Inf_1 and Inf_2. Inf_1 is the *informative content* of the word and is based on the probability for t_i to occur in any document according to a given *randomness model*. Table 4 shows the randomness models we used in this study; they all come from Amati's work (see [18] for details about them), except the hypergeometric model HG that we define ourselves. Contrary to Bernouilli models P and D, which consider a document as successive random draws of single words, the HG model considers a document as a single draw of several words: it seems more suited to image retrieval, because an object in an image corresponds to several visual words with specific frequencies (*e.g.* a face contains two eyes, one nose...), whereas, in a text, each word conveys its own meaning. Inf_2 is called *information gain* and corresponds to the increase of the weight when t_i is considered as a good descriptor of document d_j. Table 5 shows the two information gain models proposed in [18]. Amati *et al.* also add a frequency normalization to take account of the document size. Table 6 shows the normalization proposed. One of them corresponds to the local weight $l5$ that we also test as an independent weight.

Name	Inf_2
L	$\frac{1}{\mathrm{tf}_{ij}+1}$
B	$\frac{\mathrm{CF}_i+1}{n_i \cdot (\mathrm{tf}_{ij}+1)}$

Table 5: Models of divergence

H0	tf_{ij}
H1	$\mathrm{tf}_{ij} \cdot \frac{\mathrm{avg\_l}}{l_d}$
H2	$\mathrm{tf}_{ij} \cdot \log_2(1 + \frac{\mathrm{avg\_l}}{l_d})$

Table 6: tf_{ij} normalization for DFR weights

$$\begin{aligned} w(t_i, d_j) &= \mathrm{Inf}_1(t_i, d_j).\mathrm{Inf}_2(t_i, d_j) \\ w(t_i, d_j) &= -\log_2(\mathrm{Prob}_1(t_i, d_j)).(1 - \mathrm{Prob}_2(t_i, d_j)) \end{aligned} \quad (6)$$

6. EXPERIMENTS

In these experiments, we test the effectiveness of the distances and weighting schemes that we presented in Sect. 4 and 5 on several datasets.

6.1 Retrieval tasks and associated datasets

As in the case of text retrieval, we use standard image collections with groundtruth to evaluate our systems. They are widely used in the computer vision literature. Each of these collections corresponds to one of these two specific retrieval tasks: scene retrieval and object category retrieval.

6.1.1 Scene retrieval

The aim of scene retrieval is, given a query, to retrieve images containing the strictly same objects, with variations in viewpoint or illumination (see Figure 2). This task is usually used to evaluate visual word-based retrieval systems [13, 12, 19, 11, 7, 20]. We evaluate our system on this task with the two following datasets:

Nister [11]. It is composed of 2,550 scenes with 4 images each.

$l_1(t_i, d_j)$	Term Frequency TF	tf_{ij}
$l_2(t_i, d_j)$	Frequency logarithm	$\begin{cases} 1+\log(\mathrm{tf}_{ij}) & \text{if } \mathrm{tf}_{ij} > 0 \\ 0 & \text{otherwise} \end{cases}$
$l_3(t_i, d_j)$	Augmented normalized frequency	$\begin{cases} a+(1-a)\frac{\mathrm{tf}_{ij}}{\underset{t_k \in d_j(\mathrm{tf}_{kj})}{max}} & \text{if } \mathrm{tf}_{ij} > 0 \\ 0 & \text{otherwise} \end{cases}$
$l_4(t_i, d_j)$	Binary	$\begin{cases} 1 & \text{if } \mathrm{tf}_{ij} > 0 \\ 0 & \text{otherwise} \end{cases}$
$l_5(t_i, d_j)$	DFR-like normalization	$\mathrm{tf}_{ij} \cdot \frac{l_{avg}}{l_j}$
$l_6(t_i, d_j)$	Squared TF	tf_{ij}^2
$l_7(t_i, d_j)$	BM25 TF	$\mathrm{tf}_{ij} \cdot \frac{k_1+1}{\mathrm{tf}_{ij}.k1(1-b+b.\frac{l_j}{\mathrm{avg\_l}})}$ with $k_1 = 1.2, b = 0.75$

Table 2: Local weighting schemes

$g_0(t_i)$	No weigth	1
$g_1(t_i)$	Inverse document frequency (IDF)	$\log(\frac{N}{df_i})$
$g_2(t_i)$	Probabilistic IDF	$\max(0, \log(\frac{N-df_i}{df_i}))$
$g_3(t_i)$	Squared IDF	$\log(\frac{N}{df_i})^2$
$g_4(t_i)$	(Mean TF) * IDF	$\overline{tf_i}\log(\frac{N}{df_i})$
$g_5(t_i)$	Squared (mean TF) * IDF	$[\overline{tf_i}\log(\frac{N}{df_i})]^2$

Table 3: Global weighting schemes

Oxford [19]. It contains 5,062 images and provides 55 queries with groundtruth. The queries are Oxford buildings that can be partially occluded, making the problem more difficult than it was with Nister dataset.

6.1.2 Object category retrieval

This tasks does not aim at retrieving images that are exactly similar to the query, but images that contain objects from the same category than the objects appearing on the query. This task is more challenging than scene retrieval as it requires the retrieval system to be discriminant, but also to have some generalization capabilities. Visual word-based retrieval systems are very rarely evaluated on such tasks (to our knowledge, only [21] handles this task). To perform this evaluation, we use datasets designed for image classification. We consider that, given an image from one category as a query, every image from that category is relevant to this query. Here are the datasets we used:

Caltech6. We designed this dataset with six Caltech categories: airplanes, backgrounds, car rears, faces, guitars, motorbikes. It contains 5,415 images. There are few categories, but some of them have high intra-class variation.

Caltech101 [22]. It contains 8,197 images belonging to 101 categories with high intra-class variation. It is the most challenging dataset used here.

6.2 Experimental settings

6.2.1 Queries

We only use complete images as queries, even with Oxford dataset which also provides region queries. We randomly choose 200 images as queries on Caltech6, 200 on Caltech101 and 300 on Nister. 55 queries are given with Oxford dataset.

6.2.2 Visual vocabulary

We rely on standard techniques to build our visual vocabulary. Interest regions are detected by the Hessian-affine detector: it provides good performances [3] and is widely used in visual word-based studies. These regions are then described as 128-dimensions SIFT descriptors: this descriptor yields good results [4] and is the most used in the literature. The SIFT descriptors are then clustered using a hierarchical k-means algorithm [11]. The vocabulary sizes are:

Caltech6	Caltech101	Nister	Oxford
6,556	61,687	19,545	117,151

6.2.3 Evaluation

We use Mean Average Precision (MAP) and precision after 10 documents to evaluate the retrieval performance. Precision gives the rate of relevant documents among the first 10 retrieved documents, as a user would generally watch these documents first. MAP provides an overview of the performance, and especially reflects the fact that the last relevant documents retrieved are well or badly ranked. We also use Wilcoxon tests to check the statistical significance of our results, but these results are not systematically reported here by lack of space (detailled results are available in [5]).

6.3 Distance experiments

In order to evaluate the effect of the distance that matches image descriptors on the retrieval performance, we test Lk distances with several values of k. We do not report experiments using the cosine similarity as it is equivalent to $L2$ (see Sect. 4). Figure 4 details the results on the four datasets when using a TF*IDF weighting scheme, which is the most common weighting scheme in the literature. Three remarks are worth noting:

- on Caltech6 and Nister, best results are obtained when $k \approx 0.75$; the relative gain between the best and the

Name	Model of Prob_1	Approximation for Inf_1
P and D	$\left[\binom{\text{CF}_i}{\text{tf}_{ij}}\left(\frac{1}{N}\right)^{\text{tf}_{ij}}\left(\frac{N-1}{N}\right)^{\text{CF}_i-\text{tf}_{ij}}\right]$	see [18]
G and Be	$\frac{(\text{CF}_i-\text{tf}_{ij}+1)\cdot\ldots\cdot\text{CF}_i\cdot(N-1)}{(N+\text{CF}_i-\text{tf}_{ij}-1)\cdot\ldots\cdot(N+\text{CF}_i-1)}$	see [18]
In and In_e	$\left(\frac{n_j+0.5}{N+1}\right)^{\text{tf}_{ij}}$	see [18]
HG	$\frac{\binom{\text{CF}_i}{\text{tf}_{ij}}\binom{\text{CF}^*-\text{CF}_i}{l_j-\text{tf}_{ij}}}{\binom{\text{CF}^*}{l_j}}$	see [5]

Table 4: Models of randomness and their approximation

worse results ($k = 3$) is very important (for Caltech6, up to +33% in precision and +26% in MAP);

- on Caltech101, the results tend to be constant for any k;
- on Oxford, on the contrary, the best results are obtained with high values of k. Here again the relative gain between the worst and the best results are important (+18% in precision and +38% in MAP).

Some aspects of these results are discussed in Sect. 7.

6.4 Weighting experiments

Figure 5 shows the improvement provided by weighting schemes composed by the weights from Tables 2 and 3, compared to standard $l1g0$ weights. Figure 6 shows the performance improvement obtained when using DFR weights and $L1$ distance, compared to standard $l1g0$ weights. Although the results vary from one dataset to another, we can make some interesting remarks about them.

6.4.1 Local weights

On Caltech6 and Nister, the local weights providing the best improvement are $l2$, $l3$ and $l7$. This is consistant as these three local weights have a similar role: limiting the influence of high word frequencies. On Caltech101, the results are much less significative. Two reasons explain that: noise in the data and a large vocabulary. These two reasons make the mean frequency of words much smaller than in the other datasets, close to 1. In this case, local weights tend to be equivalent, and therefore do not make much difference in the results. At last, Oxford data yields opposite results, as in the case of our distance experiment: $l6$ provides the best improvement, whereas it was the worse local weight on other datasets.

6.4.2 Global weights

Squared IDF ($g3$) provides an improvement for Caltech6, Nister and Oxford. It seems therefore a good global weight in most cases, better than standard IDF. The weights we proposed, based on IDF and mean frequency ($g4$ and $g5$) perform well on two datasets, $g5$ yielding the best results. However, these results strongly depend on the dataset properties: with Nister, these weights perform bad. At last, global weighting is not effective on Caltech101: the best and most stable global weight is $g0$, *i.e.* no weighting.

6.4.3 DFR weights

The results are often close from one model to another, and the differences are not always statistically significative. However, the binomial randomness model D performs best, followed by In and Ine. Our model HG also performs well when no frequency normalization is used, since the model already handles the document length. The effectiveness of the gain model strongly depends on the data: for instance, the two gain models give opposite results on Nister and Caltech101. Globally, DFR can improve the retrieval performance, especially the precision, on some datasets, but on general datasets such as Caltech101, it worsens the standard results, like the other weights.

7. DISCUSSION

Besides the results presented in previous section, we make here three more general remarks about the visual word-based retrieval model and the use of Minkowski distances and weighting schemes.

7.1 A limit of the model for object category retrieval

The results of the distance experiments on Caltech101 show that the parameter k of Minkowski distances has a very small effect on the performances of object category retrieval when the dataset is complex. It has been shown that such behaviour is characteristic of noisy data [14]. It may reveal a weakness of the visual word model when retrieving object categories. Due to the variations between objects of a given category, and to the vocabulary size, the probability that a SIFT descriptor is assigned to one cluster instead of another is high, making the data description very noisy. The coarse clustering algorithm we use is of course partially responsible of that, but using a better clustering algorithm with more complex datasets will yield the same result. It shows that the generalization properties (induce by the descriptor quantization stage) of the visual word retrieval model is low on complex datasets, thus limiting the retrieval performance.

7.2 Influence of query properties on the system's parameters

We shown that the performance of a given weighting scheme or Minkowski distance strongly depends on the dataset we used: the best parameters for the Oxford dataset are the complete opposite of the best parameters for Caltech6 and Nister datasets. This is due to the visual properties of the queries we considered. The queries of the Oxford dataset are buildings: such objects are characterized by many repeated objects (windows, doors, arches... – see Figure 7), so giving more importance to high frequency words in the similarity measure (as done by the $l6$ local weight)

improves the performance. On the contrary, the queries of Nister and Caltech6 contain few repeated parts, so the presence or absence of a given visual word is more important than its frequency. This is why local weights that limit high word frequencies (such as $l2$, $l3$ and $l7$) can improve the performance on these datasets. However the weak performances of $l4$, and the good performances of global weights $g4$ and $g5$ show that frequency information still has some importance and must not be totally discarded. Thus, these experiments show that the optimal parameters of a visual word-based retrieval system change with the nature of the image query. Future work may include the study of techniques to automatically choose parameters that are adapted to the current query.

7.3 Relation between local weights and fractional distances

The last point is that there is a relation between local weighting and the parameter k of Minkowski distances. We observe that, on Caltech6 and Nister, using low values of k improves the performance, as well as using local weights that limit high word frequencies. On Oxford, high values of k improves the performance, as well as local weights that emphasize high word frequencies. This is consistent: low values of k tend to lower the value of local distances (we call local distance a distance computed on one dimension only), as local weights such has $l2$ do; complementary, high values of k increase local distance, as the $l6$ weight do. So tuning k or tuning the local weight has the same effect on retrieval performances. This property may be interesting, since Minkowski distances with non-integer k value have a major drawback: they do not respect triangle equality [14]. Using L1 distance instead with an appropriated local weight might provide similar results, by allowing the use of techniques that rely on triangle inequality to quicken the retrieval process, as in [23].

8. CONCLUSION AND FUTURE WORK

In this paper, we have tested several distances and weighting schemes to improve image retrieval based on bags of visual words. We have also proposed new weighting schemes suited to image retrieval. We perform these experiments on several tasks and datasets, which is generally not the case in the literature. We show that the effectiveness of a given weighting scheme or distance is strongly linked to the dataset used: a trade-off between giving more or less importance to word frequencies has to be found. Some datasets (like Caltech6 and Nister in our experiments) require to reduce word frequency influence, whereas others (like Oxford) require to emphasize it. We also show that, in the case of large and varied image collections, the noise in descriptor assignation and the need to use larger vocabularies tend to make all distances and weights equivalent. On some datasets however, the choice of a suited distance or weighting scheme can improve the performance significantly. In particular, the global weights we proposed perform best on two datasets. We also highlighted the link between local weighting and the tuning of the Minkowski distance parameter. Last, this study calls some experimental habits into question: L1 and TF*IDF are not the best parameters of visual word-based retrieval systems, and such systems need to be evaluated on several datasets to avoid bias.

The next step of this work is to study the impact of weighting schemes combined with varied vocabulary sizes. Current studies [11] tend to show that larger vocabularies provide better results. We should check whether the use of an adapted weighting scheme allows to reduce vocabulary size, hence the computational cost of the retrieval process, without losing effectiveness.

9. REFERENCES

[1] Sivic, J., Zisserman, A.: Video Google: A text retrieval approach to object matching in videos. In: Proceedings of ICCV. Volume 2., Nice, France (2003) 1470–1477

[2] Bosch, A., Zisserman, A., Munoz, X.: Scene classification via pLSA. In: Proceedings of ECCV. (2006)

[3] Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., Van Gool, L.: A comparison of affine region detectors. International Journal of Computer Vision **65**(1-2) (2005) 43–72

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

[5] Tirilly, P., Claveau, V., Gros, P.: A review of weighting schemes for bag of visual words image retrieval. Technical Report 1927, IRISA, Rennes, France (April 2009)

[6] Yang, J., Jiang, Y.G., Hauptmann, A.G., Ngo, C.W.: Evaluating bag-of-visual-words representations in scene classification. In: Proceedings of the international workshop on Multimedia Information Retrieval, New York, NY, USA, ACM (2007) 197–206

[7] Jiang, Y.G., Ngo, C.W., Yang, J.: Towards optimal bag-of-features for object categorization and semantic video retrieval. In: Proceedings of CIVR, New York, NY, USA, ACM (2007) 494–501

[8] Jiang, Y.G., Ngo, C.W.: Visual word proximity and linguistics for semantic video indexing and near-duplicate retrieval. Computer Vision and Image Understanding (2008)

[9] Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Lost in quantization: Improving particular object retrieval in large scale image databases. In: Proceedings of the international conference on Computer Vision And Pattern Recognition (CVPR). (2008)

[10] Chen, X., Hu, X., Shen, X.: Spatial weighting for bag-of-visual-words and its application in content-based image retrieval. In: Proc. of the Pacific-Asia Conference on Knowledge Discovery and Data Mining. (2009) 867–874

[11] Nister, D., Stewenius, H.: Scalable recognition with a vocabulary tree. In: Proceedings of CVPR, Washington, DC, USA, IEEE Computer Society (2006) 2161–2168

[12] Jegou, H., Douze, M., Schmid, C.: Hamming embedding and weak spatial consistency for large scale image search. In: Proceedings of ECCV. (2008)

[13] Jegou, H., Harzallah, H., Schmid, C.: A contextual dissimilarity measure for accurate and efficient image search. In: Proceedings of CVPR. (2007) 1–8

[14] Aggarwal, C.C., Hinneburg, A., Keim, D.A.: On the surprising behavior of distance metrics in high dimensional space. In: Lecture Notes in Computer Science, Springer (2001) 420–434

[15] Robertson, S.: The probability ranking principle in information retrieval. Journal of documentation **33** (1977) 294 – 304

[16] Jones, K.S., Walker, S., Robertson, S.E.: A probabilistic model of information retrieval: development and comparative experiments (part 2). Information Processing and Management **36**(6) (2000) 809–840

[17] Jones, K.S., Walker, S., Robertson, S.E.: A probabilistic model of information retrieval: development and comparative experiments (part 1). Information Processing and Management **36**(6) (2000) 779–808

[18] Amati, G., Van Rijsbergen, C.J.: Probabilistic models of information retrieval based on measuring the divergence from randomness. ACM Transactions on Information Systems **20**(4) (2002) 357–389

[19] Chum, O., Philbin, J., Sivic, J., Isard, M., Zisserman, A.: Total recall: Automatic query expansion with a generative feature model for object retrieval. In: Proceedings of ICCV, Rio De Janeiro, Brazil (2007)

[20] Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2007)

[21] Zheng, Q.F., Wang, W.Q., Gao, W.: Effective and efficient object-based image retrieval using visual phrases. In: Proceedings of ACM Multimedia, New York, USA, ACM (2006) 77–80

[22] Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories. In: CVPR, Workshop on Generative-Model Based Vision. (2004)

[23] Barros, J., French, J., Martin, W., Kelly, P., Cannon, M.: Using the triangle inequality to reduce the number of comparisons required for similarity-based retrieval. In: Proc. of SPIE/IS&T Conf. on Storage and Retrieval for Image and Video Databases IV, SPIE (1996) 392–403

Figure 2: Similar images of a scene recognition task (Nister dataset)

Figure 3: Similar images of a category retrieval task (Caltech101 dataset)

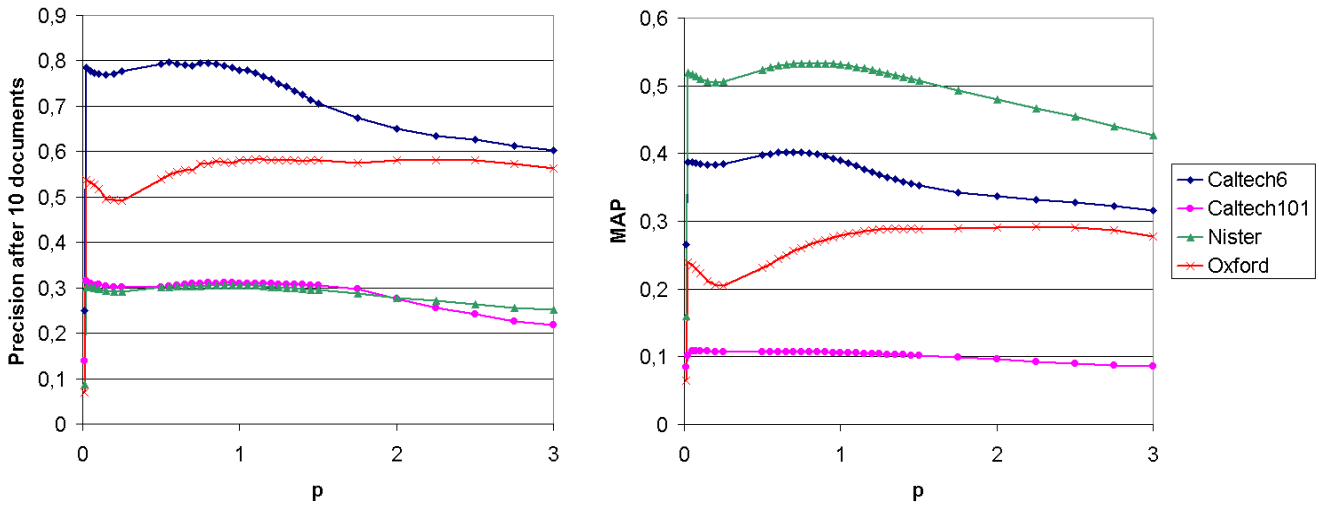

Figure 4: Retrieval performance for different values of k with a TF*IDF weighting scheme

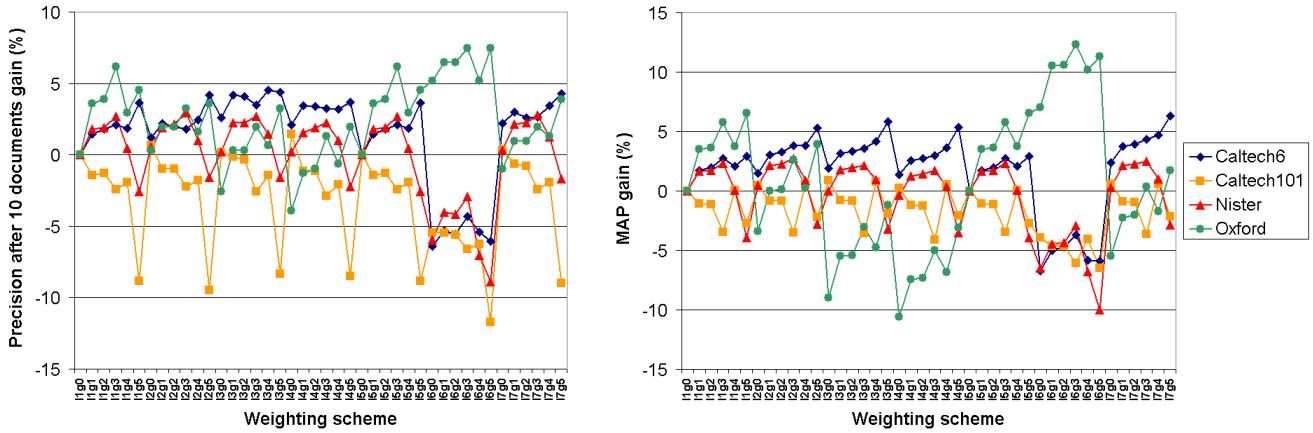

Figure 5: Results using L1 distance and the weighting schemes from Tables 2 and 3

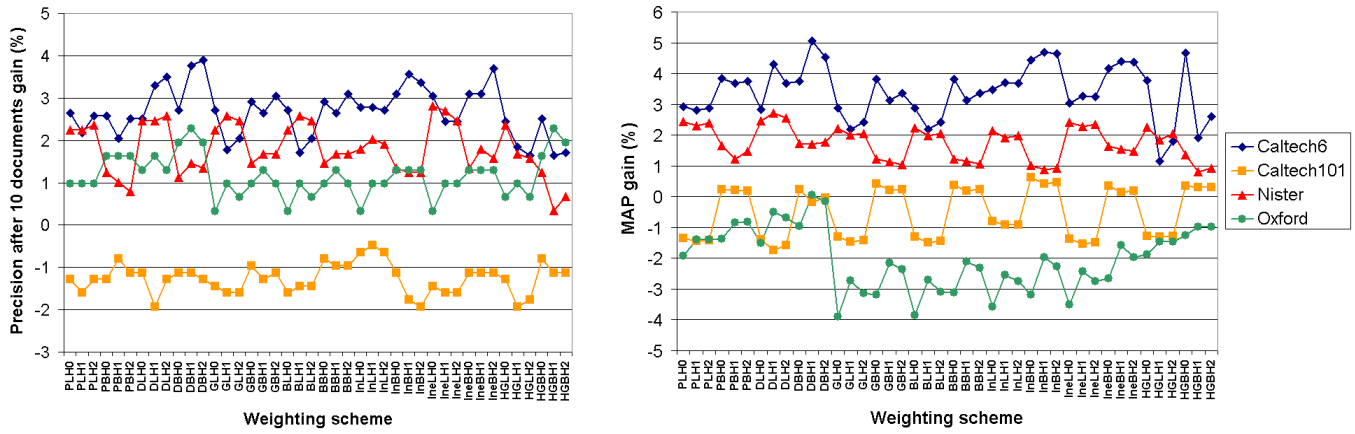

Figure 6: Results using L1 distance and the DFR weighting schemes

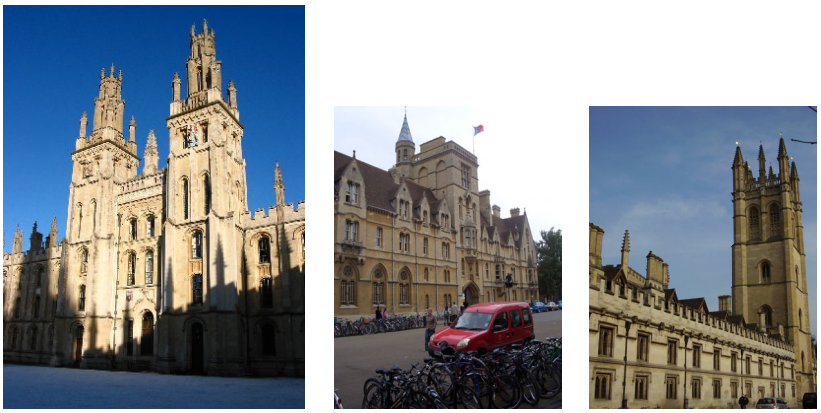

Figure 7: Some queries from Oxford dataset. They contain many repeated parts such as windows.