From Text Vocabularies to Visual Vocabularies
What Basis?

Jean Martinet
LIFL/CNRS-UMR 8022, Université Lille 1, Lille, France

Keywords: Bag-of-Features, Quantization, Vocabulary, Zipf's Law, Evaluation.

Abstract: The popular ”bag-of-visual-words” approach for representing and searching visual documents consists in describing images (or video keyframes) using a set of descriptors, that correspond to quantized low-level features. Most of existing approaches for visual words are inspired from works in text indexing, based on the implicit assumption that visual words can be handled the same way as text words. More specifically, these techniques implicitly rely on the same postulate as in text information retrieval, stating that the words distribution for a natural language globally follows Zipf's law – that is to say, words from a natural language appear in a corpus with a frequency inversely proportional to their rank. However, our study shows that the visual words distribution depends on the choice of low-level features, and also especially on the choice of the clustering method. We also show that when the visual words distribution is close to this of text words, the results of an image retrieval system are increased. To the best of our knowledge, no prior study has yet been carried out to compare the distributions of text words and visual words, with the objective of establishing the theoretical foundations of visual vocabularies.

1 INTRODUCTION

In image retrieval as in information retrieval in general, the quality of document representation is obviously central for the system efficiency. Popular bag-of-visual-words approaches consist in representing images with sets of descriptors, that correspond to quantized low-level features. The features are locally extracted from image regions, yielding a large number of feature vectors – which are the equivalent to **text words**.

A *quantification* step enables to reduce this number by gathering vectors in clusters into the feature space, and associating to each vector a representative in a discrete set – equivalent to **index terms**. Images can therefore be represented as sets, histograms, or vectors of such terms, which play the role of an indexing vocabulary.

This representation is inspired from the *bag-of-words* approach in text indexing and information retrieval, where text documents are represented as sets of terms taken from a vocabulary, that is built using the corpus.

Since almost a decade, most content-based image retrieval systems are based on this now popular representation. A number of classical text indexing techniques such as term selection or term weighting are directly applied to visual vocabularies, so is the matching method with set intersection or vector comparison.

However, visual and textual vocabularies are generated with processes of intrinsically different natures, and to our knowledge, no study has yet dealt with validating the direct application of text techniques to images. The objective of this work is to compare text and visual vocabularies, in order to study and clarify, according to some criteria, the applicability of text techniques to visual vocabularies.

The remainder of this paper is organized as follows: Section 2 presents generalities about text information retrieval, and implicit hypotheses on which traditional techniques in this domain are founded. Section 3 presents the visual bag-of-words paradigm, based on analogies between text and image. This section also gives details about widely used low-level features and clustering algorithms that we select in our work. Section 4 presents the study of visual vocabularies: it describes the vocabularies with the different generation steps, the experimental conditions, and the results. We give a conclusion to this work and mention about future research in Section 5.

Martinet J..
From Text Vocabularies to Visual Vocabularies - What Basis?.
DOI: 10.5220/0004749606680675
In *Proceedings of the 9th International Conference on Computer Vision Theory and Applications* (VISAPP-2014), pages 668-675
ISBN: 978-989-758-004-8

2 TEXT INFORMATION RETRIEVAL

Text Information Retrieval (IR) approaches are developed since about 40 years, and they are based on keywords found in the text collection(Baeza-Yates and Ribeiro-Neto, 1999).

The advantage of these approaches is namely due to their being efficient and fast, as it can be noticed with Web search engines that are able to retrieve very quickly documents among hundreds of millions (Brin and Page, 1998).

2.1 Classical Processing in Text IR

Among existing IR models, the Vector Space model (Salton, 1971) gives good results from atomic pieces of knowledge such as keywords widely used for text indexing. A document content is expressed as a group of keywords considered representative of its content, the keywords are taken from an *indexing vocabulary*. The indexing vocabulary contains all terms used for documents description; it is built from words found in the whole corpus, after possible operations and processing described below.

2.1.1 Stop-word Filtering and Anti-dictionaries

Only words conveying semantics are usually kept in the indexing vocabulary: the other so-called *stop-words* (such as ”of”, ”the”, ”for” in english), stored in an *anti-dictionary*, are not part of the indexing vocabulary. In the same way, rare words are generally removed from the indexing vocabulary.

2.1.2 Word Stemming

Besides, it is frequent to *stem* words, that is to say to detect word spelling/lexical variations, such as plural or verb conjugation, and to consider only a common root, or semantical unit (Salton and McGill, 1983) as an indexing term. For instance, the words ”writers”, ”writing”, and ”written” have the common root ”writ”, and only this root would be kept in the vocabulary.

2.1.3 Term Weighting

Document descriptors can all be considered having equal importance, i.e. being all equally representative of the document semantics: solely the occurrence (or absence) of a term in a document matters. However, the descriptors do generally not have the same importance in documents. Hence, descriptors are weighted in an index according to how well they describe the document, in order to express the relative importances and to refine the indexing.

The popular $TF \times IDF$ weighting scheme (Salton and Buckley, 1988) and its probabilistic variation $BM25$ (Jones et al., 2000) consist in giving more importance to terms that are frequent in given documents (TF) and also not frequent in the whole corpus (IDF).

2.2 Zipf’s Law and Luhn’s Model

A central aspect of text IR approaches is that they have as a foundation important characteristics of the words distribution in a natural language. Zipf’s law links word frequencies in a language to their ranks, when ordered in decreasing frequency order (Zipf, 1932). This law stipulates that that words occurrences follow this distribution model:

$$P_n = \frac{1}{n^a} \tag{1}$$

where P_n is the occurrence probability of the word at rank n, and a is a value close to 1. For instance, for the english language, the term ”the” would be the most frequent (it generally represents about 7% of the total word count) in a large text collection. The second most frequent would be ”of”, with 3,5% of the total, etc. As an illustration, Figure 1 shows the word distribution from Wikipedia articles in November 2006 in a log-log (bi-logarithmic) plot, which reveals the logarithmic distribution.

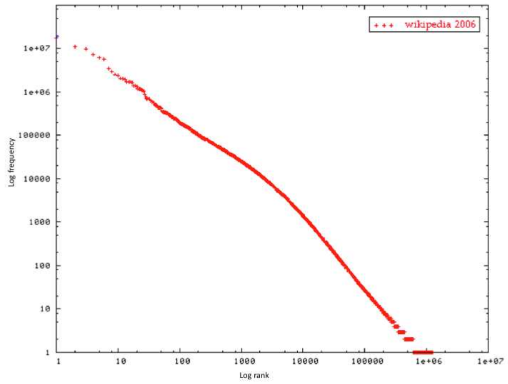

Figure 1: Word distribution from Wikipedia article in November 2006 (Source : Wikipdia).

This logarithmic distribution, interpreted through information theory (Shannon, 1948), has namely made it possible to establish the theoretical foundations of word filtering and weighting schemes. In particular, the selection techniques for significant terms are mainly grounded on hypotheses from Luhn’s model (Luhn, 1958), and this model originates from Zipf’s law.

This model indicates a relation between the rank of a word and its *resolving power*, or *discriminative power*, that is to say its capacity of identifying relevant documents (notion of recall) combined with its capacity of distinguishing non relevant documents (notion of precision); this relation is illustrated Figure 2. Hence, less discriminative words are those at low ranks – very frequent, and those at high ranks – very rare. More discriminative words are those located in-between, therefore these terms should be selected for the indexing vocabulary.

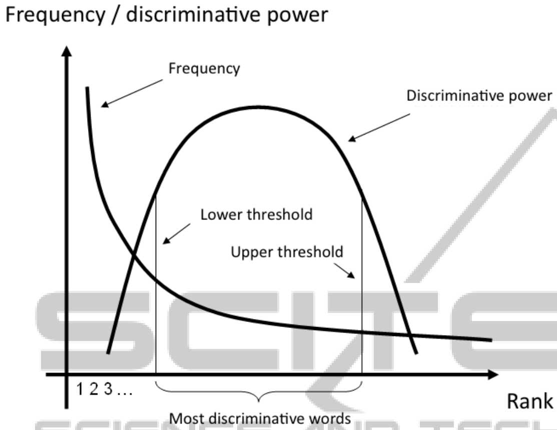

Figure 2: Zipf's law and Luhn's model.

3 BAGS OF VISUAL WORDS

In a trend similar to text documents, approaches of bags of visual words (Sivic and Zisserman, 2003; Jurie and Triggs, 2005; Grauman and Darrell, 2005; Lazebnik et al., 2006) represent images with sets of descriptors, corresponding to quantized low-level features. Such approaches are founded on the obvious analogy between text methods and visual methods.

3.1 Low-level Descriptors

Among low-level descriptors, SIFT (Scale-Invariant Feature Transform) (Lowe, 2004) and SURF (Speeded-Up Robust Features) (Bay et al., 2008) are widely used; they represent an image with a set of scale invariant local features. Both include an interest point detector, and a descriptor for such points.

3.1.1 SIFT

SIFT detector (Lowe, 2004) uses a *Difference of Gaussians* (DoG) filter to approximate the *Laplacian of Gaussians* (LoG), to accelerate processing. This interest point detector is scale invariant. The descriptor calculates a local histogram of oriented gradient in the neighborhood of the interest point, and stores the bins in a 128-dimension vector (8 orientation bins for each 4×4 position bins). A specificity of SIFT is that it generates a large number of features, spreading over the image with different scales and locations. An image with resolution 500×500 pixels typically generates about 2000 stables features.

3.1.2 SURF

The more recent detector-descriptor SURF (Bay et al., 2008) is based on a Hessian matrix for detecting interest points, that is estimated from integral images. The descriptor calculates the distribution of Haar wavelet responses in the neighborhood of interest points, stored in a 64-dimension vector.

3.2 Vector Quantization

The generation of a visual vocabulary requires to *cluster* all features extracted from the image collection. This step of *vector quantization* consists in representing each feature by the centroid of the cluster it belongs.

3.2.1 Vocabulary Size

In previous works, there has been many different attempts with several vocabulary sizes. For example, (Sivic and Zisserman, 2003) used a vocabulary size of range 6000-10000. Later, other researchers used a wide span of vocabulary sizes, e.g. (Csurka et al., 2004) who used the range 500-2500, (Nowak et al., 2006) who tried 300, 1000, and 4000, (Lazebnik et al., 2006) used 200-400, Zhang et al. (Zhang et al., 2007) used 1000, and Li et al. (Li et al., 2008) in 2008 used a vocabulary size less than 1500. Ballan et al. (Ballan et al., 2012) have addressed the problem of vocabulary generation in the context of video (i.e. with space-time dimensions) by using radius-based clustering with soft assignment. They have further performed vocabulary compression in order to avoid a too large size. There is no consensus which size or range of vocabulary size performs best. In our study, we have set the vocabulary size to be 10000.

3.2.2 Clustering with KMeans

KMeans is certainly the most widely used clustering algorithm. From a set of *k* centroids (generally initialized randomly), this iterative partitioning algorithm assigns to each point the closest centroid, and re-estimates the center of each group in order to determine new centroids. The algorithm stops when a convergence criterion is satisfied, and the final partitioning determines the *mapping* between points and centroids.

3.2.3 Clustering with SOM

Much less known are the Self-Organizing Maps (SOM) (Kohonen, 2001; Xu et al., 2007), a kind of artificial neural network, that are used for non-supervised classification. From data in a large dimension real space, SOM generates a space mapping by assigning to each point a representative after a learning step. This partitioning of data enables to gather points into clusters.

We chose to compare KMeans and SOM because of their intrinsic difference in both processing the data and behavior according to the type and size of data (Abbas, 2008).

3.3 Text Techniques Applied to Images

Once the visual vocabulary is created, images can be represented with sets, histograms, or vectors of visual terms. A great amount of work in image representation and indexing are inspired from text techniques, such as term filtering (Sivic and Zisserman, 2003; Tirilly et al., 2008), where visual terms with low interest are removed from the indexing vocabulary. Besides, visual terms can be weighted to represent their importance. For instance, the system SIMPLIcity (Wang et al., 2001) implements a weighting scheme named $RF \times IPF$ (*Region Frequency* et *Inverse Picture Frequency*), directly inspired from the widely known $TF \times IDF$ for text. Weights are given to image regions according to their frequency in the image and in the collection. Works in (Sivic and Zisserman, 2003; Tirilly et al., 2008; Philbin et al., 2008) also adapt this weighting scheme to images.

If we go further into the text-image analogy, several works (Bosch et al., ; Tirilly et al., 2008; Wu et al., 2009) have adapted *language models* to images. Some works even implement query expansion techniques (Philbin et al., 2008) (originally designed for text), by automatically adding selected visual terms to the original query in order to get better search results.

4 VISUAL VOCABULARY STUDY

This section describes the experiments made in order to compare the considered visual vocabularies.

4.1 Terminology Distinction

We start by introducing the following terminology distinction. Following the text-image analogy, we call *visual words* the image regions or local areas originating from segmentation, tessellation, or regions in the neighborhood of interest points.

Therefore, visual words correspond to text words, that is to say: occurrences of *terms*, with lexical variations. *Visual terms* correspond to text terms, that is to say: elements of the indexing vocabulary.

As a consequence, a visual word appearing in an image is an occurrence of a visual term, and the visual variations of regions assigned to a given cluster can be seen as the visual counterpart of lexical and syntactic variations for text.

4.2 Description of Considered Vocabularies

The two above-described features (SIFT and SURF), combined with the two clustering algorithms (KMeans and SOM) enable to generate four visual vocabularies: KMSIFT, KMSURF, SOMSIFT, SOMSURF. We aim at comparing these vocabularies according to their term distributions, and according to the results obtained in an image search task.

4.3 Image Collections

We selected Caltech-101 and Pascal image collections for this study. Sample images from these collections are given Figure 3. Caltech-101 database[1] contains 101 image categories, with about 50 images per category (varying from 40 to 800). We have randomly selected 10 categories in our experiments, resulting in 1345 images.

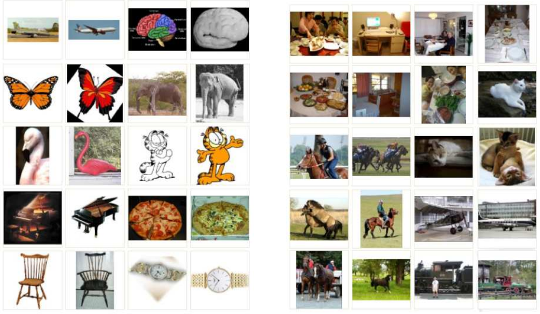

Figure 3: Image samples from Caltech-101 (left) and Pascal (right).

[1] Caltech-101, see URL : http://www.vision.caltech.edu/Image_Datasets/Caltech101/.

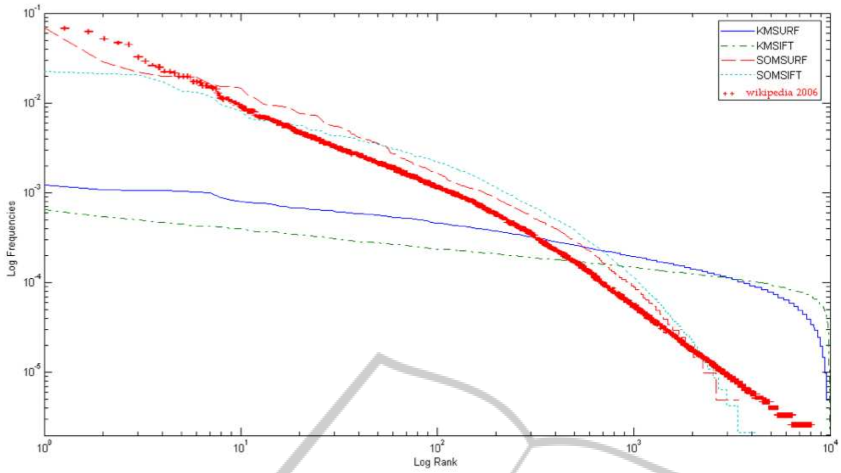

Figure 4: Vocabularies distributions for Caltech-101.

Pascal database[2] contains 20 image categories, and we selected 1000 images uniformly sampled from all categories in our experiments.

The average processing time for visual words extraction is 3h13min with a 2.93GHz Intel Xeon processor-based machine, with 4GB of RAM. The required time is slightly longer for SIFT than for SUFT. We note that SIFT generates about twice as much visual words than SURF for both databases. This is because the *Difference of Gaussians* detector extracts much more interest points than the *Hessian matrix* detector.

Table 1 shows the total number of visual words generated for each database and each descriptor, along with the average number of words per image.

Table 1: Overview of the generated visual words.

Database (# images)	Caltech-101 (1345)		Pascal (1000)	
Feature	SIFT	SURF	SIFT	SURF
# words in total	463K	204K	297K	174K
# words/image	344	151	297	174

4.4 Visual Words Distributions

Visual vocabularies have been generated for each database, using alternatively KMeans and SOM for clustering. As mentioned in Section 3.2, the obtained vocabularies have a size of 10K terms.

The generation time required to generate the vocabularies are drastically different: 4.5 days for KMeans, and 2.5h for SOM with the above-described machine, using a Matlab environment. Note that using compiled $C/C++$ implementations is likely to speed up the processing time for both algorithms.

Figures 4 and 5 show the distributions for all vocabularies for both databases, along with the words distribution from Wikipedia, as shown Figure 1. It can be easily noticed that for both databases, KMSIFT and KMSURF vocabularies have rather flat distributions, compared with SOMSIFT and SOMSURF. This indicates that KMeans generates vocabularies that are more uniformly distributed than SOM. In other words, it means that the cluster sizes are more similar to one another with KMeans than with SOM.

However, the distributions of SOM vocabularies are much similar to the english words distribution (Wikipedia articles in 2006). This tendency is the same for both databases.

We now focus on the slope of the vocabularies distributions in the log-log plot. For this purpose, we have estimated the slope of the trend line calculated using linear regression, with the least squares criterion.

The slopes are given Table 2. The results confirm that SOM-generated vocabularies are more similar to those of natural language than KMeans-generated ones. The results also indicate that the choice of detector-descriptor does not seem to greatly impact the vocabularies distribution.

Table 2: Estimated slopes (in degrees) for visual distributions (slope for Wikipedia: -24.2°).

Database	Caltech-101		Pascal	
Descriptor	SIFT	SURF	SIFT	SURF
KMeans	-7.1°	-6.3°	-8.2°	-9.8°
SOM	-19.0°	-21.5°	-18.7°	-20.1°

In order to explain this, it is useful to take a look at the different clustering processes: KMeans randomly

[2] Visual Object Classes Challenge 2009, see URL : http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2009/.

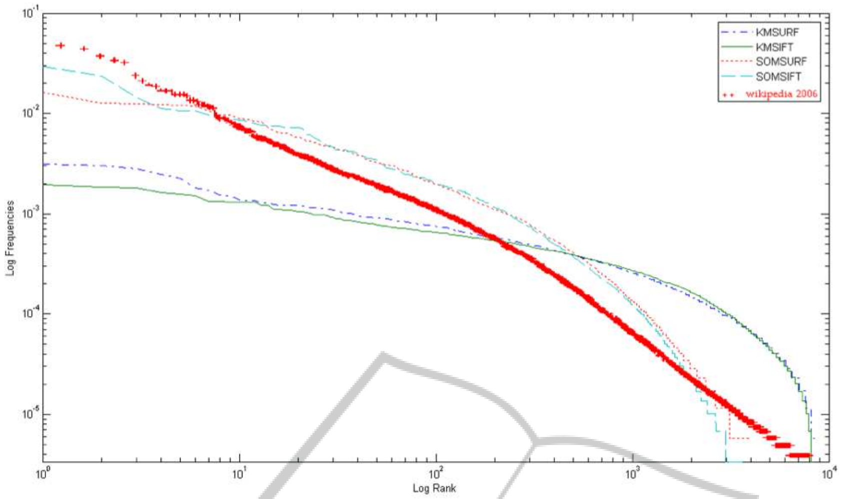

Figure 5: Vocabularies distributions for Pascal.

initializes cluster centroids, and a convergence state is searched by progressively shifting centroids. A situation in which two centroids would be both initially placed inside a given group of points close to one another (thus a good candidate for a *single* cluster) is likely to happen.

This way, instead of producing a single cluster reflecting the reality of the group of points, KMeans would artificially produce two ill-formed clusters, as illustrated Figure 6. Because the cluster generation process is fundamentally different with SOM, this algorithm does not encounter this problem.

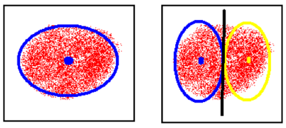

Figure 6: Illustration of clusterings: correct (left) and not correct (right) for KMeans, after different initializations.

In the remainder of this section, we focus on results obtained with the vocabularies in an image search task.

4.5 Image Search Task

The second step in our work consists in evaluating the precision that one can reach with each vocabulary in an image search task. For the purpose of this evaluation, we have implemented an image retrieval system based on the IR vector space model, integrating filtering and $TF \times IDF$ weighting scheme for visual terms. For both databases, each image is used in turn as a query, and the *ground truth* is made of all images in the same category as the query. Recall-precision curves are given Figures 7 and 8, for Caltech-101 and Pascal, respectively.

We notice from the curves that the search precision seems to be higher for Caltech-101 than for Pascal. This can be explained by the nature of images in the database: while Caltech-101 contains focused images generally showing a single well identified object with a simple background, Pascal contains less focused images often containing several objects, with more complex backgrounds.

In Figure 7, we can see noticeable differences in the precision values: SOMSURF yields a better precision than other vocabularies. Besides, KMSURF yields lower results than others. This point confirms the low influence of the detector-descriptor. The curve for SOMSIFT is slightly above KMSIFT, but the difference is too small to be significant.

In Figure 8, the results are similar for all vocabularies; the differences in the curves are much smaller than those shown Figure 7. Even though the tendency of better results with SOMSURF and lower results with KMSIFT can be noticed, the difference is not significant and too small to draw conclusions. However, we note that SIFT is less sensitive than SURF to the choice of clustering method.

As a general comment, even though KMeans is more popular than SOM, the results of our image retrieval system with SOM-generated vocabularies are better than KMeans-generated ones. According to our study, a theoretical justification could be that SOM generates vocabularies whose distributions are more similar to those of natural language words. By that very fact, the implicit hypotheses upon which text techniques are established (Zipf's law, Luhn's model) are satisfied in this case for images, which would a theoretical validation explaining the better results.

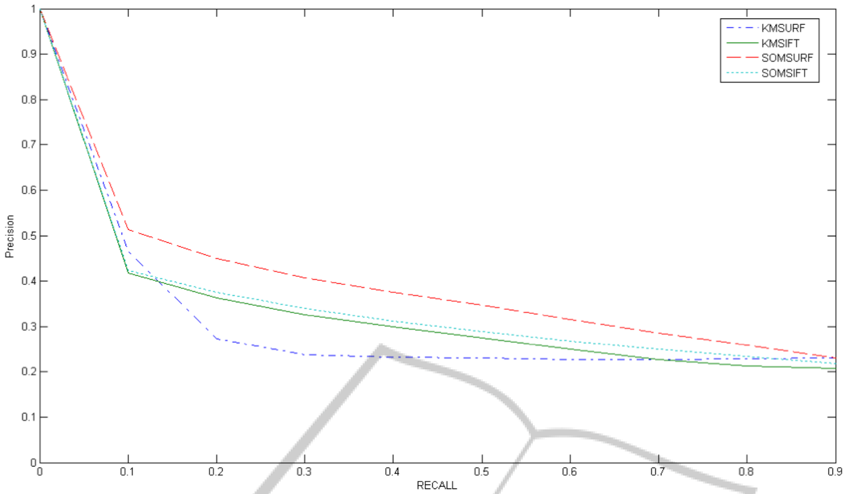

Figure 7: Recall-precision curves (Caltech-101).

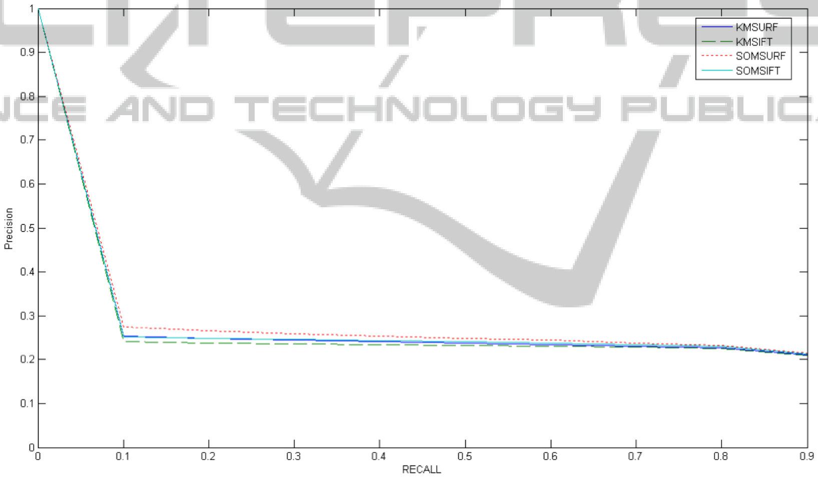

Figure 8: Recall-precision curves (Pascal).

5 CONCLUSIONS

Most of text IR techniques are implicitly founded on the postulate that natural language words globally follow Zipf's law, and on Luhn's model. Applying such techniques to visual vocabularies built from images requires prior verifications of these postulates validity.

In this paper, we have presented a study about visual vocabularies compared to text vocabularies, in order to clarify conditions for applying text techniques to images. Our study showed that visual words distributions highly depend the clustering method – the influence of the choice for low-level features is limited. Indeed, the choice for the visual terms generation method is important, and determines the vocabulary distribution. We also show that when the visual words distribution is close to text words distributions, the results of an image retrieval system are increased. Therefore, SOM clustering method, that generates distributions similar to natural languages, yields better search precision than the yet popular KMeans. To our knowledge, no prior study has yet been carried out aiming at validating the application of text techniques for images.

Our future works will focus on generalizing this comparative approach to other popular low-level descriptors in image representation. We also wish to get insights from the consequences of our results when varying the size of the vocabularies. Besides, we work toward making a more formal comparison of distributions with dedicated statistical tools. This work has been done in a global initiative of applying text techniques to images.

REFERENCES

Abbas, O. A. (2008). Comparisons between data clustering algorithms. *International Arab Journal of Information Technology (IAJIT)*, 5(3):320.

Baeza-Yates, R. A. and Ribeiro-Neto, B. A. (1999). *Modern Information Retrieval*. Addison-Wesley.

Ballan, L., Bertini, M., Del-Bimbo, A., Seidenari, L., and Serra, G. (2012). Effective codebooks for human action representation and classification in unconstrained videos. *IEEE Transactions on Multimedia*, 14(4):1234–1245.

Bay, Ess, A., Tuytelaars, T., and Gool, L. V. (2008). Surf: Speeded up robust features. *Computer Vision and Image Understanding (CVIU)*, 110(3).

Bosch, A., Zisserman, A., and Muoz, X. Scene classification via plsa. In *ECCV'06*, pages 517–530.

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. In *In journal of Computer Networks and ISDN Systems (30)*, pages 107–117, Brisbane.

Csurka, G., Dance, C. R., Fan, L., Willamowski, J., and Bray, C. (2004). Visual categorization with bags of keypoints. In *In Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22.

Grauman, K. and Darrell, T. (2005). The pyramid match kernel: Discriminative classification with sets of image features. In *ICCV*, pages 1458–1465.

Jones, K. S., Walker, S., and Robertson, S. E. (2000). A probabilistic model of information retrieval: development and comparative experiments. *Inf. Process. Manage.*, 36:779–808.

Jurie, F. and Triggs, B. (2005). Creating efficient codebooks for visual recognition. In *ICCV*, pages 604–610.

Kohonen, T. (2001). *Self-Organizing Maps*. Springer Series in Information Sciences, vol 30.

Lazebnik, S., Schmid, C., and Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR (2)*, pages 2169–2178. IEEE Computer Society.

Li, T., Mei, T., and Kweon, I. S. (2008). Learning optimal compact codebook for efficient object categorization. In *Proc. of WACV*, pages 1–6.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110.

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research Development*, 2(2):159165.

Nowak, E., Jurie, F., and Triggs, B. (2006). Sampling strategies for bag-of-features image classification. In *In Proc. ECCV*, pages 490–503. Springer.

Philbin, J., Chum, O., Isard, M., Sivic, J., and Zisserman, A. (2008). Lost in quantization: Improving particular object retrieval in large scale image databases. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1 –8.

Salton, G. (1971). *The SMART Retrieval System*. Prentice Hall.

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. In *Information Processing and Management*, pages 513–523.

Salton, G. and McGill, M. (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656.

Sivic, J. and Zisserman, A. (2003). Video google: A text retrieval approach to object matching in videos. In *ICCV*, pages 1470–1477.

Tirilly, P., Claveau, V., and Gros, P. (2008). Language modeling for bag-of-visual words image categorization. In *CIVR'08*, pages 249–258.

Wang, J., J.L., and Wiederhold, G. (2001). SIMPLIcity: Semantics-sensitive Integrated Matching for picture LIbraries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(9):947–963.

Wu, L., Hu, Y., Li, M., Yu, N., and Hua, X.-S. (2009). Scale-invariant visual language modeling for object categorization. *Multimedia, IEEE Transactions on*, 11(2):286 –294.

Xu, X., Zeng, W., and Zhao, Z. (2007). A structural adapting self-organizing maps neural network. In *ISNN (2)*, pages 913–920.

Zhang, J., Lazebnik, S., and Schmid, C. (2007). Local features and kernels for classification of texture and object categories: a comprehensive study. *International Journal of Computer Vision*, 73:2007.

Zipf, G. K. (1932). *Selective Studies and the Principle of Relative Frequency in Language*. Addison- Wesley, Cambridge, MA, USA.