Document description: what works for images should also work for text? *

Nicolas Hervé[a], Nozha Boujemaa[a], Michael E. Houle[b]

[a]INRIA Paris-Rocquencourt, Domaine de Voluceau, BP 105, 78153 Le Chesnay Cedex, France;
[b]National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430,Japan

ABSTRACT

The success of the bag-of-words approach for text has inspired the recent use of analogous strategies for global representation of images with local visual features. Many applications have been proposed for object detection, image annotation, queries-by-example, relevance feedback, automatic annotation, and clustering. In this paper, we investigate the validity of the bag-of-words analogy for image representation and, more specifically, local pattern selection for feature generation. We propose a generalized document representation framework and apply it to the evaluation of two pattern selection strategies for images: dense sampling and point-of-interest detection. We present empirical results that support our contention that text-based experimentation can provide useful insights into the effectiveness of image representations based on the bag-of-visual-words technique.

Keywords: bag-of-words, visual vocabulary, image representation, vector space model, TF-IDF, image retrieval, Reuters RCV1, ImagEVAL

1. INTRODUCTION

Datasets such as the ones managed by professional news agencies, family and holiday photo albums, and, more generally, images accessible via the Internet, all appear to have almost infinite variations in scenes, objects, and technical shooting conditions. Even in the most favorable situations, the automatic annotation of images is a very challenging task; however, for general image collections, it must be admitted that the state-of-the-art image annotation techniques are still inadequate. Much progress has been made in recent years, but the problem is still far from being solved.

We are interested here in generic frameworks for determining a global set of visual features for image data sets. Much as in the case of text keywords for documents, a collection of visual features would ideally serve as a form of visual vocabulary that would provide a common basis for description and comparison of images. These 'visual keywords' would need to relate to the visual content of individual images, and yet be sufficiently general so as to allow the discovery of similar content across subcollections of images. The first step in the development of a visual vocabulary is thus to produce candidate visual keywords by means of analysis of the images and the extraction of low-level visual features.

One can distinguish two main families of visual features. Global features are computed over the full image, and are usually well suited for describing the nature of the image (such as whether it is a photograph, a drawing, or an artistic representation, etc.) and the context of the image (for example, whether it is indoor or outdoor, day or night, natural or urban, etc.)[1] Local features are computed only on a portion of the image (usually quite small), and are generally used for object detection. Often, a region-of-interest detector is used to determine the portions of the images from which local features will be extracted. Segmentation[2] and point-of-interest detection[3,4] are the two main strategies used for this purpose. With global features, the image representation is generally straightforward. However, even when local features are to be used, applications such as learning algorithms may sometimes require a global image representation that encompasses all the local visual information. The bag-of-visual-words representation, very much inspired by the classical bag-of-words representation for text, is one of the most popular representation for images. A visual vocabulary composed of visual words

E-mail: nicolas.herve@inria.fr, nozha.boujemaa@inria.fr, meh@nii.ac.jp

is generated. An image is then represented by a coordinate vector, each value of which expresses the degree of importance of a pattern with respect to the image and/or the database as a whole.

One standard method for generating a visual vocabulary begins with the detection of points of interest within images of a training set, and the characterization of these points according to low-level visual features. Quantization of the points of interest is then performed by means of a clustering, with each cluster giving rise to a visual word. Each training set image is then associated with a histogram of the visual word frequencies. These image representations are used to train a classifier that will produce models for all the potential concepts we wish to detect. Using the vocabulary and models obtained, one is then able to predict concepts on new images.

In this paper, we will investigate the effectiveness of visual vocabularies produced by using point-of-interest detection for local feature generation. The design choices made in the creation and use of such visual vocabularies will be assessed according to the performance of analogous design choices for text data. Another motivating factor for assessing the performance relative to text is that in practice, ground truth information is generally easier and cheaper to obtain for text data than it is for images.

Applying the bag-of-word strategy is inherently easier for text than for images. For text databases, the word boundaries are well defined, the vocabulary is known in advance, and is generic across all documents. Unlike the situation with images, with text one is certain that meaningful information is carried by the words of the vocabulary — indeed, the words of any language can be regarded as having been created precisely to convey such information. If the entire vocabulary is used, the document can be viewed as being fully covered by the set of local features (the keywords).

For images on the other hand, the use of points of interest constitutes a considerable restriction on the visual content covered by the visual vocabulary. The rationale behind this limitation to small point sets is the high computational cost inherent in the generation of local descriptors that fully cover all images. Analyzing and characterizing the data in the vicinity of all pixels, over all potential scales, is too expensive on current hardware; when geometrical relationships are sought, the cost is driven even higher. Standard computers in use today are generally unable to process the full volume of information present in images, although research into scalable algorithms or the development of faster hardware may someday lead to a reduction in computational cost.

In their desire to reduce image processing costs, the computer vision community has put much effort into the development of point-of-interest detectors, beginning with their original applications for image registration. Detectors are generally attracted to specific areas of images that have high variation in the visual signal, such as the vicinity of edges and corners of regions. One of the most important considerations in point-of-interest detection is its repeatability, so that the locations of the points of interest selected from different images of a common object photographed in different conditions will be in alignment. When trying to limit the amount of information processed in the automatic annotation of images, the detection of points of interest is an attractive option, as it allows the selection of a very small proportion of image locations having the highest visual variance.

High visual variance is often considered to be associated with high semantic content. However, this assumption is not always justified, as low variance may also carry important semantic information in generic image datasets. We therefore believe that all areas of an image must be eligible to generate words of the visual vocabulary, regardless of whether they exhibit high variation or low variation. It is known that contextual information has a positive effect when performing object detection.[1,5]

This paper is organized as follows. We first introduce a generalized bag-of-words representation framework that can be applied to both image data and text data. In order to assess the likely performance of a given representation strategy for images, we construct an analogy between the image representation and a representation of a degraded text dataset within which many of the characteristics of image sets have been reproduced. We are not seeking a formal validation of the studied approaches on images, but a better understanding of the mechanisms involved. If after establishing that search performance for the degraded text representation is inherently easier than for the image representation, good performance of search for standard text representations and poor performance on the degraded text dataset would together indicate that the bag-of-words strategy for images cannot be expected to perform well. On the other hand, good performance on the degraded text dataset would indicate that the particular bag-of-words strategy has the potential to perform well for images. In this paper, we will provide experimental results with both positive and negative examples that justify the assessment of bag-of-words image representations by means of an analogy to text.

2. GENERALIZED DOCUMENT REPRESENTATION FRAMEWORK

The generalized framework proposed in this section is designed so as to facilitate the analogy between bag-of-words representations for different document types. In particular, the framework will be instantiated for text and image cases. In this section, we will use the terms 'document', 'vocabulary', 'word' and 'pattern' in a generic sense, applicable to both images and text.

2.1 TF-IDF weighting

We start with a collection C containing m documents, and assume that each document D_k is composed of s_k patterns. A vocabulary V is an ensemble of n words selected from among the union of all patterns taken over all documents.

$$C = \{D_k, k \in [1, m]\} \tag{1}$$

$$D_k = \left\{P_j^k, j \in [1, s_k]\right\} \tag{2}$$

$$V = \{W_i, i \in [1, n]\} \tag{3}$$

In vector space modeling,[6] each document of the database is associated with a vector, each coordinate of which represents a word of the vocabulary. In a boolean model, each coordinate of the vector is zero (when the corresponding word is absent) or one (when the corresponding word is present). Many refinements of the boolean model exist. The most commonly used are term weighting models that may take into account the frequency of appearance of a word, its locations within the document, or the proportion of documents that contain the word. On of the most popular weightings, TF-IDF, depends on the frequency of the word within the document, and the rarity of the word within the document set.

In our framework, we use a slightly modified variant of TF-IDF weighting. As with standard TF-IDF, for a given word W_i, we define the document frequency Df as the proportion of documents of C in which W_i appears.

$$\mathrm{Df}(W_i) = \frac{\left|\left\{D_k, k \in [1, m], \exists j \in [1, s_k] \,|W_i = P_j^k\right\}\right|}{m} \tag{4}$$

The inverse document frequency is defined as

$$\mathrm{Idf}(W_i) = \log\left(\frac{1}{\mathrm{Df}(W_i)}\right) \tag{5}$$

For a given document D_k, we define the term frequency Tf of a word W_i as the proportion of patterns of D_k being equal to W_i.

$$\mathrm{Tf}(W_i, D_k) = \frac{\left|\left\{P_j^k, j \in [1, s_k] \,|W_i = P_j^k\right\}\right|}{s_k} \tag{6}$$

We can then define the TF-IDF weighting of word W_i for document D_k to be

$$\mathrm{TfIdf}(W_i, D_k) = \mathrm{Tf}(W_i, D_k) \times \mathrm{Idf}(W_i) \tag{7}$$

Finally, each document can be represented by a vector of dimension n containing the TF-IDF measures of all the words of the vocabulary with respect to the document. In this simple retrieval system, each query is also modeled by a vector in the same manner as the documents.

2.2 Document similarity

Although global document representations have a variety of applications, for most of them some measure of similarity between documents is needed. Similarity between document vectors, or between a query vector and a document vector, is very often measured according to the angle formed by the two vectors, or by its cosine value. The classical vector angle distance measure between two documents D_a and D_b is defined as:

$$d(D_a, D_b) = \arccos\left(\frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2 \sum_{i=1}^{n} b_i^2}}\right) \tag{8}$$

An angle approaching zero, or a cosine value approaching one, indicates that the pair of documents share many important words in common.

2.3 Evaluation of representations

The quality of our representations and their associated vocabularies will be evaluated according to the *query-by-example* paradigm, in which the performance of similarity queries based at documents of the data set is measured. For each query document, a set of relevant documents is provided as a ground truth for the retrieval task, with the average precision (AP) as the evaluation measure. The precision of a query is defined as:

$$P = \frac{\text{number of relevant documents retrieved}}{\text{number of documents retrieved}} \tag{9}$$

The average precision is the average, over all relevant documents, of the precision values obtained for the minimum-size query result containing that document. For a query Q, with r the rank, N the number retrieved, $\mathrm{rel}()$ a binary function on the relevance of a given rank, and $P()$ precision at a given cut-off rank, we have :

$$\mathrm{AP}_Q = \frac{\sum_{r=1}^{N}(P(r) \times \mathrm{rel}(r))}{\text{number of relevant documents}} \tag{10}$$

The MAP is the mean of the AP taken over a set of queries. As the number of documents retrieved affects the value of the AP, we standardize the result by setting the query result size to be twice that of the relevant set.

Other than the identification of query relevant sets, in our experimental settings we avoid the reliance on ground truth information such as would be provided by a classification of the data. This to stay as close as possible to the representation evaluation, and to avoid the stacking of too many technical components. For this reason, we do not permit learning strategies for the selection of visual features, or the use of any semantic information in the creation of vocabularies. Vocabularies are suggested by the local context of the document and are not specific to the concepts covered by the data set.

2.4 Degradation of text

The text domain has a much simpler structure than the image domain: text data can be regarded as a 1-dimensional stream of symbols, whereas image data is more naturally viewed as a 2-dimensional array. In the text domain, we need not be concerned with variances in scale, angle of view, deformability of objects, lighting conditions, and other properties characteristic of images. Identification of objects is simpler, and although polysemy also causes difficulties in the text domain, the semantic meaning of a text word is easier to discern than that of an image pattern. Thus, one would clearly expect that a given representation strategy for images would perform even better if analogously applied to text. However, the differences between text and image data are so great, that a method that works poorly for images may nevertheless work well for text. For comparisons of text and image representations to be of any benefit, the advantages of text relative to images must first be neutralized.

In text, the words are very well distinguished by the spaces and other punctuation symbols. On the other hand, the identification of objects and patterns present in images is a difficult, unresolved issue in computer vision for generic databases, In order to partially eliminate this advantage held by text over images, we remove all spaces and punctuation marks from all documents of text datasets, leaving a symbol set consisting of 36 characters (26 letters and 10 digits). With the natural vocabulary no longer available, bag-of-words approaches for degraded text datasets must build up a vocabulary from patterns within the data, in a manner analogous to the techniques for building visual vocabularies. The most basic patterns in degraded text would simply be sequences of characters, without any specific semantic meaning — analogous to, but still much simpler than, patterns associated with window regions or the vicinities of points of interest in images.

3. DATASETS

3.1 The Reuters dataset

The RCV1 Reuters corpus is composed of 806,791 English-language news articles collected over a one-year period, from 20 August 1996 to 19 August 1997.[7] The original vocabulary has 435,282 words. To avoid potential biasing of the retrieval results, we make use of only main bodies of the articles. The titles, categories, and other information are omitted (treated as metadata). We first convert all alphabetic characters to lower case, and all punctuation characters to spaces. We then delete any character that is neither alphanumeric nor a space. We remove stop words and perform stemming using the Porter

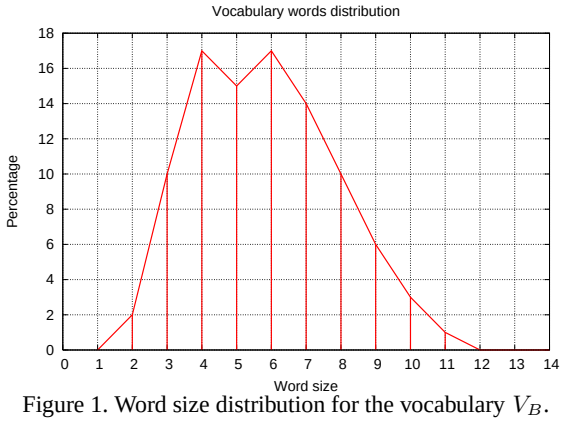

Figure 1. Word size distribution for the vocabulary V_B.

algorithm.[8] We keep numbers and 1-letter words (provided that they are not stop-words). This results in a vocabulary of 365,652 words. If we filter this vocabulary so as to retain only words present in at least 50 documents, 34,026 words remain. This vocabulary, which we will refer to as V_B, will be treated as a baseline for the experiments. For information, we show in Figure 1 the word size distribution for the vocabulary V_B.

For the experiments, we considered 7 query terms of varying lengths, selected so as to focus on a specific context or event in categories representative of the full dataset. Several of the terms were also chosen for their polysemous interpretations, so that contextual information would be required for the correct retrieval of ground truth documents. Two of the queries were composed of two keywords. One of them, the query *goldman + simpson*, focuses on a specific news event, the well-reported O.J. Simpson trial of 1995. Both of the keywords, if taken individually, have several meanings associated with them. Their polysemous nature can be seen from the low proportion of documents containing both keywords – only 4% – relative to the set of documents containing at least one of the two keywords. *Goldman* is part of the name of Goldman Sachs, a famous investment bank that often appears in the many Reuters news articles on financial matters. *Simpson* is the name of a desert in Australia, as well as the name of a Reuters journalist. The total number of documents containing the query keywords is given in Table 1. The experimental evaluation used these document subsets as the ground truth sets for the 7 queries. For each ground truth set, we perform a query by example based at each document, and compute the average precision (AP). With this treatment, the AP value measures the overall ability of the relevant documents to retrieve the remaining members of their set.

Table 1. Number of Reuters documents containing query keywords (after stemming).

Stemmed query	#Docs
nuclear	7654
goldman	6543
wto	2351
simpson	1888
greenpeac	713
arbil	587
goldman + simpson	337
zidan	194
greenpeac + nuclear	140
coulthard	125
kasparov	89

The MAP scores (average of the AP for each set) obtained using vocabulary V_B will serve as a baseline for comparison among other choices of vocabulary. All further results for a given query will be reported as a ratio of the score obtained for the same query using V_B. We also extracted two subsets of V_B for further comparisons. The first subset, V_1, is a vocabulary consisting of the 1296 words of V_B achieving the highest document frequency values. None of the query keywords is present in this small vocabulary. The second subset, V_2, is the vocabulary of all words of V_B composed of

exactly 5 characters, with the exception of the query keywords *arbil* and *zidan*. V_2 contains 5274 words. The results are reported in Table 2.

Table 2. MAP scores for text queries with respect to three vocabularies. The scores for V_1 and V_2 are expressed as proportions of the score for for V_B.

Stemmed query	V_B	V_1	V_2
goldman + simpson	0.2389	0.3110	0.3958
coulthard	0.2045	0.1284	0.5003
arbil	0.2014	0.2521	0.6577
kasparov	0.1234	0.1967	1.0969
wto	0.0897	0.2102	0.1008
zidan	0.0851	0.1131	0.2306
greenpeac + nuclear	0.0539	0.4500	0.1790
Average of 7 queries	0.1492	0.1744	0.4286

As the V_1 vocabulary contains only common words that are not keywords of any query, one can interpret the MAPs obtained as a good indication of the amount of contextual information contributing to the overall results.

3.2 The ImagEVAL dataset

The fourth task of the ImagEVAL[†] benchmark was dedicated to object detection. Ten classes of objects were proposed — *armored vehicle*, *car*, *cow*, *Eiffel tower*, *minaret and mosque*, *plane*, *road signs*, *sunglasses*, *tree* and *US flag*. The test database contains 14,000 pictures, both color and gray-scale (see Figure 2 for some examples). Some images contain objects from multiple classes, and 5000 contain no object from any class. The objects appear in a great variety of poses, contexts and sizes. The training database is composed of bounding box crops of class objects. None of the training images were extracted from the final database. The ImagEVAL database[9] is one of the most challenging image databases available.

Figure 2. Examples of ImagEVAL images containing a US flag.

Table 3 shows the number of training images provided for the ImagEVAL benchmark and the number of matching images in the full dataset. Of the training examples for the *road signs* class, only those which were true photographs of signs were used — all schematic diagrams were omitted.

For evaluations on the ImagEVAL dataset, a query was performed based on each training image, and the MAP was computed on the full dataset for the object class to which the query belonged.

[†] http://www.imageval.org

The number of relevant images is not the same for each of the 10 object categories. We compute the MAP obtained by a random ranking of the dataset images (averaged over 10 runs). We treat these results as a baseline for this dataset. All other scores are expressed as a proportion of the random baseline score, with values approaching 1.0 indicating randomness.

The results we obtain (shown in Table 8) should not to be compared with previously published results for the ImagEVAL dataset, as we have simply chosen this set as a controlled experimental setting for the comparison of pattern selection strategies. The query-by-example operation is comparable to nearest-neighbor classification using a single positive example.

Table 3. ImagEVAL training set sizes

Object class		training set	class members
Armored vehicle	AV	87	730
Car	CA	103	1651
Cow	CO	63	300
Eiffel tower	ET	38	150
Minaret and mosque	MM	82	650
Plane	PL	81	1700
Road signs	RS	31	254
Sunglasses	SU	40	1544
Tree	TR	114	2717
US flag	US	54	342

3.2.1 Low-level features

We extract very simple local features from the ImagEVAL images — color histograms weighted by the local activity of the color in a small neighborhood of each pixel.[10] These texture-weighted color histograms represent combined color and structure information. We use 64 bins to encode them and the L1 distance to compare them. We do not claim that these features are especially well-suited for object detection, but they suffice for the comparison of the different representations we study in this paper. They can easily be replaced by other types of features within the general framework we propose.

3.2.2 Quantizing features

Unlike the Reuters dataset, where due to the nature of text the patterns are already discrete, with the image data sets we instead make use of continuous local features. Each image is represented by an orderless bag of such patterns.

$$I_k = \left\{ P_j^k \in [0,1]^{64}, j \in [1, s_k] \right\} \tag{11}$$

We must therefore first quantize the patterns in order to create our vocabulary. Many algorithms may be used for this purpose. For our experiments, we have chosen the quality threshold clustering (QT) algorithm:[11]

1. A maximum cluster diameter R_{QT} is supplied as a parameter for the clustering.
2. For each point in the dataset, a tentative cluster is constructed with a range query of diameter R_{QT}.
3. The tentative cluster containing the greatest number of points is selected for the clustering.
4. All points of the selected cluster are removed from the dataset.
5. The algorithm reiterates from Step 2 until the dataset is empty.

The main advantages of QT are that it achieves good coverage of the underlying visual space in a deterministic fashion, producing the same clustering each time it is run. Each pattern is quantized by a word that is guaranteed to lie within a range of R_{QT}. QT is also easy to implement. The main disadvantage of QT is its quadratic computational cost. However, we are not concerned with this cost, as quantization is not the focus of our study. Once QT has been used to generate a vocabulary, we may quantize our images by assigning to it the closest words to each of its pattern, as follows:

$$D_k = \left\{ W_j^k, j \in [1, s_k] \,|\forall W_i, d(W_j^k, P_j^k) < d(W_i, P_j^k) \right\} \tag{12}$$

The number of clusters produced by QT, and thus the sizes of the vocabularies generated, depend on the choice of cluster diameter parameter R_{QT}. Some vocabulary sizes obtained for different values of R_{QT} are shown in Table 8.

4. PATTERN SELECTION STRATEGIES

We study two pattern selection strategies: dense sampling and point-of-interest detection. We will explain the parameter value choices and other design choices for both degraded text and image data.

4.1 Dense sampling

For image data, as dense sampling would be too computationally expensive, we restrict the sampling to locations on a uniform grid. In this way, we guarantee that detected patterns are uniformly distributed throughout the image. The support regions within which features are computed are square windows of fixed size centered at grid locations. We conduct different experiments with windows of sizes $w = 8$, 16, 32 and 64 pixels. The grid interval is chosen so as to produce approximately 1000 patterns per image.

For text, we perform dense sampling by applying a sliding window of fixed size over degraded text documents. Tests were conducted for windows of sizes 2, 3, 4 and 5 characters, each test considering every possible placement position for the fixed-size window. At every position, the text string appearing in the window was treated as a word, and the vocabulary for each test consisted of the full collection of fixed-size words encountered over all window positions over all documents in the database. The vocabulary sizes for each test are reported in Table 4, together with the proportion of the potential vocabulary (the size of which is 36^w) covered by the observed vocabulary.

Table 4. Vocabulary size depending on window size

w	Vocabulary Size	10 Most Frequent Words
2	1296 (100%)	er, es, re, on, in, at, te, an, nt, ar
3	43,700 (93.66%)	the, ing, ion, ent, and, ate, ter, for, est, day
4	666,418 (39.68%)	said, tion, nthe, dthe, ment, atio, onth, ther, inth, rthe
5	4,607,713 (7.62%)	ation, inthe, ofthe, saidt, llion, aidth, illio, tions, tiona, idthe

4.2 Point-of-interest detection

Of the many well-known point-of-interest detectors available, we chose to combine the Sift[12] detector with a color Harris detector,[13] as these detectors do not focus on the same salient visual features. Using these detector algorithms, we extracted 500 Sift points and 500 Harris points per image. As for the tests on the grid variant, we compute features within fixed-size square windows centered at detected locations, with each test involving a different choice of window size.

For the purpose of comparison, point-of-interest detection must be simulated for text in a manner analogous to that within images. For degraded text, this raises the question of what constitutes useful textual information, and what notion might be analogous to that of high local variation of signal in images? Any detection strategy for degraded text would need to be repeatable, in that the same sentence from two different documents should be characterized equivalently. Visual point-of-interest detectors concentrate on patterns having high local contrast, and avoid all others. We therefore propose detectors for text that select certain types of patterns and exclude others, namely those patterns matching a preselected subset of the vocabulary. We experimented with two strategies with the W2 dataset.

For a given subset of a text vocabulary, we define *coverage* to be the proportion of patterns in the full dataset that match a word in the subset. The strategies are :

- S1: based on document frequency scores, we form five vocabulary subsets by selecting the 10, 20, 30, 40 and 50 most frequent words.
- S2: based on cumulative inverse document frequency scores, we keep the minimum number of least-frequent words achieving 10% coverage of the data set.

S1 simulates the use of a small number of common patterns, while S2 simulates the use of a very large number of rare patterns. The choice of roughly 1000 points of interest per image suggests a text coverage limit of 10% for S2. 1000 points of interest corresponds to roughly 1% of the information within an image, which can achieved by extracting 10% of the information in each of the two image dimensions. We summarize the vocabulary statistics in Table 5.

Table 5. Vocabulary statistics for the detection strategies – W2

Strategy	Vocab. Size	Coverage
Initial W2 vocabulary	1296	100.00%
S1 - 10	10	15.37%
S1 - 20	20	25.09%
S1 - 30	30	32.48%
S1 - 40	40	38.87%
S1 - 50	50	44.44%
S2	1008	10.00%

4.3 Results and discussion

The results of testing on the Reuters dataset are shown in Tables 6 and 7. In both, the scores are expressed as a proportion of the baseline MAPs obtained using the standard approach on vocabulary V_B. Results on the ImagEVAL corpus are provided in Table 8, expressed as a proportion of the scores achieved using random rankings of database images.

Table 6. Reuters W2 - points of interest strategy

Stemmed query	S1-10	S1-20	S1-30	S1-40	S1-50	S2
goldman + simpson	0.0012	0.0046	0.0107	0.0179	0.0138	0.5142
coulthard	0.0343	0.0400	0.0681	0.0892	0.0954	0.1203
arbil	0.0005	0.0088	0.0180	0.0326	0.0575	0.6917
kasparov	0.0018	0.0051	0.0161	0.0263	0.0439	0.1777
wto	0.0054	0.0129	0.0163	0.0237	0.0264	0.0913
zidan	0.0014	0.0056	0.0095	0.0158	0.0207	0.5368
greenpeac + nuclear	0.0021	0.0030	0.0135	0.0161	0.0187	0.0202
Average of 7 queries	0.0067	0.0114	0.0217	0.0316	0.0395	0.3075

Table 7. Reuters - dense sampling strategy

Stemmed query	W2	W3	W4	W5
goldman + simpson	0.5469	0.9361	0.9974	1.0029
coulthard	0.1594	0.5845	0.9374	1.0635
arbil	0.7333	0.9779	1.1090	1.1288
kasparov	0.1729	0.3873	0.7832	1.1042
wto	0.1410	0.4168	0.5476	0.6402
zidan	0.5739	0.9074	1.0443	1.1417
greenpeac + nuclear	0.0701	0.4916	0.8101	1.0106
Average of 7 queries	0.3425	0.6716	0.8899	1.0131

On the text corpus using a window size of 2 (the W2 case), we found that both point-of-interest strategies are outperformed by dense sampling. The results indicate that the information loss associated with sparse sampling techniques such as point-of-interest detection should be avoided. It is interesting to note that the most informative 2-character words are the rarest ones. Indeed, the 50 words of highest frequency, which together cover 44.44% of the total information, performs poorly (3.95%) where the 1008 words of lowest frequency, together covering 10% of the documents, achieves a much higher score (30.75%) — only slightly worse than that of dense sampling (34.25%).

The results for the dense sampling strategy with larger window sizes are quite interesting and rather surprising: for some queries, the performance of dense sampling was better for degraded text than for the original set with differentiated words. Relative to the original vocabulary, the sliding window of dense sampling captures partial information. The majority of placements of small-sized windows capture the internal structure of words; those placements that straddle word boundaries can be regarded as capturing contextual information. For some placements, true words of the same size as the window can be captured in their entirety. Increasing the fixed window size serves leads to an increase in the proportion of contextual information captured. A small sliding window of size 2 is already able to gather very useful information. Although only 2% of the words in the original vocabulary V_B are of length 2, one third of the baseline result documents can be found using

Table 8. ImagEVAL - pattern selection strategies

	w	R_{QT}	Voc.	Avg.	AV	CA	CO	ET	MM	PL	RS	SU	TR	US
Grid	8	0.5	1390	4.12	2.31	7.80	2.30	4.05	4.05	3.67	1.36	3.57	4.16	3.00
Grid	8	0.7	469	4.91	2.05	8.00	1.85	3.90	3.95	7.35	2.06	5.81	4.78	2.68
Grid	8	0.8	276	4.43	2.69	5.66	1.72	5.27	3.60	5.91	1.49	4.32	5.20	2.27
Grid	8	1.0	115	4.75	2.40	6.00	1.70	4.01	3.99	5.52	1.27	4.60	4.81	15.34
PoI	8	0.5	1364	3.76	3.18	6.15	2.86	3.71	2.21	3.95	2.40	4.06	3.25	6.20
PoI	8	0.7	430	3.51	2.97	4.35	2.83	4.20	4.26	3.64	2.18	3.85	2.88	4.74
PoI	8	0.8	260	3.61	2.81	4.46	3.00	6.58	3.14	4.37	1.55	3.09	3.20	5.02
PoI	8	1.0	115	4.13	3.16	4.19	2.52	2.30	1.93	3.97	2.41	4.21	6.39	2.60
Grid	16	0.5	1525	5.69	1.90	7.68	1.93	4.75	3.38	8.58	2.11	4.09	7.16	7.02
Grid	16	0.7	432	5.29	2.59	5.63	1.51	3.44	3.16	7.01	2.59	3.54	6.79	14.69
Grid	16	0.8	262	5.00	2.65	6.39	1.84	4.49	2.73	9.39	2.50	4.31	4.45	10.12
Grid	16	1.0	106	3.88	2.23	5.52	1.31	3.55	2.92	7.16	3.97	3.80	3.12	3.49
PoI	16	0.5	1283	3.92	3.27	4.90	3.07	4.92	2.38	4.00	1.02	2.29	4.50	7.75
PoI	16	0.7	357	4.10	2.47	4.87	2.06	4.11	2.97	4.15	1.81	3.19	5.64	4.48
PoI	16	0.8	197	4.11	2.62	4.18	2.03	3.72	2.62	4.25	0.88	4.41	5.84	5.14
PoI	16	1.0	82	3.41	2.62	5.09	2.57	1.72	2.83	3.47	0.84	2.78	4.06	1.87
Grid	32	0.5	1536	5.30	1.91	7.37	1.90	4.02	3.72	8.63	2.66	4.38	4.11	19.77
Grid	32	0.7	472	5.65	2.11	6.79	1.91	5.85	3.11	8.94	2.51	3.49	6.42	13.96
Grid	32	0.8	280	5.09	2.37	7.83	1.20	3.99	3.29	9.17	1.58	3.74	4.07	13.35
Grid	32	1.0	112	4.08	3.04	6.92	1.43	2.97	2.64	8.01	2.21	3.29	2.85	3.20
PoI	32	0.5	1351	3.66	2.52	4.76	2.29	4.51	2.35	4.40	0.97	2.76	4.37	3.63
PoI	32	0.7	387	3.62	1.81	5.05	1.91	1.68	3.01	3.66	2.60	2.55	4.85	3.40
PoI	32	0.8	224	3.59	2.58	5.30	1.94	2.02	3.81	4.12	1.72	2.13	3.57	5.81
PoI	32	1.0	89	3.60	2.24	4.27	1.72	1.07	3.78	4.80	1.28	1.91	4.58	2.93
Grid	64	0.5	1640	4.18	2.18	6.62	1.74	4.56	3.49	6.72	2.00	2.60	3.60	5.71
Grid	64	0.7	491	4.58	2.16	6.42	2.26	2.03	4.03	8.60	2.15	2.70	4.36	4.06
Grid	64	0.8	278	3.79	1.84	6.08	1.68	4.14	2.67	5.55	2.34	1.83	3.95	4.06
Grid	64	1.0	116	4.34	2.70	7.07	1.90	4.65	5.07	4.02	2.08	2.65	4.71	3.73
PoI	64	0.5	1391	3.45	1.49	5.07	2.18	2.96	3.54	3.12	0.59	2.77	4.08	5.37
PoI	64	0.7	403	3.40	2.37	4.63	1.65	2.59	4.38	3.15	0.91	2.50	3.42	7.19
PoI	64	0.8	229	3.08	1.79	4.93	2.13	2.40	1.77	2.83	1.96	2.52	3.56	5.31
PoI	64	1.0	91	2.91	1.67	5.31	1.38	3.41	3.11	2.76	0.91	2.89	2.59	3.45

queries based on the vocabulary W2. If we compare with the results obtained for the undegraded vocabulary V_1 (which has the same size as W2), we notice that the average retrieval rates for the degraded vocabulary W2 are twice as high. We believe that a fixed choice of window size acts to some extent as a band-pass filter, focusing on true words of length of the same size as the window. It is hardly surprising that the best results were obtained for W5, as its window size is closest to the mean size of the true words in the baseline vocabulary V_B (see Figure 1). Although much noise is captured when the windows are larger (the size of W5 is approximately 12 times larger than that of V_B), noise words tend to be suppressed by low TF-IDF weightings in vector representations. Performance may also be boosted due to the increased contextual information captured with the larger window sizes. These observations led us to conduct an experiment with a vocabulary (V_2) consisting of those words of the undegraded baseline vocabulary V_B having exactly 5 characters. Except for one of the seven queries, *kasparov*, the results for V_2 are substantially worse than those of V_B. Taken together, the experiments lead us to conclude that finding precise word boundaries is not necessary, as densely sampled features are sufficient for good performance.

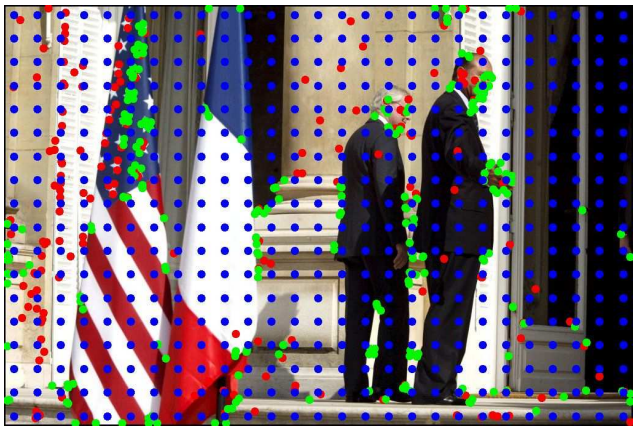

Figure 3. ImagEVAL - points of interest generated by 3 different detectors

On the ImagEVAL dataset, even if for some classes and settings the point-of-interest approach is better, globally dense sampling is more relevant and provides on average better performance for a given number of local patterns extracted. In some cases (such as the *cow* class), the point-of-interest strategy may strictly outperform the grid approach, for all windows sizes. In other cases (such as the *plane* class), we may observe the opposite. For the case of the *US flag* class, we can see that the results are better with points of interest when using small windows ($w = 8$), but the best overall results are reached with the grid approach and windows of size $w = 32$. As an example, the image of Figure 3 shows the features extracted from a picture where a flag is present — SIFT features are shown in red, Harris color features are shown in green, and the grid itself is shown in blue. We note that both SIFT and Harris points of interest are located on stars and borders of the flag, but points of interest were selected from the stripes. The striped regions of the flag provide useful information, but its visual structure is not well captured by points of interest. This example shows the advisability of using a dense sampling when little or no *a priori* information is available for the image database.

5. CONCLUSIONS AND FUTURE WORK

We introduced an original generalized framework to evaluate different representations and pattern selection strategies for documents. We proposed that techniques from the image domain be applied to a degraded version of the text documents for the purpose of assessing their effectiveness, since ground truth information is generally easier and cheaper to obtain for text than it is for images. To facilitate the comparison, we also designed a point-of-interest detection strategy for unsegmented text. We have evaluated two local patterns selection strategies used with bag-of-words representation. We have shown that on a text corpus and on an image corpus, the behaviour of dense sampling and point-of-interest sampling is similar and, in both cases, the results are generally better with dense sampling.

We believe that dense pattern sampling should be used in image representation as it leads to less information loss in comparison with point-of-interest detection. In practice, when extracting a number of patterns from an image, dense

pattern sampling allows for lower computational costs due to the relative ease of using a grid to determine their locations. Moreover, we've seen that knowledge of the exact word boundaries in the text corpus is not necessary for achieving good results. The experiments for text suggest that internal object structure and external contextual information may be helpful when annotating images, which supports our argument in favor of dense sampling of the visual content.

In future work, we plan to extend the representation and sampling strategies introduced in this paper to a full annotation framework, with several types of low-level features, multiscale combination of different window sizes in the same representation, and the inclusion of knowledge gained from vocabulary creation and learning models. We also plan to evaluate other representations for images and test them first on degraded text corpuses. This evaluation framework could also be applied to other forms of multimedia (such as sound and video) and their associated patterns.

ACKNOWLEDGMENTS

The work presented in this paper was partially supported by the European Commission under contract FP6-045389 Vitalas. All pictures are copyrighted by Bassignac-Gamma, except the top center photo in Figure 2, which is copyrighted by Keystone.

REFERENCES

[1] Hervé, N. and Boujemaa, N., "Image annotation : which approach for realistic databases ?," in [*ACM International Conference on Image and Video Retrieval (CIVR'07)*], (July 2007).

[2] Barnard, K., Duygulu, P., Forsyth, D., de Freitas, N., Blei, D. M., and Jordan, M. I., "Matching words and pictures," *Journal of Machine Learning Research* **3**, 1107–1135 (2003).

[3] Dance, C., Willamowski, J., Fan, L., Bray, C., and Csurka, G., "Visual categorization with bags of keypoints," in [*ECCV International Workshop on Statistical Learning in Computer Vision*], (2004).

[4] Zhang, J., Marszalek, M., Lazebnik, S., and Schmid, C., "Local features and kernels for classifcation of texture and object categories: An in-depth study," Tech. Rep. RR-5737, INRIA Rhône-Alpes (2005).

[5] Amores, J., Sebe, N., and Radeva, P., "Efficient object-class recognition by boosting contextual information," in [*IbPRIA*], (2005).

[6] Salton, G., [*The SMART Retrieval System—Experiments in Automatic Document Processing*], Prentice-Hall, Inc., Upper Saddle River, NJ, USA (1971).

[7] Lewis, D. D., Yang, Y., Rose, T., and Li, F., "Rcv1: A new benchmark collection for text categorization research," *Journal of Machine Learning Research* **5**, 361–397 (2004).

[8] Porter, M., "An algorithm for suffix stripping," *Program* **14(3)**, 130–137 (1980).

[9] Picault, C., "Constitution of the imageval database, en end-user oriented approach," tech. rep., Paragraphe Laboratory, Université Paris 8 (2006).

[10] Vertan, C. and Boujemaa, N., "Upgrading color distributions for image retrieval: can we do better ?," in [*International Conference on Visual Information Systems*], (2000).

[11] Heyer, L. J., Kruglyak, S., and Yooseph, S., "Exploring expression data: Identification and analysis of coexpressed genes," *Genome Research* **9**, 1106–1115 (1999).

[12] Lowe, D. G., "Object recognition from local scale-invariant features," in [*International Conference on Computer Vision*], (1999).

[13] Gouet, V. and Boujemaa, N., "Object-based queries using color points of interest," in [*IEEE Workshop CBAIVL/CVPR*], (2001).