Impact of visual information on text and content based image retrieval

Christophe Moulin, Christine Largeron, and Mathias Géry

Université de Lyon, F-42023, Saint-Étienne, France ;
CNRS, UMR 5516, Laboratoire Hubert Curien, 42023, Saint-Étienne, France ;
Université de Saint-Étienne, Jean-Monnet, F-42023, Saint-Étienne, France ;
{christophe.moulin, christine.largeron, mathias.gery}@univ-st-etienne.fr

Abstract. Nowadays, multimedia documents composed of text and images are increasingly used, thanks to the Internet and the increasing capacity of data storage. It is more and more important to be able to retrieve needles in this huge haystack. In this paper, we present a multimedia document model which combines textual and visual information. Using a bag-of-words approach, it represents a textual and visual document using a vector for each modality. Given a multimedia query, our model combines scores obtained for each modality and returns a list of relevant retrieved documents. This paper aims at studying the influence of the weight given to the visual information relative to the textual information. Experiments on the multimedia ImageCLEF collection show that results can be improved by learning this weight parameter.

1 Introduction

In order to retrieve documents in multimedia collections, especially in the context of the Web, the development of methods and tools suitable to these data types is nowadays a challenging problem in Information Retrieval (IR). Most of the current IR systems handling multimedia documents can be classified into several categories, depending on their ability to exploit textual information, visual information, or a combination of both.

In the first category, namely *Text based Image Retrieval*, an image is indexed using only the textual information related to the image (file name, legend, text surrounding the image, etc.), without taking into account the image intrinsic features. This is the case, for example, of the main commercial search engines, and also of some systems specialized in images retrieval, such as Picsearch[1].

In the second category, namely *Content Based Image Retrieval* (CBIR), only the visual content of the image, represented by local color, shape or texture features, is used [1, 2]. For example, QBIC, the IBM precursor system [3], proposes to retrieve images considering a query expressed using only those basic color, shape and texture features. The systems giving the best results are those handling a query image built by the user or an image example provided by the user ("Search by image", e.g. QBIC or more recently the search engine TinEye[2]).

[1] Picsearch: http://www.picsearch.com
[2] TinEye: http://www.tineye.com/

So, some systems propose to the user to sketch the image sought ("Search by sketch", e.g. the Gazopa and Retrievr[3] search engines) while other propose to the user to arrange on a canvas the icons corresponding to concepts that have been previously identified in the image database. But one drawback of these systems is that users do not always have a reference image, and query languages based on visual features are not always very intuitive.

Finally, the last category deals with systems handling textual and visual features simultaneously. For example, the PicHunter system [4] aims at predicting users' goal given their actions while the Picitup system[4] proposes to define a textual query and then to filter results using visual elements (a picture, a category, a color, a shape, etc.). Recently, these approaches aiming at combining textual and visual information have been encouraging [5, 6], but they have to fill the semantic gap between the objects and their visual representation [1]. A possible research direction deals with using visual ontology [7]; another one, proposed recently by Tollari, aims at associating keywords and visual information [8].

These previous works led us to propose a first approach which combines textual and visual information. Starting from a first set of documents returned for a given textual query, our system enriches the query, adding some visual terms to the original textual query in an automatic way or a semi-automatic way (i.e. asking the user for feedback on the first returned documents) [9].

Our preliminary experiments have shown the potential of combining visual and textual information. The first aim of the present work is to study how to estimate the weight of the visual information relative to the textual information. We propose to learn automatically this weight, using an IR collection as a learning set. The second aim is to check if the optimal weight accorded to each information type varies by the kind of queries, and if estimating a specific weight for each query can significantly improve the results. Indeed, the visual information is less important for concepts like e.g. "animal" or "vehicle", because these concepts can be described by very different visual features.

The next section describes the document model we proposed, combining text and images, then we present some experiments on an IR task using the ImageCLEF collection in section 3; we present the results in section 4.

2 Visual and textual document model

2.1 General framework

The figure 1 presents the global architecture of our multi-modal IR model. The first component aims at indexing the documents D and the queries Q, both composed by textual and visual information. The textual content, as well as the visual one, is represented by a bag-of-words. The second component estimates, given a query, a score for each document and for each modality (textual and visual). Finally, the last component combines linearly the score obtained for each modality, in order to retrieve the most relevant documents given a query.

[3] Gazopa: http://www.gazopa.com/, Retrievr: http://labs.systemone.at/retrievr/
[4] http://www.picitup.com/picitup

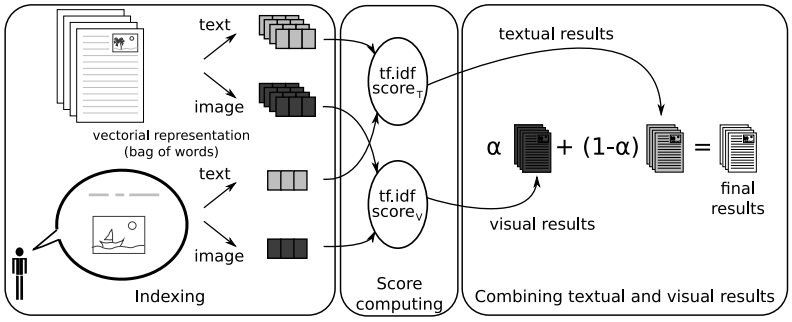

Fig. 1. Multi-modal IR model

2.2 Textual document model

Given a collection of documents D and $T = \{t_1, ..., t_j, ..., t_{|T|}\}$ the set of words occurring in the documents, each document $d_i \in D$ is represented as a vector of weights $w_{i,j}$ (vector space model [10]): $\boldsymbol{d_i} = (w_{i,1}, ..., w_{i,j}, ..., w_{i,|T|})$, with $w_{i,j}$, the weight of the term t_j in the document d_i, computed by a $tf.idf$ formula ($w_{i,j} = tf_{i,j} * idf_j$). $w_{i,j}$ is high when the term t_j is frequent in the document d_i but rare in the others.

$tf_{i,j}$ is the *term frequency* that characterizes the representativeness of the term t_j in the document d_i. We use the variant of the BM25 weighting function defined in Okapi [11] and implemented by the Lemur system [12]:

$$tf_{i,j} = \frac{k_1 \times t_{i,j}}{t_{i,j} + k_1 \times (1 - b + b * \frac{|d_i|}{d_{avg}})}$$

where $t_{i,j}$ is the number of occurrences of the term t_j in the document d_i, $|d_i|$ the size of the document d_i, d_{avg} the average size of all documents and k_1 and b two constants.

idf_j is the *inverse document frequency* which estimates the importance of the term t_j over the corpus of documents. We use also the BM25 variant implemented by Lemur:

$$idf_j = \log \frac{|D| - df_j + 0.5}{df_j + 0.5}$$

where $|D|$ is the size of the corpus and df_j the number of documents where the term t_j occurs at least one time.

If we consider a query q_k in the same way (i.e. as a short document), we can also represent it as a vector of weights. A score is then computed between the query q_k and a document d_i:

$$score_T(q_k, d_i) = \sum_{t_j \in q_k} tf_{i,j} idf_j * tf_{k,j} idf_j$$

2.3 Visual document model

In order to combine the visual and the textual information, we also represent images as vectors of weights. It is possible to use the $tf.idf$ formula in the same

way, provided we are able to extract visual words from images. It requires a visual vocabulary $V = \{v_1, ..., v_j, ..., v_{|V|}\}$, which is built in two steps using a bag of words approach [13]. In the first step, each image of the collection D is segmented into a regular grid of 16×16 cells, with at least 8×8 pixels by cell. Then, each cell is described by the visual descriptor SIFT (*Scale-Invariant Feature Transform*) based on histograms of gradient orientation [14]. SIFT converts each cell into 128-dimensional vector in such a way that each image is a collection of vectors. We have evaluated other visual descriptors, like *meanstd* [9], but only the best results, provided by SIFT, are presented in this article.

In the second step, the visual words are built by performing a k-means clustering over the visual vectors. The words of the visual vocabulary V are then defined as the centers of the clusters and the size of the visual vocabulary corresponds to the number of clusters.

This bag of visual words is analogous to the bag of textual words inasmuch as an image can then be represented by an histogram of visual words. Indeed, an image, belonging to a document or a query, can be segmented into cells described by SIFT vectors and, each vector can be assigned to the nearest cluster (i.e. visual word) according to the Euclidean distance. This way, it is possible to count the number $v_{i,j}$ of occurrences of the visual word v_j in the image, in other words the number of cells $v_{i,j}$ assigned to the cluster with the center v_j. Like in the textual model, an image is represented by a vector where the weights for the visual words are given by the $tf.idf$ formula in which $t_{i,j}$ is replaced by $v_{i,j}$ and t_j by v_j.

Finally, a visual score $score_V(q_k, d_i)$ is then computed between a document d_i and a query q_k by:

$$score_V(q_k, d_i) = \sum_{v_j \in q_k} tf_{i,j} idf_j * tf_{k,j} idf_j$$

2.4 Combining textual and visual informations

The global score for a document d_i given a query q_k is computed, combining linearly the scores computed for each modality:

$$score(q_k, d_i) = \alpha \times score_V(q_k, d_i) + (1 - \alpha) \times score_T(q_k, d_i)$$

where α is a parameter allowing to give more or less importance to the visual information relative to the textual information.

3 Experiments

In order to experiment our model, we have used the IR collection ImageCLEF [15]. Our aim is to evaluate the impact of visual information on multimedia IR: this requires to study the influence of the fusion parameter α.

3.1 ImageCLEF: IR collection

The ImageCLEF collection is composed by 151,519 XML documents extracted from Wikipedia, composed by one image (photos, drawings or painting) and a short text, which describes the image but which can also give some information related to the owner or to the copyright.

Each year, a different set of queries is delivered: in ImageCLEF 2008, used as a training collection, there are 75 queries. 42 queries contain both a textual part (a few words) and a visual part. The 33 others queries are provided with only a textual part. In order to have a visual information obtained in a similar way for all queries, the two first images ranked by a preliminary textual querying step have been used as a visual query part for all the 75 queries. In ImageCLEF 2009, used as a testing collection, there are 45 queries, containing both a textual part and a visual part (1.84 images per query).

3.2 Evaluation measures

Several evaluation measures have been used, such as MAP, $P10$ and $iP[0.1]$. Let $Q = \{q_1, ..., q_k, ..., q_{|Q|}\}$ be the set of queries and $D_k = \{d_{k,1}, ..., d_{k,i}, ..., d_{k,|D_k|}\}$ the set of relevant documents given q_k. The N_k retrieved documents for the query q_k is a list of documents ranked according to their score. In ImageCLEF competition, N_k equals to 1000. The rank r corresponds to the r^{th} document ranked by the system. Precision $P_k(N)$ is defined as the number of relevant retrieved documents given q_k divided by the N retrieved documents. Recall $R_k(N)$ is defined as the number of relevant retrieved documents divided by the number of relevant documents. AP_k is the average precision for q_k.

$$P_k(N) = \frac{\sum_{r=1}^{N} \text{rel}_k(r)}{N} \quad R_k(N) = \frac{\sum_{r=1}^{N} \text{rel}_k(r)}{|D_k|} \quad AP_k = \frac{\sum_{r=1}^{N_k} (P_k(r) \times \text{rel}_k(r))}{|D_k|}$$

where $\text{rel}_k(r)$ is a binary function which equals 1 if the r^{th} document is relevant for the query q_k and 0 otherwise.

Three evaluation measures have been used to evaluate our model. The first one (MAP: Mean Average Precision) corresponds to the average for all queries of the average precision AP_k. The second one ($P10$) is the precision at 10^{th} rank. The last one ($iP[0.1]$) is the interpolated precision at 10% recall.

$$MAP = \frac{\sum_{k=1}^{|Q|} AP_k}{|Q|} \quad P10 = \frac{\sum_{k=1}^{|Q|} P_k(10)}{|Q|} \quad iP[0.1] = \frac{\sum_{k=1}^{|Q|} iP_k[0.1]}{|Q|}$$

with:

$$iP_k[0.1] = \begin{cases} \max_{1 \leq r \leq N_k}(P_k(r) | R_k(r) \geq 0.1) & \text{if } 0.1 \leq R_k(N_k) \\ 0 & \text{otherwise} \end{cases}$$

3.3 Experimental protocol

Many experiments were conducted in order to evaluate the interest of considering visual information on an IR task, and to study the α's influence.

Learning the α parameter: firstly, queries from the ImageCLEF 2008 (resp. ImageCLEF 2009) collection are used as training set in order to calculate α_g^{2008} (resp. α_g^{2009}), the α value that globally optimize results on ImageCLEF 2008 (resp. ImageCLEF 2009). The optimal value of α correspond to the value of α that gives the best results for a given criterion, such as the MAP measure, obtained using a stepped search on the training set. We have used the MAP measure which is the main one used in the ImageCLEF competition. The learned α_g^{2008} value has been used by our system to process all the queries from the ImageCLEF 2009 collection. Our first question concerns the possibility of learning the parameter of the model on a set of queries and using it on a new set of queries: is it possible to estimate the optimized value α_g^{2009} using the ImageCLEF 2008 collection? The comparison of α_g^{2008} and α_g^{2009} will allow to conclude on the effectiveness of learning α.

Robustness of α with regard to evaluation measures: the second aim is to determine the importance of visual information relative to the textual information, depending on the use case: 1) recall-oriented (exhaustive search), retrieving a lot of documents more or less relevant, 2) precision-oriented (focused search), retrieving a smaller set of documents mostly relevant. For this purpose, we have studied the parameter α_g regarding several evaluation measures: in the first hand MAP, which focus on recall, and in the other hand $P10$ and $iP[0.1]$, which focus on precision.

Optimizing α parameter depending on the query: thirdly, we study the behavior of our model depending on the query type. Some queries seem to mainly depend on the textual information, such as *"people with dogs", "street musician"*, while others require more visual information, such as *"red fruit", "real rainbow"*. Studying how the performance of the system change depending on the kind of query is thus interesting. This local approach aims at calculating α_k, the α value optimized given a query q_k. The mean and the standard deviation of α_k will let us conclude on the variation of the α parameter depending on the query and on the interest of methods that aim at estimating the optimal α_k value for a new query. We will also study the optimization of α depending on the evaluation measures and thus, we will calculate the α_k optimized for the MAP, $P10$ and $iP[0.1]$ measures.

Global vs local approach: in the global approach, we study the variation of the α parameter in order to optimize the evaluation measure MAP_α (resp. $P10_\alpha$, $iP[0.1]_\alpha$). Let α_g be the optimal global value of the α paramater that maximizes MAP_α (resp. $P10_\alpha$, $iP[0.1]_\alpha$) on the training set:

$$\alpha_g = \alpha | MAP_{\alpha_g} = max\{MAP_\alpha, \alpha \in [0,1]\}$$

α_g is then used for all queries of the test set. During the ImageCLEF 2009 competition, α_g was obtained using all the queries of the 2008 collection and it was then used for processing the queries of the 2009 collection.

The local approach that uses a specific α per query should be the best solution. However, in practice, this local approach can not be performed since a

training set is not available for a new query. Nevertheless, in order to compare our global approach with this local approach, we have searched the α_k value that optimizes the AP_k, $P_k(10)$ and $iP_k[0.1]$ measures for each query q_k using the test set. Then the MAP_{α_l} measure, corresponding to the average of the optimized average precision AP_k, is defined by:

$$MAP_{\alpha_l} = \frac{\sum_{k=1}^{|Q|} AP_k|\alpha = \alpha_k}{|Q|}$$

3.4 Setting up of our model

The lemur software has been used with the default parameters as defined in [12]. The k_1 parameter of BM25 is set to 1. As $|d_k|$ and d_{avg} are not defined for a query q_k, b is set to 0 for the $tf_{k,j}$ computation. When the $tf_{i,j}$ is computed for a document d_i and a term t_j, this paramater b is set to 0.5. Moreover, stop-words have not been removed and a Porter stemming algorithm have been applied. The number of visual words, corresponding to the parameter k of the k-means, has been empirically set to 10,000.

4 Results

4.1 Learning parameter α

MAP is a global measure corresponding to the average of the average precision for each query. This is the official ImageCLEF measure. Table 1 summarizes the results obtained, depending on which modality is used (text, visual, text + visual), and also on the optimizing method that is used. According to the MAP measure, the visual information leads to poor results ($MAP = 0.0085$) compared to those obtained using only the text ($MAP = 0.1667$).

Table 1. Results on the ImageCLEF 2009 collection (MAP measure)

Run	MAP	Gain / text only
Text only	0.1667	
Visual only	0.0085	-94.90%
Text+Visual (α_g^{2008})	0.1903	+14.16%
Text+Visual (α_g^{2009})	0.1905	+14.28%

However, figure 2 shows that giving more importance to the visual information significantly improves the results obtained only with text, especially with α close to 0.1. Nevertheless, giving too much importance to α (i.e. $\alpha > 0.1$) reduces the results quality. The α values are not normalized: thus it is difficult to interpret them directly, and only the improvement of IR results should allow to evaluate the interest of integrating visual information.

The parameter α_g^{2008} computed with the 2008 learning collection improves the results obtained using only the text on 2009 collection (+14.16%, $MAP =$

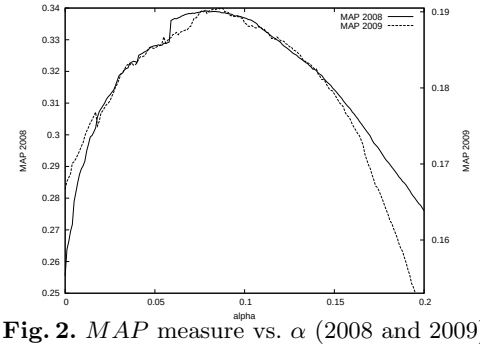

Fig. 2. MAP measure vs. α (2008 and 2009).

0.1903). This result is very interesting, particularly when it is compared to the optimal result ($MAP = 0.1905$) obtained using the α_g^{2009} value optimized on the 2009 collection itself. The MAP curves according to α, which look similar, and the values of $\alpha_g^{2008} = 0.084$ and $\alpha_g^{2009} = 0.085$, show a good robustness of the α_g parameter while changing collection (w.r.t. the MAP). Thus we think that learning α_g is possible.

4.2 Stability of parameter α_g regarding the evaluation measure

Regarding more specific evaluation measures, as for example the precision oriented measures $P10$ and $iP[0.1]$, the α parameter seems less stable than regarding the MAP measure, especially on the 2009 collection, as shown by figure 4.2 (note that $P10$ and $iP[0.1]$ are averages, while MAP is an average of averages).

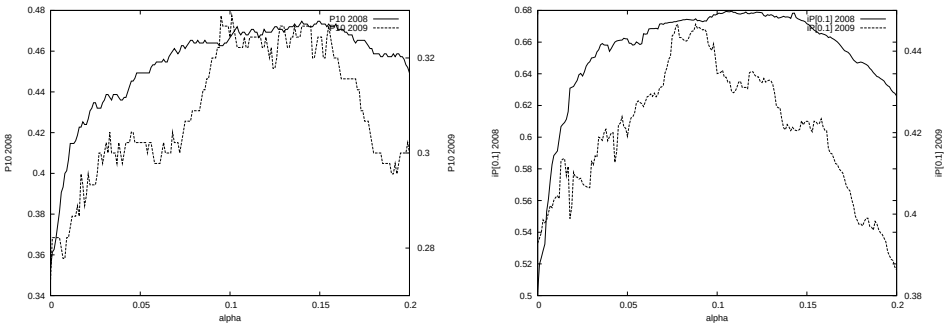

Fig. 3. $P10$ and $iP[0.1]$ measures vs. α (2008 and 2009).

For these measures, the value of the α parameter learned on 2008 ($P10$: $\alpha_g^{2008} = 0.140$; $iP[0.1]$: $\alpha_g^{2008} = 0.108$) is quite different than the optimal α value for 2009 ($P10$: $\alpha_g^{2009} = 0.095$; $iP[0.1]$: $\alpha_g^{2009} = 0.078$). Nevertheless, the weighting of the visual information through the parameter α_g^{2008}, even if relatively different than the optimal value α_g^{2009}, still allows to significantly improve the results regarding $P10$ as well as $iP[0.1]$, as shown by table 2. We observe an improvement of 19.54% regarding $P10$, and of 9.49% regarding $iP[0.1]$.

Table 2. Results on the collection ImageCLEF 2009 ($P10$ and $iP[0.1]$ measures)

Run	$P10$	Gain / text only	$iP[0.1]$	Gain / text only
Text only	0.2733		0.3929	
Visual only	0.0178	-93.49%	0.0160	-95.93%
Text+visual (α_g^{2008})	0.3267	+19.54%	0.4302	+9.49%
Text+visual (α_g^{2009})	0.3289	+20.34%	0.4466	+13.67%

4.3 Global approach vs local approach: optimizing α w.r.t. a query

The local approach, i.e. using a specific α_k parameter for each query q_k, is more challenging than the global approach, because it needs to compute a priori the value of α_k for each new query; this is an open problem. However, this approach would allow to dramatically improve the results: the potential gain is +29.99% (resp. +52.87%, +39.14%) reagrding the MAP measure (resp. $P10$, $iP[0.1]$), as shown by table 3. But implementing this local approach seems very difficult as it exists an important disparity of α_k regarding to the queries, as shown by μ_{α_l} (mean of α_k) and σ_{α_l} (standard deviation) observed for the 3 evaluation measures.

Table 3. Optimizing α_k for each query

	Run		Gain / text only	μ_{α_l}	σ_{α_l}
MAP	Text only	0.1667			
	Text+visual (α_l)	0.2167	+29.99%	0.080	0.063
$P10$	Text only	0.2733			
	Text+visual (α_l)	0.4178	+52.87%	0.055	0.058
$iP[0.1]$	Text only	0.3929			
	Text+visual (α_l)	0.5467	+39.14%	0.083	0.072

5 Conclusion and future work

In this paper, we have presented a multimedia IR model based on a bag-of-words approach. Our model combines linearly textual and visual information of multimedia documents. It allows to weight the visual information relative to the textual information using a parameter α.

Our experiments show that it is possible to learn a α_g^{2008} value for this parameter (using the ImageCLEF 2008 collection as a learning collection) and then to use it successfully on the ImageCLEF 2009 collection. This value sometimes differs compared to the optimal value α_g^{2009} (computed on the collection ImageCLEF 2009) regarding $P10$ and $iP[0.1]$, but remains relatively stable regarding MAP. However it allows to significantly improve the results regarding MAP as well as $P10$ and $iP[0.1]$.

According to our results, using a specific α_k for each query seems to be an interesting idea. In order to learn this parameter, a first approach could be to

classify the queries: visual, textual and mixed queries. Maybe it is possible for this purpose to use the length of the textual queries, which seems to be related to the queries' class. Another direction could be to analyze some visual words extracted from the first set of textual results given the query, hypothesizing that they carry some visual information about the query. Their distribution should allow to estimate a specific α_k for each query.

References

1. Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. IEEE Transactions on Pattern Analysis and Machine Intelligence **22**(12) (2000) 1349–1380
2. Lew, M.S., Sebe, N., Djeraba, C., Jain, R.: Content-based multimedia information retrieval: State of the art and challenges. ACM Transactions on Multimedia Computing, Communications, and Applications **2**(1) (2006) 1–19
3. Flickner, M., Sawhney, H.S., Ashley, J., Huang, Q., Dom, B., Gorkani, M., Hafner, J., Lee, D., Petkovic, D., Steele, D., Yanker, P.: Query by image and video content: The QBIC system. IEEE Computer **28**(9) (1995) 23–32
4. Cox, I.J., Miller, M.L., Minka, T.P., Papathomas, T.V., Yianilos, P.N.: The bayesian image retrieval system, PicHunter: theory, implementation, and psychophysical experiments. IEEE Transactions on Image Processing **9**(1) (2000) 20–37
5. Barnard, K., Duygulu, P., Forsyth, D., de Freitas, N., Blei, D.M., Jordan, M.I.: Matching words and pictures. The Journal of Machine Learning Research **3** (2003) 1107–1135
6. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Image retrieval: Ideas, influences, and trends of the new age. ACM Computing Surveys **40**(2) (2008)
7. Snoek, C.G.M., Worring, M., Gemert, J.C.V., mark Geusebroek, J., Smeulders, A.W.M.: The challenge problem for automated detection of 101 semantic concepts in multimedia. In: ACM Conference on Multimedia. (2006) 421–430
8. Tollari, S., Detyniecki, M., Marsala, C., Fakeri-Tabrizi, A., Amini, M.R., Gallinari, P.: Exploiting visual concepts to improve text-based image retrieval. In: European Conference on Information Retrieval (ECIR). (2009)
9. Moulin, C., Barat, C., Géry, M., Ducottet, C., Largeron, C.: UJM at ImageCLEFwiki 2008. In: 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008. Volume 5706 of Lecture Notes in Computer Science., Springer (2008) 779–786
10. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Communations of the ACM **18**(11) (1975) 613–620
11. Robertson, S.E., Walker, S., Hancock-Beaulieu, M., Gull, A., Lau, M.: Okapi at trec-3. In: Text REtrieval Conference. (1994) 21–30
12. Zhai, C.: Notes on the lemur TFIDF model. Technical report, Carnegie Mellon University (2001)
13. Csurka, G., Dance, C., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: ECCV'04 workshop on Statistical Learning in Computer Vision. (2004) 59–74
14. Lowe, D.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60**(2) (2004) 91–110
15. Tsikrika, T., Kludas, J.: Overview of the wikipediaMM task at ImageCLEF 2009. In: 10th Workshop of the Cross-Language Evaluation Forum, Corfu, Greece (2009)