IPAL Inter-Media Pseudo-Relevance Feedback Approach to ImageCLEF 2006 Photo Retrieval

Nicolas Maillot, Jean-Pierre Chevallet, Vlad Valea, and Joo Hwee Lim

IPAL French-Singaporean Joint Lab
Institute for Infocomm Research (I2R)
Centre National de la Recherche Scientifique (CNRS)
21 Heng Mui Keng Terrace
Singapore 119613
{nmaillot, viscjp, joohwee}@i2r.a-star.edu.sg

Abstract. This document describes the participation to ImageCLEF 2006 photographic retrieval task of the IPAL lab (Singaporean French collaboration) hosted at Institute for Infocomm Research, Singapore. This paper provides a description of the way results has been produced. The text/image database used is IAPR [1]. We have tested a *cooperative* use of a text retrieval and an image retrieval engine. We show in particular how inter-media re-ranking and pseudo-relevance feedback have been used for producing the results. We have also tested Latent Semantic Analysis (LSA) approach on visual runs. WordNet thesaurus has been used for pre-processing textual annotations within spell checking corrections. Our approach is completely automatic. A description of the runs submitted to the competition is also given.

Categories and Subject Descriptors: H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing - *Indexing methods*, *Thesauruses*; H.3.3 Information Search and Retrieval - *Clustering*, *Relevance feedback*, *Retrieval models*

1 Introduction

One of the most interesting issues in multimedia information retrieval is to use different modalities (e.g. text, image) in a cooperative way.

In this experiment, our goal is to study *inter-media pseudo-relevance feedback* (between text and image) and so to explore how the output of an image retrieval system can be used for expanding textual queries. This is motivated by the hypothesis that two images with a very strong visual similarity should share some common semantics.

We are also interested in studying how appearance-based re-ranking techniques can be used to enhance the output of a text retrieval system. This is motivated by the fact that high-level concepts have most of the time a large variety of visual appearances. In some cases, it can be useful for the end-user to obtain images of a concept which have a well-defined appearance.

This document is structured as follows. Section 2 gives a description of the system used to produce results submitted by IPAL at ImageCLEF 2006 Photo task. Section 3 provides a description of the most important runs submitted. Section 4 provides an analysis of the results.

2 System description

2.1 Overview

Our results have been produced by the cooperative use of a image indexing and retrieval system and a text indexing and retrieval system. An overview of the complete retrieval system can be found in fig. 4. The system developed contains pseudo-relevance feedback and re-ranking capabilities.

2.2 Text Indexing and Retrieval

Our goal on the text runs is to experiment a mixture of knowledge and statistical information to solve very precise short query. Knowledge comes from WordNet [2] and from the corpus it-self (for Geographical Named Entities). Statistical information comes only from the corpus.

Text indexing and retrieval were mainly achieved by the XIOTA system [3]. Before indexing, we have experimented four levels of linguistic treatment: morpho-syntactic, noun-phrase, named entity and conceptual. The morpho-syntax consists in transforming the original text into normalized word stems with part-of-speech information (POS). More information are usually available like number (singular and plural), and the stemmed form. Noun-phrase consists in grouping a word sequence that has a unique meaning like "swimming pool". In some case it includes the change of the part of speech. For example, at the morpho-syntax level, "swimming pool" is recognized as a verb (swim) followed by a noun. But at the noun phrase level, the composed noun is recognized, and identified as a unique term. Finally, at the conceptual level, all terms are replaced by a concept reference. For this last step sense disambiguation is mandatory.

Morpho-Syntax Texts have first been preprocessed in order to recognize part of speech and to correct spelling. The following steps have been followed in sequence:

- XML correction: As XIOTA relies on a correct XML data flow, an automatic XML correction is applied for correcting some closing tags.
- Part of Speech: Files are then passed through a part of speech tagger (TreeTagger[1] [4]). A correction is applied to suppress tagging from documents identifiers.

[1] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

– Unknown proper nouns: when the tagger recognizes proper nouns, it provides a unique normalize version. When the tagger does not recognize the proper noun, we assume the normalize form does not change. Other forms of unrecognized terms are supposed to be misspelled words.
– Word normalization: it consists in removing every accent, and also removing some rare character coding errors, assuming char coding is ISO-8859-1 Latin 1. This is mainly effective for foreign geographical proper nouns (e.g. Spanish).
– Spelling corrections: we make the assumption that every terms not tagged as a proper noun and unknown is misspelled (about 700 terms, ex: "buidings", "toursits"). This is false for some terms not recognized by the POS tagger because they are joint like "belltower", "blowtube"."snowcover", or because of the hyphen like "winter-jaket", "cloud-imposed". This list of unknown words is passed through aspell[2] to associate a possible correct form. When aspell proposes several choices, the first one is selected.

We think that the spelling correction is important to ensure correct index in the case of short documents. Possible misspellings are detected thanks to the part of speech step. Queries are processed in the same way. All other processing including text indexing, start from the analyzed and corrected text collection. Namely, the basic vector space indexing performs only a POS filtering before building vector indexes.

Noun Phrase We have used WordNet to detect noun phase. Candidates were selected using POS template:

– *noun (singular) + noun (singular)* (e.g. "baseball cap")
– *noun (singular) + noun (plural)* (e.g. "tennis players")
– *proper noun + proper noun* (e.g. "South America")
– *verb VG + noun* (e.g. "swimming pool")

If the two following conditions hold: the template and the presence in WordNet, we replace the word couple by one term with the correct stemmed version. It means that the two terms will be treated as only one indexing term. We have not used cooccurrence statistics because the corpus is too small.

Named Entity In this tourist image set, the location of the scene in the picture is important. That is why we have decided to detect geographic Named Entity. We have used two information sources: WordNet and the corpus itself. In fact, we have extracted information from the LOCATION tag to build a list of geographic names. Then we have tagged the rest of the corpus using first WordNet and then this list. Filtering is based on the proper noun POS, and on the lexical WordNet category "noun.location". If the location is not found in WordNet, then the location list is used. We have used this information to split the query and force the geographic information matching.

[2] http://aspell.sourceforge.net/

Concept Concept indexing seems a nice way to solve the term mismatch because all term variations is replace by one unique concept. Unfortunately the problem remains in the concept detection, because it need a disambiguation step. For this experiment, we have used the WordNet sense frequency when available. This information provides a sort of statistic on the more frequent sense of a term. Otherwise we have filtered the most frequent semantic category (lexname), and choose the most frequent one. This choice is correct most of the time for this corpus. This step produces a new corpus with WordNet concept references that enables conceptual indexing and retrieval.

2.3 Image Indexing and Retrieval

Feature Extraction. The feature extraction process is based on a tessellation of the images of the database. Each image is split into patches (fig. 1). The visual indexing process is based on patch extraction on all the images of the collection followed by feature extraction on the resulting patches. Let I be the set of the N images in the document collection. First, each image $i \in I$ is split into n_p patches p_i^k $(1 \leq k \leq n_p)$. Patch extraction on the whole collection results in $n_p \times N$ patches. Fig. 1 shows the result of patch extraction for one image.

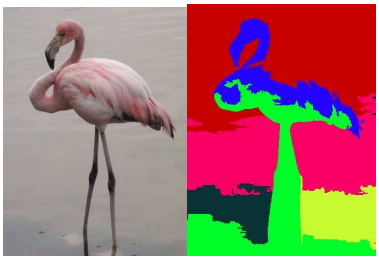
Fig. 1. Each image of the database has been split in patches. In this case the image is split in 5x5 patches.

The low-level features extracted on patches are the following:

- **Texture features** used by our system are Gabor features [5]. The resulting feature vector is of dimension 60.
- **Color Features**. Color is characterized by RGBL histograms (32 bins for each component). The resulting feature vector is of dimension 128.

For a patch p_i^k, the numerical feature vector extracted from p_i^k is noted $fe(p_i^k) \in \mathcal{R}^n$. In this case, $n = 60 + 128$.

We also define a similarity measure based on *regions* obtained by image segmentation [6]. An example of image segmentation can be found in fig. 2.



(a) (b)

Fig. 2. An image (a) segmented by the Meanshift segmentation algorithm [6]. The result is a set of regions (b). Obtaining regions that correspond to semantic entities is very challenging and remains an open-problem.

Let $R_i = \{r_i^k\}$, resp. $R_j = \{r_j^l\}$, be the set of regions obtained from segmentation of image i, resp. j. This similarity measure $dR(i,j)$ is defined as following:

$$dR(i,j) = \frac{\sum_{r_i^k \in R_i} min\{L_2(fe(r_i^k), fe(r_j^l))\}_{r_j^l \in R_j}}{card(R_i)}$$

For a region r_i^k, the numerical feature vector extracted from r_i^k is noted $fe(r_i^k) \in \mathcal{R}^n$. Additional low-level features extracted on regions are their *size* and the *position* of their centroids. This implies that in this case, $n = 60+128+1+2$.

We have also used **Local Features** to characterize *fine* details. Note that in this case, patches are not considered. We use bags of SIFT[3] [7] as explained in [8]. A visual vocabulary is built by clustering techniques (k-means). SIFT features are extracted on the whole images database. Key-points are obtained by scale-space extrema localization after Difference of Gaussian (DoG) computation. Then, the k-means algorithm is used to build the visual vocabulary. The number of clusters is set to 150. Once the visual vocabulary has been built, a bag of visterms can be associated with each image of the database. The cosine distance is used to compute the distance between two bags of visterms. The bag of visterm associated with the image i is noted $\mathbf{b_i} \in \mathcal{R}^{150}$.

Similarity Function. The visual similarity between two images i and j, $\delta_I(i,j)$, is defined as following:

$$\delta_I(i,j) = \alpha \times \frac{\sum_{k=1}^{n_p} L_2(fe(p_i^k), fe(p_j^k))}{n_p} + \beta \times \frac{\mathbf{b_i} \cdot \mathbf{b_j}}{|\mathbf{b_i}||\mathbf{b_j}|} + \gamma dR(i,j)$$

In our experiments $\alpha = 0.4$, $\beta = 0.4$, and $\gamma = 0.2$.

[3] Scale Invariant Feature Transform

2.4 Inter-Media Pseudo-Relevance Feedback

User Feedback is a basic way to solve the classic IR term mismatch problem between query and documents. User relevance is used to select relevant top retrieved document which indexing terms are injected into the initial query. This query expansion can be done automatically assuming that the k top ranked documents are relevant: this is called "pseudo-relevance feedback". Pseudo-relevance feedback has been tested for example in [9] as "local feedback" with other local term concurrence technique.

We are concerned about mixed mode queries (text + image) and interested in solving this queries using the two modalities. Other works like [10] pipeline the retrieval on one modality (text), to the other (image). Pseudo-relevance feedback for multimedia document has also been studied in [11].

Our approach is to query both modality in parallel and to apply pseudo-relevance feedback from one modality to the other. For example, the result of the image ranking drives text query expansion through documents. This information is then used to expand the textual query. We call this Inter-Media Pseudo-Relevance Feedback. As the queries contain both image and text, querying can be initiated whether by the text modality *or* by the image modality. Figure 3 illustrates this principle for text query expansion based on the image modality.

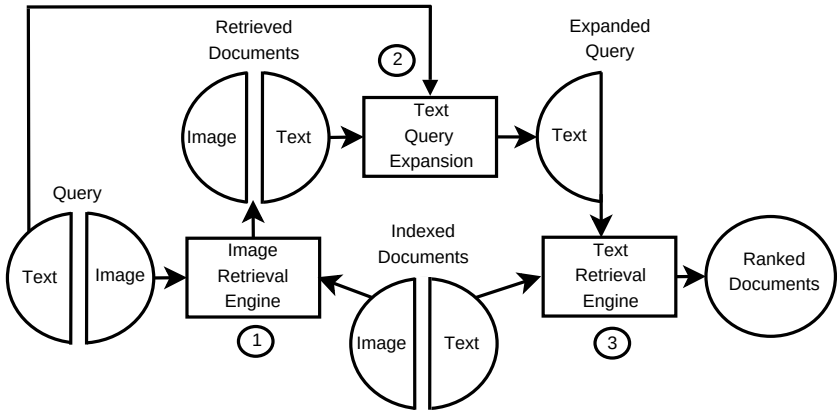


Fig. 3. Overview of *pseudo-relevance* feedback. The textual annotations associated with the top images retrieved by the image retrieval engine are used for query expansion purposes.

In this case, retrieval is achieved in 3 main steps. (1) The initial query is used as an input of the image retrieval engine. The text contained in the query is not involved in the image retrieval process. (2) The textual annotations associated with the top k documents retrieved by the image retrieval engine are then used for query expansion purposes. (3) After this expansion, the resulting

text query is processed by the text retrieval engine to produce the final set of ranked documents. In our experiments, we have set $k = 3$.

2.5 Re-Ranking

We have also integrated re-ranking mechanisms based on the visual appearance (see fig. 4) with the same hypothesis: the top k documents retrieved by text retrieval are relevant. The the k associated relevant images are used to form a class of images which hopefully corresponds to the concept represented by the query.

The goal of re-ranking is to change the rank of the images which are visually similar to the images retrieved by text retrieval. Re-ranking is used as a post-processing step (4) of the pseudo-relevance feedback described in section 2.4.

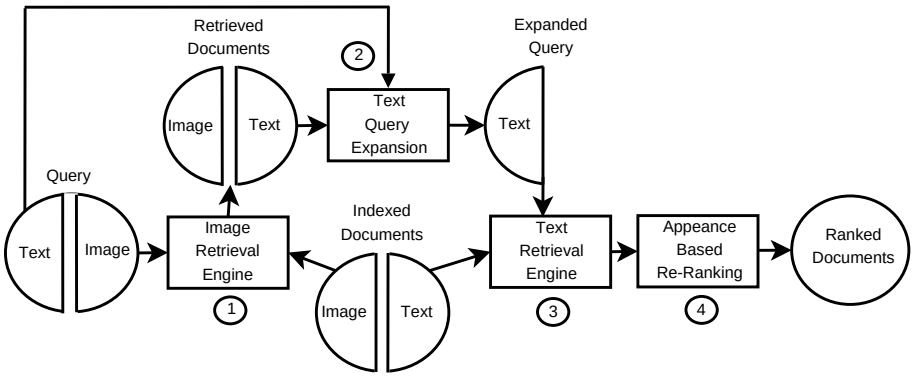


Fig. 4. Re-ranking comes as a post-process of pseudo-relevance feedback.

2.6 Implementation

Both image and text retrieval systems are implemented in C++. For text, basic IR function are part of XIOTA system, dedicated scripts are written in Perl or shell scripts. The image retrieval system heavily relies on the LTI-LIB[4] computer vision library which includes image processing algorithms (e.g. feature extraction, segmentation), machine learning algorithms, and matrix algebra functionalities. This library has a very clean object-oriented design and is very well documented.

[4] http://ltilib.sourceforge.net/

3 Description of the runs submitted

3.1 IPAL-PW

P stands for Part of Speech and W for Single word. Text are first processed as explained in section 2.2. Term filtering is done on part-of-speech. Only nouns, proper nouns, abbreviations, adjectives and verbs are kept. Stemming is provided by the POS tagger. For this run, only the document fields TITLE, DESCRIPTION and LOCATION are used. The weighting is the tf.idf, and ranking is computed with the cosine distance.

3.2 IPAL-PW-PFB3

This run results from the use of the pseudo-relevance feedback described in section 2.4. PFB3 stands for pseudo-relevance feedback involving the three top images retrieved by image retrieval. It is an extra step of the text indexing run IPAL-PW. The textual annotations associated with three images are used for expansion of the text query. It is important to note that the query is expanded from the document index with tf weighting. The query weighting is performed *after* the merge (pseudo feedback). Then it is equivalent to a merge of the original text from the document into the query text. If we consider the use of short query implies being under the Information Retrieval (IR) "subsumption matching" paradigm where relevant document is supposed to *imply* the query; building a query by merging document is closer to the IR "similarly matching" paradigm, where a relevant document is supposed to be *closed* to the query. It is also important to note that Image Retrieval Systems are quite always under the "similarly matching" paradigm.

3.3 IPAL-PW-PFB3-RR60 and IPAL-PW-PFB3-RR300

The re-ranking process described in section 2.5 has been used to produce these runs. Re-ranking was applied on the 60, resp. 300 top images in the documents retrieved by the text retrieval engine for run IPAL-PW-FB3-RR60, resp. IPAL-PW-FB3-RR300. Re-ranking is not applied on the whole set of retrieved images. The reason for that is that images which have a low ranking, share little semantics with query. Even if they are visually similar to the query, they should not be assigned a high rank.

3.4 IPAL-WN

WN stands for Indexing using WordNet concepts. Concepts are extracted from WordNet and used to expand documents and queries. Concepts and original terms are kept in the vector because of the low reliability of concept disambiguation. Hence we have not tested a real full conceptual indexing. Before query re-weighting, a classic tf.idf weighting scheme is applied to all documents and

queries. Query is then split on noun and proper noun. Weighting is then linearly rescaled to maximum 1 on these sub queries. This enables to emphasis the maximum terms in the answer as every term has the same weighting scale. As a consequence, weighting scheme is no more exact tf.idf. Nevertheless, we still use the cosine distance for ranking. Geographical Named Entities are also localized and solved apart in sub queries in the same way. Finally, top documents are those who equally match nouns with concepts, proper nouns and geographical named entities.

3.5 IPAL-WN-MF-LF

This run results from a late fusion (by a weighted-sum) of the output of the text retrieval engine and the output of the image retrieval engine. MF stands for mixed features (described in section 2.3). The principle of late fusion is depicted in fig. 5.

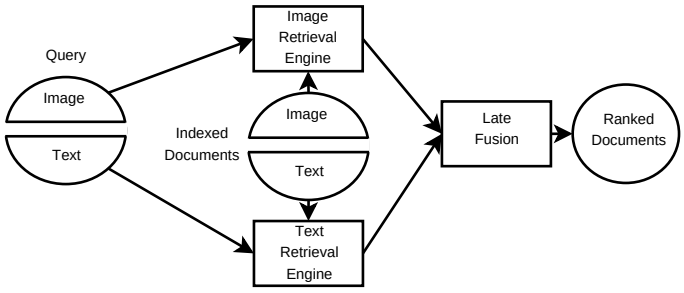


Fig. 5. Principle of late fusion.

3.6 IPAL-EI

EI stands for Equal Importance of all Nouns and Proper Nouns and Noun Phrase. Noun Phrases are computed using WordNet. This run tests the importance of Noun Phrase against other nouns. As tree-tagger does not recognize composed nouns (Noun Phrase), WordNet is used to detect composed nouns with only two nouns (e.g. tennis player, baseball cap, swimming pool) (see 2.2). Each name and proper noun produces a sub query which weight is normalized to 1 in the same way as IPAL-WN. Then sub query results are merged. Then, top documents are those who equally match nouns, proper nouns and noun phrases.

3.7 IPAL-LSA

This run results from Latent Semantic Analysis (LSA) [12] of the image patches. The role LSA is to reduce the effects of synonymy and polysemy by dimension

reduction of a term-document matrix. The resulting reduced space is called the latent space. This run does not use the same features as described in section 2.3.

Indexing is performed as following:

1. Each image is split in 16 non-overlapping patches.
2. From each patch, RGBL histogram (128 bins = 32 * 4) and edge histogram features are extracted.
3. Patches are clustered using k-means clustering algorithm (k=4000). The cluster centroids are also computed.
4. Term-document matrix is computed $A = (a_{ij})\ with\ i = 1, m\ and\ j = 1, n$, where a_{ij} is the number of patches of image j belonging to cluster i. tf-idf is computed from this term-document matrix. In our case, the size of the term-document matrix is 4000×20000.
5. Singular Value Decomposition is applied to the term-document matrix $A = USV^t$. Image coordinates matrix SV^t and a transformation matrix U^t are obtained.

Retrieval is achieved as following:

1. Images in the query are split in 16 non-overlapping patches.
2. From each patch, RGBL histogram (128 bins = 32 * 4) and edge histogram features are extracted.
3. Distance to closest clusters centroids are computed and each patch of each image in the query is assigned to the corresponding cluster. The query (Q) has the same form as a column in the term-document matrix. Tf-Idf is performed on Q.
4. The query is projected into latent semantic space by multiplication with the transformation matrix U^t, $Q_{proj} = U^t Q$.
5. Distance between the query (Q_{proj}) and all images in the database (columns of SV^t) is computed and the images are ranked.

3.8 IPAL-MF

This run was produced by the use of features described in section 2.3. The similarity distance used between two images i and j is $\delta_I(i, j)$.

4 Result Analysis

Mean Average Prevision resulting from each run is summarized in table 1.

IPAL-PW-PFB3 has produced our best Mean Average Precision. In this case, textual information extracted from the 3 top images retrieved by the image retrieval engine are used for text query expansion.

One unexpected result is the degradation of the results (compared to IPAL-PW-PFB3) when applying the appearance based re-ranking algorithm. As expected, mean average precision is lower for the run IPAL-PW-PFB3-RR300 than for the run IPAL-PW-PFB3-RR60. The difference between the two runs is of

Run ID	Run Type	Mean Average Precision
IPAL-PW-PFB3	Mixed	0.3337
IPAL-PW-PFB3-RR60	Mixed	0.2206
IPAL-PW-PFB3-RR300	Mixed	0.1409
IPAL-WN-MF-LF	Mixed	0.0568
IPAL-PW	Text	0.1619
IPAL-WN	Text	0.1428
IPAL-EI	Text	0.1362
IPAL-LSA	Visual	0.0321
IPAL-MF	Visual	0.0173

Table 1. Mean Average Precision (MAP) of submitted runs. The best run is produced by pseudo-relevance mechanisms involving the 3 first images retrieved by the image retrieval engine.

7.9%. Run IPAL-PW-PFB3-RR300 shows that when the number of images considered by re-ranking increases, MAP decreases.

For textual run only, the use of WordNet concepts decrease the MAP. We have not used any of the semantic links provided by WordNet (like hypernym between "bird" and "animal") and we have notice some problem in sense disambiguation (like "church" not recognized as a building). This may explain the lake of improvement. The role of noun phrases seems also not really crucial as IPAL-EI is lower than single terms indexing IPAL-PW. Giving equal importance to single and composed terms is hence a bad idea.

For visual only runs, Latent Semantic Analysis (LSA) leads to slightly better results compared to retrieval based on visual similarity in the feature space. However, mean average precision for visual runs remains very low. Precision at 10 documents (P10) is 0.1417 for run IPAL-LSA and 0.1050 for run IPAL-MF. Precision at 20 documents (P20) is 0.1075 for run IPAL-LSA and 0.0883 for run IPAL-MF.

5 Conclusion

Our approach to this year competition was based on a cooperative approach between an image retrieval and a text retrieval system. These experiments show that the combined use of a text retrieval and an image retrieval systems leads to better performance but only for inter media pseudo relevance feedback and not for late fusion. One surprising aspect of these results is that re-ranking based on visual appearance reduces mean average precision.

The IAPR image database is challenging. Many concepts are represented with a large variety of appearances. Query by content using a few images cannot lead to satisfactory results by using only appearance-based techniques. Indeed, a few samples a given concept cannot capture its conceptual essence.

MAP remains low and is probably still too low to be used in practical conditions. A lot of work has to be done to improve the quality of the system.

5.1 Future Work

We believe that machine learning techniques should be used to obtain an conceptual abstraction of the query images. In this case, the issue is to train the concept detectors. Providing manually a sufficient number of image samples is extremely tedious and does not really scale-up to a large number of concepts. We believe that textual annotations could help building training sets easily and to help raising low-level image features at a semantic level. One of our short-term goals is to apply Latent Semantic Analysis on both image and text modalities. In this case, the size of the resulting term-document matrix is potentially huge and technical problems related to memory management will be encountered. We plan to integrate advanced image interpretation techniques based on prior knowledge on categories of scenes of interest (e.g. indoor, outdoor). We are also interested in adding semi-automatic and ontology-driven feedback and re-ranking.

References

1. Grubinger, M., Clough, P., Mller, H., Deselaers, T.: The iapr benchmark: A new evaluation resource for visual information systems. In: LREC 06 OntoImage 2006: Language Resources for Content-Based Image Retrieval, Genoa, Italy (2006) in press
2. Miller, G.A.: Wordnet: a lexical database for english. Commun. ACM **38** (1995) 39–41
3. Chevallet, J.P.: X-iota: An open xml framework for ir experimentation. In Myaeng, S.H., Zhou, M., Wong, K.F., Zhang, H., eds.: AIRS. Volume 3411 of Lecture Notes in Computer Science., Springer (2004) 263–280
4. Schmid, H.: Probabilistic part-of-speech tagging using decision trees. In: Proceedings of International Conference on New Methods in Language Processing. (1994)
5. Manjunath, B., Ma, W.: Texture features for browsing and retrieval of image data. PAMI **18** (1996) 837–842
6. Comaniciu, D., Meer, P.: Mean shift: A robust approach toward feature space analysis. PAMI **24** (2002) 603–619
7. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision **60** (2004) 91–110
8. Csurka, G., Dance, C., Bray, C., Fan, L., Willamowski, J.: Visual categorization with bags of keypoints. In: ECCV International Workshop on Statistical Learning in Computer Vision, Prague, 2004. (2004)
9. Xu, J., Croft, W.B.: Query expansion using local and global document analysis. In: SIGIR '96: Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, ACM Press (1996) 4–11
10. Alvarez, C., Oumohmed, A.I., Mignotte, M., Nie, J.Y.: Toward cross-language and cross-media image retrieval. In: Working Notes for the CLEF 2004 Workshop. (2004)
11. Yan, R., Hauptmann, A., Jin, R.: Multimedia search with pseudo-relevance feedback. In: Intl Conf on Image and Video Retrieval. (2003) 238–247
12. Landauer, T.K., Foltz, P.W., Laham, D.: Introduction to latent semantic analysis. Discourse Processes **25** (1998) 259–284