Towards Annotation-based Query and Document Expansion for Image Retrieval

Hugo Jair Escalante, Carlos Hernández, Aurelio López, Heidy Marín,
Manuel Montes, Eduardo Morales, Enrique Sucar and Luis Villaseñor

Instituto Nacional de Astrofísica, Óptica y Electrónica,
Luis Enrique Erro No. 1, 72840, Puebla, México,
hugojair@ccc.inaoep.mx

Abstract. In this paper we report results of experiments conducted with strategies for improving text-based image retrieval. The adopted strategies were evaluated in the photographic retrieval task at *ImageCLEF2007*. We propose a Web-based method for expanding textual queries with related terms. This technique was the top-ranked query expansion method among those proposed by other *ImageCLEF2007* participants. We also consider two methods for combining visual and textual information in the retrieval process: *late-fusion* and *intermedia-feedback*. The best results were obtained by combining *intermedia-feedback* and our expansion technique. The main contribution of this paper, however, is the proposal of *"annotation-based expansion"*; a novel approach that consists of using labels assigned to images (with image annotation methods) for expanding textual queries and documents. We introduce this idea and report results of initial experiments towards enhancing text-based image retrieval via image annotation. Preliminary results show that this expansion strategy could be useful for image retrieval in the near future.

1 Introduction

Text-based image retrieval (*TBIR*) consists of using textual image annotations for obtaining images from a given annotated collection; the retrieved images should be relevant to certain user information needs (queries). Under this approach image annotations and queries are considered as small text-documents that are to be compared. Commonly, a measure based on *word matching* is used for determining similarity between query and annotations [1]. The documents that are more similar to the query are returned by the *TBIR* model. This is the predominant approach for image retrieval [2, 3], and most Web image search engines are based on this scheme.

TBIR methods can retrieve images related to high level concepts, (places, events, people and dates), taking advantage of the textual description of the image. This approach, however, is limited because usually textual annotations are very short, complicating the retrieval task. Additionally, *TBIR* methods rely on the quality of annotations, which in most of the cases are not complete. Furthermore, *TBIR* methods do not take into account information extracted from images, wasting useful information that could be useful for improving their accuracy.

This paper describes the participation of *INAOE-TIA*[1] in the photographic retrieval task at *ImageCLEF2007*. Our goal was to explore different methods that could help to improve accuracy of a baseline *TBIR* model. In this respect, we proposed an effective, yet simple, Web-based technique for expanding textual queries. Furthermore, we performed experiments with two widely used methods for combining visual and textual information. The main contribution of this paper, however, is the introduction of *annotation-based expansion* (*ABE*); a novel approach based on image annotation for expanding textual queries and documents. Experimental results show that this strategy could be useful for image retrieval in the near future, though some issues should be addressed first.

The rest of this paper is organized as follows. In the next Section we describe the techniques we considered for improving accuracy of the *TBIR* baseline. In Section 3 we introduce the *ABE* approach. Then, in Section, 4 we present experimental results of the considered methods. Finally, in Section 5 we present some conclusions and discuss future work directions.

2 Improving TBIR performance

In order to evaluate the gain we can have by using the different proposed techniques, we implemented a baseline *TBIR* model based on the *TMG* Matlab^{R} toolbox [4]. After removing meta-data and useless information, the text of the captions in the *IAPR-TC12* collection was indexed separately for the four target languages[2] (English, Spanish, German and Random). For indexing we used a *tf-idf* weighting, English stop words were removed and standard stemming was applied [1, 4]. Queries for the baseline runs were created by using the text in topics as provided by the organizers of *ImageCLEF2007* [5] (after removing meta-data). For multilingual experiments queries were translated using the online Systran[3] translation software. For retrieval we considered the cosine similarity function [1]. In the rest of this Section we present three strategies for improving accuracy of our baseline *TBIR* model.

2.1 Web based query expansion

The Web is the largest repository of information that ever existed; comprising millions of documents, the Web is a very useful source of knowledge. For this reason we consider it in this work by proposing a web-based query expansion technique. The goal is to obtain (and to incorporate) related-context terms, extracted from the Web, according to the original query. The intuitive idea is that expanded queries could be helpful for reaching relevant documents that may contain terms other than the ones contained in the original queries.

[1] Research group on machine learning, image processing and information retrieval at INAOE (http://ccc.inaoep.mx/∼tia)

[2] For further details about the collection, query-target languages and the photographic retrieval task we refer the reader to the respective overview paper [5].

[3] http://www.systranbox.com/

For each topic, we take the textual description and submitted a web-search using the Google^R search engine; the top$-k$ snippets returned by the search engine are considered for expanding a query. We tried two approaches that we called *naive* and *repetition*. The *naive* approach (*NQE*) consists of taking the snippets as they are returned by Google^R with no preprocessing. On the other hand, the *repetition* approach (*RQE*) consists of retaining the most frequent terms in the set of $k-$snippets.

2.2 Intermedia relevance feedback

Intermedia feedback[4] (*IMFB*) is a novel technique based on blind relevance feedback that has been proposed for image retrieval from annotated collections [6]. This technique consists of using a *content-based image retrieval*[5] (*CBIR*) model with a query image for retrieving documents. The top$-n$ documents returned are assumed to be relevant and the captions of such documents are combined to create a textual query. The textual query is then used with a *TBIR* model, and the documents returned by such a model are returned to the user. Note that the final textual query can be generated by considering different strategies. In this work we just concatenated the captions of the pseudo-relevant images. There are several variants of the method [6], some of which are published in this proceedings (see Chang et al and Clinchant et al). We tried combined runs of query expansion and *IMFB*, in which we applied first the query expansion technique and then the expanded queries were combined with the captions of the top$-n$ relevant documents, according to the *CBIR* model, for creating the final query for the *TBIR* model. *FIRE* was used as *CBIR* system; using the baseline run provided by the *ImageCLEF2007* organizers [5].

2.3 Late fusion of independent systems

Another way of enhancing *TBIR* accuracy is by adopting another well known mixed retrieval method, late fusion of independent retrievers (*LF*). This method consists of running two retrieval systems using a single (different) modality each. Then, the relevant documents returned by both systems are combined. For this work we adopted a fusion strategy based on the rank of documents according to two different systems we considered. Let T_R being the list of relevant documents, to a textual query, according to our *TBIR* model; documents are ranked in descending order of their relevance. Similarly, let V_R being the list of ranked relevant documents according to a *CBIR* system that uses the topic images as queries. We combined and re-ranked the documents returned by both retrieval systems, generating a new list of relevant documents $LF_R = \{T_R \cup V_R\}$; where each document $d_i \in LF_R$ is ranked according to the score formula given by Equation (1)

$$score(d_i) = \frac{\alpha \times R_{T_R}(d_i) + (1-\alpha) \times R_{V_R}(d_i)}{1_{T_R(d_i)} + 1_{V_R(d_i)}} \tag{1}$$

where $R_{T_R}(d_i)$ and $R_{V_R}(d_i)$ is the position in the ranked list of document d_i according to the *TBIR* and *CBIR* models, respectively. $1_{T_R(d_i)}$ and $1_{T_R(d_i)}$ are indicator functions

[4] Also known as *media mapping* or *transmedia re-ranking.*

[5] In a *CBIR* model, retrieval is done by considering images only. Note that the *IMFB* can start from text, obtaining query images for a *CBIR* system, as well.

that take the value 1 if document d_i is in the list of relevant documents according to the *TBIR* and *CBIR* models respectively, and zero otherwise. The denominator accounts for documents appearing in both lists of relevant documents (T_R and V_R). Documents are sorted in ascending order of their score. Intuitively with this score documents appearing in both sets (visual and textual) will appear at the top of the ranking, considering their position in the independent lists of relevant documents. We tried several values for α and the best results were obtained with $\alpha = 0.9$.

3 Annotation-based document and query expansion

The task of automatic image annotation (*AIA*) consists of assigning textual descriptors (labels) to images (or segments in images), starting from visual attributes extracted from them. *AIA* methods are well suited for un-annotated image collections, where no textual description of the images is available. Usually, after annotation, the generated labels are used for *TBIR*. In this work, however, we propose using *AIA* methods in an already annotated collection, with the goal of expanding the textual queries and/or initial annotations with labels obtained from the content of images. While manual annotations provide semantic information that may not be obtained from the visual content of the image (when/where the picture was taken?, who took the photo?, etcetera); labels obtained with *AIA* (that is, automatic annotations) can provide information about the visual content of the image that may not be explicit in the annotation (are there *sky, trees, clouds or water in the image?*). In consequence both type of annotations are complementary, and this is the basis for *ABE*.

We decided to use region-level *AIA* methods for obtaining the automatic annotations. Region-level methods can provide accurate annotations and spatial context can be used for improving annotation accuracy [7]. The process we followed for *ABE* includes: (i) segmentation and feature extraction, (ii) creating a training set of annotated regions, (iii) building a classifier, (iv) testing it and expanding queries and/or documents. For segmenting the *IAPR-TC 12 benchmark* collection we used the normalized cuts algorithm [8]; which has been used by most of the region-level annotation approaches. In Figure 1 sample images segmented with normalized cuts are shown. As we can see the algorithm works well for some images, isolating single objects; however, for other images, segmentation is not as good, partitioning single objects into several regions.

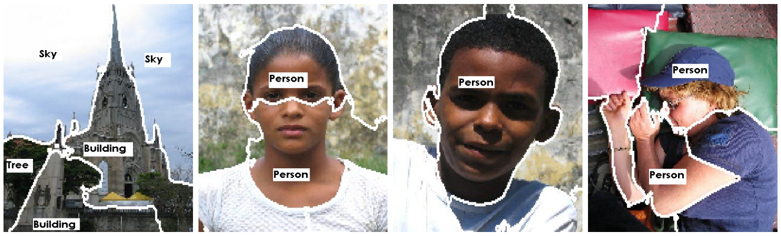

Fig. 1. Sample images from the *IAPR TC-12* collection, segmented with normalized cuts. Manual annotations are shown for each region.

After segmentation, visual attributes were extracted from each region. Attributes include color, texture and shape information of the regions (30 attributes). Each region is described by its vector of attributes. Hereafter we refer to the *attributes vector describing a region* simply by the term *region*. After feature extraction we manually annotated a set of around 2% of the total number of regions. The set of labels that can be assigned to regions was defined subjectively by the authors, by looking at the *ImageCLEF2007* textual topic descriptions. The vocabulary of labels is shown in Table 1, together with the number of regions, in our training set, annotated with each label. Some labels represent several concepts, for example, the label *water* was used for labeling regions of rivers, ocean, and sea. While other labels represent specific objects, such as *swimming-pool* and *tower*. We can see that there are several labels that have many training examples (for example, *Sky, Person*), though several other labels have only a few. This fact together with poor segmentation complicated the process of annotation.

Table 1. Vocabulary of labels considered for the annotation process. The number of instances annotated with each label in our training set is also shown.

Sky (344)	Person (285)	Building (180)	Trees (175)	Clouds (170)	Grass (138)	Water (135)	Mountain (122)
Sand (98)	Other (55)	Furniture (47)	Road (41)	Animal (28)	Snow (25)	Rock (17)	Sun (16)
Vehicle (16)	Boat (14)	Church (9)	Tower (8)	Plate (7)	Flag (4)	Statue (4)	swimming-pool (0)

The training set of region-label pairs is used with a *knn* classifier for annotating the un-annotated regions from the rest of the images. Note that the training set size is very small for achieving good results with the *knn* algorithm. In order to overcome, in part, the issues of poor segmentation and an imbalanced and small training set, we decided to apply a postprocessing to *knn* for improving annotation accuracy. Recently a method, called Markov random field improver (*MRFI*), for improving accuracy on *AIA* has been proposed [7]. *MRFI* considers a set of candidate labels for each region and selects an unique label for each region based on a Markov random field model that considers spatial information, labels association and the confidence of the *AIA* method on each label. We applied *MRFI* as postprocessing to *knn*.

For document expansion we annotated the *20,000* images, and expanded the original annotation with the automatic one. For query expansion we annotated the topic images and expanded the textual topics with the automatic annotations. In Figure 2 an expanded topic is shown (left) , as well as an expanded document (right). As we can see, some labels are repeated on the expanded topic (*sky, people and tree*); we considered repeated labels in order to have an impact in the *tf-idf* weighting, (that is, repeated terms are considered more representative of the query).

4 Experimental results

A total of 95 runs were submitted to *ImageCLEF2007* comprising all of the target languages and most of the query ones. The above described methods were tested, some

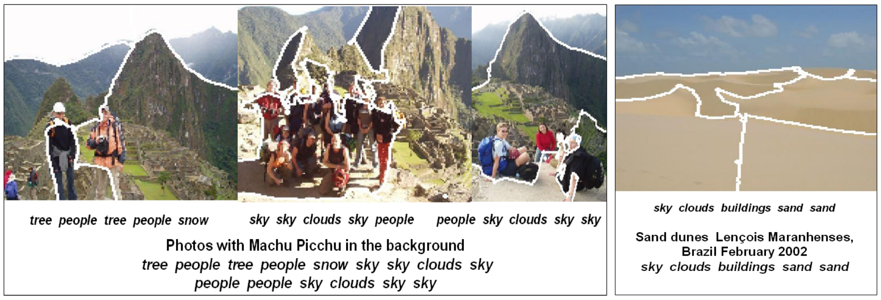
tree people tree people snow
sky sky clouds sky people
people sky clouds sky sky
Photos with Machu Picchu in the background
tree people tree people snow sky sky clouds sky
people people sky clouds sky sky
sky clouds buildings sand sand
Sand dunes Lençois Maranhenses,
Brazil February 2002
sky clouds buildings sand sand

Fig. 2. Left: expansion of the topic 36 using annotations. Right: A sample document expanded with *ABE*. Automatic annotations are shown below each segmented image. The expanded query/document is shown below images annotations.

runs are a combination of these methods. Our top ranked entries for each language configuration together with a brief description of the methods used are shown in Table 2.

Table 2. Top ranked entries for each of the query-target language configurations comprised in the *TIA's* submitted runs. In marked runs ($*$) *TIA* was the only participant group. The last column shows the percentage of improvement over the respective (monolingual) baseline *TBIR* model.

Run-ID	**Languages**	**Methods**	**Type**	**MAP**	**Ranking**	**Improvement** (%)
1	English-English	NQE+IMFB	Mixed	0.1986	22 / 142	43.3
2	Dutch-English*	NQE+IMFB	Mixed	0.1986	1 / 4	43.3
3	French-English	NQE+IMFB	Mixed	0.1986	3 / 21	43.3
4	German-English	NQE+IMFB	Mixed	0.1986	3 / 20	43.3
5	Italian-English	NQE+IMFB	Mixed	0.1986	3 / 10	43.3
6	Japanese-English	NQE+IMFB	Mixed	0.1986	2 / 6	43.3
7	Portuguese-English	NQE+IMFB	Mixed	0.1986	2 / 9	43.3
8	Russian-English	NQE+IMFB	Mixed	0.1986	2 / 6	43.3
9	Spanish-English	NQE+IMFB	Mixed	0.1986	2 / 9	43.3
10	Visual-English*	NQE+ABE+IMFB	Mixed	0.1925	1 / 1	38.9
11	German-German	NQE+LF	Mixed	0.1341	13 / 30	44.5
12	English-German	NQE+LF	Mixed	0.1113	11 / 17	19.9
13	Spanish-Spanish	NQE+LF	Mixed	0.1481	5 / 15	7.71
14	English-Spanish	NQE+LF	Mixed	0.1145	2 / 6	-16.7
15	Dutch-Random*	NQE	Text	0.0828	1 / 2	10.2
16	English-Random	NQE+IMFB	Mixed	0.1243	6 / 11	65.5
17	French-Random	NQE+IMFB	Mixed	0.1243	3 / 10	65.5
18	German-Random	NQE+IMFB	Mixed	0.1243	4 / 11	65.5
19	Italian-Random*	NQE	Text	0.0798	1 / 2	6.26
20	Portuguese-Random*	NQE	Text	0.0296	1 / 2	-60.4
21	Russian-Random*	NQE	Text	0.0763	1 / 2	1.6
22	Spanish-Random*	NQE+ IMFB	Mixed	0.1243	1 / 5	65.5

As we can see, most of the entries are ranked near the first one, and most of them outperform significantly the *TBIR* baseline (column 7). The larger improvement is of around 65%, which is a significant improvement over the *TBIR* baseline. We had some negative results, though we should emphasize that all runs (including bilingual) are compared to a monolingual *TBIR* model. For example the 14^{th} run was compared to a Spanish-Spanish *TBIR* model. It is clear that translation mistakes can degrade the performance in these runs.

The best performance overall runs was obtained by using *IMFB* together with *NQE*s. Actually the *NQE* is present in all of the top ranked runs. *NQE* outperformed *RQE* in all of the language configurations, and according to the official results *NQE* was the best technique among those other proposed for query expansion. This is a surprising result because with *NQE* several noisy terms are added to the queries. While with *RQE* only the terms that most appear among all the snippets are added. The good results of *NQE* are due to the inclusion of many highly related terms, while the insertion of some noisy terms does not affect the performance of the retrieval model.

We can observe that the runs with *IMFB+NQE* for target language English have exactly the same *MAP* value, independently of the query language. This means that the generated queries were dominated by *IMFB*. *IMFB* outperformed the *LF* method in all of the runs if we consider the *MAP*. However, an interesting finding is that *LF* obtained higher recall than any other method we tried, retrieving 16% more documents that *IMFB*. This means that the ranking strategy we adopted for *LF* should be improved.

Six runs based on *ABE* were submitted to the *ImageCLEF2007*. In these runs document and query expansion were combined with the other techniques proposed in previous sections. The descriptions of the annotation based expansion (*ABE*) runs submitted to *ImageCLEF2007* are shown in Table 3. Run 1 in Table 3 is the same as run 10 in Table 2. This is an interesting run because we start from query images only, and by *ABE* and *IMFB* we build a textual query that is used with a *TBIR* model. This approach is language independent as it starts from images only, therefore, it could be very helpful for cross-lingual image retrieval. This was the only run for the language configuration visual-English.

Table 3. Settings of the *ABE* runs. An **X** indicates that the corresponding technique is used. *ABQE* is for *annotation-based query expansion* and *ABDE* is for *annotation-based document expansion*. The ranking position is shown. Diff. is the accuracy we gain-loss with respect of using only the *methods* of column 2 without *ABE*. The last column show the percentage of improvement with respect to the *TBIR* baseline.

ID	Methods	ABQE	ABDE	Ranking	Diff. without ABE	Improvement (%)
1	Baseline,IMFB	**X**	**-**	57	-0.0061	38.9
2	Baseline,IMFB	**X**	**X**	58	-0.0061	38.9
3	NQE,LF,	**X**	**X**	84	-0.0011	22
4	NQE,Baseline	**X**	**X**	133	-0.0011	11.7
5	NQE,LF	**X**	**-**	389	-0.0927	-44.1
6	Baseline	**X**	**X**	447	-0.1115	-79.5

Results with *ABE* are mixed. The two top ranked runs with *ABE* correspond to entries that used *ABE+IMFB*. One should note that with *ABE* we have an insignificant loss of accuracy. In consequence, the favorable result is due to the *IMFB* performance instead of the *ABE* technique. The third *ABE* ranked run used *ABE* of documents and queries with *NQE+LF* which obtained a slightly lower *MAP* than *NQE+LF* without *ABE*. Therefore no gain can be attributed to the *ABE* technique. The other *ABE* runs were ranked low. We should emphasize that this was our very first effort towards developing annotation based methods for improving image retrieval. Several issues should be addressed first in order to evaluate the added value of *ABE*, these are: using better segmentation tools, creating a large and balanced training set of annotated regions, defining a better suited vocabulary for annotation and trying other *AIA* methods instead of *knn*.

5 Conclusions

We have presented experimental results obtained with different strategies for improving *TBIR* methods. An effective Web expansion method was proposed and we tried two widely known mixed retrieval methods. Furthermore, we proposed *ABE* and performed initial experiments with it. *ABE* is a novel technique that may be useful for mixing visual and textual information for image retrieval. Experimental results give evidence that most of the methods we considered improved accuracy of a *TBIR* baseline (up to 65%). The best runs were those based on *IMFB+NQE*. The *NQE* method was the top ranked query expansion method among those proposed by other participants. *IMFB* outperformed *LF* by a large margin in *MAP,* though *LF* obtained higher recall. Results with *ABE* give evidence that *AIA* methods could be helpful for image retrieval from annotated collections. This because promising results were obtained even when segmentation was poor, the training set was extremely small and imbalanced, annotations did not covered the objects present within the image collection and we used a very simple classifier. For future work we will address all of these issues and we will perform extensive experimentation for evaluating the advantages/disadvantages of *ABE*.

Acknowledgements. We would like to thank M. Grubinger, A. Harbury, T. Desealers, P. Clough and all other organizers of *ImageCLEF2007*. This work was partially supported by CONACyT under project grant 61335.

References

1. R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Pearson E. L., 1999.
2. A. Goodrum. Image information retrieval: An overview of current research. *Journal of Informing Science*, 3(2), 2000.
3. P. Clough, M. Grubinger, T. Deselaers, A. Hanbury, and H. Müller. Overview of the imageclef 2006 photographic retrieval and object annotation tasks. In *Working Notes of the CLEF*, 2006.
4. D. Zeimpekis and E. Gallopoulos. Tmg: A matlab toolbox for generating term-document matrices from text collections. In *Grouping Multidimensional Data: Recent Advances in Clustering*, pages 187–210. Springer, 2005.
5. P. Clough, M. Grubinger, T. Deselaers, A. Hanbury, and H. Müller. Overview of the imageclef 2007 photographic retrieval task. In *Working Notes of the CLEF*, 2007.
6. Y. Chang and H. Chen. Approaches of using a word-image ontology and an annotated image corpus as intermedia for cross-language image retrieval. In *Working Notes of the CLEF*, 2006.
7. H. J. Escalante, M. Montes, and L. E. Sucar. Word co-occurrence and mrf's for improving automatic image annotation. In *Proc. of the 18th British Machine Vision Conference (BMVC 2007)*, Warwick, UK, September, 2007.
8. J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.