Annotation-Based Expansion and Late Fusion of Mixed Methods for Multimedia Image Retrieval

H. Jair Escalante, J. Antonio Gonzalez, Carlos A. Hernández, Aurelio López, Manuel Montes, Eduardo Morales, Luis E. Sucar and Luis Villaseñor
{hugojair,jagonzalez,carloshg,allopez,
mmontesg,emorales,esucar,villasen}@inaoep.mx

Instituto Nacional de Astrofísica, Óptica y Electrónica
Luis Enrique Erro No. 1, 72840, Puebla, México

Abstract. This paper describes experimental results of two approaches to multimedia image retrieval: *annotation-based expansion* and *late fusion of mixed methods*. The former formulation consists of expanding manual annotations with labels generated by automatic annotation methods. Experimental results show that the performance of text-based methods can be improved with this strategy, specially, for visual topics; motivating further research in several directions. The second approach consists of combining the outputs of diverse image retrieval models based on different information. Experimental results show that competitive performance, in both retrieval and results diversification, can be obtained with this simple strategy. It is interesting that, opposed to previous work, the best results of the fusion were obtained by assigning a high weight to visual methods. Furthermore, a probabilistic modeling approach to result-diversification is proposed; experimental results reveal that some modifications are needed to achieve satisfactory results with this method.

1 Introduction

Multimedia image retrieval (MIR) is a problem that has been attracting the interest from diverse communities since the last decade [9]. The interest is increasing because of the availability of cheap devices (e.g. cell phones) able to generate large amounts of images everyday. MIR is more challenging than text-based and content-based image retrieval (TBIR and CBIR, respectively) because two modalities must be handled; the problem is further complicated because of the lack of correspondence between low-level image features (e.g. color and texture) and high-level semantics (e.g. named entities like locations or names). Nevertheless, the availability of information from different modalities, yet making reference to a common document, incite the development of methods able to exploit the diversity, redundancy and complementariness of information. This paper describes experimental results on two novel formulations that follow this line of thinking: *annotation-based expansion* (ABE) and *late fusion of heterogeneous methods* (LFHM); these approaches were developed and evaluated in the context of the photographic retrieval task at ImageCLEF2008, which is described in detail by Arni et al. [10].

The rest of this document is organized as follows. In the next section it is described the annotation-based approach. In Section 3 the LFHM formulation is presented; note that since LFHM is described in detail elsewhere [3], Section 3 is brief in its description; a probabilistic modeling approach to result-diversification is described in this section as well. In Section 4 experimental results are described and analyzed. Finally, in Section 5, the findings and contributions of this paper are summarized and future work directions are outlined.

2 Annotation-based document expansion

Automatic image annotation (AIA) is the task of assigning semantic labels to images [9]; it has been recognized as one of the *hot topics* on MIR. The ultimate goal of AIA is to allow un-annotated image collections to be searched by keywords; however, the usefulness of AIA methods should not be limited to un-annotated collections as shown in this paper and in previous work by the authors [5]. In this work region-level AIA methods were used to expand the manual annotations of images. The underlying idea is to represent documents by considering both their high-level (given by manual annotations of images) and low-level (given by labels automatically assigned to images) semantics; and then using this representation for MIR. The ABE approach is depicted in Figure 1.

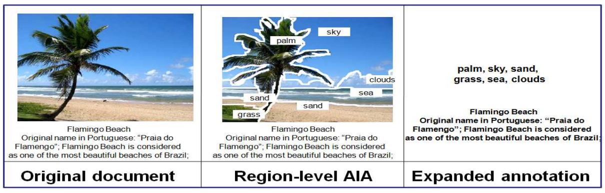

Fig. 1. Diagram of the ABE approach.

ABE was first proposed by Escalante et al. in the framework of ImageCLEF2007[1] [5]; however, the size and quality of the training data used for annotation prevented the authors of deriving concluding facts on the usefulness of automatic labels on image retrieval. In this paper we consider a much better collection to train AIA methods: a subset[2] of *the segmented and annotated IAPR-TC12 benchmark* [4]. This subset is composed of about 7,000 manually segmented and annotated images from the IAPR-TC12 collection; sample images are shown in Figure 2. Only the regions annotated with the 100 most common labels were considered; resulting in about 37,000 regions that are described by

[1] Note that some participants at ImageCLEF2008 adopted a similar approach for expanding manual annotations with visual concepts [6].

[2] This is an extension to the IAPR-TC12 benchmark that will allow to study the impact of AIA methods on MIR [4]; see, http://ccc.inaoep.mx/∼tia/saiapr.

the following features: area, boundary/area, width and height of the region, average and standard deviation in x and y, convexity, average, standard deviation and skewness in both color spaces RGB and CIE-Lab.

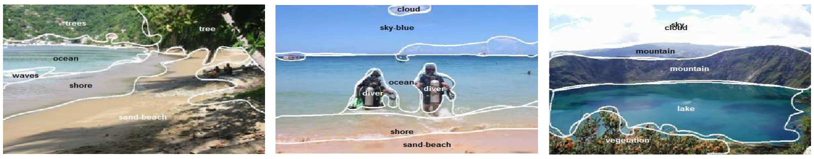

Fig. 2. Sample images from the segmented and annotated *IAPR-TC12* collection [4].

The 20,000 images in the IAPR-TC12 benchmark [10] were automatically segmented using the normalized nuts algorithm and the above features were extracted from each region. Using the subset of annotated regions together with a classifier all of the regions in the segmented collection were automatically labeled. For each image, the generated labels (manual labels were used for images in the training subset) were used as expansion of the original annotation, see Figure 1. The expanded annotation was considered a textual document and a text-based retrieval model was used for indexing the documents; the textual statement in each topic was used as query for the retrieval model. Based on previous work we selected as retrieval engine a vector space model (VSM) with a combination of augmented-normalized term-frequency and entropy for indexing/weighting documents [5, 3]. The TMG-Matlab^R toolbox was used for the implementation of all of the text-based methods considered in this work [2]. For annotation a simple knn classifier was used; additionally, it was considered a method for improving the quality of the knn annotations. This postprocessing method (referred to as MRFS) is based on a Markov random field that uses spatial relationships between connected regions for maximizing the annotation coherence for each image [1]. The energy function of this random field takes into account a relevance weight obtained from knn and probabilities that reflect association between labels and spatial relationships.

3 Late fusion of heterogeneous retrieval methods

Late fusion of independent retrieval methods is one of the simplest and most widely used approaches to combine visual and textual information for MIR [7, 3]. The approach consists of building several retrieval systems (i. e. independent retrieval models, hereafter IRMs) based on different information from the same collection of documents. At querying time, each IRM returns a list of documents relevant to a given query. The output of the different IRMs is combined to obtain a single list of ranked documents, see Figure 3 left.

In this work it was considered the combination of multiple heterogeneous IRMs through the late fusion fusion approach (i.e. LFHM). Heterogeneousness

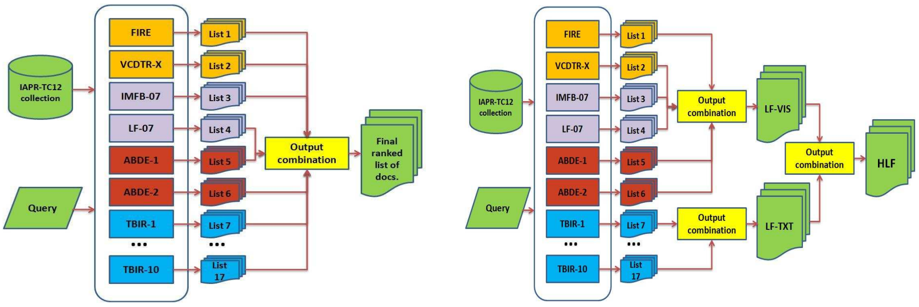

Fig. 3. Illustration of the configurations considered with LFHM. Left: straight fusion. Right: per-modality and hierarchical LFHM; LF-VIS and LF-TXT are the resultant lists of the fusion of visual and textual IRMS, respectively; HLF is the resultant list of hierarchical LFHM [3].

in IRMs has proved being important to improve the fusion results by providing complementary and diverse, yet redundant, lists of documents to the fusion [3]; the inclusion of many IRMs (the largest number considered so far to the best of our knowledge) contributed in the same directions as well, although mostly in redundancy. The lists of ranked documents are combined by assigning a score W to each document d_j as follows:

$$W(d_j) = \Big(\sum_{i=1}^{N} \mathbf{1}_{d_j \in L_i}\Big) \times \sum_{i=1}^{N} \Big(\alpha_i \times \frac{1}{\psi(d_j, L_i)}\Big) \tag{1}$$

where i indexes the N available lists of documents $L_{\{1,\ldots,N\}}$; $\psi(x, H)$ is the position of document x in ranked list H; $\mathbf{1}_a$ is an indicator function that takes the unit value when a is true and α_i ($\sum_{k=1}^{N} \alpha_k = 1$) is the relevance weighting for IRM i, when using hierarchical LFHM. Each list L_i is the output of one of the IRMs we considered, these are shown in Table 1. Documents are re-ranked in descending order of this score, and the top$-x$ documents are kept.

Different configurations for LFHM were considered: *simple* is the straight fusion of IRMs as depicted at the left of Figure 3; *per-modality* is the combination of IRMs based on the same modality; specifically, IRMs that use text only (LF-TXT) and IRMs that use images (LF-VIS) were fused separately, both configurations are shown at the right of Figure 3; finally, *hierarchical* LFHM (abbreviated HLF), is the fusion of the already fused lists LF-TXT and LF-VIS, as shown at the right of Figure 3; for HLF different weights were assigned to the textual (LF-TXT) and visual lists (LF-VIS), see Section 4.

In order to diversify the results of LFHM, an approach based on latent Dirichlet allocation (LDA) [8] was developed. LDA is a probabilistic modeling tool widely used in text analysis; it assumes documents are mixtures of unknown LDA-topics, which are nothing but (learned) probability distributions of words over documents, characterizing semantic themes. Since documents are mixtures of topics one can calculate the probability of each document given an LDA-topic $P(\mathbf{w}|z_i)$; thus, we associate each document $\mathbf{w}$ to the topic that maximizes the latter probability. In this way, each document is associated to a single LDA-topic,

ID	Name	Modality	Description
1	FIRE	IMG	The FIRE CBIR system [11] (rk. 377/474 [7])
2	VCDTR-X	IMG	Boolean TBIR build on the visual concepts provided by XRCE [6]
3	IMFB-07	TXT+IMG	Our best entry, in MAP, at ImageCLEF2007, see [5] (rk. 41/474 [7])
4	LF-07	TXT+IMG	Our best entry, in recall, at ImageCLEF2007, see [5] (rk. 82/474 [7])
5	ABDE-1	TXT+IMG	A TBIR that implements ABE as described in Section 2 (knn)
6	ABDE-2	TXT+IMG	A TBIR that implements ABE as described in Section 2 (MRFS)
7	TBIR-1	TXT	TBIR model based on the VSM representation and t/f weighting
8	TBIR-2	TXT	TBIR model based on the VSM representation and n/e weighting
9	TBIR-3	TXT	TBIR model based on the VSM representation and a/g weighting
10	TBIR-4	TXT	TBIR model based on the VSM representation and a/e weighting
11	TBIR-5	TXT	TBIR model based on the VSM representation and n/g weighting
12	TBIR-6	TXT	TBIR model based on the VSM representation and t/g weighting
13	TBIR-7	TXT	TBIR model based on the VSM representation and n/f weighting
14	TBIR-8	TXT	TBIR model based on the VSM representation and a/f weighting
15	TBIR-9	TXT	TBIR model based on the VSM representation and t/e weighting
17	TBIR-10	TXT	TBIR model based on the VSM representation and t/g weighting

Table 1. List of IRMs considered in this work . From rows 7 and on, column 4 describes the local/global weighting schemas for a VSM. Abbreviations are as follows: t, term-frequency; f, inverse document-frequency; n, augmented normalized term-frequency; e, entropy; a, alternate log; g, global-frequency/term-frequency; l, logarithmic frequency, see [3] for details. For rows 1,3 and 4 it is specified the rank (rk.) position of the respective entry at ImageCLEF2007 [7].

which can be considered a cluster. To diversify retrieval results it was considered the set of documents returned by LFHM (any configuration) to a query-topic. We used the toolbox of Steyvers et al. to obtain k LDA-topics from such set [8]; k was fixed to 20 because diversification at 20 documents was evaluated in ImageCLEF2008. Documents were grouped according the LDA-topic they belong to and a single document was selected from each LDA-topic as representative of it. The representative document was selected according its relevance weight in the list of ranked documents returned by the LFHM method. The k representative documents were placed at the top of a new list and the rest of documents were placed below them according their initial relevance weight.

4 Experimental results

This section describes results of ABE and LFHM in the photographic retrieval task at ImageCLEF2008; the following evaluation measures were considered: precision (**p20**) and cluster-recall (**c20**) at 20 documents retrieved, mean average precision (**MAP**) and number of relevant documents retrieved (**Rel-Ret**).

4.1 Annotation-based expansion

Results with different configurations of ABE are shown in Table 2. In order to illustrate the advantages of this approach, are shown results over all topics (as provided by organizers [10]) and over visual/textual[3] topics (as categorized

[3] Visual topics are those queries well suited to be answered by using information from images content (e.g. *"night shots of cathedrals"*); textual topics, on the other hand, are those queries that require using textual information to effectively retrieve documents (e.g. *"Destinations in Venezuela"*).

in ImageCLEF2006). *Baseline* is a TBIR that uses the original annotations; *Manual* uses annotations expanded with the labels from our training set (i. e. only were considered manually assigned labels); *KNN* uses annotations expanded with labels of the training set plus labels assigned with knn; *KNN-MRFS* is the same as KNN though labels assigned by knn were improved with MRFS.

Method	**MAP-A**	**P20-A**	**C20-A**	**MAP-V**	**P20-V**	**C20-V**	**MAP-T**	**P20-T**	**C20-T**
Baseline	0.2625	0.3295	0.3493	0.2916	0.3500	0.3378	**0.2334**	**0.3090**	**0.3608**
Manual	**0.2648**	0.3333	0.3510	**0.3085**	0.3700	0.3574	0.2211	0.2966	0.3446
KNN	0.2554	**0.3397**	0.3582	0.2943	**0.3775**	0.3648	0.2165	0.3019	0.3516
KNN+MRFS	0.2546	0.3295	**0.3733**	0.2971	0.3750	**0.3942**	0.2121	0.2840	0.3524

Table 2. Performance of different configurations with ABE evaluated in ImageCLEF2008. It is shown the performance over all topics (-A); over visual topics (-V) and over textual topics (-T), see the text; the best results are shown in **bold**.

From this table one can see that, as expected, the *Baseline* method obtained the best results over textual topics (columns 8–10); although over all (columns 2–4) and, mainly, over visual (columns 5–7) topics, ABE configurations consistently outperformed the baseline. Among ABE entries the best **MAP** is achieved with *Manual*; this can be due to the fact that for this setting the expanded labels were always correct. The best results on **P20** were obtained with the *KNN* technique; this means that *KNN*-labels resulted more helpful for placing relevant documents at the first positions. Finally, the best result in **c20** were obtained with the *KNN+MRFS* approach. Note that the differences may appear small in number, however, ABE can provide a significant advantages to users of MIR systems. Figure 4 shows the relevant-retrieved images to a visual-topic in the first 20 positions for ABE entries; this figure illustrates the advantages offered by using the ABE approach. It can be seen that more images are retrieved by using ABE methods; which improves the performance in all of the considered measures, note that diversity is significantly improved with *KNN-MRFS*; results were likewise for all visual topics. Results shown in Table 2 give evidence that

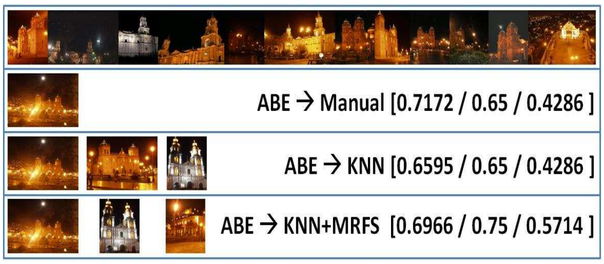
ABE → Manual [0.7172 / 0.65 / 0.4286]
ABE → KNN [0.6595 / 0.65 / 0.4286]
ABE → KNN+MRFS [0.6966 / 0.75 / 0.5714]

Fig. 4. Relevant-retrieved images at the top-20 positions for the topic #15 ("*night shots of cathedrals*"). The top row shows results when using manual annotations only; rows 2-4 show images retrieved with ABE-Manual, ABE-KNN and ABE-KNN-MRFS, respectively. *Note that ABE methods also retrieved the images from the top row.* For ABE-runs it is shown the **MAP**, **P20** and **c20** for this topic; the respective values for the baseline are the following [0.6818 / 0.6 / 0.4286].

the use of labels generated with AIA methods can be helpful to improve the performance of TBIR methods on both retrieval and diversification of results; specially for visual topics. Note that ABE is the simplest way of taking advantage of automatic labels; therefore, better results are expected by using more sophisticated strategies. Also, it is possible to notice that despite the performance of ABE entries is limited (when compared to other multi-modal methods) these methods resulted very useful when their outputs were combined with the LFHM approach, see below and refer to [3].

4.2 Late Fusion of mixed methods

Results obtained with different configurations of LFHM are shown in Table 3. It can be seen that performance of all of the configurations is quite competitive. The best result was obtained by using the HLF approach assigning a weight of 0.8 to visual methods and of 0.2 to textual ones (row 6). This is a very interesting result, opposed to previous work where higher weight to textual methods results on improved performance; this is due to the fact that visual methods are indeed a mixture of CBIR and MIR strategies. It was also interesting that the inclusion of low-performance IRMs (e.g. FIRE and VCDTR-X) resulted beneficial to the LFHM approach, see [3] for further details. Note that the recall of *HLF-0.8/0.2* was among the top-3 over all (1042) ImageCLEF2008 entries; giving evidence of the potential advantages offered by LFHM and that a better strategy for re-ranking documents is required. We would like to emphasize that the considered IRMs (shown in Table 1) are not the best methods one can try and better results are expected by using IRMs of better individual performance.

Run	**p20**	**MAP**	**c20**	**Avg.**	**Rel-Ret**	**+LDA**
Simple	0.3782	0.3001	0.4058	0.3613	1946	0.4291
LF-TXT	0.341	0.2706	0.3815	0.3311	1885	0.3335
LF-VIS	**0.4141**	0.2923	0.3864	0.3642	1966	0.3941
HLF-0.5/0.5	0.3795	0.303	0.3906	0.3577	1970	0.3721
HLF-0.8/0.2	0.391	**0.3066**	0.4033	**0.3667**	**1978**	0.3976
HLF-0.2/0.8	0.3731	0.2949	**0.4175**	0.3619	1964	0.4132

Table 3. Performance of different settings with LFHM; rows 5-7 in column 1 show the weights w_1/w_2 assigned to visual and textual lists, respectively, for HLF; column 7 shows the **c20** performance after applying the LDA diversification technique.

Column 7 in Table 3 shows the **c20** performance after applying the LDA diversification strategy. The application of such technique improved the **c20** performance of *Simple* and *LF-VIS*, although it decreased the performance of the rest. However, it is important to mention that the **MAP** and **P20** of all of the results was significantly decreased by using the LDA method. This can be due to several factors that motivate further research with this approach; namely, the top 1000 results were considered for clustering, which introduced too much noise; the ranking of LFHM is not the best approach to select representative documents; the initial list of documents may not be correct enough; and the restriction of generating k=20 clusters may be inappropriate.

5 Conclusions

We have described experimental results on two novel approaches to the MIR task: ABE and LFHM. Experimental results with ABE provide evidence that indicates the use of AIA labels can be helpful to improve the performance of TBIR methods. This is an interesting result, because even with a very simplistic approach we were able to improve both retrieval performance and diversification of results. As expected, the use of labels resulted particularly helpful for visual topics. On the other hand, results obtained with LFHM show that competitive performance can be obtained with this method; even when late fusion is the simplest approach one may try to MIR and when the considered IRMs were not the best retrieval methods one can try. Both formulations motivate further research in several aspects; namely, studying different strategies to combine manual and automatic annotations; improving the performance of AIA methods; applying LFHM with better IRMs and different fusion strategies.

Acknowledgements. The authors would like to thank the organizers of *ImageCLEF2008* for their support. This work was partially supported by CONACyT under project grant 61335 and scholarship 205834.

References

1. C. Hernández et al. Mrfs and spatial information to improve automatic image annotation. volume 4872 of *LNCS*, pages 879–892. Springer, 2007.
2. D. Zeimpekis et al. Tmg: A matlab toolbox for generating term-document matrices from text collections. In J. Kogan et al., editor, *Grouping Multidimensional Data: Recent Adv. in Clustering*, pages 187–210. Springer, 2005.
3. H. Jair Escalante et al. Late fusion of heterogeneous methods for multimedia image retrieval. In *Proc. of MIR'08*, Vancouver, BC, Canada, October 2008. ACM.
4. H. Jair Escalante et al. The segmented and annotated *IAPR-TC12* benchmark. *Submitted to CVIU*, 2008.
5. H. Jair Escalante et al. Towards annotation-based query and document expansion for image retrieval. volume 5152 of *LNCS*, pages 546–553. Springer, 2008.
6. J. Ah-Pine et al. Xrces participation to imageclef 2008. In C. Peters et al., editor, *Proc. 9th Workshop of the CLEF*, LNCS, Aarhus, Denmark, September 2008 (printed in 2009).
7. M. Grubinger et al. Overview of the ImageCLEF 2007 photographic retrieval task. In *Advances in Multilingual and Multimodal Information Retrieval, 8th Workshop of the CLEF 2007*, volume 5152, Budapest, Hungary, September 2007.
8. M. Steyvers et al. *Latent Semantic Analysis: A Road to Meaning*, chapter Probabilistic topic models. Laurence Erlbaum, 2007.
9. R. Datta et al. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*, 40(2), 2008.
10. T. Arni et al. Overview of the 2008 photographic retrieval task. In C. Peters et al., editor, *Proc. 9th Workshop of the CLEF*, LNCS, Aarhus, Denmark, September 2008 (printed in 2009).
11. T. Gass et al. Fire in imageclef 2007: Svms and logistic regression to fuse image descriptors in for photo retrieval. volume 5152 of *LNCS*. Springer, 2008.