Late Fusion of Heterogeneous Methods for Multimedia Image Retrieval*

Hugo Jair Escalante
National Institute of Astrophysics, Optics and Electronics
Luis Enrique Erro # 1
72840, México
hugojair@ccc.inaoep.mx

Carlos Hérnadez
National Institute of Astrophysics, Optics and Electronics
Luis Enrique Erro # 1
72840, México
carloshg@ccc.inaoep.mx

L. Enrique Sucar
National Institute of Astrophysics, Optics and Electronics
Luis Enrique Erro # 1
72840, México
esucar@inaoep.mx

Manuel Montes
National Institute of Astrophysics, Optics and Electronics
Luis Enrique Erro # 1
72840, México
mmontesg@inaoep.mx

ABSTRACT

Late fusion of independent retrieval methods is the simpler approach and a widely used one for combining visual and textual information for the search process. Usually each retrieval method is based on a single modality, or even, when several methods are considered per modality, all of them use the same information for indexing/querying. The latter reduces the diversity and complementariness of documents considered for the fusion, as a consequence the performance of the fusion approach is poor.

In this paper we study the combination of multiple heterogeneous methods for image retrieval in annotated collections. Heterogeneousness is considered in terms of *i*) the modality in which the methods are based on, *ii*) in the information they use for indexing/querying and *iii*) in the individual performance of the methods. Different settings for the fusion are considered including weighted, global, permodality and hierarchical. We report experimental results, in an image retrieval benchmark, that show that the proposed combination outperforms significantly any of the individual methods we consider. Retrieval performance is comparable to the best performance obtained in the context of ImageCLEF2007. An interesting result is that even methods that perform poor (individually) resulted very useful to the fusion strategy. Furthermore, opposed to work reported in the literature, better results were obtained by assigning a low weight to text-based methods. The main contribution of this paper is experimental, several interesting findings are reported that motivate further research on diverse subjects.

Categories and Subject Descriptors

H.3.3 [**Information Systems Applications**]: Information Search and Retrieval—*information filtering, selection process, retrieval models*

General Terms

Experimentation, Performance

Keywords

Late fusion, Image retrieval.

1. INTRODUCTION

Multimedia image retrieval is a challenging task because it requires effectively processing information in two modalities: textual and visual [14, 4]. Since it is not easy to effectively process images and text in such a way that retrieval performance is acceptable, most methods (e. g. $ImageGoogle^R$) have considered a single modality for indexing and searching for images [7, 14, 4]. However, methods based on either text or images perform well only for a certain sort of queries. For this reason, research on multimedia image retrieval has become a very active and relevant research area [7, 14, 4]. Several methods have been proposed so far, including the fusion of retrieval methods [15], techniques that use relevance feedback [3], indexing of heterogeneous vectors of features [16], and a suite of ad-hoc and very complex methods for specific collections of images [14, 4]. The problem with most of these methods is that, despite being very competitive, they are usually very difficult to implement/reproduce and to use them in practice.

Late fusion of independent retrieval models (LFIRM) is one of the simplest and most widely used approaches for combining visual and textual information in the retrieval

*This work was supported by CONACyT under project grant 61335.

MIR '08 Vancouver, BC, CANADA

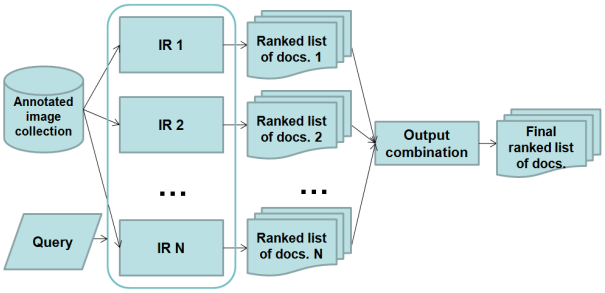

Figure 1: Graphical diagram of the LFIRM approach. The output of different IRMs is combined for obtaining a single list of ranked documents.

process [15, 4, 11, 2, 10]. This approach consists of building several retrieval systems (i. e. independent retrieval models, hereafter IRM) using subsets of the same collection of documents. At querying time, each IRM returns a list of documents tentatively relevant to a given query. The output of the different IRMs is then combined for obtaining a single list of ranked documents, see Figure 1. A common problem with this approach is that often accuracy of the LFIRM method is only slightly superior to the best IRM [15, 4, 2, 10]; also, the output of the best IRM is weighted much more higher to obtain acceptable performance and the performance of the LFIRM will depend on the best IRM [13, 15, 4, 2, 10]. Furthermore, usually in LFIRM methods a single IRM is considered for each modality [13, 15, 4, 2, 10, 17]. The latter fact limits the performance of LFIRM because, despite the potential diversity of documents due to the IRMs, there is little, if any, redundance through the IRMs and therefore the combination is not effective [15, 4]. Some LFIRM systems consider multiple IRMs for each modality, however, most of these IRMs are very homogeneous. That is, these methods are variations of a same retrieval model using different parameters or meta-data for indexing [13, 4, 2, 10].

In this paper we study the combination of heterogeneous IRMs through the LFIRM approach for multimedia image retrieval. Opposed to previous work reported in the literature, our study considers a variety of IRMs that include uni-modal and multi-modal search methods, as well as methods that are different in nature. The IRMs we consider include retrieval methods based on text, image-content, visual-concept detection, region-level image annotation, Web-based query expansion and inter-media relevance feedback. Heterogeneousness is important because it can be useful for providing diverse, complementary and redundant lists of documents to the LFIRM approach, reducing the retrieval problem to that of effectively combining lists of ranked documents. For merging the lists we assign a score to each document in the lists and rank them in descending order of this score. The combined list is formed by keeping the top-k ranked documents. The score we propose is a simple linear combination of the documents positions and the number of lists in which each document appears.

The main contribution of this work is experimental, in consequence our focus is on describing the results of extensive experimentation and highlighting interesting findings that can motivate further research. Our experiments aim to empirically prove the validity of the approach by using the IAPR-TC12 collection [8], an established image retrieval benchmark used in the photographic retrieval track of ImageCLEF[1]. Experimental results show that the heterogeneous LFIRM (HLFIRM) approach significantly outperform any of the IRMs used; using uniform weights for the IRMs gives comparable performance to that of weighting higher the best IRMs. An interesting result is that even methods that perform poorly individually are useful for improving the performance of the HLFIRM. Also, opposed to most previous work on LFIRM for image retrieval, our results give evidence that the best performance of the HLFIRM method is achieved by assigning higher weight to multi-modal IRMs. Our best result is obtained by hierarchically applying the HLFIRM approach, with a comparable performance to that of the best entrant of ImageCLEF2007, even when that entrant included an user-in-the-loop into the retrieval process (Note that we have used the 2008 topics, which are more difficult than those used in 2007).

The rest of this paper is organized as follows. In the next Section we review related work on LFIRM for image retrieval. Then, in Section 3 we describe the IRMs we used and the merging strategy for HLFIRM. Next, in Section 4 we present experimental results that show the validity of HLFIRM. Finally, in Section 5, we discuss the results of this work and outline current and future work directions.

2. RELATED WORK

In multimedia image retrieval, the sources of information are visual features extracted from the image and textual features in the form of associated captions. These sources of information have been mostly used individually and separately. Textual features have proved to be more effective for this task than their visual counterpart, and systems based only on these features tend to significantly outperform systems based merely on visual features, which perform poorly [7, 4]. However, a problem generally found in both cases is the lack of generalization, which makes systems fail with varied sets of queries, see Section 4.1. A recent tendency to fuse visual and textual features has been observed in different evaluation tracks such as TRECVID [12] and ImageCLEF [4], with the belief that these sources of information more than competing are complementary, and that the actual problem may be reduced to finding a way of adequately fusing them.

This kind of fusion is a sort of multimedia information retrieval, and it can be performed either as early or late fusion, in this work we focused on the latter approach. Research on these two directions has already been developed, but current performance of these methods remains poor, showing the need of research to find better fusion alternatives and to select better individual IRMs.

In information retrieval the late fusion approach has been widely used for combining the output of several textual retrieval systems [1, 10, 13]. For multimedia image and video retrieval this approach has been also explored already [11, 17, 15, 4, 10, 13]. However, most approaches have focused on using a single IRM for each modality [4, 10, 13]. Most of these approaches build a text-based IRM and a CBIR

[1]ImageCLEF is a forum for the evaluation of image retrieval systems with emphasis on methods that combine text and images [4].

system separately. At querying time, both systems are run and the output of these methods is merged by using fusion operators (e. g. CombSum, Round Robin, etcetera) well known in the information retrieval community [4, 10, 13]. Usually, the weight of the textual IRM is much higher than that of the visual IRM [15, 4, 10, 13]. However, using a single IRM for each modality does not allow exploiting redundancy among IRMs. This is because IRMs of the different modalities retrieve generally different documents as relevant. Some methods have proposed the use of several IRMs for each modality [15, 13], however all of the IRMs per-modality are homogeneous (in the sense that they use the same retrieval model with variations in its parameters). These approaches do not exploit effectively the diversity we would have by having more heterogeneous IRMs, even for the same modality. As we will see in Section 4, having many homogeneous IRMs does not help the fusion approach because documents retrieved are so similar among them.

In order to overcome the above limitations, in this work we propose the late fusion of heterogeneous IRMs. The advantages of our work is that by using heterogenous IRMs we can exploit more effectively the diversity of retrieval results. Our approach includes the use of several IRMs per-modality (that are also heterogeneous among them) to exploit redundancy in the retrieval results. Opposed to previous work, the IRM that best performs individually does not need to be weighted high. Further, the combination we propose outperforms significantly the performance of the best IRM.

3. PROPOSED APPROACH

The proposed approach to image retrieval in annotated collections is graphically depicted in Figure 1. Each IRM is an independent retrieval method that uses subsets of the document collection for indexing/searching. Because we are interested in providing diverse, complementary and redundant lists of documents to the HLFIRM approach, we consider text-based, image-based and multi-modal retrieval methods. These methods are described in Table 1. For each query, each of these IRMs is run independently. The result is a ranked list of documents for each IRM. Ranked lists of documents are then merged by using the strategy described below. In the rest of this section we describe the IRMs we considered and the combination strategy we propose for HLFIRM.

3.1 Independent retrieval methods

3.1.1 Text-based IRMs

Text-based IRMs (rows 7-15 in Table 1) are variants of the vector space retrieval model (VSM) using different weighting schemas. All of these methods index the available text in image annotations by using different weighting strategies (see Table 1). For querying, these methods use the textual statements of topics (see Section 4.1). For building the textual methods we used the TMG $Matlab^R$ toolbox [18]. We consider ten textual IRMs because in the collection we used it is supposed that text-based methods perform better than methods that use images. However, as we will see in Section 4, this is true only to some extent.

3.1.2 Image-based IRMs

Two image-based methods are considered for HLFIRM, these are FIRE and VCDTR-X. FIRE is a content-based image retrieval (CBIR) system that works under the query-by-example formulation [6]. FIRE uses the sample images from the topics for querying. Since we are only interested on the output of the IRMs we used the FIRE baseline run provided by ImageCLEF2007 organizers [4].

VCDTR-X is a novel IRM that uses image-level annotations assigned to images by using a method developed for ImageCLEF2008. All images (including topic images) are automatically annotated by using this method. The assigned annotations are then used for building a retrieval model with boolean weighting. Queries for VCDTR-X are the automatical annotations assigned to topic images. No information from manual annotations of images and topics was considered. The annotation vocabulary is composed of 17 keywords that describe visual aspects of the images. The annotation method was developed by the Xerox Research Center Europe group (XRCE) for the visual concept detection (VCDT) track at ImageCLEF2008 [to be published]. XRCE kindly provided the image-level annotations to the authors of this paper.

ID	Name	Modality	Description
1	LF-07	TXT+IMG	WQE+LF
2	IMFB-07	TXT+IMG	WQE+IMFB
3	FIRE	IMG	CBIR
4	VCDTR-X	IMG	VCDT
5	ABDE-1	TXT+IMG	ABIR
6	ABDE-2	TXT+IMG	ABIR
7	TBIR-1	TXT	VSM t/f
8	TBIR-2	TXT	VSM n/e
9	TBIR-3	TXT	VSM a/g
10	TBIR-4	TXT	VSM a/e
11	TBIR-5	TXT	VSM n/g
12	TBIR-5	TXT	VSM t/g
13	TBIR-6	TXT	VSM n/f
14	TBIR-7	TXT	VSM a/f
15	TBIR-8	TXT	VSM t/e
17	TBIR-9	TXT	VSM t/g

Table 1: Description of the considered IRMs. From rows 7 and on, column 4 describes the local/global weighting schemas for a VSM. Abbreviations are as follows: WQE, web-based query expansion; IMFB, inter-media relevance feedback; LF, Late fusion; t, term-frequency; f, inverse document-frequency; n, augmented normalized term-frequency; e, entropy; a, alternate log; g, global-frequency/f; l, logarithmic frequency.

3.1.3 Multi-modal IRMs

Four multi-modal IRMs (rows 1-2, 5-6 in Table 1) of different nature were considered for HLFIRM. ABDE methods (rows 5-6) are annotation-based document expansion search techniques [5, 4]. Under this formulation the entire collection of images is segmented and visual features are extracted from the regions. Using a training set of annotated regions all of the regions in the collection are annotated, considering a vocabulary of 222 words arranged in a conceptual hierarchy. The annotation method is based in a Markov random field model that attempts to assign to each region the annotation that minimizes an energy function that considers spatial relationships between regions [9]. The generated labels are then used for expanding the manual annotations of images. The expanded annotations are indexed using a VSM with tf-idf weighting. For querying, ABDE methods use the textual statement of topics. Two variants of the method are considered that differ in the parameters of the annotation

model.

IRMs in rows 1 and 2 of Table 1 are multi-modal methods proposed for the ImageCLEF2007 competition [4, 5]. The first IRM applies inter-media relevance feedback, a technique where the input for a text-based system is obtained from the output of a CBIR system combined with the original textual query [3, 5]. This was our best-ranked entry for ImageCLEF2007, and for that reason we consider it for this work. The second IRM is a LFIRM system that combines the outputs of a textual method and a CBIR system [5]. The textual-method performs Web-based query expansion, a technique in which each topic-statement is used as a query for $Google^R$, the top-20 snippets are attached to the original query. The CBIR system was the FIRE run described in the latter section. This was the run of our group with the highest recall, and that is why we considered for this work.

As we can see we have considered a variety of methods that can offer diversity, redundancy and complementariness of documents, opposed to previous work on LFIRM that use single-modality IRMs.

All of the IRMs are build by the authors, although some of them are based on methods developed by other research groups. Note that methods that have been already evaluated in ImageCLEF2007 are very useful because in this way we can compare our experimental results based on ImageCLEF2008 data.

3.2 Heterogenous late fusion of IRMs

Each time a query q is sent to the HLFIRM method the $N-$IRMs are run separately. The result is a set of N ranked lists (in decreasing order of relevance) of documents. The information of the N ranked lists is used for obtaining a single list of ranked documents, which is returned to the user in response the the query q. The final list is obtained by assigning a score to each document appearing in at least one of the N lists. A high score value indicates that the document is more likely to be relevant to query q. Documents are sorted in decreasing order of their score and the top$-k$ documents are considered for the final list of ranked documents.

For this work we consider a simple (yet very effective) score based on a weighted linear combination of the documents rank through the different lists. The proposed score takes into account redundancy of documents and the individual performance of each retrieval method. Diversity and complementariness are bring to play by the heterogeneousness of the considered IRMs, while redundancy is considered through the use of several IRMs per modality. We assign a score S_{HLFIRM} to each document d_j in at least one of N lists $L_{\{1,\ldots,N\}}$ as described by Equation (1):

$$S_{HLFIRM}(d_j) = \big(\sum_{i=1}^{N} \mathbf{1}_{d_j \in L_i}\big) \times \sum_{i=1}^{N} \big(\alpha_i \times \frac{1}{\psi(d_j, L_i)}\big) \qquad (1)$$

where i indexes the N available lists of documents; $\psi(x, H)$ is the position of document x in ranked list H; 1_a is an indicator function that takes the unit value when a is true and α_i, with $\sum_{k=1}^{N} \alpha_k = 1$, is the importance weighting for IRM i. α_i's allow including prior knowledge into the retrieval process in the form of the confidence we have on each IRM.

Documents appearing in several lists at the top positions will receive a higher score, while documents appearing in a few list or appearing at the bottom positions most of the times will be scored low. Eventually, only relevant documents will be kept. The more relevant a document is to a query the higher will be its position in the final list. As we can see this is a simple and intuitive way of merging the output of IRMs proved to be very useful in practice.

4. EXPERIMENTAL RESULTS

In this section we present experimental results that give empirical evidence of the validity of the HLFIRM method for multimedia image retrieval. Retrieval performance is evaluated by using the following standard measures from information retrieval, mean average precision (MAP), precision at 20 documents and total recall. The top 1000 documents are used for evaluating MAP and total recall. For all of the experiments it is reported the average of these measures over all the queries considered, see Section 4.1.

4.1 Image collection

For our experiments we use the image collection and ground truth data used in ImageCLEF2008. The use of these resources allows a direct comparison of our method with state-of-the-art retrieval systems. This collection is the IAPR-TC12, an established benchmark for the evaluation of image retrieval systems [8]. The collection is composed of around $20,000$ real-images taken from locations around the world and comprising a varying cross-section of still natural images. Each image has an associated (manually assigned) textual annotation that describes, to some extend, the visual and semantic content of the image. Annotations are available in English, German and Spanish (we used English annotations in our experiments).

For querying, multimedia topics are provided, these consist of a textual query statement and three sample images. A sample topic from this collection is shown in Figure 2, and relevant images for this topic are shown in Figure 3. For illustration we show in bold-uppercase the textual statement used for ImageCLEF2007 in Figure 2, for ImageCLEF2008 the full text is used. Topics are used to build queries, for textual queries we use the topic statement as it, while for visual queries we use the images or information from their content (see Section 3.1). Note that since more text is provided in 2008 topics, text-based methods are supposed to perform better than visual-based approaches.

Relevant images shown in Figure 3 give an idea of the difficulty of the task. The left image can be retrieved by using a good CBIR. While the middle image can be retrieved by using a text-based system or a multi-modal one. The rightmost image is a much harder image to retrieve because neither the annotation nor the image provide clues that can link the image to the the topic stated in Figure 2. The difficulty of the task, therefore, requires taking into account both, visual and textual information, in order to achieve acceptable performance.

For the evaluation of an arbitrary retrieval system one should run it over the total number of topics (39 are used for ImageCLEF2008 and 60 were used in ImageCLEF2007) and evaluating the list of retrieved documents using the above described measures. We used the 2008 version because we want to compare our method against *cutting-edge* multimedia image retrieval systems. Further, one should note that, from the multimedia point of view, the collection we use in this work is more challenging than that used in 2007. For illustration we show a performance comparison of the three

Figure 2: Textual statement for topic 2: CHURCH WITH MORE THAN TWO TOWERS, ***Relevant images will show a church, cathedral or a mosque with three or more towers. Churches with only one or two towers are not relevant. Buildings that are not churches, cathedrals or mosques are not relevant even if they have more than two towers.***

Figure 3: Relevant images for the topic 2 (see Figure 2). The titles of the images are, from left to right. Left: ***The St. Patrick's Cathedral*****, middle:** ***The Church of the Savior on Blood*****, right:** ***View from the Sydney Sky Tower*****.**

IRMs that were evaluated at ImageCLEF2007 and that are also considered for HLFIRM with ImageCLEF2008 data. Results of this comparison are shown in Table 2. We can clearly appreciate that the MAP performance of the IRMs is always superior in the 2007 version, the same pattern is observed for precision and relevant-retrieved documents. This result gives evidence that the results we report on the 2008 version are a pessimistic estimate of the performance we would have on the 2007 collection. This means that our results are (in the worst case) comparable to those obtained in 2007. This is a highly useful result since ImageCLEF2008 results are not available yet. Also, it is interesting to note that the IRMs in Table 2 were not ranked at the top positions of the ImageCLEF2007 results [4], and yet resulted very useful for HLFIRM.

4.2 Individual performance of IRMs

In the first experiment we analyzed the individual performance of IRMs in order to measure the advantages of using the HLFIRM method instead of a good single IRM. In Figure 4 it is shown the retrieval performance for each of the IRMs described in Table 1. In average text-based

IRM	M-07	M-08	P-07	P-08	R-07	R-08	Rk
LF-07	0.1701	0.1598	0.2242	0.2321	58.55	57.79	41
IMFB-07	0.1986	0.1825	0.2917	0.3205	50.32	42.71	82
FIRE	0.1172	0.0939	0.2558	0.2282	36.56	32.42	300

Table 2: Comparison of IRMs (rows 1-3 in Table 1) for the 2007 and 2008 ImageCLEF collection. Second and third column show the MAP; columns 4 and 5 show precision at 20 documents; columns 6 and 7 show the percentage of relevant-retrieved documents (i. e. recall); the last column shows the position of each IRM in the general list of ranked participants at ImageCLEF2007.

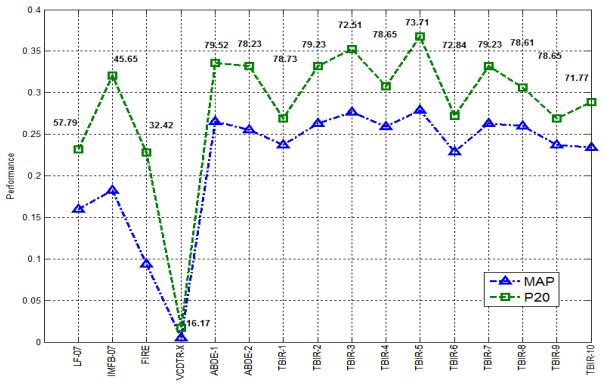

Figure 4: Individual performance of the IRMs described in Table 1: the MAP (solid blue line) and precision at 20 documents (dashed green line). The number above plotted value is the percentage of relevant documents retrieved by each IRMs at the first 1000 documents.

methods obtained a MAP of 0.25385, compared to 0.21575 and 0.04955 obtained by multi-modal and visual IRMs respectively. This result shows the superior (individual) performance of textual methods. Different combinations of local/global weighings for indexation produce slightly different retrieval results. The latter result motivates research on model selection for information retrieval. The IRM with lowest performance is VCDTR-X.

The better performance of textual methods is due to the quantity of text available in the collection. The average length of annotations is 21.5 words (in 2007 the average was of about 6.6 words per document). Looking at the number of relevant documents retrieved (bold numbers above plotted values), the ABDE-1 method retrieved the largest number of relevant documents, 1918 out of 2412 representing 79.6%, compared to 74.3% for the best MAP performer (i. e. TBIR-5). This result uncover the apparent advantage of using text-based methods only, and shows that better relevance ranking strategies are needed for the ABDE-1 approach. From Figure 4 we also can see the diversity, redundancy and complementariness of IRMs we consider; in terms of MAP performance, number of retrieved documents, modalities and heterogeneity of indexing/retrieval methods. All of these are desired properties when designing HLFIRM.

4.3 Per-modality performance of HLFIRM

In the next experiment we applied the HLFIRM approach to IRMs of the same modality. The result of this experiment will help us to determine the importance of each modality in terms of retrieval performance; it will also be useful for comparing HLFIRM when using as input homogeneous-like (text-based) and heterogeneous[2] (image-based and multi-modal) IRMs as input. For this experiment we run HLFIRM three times, each time using as input a lists of documents from IRMs of common modalities. Equal weights are as-

[2]Image-based and multi-modal IRMs are heterogeneous because they are based on retrieval strategies that are different in nature. Further, different information is considered by each of these IRMs. On the other hand, textual IRMs are more homogeneous because all of them use the same information, they only differ in the way that a term-document matrix is build.

ID	Modality	MAP	P20	R	DR
TBIR-5	Textual	0.2788	0.3679	73.71	73.71
TXT	Textual	0.2656	0.3295	80.55	81.1
IMG	Visual	0.0626	0.1731	30.01	39.4
TXT+IMG	Multi-Modal	0.2882	0.4026	82.50	83.8

Table 3: Performance of LFIRM by grouping IRMs per modality. For reference, are shown the results of the IRM with the best individual performance (first row). Column 5 (DR) shows the number of relevant documents in the union of the relevant documents retrieved by the IRMs per modality.

signed to each IRM (i. e. $\alpha_1 = \ldots \alpha_N$). The result are three lists (one per modality) of ranked documents, identified by TXT, IMG, TXT+IMG for the textual, visual and multi-modal modalities, respectively. Results of this experiment are shown in Table 3.

From Table 3 we can clearly appreciate that HLFIRM help significantly to improve the MAP performance of the multi-modal IRMs. The MAP performance of TXT+IMG is superior to the best IRM individually. This is an interesting result because most multi-modal IRMs performed poor individually in MAP. On the contrary, the performance of TXT (i. e. HLFIRM with text-based IRMs as input) is lower than the best IRM and therefore, HLFIRM is not helping here. This can be due to the lack of diversity through lists from text-based IRMs. Because, even when different retrieval performances are achieved by using different weighting strategies, the diversity and complementariness of retrieved documents is limited and therefore the HLFIRM method does not help them. This gives evidence that HLFIRM works well when one considers retrieval methods that are different in nature, or use different information from the same data. Visual methods do not perform well because, despite being different, the base IRMs perform poorly. Column 5 in Table 3 is an indicator of the, per-modality, diversity of documents. Note that diversity of documents in four multi-modal IRMs (TXT+IMG) is superior to that in ten textual IRMs (TXT), a result that clearly illustrates why HLFIRM works well for TXT+IMG.

Results from this section show that the use of heterogeneous IRMs instead of homogeneous ones for HLFIRM results in a better retrieval performance. However, another important result is that the performance of HLFIRM strongly depends on having a balance between the number of relevant documents retrieved (quantity) by IRMs and their individual performance (quality).

4.4 Global HLFIRM

In the rest of this section we analyze the performance of HLFIRM under different settings. The number of relevant documents in the union of all of lists from the IRMs is 2118 out of 2412, which represents the 87.8% of the total of relevant documents. This number is an upper bound on the maximum of documents that we can retrieve (**R**).

First, we evaluate the performance of the HLFIRM method globally, by considering the lists of each IRM as input for the HLFIRM (see Figure 1 and Equation (1)). Note that under this setting text-based methods have an implicit preference because they represent the 62.5% (10 out 16) of the total IRMs. Rows 2-4 in Table 4 show the results of the latter experiment. The following strategies are considered for weighting the contribution of each IRM (i. e. α_i values

Weighting	MAP	P20	R
Uniform	0.2967	0.3705	81.01
Performance	0.2664	0.3397	80.01
Modality	0.2657	0.3385	80.01
0.3/0.3/0.3	0.2884	0.3872	81.13
0.1/0.1/0.8	0.3	0.41923	82.21
0.1/0.8/0.1	0.2726	0.3859	81.26
0.8/0.1/0.1	0.3024	0.4192	82.21

Table 4: HLFIRM performance by merging the lists of IRMs from Table 1 (rows 2-4); weights assigned uniformly, according individual performance of IRMs and according the modality of each IRM. Also the performance by merging the lists evaluated in Table 3 are shown (rows 6-9). Each configuration $w_1/w_2/w_3$ indicates the weight assigned to each modality txt / img /txt+img.

in Equation (1)). In *uniform* weighting (row 2), an equal weight is assigned to every IRM. With *performance* (row 3) weights are assigned proportional to the individual performance of IRMs, see Figure 4. While with *modality* (row 4) weights are assigned proportional to the per-modality performance from table 3.

We can see that there is an improvement in MAP of 6% over the best single IRM by equally weighting to all IRMs; also there is a slight improvement of around 3.2% over the best per-modality result (TXT+IMG in Table 3). Using the other weighting strategies (rows 3-4) do not helped the HLFIRM method, the performance under such settings is lower than most of the IRMs independently. This is another interesting result, because it gives evidence that the MAP/precision/recall performance are not reliable estimators of the *goodness* of IRMs for HLFIRM (i. e. being individually-good or per-modality good does not implies fused-good). Therefore, alternative strategies for evaluating the goodness of IRMs for HLFIRM are needed.

4.5 Hierarchical HLFIRM

Rows 5-10 in Table 4 show the results of applying HLFIRM to the per-modality (already fused) lists described in Table 3. This experiment is a two-stage hierarchical application of HLFIRM. In the first stage, IRMs of common modalities are used with HLFIRM for obtaining three lists (TXT, IMG, TXT+IMG), equal weighing is used at this phase. At a second stage these three lists are used again with HLFIRM for obtaining a final list of documents. Four weighting combinations are used for obtaining the final list.

This time the best performance of HLFIRM is achieved by giving preference to either textual or multi-modal methods (rows 6 and 8), the difference is not significant between these two configurations. On the other hand, we can see that giving preference to visual methods is not a good weighting strategy. Therefore, it seems that by using textual and multi-modal methods is enough to achieve an acceptable retrieval performance. However, given that only 2 out of 16 IRMs (i. e. 12.5%) are visual-based and given the poor performance of these methods, it may be possible that we are not effectively taking advantage of the positive impact that visual-methods can provide to HLFIRM.

In order to verify the latter statement we conducted the following experiment. First, we applied HLFIRM taking as input the lists of IRMs that make use of images in any form (i. e. visual and multi-modal IRMs), the resultant list is (MM+IMG). The experiment consists of comparing

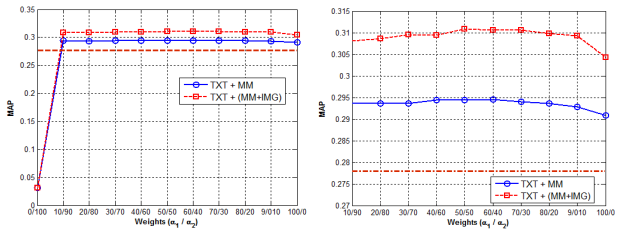

Figure 5: MAP of HLFIRM using as input the lists TXT + MM (blue-solid line) and TXT + (MM+IMG) (dashed-red line). The horizontal dashed-dotted line represents the MAP of the best individual IRM. Different weighting combinations are considered in the X-axis.

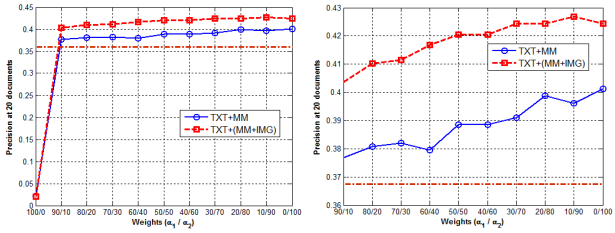

Figure 6: Precision at 20 documents of HLFIRM using as input the lists TXT + MM (blue-solid line) and TXT + (MM+IMG) (dashed-red line). The horizontal dashed-dotted line represents the Precision of the best individual IRM. Different weighting combinations are considered in the X-axis.

HLFIRM performance by using the TXT list combined with MM vs combined with (MM + IMG). Therefore, HLFIRM is applied twice: first using as input TXT + MM and later using TXT + (MM + IMG). Different weight values are considered for each list. Results of this experiment are shown in Figures 5 and 6.

In Figure 5 it is shown the MAP for each setting by using different weighting strategies. The left plot illustrates the results of the entire experiment, while the right one omits the first x-value (a very small value that difficult visualization) in order to analyze the results with more detail. Weight values are shown in the x-axis: α / $1-\alpha$ indicates that a weight of α is assigned to the TXT list and a weight of $1-\alpha$ is used with either the MM+IMG or IMG lists.

The first interesting result is that by using only the TXT list (100/0 setting) there is a significant decrease in MAP with HLFIRM. This is due to the fact that documents appearing in the two lists (TXT + MM or TXT + MM+IMG) receive a weight proportional to the double of the weight they have in the TXT list. From the right plot, we can see that better results are obtained with HLFIRM using the (MM+IMG) list instead of MM. Both settings outperform the best single-IRM, however the improvement of the best entry (50/50) using TXT + (MM + IMG) is of around 10.35% with respect to the best IRM performer; while the best result weights-combination of TXT + MM outperforms the best IRM by only 5.36%.

Note that the best weighting configuration (50/50) is, again, an interesting result. This is because in most previous work reported on the subject, the best performance of LFIRM methods is obtained by giving preference to the textual IRMs (i. e. the methods that individually use to perform better). In Figure 6 it is shown the precision at 20 documents for the same entries compared in Figure 5. We can see that the same pattern is observed, though this time the improvement in precision with respect to the best single IRM is of 13.82% and 5.30%, for TXT + (MM+IMG) and TXT + MM respectively. The best result in precision is obtained with TXT + (MM+IMG) and the weighting configuration 10/90, which almost discards the contribution of the TXT list. Again, this is an interesting result that differs from previous work on the subject.

Experimental results from this section show that even IRMs that perform poor both individually and per-modality can be useful for HLFIRM if they are properly combined (in our case the hierarchical HLFIRM combination performed very well). Therefore, no IRM nor no modality should be discarded because of their (apparent) poor individual performance. Ranked lists from low-performance IRMs can be useful for assigning a low-rank to irrelevant documents and improving retrieval performance.

ID	Run	MAP	P20	R
1	TXT+MMIMG	0.311	0.4205	81.88
2	Cut01	0.3175	0.4592	65.89
3	XRCE-1	0.3168	0.4167	75.55
4	XRCE-2	0.3020	0.3733	75.96
5	Cut-2	0.2846	0.5283	64.12
6	IPAL	0.2833	0.4867	70.01

Table 5: Comparison of the entry with highest MAP that we obtained vs the top-5 ranked entries in ImageCLEF2007. The configuration with the highest MAP we obtained was the combination TXT + (MM + IMG) with equal weights (50/50).

4.6 ImageCLEF2007 Comparison

In order to compare our results to other successful multimedia image retrieval methods, in Table 5 it is shown a comparison of our best result in MAP against the top-5 entries in ImageCLEF2007.

As we can see, the results of HLFIRM are comparable to best ImageCLEF2007 entry, even when such entry required the intervention of an user for providing relevance feedback. We should emphasize, however, that the 2008 collection is more difficult that that used in 2007, see Section 4.1; and therefore, our results can improve if we run TXT+MMIMG over the ImageCLEF2007 topics (this is work in progress). The performance of HLFIRM is superior to entries with ID 4-6. An important result is that using HLFIRM taking as input IRMs that were ranked low at ImageCLEF2007 (see Table 2) we can obtain results comparable to those achieved by the best methods. This confirms again that by using middle-low performance IRMs we can obtain a superior combined performance with HLFIRM. Also, IRMs with high recall and of diverse nature can be more helpful than IRMs with high individual recall.

The experimental results presented along this section give empirical evidence of the validity of the HLFIRM approach and motivate further research on diverse areas.

5. CONCLUSIONS

We have presented experimental results on the late fusion of heterogeneous retrieval methods for multimedia im-

age retrieval. LFIRM is a widely used approach to combine information from multiple retrieval systems. For multimedia image retrieval, this approach has been already successfully used. However, most research on LFIRM has focussed on a single-homogeneous retrieval method for each modality involved. Further, the performance obtained by the combined LFIRM method is only slightly superior to that achieved by a single retrieval method.

We conducted experiments with heterogeneous retrieval methods combined with the LFIRM approach. To the best of our knowledge this issue has not been explored before for multimedia image retrieval. Experimental results on benchmark data show the validity of the approach and its comparable performance to state-of-the-art methods. An improvement of around 13% is reported by using the HLFIRM method instead of the best single IRM performer. Further, retrieval methods of the modality with highest individual and per-modality performance do not need to be weighted high in order to achieve acceptable performance.

The following interesting findings are result of our experimentation and therefore our contributions. *i*) The use of heterogeneous IRMs in HLFIRM outperforms fusion of homogeneous retrievers. *ii*) The simple fusion score we considered resulted very useful for list merging. *iii*) IRMs with high (individual or per-modality) performance, in terms of MAP and precision at 20 documents, are not so useful as IRMs that offer diversity and high recall. *iv*) The quantity and quality of lists from IRMs is crucial for HLFIRM, IRMs with a balance between these two properties are more useful to HLFIRM instead of IRMs that either (but not both) have high quality or there are many of them available. *v*) Opposed to previous work, by giving preference to the modality/IRMs that are majority or perform better individually/per-modality does not results in better performance. *vi*) The hierarchical application of HLFIRM significantly outperforms a global or per-modality application of HLFIRM. *vii*) Results with hierarchical HLFIRM are comparable and presumably better than that of methods proposed for ImageCLEF2007.

The variety of individual retrieval methods we considered resulted sufficient and very useful for reaching state-of-the-art performance. However, it would be necessary to state necessary and sufficient conditions for obtaining good performance with HLFIRM. The latter is the main future work direction we are following. Although we think that the present paper will motivate further research in several other directions. Including, *i*) the development of IRMs that maximize diversity and recall instead of MAP or precision; *ii*) studying different ways to reliably estimate the goodness of IRMs for HLFIRM; *iii*) developing measures for determining the diversity, complementariness and redundancy for sets of IRMs; *iv*) studying and experimenting with more sophisticated fusion strategies; *vi*) developing new strategies for the application of HLFIRM just like iterative HLFIRM.

6. REFERENCES

[1] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Pearson E. L., 1999.

[2] R. Besancon and C. Millet. Merging results from different media: Lic2m experiments at imageclef 2005. In *Working notes of the CLEF 2005*. CLEF.

[3] 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*. CLEF, 2006.

[4] P. Clough, M. Grubinger, T. Deselaers, A. Hanbury, and H. Müller. Overview of the imageclef 2007 photographic retrieval task. In *CLEF 2007*, volume 5152 of *LNCS*. CLEF, Springer-Verlag, 2008.

[5] H. J. Escalante and et al. Towards annotation-based query and document expansion for image retrieval. In *CLEF 2007*, volume 5152 of *LNCS*, pages 546–553. Springer-Verlag, 2008.

[6] T. Gass, T. Weyand, T. Deselaers, and H. Ney. Fire in imageclef 2007: Support vector machines and logistic regression to fuse image descriptors in for photo retrieval. volume 5152 of *LNCS*. Springer-Verlag, 2008.

[7] A. Goodrum. Image information retrieval: An overview of current research. *Journal of Informing Science*, 3(2), 2000.

[8] M. Grubinger, P. Clough, H. Müller, and T. Deselaers. The iapr tc-12 benchmark: A new evaluation resource for visual information systems. In *Proc. of the Intl. Workshop OntoImage'2006 Language Resources for CBIR*, Genoa, Italy, 2006.

[9] C. Hernández and L. E. Sucar. Markov random fields and spatial information to improve automatic image annotaion. In *Proc. of the the 2007 Pacific-Rim Symposium on Image and Video Technology*, volume 4872 of *LNCS*, pages 879–892. Springer, 2007.

[10] R. Izquierdo-Beviá, D. Tomás, M. Saiz-Noeda, and J. L. Vicedo. University of alicante in imageclef2005. In *Working Notes of the CLEF*. CLEF, 2005.

[11] M. M. Rautiainen and T. Seppdnen. Comparison of visual features and fusion techniques in automatic detection of concepts from news video. In *Proceedings of the IEEE ICME*, pages 932–935, 2005.

[12] P. Over and A. F. Smeaton., editors. *Proc. of the international workshop on TRECVID video summarization.*, Augsburg, Bavaria, Germany., 2007.

[13] V. Peinado, F. López-Ostenero, and J. Gonzalo. Uned at imageclef 2005: Automatically structured queries with named entities over metadata. In *Working Notes of the CLEF*. CLEF, 2005.

[14] J. L. R. Datta, D. Joshi and J. Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*, to appear, 2008.

[15] M. Rautiainen, T. Ojala, and S. Tapio. Analysing the performance of visual, concept and text features in content-based video retrieval. In *MIR '04: Proc. of the 6th ACM workshop on Multimedia information retrieval*, pages 197–204, New York, NY, USA, 2004. ACM Press.

[16] S. Sclaroff, M. L. Cascia, and S. Sethi. Unifying textual and visual cues for content-based image retrieval on the world wide web. *Computer Vision*, 75(1/2):86–98, July/August 1999.

[17] C. Snoek, M. Worring, and A. Smeulders. Early versus late fusion in semantic video analysis. In *Proc. of the 13th Annual ACM Conference on Multimedia*, pages 399–402, Singapore, 2005. ACM.

[18] D. Zeimpekis and E. Gallopoulos. Tmg: A matlab toolbox for generating term-document matrices from text collections. In *Recent Advances in Clustering*, pages 187–210. Springer, 2005.