MIRACLE's Combination of Visual and Textual Queries for Medical Images Retrieval

Julio Villena-Román[1,3], José Carlos González-Cristóbal[2, 3]
José Miguel Goñi-Menoyo[2], José Luís Martínez-Fernandez[1, 3]
Juan José Fernández[2]

[1] Universidad Carlos III de Madrid
[2] Universidad Politécnica de Madrid
[3] DAEDALUS - Data, Decisions and Language, S.A.

jvillena@daedalus.es, jgonzalez@dit.upm.es,
josemiguel.goni@upm.es, jmartinez@daedalus.es
jjfernandez@isys.dia.fi.upm.es

Abstract

This paper presents the 2005 MIRACLE's team participation in the ImageCLEFmed task of ImageCLEF 2005. This task certainly requires the use of image retrieval techniques and therefore it is mainly aimed at image analysis research groups. Although our areas of expertise don't include image analysis research, we decided to make the effort to participate in this task to promote and encourage multidisciplinary participation in all aspects of information retrieval, no matter if it is text or content based. We resort to a publicly available image retrieval system (GIFT [1]) when needed.

Categories and Subject Descriptors

H.3 [Information Storage and Retrieval]: H.3.1 Content Analysis and Indexing; H.3.2 Information Storage; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software. E.1 [Data Structures]. E.2 [Data Storage Representations]. H.2 [Database Management].

Keywords

Linguistic Engineering, Information Retrieval, image retrieval, cross-lingual, content-based retrieval, visual, text-based retrieval, textual, relevance feedback, GIFT, KSite, Lucene, IRMA project.

1 Introduction

ImageCLEF is the cross-language image retrieval track which was established in 2003 as part of the Cross Language Evaluation Forum (CLEF), a benchmarking event for multilingual information retrieval held annually since 2000. Images are language independent by nature, but often they are accompanied by texts semantically related to the image (e.g. textual captions or metadata). Images can then be retrieved using primitive features based on its contents (e.g. visual exemplar) or abstract features expressed through text or a combination of both.

Originally, ImageCLEF focused specifically on evaluating the retrieval of images described by text captions using queries written in a different language, therefore having to deal with monolingual and bilingual image retrieval (multilingual retrieval was not possible as the document collection is only in one language) [13],[14]. Later, the scope of ImageCLEF widened and goals evolved to investigate the effectiveness of combining text and image for retrieval (text and content-based) [8], collect and provide resources for benchmarking image retrieval systems and promote the exchange of ideas which will lead to improvements in the performance of retrieval systems in general.

With this objective in mind, a medical retrieval task was included in 2004 campaign [8] and continued this year. In this task (referred as ImageCLEFmed), example images are used to perform a search against a medical image database consisting of images such as scans and x-rays [4] to find similar images. Each medical image or a group of images represents an illness, and case notes in English or French are associated with each illness to be used for diagnosis or to perform a text-based query. Some of the queries are rather based on visual characteristics and responses with a content-based retrieval system may deliver satisfying results. Other queries cannot be solved with visual characteristics alone; thus they may seem very hard for visual-only retrieval researchers.

The MIRACLE team is made up of three university research groups located in Madrid (UPM, UC3M and UAM) along with DAEDALUS, a company founded in 1998 as a spin-off of two of these groups. DAEDALUS is a leading company in linguistic technologies in Spain and is the coordinator of the MIRACLE team. This is the third participation in CLEF, after years 2003 and 2004 [14],[10],[9],[3],[2]. As well as bilingual, monolingual and cross lingual tasks, the team has participated in the ImageCLEF, Q&A, WebCLEF and GeoCLEF tracks.

This paper describes our participation in the ImageCLEFmed task of ImageCLEF 2005. This task certainly requires the use of image retrieval techniques and therefore it is mainly aimed at image analysis research groups. Although our areas of expertise don't include image analysis research, we decided to make the effort to participate in this task to promote and encourage multidisciplinary participation in all aspects of information retrieval, no matter if it is text or content based. We resort to a publicly available image retrieval system (GIFT [1]) when needed.

2 Task goals

Image or multimedia retrieval is interesting for the domain of cross-language information retrieval as the media such as images are inherently almost insensitive to language. Many collections exist on the Internet which contain images as well as multilingual texts. However, the retrieval of images is an often-neglected topic in the information retrieval domain. In particular, hospitals produce an enormous amount of visual data but tools to manage these images and videos are scarce and exist currently only as research prototypes [4].

The main goal of ImageCLEFmed task is to improve the retrieval of medical images from heterogeneous and multilingual document collections containing images as well as text. The task is somewhat similar to the classic TREC ad hoc retrieval task, with a scenery in which a system knows the set of documents to be searched, but cannot anticipate the particular topic that will be investigated (i.e., topics are not known to the system in advance).

ImageCLEFmed 2005 extends the 2004 experiments with a larger database and more complex queries. The database consists of images [5] from the Casimage (Radiology and pathology), MIR (Mallinckrodt Institute of Radiology, nuclear medicine), PEIR (Pathology Education Instructional Resource, Pathology and radiology) and PathoPIC (Pathology) datasets, with about 50,000 images in all. The collection also contains about 50,000 annotations in XML format. While the majority are written in English (over 40,000), there is a significant number in French (over 1,800) and German (over 7,800), and a few cases with no annotation at all. The quality of the texts is variable between collections and even within the same collection.

Query tasks have been formulated with example images and a short textual description explaining the research goal. The task organizers provide a list of topic statements in English, French and German, and a collection of images for each topic. Normally one or two example images for the desired result are supplied. One query also contains a negative example as a test. The goal of ImageCLEFmed is to retrieve as many relevant images as possible from the given visual and multilingual topics.

The task organizers have also made available results from a state-of-the-art image retrieval system (medGIFT) and a state-of-the-art text engine (Lucene [7]).

The next section is devoted to the description of the different experiments which were carried out.

3 Description of experiments

We focused our experiments to fully automatic retrieval, avoiding any manual feedback, and submitted runs both using only visual features for retrieval (content-based retrieval) and also runs using visual features and text (combination of content-based and text-based retrieval).

To isolate from the content-based retrieval part of the process, we resorted to GIFT (GNU Image Finding Tool) [1], a publicly available content-based image retrieval system which was developed under the GNU license and allows to perform query by example on images, using an image as the starting point for the search process. GIFT relies entirely on the image contents and thus it doesn't require the collection to be annotated. It also provides a mechanism to improve query results by relevance feedback.

Our approach is based on the multidisciplinary combination of GIFT content-based searches with text-based retrieval techniques. Our system consists of three parts: the content-based retrieval component (mainly GIFT), the text-based search engine and the merging component, which combines the results from the others to provide the final results.

We finally submitted 13 different runs to be evaluated by the task coordinators, which are explained in the following section.

Content-Based Retrieval

Without feedback

This experiment consists on a content-based-only retrieval using GIFT. Initially the complete image database was indexed in a single collection using GIFT, down-scaling each image to 32x32 pixels. For each ImageCLEFmed query, a visual query is made up of all the images contained in the ImageCLEFmed query. Then, this visual query is introduced into the system to obtain the list of more relevant images (i.e., images which are more similar to those included in the visual query), along with the corresponding relevance values.. Although different search algorithms can be integrated as plug-ins in GIFT, only the provided separate normalisation algorithm has been used in our experiments.

There is only one submission, with "mirabase.qtop" as its run identifier.

With feedback

These experiments are similar to the preceding one, also a content-based-only retrieval using GIFT, but incorporating relevance feedback. Each visual query is introduced into the system to obtain the list of images which are more similar to the visual query. Then the top N results are added to the original visual query to build a new visual query which is again introduced into the system to obtain the final list of results. In addition, GIFT allows to build a weighted visual query in which a relevance value may be associated to each included image.

There are 3 submissions:

- mirarf5.qtop
  This run takes the 5 most relevant images for feedback, each one with a value of 1 for its relevance in the visual query. The relevance in the visual query for the original images remains 1.
- mirarf5.1.qtop
  The same as mirarf5.qtop but using a value of 0.5 for the relevance in query of feedback images. The relevance in query for the original images remains 1.
- mirarf5.2.qtop
  The same as mirarf5.qtop but using a value of 0.5 for the relevance in query of the original images.

Finally, the different content-based runs are shown in Table 1.

Table 1: Summary of content-based runs

Run id	Feedback	Number of feedback images	Relevance of original images	Relevance of feedback images
mirabase.qtop	NO	Not applicable	Not applicable	Not applicable
mirarf5.qtop	YES	5	1	1
mirarf5.1.qtop	YES	5	1	0.5
mirarf5.2.qtop	YES	5	0.5	1

Content-Based and Text-Based Mixed Retrieval

Our main interest is not in experiments where only image content is used in the retrieval process. Instead, our challenge was to test whether the text-based image retrieval could improve the analysis of the content of the image, or vice versa. We like to determine how text and image attributes can be combined to enhance image retrieval, in this case, in the medical domain.

First all the case annotations are indexed using a text-based retrieval engine (explained later). Natural language processing techniques are applied before indexing. An adhoc language-specific (for English, German and French) parser is used to identify different classes of alphanumerical tokens such as dates, proper nouns, acronyms, etc., as well as recognising common compound words. Text is tokenized, stemmed [11][12] and stop word filtered (for the three languages). Only one index is created, combining keywords in the three different languages.

Two different text-based retrieval engines were used. One was Lucene [7], with the results provided by the task organizers. The other engine was KSite [6], fully developed by DAEDALUS, which offers the possibility to use a probabilistic (BM25) model or a vector space model for the indexing strategy. Only the probabilistic model was used in our experiments.

The combination strategy consists on reordering the results from the content-based retrieval using a text-based retrieval. For each ImageCLEFmed query, a multilingual textual query is build with the English, German and French queries (first processing each one with the language depending parser and then concatenating the three lists), and executed in the search engine to obtain the list of top-1000 cases which are more relevant to the textual query.

The list of relevant images from the content-based retrieval is reordered, moving to the beginning of the list those images which belong to a case that is in the list of top-1000 cases. The rest of the images remain in the end of the list.

There are 10 different submissions:

- mirabasefil.qtop, mirarf5fil.qtop, mirarf5.1fil.qtop, mirarf5.2fil.qtop
  These runs consist on the combination as previously described of content-based-only runs with the text-based retrieval obtained with KSite.
- mirabasefil2.qtop, mirarf5fil2.qtop, mirarf5.1fil2.qtop, mirarf5.2fil2.qtop
  The same experiment, but using Lucene.
- Other runs
  Two other experiments were developed to test if there was any difference in results when using our own content-based GIFT index or using the medGIFT results provided by the task organizers. So, medGIFT was used as the starting point and then the same combination method as described before was applied.
    - mirabase2fil.qtop
      medGIFT results filtered with text-based KSite results
    - mirabase2fil2.qtop
      medGIFT results filtered with Lucene results

Finally, the different mixed retrieval runs are shown in Table 2.

Table 2: Summary of mixed retrieval runs

Run id	Feedback	Number of feedback images	Relevance of original images	Relevance of feedback images	Text retrieval engine
mirabasefil.qtop	NO	Not applicable	Not applicable	Not applicable	KSite
mirarf5fil.qtop	YES	5	1	1	KSite
mirarf5.1fil.qtop	YES	5	1	0.5	KSite
mirarf5.2fil.qtop	YES	5	0.5	1	KSite
mirarf5fil2.qtop	YES	5	1	1	Lucene
mirarf5.1fil2.qtop	YES	5	1	0.5	Lucene
mirarf5.2fil2.qtop	YES	5	0.5	1	Lucene
mirabase2fil.qtop	?	?	?	?	KSite
mirabase2fil2.qtop	?	?	?	?	Lucene

4 Evaluation

Relevance assessments have been performed by experienced medical students and medical doctors at OHSU and the University hospitals of Geneva. Submissions from all groups are used to create image pools, which are judged for relevance by assessors, based on using a ternary classification scheme: (1) relevant, (2) partially relevant and (3) not relevant. The aim of the ternary scheme is to help assessors in making their relevance judgments more accurate (e.g., an image is definitely relevant in some way, but maybe the query object is not directly in the foreground; it is therefore considered partially relevant).

The pools are assessed and the end result is a set of relevance assessments called qrels, which are then used to evaluate system performance and compare submissions from different groups.

Content-Based runs

The results of the content-based runs are shown in the next table, ordered by its mean average precision value.

Table 3: Evaluation of content-based runs

Run	Average Precision	%
mirabase.qtop	0.0942	
mirarf5.1.qtop	0.0942	100.0%
mirarf5.qtop	0.0941	99.8%
mirarf5.2.qtop	0.0934	99.1%

As shown in Table 3, the best result was obtained with the base experiment, which means that relevance feedback has failed to improve the results (neither to worsen them). This may be due to an incorrect choice of the parameters, but this has to be further studied.

Apart from MIRACLE, other 8 groups participated in this year's evaluation in the content-based-only runs. Table 4 compares each group's best submission.

Table 4: Comparison of content-based runs from different groups

Group	Average Precision	%
I2R	0.1455	
MIRACLE	0.0942	64.7%
GE	0.0941	64.7%
rwth	0.0751	51.6%
i6	0.0713	49.0%
nctu	0.0672	46.2%
<unknown>	0.0637	43.8%
ceamdI	0.0465	32.0%
cindi	0.0072	4.9%

Only one group is above us in the group ranking, although their average precision is much better than ours. Our pragmatic approach using a "standard" publicly available content-based retrieval engine such as GIFT has proved to be a better approach than other presumably more complex techniques. We still have to test if another selection of indexing parameters (different from image down-scaling to 32x32 pixels and separate normalization algorithm) may provide better results.

Content-Based and Text-Based Mixed Retrieval

The results of the content-based and text-based mixed retrieval runs are shown in Table 5, ordered by its mean average precision value. In this case, using relevance feedback provides slightly better precision values. Considering the best runs, the optimum choice seems to be to assume 1.0 for the relevance of the top 5 results and reduce the relevance of the images in the original query.

Table 5: Comparison of mixed retrieval runs

Run	Average Precision	%
mirarf5.2fil.qtop	0.1173	
mirarf5fil.qtop	0.1171	99.8%
mirabasefil.qtop	0.1164	99.2%
mirabase2fil.qtop	0.1162	99.0%
mirarf5.1fil.qtop	0.1159	98.8%
mirarf5fil2.qtop	0.1028	87.6%
mirarf5.2fil2.qtop	0.1027	87.6%
mirarf5.1fil2.qtop	0.1019	86.9%
mirabasefil2.qtop	0.0998	85.1%
mirabase2fil2.qtop	0.0998	85.1%

Table 5 also shows that the results are better with our own text-based search engine than using Lucene (all runs offer better precision values), at least with the adopted combination strategy. This difference could be attributed to a better language dependent pre-processing and removal of stop words.

It is interesting to observe that the worst combination is to take both results provided by the task organizers (content-based medGIFT results and text-based Lucene results), with a performance decrease of 15%.

Comparing content-based runs with the mixed runs, Table 6 shows that the combination of both types of retrieval offers better performance and even the worst mixed run is better than the best content-based only run. This actually proves that text-based image retrieval can be used to improve the content-based only retrieval, with much superior performance.

Table 6: Comparison of content-based and mixed retrieval strategies

Run	Average Precision	%
mirarf5.2fil.qtop	0.1173	
mirabase2fil2.qtop	0.0998	85.1%
mirabase.qtop	0.0942	80%

Apart from MIRACLE, other 6 groups participated in this year's evaluation in the content-based and text-based runs. Table 7 shows the results for each group's best submission.

Table 7: Comparison of mixed retrieval runs from other groups

Group	Average Precision	%
I2R	0.2821	
Nctu	0.2389	84.7%
Ubimed	0.2358	83.6%
MIRACLE	0.1173	41.6%
GE	0.0981	34.8%
i6	0.0667	23.6%
ceamdl	0.0538	19.1%

In this case, our position in the table shows that the submissions from other groups clearly surpassed our results. Anyway, these results are very satisfying for us, considering that we are not a group with expertise in image analysis research.

It is also interesting to note that most groups managed to improve their results with mixed approaches over the content-based only runs. This is especially visible for the NCTU group, with an improvement from 0.06 to 0.23 (+355%) in the average precision.

5 Conclusions and Future Work

Our main interest is not in experiments where only image content is used in the retrieval process. Instead, our challenge was to test whether the text-based image retrieval could improve the analysis of the content of the image, or vice versa. Results show that this hypothesis was right. Our combination of a "black-box" search using a publicly accessible content-based retrieval engine with a text-based search has turned to provide comparable results to other presumably "more complex" techniques. This simplicity may be a good starting point for the implementation of a real system.

We think that there still may be some space for improvement with a more careful study of the parameters for the relevance feedback and the combination strategy.

Acknowledgements

This work has been partially supported by the Spanish R+D National Plan, by means of the project RIMMEL (Multilingual and Multimedia Information Retrieval, and its Evaluation), TIN2004-07588-C03-01 and also by the European Union with the funding of NEDINE project in the e-Content programme.

Special mention to our colleagues of the MIRACLE team should be done (in alphabetical order): Ana María García-Serrano, Ana González-Ledesma, José José Mª Guirao-Miras, Sara Lana-Serrano, Paloma Martínez-Fernández, Ángel Martínez-González, Antonio Moreno-Sandoval and César de Pablo Sánchez.

References

[1] GIFT: The GNU Image-Finding Tool. On line http://www.gnu.org/software/gift/ [Visited 18/07/2005]

[2] Goñi-Menoyo, José M; González, José C.; Martínez-Fernández, José L.; and Villena, J. MIRACLE's Hybrid Approach to Bilingual and Monolingual Information Retrieval. CLEF 2004 proceedings (Peters, C. et al., Eds.). Lecture Notes in Computer Science, vol. 3491, pp. 188-199. Springer, 2005 (to appear).

[3] Goñi-Menoyo, José M.; González, José C.; Martínez-Fernández, José L.; Villena-Román, Julio; García-Serrano, Ana; Martínez-Fernández, Paloma; de Pablo-Sánchez, César; and Alonso-Sánchez, Javier. MIRACLE's hybrid approach to bilingual and monolingual Information Retrieval. Working Notes for the CLEF 2004 Workshop (Carol Peters and Francesca Borri, Eds.), pp. 141-150. Bath, United Kingdom, 2004.

[4] Goodrum, A.A.: Image Information Retrieval: An Overview of Current Research. Informing Science, Vol 3(2):63-66 (2000).

[5] IRMA project: Image Retrieval in Medical Applications. On line http://www.irma-project.org/ [Visited 18/07/2005]

[6] KSite [Agente Corporativo]. On line http://www.daedalus.es/ProdKSiteAC-E.php. [Visited 13/07/2005]

[7] Lucene. On line http://lucene.apache.org/. [Visited 13/07/2005]

[8] Martínez-Fernández, José L.; García-Serrano, Ana; Villena, J. and Méndez-Sáez, V.; MIRACLE approach to ImageCLEF 2004: merging textual and content-based Image Retrieval. CLEF 2004 proceedings (Peters, C. et al., Eds.). Lecture Notes in Computer Science, vol. 3491. Springer, 2005 (to appear).

[9] Martínez, José L.; Villena, Julio; Fombella, Jorge; G. Serrano, Ana; Martínez, Paloma; Goñi, José M.; and González, José C. MIRACLE Approaches to Multilingual Information Retrieval: A Baseline for Future Research. Comparative Evaluation of Multilingual Information Access Systems (Peters, C; Gonzalo, J.; Brascher, M.; and Kluck, M., Eds.). Lecture Notes in Computer Science, vol. 3237, pp. 210-219. Springer, 2004.

[10] Martínez, J.L.; Villena-Román, J.; Fombella, J.; García-Serrano, A.; Ruiz, A.; Martínez, P.; Goñi, J.M.; and González, J.C. (Carol Peters, Ed.): Evaluation of MIRACLE approach results for CLEF 2003. Working Notes for the CLEF 2003 Workshop, 21-22 August, Trondheim, Norway.

[11] Porter, Martin. Snowball stemmers and resources page. On line http://www.snowball.tartarus.org. [Visited 13/07/2005]

[12] University of Neuchatel. Page of resources for CLEF (Stopwords, transliteration, stemmers …). On line http://www.unine.ch/info/clef/. [Visited 13/07/2005]

[13] Villena, Julio; Martínez, José L.; Fombella, Jorge; G. Serrano, Ana; Ruiz, Alberto; Martínez, Paloma; Goñi, José M.; and González, José C. Image Retrieval: The MIRACLE Approach. Comparative Evaluation of Multilingual Information Access Systems (Peters, C; Gonzalo, J.; Brascher, M.; and Kluck, M., Eds.). Lecture Notes in Computer Science, vol. 3237, pp. 621-630. Springer, 2004.

[14] Villena-Román, J.; Martínez, J.L.; Fombella, J.; García-Serrano, A.; Ruiz, A.; Martínez, P.; Goñi, J.M.; and González, J.C. (Carol Peters, Ed.); MIRACLE results for ImageCLEF 2003. Working Notes for the CLEF 2003 Workshop, 21-22 August, Trondheim, Norway.