The ImageCLEF Medical Retrieval Task at ICPR 2010 Information Fusion

Henning Müller
University of Applied Sciences Western Switzerland (HES–SO)
henning.mueller@sim.hcuge.ch

Jayashree Kalpathy–Cramer
Oregon Health and Science University, Portland, OR, USA
kalpathy@ohsu.edu

Abstract

An increasing number of clinicians, researchers, educators and patients routinely search for medical information on the Internet as well as in image archives. However, image retrieval is far less understood and developed than text–based search. The ImageCLEF medical image retrieval task is an international benchmark that enables researchers to assess and compare techniques for medical image retrieval using standard test collections. Although text retrieval is mature and well researched, it is limited by the quality and availability of the annotations associated with the images. Advances in computer vision have led to methods for using the image itself as search entity. However, the success of purely content–based techniques has been limited and these systems have not had much clinical success. On the other hand a combination of text– and content–based retrieval can achieve improved retrieval performance if combined effectively. Combining visual and textual runs is not trivial based on experience in ImageCLEF. The goal of the fusion challenge at ICPR is to encourage participants to combine visual and textual results to improve search performance. Participants were provided textual and visual runs, as well as the results of the manual judgments from ImageCLEFmed 2008 as training data. The goal was to combine textual and visual runs from 2009. In this paper, we present the results from this ICPR contest.

1. Introduction

Image retrieval is a burgeoning area of research in medical informatics [6, 9, 2]. With the increasing use of digital imaging in all aspects of health care and medical research, there has been a substantial growth in the number of images being created every day in healthcare settings. An increasing number of clinicians, researchers, educators and patients routinely search for relevant medical information on the Internet as well as in image archives and PACS (Picture Archival and Communication Systems) [6, 2, 7]. Consequently, there is a critical need to manage the storage and retrieval of these image collections. However, image retrieval is far less understood and developed than text–based searching. Text retrieval has a long history of evaluation campaigns in which different groups use a common test collection to compare the performance of their methods. The best known such campaign is the Text REtrevial Conference (TREC[1], [1]), which has been running continuously since 1992. There have been several offshoots from TREC, including the Cross–Language Evaluation Forum (CLEF[2]). CLEF operates on an annual cycle, and has produced numerous test collections since its inception in 2000 [8]. While CLEFs focus was originally on cross–language text retrieval it has grown to include multimedia retrieval tracks of several varieties. The largest of these, ImageCLEF[3] , started in 2003 as a response to the need for standardized image collections and a forum for evaluation. It has grown to become todays pre–eminent venue for image retrieval evaluation.

2. The Annual ImageCLEF Challenge

ImageCLEF is an international benchmark that includes several sub–tracks concerned with various aspects of image retrieval [3]; one of these tracks is the medical retrieval task run since 2004. This task within ImageCLEF enables researchers to assess and compare techniques for medical image retrieval using

[1] http://trec.nist.gov/
[2] http://www.clef-campaign.org/
[3] http://www.imageclef.org/

standard collections. ImageCLEFmed uses the same methodology as information retrieval challenges including TREC. Participants are given a set of topics that represent information needs. They submit an ordered list of runs that contain images that their system believe best meet the information need. Manual judgments using domain experts, typically clinicians, are used to create ground truth. The medical image retrieval tracks test collection began with a teaching database of 8,000 images. Since then, it has grown to a collection of over 74,000 images from the scientific literature, as well as a set of topics that are known to be well–suited for textual, visual or mixed retrieval methods. A major goal of ImageCLEF has been to foster development and growth of multimodal retrieval techniques: i.e., retrieval techniques that combine visual, textual, and other methods to improve retrieval performance.

Traditionally, image retrieval systems have been text–based, relying on the textual annotations or captions associated with images. Several commercial systems, such as Google Images[4] and Yahoo! images[5], employ this approach. Although text–based information retrieval methods are mature and well researched, they are limited by the quality of the annotations applied to the images. Advances in techniques in computer vision have led to a second family of methods for image retrieval: content–based image retrieval (CBIR). In a CBIR system, the visual contents of the image itself are represented by visual features (colors, textures, shape) and compared to similar abstractions of all images in the database. Typically, such systems present the user with an ordered list of images that are visually most similar to the sample (or query) image. The text–based systems typically perform significantly better than purely visual systems at ImageCLEF.

Multimodal systems combine the textual information associated with the image with the actual image features in an effort to improve performance, especially early precision. However, our experience from the ImageCLEF challenge, especially of the last few years has been that these combinations of textual and visual systems can be quite fragile, with the mixed runs often performing worse than the corresponding textual run. We believe that advances in machine learning can be used more effectively to learn how best to incorporate the multimodal information to provide the user with search results that best meet their needs [4]. Thus, the goal of the fusion challenge at ICPR is to encourage participants to effectively combine visual and textual results to improve search performance. Participants were provided textual and visual runs, as well as the results of the manual judgments from the ImageCLEFmed 2008 challenge as training data. The goal was to combine similar textual and visual runs from 2009 challenge for testing. In this paper, we present the preliminary results from this ICPR competition

3. The ImageCLEF Fusion Challenge

In both 2008 and 2009, the Radiological Society of North America (RSNA) made a subset of its journals image collections available for use by participants in ImageCLEF. The 2009 database contains 74,902 images, the largest collection yet [5]. The organizers created a set of 25 search topics based on a user study conducted at Oregon Health & Science University (OHSU) in 2009 [7]. These topics consisted of 10 visual, 10 mixed and 5 semantically oriented topics, as categorized by the organizers based on past experience and nature of the query. During 2008 and 2009, a panel of clinicians, using a web–based interface, created relevance judgments. The manually judged results were used to evaluate the submitted runs using the trec_eval software package. This package provides commonly used information retrieval measures including mean average precision (MAP), recall as well as precision at various levels for all topics.

For the ICPR fusion contest, the goal was to combine the best visual and textual runs that had been submitted previously to improve performance over the purely visual and purely textual runs. After participants registered they were provided access to the training data in early November 2009. The training set consisted of the four best textual and visual runs from different groups in 2008. These runs were anonymized to remove information about the group. We also provided the qrel, the file that contained the output for the manual judgments as well as the results obtained by the training runs using the trec_eval package. Participants could create fusion runs using combinations of the provided training runs and evaluate the performance using the trec_eval along with the abovementioned qrel file as well as the results of the evaluation measures for the runs. We released the test runs two weeks later. Again these consisted of the four best textual and four best visual runs, this time from 2009. The ground truth in the form of qrel was not provided at this time. The judgments were released in early January so that the participants could evaluate their runs in time for submission to ICPR 2010. To summarize, the timeline for this contest was as follows:

- 16.11.2009 Release of training data
- 30.11.2009 Release of test data

[4] http://images.google.com/

[5] http://images.yahoo.com/

- 04.01.2010 Submission of results
- 10.01.2010 Release of ground truth data
- 15.01.2010 Conference paper submission

4. Results

Table 1 contains the performance of the training runs that were provided. As can be seen, the textual runs perform significantly better than the visual runs for all measures.

Table 1. Results of the training runs.

Run	Recall	MAP	P5	P10
Text1	0.63	0.29	0.49	0.46
Text2	0.65	0.28	0.51	0.47
Text3	0.54	0.27	0.51	0.47
Text4	0.61	0.28	0.44	0.41
Visual1	0.06	0.028	0.15	0.13
Visual2	0.24	0.035	0.17	0.17
Visual3	0.17	0.042	0.22	0.17

This performance gap was similarly true for the test runs (Table 2).

Table 2. Results of the test runs.

Run	Recall	MAP	P5	P10
Text1	0.73	0.35	0.58	0.56
Text2	0.66	0.35	0.65	0.62
Text3	0.77	0.43	0.70	0.66
Text4	0.80	0.38	0.65	0.62
Visual1	0.12	0.01	0.09	0.08
Visual2	0.12	0.01	0.08	0.07
Visual3	0.11	0.01	0.09	0.07
Visual4	0.11	0.01	0.09	0.08

Participants were successful in creating fusion runs that were better than the original text and visual runs, as well being substantially better than the official mixed runs that had been submitted to ImageCLEFmed 2009. We received 49 runs from five groups. Of the 35 mixed runs that were submitted, 18 had higher MAP compared to the best textual training run and interestingly, 25 had higher MAP compared to the best official mixed run in 2009 as seen in Figure 1. This shows the potential performance gains through fusing varying techniques.

Figure 2 shows the precisions of the best original runs and the best fusion runs. There is a slight improvement in early precision with the best fusion runs both textual and mixed. However, the fusion runs created using only visual runs performed quite poorly, which is not surprising. Although there was little difference between the best fusion mixed and textual runs for the MAP, the runs with highest early precision used the visual runs in combination with the textual runs.

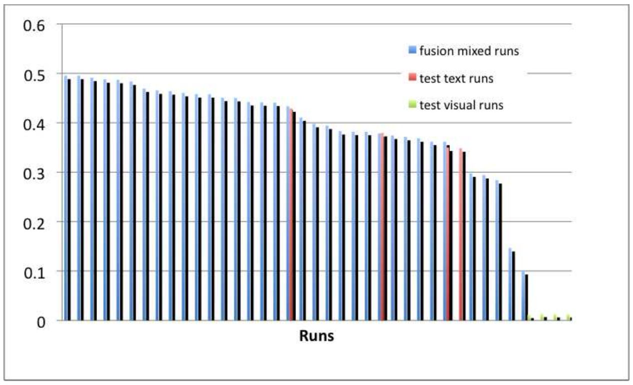

Figure 1. MAP of all fusion runs and test runs.

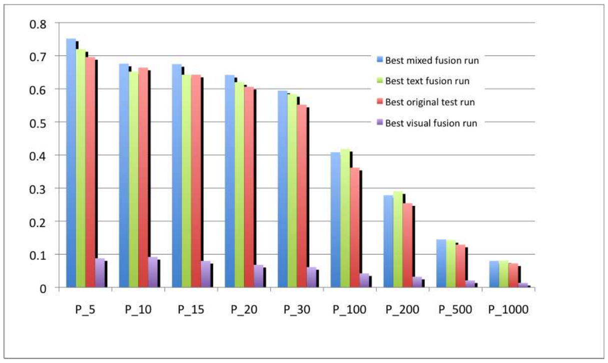

Figure 2. Early precision of original text runs and fusion runs.

Table 3 displays the best two and the worst runs for each group in the mixed fusion category, which was the category with the largest number of submissions and the best results of all groups. The early precision and the MAP of these runs are clearly superior to all the text runs shown in Table 2.

The detailed techniques used by the participants are not explained in this paper as not all participants submitted a detailed description to us. At the ICPR special session for the contest, these techniques will be compared against each other.

Combinations of only the textual runs delivered similar results to the mixed runs with the best technique (SIFT group) obtaining 0.487, so slightly lower than

Table 3. Performance metrics for fusion mixed runs.

Group	MAP	P5	P10
SIFT, Ireland	**0.495**	0.712	0.660
SIFT, Ireland	0.495	0.712	0.660
PRISMA, Chile	0.491	**0.760**	**0.696**
MedGIFT, CH	0.488	0.712	0.672
MedGIFT, CH	0.487	0.712	0.672
PRISMA, Chile	0.466	0.752	0.676
OHSU, USA	0.458	0.752	0.676
MedGIFT, CH	0.441	0.720	0.656
SIFT, Ireland	0.434	0.696	0.652
ISDM, Spain	0.383	0.688	0.668
ISDM, Spain	0.382	0.696	0.652
PRISMA, Chile	0.284	0.496	0.504
ISDM, Spain	0.100	0.352	0.292

the combination of the mixed runs. Other groups similarly had slightly better results using the mixed combinations compared to only comparing the text runs. For early precision this was similar, obtaining 0.72 compared to 0.76 for the best mixed combination run, with most other groups having a slightly lower early precision for the text only runs.

Combinations of only visual runs delivered a best MAP of 0.179 and a P5 of 0.088, both better than the results of any of the visual runs submitted but far from satisfying. As the topics were rather oriented towards semantics this was expected, though.

5. Conclusions

The first fusion challenge to combine visual and textual runs from medical image retrieval was organized for ICPR 2010. The goal of this context was to encourage participants to explore machine learning and other advanced techniques to effectively combine runs from the ImageCLEFmed challenge given a set of training runs and their performance metrics. Five groups submitted a total of 49 runs, many of which demonstrated the effectiveness of a multimodal approach to image retrieval. It was encouraging to note that about half of the submitted runs performed better than all the test runs. On the other hand, a few of the mixed runs that we submitted performed poorly, possibly due to the really poor performance of the visual test runs. The best runs obtained a MAP of 0.495 compared to the best run in the ImageCLEF of 0.43 and the best combined run in ImageCLEF 2009 of even 0.41. Such gains of over 20% show the potential of well combining visual and textual cues for medical image retrieval. The focus of ImageCLEF should be on fostering such developments In the past, particularly the combination of media has been of limited effectiveness in ImageCLEF as most research groups work on either visual or textual retrieval but not the two. The small participation of only five research groups on the other hand also showed that there might be even more potential if successful techniques for fusion are consistently applied and tested.

6. Acknowledgements

We would like to acknowledge the support of the National Library of Medicine grant 1K99LM009889 and of the BeMeVIS project of the HES–SO.

References

[1] D. Harman. Overview of the first Text REtrieval Conference (TREC–1). In *Proceedings of the first Text REtrieval Conference (TREC–1)*, pages 1–20, Washington DC, USA, 1992.

[2] W. Hersh, H. Müller, J. Jensen, J. Yang, P. Gorman, and P. Ruch. Advancing biomedical image retrieval: Development and analysis of a test collection. *JAMIA*, 13(5):488–496, September/October 2006.

[3] H. Müller, T. Deselaers, E. Kim, J. Kalpathy-Cramer, T. M. Deserno, P. Clough, and W. Hersh. Overview of the ImageCLEFmed 2007 medical retrieval and annotation tasks. In *CLEF 2007 Proceedings*, volume 5152 of *Lecture Notes in Computer Science (LNCS)*, pages 473–491, Budapest, Hungary, 2008. Springer.

[4] H. Müller and J. Kalpathy-Cramer. Analyzing the content out of context — features and gaps in medical image retrieval. *International Journal on Healthcare Information Systems and Informatics*, 4(1):88–98, 2009.

[5] H. Müller, J. Kalpathy-Cramer, I. Eggel, S. Bedrick, R. Said, B. Bakke, C. E. Kahn Jr., and W. Hersh. Overview of the CLEF 2009 medical image retrieval track. In *Working Notes of CLEF 2009*, Corfu, Greece, September 2009.

[6] H. Müller, N. Michoux, D. Bandon, and A. Geissbuhler. A review of content-based image retrieval systems in medicine–clinical benefits and future directions. *IJMI*, 73(1):1–23, February 2004.

[7] S. Radhouani, J. Kalpathy-Cramer, S. Bedrick, and W. Hersh. Medical image retrieval, a user study. Technical report, Medical Inforamtics and Outcome Research, OHSU, Portland, OR, USA, June 2009.

[8] J. Savoy. Report on CLEF–2001 experiments. In *Report on the CLEF Conference 2001 (Cross Language Evaluation Forum)*, pages 27–43, Darmstadt, Germany, 2002. Springer LNCS 2406.

[9] H. D. Tagare, C. Jaffe, and J. Duncan. Medical image databases: A content–based retrieval approach. *JAMIA*, 4(3):184–198, 1997.