

Evaluating Multimodal Relevance Feedback Techniques for Medical Image Retrieval

Dimitrios Markonis · Roger Schaer ·
Henning Müller

Received: date / Accepted: date

Abstract Medical image retrieval can assist physicians in finding information supporting their diagnosis and fulfilling information needs. Systems that allow searching for medical images need to provide tools for quick and easy navigation and query refinement as the time available for information search is often short. Relevance feedback is a powerful tool in information retrieval.

This study evaluates relevance feedback techniques with regard to the content they use. A novel relevance feedback technique that uses both text and visual information of the results is proposed. The two information modalities from the image examples are fused either at the feature level using the Rocchio algorithm or at the query list fusion step using a common late fusion rule.

Results using the ImageCLEF 2012 benchmark database for medical image retrieval show the potential of relevance feedback techniques in medical image retrieval. The mean average precision (mAP) is used as the evaluation metric and the proposed method outperforms commonly-used methods. The baseline without feedback reached 16% whereas the relevance feedback with 20 images reached up to 26.35% with three steps and when using 100 images up to 34.87% in four steps. Most improvements occur in the first two steps of relevance feedback and then results start to become relatively flat. This might also be due to only using positive feedback as negative feedback often also improves results after more steps. The effect of relevance feedback in automatically spelling corrected and translated queries is investigated as well. Results without mistakes were better than spell-corrected results but the spelling correction more than double results over non-corrected retrieval.

Multimodal relevance feedback has shown to be able to help visual medical information retrieval. Next steps include integrating semantics into relevance

D. Markonis, R. Schaer, H. Mueller
HES-SO, Information Systems Institute
Rue du TechnoPole 3
3960 Sierre, Switzerland
E-mail: firstname.lastname@hevs.ch

feedback techniques to benefit of the structured knowledge of ontologies and experimenting on the fusion of text and visual information.

Keywords relevance feedback · content-based image retrieval · medical image retrieval

1 Introduction

Searching for images is a daily task for many medical professionals, especially in image-oriented fields such as radiology [9]. However, the huge amount of visual data in hospitals and the medical literature is not always easily accessible and physicians have generally little time for information search as they need to diagnose an increasing number of cases with increasing image details in a limited amount of time.

Therefore, medical image retrieval systems need to return information adjusted to the knowledge level and expertise of the user in a quick and precise fashion. A well known technique trying to improve search results by user interaction is relevance feedback [15]. Relevance feedback allows the user to mark results returned in a previous search step as relevant or irrelevant to refine the initial query. The concept behind relevance feedback is that though users may have difficulties in formulating a precise query for a specific task, they generally see quickly whether a returned result is relevant to the information need or not. This technique found use in image retrieval particularly with the emergence of content-based image retrieval (CBIR) systems [20–22]. Following the CBIR mentality, the visual content of the marked results is used to refine the initial image query. With the result images represented as a grid of thumbnails and limited metadata, relevance feedback can be applied quickly to speed up the search iterations and refine results. Recent user-tests with radiologists on a medical image search system also showed that this method is intuitive and straightforward to learn [8].

Depending on whether the user manually provides the feedback to the system (e.g. by marking results) or the system obtaining this information automatically (e.g. by log analysis) relevance feedback can be categorized as explicit or implicit. Moreover, the information obtained by relevance feedback can be used to affect the general behaviour of the system (long-term learning). In [13] a market basket analysis algorithm is applied in image retrieval using long-term learning. A recent review of short-term and long-term learning relevance feedback techniques in CBIR can be found in [7]. An extensive survey of relevance feedback in text-based retrieval systems is presented in [17] and for CBIR in [16]. Another survey [?] gives a good overview on key aspects of image retrieval relevance feedback such as the objectives of image retrieval the main relevance feedback mechanisms and the different evaluation strategies, using real users or pseudo-relevance feedback.

Strategies of simulated feedback are also presented in [?] where relevance feedback is divided into two main strategies: at the level of results of multiple image queries and at the feature level, creating a pseudo-image out of the set

of image queries. The challenges of adding negative feedback to the mechanism are also discussed, as negative examples are often much more than the positives ones, "destroying" thus the query. In [?] several alternating feature spaces are presented for relevance feedback showing to improve results as exploring new areas of the feature space in each iteration. In [?] a Bayesian approach for relevance feedback is proposed following a target search algorithm. The same authors also explore relevance feedback techniques on small displays [?].

In the medical informatics field, [1] applies CBIR with relevance feedback on mammography retrieval. In [14], an image retrieval framework using relevance feedback is evaluated on a dataset of 5000 medical images that uses support vector machines to compute the refined queries.

There are many existing medical retrieval systems that combine text and visual information: such as NovaSearch [?], Open-I¹ and others. Relevance feedback is not always available in these system and often only evaluated in a qualitative manner. In [?] an approach for relevance feedback using similarity fusion is shown to improve the retrieval performance in two iterations of medical image search. However it is only evaluated with respect to the size of the shortlist used for the pseudo-relevance feedback and not against other relevance feedback techniques.

In this paper we evaluate different explicit, short-term relevance feedback techniques using visual content or text for medical image retrieval. We propose a technique that combines visual and text-based relevance feedback and show that it achieves a competitive performance to the state-of-the-art approaches.

2 Methods

In this study the same categorization as in [?] is followed. Two main feedback strategies with respect to the retrieval stage where the relevance feedback information is added to the new query are examined. In Figure 1 the image retrieval pipeline is shown as well as the steps where relevance feedback can be used. The first strategy is at the feature level, resulting into a single feature representation for all the query images and consequently a single result list. There is no need for result list fusion in this strategy. The second one is performed at the result list fusion step, where each image query has returned a different result list. In the multi-modal approaches the combination of visual and textual information is obtained at the Result list fusion stage for both strategies.

¹ <http://openi.nlm.nih.gov/>

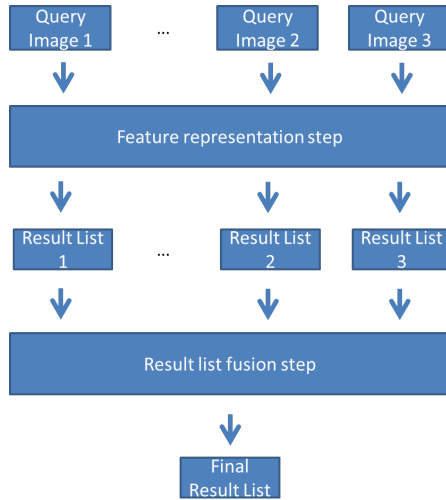


Fig. 1 Block diagram of the image retrieval pipeline. Relevance feedback strategies can be applied at the Feature representation and Result list fusion steps.

2.1 Rocchio algorithm

One of the most well known relevance feedback techniques is Rocchio's algorithm [15]. Its mathematical definition is given below:

$$\mathbf{q}_m = \alpha \mathbf{q}_o + \beta \frac{1}{|D_r|} \sum_{\mathbf{d}_j \in D_r} \mathbf{d}_j - \gamma \frac{1}{|D_{nr}|} \sum_{\mathbf{d}_j \in D_{nr}} \mathbf{d}_j \quad (1)$$

where \mathbf{q}_m is the modified query,
 \mathbf{q}_o is the original query,
 D_r is the set of relevant images,
 D_{nr} is the set of non-relevant images and
 α, β and γ are weights.

Typical values for the weights are $\alpha = 1, \beta = 0.8$ and $\gamma = 0.2$. Rocchio's algorithm is typically used in vector space models and also for CBIR. Intuitively, the original query vector is moved towards the relevant vectors and away from the irrelevant ones. By giving a weight to the positive and negative parts a problem of CBIR can be avoided that when more negative than positive feedback exists that also many relevant images disappear from the results set (basically leaving images with few features or features not present in the initial set of images).

2.2 Late fusion

Another technique that showed potential in image retrieval [5] is late fusion. Late fusion [2] is used in information retrieval to combine result lists. It can

be applied for fusing multiple queries and for multi-modal techniques where results of text and visual retrieval are for example combined. It can also be used for fusing multiple features, even though early fusion is more commonly chosen for this purpose. The concept behind this method is to merge the result lists into a single list while boosting common occurrences using a fusion rule.

For example, the fusion rule of the score-based late fusion method CombMNZ [19] is defined as:

$$S_{\text{combMNZ}}(i) = F(i) * S_{\text{combSUM}}(i) \quad (2)$$

where $F(i)$ is the number of times an image i is present in retrieved lists with a non-zero score, and $S(i)$ is the score assigned to image i . CombSUM is given by

$$S_{\text{combSUM}}(i) = \sum_{j=1}^{N_j} S_j(i) \quad (3)$$

where $S_j(i)$ is the score assigned to image i in retrieved list j .

2.3 Multi-modal relevance feedback

Most of the techniques use vectors either from the text or the visual models. However, it has been shown that approaches that use both text and visual information can outperform single-modal ones in image retrieval if performed carefully [11]. We propose the use of multi-modal information for relevance feedback to enhance the retrieval performance. This is, to the best of our knowledge, the first time that such a technique is proposed in image retrieval. As late fusion is applied on result lists, it is straightforward to use for combining results from visual and text queries.

2.4 Relevance feedback in multi-lingual queries

Another experiment run in this study is to investigate how the RF performs in more realistic scenarios when automatic spelling correction and language translation may have been applied to the query. For this, an even distribution of spelling errors across the text queries were introduced: diacritics omission, leaving out white space, character omission, character insertion, character replacement and character swapping. Automatic spelling correction was then applied to the queries, while queries in German, French and Czech were automatically translated into English.

2.5 Experimental setup

For evaluating the relevance feedback techniques the following experimental setup was followed: The n search iterations are initiated with a text query in

iteration 0. The relevant results from the top k results of iteration i were used in the relevance feedback formula of the iteration $i + 1$ for $i = 0 \dots n - 2$.

The image dataset, topics and ground truth of ImageCLEF 2012 medical image retrieval task [10] were used in this evaluation. The ad-hoc image based topics were used in our study. The dataset contains more than 300,000 images from the medical open access literature (subset of PubMed Central).

The image captions were accessed by the text-based runs and indexed with the Lucene² text search engine. A vector space model was used along with tokenization, stopword removal, stemming and inverse document frequency-term frequency weighting. The Bag-of-visual-words model described in [3] and the bag-of-colors model appearing in [4] were used for the visual modelling of the images. Since only positive feedback was used in this study, no weights were used for the Rocchio algorithm. In multi-modal runs, the fusion of the visual and text information is performed only for the 1000 top results as in the evaluation of ImageCLEF only the top 1000 documents are taken into account in any case.

Five techniques were evaluated in this study:

1. **text**: text-based RF using vector space model. Word stemming, tokenization and stopword removal is performed in both text and multi-modal runs.
2. **visual_rocchio**: visual RF using Rocchio to fuse the relevant image vectors and CombMNZ fusion to fuse the original query results with the visual results.
3. **visual_lf**: visual RF using late fusion (and the CombMNZ fusion rule) to fuse the relevant image results and the original query results with the visual ones.
4. **mixed_rocchio**: multi-modal RF using Rocchio to fuse the relevant image vectors and CombMNZ fusion to fuse the original query results with the relevant caption results and relevant visual results.
5. **mixed_lf**: multi-modal RF using late fusion (and the CombMNZ fusion rule) to fuse the relevant image results and the original query results with the caption text results and relevant visual results.

Regarding the experiment for relevance feedback in combination with the machine translation and automatic spelling-correction, the Health on the Net (HON) spell checker³ was used. The number of spellchecked recommendations to return was set to 25. The decision whether to take the highest frequency spelling suggestion (or keep the existing user query as being spelt correctly) as the correctly spelt term was made based on the ratio of spell suggestion frequency to query term frequency in the collection (threshold set to $\geq 8:1$ experimentally).

The MOSES system [6] was used to automatically translate German, French and Czech queries into English. The ImageCLEF dataset contains translations

² <http://lucene.apache.org/>

³ <https://github.com/healthonnet/HonSpellcheckerClient>

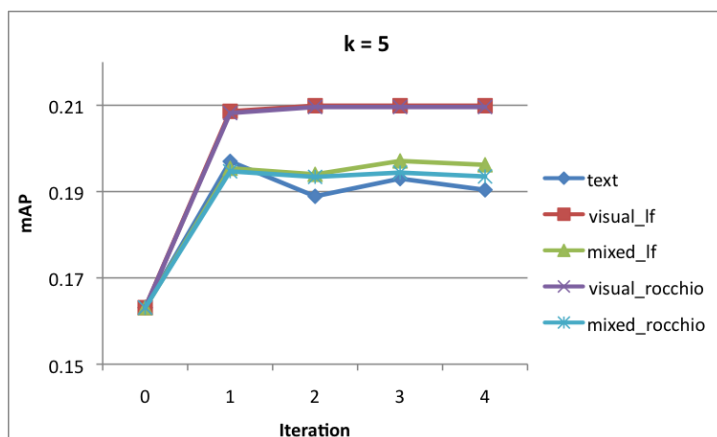


Fig. 2 Mean average precision per search iteration for $k = 5$.

of the queries for German and French while manual translation was used to translate the queries in Czech.

Three main runs were evaluated for each language:

- The first run used the queries after being translated into English by the query translation service.
- The second run used the queries after being translated into English by the query translation service and spelling errors were artificially introduced.
- The third run used the translated queries as the two runs above with the spelling errors corrected by the spelling correction service.

All the above queries were used as input to the experiment described in the beginning of this section using the best performing RF technique and $k = 100$.

3 Results

The evaluation of the five techniques was performed for $k = 5, 20, 50, 100$ and $n = 5$. Results of the mean average precision (mAP) of each technique per iteration are shown in Figures 2, 3, 4, 5.

Table 1 gives the best mAP scores of each run. The numbers in parentheses are the number of the iteration when this score was achieved. For scores that were the same in multiple iterations of the same run, the iteration closer to the first is used.

Table 2 shows the effect of the translation (row 1), the introduction of spelling errors (row 2), and the automatic spelling correction (row 3) to the retrieval performance before applying any RF.

Figure 6 demonstrates the mean average precision at iteration 4 of the *mixed_lf* technique.

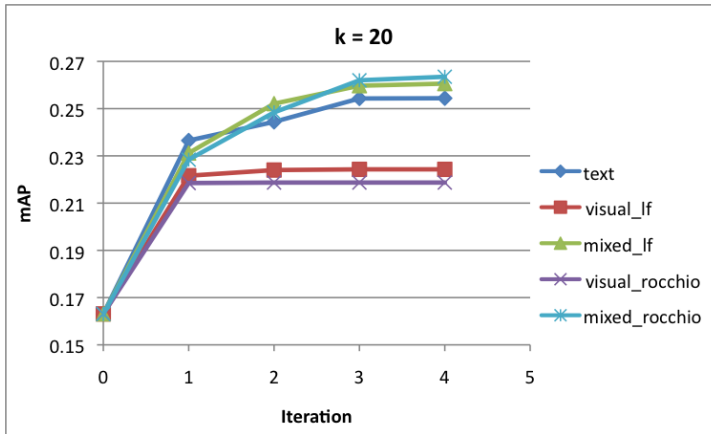


Fig. 3 Mean average precision per search iteration for $k = 20$.

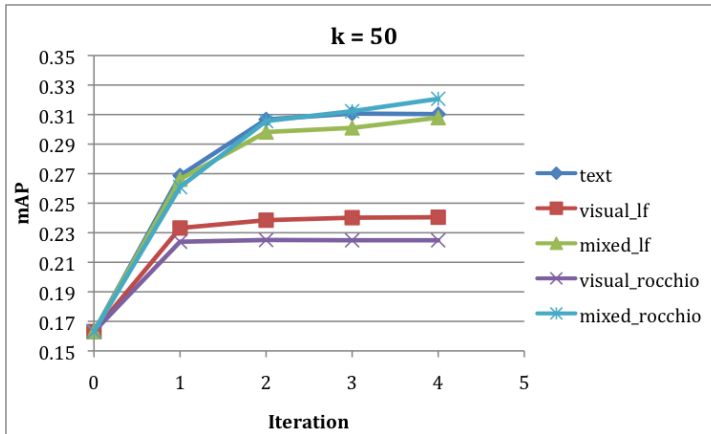


Fig. 4 Mean average precision per search iteration for $k = 50$.

Table 1 Best mAP scores

Run	k = 5	k = 20	k = 50	k = 100
text	0.197 (1)	0.2544 (4)	0.3107 (3)	0.3349 (4)
visual_lf	0.2099 (2)	0.2243 (3)	0.2405 (4)	0.2553 (3)
visual_roc	0.2096 (2)	0.2187 (2)	0.2249 (3)	0.2268 (2)
mixed_lf	0.1971 (3)	0.2606 (4)	0.3079 (4)	0.3487 (3)
mixed_roc	0.1947 (1)	0.2635 (4)	0.3207 (4)	0.3466 (4)

Table 3 shows the results of a sample query for different categories of relevance feedback methods, illustrating the differences in the results when using text, visual and mixed relevance feedback.

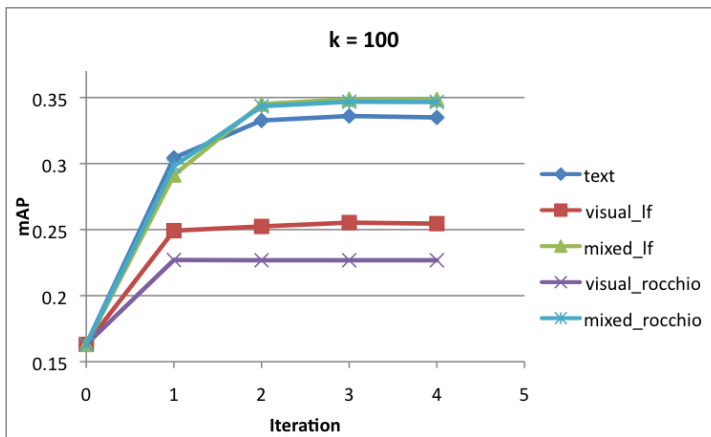


Fig. 5 Mean average precision per search iteration for $k = 100$.

Table 2 Iteration 0 mAP for each of the languages. The values are for, respectively, no spelling error, spelling error introduced and spelling correction applied. Machine translation into English is applied for the French, Czech and German queries.

	English	French	Czech	German
No spelling error	0.1783	0.1683	0.1381	0.1270
Spelling error	0.0468	0.0904	0.0321	0.0448
Spelling correction	0.1349	0.0973	0.0995	0.1023

4 Discussion

Medical image retrieval from articles in the literature is a challenging task as the image datasets from the biomedical literature are quite noisy (containing many graphs, diagrams and non-medical pictures). Moreover the areas in the images containing pathology related information are small and the differences are quite subtle with usually only a very small portion of the image being relevant.

All of the evaluated techniques improve retrieval after the initial search iteration. This demonstrates the potential of relevance feedback for refining medical image search queries. Relevance feedback using only visual appearance models, even though improving the retrieval performance after the first iteration, performed worse than the text-based runs in most cases. Visual features still suffer from the semantic gap between the expressiveness of visual features and our human interpretation. Still, this shows their usefulness in image datasets where no or few text meta-data are available. Moreover, when combined with the text-information in the proposed method, they improve the text-only baseline. Recently introduced higher-order representations, such as Fisher Vectors or Vectors of locally aggregated descriptors (VLADs) may further improve the retrieval results in this scenario.

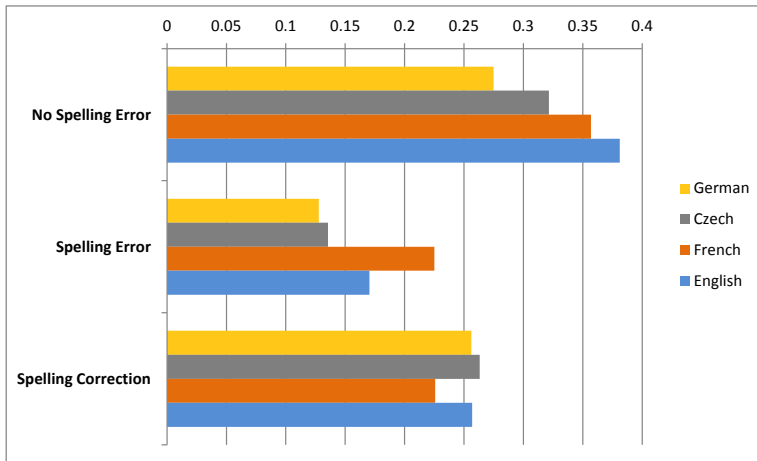


Fig. 6 Plot of the mixed.lf values for iteration 4 for each of the languages. The values are for, respectively, no spelling error, spelling error introduced and spelling correction applied. Machine translation into English is applied for the French, Czech and German queries.


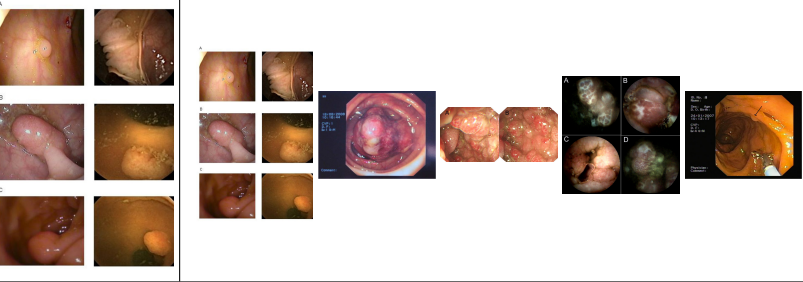
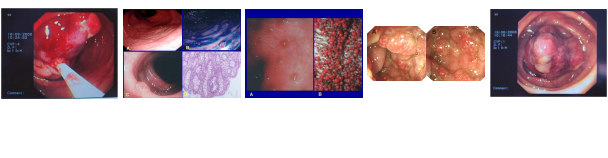
The proposed multi-modal runs provide the best results in all the cases except for case $k = 5$. Surprisingly, the visual runs perform slightly better than the text and the multi-modal approaches for this case. However, assuming independent and normal distributed average precision values the significance tests show that the difference is not statistically significant.

We consider the case $k = 20$ as the most realistic scenario since users do not often inspect more than 2 pages of results and most actually stay only on the initial results page. Especially for grid-like result interface views, where each page can contain 20 to 50 results, we consider $k = 20$ more realistic than $k = 5$. In this case the proposed methods achieve the best performance with 0.2606 and 0.2635 respectively. Again, the significance tests do not find any significant difference between the three best approaches. However, applying different fusion rules for combining visual and text information (such as linear-weighting) could further improve the results of the mixed approaches. It can be noted that as the k increases, the performance improvement also increases, highlighting the added value of relevance feedback. Larger values of k than 100 were not explored as these scenarios were judged as unrealistic.

In the visual runs using Rocchio for combining the visual queries is performing worse than late fusion. This comes in accordance with the findings in [3]. The reason behind this could be that the large visual diversity of relevant images in medicine and the curse of dimensionality cause the modified vector to behave as an outlier in the high dimensional visual feature space. In the mixed runs the difference between the two methods is not statistically significant with Rocchio performing slightly better than the late fusion.

Relevance feedback is shown to be able to improve the retrieval performance in difficult real world scenarios where spelling errors are introduced and corrected as well as machine translation is applied to the queries.

Table 3 Sample query : top 5 results of iteration 4 ($k = 100$) for the 3 categories of methods (text, visual and mixed). The late fusion (lf) variants of the visual and mixed methods are shown.

	query	results
text	Endoscopy images showing colon polyps	
visual_lf		
mixed_lf	Endoscopy images showing colon polyps	

Irrelevant results were ignored, as they often have little or no impact on the retrieval performance [12, 18]. More importantly, the ground truth of the dataset used contains a much larger portion of annotated irrelevant results than relevant ones. This was considered to potentially simulate an unrealistic scenario, as users do not usually mark many results as negative examples (or only very few). Having too many negative examples could also cause the modified vector to follow an outlier behaviour. Preliminary results confirmed this hypothesis, where the use of negative results for relevance feedback can decrease performance after the first iteration if not handled in a careful way.

A larger number of steps could be investigated but this might be unrealistic, given the fact that physicians have little time and stop after a few minutes of search [9]. Often users will only test a few steps of relevance feedback at the most.

In the evaluation of our relevance feedback mechanism we assume a perfect user who marks all relevant items as relevant. As past literature shows [?], by selecting not all relevant items but those adding most information, a human user familiar with the system can potentially achieve better results than the automatic system when adding positive and negative feedback, so we feel that

this is a good approximation. For novice users this is maybe unlikely and a novice user might thus simply add all relevant items which is the scenario we take for our feedback evaluation.

The advantage of perfect positive feedback is also that results become reproducible, so the exact same conditions can be reproduced for other techniques whereas manual user tests can depend strongly on the person supplying the feedback and thus do not allow comparing performance over several systems in a homogeneous way. User tests with a very similar system for several users have been published [8], and we feel for this article a reproducible way for supplying relevance feedback is better. This does not replace user tests and more on interaction and interface design can be learned via such user tests.

5 Conclusions

This paper proposes the use of multi-modal information when applying relevance feedback to medical image retrieval. An experiment was set up to simulate the relevance feedback of a user on medical topics from ImageCLEF 2012.

In general, all the techniques evaluated in this study improve the performance, which shows the added value of relevance feedback. Text-based relevance feedback showed consistently good results. Visual techniques showed competitive performance for a small number of steps, underperforming in the remaining cases. The proposed multi-modal approaches show promising results slightly outperforming the text-based one but without statistical significance.

More fusion techniques are going to be evaluated in the future. Comparison to manual query refinement by users is considered in future plans to assess relevance feedback as a concept in medical image retrieval. The addition of semantic search is also of interest, to take advantage of the structured knowledge of the medical ontologies such as RadLex⁴ (Radiology Lexicon) [?] and MeSH⁵ (Medical Subject Headings) [?] or other medical ontologies to model semantic knowledge.

6 Acknowledgements

This work was supported by the EU 7th Framework Program in the context of the Khresmoi project (grant 257528).

References

1. Chen, C.C., Huang, P.J., Gwo, C.Y., Li, Y., Wei, C.H.: Mammogram retrieval: Image selection strategy of relevance feedback for locating similar lesions. *International Journal of Digital Library Systems (IJDLIS)* **2**(4), 45–53 (2011)

⁴ <http://www.radlex.org/>

⁵ <http://www.nlm.nih.gov/mesh/MBrowser.html>

2. Depeursinge, A., Müller, H.: Fusion techniques for combining textual and visual information retrieval. In: H. Müller, P. Clough, T. Deselaers, B. Caputo (eds.) *ImageCLEF, The Springer International Series On Information Retrieval*, vol. 32, pp. 95–114. Springer Berlin Heidelberg (2010)
3. García Seco de Herrera, A., Markonis, D., Eggel, I., Müller, H.: The medGIFT group in ImageCLEFmed 2012. In: *Working Notes of CLEF 2012* (2012)
4. García Seco de Herrera, A., Markonis, D., Müller, H.: Bag of colors for biomedical document image classification. In: H. Greenspan, H. Müller (eds.) *Medical Content-based Retrieval for Clinical Decision Support, MCBR-CDS 2012*, pp. 110–121. *Lecture Notes in Computer Sciences (LNCS)* (2013)
5. García Seco de Herrera, A., Markonis, D., Schaer, R., Eggel, I., Müller, H.: The medGIFT group in ImageCLEFmed 2013. In: *Working Notes of CLEF 2013 (Cross Language Evaluation Forum)* (2013)
6. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E.: Moses: Open source toolkit for statistical machine translation. In: *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, ACL '07*, pp. 177–180. Association for Computational Linguistics, Stroudsburg, PA, USA (2007)
7. Li, J., Allinson, N.M.: Relevance feedback in content-based image retrieval: a survey. In: *Handbook on Neural Information Processing*, pp. 433–469. Springer (2013)
8. Markonis, D., Baroz, F., Ruiz de Castaneda, R.L., Boyer, C., Müller, H.: User tests for assessing a medical image retrieval system: A pilot study. In: *MEDINFO 2013* (2013)
9. Markonis, D., Holzer, M., Dungs, S., Vargas, A., Langs, G., Kriewel, S., Müller, H.: A survey on visual information search behavior and requirements of radiologists. *Methods of Information in Medicine* **51**(6), 539–548 (2012)
10. Müller, H., García Seco de Herrera, A., Kalpathy-Cramer, J., Demner Fushman, D., Antani, S., Eggel, I.: Overview of the ImageCLEF 2012 medical image retrieval and classification tasks. In: *Working Notes of CLEF 2012 (Cross Language Evaluation Forum)* (2012)
11. Müller, H., Kalpathy-Cramer, J.: The ImageCLEF medical retrieval task at icpr 2010 — information fusion to combine visual and textual information. In: *Proceedings of the International Conference on Pattern Recognition (ICPR 2010)*, *Lecture Notes in Computer Science (LNCS)*. Springer, Istanbul, Turkey (2010)
12. Müller, H., Müller, W., Squire, D.M., Marchand-Maillet, S., Pun, T.: Strategies for positive and negative relevance feedback in image retrieval. Tech. Rep. 00.01, Computer Vision Group, Computing Centre, University of Geneva, rue Général Dufour, 24, CH-1211 Genève, Switzerland (2000)
13. Müller, H., Squire, D.M., Pun, T.: Learning from user behavior in image retrieval: Application of the market basket analysis. *International Journal of Computer Vision* **56**(1–2), 65–77 (2004). (Special Issue on Content-Based Image Retrieval)
14. Rahman, M.M., Bhattacharya, P., Desai, B.C.: A framework for medical image retrieval using machine learning and statistical similarity matching techniques with relevance feedback. *Information Technology in Biomedicine, IEEE Transactions on* **11**(1), 58–69 (2007)
15. Rocchio, J.J.: Relevance feedback in information retrieval. In: *The SMART Retrieval System, Experiments in Automatic Document Processing*, pp. 313–323. Prentice Hall, Englewood Cliffs, New Jersey, USA (1971)
16. Rui, Y., Huang, T.S., Mehrotra, S.: Relevance feedback techniques in interactive content-based image retrieval. In: I.K. Sethi, R.C. Jain (eds.) *Storage and Retrieval for Image and Video Databases VI, SPIEProc*, vol. 3312, pp. 25–36 (1997)
17. Ruthven, I., Lalmas, M.: A survey on the use of relevance feedback for information access systems. *The Knowledge Engineering Review* **18**(02), 95–145 (2003)
18. Salton, G., Buckley, C.: Improving retrieval performance by relevance feedback. *Readings in information retrieval* **24**, 5 (1997)
19. Shaw, J.A., Fox, E.A.: Combination of multiple searches. In: *TREC-2: The Second Text REtrieval Conference*, pp. 243–252 (1994)

-
20. Squire, D.M., Müller, W., Müller, H., Pun, T.: Content-based query of image databases: inspirations from text retrieval. *Pattern Recognition Letters (Selected Papers from The 11th Scandinavian Conference on Image Analysis SCIA '99)* **21**(13-14), 1193-1198 (2000). B.K. Ersboll, P. Johansen, Eds.
 21. Taycher, L., Cascia, M.L., Sclaroff, S.: Image digestion and relevance feedback in the ImageRover WWW search engine. pp. 85-94 (1997)
 22. Wood, M.E., Campbell, N.W., Thomas, B.T.: Iterative refinement by relevance feedback in content-based digital image retrieval. pp. 13-20 (1998)