Comparing feature sets for content–based image retrieval in a medical case database

Henning Müller[a], Antoine Rosset[a], Jean–Paul Vallée[a], Antoine Geissbuhler[a]

[a]University Hospitals of Geneva, Service of Medical Informatics, 24, Rue Micheli-du-Crest, CH-1211 Geneva 14, Switzerland;

ABSTRACT

Content–based image retrieval systems (CBIRSs) have frequently been proposed for the use in medical image databases and PACS. Still, only few systems were developed and used in a real clinical environment. It rather seems that medical professionals define their needs and computer scientists develop systems based on data sets they receive with little or no interaction between the two groups. A first study on the diagnostic use of medical image retrieval also shows an improvement in diagnostics when using CBIRSs which underlines the potential importance of this technique.

This article explains the use of an open source image retrieval system (GIFT – GNU Image Finding Tool) for the retrieval of medical images in the medical case database system CasImage that is used in daily, clinical routine in the university hospitals of Geneva. Although the base system of GIFT shows an unsatisfactory performance, already little changes in the feature space show to significantly improve the retrieval results. The performance of variations in feature space with respect to color (grey level) quantizations and changes in texture analysis (Gabor filters) is compared. Whereas stock photography relies mainly on colors for retrieval, medical images need a larger number of grey levels for successful retrieval, especially when executing feedback queries. The results also show that a too fine granularity in the grey levels lowers the retrieval quality, especially with single–image queries

For the evaluation of the retrieval performance, a subset of the entire case database of more than 40,000 images is taken with a total of 3752 images. Ground truth was generated by a user who defined the expected query result of a perfect system by selecting images relevant to a given query image. The results show that a smaller number of grey levels (32–64) leads to a better retrieval performance, especially when using relevance feedback. The use of more scales and directions for the Gabor filters in the texture analysis also leads to improved results but response time is going up equally due to the larger feature space.

CBIRSs can be of great use in managing large medical image databases. They allow to find images that might otherwise be lost for research and publications. They also give students students the possibility to navigate within large image repositories. In the future, CBIR might also become more important in case–based reasoning and evidence–based medicine to support the diagnostics because first studies show good results.

1. INTRODUCTION

Content–based image retrieval (CBIR) is one of the most active research areas within the computer vision field. The spread of digital cameras at consumer prices and especially the availability of large amounts of multimedia information via the Internet has created the need to develop tools to manage these data. CBIR tries to solve this problem without the (expensive) use of manual annotation. Visual image features are extracted from the images automatically and indexed into a database. This means that semantic information about the image content is frequently lost (*semantic gap*).[1] The most frequent query paradigm is query by example where the system returns the most similar images from a database with respect to (an) example(s) provided by the user. Some of the well known systems are QBIC[*2] and Virage[†3] in the commercial field and Blobworld[4] or Photobook[5] as research systems. Good overview articles of CBIR are.[1, 6–8]

Further author information: (Correspondence to Henning Müller) henning.mueller@dim.hcuge.ch
, tel. ++41 22 372 61 75, fax ++41 22 372 8680

[*] `http://wwwqbic.almaden.ibm.com/`

[†] `http://www.virage.com/`

Medical image retrieval based on the image content has been proposed several times for the use in PACS systems[9, 10] or case databases[11, 12] and in an even more general sense.[13–15] Still, only few system such as the IRMA[‡] system on case classification[16] and the ASSERT (Automatic Search and Selection Engine with Retrieval Tools) system on High Resolution Lung CT images (HRCT)[17] have been implemented and evaluated. The ASSERT project has also performed a study on the use of content-based image retrieval as a diagnostic aid.[18] This study showed an improvement in diagnostic quality, especially for less experienced radiologists.

Even fewer systems seem to be used in clinical practice. Most of the systems stay research prototypes in a variety of medical departments from the analysis of tumor shapes,[19] neurology[20] to pathology images.[21] Other articles show ideas for architectures of retrieval system integration[10, 22] but no or only a partial implementation seems to have taken place. Rarely, implementation details on the used visual features are given and sometimes only text is used as a content description for the (visual) images.[10, 23] Two review articles present some of the techniques and currently employed systems in the domain.[24, 25]

2. AVAILABLE DATA FOR SYSTEM EVALUATION

The case database system CasImage[§][11] has been developed at the University Hospitals of Geneva where it is used in daily practice by the MDs to store interesting or typical cases.

2.1. Images

The entire database of CasImage currently contains more than 40,000 images, with new images being added daily. For the performance comparison in this paper, we selected a subset of these images to limit the generation of ground truth and shorten the time to execute the queries and index the data with varying color schemes. We choose a total of 3752 images from the center of the database containing a large variety of differing images from CTs, MRIs, to radiographs but also photos. Most images are grey level images but there are several colored images as well. Figure 1 shows four image examples from the database. We are currently underway to generate a larger image dataset that is anonymized and can be distributed freely to foster the comparison of various retrieval systems.

As CasImage uses a web interface, the images are converted from the DICOM format or other formats into JPEG. The level/window settings can be chosen by the user on insertion of an image. This means an information loss and a maximum grey level depth of 256 values. For real clinical studies, the use of DICOM images with the full grey level information is necessary.

2.2. Ground truth

When evaluating information retrieval systems, there is always the problem of defining what a perfect system response would be like. This process of finding ground truth can be done based on user judgments who are given a query image and then have to find out what, in their opinion, needs to be returned by the system. Such ground truthing is used for textual information retrieval (TREC[¶],[26]) as well as for image retrieval.[27] When ground truth in form of diagnostics or a classification of body parts is available, it is much easier to automatically define the ground truth. Due to the extremely varying quality of annotation in CasImage, this was not possible. Another possibility for ground truthing is annotation or coding and the automatic generation of ground truth from annotation.[28]

The problem with manual ground truthing or coding is that large databases take a very long time to be analyzed. To reduce the number of documents to be controlled, a pooling technique was introduced in[29] and is used in all TREC ground truthings as well as in other text retrieval evaluation tasks. This technique means that only the highest–ranked documents of each system in the evaluation have to be observed. This technique is known to influence the absolute results of system performance[30] which can mean that for TREC the recall results are incorrect starting from 50–60% recall.[26] Still, the studies show that the results are not biased and rather small.

[‡] http://www.irma-project.org/

[§] http://www.casimage.com/

[¶] http://trec.nist.gov/

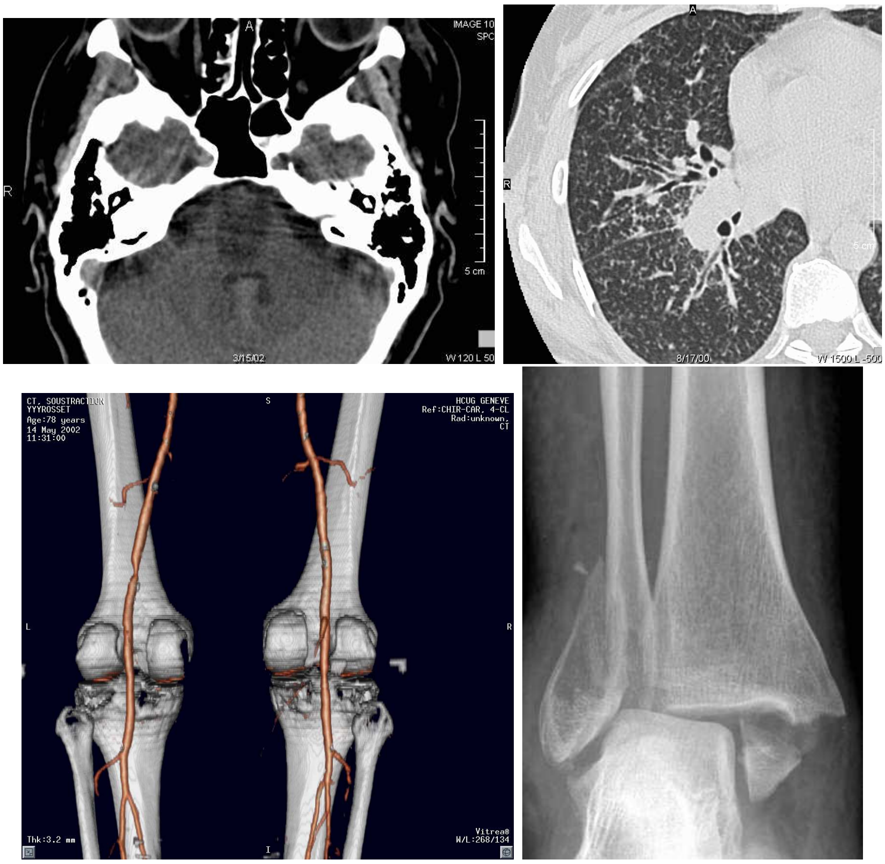


Figure 1. Some example images of the CasImage system.

Ten images from the CasImage collection were chosen as query images for the evaluation of the system. They are from various anatomic regions, have varying difficulties and represent mostly grey scale but also a few queries with colors. In our tests, one users was ground truthing the database. We used pooling as in TREC where the user had to watch the highest–ranked 200 documents of each system, in our case of each color quantization. For the seven different systems, around 500 different images had to be watched per query compared to 3752 if the entire database would have to be checked. As relevant images, an average of 50 images was chosen with 115 being the highest and 12 the lowest. An average of 35 were chosen from the 200 response images of the first system and 15 from the additional images that were supplied by the other systems. This shows that it is important to use pooling and not simply the highest–ranked replies from the one system for ground truthing.

2.3. Evaluation of the retrieval systems

We use an automatic benchmark based on perl scripts that uses the performance measures commonly defined in the image and text retrieval literature.[26, 31] Most measures are based on precision and recall. that are defined as follows:

$$precision = \frac{number\ of\ relevant\ items\ retrieved}{number\ of\ all\ items\ retrieved} \tag{1}$$

$$recall = \frac{number\ of\ relevant\ items\ retrieved}{number\ of\ all\ relevant\ items} \tag{2}$$

The measures used are easy to interpret, show different system aspects and are standard in the field of (visual) information retrieval:

- Basic precision and recall measures after 20, 50 and 100 images;
- precision vs. recall (PR) graphs;
- rank of the first relevant image and average rank measures;
- query response time.

As most important performance measures we see the precision after 20 and 50 images as this corresponds to the number of images a user watches on screen. A good measure for overall system performance is the normalized average rank:

$$\widetilde{Rank} = \frac{1}{NN_R}\left(\sum_{i=1}^{N_R} R_i - \frac{N_R(N_R+1)}{2}\right) \tag{3}$$

where R_i is the rank at which the ith relevant image is retrieved, N the number of images in the database and N_R the number of relevant images. This measure is 0 for perfect performance, and approaches 1 as performance worsens. For random retrieval the result would be 0.5. This measure is very similar to the one used in MPEG–7 and BIRDS-I.[32, 33]

The most powerful measure to compare several systems is the PR graph as it contains much information and is easy to interpret. A perfect system response would result in a straight line at precision 1 up to recall 1. Curves that are higher up are better than curves that are further down. The beginning of the graph is most important as these are the first images that are retrieved (and by consequence the ones that are shown on screen)).

The communication with the retrieval system is done via the Multimedia Retrieval Markup Language (MRML[‖]) to have a completely automated access.

The time measured for the retrieval includes the transmissions of URLs for all 3752 images in order of similarity. which means that a user looking at 20-50 images on screen will have a much faster reply (around 2–3 seconds faster). Response time was taken on a slow Pentium III-500 machine with 256 MB RAM.

All queries were executed for each retrieval system. Based on the first 20 result images, positive and negative relevance feedback (RF) is generated for each system and the performance with RF is evaluated. Often, the performance with RF is seen to be much more important than in the first query step where the retrieval system had less information.[34] Results are finally averaged over all the images that was queried with.

3. MEDGIFT – A CONTENT–BASED IMAGE RETRIEVAL SYSTEM

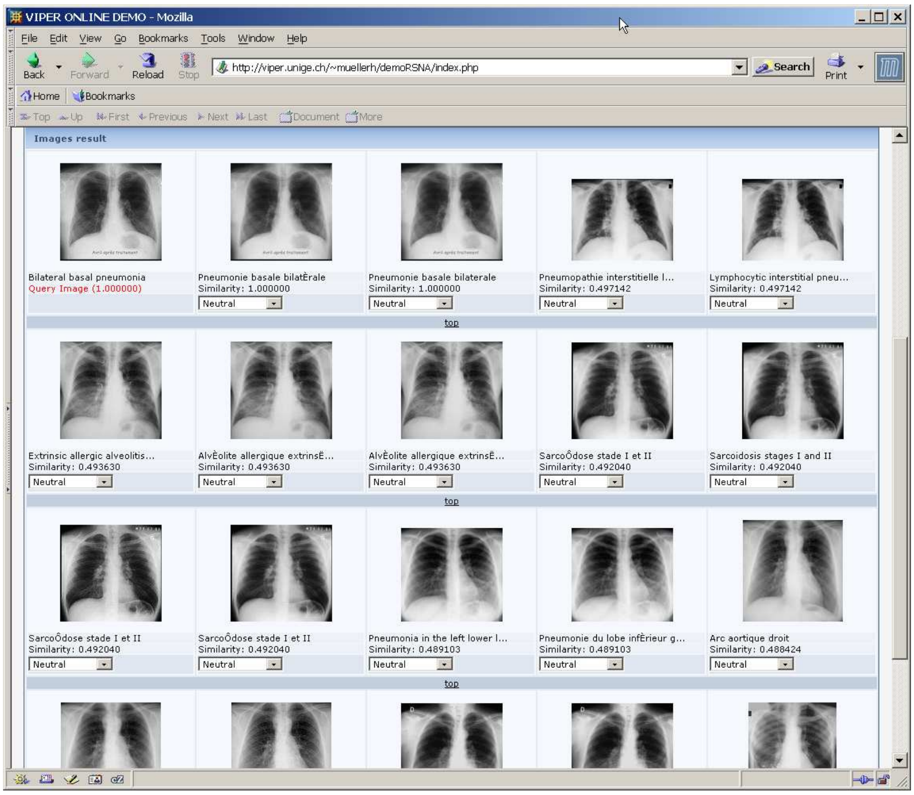
The medGIFT[**] system is an adaptation of the GIFT[††] (GNU Image Finding Tool) that is the result of the Viper[‡‡] project (Visual Information Processing for Enhanced Retrieval). For using medical images, several changes in feature space were implemented. The interface with an example query is shown in Figure 2. The addition of more features appropriate for the use with medical images is in process. These features will include mainly texture measures such as those based on cooccurence matrices, edge histograms and also local features based on image segmentation. It still needs to be analyzed whether this does add information over the currently used responses of Gabor filters.

[‖] `http://www.mrml.net/`

[**] `http://www.sim.hcuge.ch/medgift/`

[††] `http://www.gnu.org/software/gift/`

[‡‡] `http://viper.unige.ch/`


VIPER ONLINE DEMO - Mozilla
File Edit View Go Bookmarks Tools Window Help
Back
Forward
Reload
Stop
http://viper.unige.ch/~muellerh/demoRSNA/index.php
Search
Print
Home
Bookmarks
Top
Up
First
Previous
Next
Last
Document
More
top

Images result

Bilateral basal pneumonia Query Image (1.000000)	Pneumonie basale bilatÈrale Similarity: 1.000000 Neutral	Pneumonie basale bilaterale Similarity: 1.000000 Neutral	Pneumopathie interstitielle l... Similarity: 0.497142 Neutral	Lymphocytic interstitial pneu... Similarity: 0.497142 Neutral
Extrinsic allergic alveolitis... Similarity: 0.493630 Neutral	AlvÈolite allergique extrinsË... Similarity: 0.493630 Neutral	AlvÈolite allergique extrinsË... Similarity: 0.493630 Neutral	SarcoÔdose stade I et II Similarity: 0.492040 Neutral	Sarcoidosis stages I and II Similarity: 0.492040 Neutral
SarcoÔdose stade I et II Similarity: 0.492040 Neutral	SarcoÔdose stade I et II Similarity: 0.492040 Neutral	Pneumonia in the left lower l... Similarity: 0.489103 Neutral	Pneumonie du lobe infÈrieur g... Similarity: 0.489103 Neutral	Arc aortique droit Similarity: 0.488424 Neutral

Figure 2. A screen shot of the medGIFT system.

3.1. The GNU Image Finding Tool (GIFT)

The technology of the GIFT is described in much detail in.[27, 35] GIFT is free of charge, the source code can be downloaded for changes and its usage in research projects is thus fairly easy. It has already been used in several research projects other than those of the original system developers.[36, 37] Its main techniques are based on experiences from text retrieval using frequency–based feature weights based on two principles:

- Features that are frequent in an image describe this image well;
- features that are frequent in the collection do not distinguish images well from each other.

As features, the GIFT uses global and local color and texture features. The color features are calculated in HSV (Hue, Saturation, Value) space which is closer to human perception than spaces such as RGB (Red, Green, Blue) or CMY (Cyan, Magenta, Yellow).[38] It uses 166 colors, including 4 grey levels. The texture features are based on Gabor filter responses in four directions and at 3 different scales that are classified into ten different strengths. All features are computed offline and stored in an inverted file structure, well–known from text retrieval, which gives efficient access to the very large number of possible features ($> 85,000$). Each image contains 1,000-2,000 features.

3.2. Changes for the use with medical images

To adapt the GIFT for the use with medical images, mainly changes in the color quantization were done, reducing the number of colors and rising the number of grey levels. These small changes already lead to much better results.

The basic GIFT system uses a quantization of the HSV space into 16 hues, 3 saturations and 3 values, plus 4 levels of grey. This had to be changed to add more levels of grey and reduce the number of colors. We also added more directions and more scales to the Gabor filters to evaluate these effects on the system performance. The following system configurations were tested:

Parameter for variation	gift1	gift2	gift3	gift4	gift5	gift6	gift7
Hue	18	6	6	6	6	6	6
Saturation	3	3	3	2	2	3	3
Value	3	3	3	3	2	3	3
Grey levels	4	32	64	128	200	64	64
Gabor directions	4	4	4	4	4	6	6
Gabor scales	3	3	3	3	3	3	5

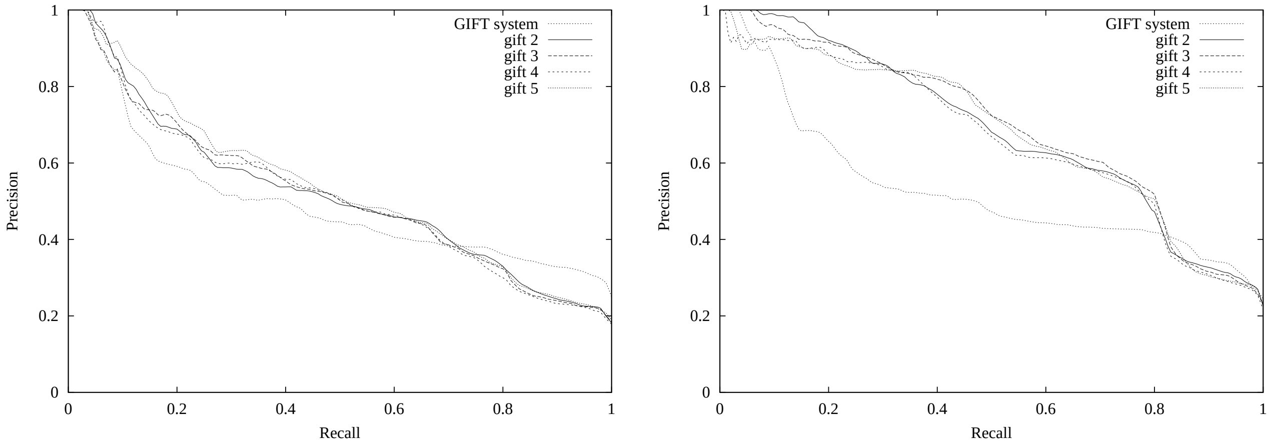
Table 1. These configurations were tested with the GIFT.

The reduction of saturation and hue was chosen for the systems using a large number of grey levels because of a maximum of 256 colors in the current GIFT system. As the images are in JPEG, this should be sufficient, and the results show that a smaller number of grey levels actually leads to better results. Still, the number of colors for retrieval does not need to be as high as in the images because the quantization can otherwise get too specific for retrieval.

A larger number of Gabor directions leads to even better results but the time to execute a query and to store the features will also rise enormously. More scales of the Gabor filters deliver equally better results and longer execution times.

4. RESULTS

This section presents the results of the evaluation procedure in the graphical form of PR graphs and as several numerical measures. The results of the performance measures are subsequently analyzed to define the techniques that are best suited for medical image retrieval.



(a) first query step

(b) with relevance feedback

Figure 3. PR graphs without and with feedback using various color quantizations.

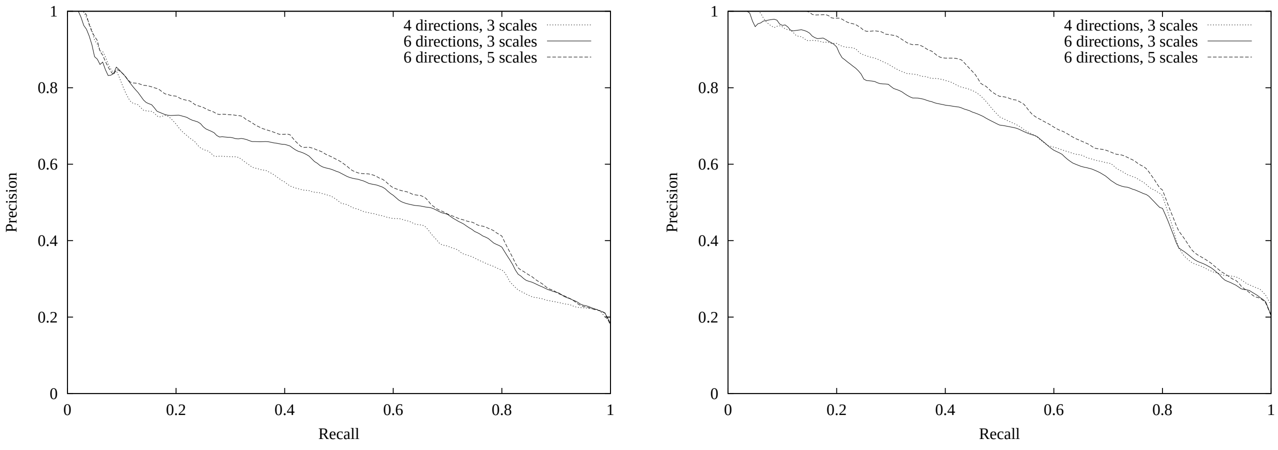
Figure 3 shows the comparison of the color quantizations mentioned in Table 1. In the first query step, the results are fairly similar with the standard GIFT being the worst. Higher numbers of grey levels get significantly better results with the system using 200 grey levels being slightly better than the others. When analyzing the second graph that shows the results with RF, the picture changes. The base GIFT gets by far the worst results whereas the system with fewer grey levels is best. When using several steps of RF these results continue.

The fact that 32 and 64 grey levels lead to the best results when using RF means that a larger number of grey levels is too specific for CBIR.

	without feedback					with feedback				
Measure	gift1	gift2	gift3	gift4	gift5	gift1	gift2	gift3	gift4	gift5
No. relevant Images	49.8	49.8	49.8	49.8	49.8	49.8	49.8	49.8	49.8	49.8
Query duration	14.0	7.9	8.9	8.7	8.15	12.3	10.1	11.1	11.5	10.8
Pos. of first relevant	4.7	1.1	1.3	1.1	1.2	4.3	1	1	1	1
Recall at 0.5 precision	.209	.398	.333	.420	.370	.310	.467	.475	.445	.472
Average Rank	251	202	223	225	242	350	227	253	249	255
$\widetilde{Rank}$	.060	.047	.053	.053	.058	.087	0.054	.061	.059	.061
Precision at 20	.535	.595	.615	.61	.645	.59	0.78	0.79	0.73	.755
Precision at 50	.382	.402	.42	.408	.402	.374	.482	0.48	.442	.462
Precision at Nr	.467	.498	.490	.491	.515	.497	.644	.640	.622	.637
Recall at 100	.578	.639	.641	.632	.635	.576	.728	.726	.711	.711

Table 2. Performance measures for various color quantizations with and without RF.

The values in Table 2 underline the results from the PR graphs. Without RF, the rank results are best for a small number of colors whereas the precision and recall values are very similar for all systems but the basic GIFT. With RF, the results are mostly the best when using only 32 grey levels. Only early precision and precision where the recall drops below 0.5 are better with 64 grey levels. More levels of grey lead to worse results. This observation continues with RF.



(a) first query step

(b) with relevance feedback

Figure 4. PR graphs without and with RF using varying numbers of directions and scales.

Figure 4 shows a comparison of three settings for the Gabor filters with and without RF. The results show clearly that more directions of the Gabor filters improve the results significantly. Also, the addition of more scales leads to a significant improvement. With one step of RF, the system with 6 and with 4 directions lead to very similar results whereas the system with 6 directions and 5 scales achieves the best results.

Table 3 shows how much longer the query evaluation takes for the systems with more scales and directions. This is significant and in certain domains a less perfect result can be more desirable when it can be received so much faster. In,[39] a 10 second response time is seen as the limit to keep the user focused on the dialog. The best performing system in this test is strongly above this limit. Search pruning techniques can be one way to reduce this response time.[40]

	without feedback			with feedback		
Measure	gift3	gift6	gift7	gift3	gift6	gift7
No. relevant Images	49.8	49.8	49.8	49.8	49.8	49.8
Query duration	8.9	18.1	47.4	11.1	27.1	96.8
Pos. of first relevant	1.3	1.6	1.6	1	1	1
Recall at 0.5 precision	.333	.389	.486	.475	.456	.541
Average Rank	223	192	185	253	226	190
$\widetilde{Rank}$	.053	.044	.042	.061	.054	.044
Precision at 20	.615	.615	.660	.79	.745	.835
Precision at 50	.420	.45	.460	.48	.468	.52
Precision at Nr	.490	.548	.570	.640	.616	.673
Recall at 100	.641	.696	.703	.726	.705	.743

Table 3. Performance measures for varying Gabor features with and without RF.

On the other hand, the results for the system with more directions and especially for the system with more directions and scales are significantly better. The precision and recall values are more than 5% higher and the average rank measures are even stronger so. Remarkable is the fact that the average rank values with RF are worse than in the initial query step. This is due to negative RF that also moves relevant images away from the query result as the images used for positive and negative RF are very similar in feature space. The high precision values after 20 and 50 images show that this happens at lower ranks.

The results show that the basic GIFT is not perfect for medical image retrieval but already small changes significantly enhance the performance. Without feedback the system still delivers acceptable results but with feedback the results drop dramatically. Other test with a similar database even show a more surprising results, that in the first feedback step a system with four grey levels delivers best results. Again, the quality changes with feedback where 32 gray levels seem to deliver the best results. It can be interesting to explore this effect further and use technologies such as[41] where a different feature space is used for every. Is the case of medical, this can be used for different grey level quantizations to best explore the image space and improve precision and recall.

5. CONCLUSIONS

The GIFT retrieval system has shown to be easily adaptable for the use in medical applications. It is free of charge and the source code is available and can easily be adapted. The base system can surely not be used for image retrieval in a clinical setting but with a few small changes the retrieval performance improves significantly. The retrieval quality obtained is high enough for the use in a case database such as CasImage to complement the normal text–based search, especially for teaching and finding interesting cases. Students can also profit from the technology when exploring large image repositories. For the use in systems for case–based reasoning or in evidence–based medicine, a more detailed clinical evaluation in specialized domains will be necessary and more specific features can become important.

An optimal value for the number of grey level for retrieval seems to be far lower than the maximum resolution. Using a too large number of grey levels can make retrieval too specific, so that similar cases with small variations of the grey levels are not found. A reasonable value seems to be 32 or 64 grey levels for the retrieval of varied medical images such as in CasImage. This might change when using the full grey level resolution of the DICOM images in a limited domain such as CT images because the colors correspond in this case exactly to certain density values. A promising technique might be to switch feature spaces between query steps to reach various parts of the feature space and find a maximum of potentially interesting cases. When using Gabor filters, the use of additional orientations and scales also improves the performance. On the other hand the space for storage and the response times go up significantly. A good compromise needs to be found, depending on the domain. Using 6 direction but only 3 scales seems like a reasonable option. Still, the use of more scales can be achieved with reasonable computing times when pruning is used and fast computers are accessible. It is planned to implement a large number of different features and especially higher level features such as shapes and evaluate them with respect to medical image retrieval.

Besides the small changes in the GIFT system to adapt it to medical image retrieval, it is planned to implement a larger number of possible features and evaluate the performance with respect to medical images. This can create a tool box to adapt the image retrieval system to certain specific needs in various application domains such as dermatology images or HRCTs of the lung.

REFERENCES

1. A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content–based image retrieval at the end of the early years," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22 No 12**, pp. 1349–1380, 2000.
2. M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker, "Query by Image and Video Content: The QBIC system," *IEEE Computer* **28**, pp. 23–32, September 1995.
3. J. R. Bach, C. Fuller, A. Gupta, A. Hampapur, B. Horowitz, R. Humphrey, R. Jain, and C.-F. Shu, "The Virage image search engine: An open framework for image management," in *Storage & Retrieval for Image and Video Databases IV*, I. K. Sethi and R. C. Jain, eds., *IS&T/SPIE Proceedings* **2670**, pp. 76–87, (San Jose, CA, USA), March 1996.
4. C. Carson, M. Thomas, S. Belongie, J. M. Hellerstein, and J. Malik, "Blobworld: A system for region–based image indexing and retrieval," in *Third International Conference On Visual Information Systems (VISUAL'99)*, D. P. Huijsmans and A. W. M. Smeulders, eds., *Lecture Notes in Computer Science*, pp. 509–516, Springer–Verlag, (Amsterdam, The Netherlands), June 2–4 1999.
5. A. Pentland, R. W. Picard, and S. Sclaroff, "Photobook: Tools for content–based manipulation of image databases," *International Journal of Computer Vision* **18**, pp. 233–254, June 1996.
6. A. Gupta and R. Jain, "Visual information retrieval," *Communications of the ACM* **40**, pp. 70–79, May 1997.
7. J. P. Eakins and M. E. Graham, "content–based image retrieval," Tech. Rep. JTAP–039, JISC Technology Application Program, Newcastle upon Tyne, 2000.
8. Y.-C. Chang, L. Bergmann, J. R. Smith, and C.-S. Li, "Query taxonomy of multimedia databases," in *Multimedia Storage and Archiving Systems IV (VV02)*, S. Panchanathan, S.-F. Chang, and C.-C. J. Kuo, eds., *SPIE Proceedings* **3846**, (Boston, Massachusetts, USA), September 20–22 1999. (SPIE Symposium on Voice, Video and Data Communications).
9. H. Qi and W. E. Snyder, "Content–based image retrieval in PACS," *Journal of Digital Imaging* **12**(2), pp. 81–83, 1999.
10. E. El-Kwae, H. Xu, and M. R. Kabuka, "Content–based retrieval in picture archiving and communication systems," *Journal of Digital Imaging* **13**(2), pp. 70–81, 2000.
11. A. Rosset, O. Ratib, A. Geissbuhler, and J.-P. Vallée, "Integration of a multimedia teaching and reference database in a PACS environment," *RadioGraphics* **22**(6), pp. 1567–1577, 2002.
12. G. Bucci, S. Cagnoni, and R. De Domicinis, "Integrating content–based retrieval in a medical image reference database," *Computerized Medical Imaging and Graphics* **20**(4), pp. 231–241, 1996.
13. H. D. Tagare, C. Jaffe, and J. Duncan, "Medical image databases: A content–based retrieval approach," *Journal of the American Medical Informatics Association* **4**(3), pp. 184–198, 1997.
14. H. J. Lowe, I. Antipov, W. Hersh, and C. Arnott Smith, "Towards knowledge–based retrieval of medical images. the role of semantic indexing, image content representation and knowledge–based retrieval," in *Proceedings of the Annual Symposium of the American Society for Medical Informatics (AMIA)*, pp. 882–886, (Nashville, TN, USA), October 1998.

15. W. W. Chu, A. F. Cárdenas, and R. K. Taira, "KMED: A knowledge–based multimedia distributed database system," *Information Systems* **19**(4), pp. 33–54, 1994.

16. D. Keysers, J. Dahmen, H. Ney, B. B. Wein, and T. M. Lehmann, "A statistical framework for model–based image retrieval in medical applications," *Journal of Electronic Imaging* **12**(1), pp. 59–68, 2003.

17. C.-R. Shyu, C. E. Brodley, A. C. Kak, A. Kosaka, A. M. Aisen, and L. S. Broderick, "ASSERT: A physician–in–the–loop content–based retrieval system for HRCT image databases," *Computer Vision and Image Understanding (special issue on content–based access for image and video libraries)* **75**, pp. 111–132, July/August 1999.

18. A. M. Aisen, L. S. Broderick, H. Winer-Muram, C. E. Brodley, A. C. Kak, C. Pavlopoulou, J. Dy, C.-R. Shyu, and A. Marchiori, "Automated storage and retrieval of thin–section CT images to assist diagnosis: System description and preliminary assessment," *Radiology* **228**, pp. 265–270, 2003.

19. P. Korn, N. Sidiropoulos, C. Faloutsos, E. Siegel, and Z. Protopapas, "Fast and effective retrieval of medical tumor shapes," *IEEE Transactions on Knowledge and Data Engineering* **10**(6), pp. 889–904, 1998.

20. Y. Liu, A. Lazar, W. E. Rothfus, M. Buzoiano, and T. Kanade, "Classification-driven feature space reduction for semantic–based neuroimage retrieval," in *Proceedings of the International Syposium on Information Retrieval and Exploration in Large Medical Image Collections (VISIM 2001)*, (Utrecht, The Netherlands), October 2001.

21. *Pathfinder: Region–based searching of Pathology Images using IRM*, (Los Angeles, CA, USA), November 2000.

22. M. Tsiknakis, D. Katehakis, and C. Orphanoudakis, Stelios, "Intelligent image management in a distributed PACS and telemedicine environment," *IEEE Communications Magazine* **34**(7), pp. 36–45, 1996.

23. C. Le Bozec, E. Zapletal, M.-C. Jaulent, D. Heudes, and P. Degoulet, "Towards content–based image retrieval in HIS–integrated PACS," in *Proceedings of the Annual Symposium of the American Society for Medical Informatics (AMIA)*, pp. 477–481, (Los Angeles, CA, USA), November 2000.

24. H. Müller, N. Michoux, D. Bandon, and A. Geissbuhler, "A review of content–based image retrieval systems in medicine – clinical benefits and future directions," *International Journal of Medical Informatics* , 2004 (accepted).

25. L. H. Y. Tang, R. Hanka, and H. H. S. Ip, "A review of intelligent content–based indexing and browsing of medical images," *Health Informatics Journal* **5**, pp. 40–49, 1999.

26. D. Harman, "Overview of the first Text REtrieval Conference (TREC–1)," in *Proceedings of the first Text REtrieval Conference (TREC–1)*, pp. 1–20, (Washington DC, USA), 1992.

27. D. M. Squire, H. Müller, W. Müller, S. Marchand-Maillet, and T. Pun, "Design and evaluation of a content–based image retrieval system," in *Design & Management of Multimedia Information Systems: Opportunities & Challenges*,[42] ch. 7, pp. 125–151.

28. C. Jörgensen and P. Jörgensen, "Testing a vocabulary for image indexing and ground truthing," in *Internet Imaging III*, G. Beretta and R. Schettini, eds., *SPIE Proceedings* **4672**, pp. 207–215, (San Jose, California, USA), January 21–22 2002. (SPIE Photonics West Conference).

29. K. Sparck Jones and C. van Rijsbergen, "Report on the need for and provision of an ideal information retrieval test collection," British Library Research and Development Report 5266, Computer Laboratory, University of Cambridge, 1975.

30. J. Zobel, "How reliable are the results of large–scale information retrieval experiments?," in *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, and J. Zobel, eds., pp. 307–314, ACM Press, New York, (Melbourne, Australia), August 1998.

31. H. Müller, W. Müller, D. M. Squire, S. Marchand-Maillet, and T. Pun, "Performance evaluation in content–based image retrieval: Overview and proposals," *Pattern Recognition Letters* **22**, pp. 593–601, April 2001.

32. N. J. Gunther and G. Beretta, "A benchmark for image retrieval using distributed systems over the internet: BIRDS–I," tech. rep., HP Labs, Palo Alto, Technical Report HPL–2000–162, San Jose, 2001.

33. P. S. Salembier and B. S. Manjunath, "Audiovisual content description and retrieval: Tools and MPEG–7 standardization techniques," in *IEEE International Conference on Image Processing (ICIP 2000)*, (Vancouver, BC, Canada), December 2000.

34. Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra, "Relevance feedback: A power tool for interactive content–based image retrieval," *IEEE Transactions on Circuits and Systems for Video Technology* **8**, pp. 644–655, September 1998. (Special Issue on Segmentation, Description, and Retrieval of Video Content).

35. D. M. Squire, W. Müller, H. Müller, and T. Pun, "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), pp. 1193–1198, 2000. B.K. Ersboll, P. Johansen, Eds.

36. S. Chen and L. Li, "Build a content based image retrieval system of endoscopic image," in *Proceedings of the Medical Informatics Symposium Taiwan (MIST 2003)*, (Taipei, Taiwan), September 2003.

37. S. Lim and G. Lu, "Effectiveness and efficiency for six color spaces for content–based image retrievl," in *Proceedings of the Internation Workshop on Content–Based Multimedia Indexing (CBMI 2003)*, pp. 125–222, (Rennes, France), September 2003.

38. A. Vellaikal and C.-C. J. Kuo, "content–based image retrieval using multiresolution histogram representation," in *Digital Image Storage and Archiving Systems*, C.-C. J. Kuo, ed., *SPIE Proceedings* **2606**, pp. 312–323, (Philadelphia, PA, USA), October 1995.

39. J. Nielsen, *Usability Engineering*, Academic Press, Boston, MA, 1993.

40. H. Müller, D. M. Squire, W. Müller, and T. Pun, "Efficient access methods for content–based image retrieval with inverted files," in *Multimedia Storage and Archiving Systems IV (VV02)*, S. Panchanathan, S.-F. Chang, and C.-C. J. Kuo, eds., *SPIE Proceedings* **3846**, pp. 461–472, (Boston, Massachusetts, USA), September 20–22 1999.

41. F. Qian, M. Li, H.-J. Zhang, W.-Y. Ma, and B. Zhang, "Alternating feature spaces in relevance feedback," *International Journal on Multimedia Tools and Applications* **21**, pp. 35–54, 2003. (Special Issue on Multimedia Information Retrieval).

42. S. M. Rahman, *Design & Management of Multimedia Information Systems: Opportunities & Challenges*, Idea Group Publishing, London, 2001.