A Probabilistic Approach to Medical Image Retrieval

Koen Lubbers[1], Arjen P. de Vries[2], Theo Huibers[1], and Paul van der Vet[1]

[1] University of Twente, Enschede, The Netherlands
{k.f.lubbers, t.w.c.huibers, p.e.vandervet}@ewi.utwente.nl
[2] Centrum voor Wiskunde en Informatica (CWI), Amsterdam, The Netherlands
arjen@acm.org

Abstract. We present a probabilistic approach to the medical retrieval task. We experimented with the Westerveld method [1] to obtain our results for ImageCLEF. In addition to these results we describe our findings of involving a medical expert in our research. The expert helped us identifying useful image retrieval applications and reflected upon the setup of ImageCLEF's medical task. Finally we describe the evaluation of an interactive implementation of the probabilistic approach.

1 Introduction

The amount of information available through all kinds of sources is growing larger and larger. The goal of information retrieval systems is to help a user in efficiently finding relevant information. Image retrieval is a sub domain of information retrieval. This relatively new research area is about gaining access to images that match a query. Apart from text, such a query can consist of a sketch or an actual image.

Several information retrieval techniques have been applied to the image retrieval field lately [2]. Although probabilistic methods are often used to determine the relevance of textual documents, they have hardly been applied to image retrieval tasks. The goal of our work is to explore the possibilities of the probabilistic Westerveld method [3, 1].

In recent years, much research has been done into specific medical image retrieval systems [4, 5, 6, 7, 8]. For comparison reasons, we have chosen to test the *generic* Westerveld method in a medical environment. A part of testing a method is to compare it to other (specific) systems. Until recently, a fair comparison of content-based image retrieval methods under similar circumstances was lacking [9]. The ImageCLEF medical retrieval task [10] is an evaluation that tries to change this. We have participated in CLEF to experiment with a medical image collection and to be able to compare our results with other systems.

Few studies are known in which medical experts have participated in the evaluation of medical retrieval systems [11]. Therefore, in addition to our participation in CLEF, we have involved a medical physicist from the Academic Medical Centre (AMC) in Amsterdam in our research. We have asked the expert

C. Peters et al. (Eds.): CLEF 2004, LNCS 3491, pp. 761–772, 2005.

to identify useful applications of image retrieval techniques within the medical domain, and to reflect upon the setup of ImageCLEF's medical search task.

1.1 Image Retrieval in a Medical Environment

Researchers from the University of Berkeley estimate that about 2 billion X-rays are produced in hospitals worldwide each year [12] (this corresponds to approximately 5.5 million new medical images every day!). A growing number of hospitals is switching to handling their image data in digital format. Current Picture Archiving and Communication Systems (PACS) offer the possibility to save images with additional relevant information, like a patient name or number, and additional information from a medical case. Subsequently, all this data will be available from the different workstations throughout the hospital.

To identify useful applications of image retrieval systems, we first looked at the present situation with the PACS in the AMC. When images are produced they will be stored automatically with information like patient name, number, body region, and modality as metadata. This metadata is available because of the electronic request a doctor has to submit before the image is produced. This means that searching by body part or modality with a content-based retrieval method will often not be useful, because most of the time the correct modality and body part are available in text.

However, an image retrieval system could serve as a control tool. People do make mistakes, and images could, for example, end up at the wrong patient or a doctor who produces an image of the left knee is actually supposed to deliver an image of the right knee. Furthermore, error rates with respect to automatically stored anatomical regions seem to be very high: about 15 to 20% [13]. This is where a retrieval system could be convenient: on a basis of already classified images it can determine how much the new image differs from the expected visual features.

An important finding in this study is that the PACS used at AMC does not associate images and pathology. When a medical doctor wants to look at images with the same or similar pathology, for example for comparison to the image shown on his screen, no suitable solution exists. The AMC medical experts therefore indicated three particularly useful fields for application of image retrieval tools: education, research and diagnosis.

For educational purposes, a medical doctor would like to find images in a corresponding field of pathology. These images could serve as cases for medical students. In the research area, image retrieval could be used to analyze the visual features of clusters of images with corresponding syndromes. This could result in a thesaurus of visual features connected to different kinds of images and syndromes. The third application is the diagnosis of problematic cases. When a medical doctor is not sure about a certain image, he would like to be able to use a retrieval method to find other images of the same kind. In this way, he will find useful information in the cases connected to the retrieved images.

Apart from identifying useful applications, image retrieval research in a medical environment shows medical experts a way in which technology can support

their daily activities. Medical doctors do not always believe in the abilities of computer systems to offer added value to their work. By involving them in image retrieval research, the technological frontiers of the medical sector are explored.

2 Background

The Westerveld image retrieval approach [3, 1] has not been designed for specific images. It has been tested mainly on collections with a large variety in images. Westerveld, following Vasconcelos [14], models the visual features by using Gaussian Mixture Models (GMMs). The basic idea is that an image consists of a certain number of 'aspects', where each of these aspects can be described in one component of the GMM. Each sample that is taken from an image is assumed to have been generated by one of these components. A Gaussian Mixture Model (GMM) is a weighted sum of multivariate Gaussian distributions, where the weights are considered as prior probabilities of the different components. We will explain briefly what happens when the parameters for a GMM are estimated. For a more detailed explanation of the generative probabilistic retrieval model the reader is referred to [3, 1].

The steps of creating a probabilistic image model are shown in Figure 1. First, the RGB representation of the image is converted into YCbCr colour space. Next, each of the colour channels of the image is divided into samples of 8 by 8 pixels. Then, a discrete cosine transform (DCT) is performed on every sample. By default, the different samples are described by 14-dimensional vectors. Each vector consists of the first 10 DCT coefficients from the Y channel, the DC coefficient of both the Cb and the Cr channel, and the x and y position of the sample in the image.

The feature vectors of an image are fed to the EM algorithm to find the parameters of the mixture models. The algorithm starts with introducing a given number of components by grouping the samples randomly. This is the first expectation step. In the maximization step, the parameters of each component are calculated, based on the samples assigned to that component. A component represents the average colour and texture of the samples assigned to it. In the second expectation step, the samples are regrouped. For example: a sample of a blue sky will be assigned to the component that explains best the visual characteristics of the blue sky. The E-step and the M-step iterate until the algorithm converges.

A collection of images can be indexed by estimating the GMM for each of the images. Query images are represented as a collection of samples. The basis of the retrieval step is to estimate, for each model of the collection images, the probability that the query samples could be observed given that collection image model. In other words, the goal is to find the document that is most likely to have produced a certain query. The joint probability of a document producing this certain query is calculated by multiplying the probabilities for each individual sample of the query.

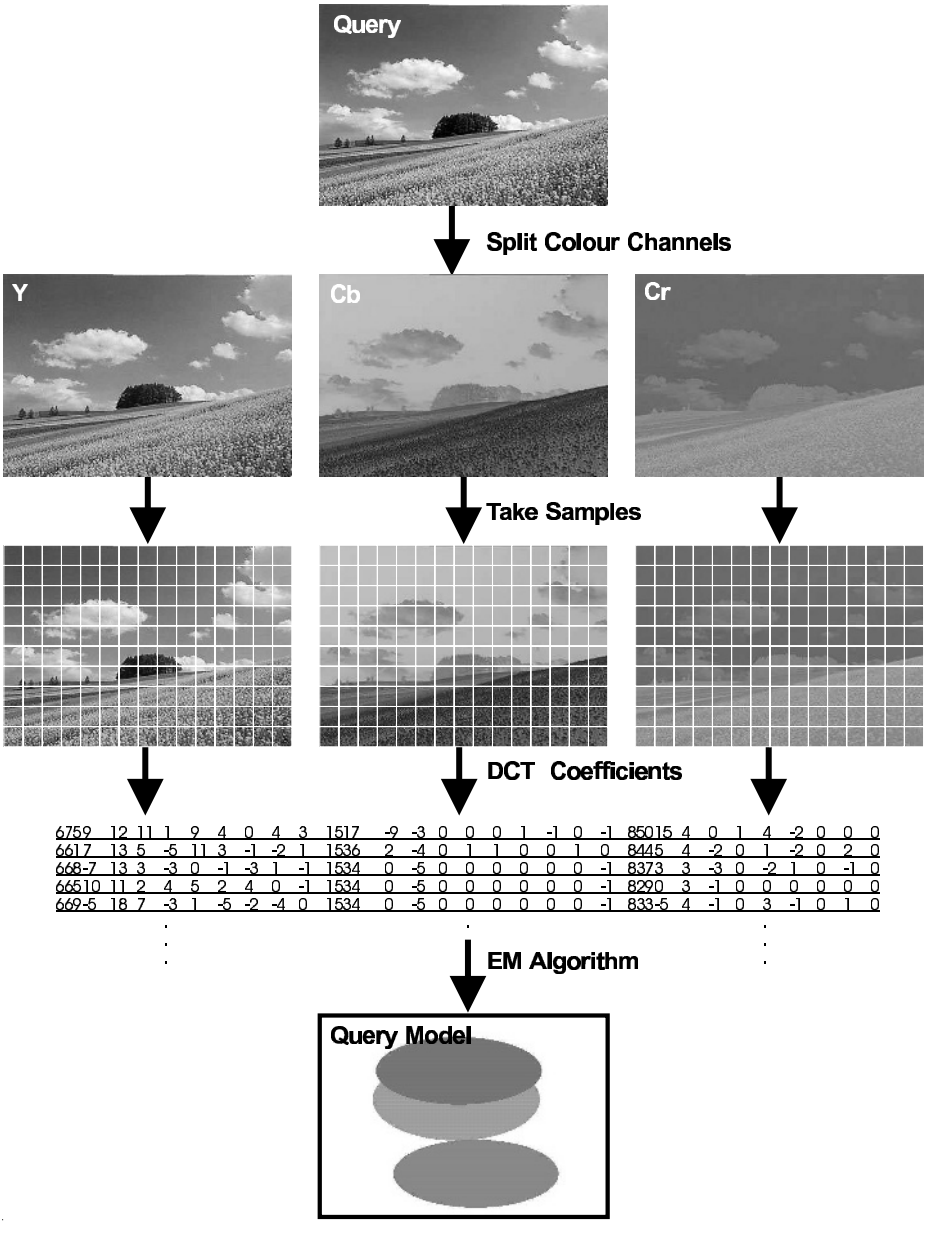

Fig. 1. Building a Gaussian Mixture Model from an image [15]

3 Experimental Setup

The main research question in our ImageCLEF experiments is how a generic image retrieval system would perform on a domain-specific retrieval problem.We

decided to ignore the textual information in the medical cases, to provide a solid basis to judge the possible merits of content-based retrieval techniques for search in medical image archives. The combination with textual information is postponed to future research.

Table 1. Standard settings of the Westerveld image retrieval method

Parameter	Default	Description
blocksize	8	size of the samples in pixels
C	8	number of mixture components
convert	1	binary, convert image from RGB to YCbCr colour space
imagesize	240x352	size to which an image is scaled before samples are taken
ncoeffcbcr	1	number of DCT coefficients from Cb and Cr channel
ncoeffy	10	number of DCT coefficients from Y channel
overlap	0	samples will overlap or not
Scale	1	image is scaled before samples are taken or not
XYpos	1	x and y position of a sample are used in feature vector

The default values of the method (see Table 1) are the point of departure of testing with different parameters. During the process of testing with different parameter settings, we varied one parameter at a time. We have tested with both values for each of the binary parameters. The basic rule for adjusting the other values is that we will never reduce the information represented below the default settings.

First, we indexed a sub-collection of the medical CLEF collection to find out which parameters would qualify to be used to get the results for the submission. The selected settings from this experiment were used to build eight different indices of the whole medical collection. We then chose the four best indices by ranking all retrieval results with all queries, based on an 'educated guess' of the precision at a document cut-off level of 20 (doing manual assessments ourselves). We distinguished precision A and precision B. The first value is based on an image being relevant or not according to the CLEF task (image being relevant on both body part and modality) and the second one is only based on the modalities of the images. A modality describes the way in which medical images are produced: MRI, CT, etc.

After the submission of the runs, we have performed more experiments with the system. Several new experiments indicated that the conversion to YCbCr affected the performance of the system negatively. These new experiments were performed with a new sub-collection, which consisted of ten relevant images per query. The relevant images were manually selected from the medical CLEF collection with the help of the medical expert from the AMC.

Because we knew the number of relevant images for each query in de sub collection, we were able to follow Kraaij [16] and compare the retrieval results with R-recall. This means that recall is measured at a document cut-off level, which equals the number of relevant images for a certain query.

Because of the new findings with the second sub-collection, we indexed the whole collection with parameter convert=0 in order to create a new run. Furthermore, we used the setting without conversion as a new basic state and started varying the other parameters to find another way to improve retrieval results.

4 Analysis

The results of the experiment we used to select four out of eight runs for submission are shown in Table 2.

We submitted the first four runs. Since new experiments showed that results were far better when conversion was not applied, we did not expect very good results from the official medical evaluation. After indexing the medical CLEF collection without conversion, retrieval with the queries proved that results with the whole collection were indeed far better: the average precision A equals 0.47.

Table 2. Qualifying runs for submission

Rank	Parameter	Avg precision A	Avg precision B	Avg rank
1	ncoeffy=20	0.22	0.57	3.8
2	default	0.20	0.58	3.8
3	c=16	0.24	0.56	4.0
4	c=4	0.20	0.55	4.1
5	XYpos=0	0.18	0.55	4.9
6	ncoeffcbcr=2	0.18	0.54	5.0
7	imagesize=300x440	0.17	0.50	6.5
8	overlap=1	0.18	0.46	6.6

Further experiments with the second sub-collection showed that there were no parameter settings that improved the retrieval results of the new basic state with convert=0. We concluded that the best way to use the current version of the Westerveld method with the medical CLEF collection is with only one adjustment: disable the conversion to the YCbCr colour space.

We found that R-recall in the experiments with the second sub collection varied from 0.41 to 0.48. We got these results by testing with the fixed settings convert=0, while varying the other parameters one by one. After the release of the judgements from the CLEF medical task (the so-called qrels), we were able to calculate R-recall values for the results we found after retrieval with the total medical image collection. The average R-recall value over the 26 queries equals 0.29. This means that our sub-collection may have been a more ideal test environment than the whole CLEF collection, but it can also imply that we evaluated the results less strictly than the CLEF assessors did.

The official results are expressed in Mean Average Precision (MAP). The best result from the runs we submitted has a MAP of 0.1069. The use of the new parameter settings showed the improvement we expected: the Westerveld method performs about twice as good when the colour space is not converted. Using the RGB representation of the images, the systems scores a MAP of 0.2359, which is a satisfying initial retrieval result.

4.1 Conversion of Colour Spaces

Based on our experience with the retrieval model on other image retrieval tasks, we expected that indexing the collection without conversion to the YCbCr colour space would have given inferior results. The results after the submission of the runs however, showed that without conversion the retrieval method performed about twice as good. This finding proved to be reproducible.

Since earlier testing with the Westerveld method turned out that better results were obtained when working with YCbCr colour space, the following question remains: why does conversion perform less well with the medical collection? We have not yet found a perfect explanation for the degraded retrieval effectiveness after conversion to YCbCr colour space. We believe that the cause of the observed change in performance is to be found in the difference between the medical collection and the previously used testing collections: the medical collection consists almost completely of greyscale images.

In colour images, the three channels in RGB all contain information on both intensity and colour, so the different dimensions are correlated. The motivation for conversion is that in YCbCr colour space, the intensity channel (Y) is separated from the colour channels (Cb and Cr), and the information in each channel is independent from the information in the other channels. In a greyscale situation however, there is no colour information, and the three channels represent the same amount of intensity: R=G=B. Given a greyscale image, Y will be created as usual, but the Cb and the Cr channel both equal 128 in every possible greyscale situation.

Now, recall that the feature vectors to represent the image samples are computed from the DCT transformation over 8x8 pixel blocks. In the feature vectors for an RGB image, the first DCT coefficient (corresponding to the average intensity in the pixel block) is represented in three dimensions. In the YCbCr case, this information is only represented in one dimension. Theoretically, because we assume a diagonal covariance matrix, the complete correlation between the three dimensions in the RGB case (those corresponding to the first DCT coefficient of the three (identical) colour channels) should however affect retrieval negatively rather than improve its results. Yet, the experiments proof otherwise.

Our current intuition is that the duplicated information separates, in feature space, the intensity information more than the textural information (which is represented in the higher coefficients of the DCT transformation). This 'encourages' the EM algorithm model during training to prefer textural information over the intensity information in the image samples. For medical images, textural information seems more important than the intensity information, so this could explain the improved effectiveness of the model. This hypothesis is further supported by observations in earlier experiments (on TRECVID data) [17], where we demonstrated that the textural information in images was dominated by colour information (on YCbCr colour space). Further research is however needed to (in)validate this explanation of the experimental results.

5 Interactive Experiments

After identifying useful applications of medical image retrieval systems, we applied the probabilistic approach in an interactive retrieval system. This system tries to learn from the relevance feedback given by the user [18], attempting to reduce the semantic gap by inserting a human 'in-the-loop'. More information about this research activity can be found in [19]. In order to realise a suitable system, we had to shorten the retrieval time and make the method user-friendly. Again, since we want to learn the strengths of the content-based image retrieval method, we did not use the text in the medical case descriptions. Note that Smeulders describes two other ways to deal with semantics: interpretation and similarity between features [9].

After a medical doctor of the AMC uploads a query image, the system estimates the parameters of its GMM. It then compares the query model to the GMMs of the images in the CLEF collection and presents an initial retrieval result. For efficiency reasons, an approximation of the Kullback Leibler distance between the image models is used as an alternative to the likelihood of observing the query image samples. The results obtained are very similar to those of the original system. After this initial retrieval step, the medical doctor marks retrieved images as relevant or irrelevant; the next iteration takes the feedback into account to re-rank the remaining images.

The interactive system turned out to be very intuitive and easy to use, partially because the doctors in the AMC are already used to a web-based interface for accessing the PACS system. After a query has been uploaded the system is sufficiently fast in presenting the retrieval results. Within a minute, a medical doctor can go through about five iterations. Figure 2 shows a screenshot of the interactive retrieval system (it shows the results after uploading topic 24 of the medical CLEF collection). When a query has been posted the results are displayed within a second.

The interactive experiment pointed out two possible improvements for our retrieval system. First, although the medical CLEF collection is representative for the type of images encountered in the AMC, two main differences are observed in relation to the background and the greyscale representation of the images. When we save an AMC image as JPEG and make it anonymous, all greyscale images are represented as greyscale instead of RGB. Of course, only a minor modification fixes this. A more significant difference is that the AMC data consist for a large part of the image of black background only. The subjects within the images of the CLEF collection seem to have been cropped cleverly.

Finally, explaining the search task applied at the ImageCLEF medical retrieval task to the medical expert has raised some issues with the task evaluated at this first medical image CLEF evaluation, and also demonstrated clearly the existence of 'the semantic gap'. From the system point of view, the results did not look bad, and any mistakes could be easily explained from its inner workings. The system performs well at retrieving images with the same kind of visual features, which often means the same modality. However, medical doctors are interested in finding images with corresponding syndromes, or at least corre-

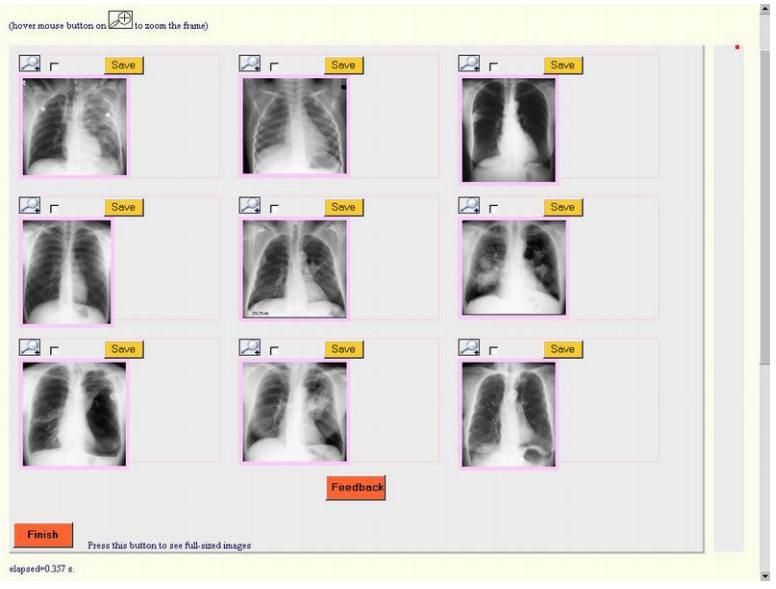

Fig. 2. Screenshot of the initial retrieval results after searching with query 24

sponding body parts. It is far more interesting to retrieve a CT of the brain with an MRI of the brain as a query, than to find an abdomen MRI with it. It may be more useful to measure the performance of retrieval systems using body part only (as opposed to the performance on modality and body part).

6 Conclusions and Future Work

The main goal of our research was to investigate if a generic image retrieval model could also be applied to a domain-specific task such as the retrieval of medical images. We have tested the probabilistic image retrieval model developed by Westerveld using the CLEF medical image test collection, which allows the objective comparison of different approaches to the retrieval problem. We also evaluated an interactive version of our system with a medical expert from the AMC.

The best performance of the Westerveld method has been obtained after adjusting one of the parameters in the representation of the image data. When the medical images are not converted from RGB to YCbCr colour space, the Mean Average Precision in our runs equals 0.2359. This is a satisfying result, especially when considering that we have not used the text of the medical cases in our system.

It is essential that medical doctors - the future users of image retrieval systems - are involved in image retrieval research. With the help from the AMC we identified a number of useful medical retrieval applications. Evaluating the

CLEF images with a medical expert showed that the collection seems to be a rather ideal representation of the images present in the hospital. Furthermore, an experiment with the probabilistic Westerveld method indicated the semantic gap. Retrieval results are most likely to be useful when a system can deal with this gap.

Since we neglected text in our approach, we tried to apply the retrieval method in an interactive system. This system proved to be easy to use and to work fast. However, it still needs to learn from the relevance feedback of experts. Improvements of the Westerveld method itself and allowing the interactive system to learn from medical doctors can lead to adequate support of the daily activities in medical practise.

The AMC image collection showed that an image retrieval method needs to be able to work with greyscale images. Furthermore, it seemed that images from this hospital contained a large black background. An experiment with the smoothing function of the Westerveld can show if the system can automatically neglect this background.

To obtain better retrieval results, we have to deal with the semantic gap. The interactive system will only improve when real users give relevance feedback to initial results. Further research should point out if the system is really able to learn from experience.

Another way to deal with semantics is to embrace a text retrieval method. The Westerveld method has already been tested in combination with a probabilistic text retrieval approach [3].

During a next medical retrieval task it may be possible to increase the performance of retrieval systems through interpretation and similarity between features. The clusters of relevant images per query offer the possibility to create a sort of medical thesaurus, which consists of visual features of certain modalities, body parts, or even syndromes.

Evaluation with the AMC showed that searching for images with identical modality and body part is not a useful task for image retrieval systems. Medical doctors will be interested in a certain pathology: they want to find images with corresponding syndromes. It would be useful if the next medical CLEF collection contained a number of sub-collections. A sub collection can, for example, contain images with corresponding body parts. A challenge for image retrieval systems is to distinguish the visual features of images that do contain a certain abnormality, and images that do not.

Finally, we would like to add another challenge for image retrieval research. The basis of an image retrieval method is a certain image collection that can be indexed. However, when a medical doctor wants to use an application to search for clues regarding the diagnosis of his query image, he might not find satisfying results in the image collection at his own hospital. Retrieval systems can really add value when experts from several hospitals can learn from each others experience. This implies the need for a standard way of indexing and searching. Such a standard can only be reached when different research groups meet to evaluate their results together. This shows the importance of evaluations like ImageCLEF in the future.

Acknowledgements

Without the help of Thijs Westerveld we would not have been able to participate and test with a probabilistic approach to image retrieval. We would also like to thank Lioudmila Boldareva for making it possible to combine the used method with the interactive relevance feedback system. Finally, we thank Jan Habraken, who helped us evaluating our image retrieval work from a medical point of view.

References

1. Westerveld, T.: Using generative probabilistic models for multimedia retrieval. PhD thesis, University of Twente, CTIT Ph.D.-thesis series, ISSN 1381-3617; No. 04-67, Enschede, The Netherlands (2004)
2. Squire, D.M., Müller, W., Müller, H., Pun, T.: Content-based query of image databases: inspirations from text retrieval. Pattern Recognition Letters **21** (2000) 1193–1198
3. Westerveld, T., de Vries, A., van Ballegooij, A., de Jong, F., Hiemstra, D.: A probabilistic multimedia retrieval model and its evaluation. EURASIP Journal on Applied Signal Processing **2003** (2003) 186–198
4. Korn, P., Sidiropoulos, N., Faloutsos, C., Siegel, E., Protopapas, Z.: Fast and effective retrieval of medical tumor shapes. IEEE Transactions on Knowledge and Data Engineering **10** (1998) 889–904
5. Lehmann, T., Guld, M., Thies, C., Fischer, B., Spitzer, K., Keysers, D., Ney, H., Kohnen, M., Schubert, H., Wein, B.: Content-based image retrieval in medical applications. Methods of Information in Medicine **43** (2004) 354–361 In press.
6. Mattiea, M., Staib, L., Stratmann, E., Tagare, H., Duncan, J., Miller, P.: Pathmaster: Content-based cell image retrieval using automated feature extraction. Journal of the American Medical Informatics Association **7** (2000) 404–415
7. Müller, H., Rosset, A., Vallée, J., Geissbuhler, A.: Comparing feature sets for content-based medical information retrieval. In: SPIE Medical Imaging, San Diego, CA, USA (2004)
8. Shyu, C., Brodley, C., Kak, A., Kosaka, A., Aisen, A., Broderick, L.: Assert – a physician-in-the-loop content-based retrieval system for hrct image databases. Computer Vision and Image Understanding **75** (1999) 111–132
9. Smeulders, A., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval: the end of the early years. IEEE transactions on Pattern Analysis and Machine Intelligence **22** (2000) 1349–1380 invited review.
10. Clough, P., Müller, H., Sanderson, M.: The CLEF Cross Language Image Retrieval Track (ImageCLEF) 2004. In Peters, C., Clough, P., Gonzalo, J., Jones, G., Kluck, M., Magnini, B., eds.: Fifth Workshop of the Cross–Language Evaluation Forum (CLEF 2004), Lecture Notes in Computer Science (LNCS), Springer, Heidelberg, Germany (in print) (2005)
11. Müller, H., Michoux, N., Bandon, D., Geissbuhler, A.: A review of content-based image retrieval systems in medicine - clinical benefits and future directions. International Journal of Medical Informatics **73** (2004) 1–23
12. Lyman, P., Varian, H.R.: "how much information", 2003 (2003) Retrieved from http://www.sims.berkeley.edu/how-much-info-2003 on 1/11/2004.

13. Güld, M., Kohnen, M., Keysers, D., Schubert, H., Wein, B., Bredno, J., Lehmann, T.: Quality of dicom header information for image categorization. In: Proceedings SPIE. Volume 4685. (2002) 280–287
14. Vasconcelos, N.: Bayesian Models for Visual Information Retrieval. PhD thesis, Massachusetts Institut of Technology (2000)
15. Westerveld, T., de Vries, A.: Experimental evaluation of a generative probabilistic image retrieval model on 'easy' data. In: Proceedings of the Multimedia Information Retrieval Workshop 2003. (2003) in conjunction with the 26th annual ACM SIGIR conference on Information Retrieval.
16. Kraaij, W.: Variations on language modeling for information retrieval. PhD thesis, CTIT PhD thesis series No. 04-62, Neslia Paniculata (2004)
17. Westerveld, T., de Vries, A.: Experimental result analysis for a generative probabilistic image retrieval model. In: Proceedings of the 26th ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'03). (2003) 135–142
18. Boldareva, L.: Improving objects similarities with relevance judgements from the searchers. In: 27th European Conference on Information Retrieval (ECIR '05), Santiago de Compostela, Spain (2005) To appear (poster).
19. Lubbers, K.: Image retrieval in de medische praktijk: mogelijkheden van een probabilistische aanpak. Master's thesis, University of Twente (2004) In Dutch.