

Medical Image Retrieval Using Texture, Locality And Colour

Peter Howarth, Alexei Yavlinsky, Daniel Heesch, and Stefan Ruger

Multimedia Information Retrieval, Imperial College London, UK.

<http://km.doc.ic.ac.uk>

{[peter.howarth](mailto:peter.howarth@imperial.ac.uk), [alexei.yavlinsky](mailto:alexei.yavlinsky@imperial.ac.uk), [daniel.heesch](mailto:daniel.heesch@imperial.ac.uk),
[s.rueger](mailto:s.rueger@imperial.ac.uk)} @imperial.ac.uk

Abstract. We describe our experiments for the Image CLEF medical retrieval task. Our efforts were focused on the initial visual search. A content based approach was followed. We used texture, localisation and colour features that have been proven by previous experiments. The images in the collection had specific characteristics. Medical images have a formulaic composition for each modality and anatomic region. We were able to choose features that would perform well in this domain. Tiling a Gabor texture feature to add localisation information proved to be particularly effective. The distances from each feature were combined with equal weighting. This smoothed the performance across the queries. The retrieval results showed that this simple approach was successful, with our system coming third in the evaluation.

1 Introduction

Content based image retrieval (CBIR) aims to provide a way to search generic image collections. Traditionally, for highly constrained domains, such as medical images, CBIR has been viewed as being too imprecise. Our aim was to determine if the CBIR approach could be viable for an initial search or filtering step in a medical image collection. We focused on choosing high quality visual features that have good discriminatory power for the collection.

In this paper we first present a brief overview of our system. Section 3 explains the rationale for using specific visual features and details how they are computed. Results of our run are presented in Section 4, followed by a postmortem analysis. Some of our ideas for future work are presented in Section 5. We would like to have applied classification methods to the collection but the lack of training data precluded this. We also discuss the use of a browsing paradigm for this type of collection. Our conclusions round off the paper.

2 System Overview

The initial visual search task was very straightforward, with 26 single image queries, it is described in [1]. With no training data it was not possible to use

any learning classifiers. We therefore used a simple system to tackle the retrieval task, using the features described in the next section. The following steps were carried out:

1. Features were generated for the test collection and query images;
2. For each feature the Manhattan distance between the query images and the test set was calculated;
3. The set of distances from each feature was normalised by dividing by their median. This ensured that each feature would have an equal weighting;
4. The distances for each query were summed over all features. This gave the overall distance from each query image to the test set. These were then sorted to produce a ranked list of retrieval results.

3 Features

The initial step in our work was to look at the collection and determine its characteristics. As a relatively specific domain it displayed a large degree of homogeneity. We realised this could be exploited by choosing features that would differentiate the image types.

The collection contained a large number of monochrome images, such as x-rays and CT scans, with very specific layout. The patients are positioned very precisely to show the area under investigation at the centre of the image. The layout can be used to indicate both modality and anatomic region. For this reason a localisation feature, thumbnail, was used to detect images with similar layouts. Within the modalities the images could be discriminated by structure and texture. We therefore chose to use a convolution feature to discriminate structure and two texture features, co-occurrence matrices and Gabor filters. The two texture features were applied to non-overlapping image tiles. This adds some locality discrimination to the feature. Finally, for the relatively small number of colour images we deployed a colour structure descriptor.

3.1 Thumbnail

This is perhaps the simplest feature in our feature set, yet it is highly effective in detecting images with a near identical layout. Each image is converted to grey scale and then scaled down to a thumbnail of fixed size. For these experiments we used 40×30 pixels. The pixel values of this new image then make up the feature vector.

3.2 Convolution

This feature is based on Tieu and Viola's method [2]. It relies on a large number of highly selective features that can determine structure within an image and capture information about texture and edges. A vast set of features are defined such that each feature will have a high value for only a small proportion of

images. This enables an effective search by matching the features that are defined by the query. Due to the nature of the image collection we applied the feature to grey level images rather than RGB.

The feature generation process starts with a set of 25 primitive features (eg, edge detectors) that are applied to the grey level image. This generates 25 feature maps. Each of these is rectified and down-sampled before being filtered again by each of the 25 primitive filters. This gives 625 feature maps. The second stage of the process discovers arrangements of features in the previous levels. The values of each feature map are summed to give a single number. These are combined into a feature vector of 625 values.

3.3 Co-occurrence

Haralick [3] suggested the use of grey level co-occurrence matrices (GLCM) to extract second order statistics from an image. They have been used very successfully for texture classification. The GLCM of an image is defined as a matrix of frequencies at which two pixels, separated by a certain vector, occur in the image. The distribution in the matrix will depend on the angular and distance relationship between pixels. Varying the vector used allows the capturing of different texture characteristics. Once the GLCM has been created, various features can be computed from it. These have been classified into four groups: visual texture characteristics, statistics, information theory and information measures of correlation [3, 4].

Using the results of our recent evaluation [5] we chose the following configuration for creating the GLCM:

- The original image was split into 7×7 non-overlapping tiles and the feature run for each of these;
- The colour image was quantised into 64 grey levels;
- 16 GLCMs were created for each image tile using vectors of length 1, 2, 3, and 4 pixels and orientations 0 , $\pi/4$, $\pi/2$ and $3\pi/4$;
- For each normalised co-occurrence matrix $P(i, j)$ we calculated a homogeneity feature H_p ,

$$H_p = \sum_i \sum_j \frac{P(i, j)}{1 + |i - j|} .$$

This feature was chosen as it had performed consistently well in previous evaluations.

3.4 Gabor

One of the most popular signal processing based approaches for texture feature extraction has been the use of Gabor filters. These enable filtering in the frequency and spatial domain. It has been proposed that Gabor filters can be used to model the responses of the human visual system. Turner [6] first implemented

this by using a bank of Gabor filters to analyse texture. A range of filters at different scales and orientations allows multichannel filtering of an image to extract frequency and orientation information. This can then be used to decompose the image into texture features.

Our implementation is based on that of Manjunath et al [7]. The feature is built by filtering the image with a bank of orientation and scale sensitive filters and computing the mean and standard deviation of the output in the frequency domain.

Filtering an image $I(x, y)$ with Gabor filters g_{mn} designed according to [7] results in its Gabor wavelet transform W_{mn} ,

$$W_{mn}(x, y) = \int I(x_1, y_1)g_{mn}^*(x - x_1, y - y_1)dx_1dy_1$$

The mean and standard deviation of the magnitude $|W_{mn}|$ are used for the feature vector. The outputs of filters at different scales have different ranges. For this reason each element of the feature vector is normalised using the standard deviation of that element across the entire database.

From our evaluation [5] we found that a filter bank with 2 scales and 4 orientations gave the best retrieval performance. We used this configuration and applied it to 7×7 non-overlapping tiles created from the original image.

3.5 Colour Structure Descriptor HDS-S

For the colour images in the collection we used a feature that is good at capturing local colour image structure. It is defined in the HMMD (hue, min, max diff) colour space. This is used in the MPEG-7 standard and is derived from both RGB and HSV spaces. The hue component is taken from HSV and the min and max components are from the maximum and minimum values in the RGB space. The diff component is the difference between min and max. We follow the MPEG-7 standard and quantise this space non-uniformly into 184 bins in the 3 dimensional hue, diff and sum (HDS) colour space, see Manjunath and Ohm [8] for details of the quantisation.

To calculate the colour structure descriptor an 8×8 window is slid over the image. Each of the 184 bins of the HDS histogram contains the number of window positions for which there is at least one pixel falling into the bin under consideration. This feature, which we call HDS-S, is capable of discriminating between images with the same global colour distribution but differing local colour structures. For this reason it is suited to colour medical images which tend to have similar overall colour but differing structure depending on the detail of the photograph.

4 Results

Fig 1 shows the precision recall graph for our run. Our system achieved 37.8% mean average precision (m.a.p.) retrieval across all queries. This put us in third

place for the automatic retrieval task. The best performance was achieved by Buffalo [9] with a m.a.p. of 39.0%. The median was 28.8%, with 34 runs submitted. All the runs primarily used visual features. In addition 11 used text and 12 query expansion. A summary of all the results is given in [1]. The performance of our system shows that our simple approach, using good quality visual features, produced results comparable with the top systems.

4.1 Analysis of results

With the relevance judgements available it was possible to look at how individual features had performed. Fig. 1 shows the precision recall graph for the individual features together with that for the combined features of the submitted run. Table 1 shows the mean average precisions for the same features.

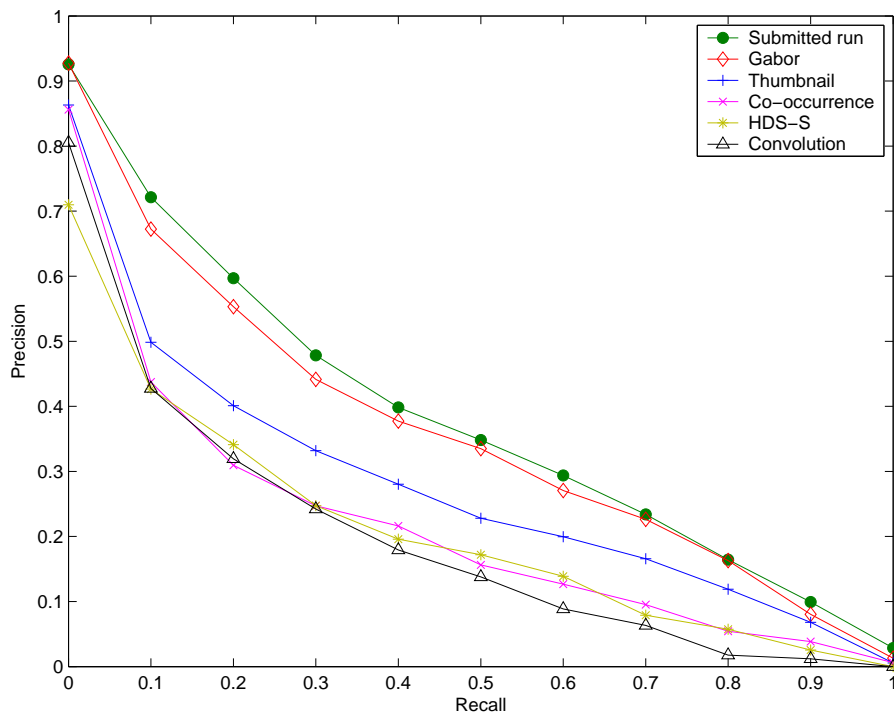


Fig. 1. Precision-recall graph for combined and individual features

From these results it is clear that all features performed reasonably well. Considering individual features, Gabor performed best, with thumbnail a clear second and the remaining 3 closely grouped. Some additional feature combinations were tested, including adding the Gabor feature to each of the others in

Table 1. Mean average precision for combined and individual features

Feature	Mean average precision
Submitted run (combined)	37.8%
Gabor	35.3%
Thumbnail	26.3%
Co-occurrence	19.8%
HDS-S	19.5%
Convolution	18.1%

turn. However, none of these improved the mean average precision above that of the submitted run.

To get further insight into the results we looked at average precision by query. Fig. 2 shows the maximum and median average precision together with the results for our run. The queries are ordered by maximum precision to sort them by difficulty. It is clear from this graph that our system performed consistently well across all the queries. It was above median for 24 of the 26 queries and performed the best for one query.

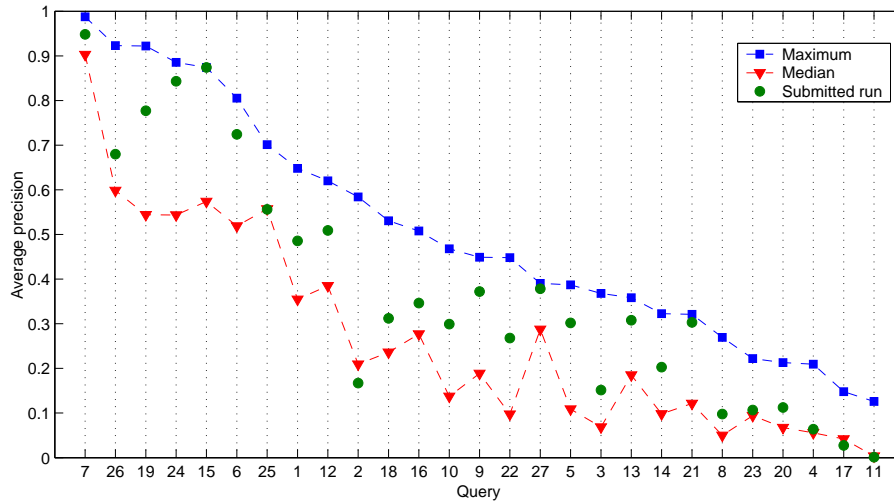


Fig. 2. Average precision by query

To determine the reason for our consistency we looked at the performance of individual features by query. Fig 3 shows the average precision for each feature and the combination of features. We can pull several interesting facts from the figure:

- The submitted run outperformed all individual features for 16 of the 26 queries.

- In all cases the submitted run was better than the mean and median of the individual features.
- The most consistent feature was Gabor. It was top for 14 queries.
- HDS-S (colour feature) showed the most variation. It was the worst for 13 queries and best for 4. Of these 4, half of the query images were colour.
- Thumbnail beat the maximum (of all submitted runs) for 3 queries.

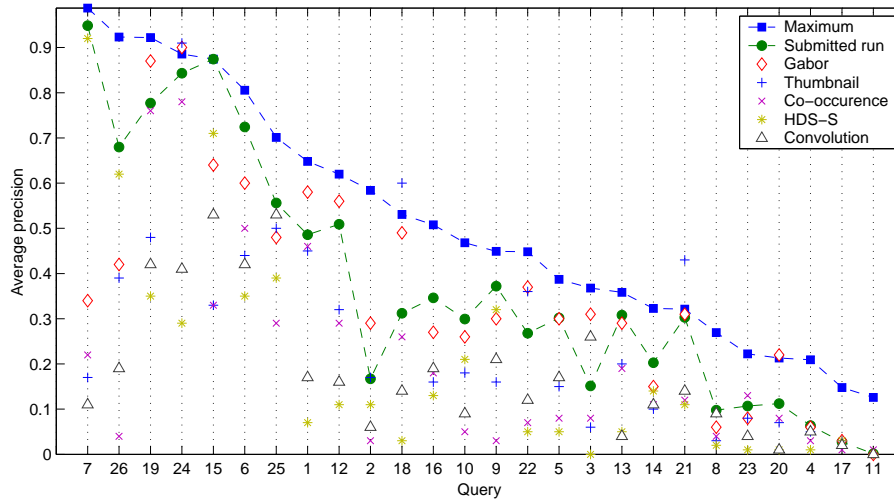


Fig. 3. Average precision by query for individual features

It appears from these facts that the main reason for our consistency was the good performance by all the features used coupled with the effect of summing the features. Combining distances using equal weighting evens out the performance variation from each feature across the queries.

Individually, Gabor and thumbnail performed best. This was expected and is almost certainly due to the locality information within both the features. This is an effective discriminator due to the characteristic composition of images within the collection. HDS-S gave good discrimination for the colour queries but otherwise was poor. The feature is a histogram and contains no locality information.

The variation in average precision indicates that it would be possible to improve performance by weighting features differently depending on the query. However, this is not a trivial problem. Other than increasing the importance of the colour feature for colour images there are no obvious links between image types and feature performance. To tackle this problem we would need to apply learning methods.

4.2 Comparison with other systems

We also compared our approach to the methods used by the other top performing systems. The top seven systems were from Buffalo, Aachen and us.

Aachen [10] used a similar approach with visual features. They used a wide selection of features and optimised the weightings to the collection. They did this by creating their own relevance judgements and then evaluating different weightings to find the best. They also employed a simple query expansion method, using the query image and its nearest neighbour to query the collection.

Buffalo [9] used a different approach, combining visual and text retrieval. An initial visual query was used to rank the images. The text associated with the top images was then used to generate a text query. Finally the text and visual results were combined linearly.

Our approach was simpler than those above yet it gave similar performance. We believe that this was due to the quality of features used.

5 Future work

In addition to the search task we also used put the data set into a novel browsing network, NN^k , developed in our group [11]. Although we did not carry out a formal evaluation we found that through browsing it was possible to rapidly access similar images in the collection. A medical expert would be presented with a range of images to review and identify those that they were particularly interested in. It is clear that the browsing paradigm is an effective way of searching data collections of this size and complexity.

This system can be accessed at <http://km.doc.ic.ac.uk>, via the demo page. Open the iBase application and then select the CasImage collection on the settings tab. A screen shot of the application is shown in Fig 4.

The experiments carried out used very effective features. However, they were combined in a simple way, by summing the distances obtained from each feature. As shown shown in the postmortem analysis there was a wide variation in performance of features for different queries. This would indicate that when querying for certain modalities or anatomic regions different combinations of features may perform better. By varying the weights applied to features we can introduce a degree of plasticity into our system and then use machine learning techniques to improve retrieval performance.

Given training data we would like to train a support vector machine as a meta classifier. We have deployed this technique in other contexts, see Yavlinsky et al. [12]. We propose that it would be possible to learn the optimal weights for retrieving a specific modality, such as CT scan or x-ray.

6 Conclusion

Our experiments showed that it is possible to achieve good retrieval performance on a medical image collection using a CBIR approach. We used global features,

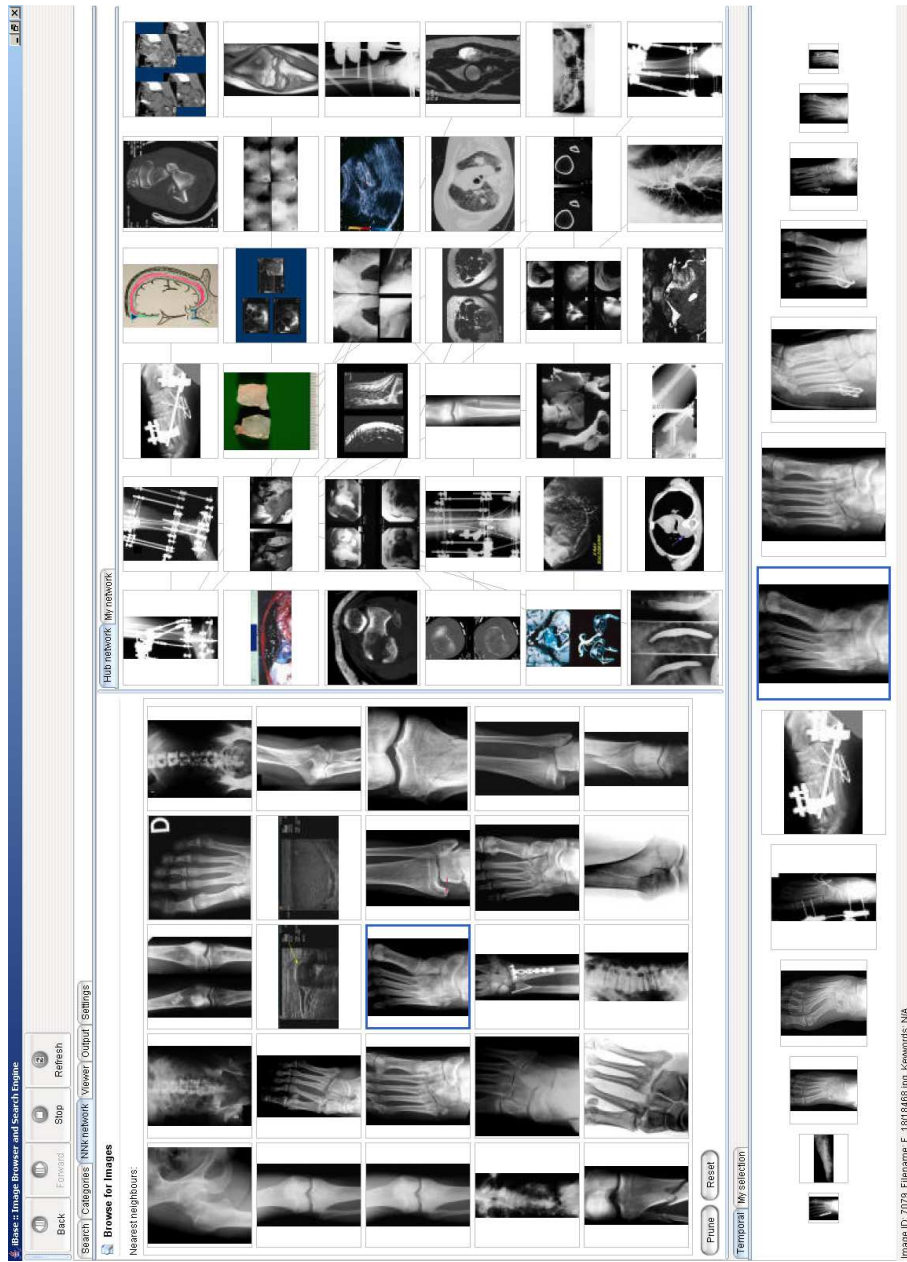


Fig. 4. iBase showing the NN^k browsing window

which is in contrast to the highly specialised methods normally used for medical imaging. Given the constrained domain we were able to choose visual features that had good discriminatory power for the collection. We identified texture and locality as the key discriminators. Correspondingly, we predicted that a tiled Gabor feature would be an ideal feature for the dataset. The analysis of our results subsequently showed this to be the case.

Combining features using equal weighting was beneficial across the query set. It smoothed out the effect of individual features and gave the maximum retrieval performance. In addition, the analysis of individual features indicates that there is scope for applying learning methods to fuse features with optimised weights and further improve retrieval performance.

References

1. 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)
2. Tieu, K., Viola, P.: Boosting image retrieval. In: International Conference on Spoken Language Processing. (2000)
3. Haralick, R.: Statistical and structural approaches to texture. *Proceedings of the IEEE* **67** (1979) 786–804
4. Gotlieb, C.C., Kreyszig, H.E.: Texture descriptors based on co-occurrence matrices. *Computer Vision, Graphics and Image Processing* **51** (1990) 70–86
5. Howarth, P., Rüger, S.: Evaluation of texture features for content-based image retrieval. In: *Proceedings of the International Conference on Image and Video Retrieval*, Springer-Verlag (2004) 326–324
6. Turner, M.: Texture discrimination by Gabor functions. *Biological Cybernetics* **55** (1986) 71–82
7. Manjunath, B., Ma, W.: Texture features for browsing and retrieval of image data. *IEEE Trans on Pattern Analysis and Machine Intelligence* **18** (1996) 837–842
8. Manjunath, B.S., Ohm, J.R.: Color and texture descriptors. *IEEE Trans on circuits and systems for video technology* **11** (2001) 703–715
9. Ruiz, M., Srikanth, M.: UB at CLEF2004: Part 2 – cross language medical image retrieval. *CLEF Workshop* (2004)
10. Deselaers, T., Keysers, D., Ney, H.: FIRE - flexible image retrieval engine: Image-CLEF 2004 evaluation. *CLEF Workshop* (2004)
11. Heesch, D., Rüger, S.: NN^k networks for content-based image retrieval. In: 26th European Conference on Information Retrieval, Springer-Verlag (2004) 253–266
12. Yavlinsky, A., Pickering, M., Heesch, D., Rüger, S.: A comparative study of evidence combination strategies. In: *IEEE International Conference on Acoustics, Speech, and Signal Processing. Volume III.* (2004) 1040–1043