Combining Multilevel Visual Features for Medical Image Retrieval in ImageCLEFmed 2005

Wei Xiong[1], Bo Qiu[1], Qi Tian[1], Changsheng Xu[1], S.H. Ong[2], and Kelvin Foong[3]
[1]Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613
{wxiong,visqiu,tian,xucs}@i2r.a-star.edu.sg
[2]Department of Electrical and Computer Engineering, National University of Singapore,
Singapore 117576, eleongsh@nus.edu.sg
[3]Department of Preventive Dentistry, National University of Singapore,
Singapore 119074, pndfwc@nus.edu.sg

Abstract

In this paper we report our work on the fully automatic medical image retrieval task in ImageCLEFmed 2005. First, we manually identify visually similar sample images by visual perception for each query topic. These help us understand the variations of the query topic and form templates for similarity measure. To achieve higher performance, two similarity measuring channels are used with each using different sets of features and operating in parallel. Their results are then combined to form a final score for similarity ranking. To improve efficiency, a pre-filtering process using other features is utilized to act as a coarse topic image filtering before the two similarity measures for fine topic retrieval. During retrieval, no relevance feedback is used. Only visual features are used in our experiments for all the topics including visually possible queries (topics 1–11), mixed visual/semantic queries (topics 12-22) and semantic (rather textual) queries (topics 23-25). Over 50,000 medical images our approach achieved a mean average precision of 14.6% for all 25 topics, ranked as the best-performance run for the automatic medical image retrieval task in the ImageCLEFmed 2005.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval;

General Terms

Measurement, Performance, Experimentation

Keywords

medical, image retrieval, ImageCLEFmed, hierarchical,multiple, feature, fusion, performance, mean average precision, experiments

1 Introduction

ImageCLEFmed is a subtask of ImageCLEF. In 2005, it offers tasks for both system-centered and user-centered evaluation of image retrieval systems. One of the four tasks offered is the ImageCLEFmed which contains two medical-related sub tasks: medical image retrieval and automatic annotation for medical images. This paper reports our work in the sub-task of medical image retrieval. We confine our efforts in automatic runs using visual features only. The test data for this sub-task consist of four components: Casimage (containing 8725 radiology pathology images), MIR (containing 1177 nuclear medicine images from Mallinckrodt Institute of Radiology), PEIR (containing 32319 pathology and radiology images from Pathology Education Instructional Resource) and PathoPIC (containing 7805 pathology images). Besides the different imaging modalities and anatomic regions, these 50k+ images are of various sizes and image qualities. Some of them use 1-color channel while others 3-color channels. Results of ImageCLEFmed 2004 [4] with 26 retrieval topics on the Casimage collection are provided as training data for ImageCLEFmed 2005. 25 query topics are provided in this track. Each of them contains topic statements in English, French and German, and a collection of images for each topic. Normally one or two example images for the desired result for the topic are supplied. One query also contains a negative example as a test. These queries are divided into visually possible queries (topics 1–11), mixed visual/semantic queries (topics 12-22) and semantic (rather textual) queries (topics 23-25). Since we will use visual characteristics alone, it would be very hard for us to handle topics 12-25. This has been proved by the submitted results of this forum: the best run of the mixtures of textual and visual retrievals is almost twice good as that of runs using visual-only retrievals.

Content-based image retrieval (CBIR) retrieves images in terms of their visual contents [8, 9]. We apply it into this medical image retrieval campaign. The low-level raw features we used include pixel-level features such as color (which are very local), regional-level features such as regional color, shape and texture (which combine local and global information), and image-level features such as color and texture statistical property (which are global). Before retrieval, we manually identify more visually similar sample images by visual perception for each query topic. We can also generate some synthetic images from these chosen images with reasonable visual varieties. This helps us understand the visual variations of the query topic. Take the example of topic 3 provided in this track which asks to show pathology images of an entire kidney. Figure 1 shows the example image provided for this query topic together with three of many other instances of the query. They all contain an entire kidney but they differ in almost all visual properties. In fact, there are a lot of variations of an entire kidney in size, shape, color, and/or skin texture. It shows that one should not choose visual features from a single example alone. It also shows that the retrieval task in this campaign is rather challenging. In some sense, retrieval of each topic is like a task of face recognition which has been proven to be very difficult [14].

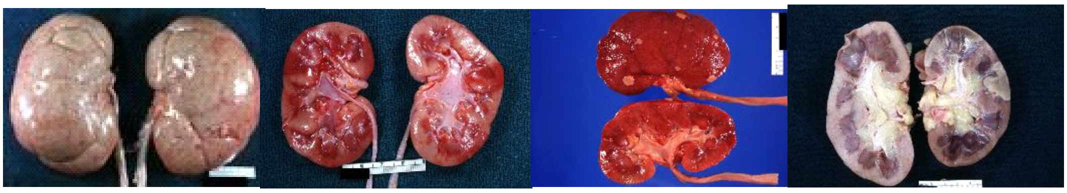

Figure 1: A query topic has many instances visually different. The leftmost is provided as the query example while others are the instances chosen.

The sample images chosen for each topic are used as training data to design similarity measuring functions. We have used three functions η_1, η_2 and η_3 for each topic with three different sets of features, respectively. The first function η_1 serves as a filter to remove those dissimilar images. This improves efficiency as many images can be excluded in the next comparison stage. Next, η_2 and η_3 operates in parallel [12] to yield two similarity measures. Then they are combined to produce a final score for image ranking. During retrieval, no relevance feedback is used anymore.

We submitted seven runs of experiments. All use visual features only but cover all the 25 topics including the visually possible queries (topics 1–11), mixed visual/semantic queries (topics 12-22) and semantic (rather textual) queries (topics 23-25). Five runs achieved top performance in this sub-task. We covered the top five runs in this subtask.

In the following sections, we introduce our work in more detail. The features used are explained in Section 2, followed by the retrieval methodology presented in Section 3. Section 4 explains our experiments. Conclusions are then drawn.

2 Multiple Feature Descriptions

In this section, we describe the visual features used in our work. A survey of visual features useful for general CBIR can be found in [13]. We have employed color, shape, texture characteristics at the pixel level, the region level and the entire image level.

2.1 Global Color and Layout Property

We have noticed that some of the query examples provided are colorful while others are gray. Those color images use three color channels. Most of gray images use one channel only. However there are some that still employ three channels. This channel information can be used directly to classify images. Furthermore, the layouts of images are different consistently. For example, the ultrasound images are almost triangles. These features form description set $\mathbf{F}_1$.

2.2 Low Resolution Pixel Map at Pixel Level

Images in the database, even in the same class, vary in size and may have translations. Resizing them into a thumbnail [2, 5] of a fixed size, through introducing distortions, may overcome the above difficulties in representing the same class of images in the database. It is a reduced and low-resolution version from the original image in the database ignoring its original size. Thus this can also remove noise in the high-frequency band. A 16-by-16 image pixel map, called an "icon", is used. Examples are shown in Figure 2. Here the left three images are original ones in the same class and the right three images are their respective reduced versions. They look more similar visually than their original versions. These so-called "icons" are extensively used in face recognition and have proven to be effective [2]. We have also applied them to medical image retrieval [12]. We call them description set $\mathbf{F}_2$.

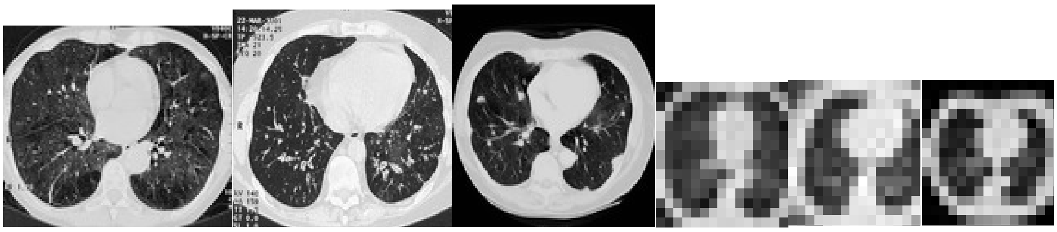

Figure 2: Examples of original images and their respective low-resolution maps

2.3 Blob Feature at Object/Image Level

We consider both regional and image-wide color, texture and shape features. Besides the pixel color, local contrast, anisotropy and polarity are captured to form a joint color-texture-spatial feature vector space. Gaussian-mixture models are built locally. The purpose is to define homogeneous regions in the feature space quantitatively. Numbering less than 5, the number of Gaussians for each region may be different from each other. All the model parameters are found using the EM algorithm. These regions are extracted by a region merging process. The merged regions are segmented and referred to as meaningful local objects. The color, texture and shape properties of the regions are computed. The largest 10 regions are identified and obtained separately. Mean values and statistical properties of color, texture features and area are counted. The contour of each region is represented by elliptical Fourier expansion descriptors. We use the first 20 coefficients. Each region can be represented by the ellipse reconstructed from the first order of the expansion. These are the so-called blob representation [3]. Three pairs of examples are shown in Figure 3 where for each pair the left one is the original image and the right one is its blob representation. We have also included the global color histogram and texture histogram over the whole image. The above regional features and the global features form feature set $\mathbf{F}_3$, called as "blob" in this paper. The feature vector is of 352-dimension. For a more detailed description of the specific usage, see our previous work [11, 12].

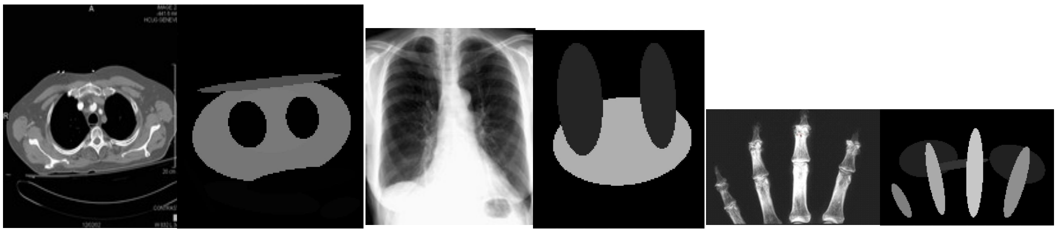

Figure 3: Examples of original images and their respective blob representations

3 Retrieval Methodology

The section presents the retrieval methodology used in this work. The basic processing flow of our approach is illustrated in Figure 4. Above the dashed horizontal line are processing procedures for any given query and below the line are procedures for the images in the test databases.

Before retrieval, we browse the four databases provided especially the CASImage database used for training. For j-th query topic, j=*1,...,25*, some more semantically similar images (which may be visually different) are chosen to form a set Q_n^j of n training images. For each image, some raw features, such as color, geometrical and texture properties, are extracted to form a $p \times 1$ feature vector $\mathbf{x_i} = (x_{i1}, x_{i2}, \cdots, x_{ip})$ for image i.

Principal components are analyzed upon a set of such features. (We will provide more details later.) An eigenspace E_j is set up for each query topic j, $j = 1, \ldots, 25$. Feature dimension may also be reduced, which is illustrated as a dashed box. The feature vector of a test image $\mathbf{x_i} = (x_{i1}, x_{i2}, \ldots, x_{ip})$ is then projected to E_j. The similarity is measured in E_j.

This procedure is repeated for all test images to generate a similarity ranked list for them. For the test image, a pre-filtering is introduced using $\mathbf{F}_1$. Those images that are impossible to be similar are excluded earlier (denoted by "N"). Only those which can pass (indicated by "Y") will go to the final comparison stage. In this final stage, two parallel engines are introduced for similarity measures. They use independent sets of features "icon" and "blob", representing local and global characteristics, respectively.

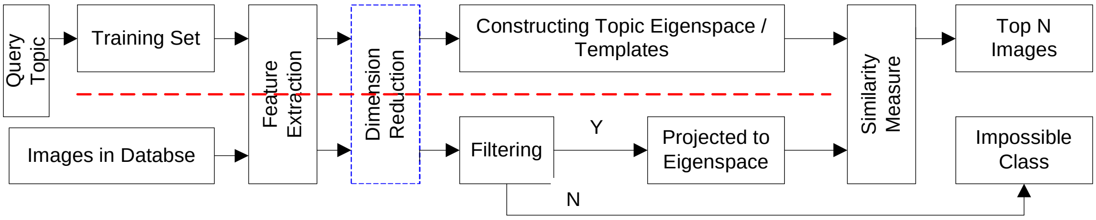

Figure 4: A diagram of the processing flow

More specifically, we analysis principal components and utilize them in two ways. The first is for feature dimension reduction. The second is used to design similarity measuring functions [10, 12]. Given a training dataset Q_n^j with n images: $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n)^T$, the generating matrix can be constructed as [10, 12]

$$\mathbf{C}_1 = \mathbf{X}\mathbf{X}^{\mathrm{T}}, \quad \text{or}, \quad \mathbf{C}_2 = \mathbf{X}^{\mathrm{T}}\mathbf{X}. \tag{1}$$

Here $\mathbf{C}_1$ is of $n \times n$ and $\mathbf{C}_2$ is of $p \times p$. As mentioned before, n is the number of images/vectors and p is the number of features. $\mathbf{C}_1$ is used to generate templates of this dataset and $\mathbf{C}_2$ is used to reduce the dimension of the feature space when necessary. Supposing m out of n eigenvalues (λ_i) and their eigenvectors ($\mathbf{u}_i$) are chosen based on $\mathbf{C}_1$. From the eigenvector matrix $\mathbf{U} = (\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_m)^T$, the template vectors $\mathbf{V} = (\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_m)^T$ are given by using

$$\mathbf{v}_i = \frac{1}{\sqrt{\lambda_i}} \mathbf{X}^T \mathbf{u}_i, \qquad i = 1, ..., m. \tag{2}$$

Given a test image set, its feature matrix $\mathbf{Y}$ is reconstructed by

$$\mathbf{Y}' = \mathbf{V}\mathbf{V}^T\mathbf{Y}, \tag{3}$$

with a least square error

$$s = \|\mathbf{Y} - \mathbf{Y}'\|. \tag{4}$$

The similarity-measuring functions η_2 and η_3 have the same form of this error but with different feature sets as parameters.

The weighted summation rule [1, 6] is used to fuse the two similarity measures:

$$d = w_1 s_1 + w_2 s_2, \tag{5}$$

where s_1 and s_2 are the similarity computed above using different feature sets $\mathbf{F}_1$ and $\mathbf{F}_2$ whereas w_1 and w_2 are weighted coefficients subject to $0 \leq w_1, w_2 \leq 1$, $w_1 + w_2 = 1$. The resulting distance d serves as the final score for ranking: the larger the score is, the less similar the query and the test image are.

4 Experiments

In this campaign, we use visual features alone for all topic retrieval tasks including those that require text information. Experiments start from choosing training data. For each topic, we manually choose some images in the test database to represent the visual varieties of the query topic. Three undergraduate engineering students without medical background select these images. The only criteria are the visual appearance of the images. Consequently, it is no doubt there are many images wrongly chosen and the numbers of images are larger. The more correct the visual varieties of the query topic we can collect into the training set, the better the representation is semantically. This is done offline and before the retrieval.

We have also referred to the results from the baseline work from medGIFT [7]. Table 2 lists the number of images collected for each query topic. Here "q", "a" and "b" refers to the query topic and the number of images for two sets of training data, respectively. Total number of images in the training set "a" is 4789 with a mean 191.56 for each topic. In other words, 9.573% of the 50026 test images are used for training, which is a small portion. For training set "b", total there are 3874 images (i.e., 7.744% of the 50026 test images) with a mean 154.96 for each topic.

Next, we compute all the three feature sets $\mathbf{F}_1$, $\mathbf{F}_2$ ("icon") and $\mathbf{F}_3$ ("blob") for all images including those for training and testing. The similarity measuring function η_1 is a unit function for binary classification in terms of $\mathbf{F}_1$. Design η_2 and η_3 according to Equations (1) to (4) using $\mathbf{F}_2$ and $\mathbf{F}_3$ respectively. We combine their results according to Equation (5) with the same coefficients ($w_1 = w_2 = 0.5$).

Table 1: Number of images in the training set for each of 25 query topics

q	1	2	3	4	5	6	7	8	9	10	11	12	13
a	460	161	79	142	146	194	19	9	107	257	33	418	382
b	457	161	79	117	96	24	19	9	107	257	39	360	371
q	14	15	16	17	18	19	20	21	22	23	24	25	
a	420	105	40	316	161	181	190	149	44	23	571	179	
b	140	167	26	286	141	176	150	56	44	10	468	117	

We submitted seven runs of retrievals. Table 2 lists their labels and their performance in terms of mean average precision (MAP). They are divided into 3 groups as shown in Table 2. Only Group 3 “I2Rfus” utilizes all techniques mentioned above. Other runs in Groups 1 and 2 use parts of the techniques for comparison. In Group 1, two subgroups are further divided in terms of the feature sets used. Subgroup 1 uses “blob” and Subgroup 2 uses “icon”. In each subgroup, we have two members, one using pre-filtering (“I2RbPBcf” and “I2RcPBcf”) while others (“I2RbPBnf” and “I2RcPBnf”) do not.

Table 2: MAPs of seven runs submitted to ImageCLEFmed 2005

Group	Run	MAP	Group	Run	MAP
1 (set “a”)	I2RbPBnf	0.1067	2 (set “a”)	I2Rfus	0.1455
1 (set “a”)	I2RcPBnf	0.1114	3 (set “b”)	I2RbP1nf	0.0928
1 (set “a”)	I2RbPBcf	0.1068	3 (set “b”)	I2RcP1nf	0.0934
1 (set “a”)	I2RcPBcf	0.1188			

We observe that using the “icon” feature set gives normally slightly higher MAP than using the “blob” feature set. This is clear by comparing “I2RbPBnf” (0.1067) against “I2RcPBnf” (0.1068), and “I2RbPBcf” (0.1114) against “I2RcPBcf” (0.1188), respectively. The binary classifier in Stage 1 improves the entire system performance. To see this effect, we can compare “I2RbPBnf” (0.1067) against “I2RbPBcf” (0.1114) and “I2RcPBnf” (0.1068) against “I2RcPBcf” (0.1188), respectively. The improvement is more significant when using the “icon” feature set (11.24%) than using the “blob” set (with 4.4%). Group 2 is the fusion of “I2RcPBnf” and “I2RcPBcf” where the weights are equal. It achieves the best results (MAP=14.55%).

It is important to select more examples to form a training set for each query topic before retrieval. To have a comparison, some of images (the underlined numbers in Table 1) are removed from the representation sets of some topics. We repeat experiments “I2RbPBnf” (0.1067) against “I2RcPBnf” (0.1068) but using these new training sets (Set “b”). This results in “I2RbP1nf” (using “blob” with MAP=0.0928) and “I2RcP1nf” (using “icon” with MAP=0.0934) in Group 3. Again, “icon” features have slightly better precision performance. Comparing experiments vertically using the two training sets, one finds that performance of Group 3 drops down using either feature set. This shows that the representation of the query topic using the training set is indeed important.

5 Discussion and conclusion

We have reported our efforts to the medical image retrieval task in the ImageCLEFmed 2005. We analyzed the contents of images and employ three sets of visual features at different levels to represent each image. We start from manual selecting some training images for each topic before retrieval. These images construct a training set for us to span an eigenspace for the topic and to define similarity metrics for it. A pre-filtering process is used to act as a coarse topic image filtering before the two similarity measures for fine topic retrieval. The features are simple and the comparison is easy and fast. Many test images can be simply classified into impossible class reliably. To achieve higher performance, two similarity measuring channels are used. They use different sets of features and operate in parallel. Their results are then combined to form a final score for similarity ranking. We have not used relevance feedback during the retrieval.

In our experiments, only visual features are applied to not only the 11 visual-retrieval-possible topics, but also those 13 topics needing rather textual information. We have submitted seven runs in this track. Our best approach utilizes multiple sets of features with pre-filtering and fusing strategies, which enables us to achieve a very good performance in the visual-only group.

It should be noted that our work is based on the pre-selection of more example images for the query topic. The training data were chosen offline and it may be inconvenient for online applications. Both the qualities and the number of these images influence the retrieval performance. In our current efforts, they are still large for some topics and yet we have not refined them. Our future efforts would refine the sets and combine some machine learning techniques to facilitate the selection.

References

[1] F. Alkoot and J. Kittler. Experimental evaluation of expert fusion strategies. *Pattern Recognition Letters*, 20:1361–1369, 1999.

[2] Peter N. Belhumeur, J. P. Hespanha, and David J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on pattern analysis and machine intelligence*, 19(7):711–720, 1997.

[3] C. Carson, S. Belongie, H. Greenspan, and J. Malik. Recognition of images in large databases using color and texture. *IEEE Transactions on pattern analysis and machine intelligence*, 24(8):1026–1038, 2002.

[4] Paul Clough, Mark Sanderson, and Henning Müller. The clef 2004 cross language image retrieval track. In C. Peters, J. Gonzalo P. Clough, G. Jones, M. Kluck, and B. Magnini, editors, *Lecture Notes in Computer Science (LNCS)*, Multilingual Information Access for Text, Speech and Images: Results of the Fifth CLEF Evaluation Campaign, Heidelberg, Germany, 2005. Springer. in print.

[5] P. Howarth, A. Yavlinsky, D. Heesch, and S. Rüger. Visual features for content-based medical image retrieval. In *Proceedings of Cross Language Evaluation Forum (CLEF) Workshop 2004*, Bath, UK, 2004.

[6] L. I. Kuncheva. A theoretical study on six classifier fusion strategies. *IEEE Transactions on pattern analysis and machine intelligence*, 24(2):281–286, Feburary 2002.

[7] Henning Müller, A. GeissbMühler, and P. Ruch. Report on the clef experiments: Combining image and multi-lingual search for medical image retrieval. In C. Peters, J. Gonzalo P. Clough, G. Jones, M. Kluck, and B. Magnini, editors, *Lecture Notes in Computer Science (LNCS)*, Multilingual Information Access for Text, Speech and Images: Results of the Fifth CLEF Evaluation Campaign, Heidelberg, Germany, 2005. Springer. in print.

[8] Henning 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*, 73:1–23, 2004.

[9] 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(12):1349–1380, 2000.

[10] Wei Xiong, Bo Qiu, Qi Tian, , Changsheng Xu, Sim Heng Ong, and Kelvin Foong. Content-based medical image retrieval using dynamically optimized regional features. In *The IEEE International Conference on Image Processing 2005*, Genoa, Italy, September 2005. to appear.

[11] Wei Xiong, Bo Qiu, Qi Tian, Henning Müller, and Changsheng Xu. A novel content-based medical image retrieval method based on query topic dependent image features (QTDIF). *Proceedings of SPIE*, 5748:123–133, 2005.

[12] Wei Xiong, Bo Qiu, Qi Tian, Changsheng Xu, Sim Heng Ong, Kelvin Foong, and Jean-Pierre Chevallet. Multipre : A novel framework with multiple parallel retrieval engines for content-based image retrieval. In *ACM Multimedia 2005*, Hilton, Singapore, November 2005. to appear.

[13] Zijun Yang and C.-C. Jay Kuo. Survey on image content analysis, indexing, and retrieval techniques and status report of mepg-7. *Tamkang Journal of Science and Engineering*, 2(3):101–118, 1999.

[14] W. Zhao, R. Chellappa, A. Rosenfeld, and P. Phillips. Face recognition: A literature survey, 2000.