Content-based Retrieval of Medical Images with Relative Entropy

Mehran Moshfeghi[a], Craig Saiz[a], and Hua Yu[b]
[a]Electrical Engineering and Computer Science, University of California, Irvine, CA 92697-3425
[b]Information and Computer Science, University of California, Irvine, CA 92697-3425

ABSTRACT

Medical image databases are growing at a rapid rate because of the increase in digital medical imaging modalities and the deployment of Picture Archiving and Communication Systems (PACS), Electronic Medical Records (EMR) and telemedicine applications. There is growing research interest in Content-Based Image Retrieval (CBIR) of medical images from such digital archives. A new distance function for CBIR is presented for measuring the similarity between two images. The distance function is a variant of relative entropy, or the Kullback-Liebler distance. The new distance is the sum of the relative entropy of the two images to each other. The latter is a symmetric non-negative function and is only zero when the two images have identical probability distributions. This method has been implemented in a prototype system and has been applied to a database of medical images. Initial results demonstrate improvements over L_1-norm and L_2-norm histogram matching. The method is computationally simple since it does not require image segmentation. It is invariant to translation, rotation and scaling. The method has also been extended to support retrieval based on Region-Of-Interest (ROI) queries.

Keywords: Content-Based Image Retrieval (CBIR), relative entropy, Region-Of-Interest (ROI), medical image archive, medical imaging, PACS

1. INTRODUCTION

Image retrieval with traditional Picture Archiving and Communication Systems (PACS) and Electronic Medical Records (EMR) systems is based on textual descriptions. Information such as filenames, keywords and patient identification numbers are typically stored with the images and are used to retrieve images. Correct retrieval from such systems typically requires exact keywords. Doctors, however, may use different keywords to describe the same image because of the difficulty in interpreting some images and the lack of a unified lexicon. There is therefore a need for adding Content-Based Image Retrieval (CBIR) capability in PACS systems in order to provide image indexing and retrieval based on content [1]. In contrast to text-based techniques CBIR systems retrieve images by comparing features extracted from the images themselves. CBIR techniques can be combined with text-based techniques to provide powerful image retrieval systems. CBIR systems can also assist in the construction of teaching files or be used to reduce the cost of image analysis by assisting in the classification of routine images.

CBIR systems offer different kinds of searches. A category search looks for arbitrary images of a certain class such as chest x-ray images. A target search aims to find a specific image such as the query image or another image of the same object. There has been a great deal of interest in the research community to develop CBIR algorithms for non-medical image and video collections [2-4]. With the deployment of digital medical archives and EMRs, there has also been growing interest in indexing biomedical images [5-8]. Medical images have special characteristics and often require retrieval algorithms that are different from those used for non-medical images. Several groups have reported algorithms that are customized for specific applications and modalities, such as mammography [9], Computerized Tomography (CT) for neuroradiology [10], and CT for lung imaging [11,12]. Systems have also been designed recently that can handle general collections of medical images by calculating regional image characteristics and by using a-priori anatomic and physiologic knowledge [13,14].

Figure 1 illustrates a block diagram of a typical CBIR system. Images are input into a database and for each image a feature vector is calculated. Color, texture, and shape are typical features used in many CBIR systems. Feature vectors are stored as an index for the image database. At the query stage the feature vector of the query image is extracted and

its distances with the feature vectors of the index are calculated and compared. Images are then ranked in similarity based on the value of this distance. Color histogram feature matching is popular because it is easy to compute. Histogram matching also eliminates the need for image registration because histograms are invariant to scaling, translation and rotation. In this paper a new distance metric for measuring the similarity between two histograms is reported. The method is tested on medical images and its performance is compared with conventional distance metrics. The method is also extended the method to support Region-Of-Interest (ROI) queries.

2. METHODS

Color histogram matching techniques typically use the L_1-norm [15] or the L_2-norm [2] distance. Let p(x) be the histogram of a search image and q(x) be the histogram of a query image. Thus, p(x) represents the frequency of occurrence of pixels in a search image with pixel value x. Likewise, q(x) represents the number of pixels in a query image with pixel value x. Then the L_1-norm distance is defined as:

$$L_1(p,q) = \sum |p(x) - q(x)| \quad (1)$$

and the L_2-norm distance is defined as:

$$L_2(p,q) = [\, \sum(p(x) - q(x))^2 \,]^{1/2} \quad (2)$$

where the summations in Eqn.1 and Eqn.2 are over the histogram bin values. The L_1-norm distance between two image histograms is always less than twice the number of pixels per image, while the L_2-norm distance is always less than $\sqrt{2}$ times the number of pixels per image.

This paper reports a new distance metric that can be used for CBIR color matching methods. The new distance uses a variant of relative entropy to measure similarity between two images. The entropy of a random variable is a measure of the average number of bits needed to represent it. The relative entropy or Kullback-Liebler distance, D(p||q), represents the additional number of bits needed on average, as compared to the entropy limit, to represent a random variable. Relative entropy is used in information theory to measure the distance between two random variable probability distributions and is a measure of the inefficiency of assuming that the probability distribution is q(x) when the true distribution is p(x) [16]. In statistics it arises as an expected logarithm of the likelihood ratio and is defined as:

$$D(p||q) = \sum p(x)\log[p(x)/q(x)] \quad (3)$$

where the summation is over x. Relative entropy is a convex function of p(x), is always non-negative, and equals zero only if p(x)=q(x). Relative entropy is not a true distance because it is not symmetric; that is $D(p||q) \neq D(q||p)$. Therefore, a new distance, $D_s(p,q)$, is introduced that is the sum of the relative entropy of the two images to each other:

$$\begin{aligned} D_s(p,q) &= D(p||q)+D(q||p) \\ &= \sum p(x)\log[p(x)/q(x)] + \sum q(x)\log[q(x)/p(x)] \end{aligned} \quad (4)$$

where the summations are over x. D_s is a symmetric non-negative function and can be used to measure the similarity between two images. Smaller values of D_s represent greater similarity between two images. D_s is only zero when the probability distributions of the two images are identical. In this paper the histograms of the search image and the query image are used to represent the probability distributions p(x) and q(x), respectively. Other authors have used mutual information for medical image retrieval [17]. Mutual information is a measure of the amount of information that a random variable contains about another random variable. It represents the reduction in the uncertainty of one random variable due to knowledge of another random variable. However, mutual information of two independent random variables is zero [16] and cannot be used as a reliable similarity metric for images.

A typical query image includes relevant areas as well as irrelevant areas. Irrelevant areas reduce the accuracy of CBIR systems that use global features. It is therefore advantageous to develop systems that have the ability to determine similarity based on relevant regions [18,19]. The relative entropy sum technique has been extended to support user-defined ROI queries, where the user identifies the relevant regions in the query image. Local features are then extracted by limiting the calculations to the outlined ROIs in the query image and the search images. An alternative approach is to divide the images into smaller blocks and extract features for each of the blocks. The similarity between an ROI-based query image and a database image can then be calculated by a weighted combination of the individual image block similarity distances, where the weights are the ratio of the ROI area overlap with the blocks [18].

3. RESULTS

A prototype CBIR system has been implemented that uses the sum of the relative entropy of two images to each other to measure their similarity. Figure 2 shows a screenshot of the system. The interface allows the user to choose an image as a query image. The search options allow the user to specify the distance measure, the histogram bin size, and the number of top matches to display. The system has been tested on a database of 120 digital medical images. The database includes images from different modalities such as x-ray, ultrasound, CT and MRI. For a given query image, the system calculates the similarity distance for the database images and orders the results from best to worst.

For implementation purposes the probability distributions p(x) and q(x) of Eqn.4 are represented by the histograms of the search image and the query image, respectively. The probability density functions are therefore represented with histogram bin frequencies. Images are scaled so that they have the same number of pixels. If the histogram bin size is 1 and the images have a range of n intensity values, then the summations in Eqn.4 are for x = 1 to n. The histograms are often calculated with coarser bin sizes to smooth the histograms and reduce computations. Thus, if the bin size is equal to m the summations in Eqn.4 are for x = 1 to n/m. Experiments showed that bin sizes of 8, 16 or 32 were a good compromise between smoothing out histogram noise and retaining histogram features.

Experiments were carried out to compare the performance of the D_s distance with the L_1-norm and L_2-norm distances. Figure 3 illustrates a chest x-ray image retrieval example. Figure 3(a) shows the top ten retrieved images with the relative entropy sum method. Image labels are the values for the relative entropy sum and the image filename. The query image is 001.jpg and histogram bin size is 16. Figure 3(b) illustrates the retrieval results for the same query image, but using the L_1-norm histogram distance. Image labels are the values for the L_1-norm distance and the image filename. In this case the ninth retrieved image is erroneous since it is not a chest x-ray. Figure 3(c) demonstrates retrieval results for the 001.jpg query image with the L_2-norm histogram distance. Image labels are the values for the L_2-norm distance and the image filename. The tenth image is erroneously retrieved since it is not a chest x-ray. The ranking order of the retrieved chest x-ray images also varies in Figure 3 depending on the distance metric used. In this example the relative entropy sum method performs better than the L_1-norm and L_2-norm distances since all its retrieved images are chest x-rays.

Figure 4 shows a lung CT image retrieval example. Figure 4(a) demonstrates the top ten retrieved images with the relative entropy sum method. The query image is 002.jpg and histogram bin size is 16. Figure 4(b) illustrates the retrieval results for the same query image, but using instead the L_1-norm histogram distance. The ninth retrieved image is erroneous since it is not a lung CT image. Figure 4(c) demonstrates retrieval results for the 002.jpg query image with the L_2-norm histogram distance. The seventh image is erroneously retrieved since it is not a lung CT image. The relative entropy method performs better than the L_1-norm and L_2-norm distance in this example as well.

Figure 5 illustrates a ROI-based image retrieval example, where the query is the marked rectangular region in image 021.jpg. In this example the user is querying for chest x-ray images that have a rib structure on the left side. The top ten retrieved images for each distance function are illustrated in Figure 5. Image labels are the distance function values and image filenames. Histogram bin size is 16 for all the figures. Figure 5(a) shows the top ten retrieved images with the relative entropy sum method. Figure 5(b) demonstrates retrieval results with the L_1-norm histogram distance. The fourth image is erroneous since it is not a chest x-ray with a rib structure on the left side. Figure 5(c) shows the retrieval results for the L_2-norm histogram distance. The fourth, eighth and tenth retrieved image are erroneously retrieved since they are

not chest x-rays with rib structures on the left side. The relative entropy method also performs better than the L_1-norm and L_2-norm distance in this ROI-based query example.

When the relative entropy sum method is applied globally to the entire image, as in Figures 3 and 4, the method is invariant to translation, rotation and scaling transformations because these operations do not change the histogram distribution. As a result, there is no need for image registration prior to image feature comparison and retrieval. The ROI query approach of Figure 5, however, has the disadvantage that it introduces dependencies on translation, rotation and scaling transformations. For example, consider the case where the query ROI outlines a tumor in the center of the image. A similar image with the tumor at the bottom of the image will not be ranked as similar because the calculations exclude regions outside of the query ROI. One approach to overcome this limitation is to do region-matching where the ROI coordinates in the search image are shifted to different positions and the similarity distance is calculated for each location. The smallest similarity distance is then assigned to the image. The computation complexity of this approach, however, is much greater since calculations are repeated for every ROI shift.

4. DISCUSSION

The proposed relative entropy sum CBIR method has low complexity since it does not require image segmentation. The method is invariant to translation, rotation and scaling operations when it is applied to the entire image. A fully automated prototype has been developed and preliminary tests indicate improved results compared to L_1-norm and L_2-norm histogram matching. Precision and recall experiments on larger databases are planned for future work. Relative entropy sum can be used in CBIR systems either as stand-alone, or it can be combined with other features such as shape and texture [20]. Methods that rely solely on histogram matching have the limitation that different images can have histogram distributions that are very similar. The addition of other features can reduce the effects of this limitation. A technique based on multiple features can also apply relevance feedback to the query results to modify the feature weights and improve the system accuracy.

ACKNOWLEDGEMENTS

We would like to thank useful discussions with Dr. Hamid Jafarkhani.

REFERENCES

1. T.M. Lehmann, B.B. Wein, H. Greenspan, “Integration of content-based image retrieval to picture archiving and communication systems, Proc. of Medical Informatics Europe (MIE 2003), 2003
2. W. Niblack, R. Barber, W. Equitz, M. Flickner, D. Glasman, D. Petrovic and P. Yanker, “The QBIC project – querying images by content using color, texture and shape, Proc. of SPIE, vol. 1908, pp. 173-187, 1993
3. A. Pentland, R. Pickard, S. Sclaroff, “Photobook – tools for content-based manipulation of image databases”, Proc. of SPIE, vol. 2185, pp. 34-47, 1994
4. S. Antani, R. Kasturi, and R. Jain, “A survey on the use of pattern recognition methods for abstraction, indexing and retrieval of images and video”, Pattern Recognition, vol. 35, no. 4., pp. 945-965, 2002
5. H. D. Tagare, C.C. Jaffe, and J. Duncan, “Medical image databases: A content-based approach”, J. of the American Medical Informatics Association (JAMIA), vol. 4, no. 3, pp. 184-198, 1997
6. T.M. Lehmann, B.B. Wein, J. Dahmen, J. Bredno, F. Vogelsang, and M. Kohnen, “Content-based image retrieval in medical applications: A novel multistep approach”, Storage and Retrieval for Media Databases, Proc. of SPIE vol. 3972, pp. 312-320, 2000
7. S. Batty, A. Blandford, J. Clark, T. Fryer, and X. Gao, “Content based retrieval of lesioned brain images”, Medical Imaging 2002, Proc. of SPIE vol. 4685, pp. 128-136, 2002
8. S. Antani, L.R. Long, G.R. Thoma, and D.J. Lee, “Evaluation of shape indexing methods for content-based retrieval of X-ray images”, Storage and Retrieval for Media Databases, Proc. of SPIE vol. 5021, pp. 405-416, 2003

9. P. Korn, N. Sidiropoulous, C. Faloutsos, E. Siegel, and Z. Protopapas, “Fast and effective retrieval of medical tumor shapes, IEEE Trans. on Knowledge and Data Engineering, vol. 10, pp. 889-904, 1998
10. Y. Liu, W.E. Rothfus, T. Kanade, “Content-based 3D neuroradiologic image retrieval – preliminary results, Technical Report CMU-RI-TR-98-04, Carnegie Mellon University, Pittsburg, PA, 1998
11. A.F. Abate,M. Nappi, G. Tortora, M. Tucci, “IME – an image management environment with content-based access, Image and Vision Computing, vol. 17, pp. 967-980, 1999
12. C.R. Shyu, C.E. Brodley, A.C. Kak, A. Kosaka, A.M. Aisen, L.S. Broderick, “ASSERT – a physician-in-the-loopcontent-based retrieval system for HRCT image databases, “Computer Vision and Image Understanding, vol. 75, pp. 111-132, 1999
13. H. Greenspan, J. Goldberger, L. Ridel, “A continuous probabilistic framework for image matching, J. of Computer Vision and Image Understanding, vol. 84, no. 3, pp. 384-406, 2001
14. D. Keysers, J. Dahmen, H. Ney, B.B. Wein, T.M. Lehmann, “A statistical framework for model-based image retrieval in medical application, J. of Electronic Imaging, vol. 12, no. 1, pp. 59-68, 2003
15. M.J. Swain and D.H. Ballard, “Color indexing”, Intern. Journal of Computer Vision, vol. 7, no. 1, pp. 11-32, 1991
16. T.M. Cover and J.A. Thomas: Elements of Information Theory, Wiley-Interscience Publication, New York, 1991
17. L. Shunshan, Z. Tiange, and Z. Hong, “Medical images retrieval based on mutual information”, SPIE Medical Image Acquisition and Processing, Proc. of SPIE vol. 4549, pp. 119-125, 2001
18. Q. Tian, Y. Wu, T.S. Huang, “Combine user defined region-of-interest and spatial layout for image retrieval, International Conference on Image Processing (ICIP2000), vol. 3, pp. 746-749, 2000
19. K. Vu, K.A. Hua, W. Tavanapong, “Image retrieval based on regions of interest”, IEEE Transaction on Knowledge and Data Engineering, vol. 15, no. 3, pp. 1045-1049, 2003
20. H. Tamura, S. Mori, and T. Yamawaki, “Textural features corresponding to visual perception”, IEEE Transactions on Systems, Man, and Cybernetics, vol. SMC-8, no. 6, pp. 460-473, June 1978

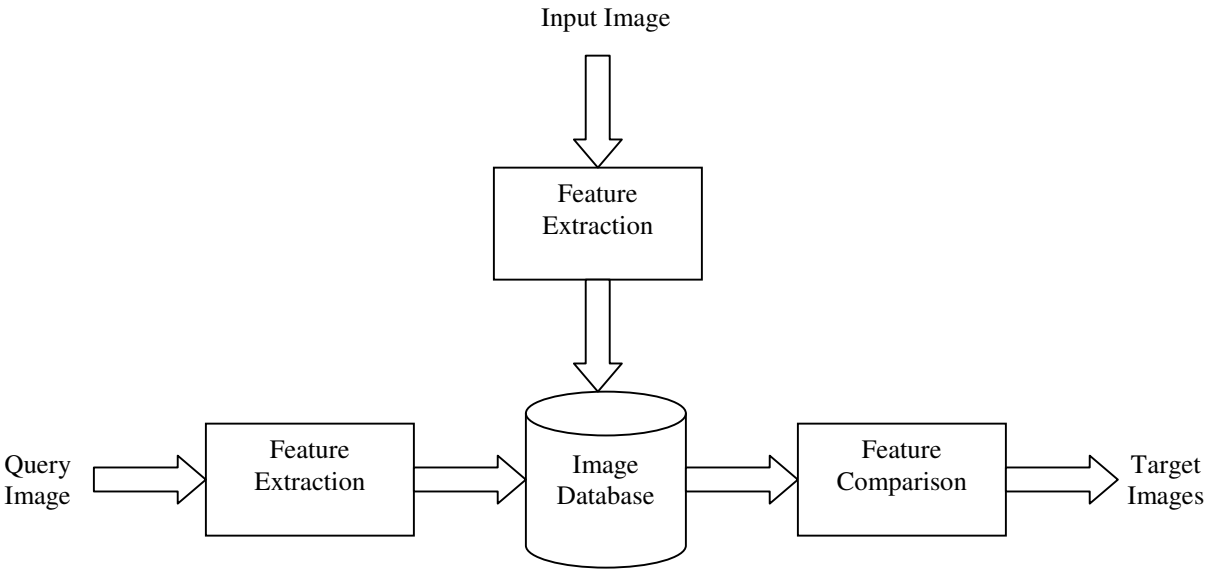


Figure 1. Block diagram of a typical Content-Based Image Retrieval (CBIR) system.

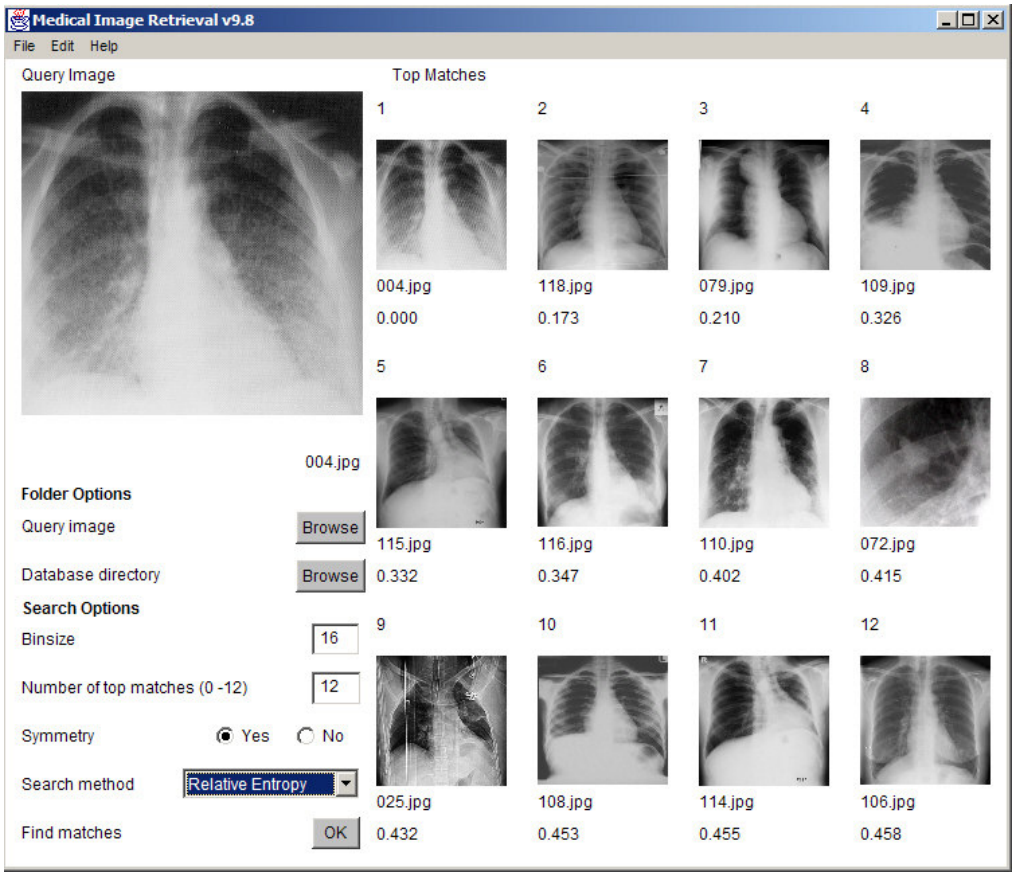
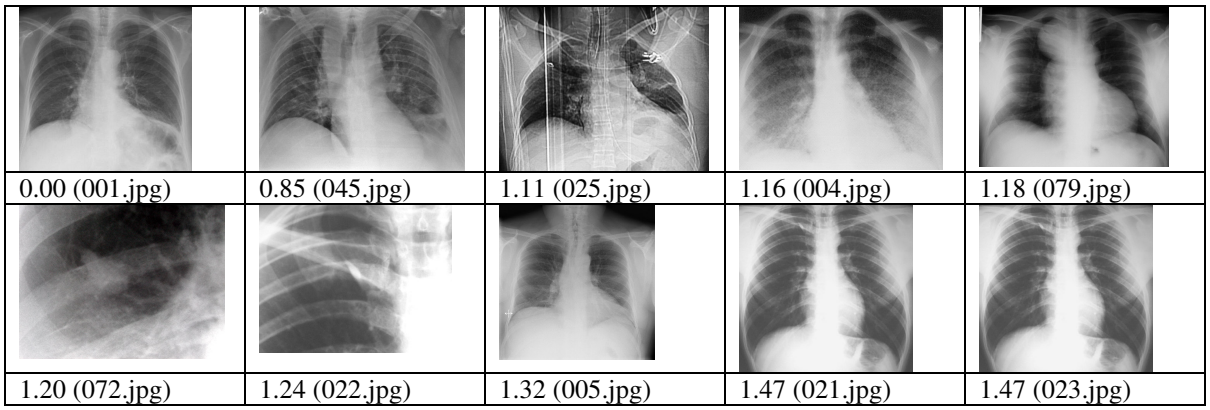
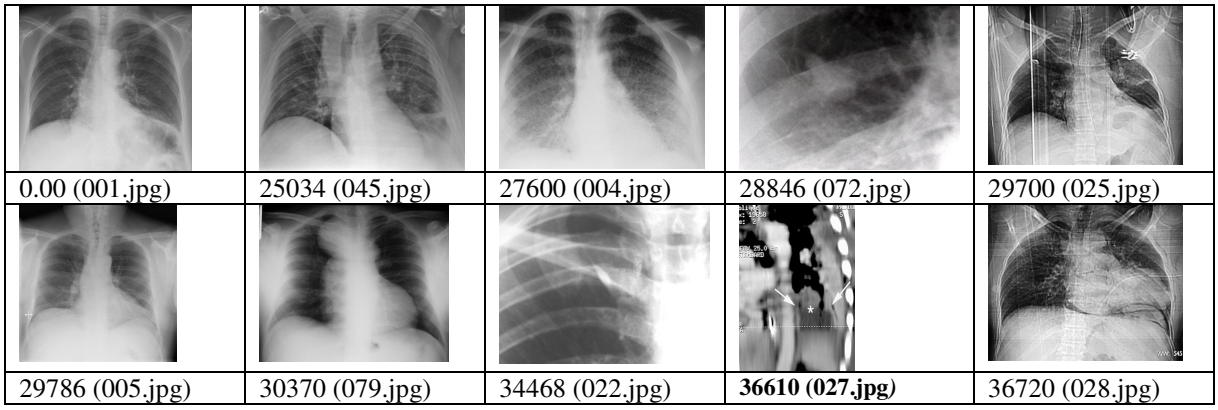


Figure 2. User interface of the developed CBIR system.



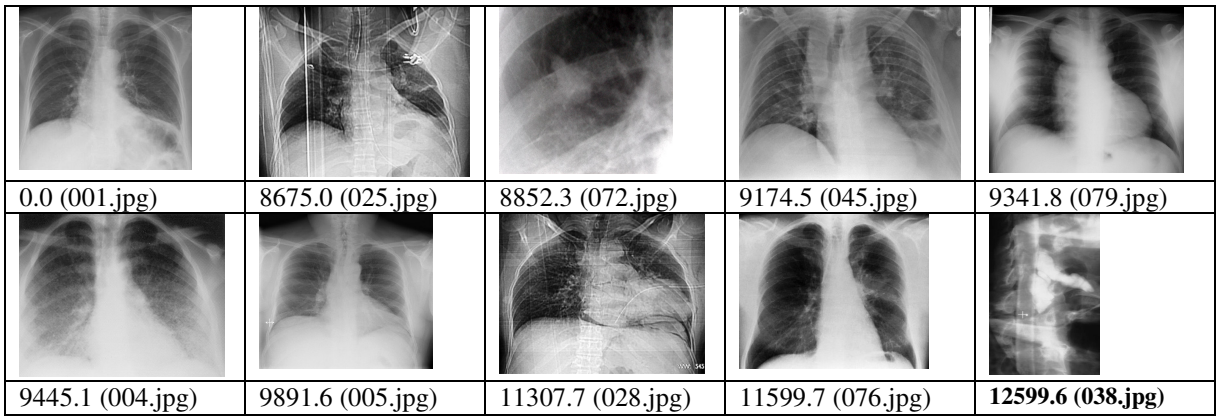
0.00 (001.jpg)	0.85 (045.jpg)	1.11 (025.jpg)	1.16 (004.jpg)	1.18 (079.jpg)
1.20 (072.jpg)	1.24 (022.jpg)	1.32 (005.jpg)	1.47 (021.jpg)	1.47 (023.jpg)

(a) Relative entropy sum distance



0.00 (001.jpg)	25034 (045.jpg)	27600 (004.jpg)	28846 (072.jpg)	29700 (025.jpg)
29786 (005.jpg)	30370 (079.jpg)	34468 (022.jpg)	**36610 (027.jpg)**	36720 (028.jpg)

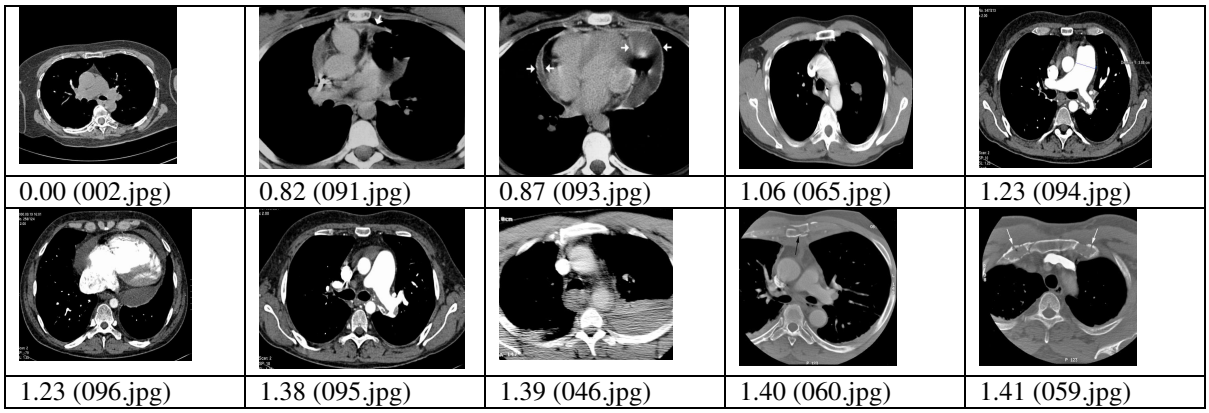
(b) L_1-norm histogram distance



0.0 (001.jpg)	8675.0 (025.jpg)	8852.3 (072.jpg)	9174.5 (045.jpg)	9341.8 (079.jpg)
9445.1 (004.jpg)	9891.6 (005.jpg)	11307.7 (028.jpg)	11599.7 (076.jpg)	**12599.6 (038.jpg)**

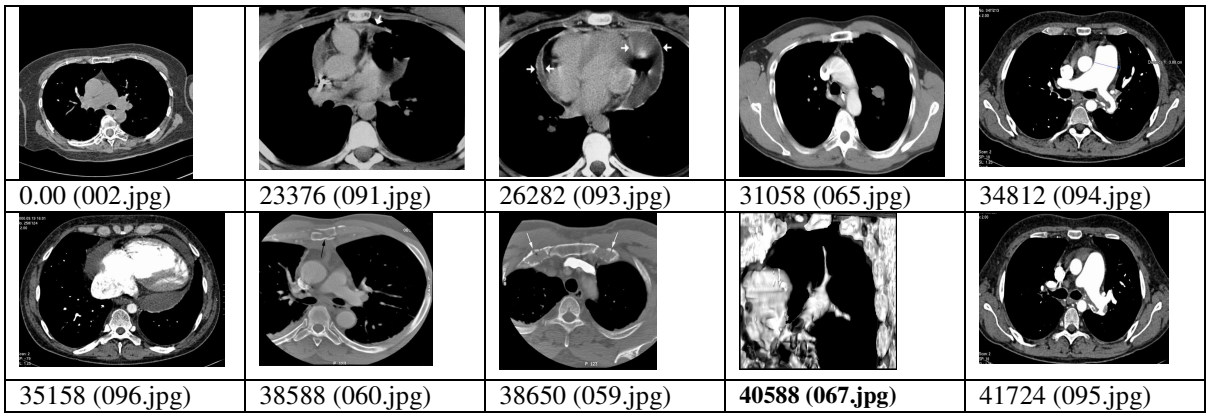
(c) L_2-norm histogram distance

Figure 3. Chest x-ray image retrieval example, where the query image is 001.jpg and histogram bin size is 16. The top ten retrieved images for each distance function are illustrated. Image labels are the distance function values and image filenames. (a) Relative entropy sum distance. (b) L_1-norm histogram distance. The ninth retrieved image is not a chest x-ray. (c) L_2-norm histogram distance. The tenth retrieved image is not a chest x-ray.



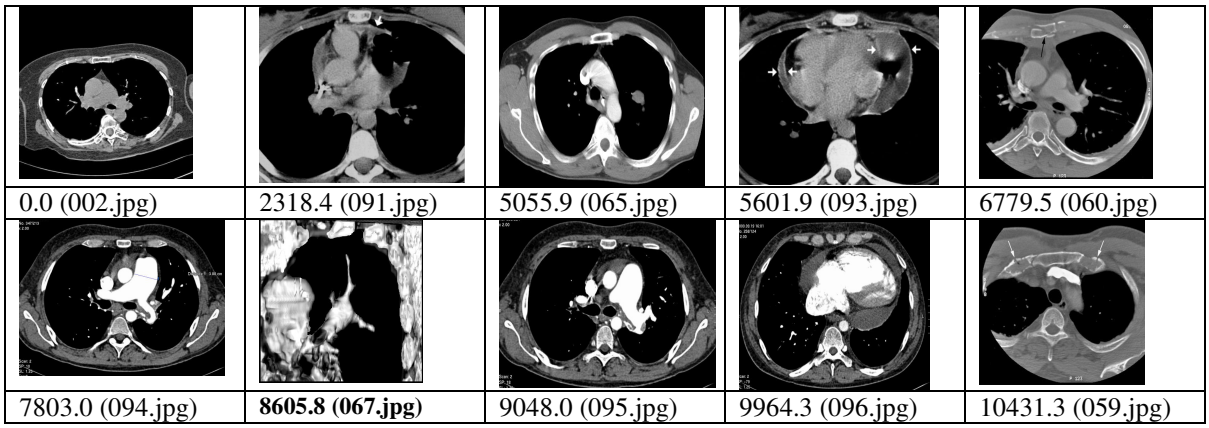
0.00 (002.jpg)	0.82 (091.jpg)	0.87 (093.jpg)	1.06 (065.jpg)	1.23 (094.jpg)
1.23 (096.jpg)	1.38 (095.jpg)	1.39 (046.jpg)	1.40 (060.jpg)	1.41 (059.jpg)

(a) Relative entropy sum distance



0.00 (002.jpg)	23376 (091.jpg)	26282 (093.jpg)	31058 (065.jpg)	34812 (094.jpg)
35158 (096.jpg)	38588 (060.jpg)	38650 (059.jpg)	**40588 (067.jpg)**	41724 (095.jpg)

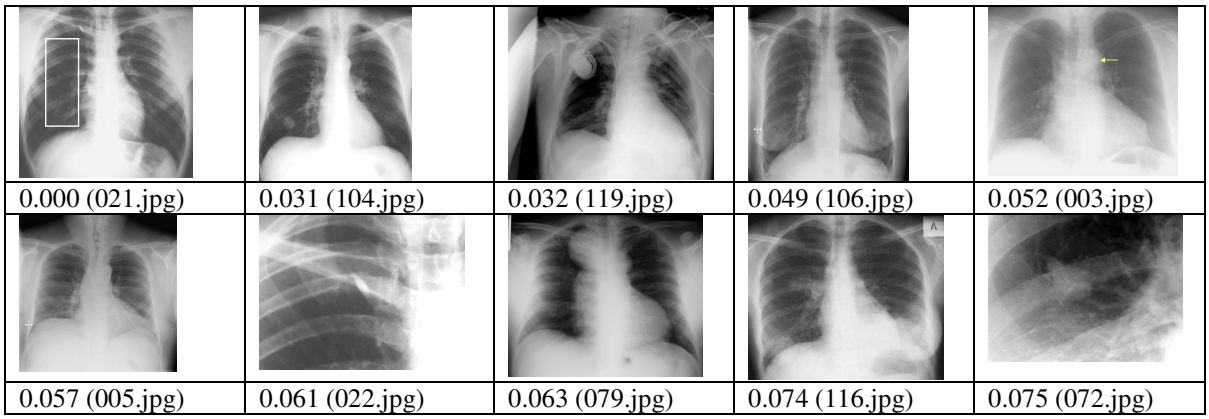
(b) L_1-norm histogram distance



0.0 (002.jpg)	2318.4 (091.jpg)	5055.9 (065.jpg)	5601.9 (093.jpg)	6779.5 (060.jpg)
7803.0 (094.jpg)	**8605.8 (067.jpg)**	9048.0 (095.jpg)	9964.3 (096.jpg)	10431.3 (059.jpg)

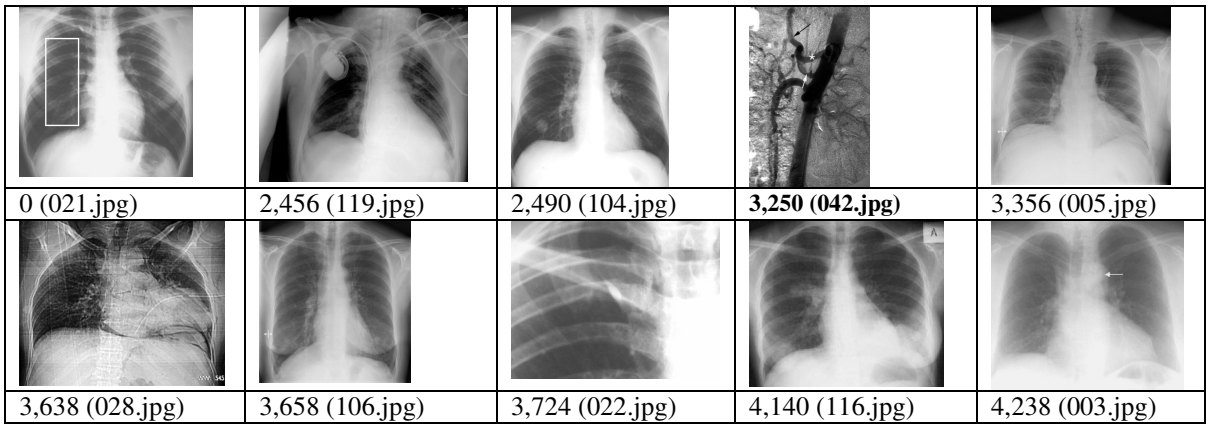
(c) L_2-norm histogram distance

Figure 4. Lung CT image retrieval example, where the query image is 002.jpg and histogram bin size is 16. The top ten retrieved images for each distance function are illustrated. Image labels are the distance function values and image filenames. (a) Relative entropy sum distance. (b) L_1-norm histogram distance. The ninth retrieved image is not a lung CT image. (c) L_2-norm histogram distance. The seventh retrieved image is not a lung CT image.



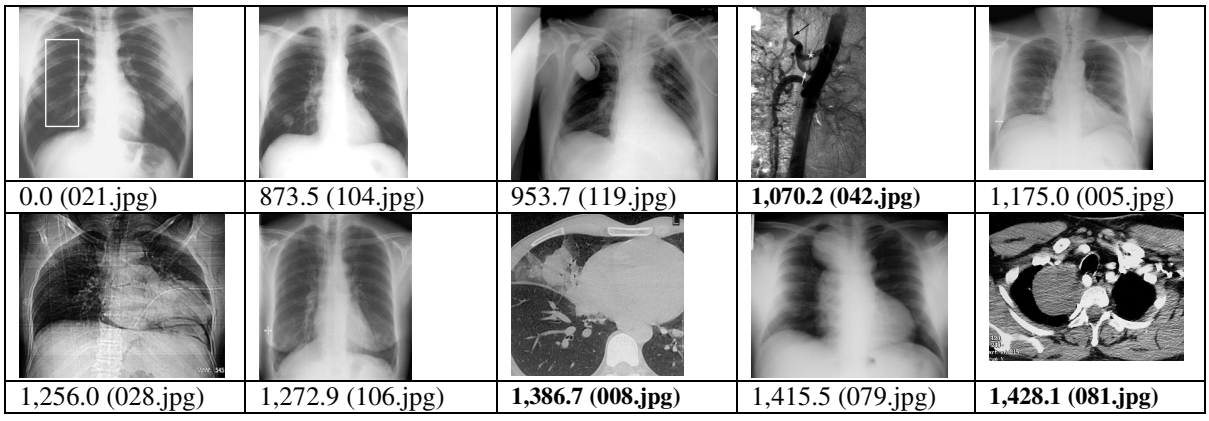
0.000 (021.jpg)	0.031 (104.jpg)	0.032 (119.jpg)	0.049 (106.jpg)	0.052 (003.jpg)
0.057 (005.jpg)	0.061 (022.jpg)	0.063 (079.jpg)	0.074 (116.jpg)	0.075 (072.jpg)

(a) Relative entropy sum distance



0 (021.jpg)	2,456 (119.jpg)	2,490 (104.jpg)	**3,250 (042.jpg)**	3,356 (005.jpg)
3,638 (028.jpg)	3,658 (106.jpg)	3,724 (022.jpg)	4,140 (116.jpg)	4,238 (003.jpg)

(b) L_1-norm histogram distance



0.0 (021.jpg)	873.5 (104.jpg)	953.7 (119.jpg)	**1,070.2 (042.jpg)**	1,175.0 (005.jpg)
1,256.0 (028.jpg)	1,272.9 (106.jpg)	**1,386.7 (008.jpg)**	1,415.5 (079.jpg)	**1,428.1 (081.jpg)**

(c) L_2-norm histogram distance

Figure 5. ROI-based image retrieval example, where the query is the marked rectangular rib region in the left side of image 021.jpg. Histogram bin size is 16. The top ten retrieved images for each distance function are illustrated. Image labels are the distance function values and image filenames. (a) Relative entropy sum distance. (b) L_1-norm histogram distance. The fourth retrieved image is incorrect. (c) L_2-norm histogram distance. The fourth, eighth and tenth retrieved image are incorrect.