International Journal of Soft Computing 9 (6): 386-390, 2014
ISSN: 1816-9503


# Content Based Image Retrieval for CT Images of Lungs

Jinsa Kuruvilla and K. Gunavathi
Department of ECE, PSG College of Technology, 641004 Coimbatore, India

**Abstract:** In this study, researchers present Content Based Image Retrieval (CBIR) System for Computed Tomography (CT) images of lungs. When a query image is given to the system, the system will retrieve the images which are similar to the query image from a database of cancerous and non-cancerous images. In CBIR Systems, the visual contents of the images in the database are extracted and stored as feature vectors to form a feature database. The Gray Level Co-occurrence Matrix (GLCM) parameters and statistical parameters are used to form the feature vectors. The parameters which are most relevant for retrieval process are found by artificial neural network classifier. Similarity measure plays an important role in CBIR Systems. The similarity comparison is done by different distance measures like Euclidean, Cityblock, Chebychev, Tversky, Manhattan, Canberra, Bray-Curtis, Squared Chord and Chi Squared. They calculate the similarities between the query image and images in the database. Different similarity measures have different effects in an Image Retrieval System. It is important to find the best similarity measure for CBIR System. The performance of the system is evaluated by Precision Rate (PR). The maximum retrieval rate obtained for cancerous images is 95% by GLCM parameters contrast and dissimilarity with modified Bray-Curtis distance.

**Key words:** Content based image retrieval, computed tomography, gray level co-occurrence matrix, artificial neural network, India

## INTRODUCTION

Lung cancer is the leading cause of cancer deaths in both women and men. Computer-Aided Diagnosis System is very helpful for radiologist in detection and diagnosing abnormalities earlier and faster than other screening programs. The computer aided diagnosis is a second opinion for radiologists before suggesting a biopsy test (Muller *et al*., 2004; Quellec *et al*., 2010). Researchers developed a content based image retrieval system which helps radiologist in identifying suspicious images by providing a visual comparison of a given image to a collection of similar images of known pathology (Bulo *et al*., 2011; Welter *et al*., 2012; Wong and Hsu, 2006). Few works based on content based image retrieval have been reported in the literature. Lam *et al*. (2007) developed a CBIR System based on Haralick, Gabor and Markov features. The distances used are Euclidean, Manhattan and Chebychev distances. The retrieval rate obtained is 88%. Wei *et al*. (2009) developed a CBIR System for mammogram images and the retrieval rate obtained is 82%. Wei *et al*. (2012) proposed a CBIR System based on Zernike moments for retrieval of mammogram images. The shape feature like compactness, fractional concavity and speculation index are used as features. The retrieval rate obtained is 90%.

De Oliveira *et al*. (2010) proposed a CBIR System using breast density patterns. The average precision rate is 90%.

## MATERIALS AND METHODS

The images are collected from a database of Lung Image Database Consortium (LIDC) and also from reputed hospitals. CT images of 180 patients are collected including both men and women. The average age of the patients considered is 64.2 years (age of the youngest patient is 18 years and the oldest patient is 85 years). Figure 1 shows the CBIR System used in this study.

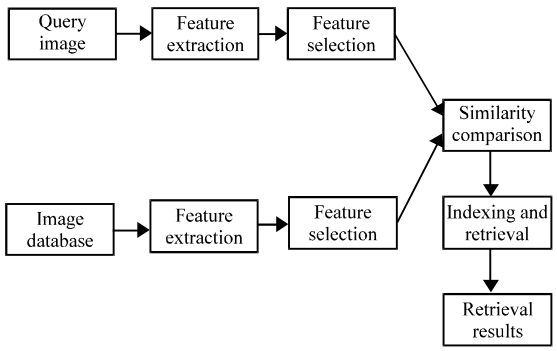


Fig. 1: CBIR System

**Corresponding Author:** Jinsa Kuruvilla, Department of ECE, PSG College of Technology, 641004 Coimbatore, India

**Feature extraction and feature selection:** Feature extraction is the basis and most important component of the CBIR System (Chun *et al*., 2008). GLCM features and statistical features are extracted from the images (Mabrouk *et al*., 2013). The GLCM features extracted are energy, entropy, dissimilarity, contrast, inverse difference, correlation, homogeneity, autocorrelation, cluster shade, cluster prominence, maximum probability, sum of squares, sum average, sum variance, sum entropy, difference variance, difference entropy, information measures of correlation, information measures of correlation, maximal correlation coefficient, Inverse difference normalized and Inverse difference moment normalized (Albregtsen, 2008; Kinoshita *et al*., 2007; Al-Kadi and Watson, 2008; Huang and Dai, 2003; Caicedo *et al*., 2011). The statistical parameters extracted are mean, standard deviation, skewness and kurtosis. The features which are relevant for retrieval are found by giving the parameters to classifier and finding the classification accuracy. The classifier used is artificial neural network classifier. The back propagation network with a new training function is used for classification (Park *et al*., 2004). In the training function, each variable is adjusted according to the gradient descent with momentum given by:

$$dX = 3.5 \times mc \times dXprev + lr \times (1 - mc) \times mc \times \frac{dpref}{dX} \quad (1)$$

Where:

dXprev = The previous change to the weight or bias
mc = Momentum constant
lr = Learning rate
dperf = The derivative of performance with respect to the weight and bias variables X

The classification results show that the relevant features from GLCM are autocorrelation, contrast, correlation, cluster shade, cluster prominence, dissimilarity, energy, entropy, homogeneity and sum variance. The relevant feature from statistical features is skewness.

**Similarity measures:** Selection of similarity measures has a direct impact on the performance of a CBIR System (Arevalillo-Herraez *et al*., 2008). The retrieval result is not a single image but a list of images ranked by their similarities with the query image. The similarity measures used are:

**Euclidean distance:** Euclidean distance between two images is the square root of the sum of the squares of the differences between the feature values. The Euclidean distance E(p, q) between points p and q, if p = (p1, p2, ..., pn) and q = (q1, q2, ..., qn) is given by:

$$E(p, q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \quad (2)$$

**Manhattan distance:** The Manhattan distance between two images is the sum of the differences of their corresponding features. The Manhattan distance M(p, q) between points p and q, if p = (p1, p2, ..., pn) and q = (q1, q2, ..., qn) is given by:

$$M(p, q) = \sqrt{\sum_{i=1}^{n} |p_i - q_i|} \quad (3)$$

**City Block distance:** The City Block distance CB(p, q) between points p and q, if p = (p1, p2, ..., pn) and q = (q1, q2, ..., qn) is given by Mukhopadhyay *et al*. (2013):

$$CB(p, q) = \sum_{i=1}^{n} |p_i - q_i| \quad (4)$$

**Chebychev distance:** The Chebychev distance CH(p, q) between points p and q, if p = (p1, p2, ..., pn) and q = (q1, q2, ..., qn) is given by:

$$CH(p, q) = \max_{i=1 \text{to} n} |p_i - q_i| \quad (5)$$

**Tversky distance:** The Tversky distance TD(p, q) between points p and q, if p = (p1, p2,..., pn) and q = (q1, q2,..., qn), is given by:

$$TD(p, q) = \frac{p_i - q_i}{(p_i - q_i) + 2(\min(p_i, 1 - q_i)) + 2(\min(1 - p_i, q_i))} \quad (6)$$

**Canberra distance:** The Canberra distance CD(p, q) between points p and q, if p = (p1, p2, ..., pn) and q = (q1, q2, ..., qn) is given by:

$$CD(p, q) = \sum_{i=1}^{n} \frac{|p_i - q_i|}{|p_i + q_i|} \quad (7)$$

**Bray-Curtis:** It is a statistic used to quantify the compositional dissimilarity between the two sets. The Bray-Curtis distance BC(p, q) between points p and q, if p = (p1, p2, ..., pn) and q = (q1, q2, ..., qn), is given by:

$$BC(p, q) = \sum_{i=1}^{n} \frac{(p_i - q_i)}{(p_i + q_i)} \quad (8)$$

**Chi Squared distance:** The Chi-Squared distance CS(p, q) between points p and q, if p = (p1, p2, ..., pn) and q = (q1, q2, ..., qn) is given by:

$$CS(p, q)=\sqrt{\sum_{i=1}^{n}\left(\sqrt{p_i}-\sqrt{q_i}\right)^2} \quad (9)$$

**Squared Chord distance:** The Squared Chord distance SC(p, q) between points p and q, if p = (p1, p2, ..., pn) and q = (q1, q2, ..., qn) is given by Cai *et al*. (2000).

$$SC(p,q)=\sum_{i=1}^{n}\left(\sqrt{p_i}-\sqrt{q_i}\right)^2 \quad (10)$$

## RESULTS AND DISCUSSION

The retrieval results of 10 GLCM parameters namely autocorrelation, contrast, correlation, cluster shade, cluster prominence, dissimilarity, energy, entropy, homogeneity and sum variance and the statistical parameter skewness using different distance measures are analyzed. The performance is measured by precision rate (Ayyachamy and Manivannan, 2013). The Precision Rate (PR) or retrieval rate is given by:

$$\text{Precision rate} = \frac{\text{Number of relevant images retrieved}}{\text{Number of image retrieved}} \quad (11)$$

The results show that Euclidean, Cityblock, Chebychev and Manhattan distance give the same precision rate for the parameters. Squared Chord and Chi-Quared distance also give the same precision rate for the parameters. So, the comparison is done between Euclidean, Squared Chord, Canberra, Bray-Curtis and Tversky distances. The retrieval rate for skewness is less compared to GLCM parameters. The maximum retrieval rate of 92.5% is obtained for contrast and dissimilarity by using Bray-Curtis distance. The performance of the five distance measures are shown in Table 1.

The retrieval rate is increased to 95% by modified Bray-C urtis distance. The modified Bray-Curtis distance is given by:

$$MBC(p, q)=\sum_{i=1}^{n}\frac{\left(p_i-q_i\right)^2}{\left(p_i+q_i\right)^2} \quad (12)$$

The query image 1 and the retrieved images for contrast using modified Bray-Curtis distance is shown in Fig. 2 and 3 (for clarity only six retrieved images are shown).

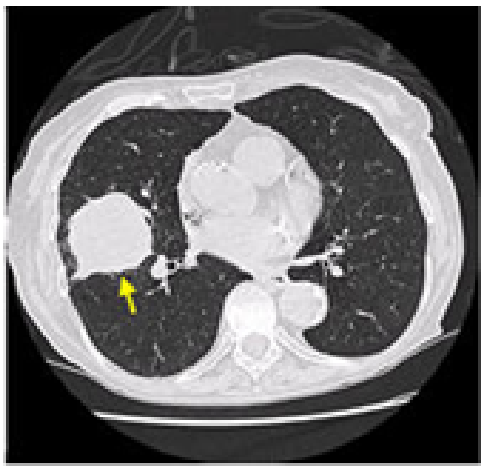

Fig. 2: Query image 1

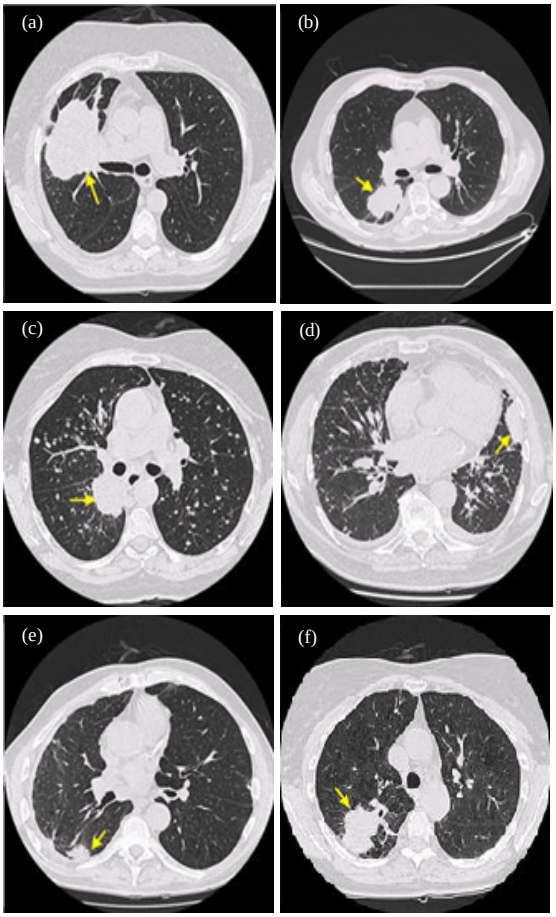


Fig. 3: Retrieved images of query image 1 with contrast by modified Bray-Curtis distance. a) Image 98, b) image 69, c) image 100, d) image 94, e) image 97 and image 23

Fig. 4: Query image 2

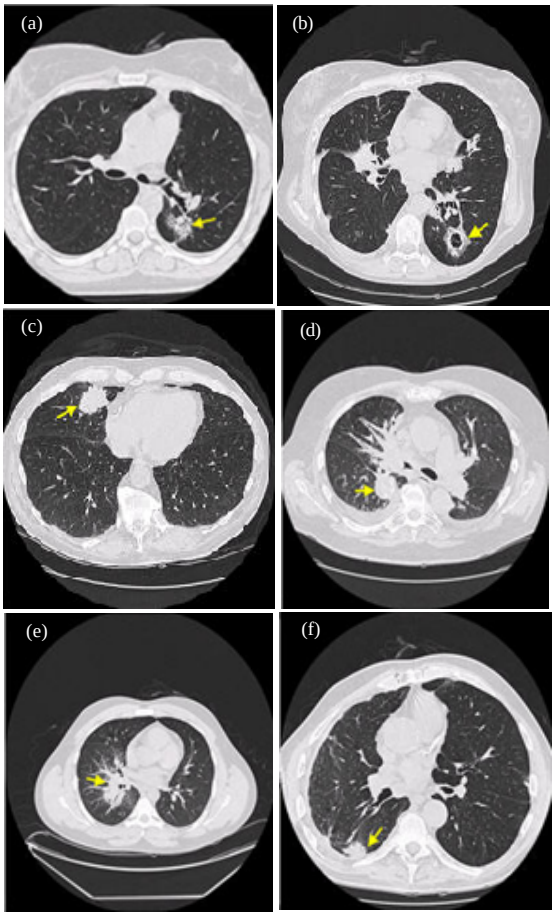


Fig. 5: Retrieved images of query image 2 with dissimilarity by modified Bray-Curtis distance. a) Image 15, b) image 36, c) image 43, d) image 65, e) image 60 and f) image 97

Table 1: Performance of different distance measures

| Distance measure | Average precision rate (%) |
|---|---|
| Euclidean | 88.0 |
| Squared Chord | 86.5 |
| Canberra | 89.5 |
| Bray-Curtis | 92.5 |
| Tversky | 91.0 |

The query image 2 and the retrieved images for dissimilarity using modified Bray-Curtis distance is shown in Fig. 4 and 5.

## CONCLUSION

In this study, content based image retrieval system is proposed which helps in computer aided diagnosis for lung cancer. The GLCM parameters and statistical parameters are used as feature vectors for retrieval process. Out of 22 GLCM parameters, the relevant 10 parameters are selected artificial neural network classifier. Skewness is the statistical parameter selected by the classifier for retrieval. Compared to GLCM parameters, the retrieval rate of statistical parameters are less. For retrieving cancerous images the average precision rate obtained is 95% with the parameters contrast and dissimilarity using modified Bray-Curtis distance.

## REFERENCES

Al-Kadi, O.S. and D. Watson, 2008. Texture analysis of aggressive and non aggressive lung tumor CE CT images. IEEE Trans. Biomed. Engin., 55: 1822-1830.

Albregtsen, F., 2008. Statistical texture measures computed from gray level coocurrence matrices. Technical Note, November 5, 2008. http://www.uio.no/studier/emner/matnat/ifi/INF4300/h08/undervisningsmateriale/glcm.pdf.

Arevalillo-Herraez M., J. Domingo and F.J. Ferri, 2008. Combining similarity measures in content-based image retrieval. Pattern Recogn. Lett., 29: 2174-2181.

Ayyachamy, S. and V.S. Manivannan, 2013. Distance measures for medical image retrieval. Int. J. Imag. Syst., 23: 9-21.

Bulo, S.R., M. Rabbi and M. Pelillo, 2011. Content-based image retrieval with relevance feedback using random walks. Pattern Recogn., 44: 2109-2122.

Cai, W., D. Feng and R. Fulton, 2000. Content-based retrieval of dynamic PET functional images. IEEE Trans. Inform. Technol. Biomedic., 4: 152-158.

Caicedo, J.C., F.A. Gonzalez and E. Romero, 2011. Content-based histopathology image retrieval using a kernel-based semantic annotation framework. J. Biomed. Inform., 44: 519-528.

Chun, Y.D., N.C. Kim and I.H. Jang, 2008. Content-based image retrieval using multiresolution color and texture features. IEEE Trans. Multimed., 10: 1073-1084.

De Oliveira, J.E.E., A.M.C. Machado, G.C. Chaveza, A.P.B. Lopes, T.M. Deserno and A.D.A. Araujo, 2010. MammoSys: A content-based image retrieval system using breast density patterns. Comput. Methods Programs Biomedic., 99: 289-297.

Huang, P.W. and S.K. Dai, 2003. Image retrieval by texture similarity. Pattern Recogn., 36: 665-679.

Kinoshita, S.K., P.M. de Azevedo-Marques, R.R. Pereira Jr., J.A.H. Rodrigues and R.M. Rangayyan, 2007. Content-based retrieval of mammograms using visual features related to breast density patterns. J. Digital Imaging, 20: 172-190.

Lam, M.O., T. Disney, D.S. Raicu, J. Furst and D.S. Channin, 2007. BRISC-an open source pulmonary nodule image retrieval framework. J. Digit. Imag., 20: 63-71.

Mabrouk, M., A. Karrar and A. Sharawy, 2013. Support vector machine based computer aided diagnosis system for large lung nodules classification. J. Med. Imaging Health Inform., 3: 214-220.

Mukhopadhyay, S., J.K. Dash and R. Das Gupta, 2013. Content-based texture image retrieval using fuzzy class membership. Pattern Recogn. Lett., 34: 646-654.

Muller, H., N. Michoux, D. Bandon and A. Geissbuhler, 2004. A review of content-based image retrieval systems in medical applications-clinical benefits and future directions. Int. J. Med. Inform., 73: 1-23.

Park, S.B., J.W. Lee and S.K. Kim, 2004. Content based image classification using neural network. Pattern Recogn. Lett., 25: 287-300.

Quellec, G., M. Lamard, G. Cazuguel, B. Cochener and C. Roux, 2010. Wavelet optimization for content-based image retrieval in medical databases. Med. Image Anal., 14: 227-241.

Wei, C.H., S.Y. Chen and X.H. Liu, 2012. Mammogram retrieval on similar mass lesions. Comput. Methods Programs Biomedic., 106: 234-248.

Wei, L., Y. Yang and R.M. Nishikawa, 2009. Microcalcification classification assisted by content-based image retrieval for breast cancer diagnosis. Pattern Recognit., 42: 1126-1132.

Welter, P., B. Fischer, R.W. Gunther and T.M. Deserno ne Lehmann, 2012. Generic integration of content-based image retrieval in computer-aided diagnosis. Comput. Methods Programs Biomed., 108: 589-599.

Wong, W.T. and S.H. Hsu, 2006. Application of SVM and ANN for image retrieval. Eur. J. Operat. Res., 173: 938-950.