New Discrete Cosine Transformation Algorithm in Image Retrieval

Dai Yin*[12], Xiao-wen Chen[2],Yu-yi Wang[2], Yun-long Zheng[2]

1. College of information and science
2. school of Sino-Dutch Biomedical &Information Engineering
Northeastern University
Shenyang, 110000, China
E-mail: daiyin@bmie.neu.edu.cn

Abstract-To improve the efficacy of traditional medical image retrieval, this paper presents a new medical algorithm that uses discreet cosine transformation innovatively. This paper chose 42 pictures from upper torso CT images obtained from patients with hepatocellular carcinoma in ShengJing Hospital to serve as databases. Then by comparing the retrieval results, 6 pictures' similarity to the given query image, gotten by different retrieval algorithms to judge the efficacy of different algorithms. In the results, DCT algorithms display better retrieval performance than traditional grayscale histogram cumulative algorithm, extracting more similar images obviously. If scholarships overcome some drawbacks of this algorithms, this technology would have promising prospect to be applied for medical system in hospital.

Keywords-content based retrieval; medical image retrieval; grayscale histogram; DCT; similarity ;

I. INTRODUCTION

Medical retrieval technology is an approach that uses existed features to go on medical image retrieval. Nowadays, most researches concerning medical retrieval focus on content-based medical image aspect. As a kind of special image, medical image has its own unique characteristics: high gray resolution, high spatial resolution, similar images and so on. Thus, the common extracting features like texture, shape to retrieve images is useless.

Discrete Cosine Transformation (DCT) theory, which is brought up in 1974, is applied to signal process generally, being regarded as the best transformation on dealing with image and sound signal. Compared with Discrete Fourier Transformation (DFT), DCT can simplify calculation and delete imaginative part. In this paper, it used an image energy-based retrieval method. Compared to traditional grayscale histogram algorithm, energy-concentrating DCT can extract the features based on the energy distribution of images, by doing so, there is no need to think about the process of the imaginary in the result. It also decrease the deviation when extracting the original features of the images, getting the better effect.

II. IMAGE RETRIEVAL BASED ON CONTENT

In recent years, the number of digital images all over the world has been increasing sharply and quickly. And image retrieval technology becomes a very active research field. On the early stage of image retrieval, it depends on the word and comment to retrieve without thinking of the information that the image brought itself. Besides, it lacks the standard, generally accepted judge language. So in the late 19th century, scholarships shifted their focus to content-based image retrieval.

The process of content-based image retrieval can be divided into several steps: detecting the image, extracting the quantitative features of image, finding the images that match the original image, achieving image retrieval finally.

The features of image mainly are distinguished into two categories, physical features and semantic features[1]-[6] In most cases, it's physical features that are used widespread in content-based image retrieval. They include color, texture, shape, spatial relation, etc. Compared to the understanding of humanity to images, all physical features are extracted according to the primitive data and they are actually relatively basic features. Semantic characteristic possesses more meaning in philosophy sense. This also involves debate about judgment of subjectivity and objectivity. Besides, there is no clear standard for the description of semantic characteristic.

III. TRADITIONAL ALGORITHM IN MEDICAL IMAGE RETRIEVAL

A. Basic principle

In medical image retrieval, given the images are grayscale image, it's normal to use grayscale as the feature to retrieve relevant images. And the most basic and common approach is traditional grayscale histogram algorithm[7], which is to go through every pixel in the image, clarify its gray value and calculate the number of respective gray value. The number is shown as the form of histogram. Finally to calculate the Euler distance value based on the histogram to get similarity. Every value in grayscale histogram represents the number of pixels which have the same grayscale value. Grayscale histogram can avoid the influence brought by images' rotation, size transformation and translation.

B. Drawbacks of traditional histogram cumulative algorithm

Apparently, traditional algorithm has many drawbacks that will worsen retrieval effect. For example, cumulative grayscale histogram neglects totally the spatial relation of images, only noting the sum of pixels that possess the same grayscale value. This will cause a possibility that two clearly different images being clarified as the same,

just like the example of Fig ..1. Furthermore, the intensity of image will effect grayscale value. The same object's histogram with different intensity will make difference.

Figure 1. Different images with the same similarity

IV. DCT ALGORITHM BASED ON SEGMENTED GRAYSCALE HISTOGRAM ALGORITHM

Because of the disadvantages of traditional cumulative grayscale histogram algorithm, this paper uses segmentation as pre-procession, it can filter some images that have dramatic differences (like images above), then using DCT algorithm as retrieval approach to get the similarity. The calculation of similarity is by Euler distance function. The equation of Euler distance:

$$D(I,J) = \left(\sum_j |f_i(I) - f_i(J)|^2 \right)^{\frac{1}{2}} \tag{1}$$

$D(I,J)$ is the distance between the given query image and image in databases, $f_i(I)$ is the feature value of the component of feature vector, which represents image I as a whole.

A. Segmented histogram cumulative algorithm

Considering the ignorance of traditional cumulative grayscale histogram to spatial features, in this paper, segmentation was used as pre-procession. The step was firstly equally dividing the image into 4 parts, then respectively calculating the grayscale histogram, finally getting average value of 4 parts' similarity as the final similarity. This algorithm, in a way, take spatial relation into consideration, improving traditional algorithm. For the evaluation of similarity, in order to overcome the influence of the intensity of images to retrieval result, this paper adopt percentage-most principle, using the percentage of pixels out of the whole pixels to represent the similarity.

The equation of similarity:

$$Sim(M,N) = \sum_i^N \frac{\min(P_i(N), P_i(M))}{\max(P_i(N), P_i(M))} \times \frac{P_i(N)}{\sum P_i(N)} \tag{2}$$

$Sim(M, N)$ is the similarity of M and N, $P_i(N)$ is the number of pixels whose gray value is i. Just as reminder, these similarities presents the respective similarities after segmentation. The total similarity is the average value of 4 sub-similarity [8].

DCT algorithm

Discrete Cosine Transformation [9-10] shares many similarities with Discrete Fourier Transformation. In 1974, Ahmed and Rao firstly propose DCT. As known to all, the DCT of real even function is real function. The property of DCT is to turn function into even function, cutting out the imaginary in frequency domain. Thus, the length of result after DCT is much shorter than the length after DFT, greatly declining calculation amount. So does the time. What's more, the DFT of real signal possess conjugate symmetry, so there will be half data redundancy in the result. Discrete Cosine Transform is defined as a transformation of real signal. The consequence is also a real signal in the frequency domain obtained by converting. As compared with DFT, DCT enable calculation to be reduced by more than half. DCT has also a very important property (characteristic energy concentration): Most natural signals (sound, images)'s energy concentrated in the low frequency portion after a discrete cosine transform. Therefore, DCT has been widely used in the (voice, image) data compression. Since the DCT is derived from DFT transform, thus many of the properties of the DFT are still retained in DCT. In image processing, each picture will be cut into small pieces 8×8 block and size can be random. But for historical reasons people used to cut blocks of 8×8 [11]. Two-dimensional image processing and one-dimensional signal processing's principles are the same, and some formulas are not the same. In two-dimensional image, DCT has eight kinds of types, of which four are common. A one-dimensional discrete cosine transform (DCT-2 type) is the following formula:

$$\hat{X}[n] = \sqrt{\frac{N}{2}} \sum_{m=0}^{n-1} X[m] \cos\left(\frac{\pi}{n}\left(m + \frac{1}{2}\right)n\right) \tag{3}$$

$X[m]$ is one-dimensional function, $\hat{X}(n)$ is frequency function.

The DCT of two-dimensional image is the same in terms of principle with the one-dimensional images', it is:

$$G(u,v)$$
$$= c(u)c(v)\frac{2}{\sqrt{MN}} \sum_{n=0}^{N-1}\sum_{m=0}^{M-1} f[x,y]\left[\frac{\pi u(2x+1)}{2M}\right]\cos\left[\frac{\pi v(2y+1)}{2N}\right]$$
$$c(u) = \begin{cases} \frac{1}{\sqrt{2}} & u = 0 \\ 0 & others \end{cases} \tag{4}$$

The advantages of DCT is that you don't need to consider the imaginary part any more, decreasing the deviation of results as well as making the features more accurate and better represents the image. The process of the retrieval algorithm is to obtain the pointer of the set of pixels, then manage DCT-2 mode transformation. The result set is obtained by getting Euclidean distance and take the similarity between the images as the average value (the smaller the closer to the original image), then sorting according to similarity.

.

B. Retrieval algorithm result

During the Exploration and research on algorithms, this paper used upper torso CT images obtained from patients with hepatocellular carcinoma in Shengjing Hospital in Shenyang. And a series of experiments were

carried out to assess the retrieval algorithm effects. This paper chose medical image mentioned before to do comparison experiments. On the condition of controlling all variables this paper conducted retrieval experiments about both traditional histogram cumulative algorithm, DCT algorithm. First, the specific process is to upload 42 medical images mentioned above into database. Second, to calculate each image's similarity according to all kinds of algorithm. Finally, to assess these algorithm's efficacy according to outputting the 6 images that possess the highest similarities.

The results of experiments are listed as Fig .2, Fig .3.
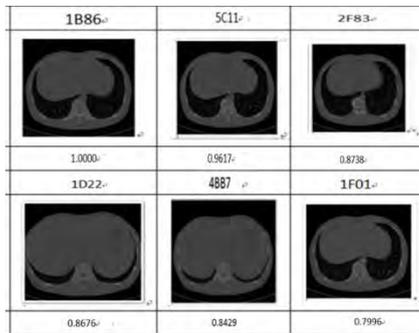

Figure 2. Retrieval result of traditional histogram algorithm
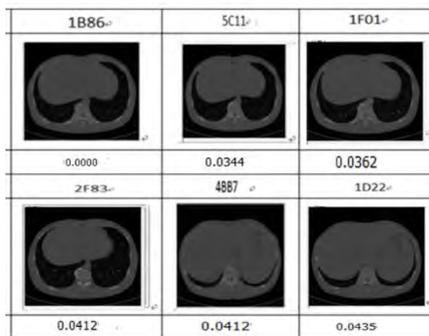

Figure 3. Retrieval result of DCT algorithm

Because of the difference of similarity calculation method between traditional histogram algorithm and DCT algorithm, similarity can't be compared between each other. But the results can be clearly seen in the figure—the effect of DCT retrieval algorithm is better than using only histogram retrieval. The algorithm that firstly to set 0.75 as a threshold value to filter irrelative images, and then the application of DCT greatly saved the time required for retrieval and computer memory. Seeing from the results, you can find discrete cosine transformation algorithm for image evaluation is very close to human criteria and more suitable than cumulative grayscale histogram. Therefore, the system which uses two modes retrieval can not only improve the efficiency, but also improve the speed of retrieval. The secondary mode is successful.

C. Evaluation of retrieval algorithms

The assessments of retrieval algorithms also have a variety of ways. Some researchers only take the screenshot of retrieval results as the algorithm evaluation methods, which largely is not objective and likely to be the best results chosen by author.

The common assessment indicators are sensitivity and specificity.

$$sensitivity = \frac{marked\ positive\ elements}{all\ elemets} \quad (5)$$

$$specificity = \frac{marked\ irrelative\ elements}{all\ elemets} \quad (6)$$

These two values can also show receiver operating characteristic curve (ROC) which contain more useful information for recipients to express. Another common value is accuracy: Besides, no one can deny the importance of speed for retrieval algorithm, especially for dealing with large amount of data.

$$accuracy = \frac{marked\ positive\ elements}{all\ marked\ elemets} \quad (7)$$

Based on the experimental results of Fig .2 and Fig .3, several algorithms mentioned above were evaluated about their retrieval performance. Standard assessments of algorithms is the three indicators mentioned in the retrieval algorithm performance evaluation, namely specificity, accuracy, and sensitivity. See detailed definition in (5), (6), (7).

For the assessments of algorithms mentioned above, this paper has had experiments. Images in experiments are 42 standard DICOM images, of which there are 6 similar images retrieved. The results are recorded below.

TABLEI ASSESSMETN RESULTS

	Specificity	Accuracy	Sensitivity
Traditional grayscale histogram	60.7%	38.4%	52.5%
DCT algorithm	94.6%	66.7%	80.0%

As seen, in terms of specificity, accuracy or sensitivity, DCT algorithm is much better than traditional grayscale histogram algorithm.

V. CONCLUSION

In today when a large number of cutting-edged medical technologies is appearing and getting used, the number of medical images exponentially grows, social's need for medical image retrieval and its system are also increasing. In order to quickly and efficiently obtain relevant images in the mass of medical images, medical image retrieval technology is constantly improving and making innovation, and continue to expand their own application scope and potential. Under this background, the paper launched research on medical image retrieval techniques, carrying out an innovative application, as well as practical design retrieval system, and achieved good results. This paper completed the following tasks:

1. For the cumulative grayscale histogram algorithm, since this algorithm will ignore image's spatial relation. On the basis of the original algorithm, firstly take pre-segmentation, and finally get the average similarity of the sub-block as the total similarity. The new algorithm to some extent, improve the accuracy of the original algorithm, improving the similarity of similar images, decreasing the similarity of non-similar image.
2. For the overall design of the search algorithm, this paper use two algorithm modes. Namely, using segmentation cumulative grayscale histogram as

first mode, DCT as second mode, enhancing the overall accuracy and it's needless to handle the imaginary part of DFT. For discrete cosine transform' application in medical image retrieval, It's an innovation.

Although the basic function of system and progression of algorithm have been achieved, there are still many places necessarily to improve:

1. DCT algorithm's computing speed is still very slow. On the basis of a fast algorithm has been applied, a rapid response is still not done. The waiting time is long. How to reduce time is one of the focus of future research.

2. For a large number of image processing this system still cannot deal with. In practice, the system of medical image retrieval must apply to handle great amount of data, how to design the system so that it can quickly and efficiently handle enormous data is an important direction for future research.

3. In the course of the assessment for the algorithm, the paper found in the existing evaluation index still contains too many subjective factors, such as the decision for similar images and so on. These factors will directly affect the value of the algorithm evaluation indicators, thus further affecting the researchers to evaluate the algorithm. Therefore, further research and exploration of algorithm assessment tools are also very necessary.

REFERENCES

[1] G. Eason, B. Noble, and I. N. Sneddon, ―On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)* Syam, B. Victor, J.S.R. , Rao, Y.S. ―Efficient similarity measure via Genetic algorithm for content based medical image retrieval with extensive features," IEEE international conference, Kottayam, March 2013,pp. 704-711.

[2] J.P. Eakins, M.E. Graham, ―Content-based image retrieval," Tech. Rep. JTAP-039, JISC Technology Application Program, Newcastle upon Tyne, 2000.

[3] Muller H, Michoux N, Bandon Duet al. A review of content-based image retrieval systems in medical applications-clinical benefits and future directions [J]. Medical Informatics, 2004, 73(1): 1-23.

[4] Henning Müller, Nicolas Michoux, David Bandon, et al. A review of content-based image retrieval systems in medical applications——clinical benefits and future directions [J]. International Journal of Medical Informatics, 2004, 73: 1-23.

[5] Payel Ghosh, Sameer Antani, L. Rodney Long, George R. Thoma, ―Review of Medical Image Retrieval Systems and Future Directions" [J], Computer-Based Medical Systems (CBMS). 2011

[6] J. S. Weszka, C. R. Dyer, A. Rosenfeld, ―A comparative study of texture measures for terrain classification," IEEE Trans. Sys. Man Cybernetics 6 (4) (1976) 269-285.

[7] Li Shun-shang, Zhuang Tian-Ge, Chen Hui―Medical image retrieval based on Gray Histogram and Correlation," Shanghai Jiaotong University Newspaper, 2001, 35(5): pp695-698.

[8] Rahman M M, Bhattacharya P, Desai B C. A framework for medical image retrieval using machine learning and statistical similarity matching techniques with relevance feedback [J]. IEEE Transactions on Information Technology in Biomedicine, 2007, 11(1): 58-69.

[9] Lu Ye-Ping, Li Feng-Ting, Chen Zhao-Long, ―Research on discrete cosine transformation coding and its present and future," Journal of China Institute Of Communications, 2004, 25（2）: 106-117

[10] K. Rajakumar, Dr. S. Muttan. ―Medical image retrieval using modified DCT [J]. Proceida Computer Science, 2010, 2: 298–302.

[11] Rafael C. Gonzalez, Richard E. Woods, Steven L. Eddins. Digital Image Processing Using MATLAB [M]. Beijing: Publishing House of Electronics Industry, 2004. R. Nicole, ―Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.