Lung Cancer Diseases Diagnostic Asistance Using Gray Color Analysis

Paulus
Graduate Program in Informatics Engineering,
Department of Computer Science,
Bina Nusantara University,
Jakarta, Indonesia.
aizu_120186@yahoo.com

Ford Lumban Gaol
Faculty of Computer Science,
Bina Nusantara University.
Jakarta, Indonesia.
fgaol@binus.edu

Abstract— **Errors in diagnosing the disease is a critical risk that must be faced by any person giving treatment to the hospital. Medical treatment can not always be done with perfect accuracy. Lung cancer is one of the most deadly disease that prone to misdiagnose. In general, some practitioners tend to "read" cancer in x-ray rontgen image as tumor this could be fatal. To generate a diagnose, a general practitioner use three kind of examination i.e : patient History, Radiologic examination, phisical examination. In this paper, Gray color for image indexing and retrieval are investigated. The features are derived based on the statistical distribution of Harralick feature from image sample. By utilizing the proposed invariant features, the similarity measure between query and database images provides reliable retrieval results.**

.**Keywords-component; Harralick feature, Gray color analysis, Correlogram Method, Gray Level Co-Matrix**

I. INTRODUCTION

Errors in diagnosing the disease is a critical risk that must be faced by any person giving treatment to the hospital. Medical treatment can not always be done with perfect accuracy (100%). This is due to certain medications standard that set by each medical staff. There are many times the standard is not uniform. There is sometimes the medical staff, in this case the doctor, gives the wrong diagnosis to patients and this could be fatal [9]. In addition errors in diagnosis can also be caused by medical doctor bias. Medical doctor bias can occur when a doctor gives a diagnosis of certain diseases where the diagnosis is made has a tendency that led to the diagnosis of a disease that is often given by the doctor. This is fatal because of errors in diagnosing the disease can lead to errors in the provision of healing methods and treatments will also be given to the patient.

According to data that obtained from the Patient Safety Incident, the number of patients who experienced an error in diagnosis is one of the 155 patients who were hospitalized in 1000 hospitals [10]. There are five types of the top diseases that are often experienced failure in terms of diagnoses including: myocardial infarction, Breast Cancer, Lung Cancer - Lung, appendisities, Colon Cancer [7].

Lung cancer - Lung disease is the most frequent errors in diagnosis [5]. Errors in diagnosis generally occurs when:

- Failure to know the symptoms of lung cancer - lung.
- Diagnose a tumor as benign.
- Error in reading lab results.

Lung cancer is one of the most deadly disease that prone to misdiagnose. In general, some practitioners tend to "read" cancer in x-ray rontgen image as tumor this could be fatal. To generate a diagnose, a general practitioner use three kind of examination i.e : patient History, Radiologic examination, phisical examination [1].

Graycomatrix will be used in this paper to diagnose Lung cancer disease using radiologic examination. In this case we use X ray Rontgen.. The ability to indexing and retrieve gray color of an image is the main reason we use this technique since the x-ray rontgen color appears to be gray. X-ray rontgen is one of many radiologic examination that used to generate a supporting diagnose from certain disease.

In this study, the authors propose the application of the method to determine image Correlogram lung cancer based on a comparison between the results of x-rays with an existing database to diagnose the disease.

II. THE CONCEPT OF CORRELOGRAM METHOD

Correlogram color of an image is a table that indexed by using color pairs where *k* is the *h* is an initial point of entry (*i, j*) is to specify the possible discovery of pixels of color *j* at distance *k* from pixel color *i* of the image in question [4]. Features of such images provides a high level of tolerance on the change in the image of the scheme or the same topic due to the viewpoints of different shots [4].

Correlogram gives an idea of how the relationship of information with the distance from the pair of color-varying distance. Assume that *I* is the *n*x*n* sized pictures, the colors of *I* we quantify into *m* color $c_{1,}c_2c_{1,}c_2$, ..., c_mc_m. *m* is the number constant. For the pixel, *I*(*p*) will declare its color so as to produce equality.

We will use the - norm to calculate the distance between pixels, for pixels

$$p = (x,y) \epsilon\, Ip = (x,y) \epsilon\, I$$

I(*p*) will be set into:

$$I_c \triangleq \{p \mid I(p) = c\}.$$

978-0-7695-4262-1/10 $26.00 © 2010 IEEE
DOI 10.1109/CIMSiM.2010.79

Therefore notation $p \in I, I(p) = c$. We will use the $L_\infty L_\infty$ - norm to calculate the distance between pixels, for pixels $p1 = (x_1, y_1)$, $p_2 = (x_2, y_2)$, then set

$$|p_1 - p_2| \triangleq \max\{|x_1 - x_2|, |y_1 - y_2|\}.$$

After that a set of numbers (1,2,3,4,5,6 ,..., n) will be declared to be n. Thus obtained the following equation:

$$Pr_{p1\in I_{ci}, p2\in I}[p2 \in I_{cj} || p1 - p2 = k]$$

Where every pixel on the color c_i of the image, $\gamma^{(k)}_{ci,cj}$ providing the possibility that the pixel at a short distance to the k of the pixel is the color of the unknown.

When we use Correlogram we must remember that there are problems concerning the size of d that will be used to perform computation of Correlogram. Next we will perform calculations to deal with a case when the value of d that is used to calculate the value of Correlogram.

To calculate Correlogam we first perform the following computation, where calculations are similar to the calculations performed for co-occurance matrix method in the analysis of texture and color – gray.

The formula to calculate the image Correlogram method is as follows:

$$\gamma^{(k)}_{c_i,c_j}(I) = \frac{\Gamma^{(k)}_{c_i,c_j}(I)}{h_{c_i}(I) \cdot 8k}$$

We will conduct the first CCM calculations, after getting the result of CCM calculations we will do the division with the results of those calculations by using the histogram after it was multiplied by 8k, 8k factor due to the eight neighboring pixels of the pixel center.

Correlogram method will serve as the basis for the theory of image Correlogram method capable of providing spatial information using the calculation formula described previously, so as to provide the level of the high-accuration to search images.

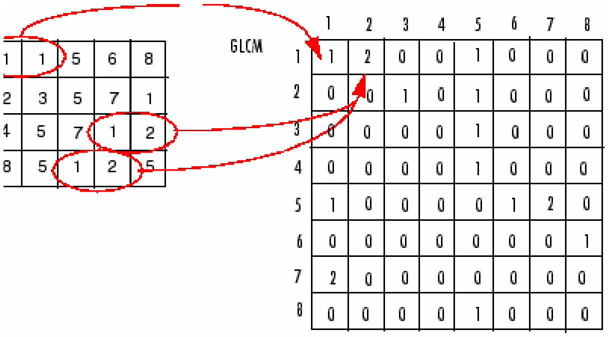
III. Gray Level Coocurance Matrix

Gray Level Co-Matrix occurrance was discovered by Robert Haralick [2], Gray Level Coocurrance matrix (which is better known as *glcm*) was used to measure the intensity of the gray level of an image by counting the number of the emergence of a pixel with a pixel with color - gray with a value of i appear horizontally by pixels that have a value of j in which the two pixels are adjacent, if the image is binary image then these pictures menskalakan glcm will be two levels of gray if the image has the intensity of the the image will be scaled to eight levels of gray.

glcm = graycomatrix(I) creates a gray-level co-occurrence matrix (GLCM) from image I. Graycomatrix creates the GLCM by calculating how often a pixel with gray-level (grayscale intensity) value i occurs horizontally adjacent to a pixel with the value j Each element (i,j) in glcm specifies the number of times that the pixel with value i occurred horizontally adjacent to a pixel with value j.

graycomatrix calculates the GLCM from a scaled version of the image. By default, if I is a binary image, graycomatrix scales the image to two gray-levels. If I is an intensity image, graycomatrix scales the image to eight gray-levels. We can specify the number of gray-levels graycomatrix uses to scale the image by using the 'NumLevels' parameter, and the way that graycomatrix scales the values using the 'GrayLimits' parameter.

The following figure shows how graycomatrix calculates several values in the GLCM of the 4-by-5 image I. Element (1,1) in the GLCM contains the value 1 because there is only one instance in the image where two, horizontally adjacent pixels have the values 1 and 1. Element (1,2) in the GLCM contains the value 2 because there are two instances in the image where two, horizontally adjacent pixels have the values 1 and 2. graycomatrix continues this processing to fill in all the values in the GLCM.



1	1	5	6	8
2	3	5	7	1
4	5	7	1	2
8	5	1	2	5

	1	2	3	4	5	6	7	8
1	1	2	0	0	1	0	0	0
2	0	0	1	0	1	0	0	0
3	0	0	0	0	1	0	0	0
4	0	0	0	0	1	0	0	0
5	1	0	0	0	0	1	2	0
6	0	0	0	0	0	0	0	1
7	2	0	0	0	0	0	0	0
8	0	0	0	0	1	0	0	0

Figure 1

By default, the graycomatrix function creates a single GLCM, with the spatial relationship, or *offset*, defined as two horizontally adjacent pixels. However, a single GLCM might not be enough to describe the textural features of the input image. For example, a single horizontal offset might not be sensitive to texture with a vertical orientation. For this reason, graycomatrix can create multiple GLCMs for a single input image.

To create multiple GLCMs, specify an array of offsets to the graycomatrix function. These offsets define pixel relationships of varying direction and distance. For example, we can define an array of offsets that specify four directions (horizontal, vertical, and two diagonals) and four distances. In this case, the input image is represented by 16 GLCMs. When we calculate statistics from these GLCMs, we can take the average.

We specify these offsets as a p-by-2 array of integers. Each row in the array is a two-element vector, [row_offset, col_offset], that specifies one offset. row_offset is the number of rows between the pixel of interest and its neighbor. col_offset is the number of columns between the pixel of interest and its neighbor. This example creates an offset that specifies four directions and 4 distances for each direction

```
offsets = [ 0 1; 0 2; 0 3; 0 4;...
          -1 1; -2 2; -3 3; -4 4;...
          -1 0; -2 0; -3 0; -4 0;...
          -1 -1; -2 -2; -3 -3; -4 -4];
```

The figure illustrates the spatial relationships of pixels that are defined by this array of offsets, where D represents the distance from the pixel of interest.

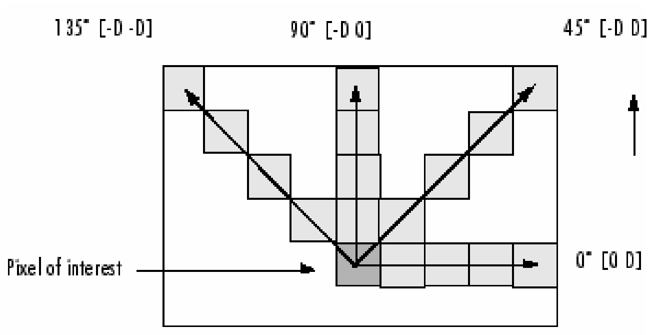


Figure 2. Spatial Relationshsip of Pixels

There are four main statistical feature from graycomatrix widely known as Harralick feature [2]. The definition and the feature will be describe in table below.

Table 1. The Definition and Description for Feature

Statistic	Description
CONTRAST	Mesures the local varuatiobs in the gray level coocurrent matrix.
CORRELATION	Measures the joint probability occuranece of the specified pixel pairs.
EBERGY	Provides the sum of squeraed elements in the GLCM. Also know as uniformity or the angualr dcond moment .
HOMOGENITY	Measures the closeness of the distribution of elemtns in the GLCM to the GLCM diagonal.

IV. PERFORMANCE MEASUREMENT

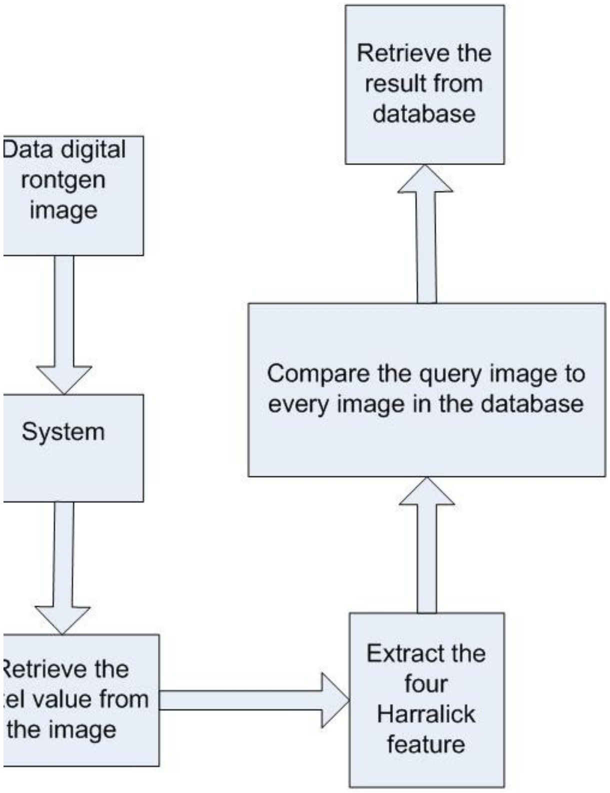
To measure the effciency and precision of the sistem researcher will use recall and precision. Effiiency means the speed during the retrieval of the query result, precision means the accuracy of retrieval of the query result. Recall and precision are collectively used to measure the effectiveness of a retrieval system. Recall measures the capacity to retrieve relevant information items from the database. It define as the ratio between the number of relevant items retrieved and total number of relevant items in the database. During performance testing, the total items number of relevant items in the database for each testing query should be determined by an expert in the domain. The higher the recall, the better the performance. Precision measures the retrieval accuracy. It is defined as the ratio between number of relevant items retrieved and the total number of retrieved items. The higher the precision, the higher the retrieval performance.[3]

V. METHODOLOGY.

This research used sample image as a query and then extract Harralick feature of the image and the same method will be apply to every image in database.

The next step we use distance measurement to compare the Harralick features of the query image and every image in database.

The Sistem will show the 5 smallest distance of the image. Eucludian distance will be used as the method to calculate distance between the query image and every image in database. The flow of the methodology will be shown in figure 3


Data digital rontgen image
System
Retrieve the
el value from
the image
Extract the four Harralick feature
Compare the query image to every image in the database
Retrieve the result from database

\

Figure3. Spatial Relationshsip of Pixels

VI. RESULT

First retrieve a digital image of cancer infected Lung Radiology picture as a query image extract four Harralick.The Sample Image that shown in fiure 5.1 will be use as a sample datum :

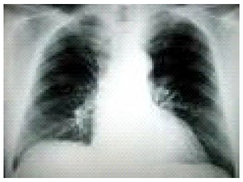


Figure 4. Sample Lung Image

From the above image the statistical feature of an image which is know for Harralick Feature would be extracted, the result of the ectraction will b e shown in below graphic

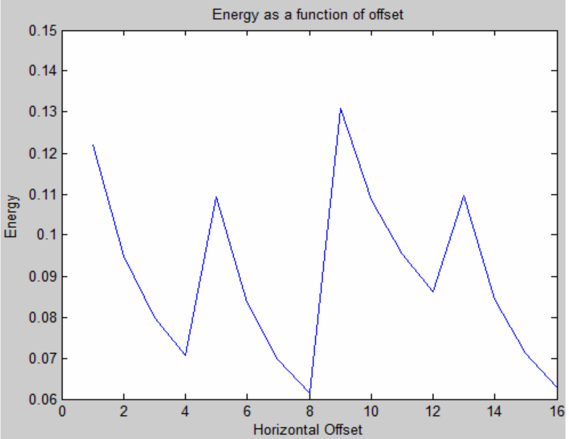


Figure 5 Energy from query image

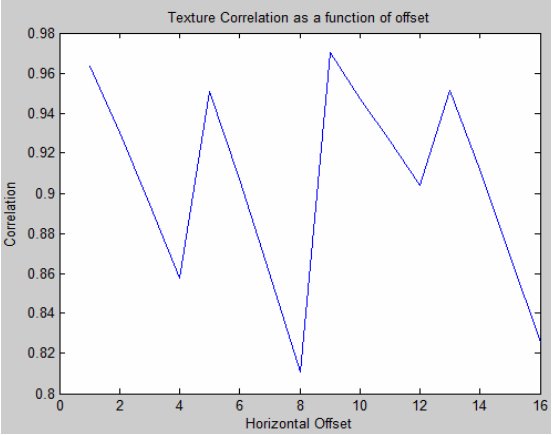


Figure 6 Correlation from query image

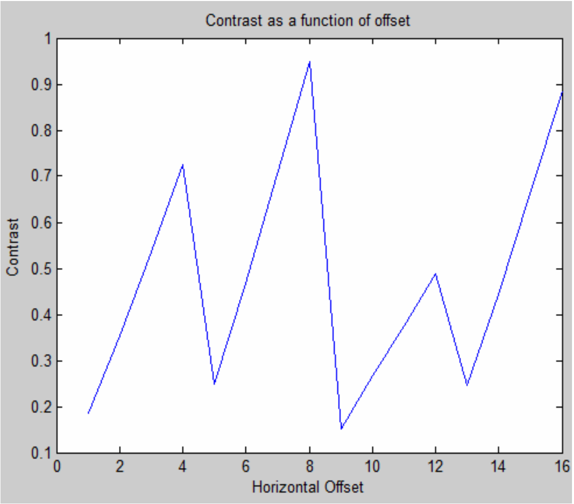


Figure 7 Contrast from Query image

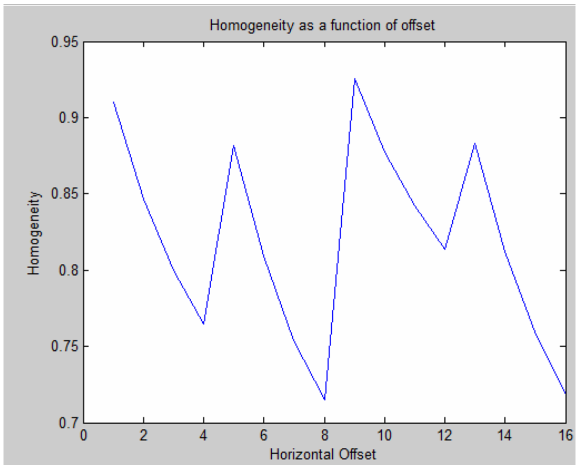


Figure 8. Homogeneity from Query Image

Then we use the four offset from GLCM to compare the pixel with the neghbouring pixel from Query image. From the retrieval the four Harralick feature using the offsetes that mention in section 2 resulting in matrix [1X16] and then the result will be concat to matrix[1X64], and then the Eucludian distance describe in equation below.

$$\mathrm{d}(\mathbf{p},\mathbf{q}) = \sqrt{(p_1-q_1)^2 + (p_2-q_2)^2 + \cdots + (p_n-q_n)^2} = \sqrt{\sum_{i=1}^{n}(p_i-q_i)^2}.$$

.

will be used to calculate the distance of Query image with every image in database .

In Figure 9. shows the query result from Figure 4.

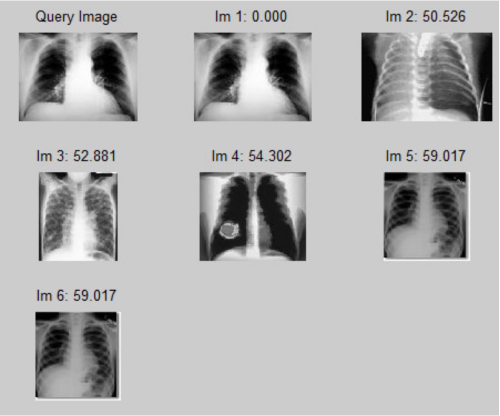


Figure 9. Query result

This Research used 20 Sample image, the following table will show the Performance of the System :

Table 2. System Performance

Performance Measurement		
Number of Items Return	***Recall***	***Precision***
1	1/3	1/1
2	2/3	2/2
3	3/3	3/3
4	3/3	3/4
5	3/3	3/5
6	3/3	3/6
7	3/3	3/7
8	3/3	3/8
9	3/3	3/9
10	3/3	3/10
11	3/3	3/11
12	3/3	3/12
13	3/3	3/13
14	3/3	3/14
15	3/3	3/15
16	3/3	3/16
17	3/3	3/17
18	3/3	3/18
19	3/3	3/19
20	3/3	3/20

As we can see from the recall of the third item the system already return 3 exact or relatively similar to the query image. This system has high precision and performance in Rontgen image retrieval, by implementing this system than misdiagnoses could be avoid For the future references researcher suggests to developed biomedical image clustering to enhance the efficiency of the system purposed.

VII. Conclusion

This Research provides the analysis of gray color images of x-ray Rontgen, as we could see from the result from the recall and precision method the system return exact or relatively similar to the query image on return item 3. This system has high precision and performance in Rontgen image retrieval, this system could be used in medical area to generate a supporting diagnoses. The accuracy of the system already shown in table 5.1 For the Future Refrences researcher suggest the classification in order to increase the performance of the system.

References

[1] Peterson MC, HolbrookJH, Hales D, Smith NL, Staker LV: Contributions of the history, physical examination, and laboratory investigation in making medical diagnoses. West J Med 2008 Feb; 156:163-165

[2] Haralick, R.M., K. Shanmugan, and I. Dinstein, "Textural Features for Image Classification", IEEE Transactions on Systems, Man, and Cybernetics, Vol. SMC-3, 1973, pp. 610-621

[3] L. GuoJun, “Multi Media Data Base Management System ”, Lu, Guojun. 1999. *Multimedia Database Management Systems.* Artech House, Inc

[4] H. Jing, K. Ravi S., M. Mandar, Z. Wei-Jing, Z. Rabih, “Image Indexing Using Color Correlogram”, 2005, Cornell University Ithaca, NY 14853.

[5] Patient Safety in American Hospitals, July 2004, HealthGrades Quality Study, Health Grades, http://www.healthgrades.com/media/english/pdf/HG_Patient_Safety_Study_Final.pdf

[6] National Patient Safety Foundation at the AMA: Public Opinion of Patient Safety Issues, Louis Harris & Associates, September 2007.

[7] John Davenport, MD, JD, Documenting High-Risk Cases to Avoid Malpractice Liability, Family Practice Management, October 2009.

[8] R. Patel, Pradip, “ Seelcted topics in MRI and Radiologi ” , Springer 2006.

[9] http://www.medicalmalpractice.com (accesed on 23 June 2010)

[10] http://www.wrongdiagnosis.com (Accesed on 10 May2010)

.