International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

Detection of Lung Diseases using Nearest Neighbor Classification Technique

Alyaa Hussein Ali[1], Zahra Jaafer Abod Jasim[2]

[1]University of Baghdad, College of Science for Women

[2]University of Baghdad, College of Science for Women

Abstract: ***Lung diseases can be defined as the disorder that affects the lungs. The Lang is the organs that allow the human to breathe; the lung diseases can be detecting using CT images and image processing techniques. The first step in image enhancement is using the wiener filter as a noise remover filter; the segmentation process which is essentially used in detecting the defect in lung by using optimized Otsu threshold method which is based on improved the thresholding algorithm. The Nearest neighbor supervised classification method is used for classifying the lung diseases with the use of the histogram statistical features (first order statistical features).***

Keywords: CT images, wiener filter, O'tsu threshold, statistical, features, nearest neighbor

1. Introduction

The lung disease can be consider as many disorders that affecting the lungs such as asthma, chronic obstructive pulmonary (COPD) disease, infections such as tuberculosis, influenza, lung cancer, pneumonia and other breathing problems[1]. The lung diseases can differ from person to another depending on their type. The first signs that there is a diseases in the lung is the trouble in breathing, shortness in breathing (not getting enough air), pain or discomfort when breathing. The chest computerized tomography (CT) is considered in this search to detecting the lung tumors. In 2008 Diciotti proved that the CT is the most sensitive imaging technique for detecting lunge nodules, while in 2011 K.A.G.Udeshani succeed in putting a successful solution for detecting the lung cancer nodules using image processing and neural networks. in 2012 Kuchanur discover a fast method to detect the lung nodules and separated the cancer image from the other lung diseases. In 2013 C. Chandrasekar examined a segmentation technique which based on fuzzy possibility with the classification of lung nodules as normal and abnormal using (SVM). Different methods are used to produce the best result. In this search the first step in detecting the lung disease is to remove the noise from the CT image open, close and Wiener filter are used as an enhancement methods. The Otsu's threshold is applied as a segmentation method [1]. "O'tsu thresholding" performs better results than the traditional threshold it produces suitable binary images. The "nearest neighbor" which is the supervised classification technique is applied to classify the lung diseases and the statistical method is applied to identify the result.

2. Test Images

The images which are used in the search are (CT) images. The CT scan provide the ability to distinguish between tumor and normal tissues with the help of digital image processing [2]. The computed tomography gives image that can scan through breathing which make the (CT) tube preferred way to describe the rib cage [3], the test images are three for cancer case, three for Echinococcosis and three for Tuberculosis.

3. Thresholding

The thresholding is the oldest and the simplest segmentation methods. Its idea is based on the extract of object from its background by gathering the intensity value of each pixel according to the threshold value [4].

4. Otsu Thresholding

Otsu's threshold is the best thresholding methods used for the real world images with regarding the uniformity and shape measures. The Otsu's threshold takes too much time to be practical for multilevel threshold selection [5]. Otsu's used to automatically perform clustering-based image thresholding by using binary images, this can obtain by transforming the gray image to a binary image. The Otsu's threshold algorithm proposed that the test image must contain two classes of pixels following bimodal histogram "foreground pixels and background pixels", then it evaluate the best threshold value "optimum threshold" which can separate these classes so that, their combined spread "intra-class variance" is minimum. The Otsu's method is completely a method for finding the threshold that minimizes the "intra-class variance". The variance with in class is [6].

$$\sigma_w^2(t) = q_{1(t)}\sigma_{1(t)}^2 + q_{2(t)}\sigma_{2(t)}^2 \quad (1)$$

$$q_{i(t)} = \sum_{i=1}^{t} p_{(i)} \quad (2)$$

(t) Threshold value and (i) gray level. The class means can be obtained by the following equation [6]:

$$\mu_{i(t)} = \sum_{i=1}^{t} \frac{i\, p_{(i)}}{q_{i(t)}} \quad (3)$$

$$\sigma^2{}_{(t)} = \sigma^2 - \sigma^2{}_{w(t)} \quad (4)$$

$$\sigma^2 = \sigma_{w(t)}^2 + q_{1(t)}\left[1 - q_{1(t)}\right]\left[\mu_{1(t)} - \mu_{2(t)}\right]^2 \quad (5)$$

5. Open and Close

Opening process deals with the edges of the object, it makes them smoother by eliminating the thin edges. It is erosion followed by dilation, the closing process deals with the object edges, they merge the narrow lines and break the thin lines. They eliminates small gabs and fills the holes in the counters.

6. Wiener filter

The wiener filter is an optimal filter for remove noise from images, it remove the noise that corrupted the signal. It is the filter whose output would come as close as to the original signal as showing in equation (6) [7].

$$\hat{F}_{(u,v)} = \left[\frac{H_{(u,v)^*}}{|H_{(u,v)}|^2 + \left[S_{n(u,v)} / S_{f(u,v)} \right]} \right] G_{(u,v)} \quad (6)$$

The H_((u,v))is the degradation function (* indicates complex conjugate) and G_((u,v)) is the degraded image. The ⟦S_f⟧_((u,v)) and ⟦S_n⟧_((u,v)) are the power spectra of the original image (prior to degradation) and the noise if the noise is zero.

7. First-Order Statistics Features

It is one of the texture feature properties study, it can be calculated from the histogram representation of pixel intensities which represent the image. It count on only the "individual pixel" values and not on the occurrence of "neighboring pixel value". "The First-order statistics feature" measure the likelihood of a gray value in the observed data randomly. The statistical representation of the image features by measuring the properties of the texture direct from the histogram of the images are.

•Energy: it represent the sum of squared elements. It give information about the gray level distribution. Its range is from 0 to 1 [8].

$$E = \sum_{i=0}^{G-1} (p(i))^2 \quad (7)$$

p(i) is the probability density distribution of occurrence of the intensity, it determined from the histogram where the total number of pixels in the image is given by[9].

$$p(i) = H(i)/NM \quad (8)$$

i= 0,1,2……. G-1. N=number of cell in horizontal domain. M= number of cell in vertical domain. G=gray level of an image (255).

• Entropy: It Measure the randomness of a gray-level distribution in the texture images. It expected to be high value if the gray levels are randomly distributed throughout the image otherwise it is low. It's inversely proportional to the energy value. It can be represented by [10]:

$$H = -\sum_{i=0}^{G-1} p(i) \log_2 [p(i)] \quad (9)$$

•Mean: It is the average value of the data or sometimes called the mean value of the gray levels in the image. The mean value is large if the sum of the gray levels of the image is high. It equation given by [9].

$$\mu = \sum_{i=0}^{G-1} i\, p(i) \quad (10)$$

•Variance: The value of the variance expected to be large if the gray levels of the image are spread out greatly. It is can be given by the following equation [9].

$$\sigma^2 = \sum_{i=0}^{G-1} (i-\mu)^2 p(i) \quad (11)$$

•Standard Deviation: It represent the mean square root of the variance, the standard deviation shows much 'variation' or 'dispersion' exists from the average [9].

$$Std = \sqrt{\sigma^2} \quad (12)$$

•Skewness: The Skewness represent the (third moment). The value is measure to be "0 for symmetric histograms representation", "positive value when the histograms skewed to the right" and "negative value for histograms skewed to the left". The skewness can be represented by the following equation [9,10].

$$\mu_3 = \sigma^{-3} \sum_{i=0}^{G-1} (i-\mu)^3 p(i) \quad (13)$$

Kurtosis: The kurtosis value represent the flatness of histogram. Kurtosis value is the degree of peachiness of a distribution, it can be defined as the normalized form of the fourth central moment of the distribution[10,11].

$$\mu_4 = \sigma^{-4} \sum_{i=0}^{G-1} (i-\mu)^4 p(i) \quad (14)$$

8. The Nearest Neighbor Classification

The Nearest Neighbor Classification technique is one of the most important supervised learning process algorithm. It predicts the propose test sample's with respect to the training "K "samples which are the closest neighbors to the proposed sample, it judge to that category which has the largest category probability. The process of Nearest Neighbor to classify the test sample represent by the following step [12, 13]:

1) Determine the number nearest neighboring "K" pixels to be consider.
2) Measure the distance from chosen sample and the all the remains training samples using the following equation

$$d_{ij} = \sum_{k=1}^{n} (x_{ik} - x_{jk})^2 \quad (15)$$

In which the (x_ik) is refers to the pixel points, and the (x_jk) is the values of training samples. Sort the distance and determine nearest neighbor based on the minimum distance of k-pixel. Assign the majority class among the nearest neighbors to the accreditation on the values minimum square distances.

9. Hot Color

Color map "Hot" which also known as "Warm Color" refer to the smoothly changes in the color from the black, passing into shades of red color, orange color the yellow color, and ending at the white color. The adjacent color in this model

has equal distance, a 256 colors scale are implemented as an extension of the 16-step color scale [14].

Let the color "R_i , G_i , B_i" and "R_i+1 , G_i+1 , B_i+1" represent any two adjacent base colors and I_i and I_i+1 denote their corresponding gray levels. The gray level I ($I_i < I < I_i+1$ for $1 \le i \le 15$), associated with the color "*R*, *G*, *B*" represented by the following Equation[14].

$$R = (R_{i+1} - R_i)\left(\frac{I-I_i}{I_{i+1}-I_i}\right) + R_i \quad (15)$$

$$G = (G_{i+1} - G_i)\left(\frac{I-I_i}{I_{i+1}-I_i}\right) + G_i \quad (16)$$

$$B = (B_{i+1} - B_i)\left(\frac{I-I_i}{I_{i+1}-I_i}\right) + B_i \quad (17)$$

10. Methodology

The CT images is taken from the net, the O'tsu threshold is used with the help of open and close process to remove the noise and finding the region of interest, the Winner filter is used to remove the nose and restoration which is funding in the CT images, the Hot color is applied preparing the image to the nearest neighbor classification, as a finally step the statistical features is determined which is calculated from the first order statistical feature.

11. Discussion

Figure (1) shows the test CT images for Cancer Echinococcosis and Tuberculosis cases. Figures (2,3,4), represented the Otsu thresholding for cancer, Echinococcosis and Tuberculosis cases, it is the traditional way for segmentation, which gives separation for Echinococcosis , cancer Tuberculosis tumors, it is not perfect way to separate the tumors from the health tissue so that the images need enhancement technique such that, the open and close with the winner filter is applied to smooth the image and to remove the noise from it, the Hot color which is a color representation helps to separate the texture classes from each other. In the Nearest Neighbor supervised classification technique each part in the texture has specific color, this helps to separate the texture, the normal from abnormal color in a separately classes. As the Cancer ,the Echinococcosis and Tuberculosis part separated from the rest image as shown in figure(5,6,7) the first order statistical feature is determined and six features are calculated each one represent an interpretation of the texture behavior Table(1,2,3) shows that.

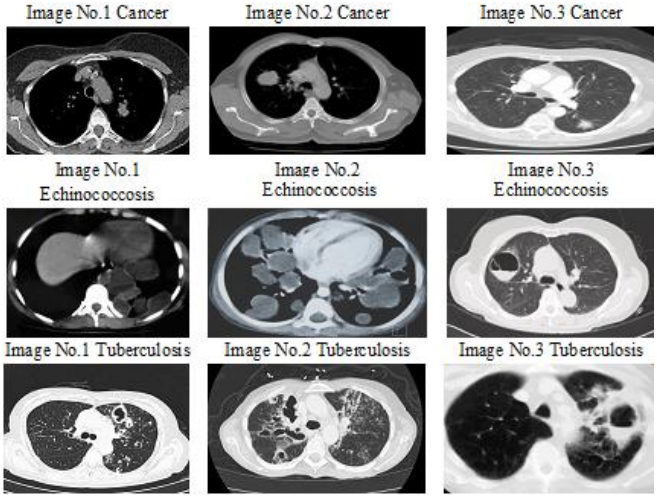

Figure 1: The test CT images for Cancer, Echinococcosis and Tuberculosis cases.

Figure 2: The Otsu thresholding for the cancer case images.

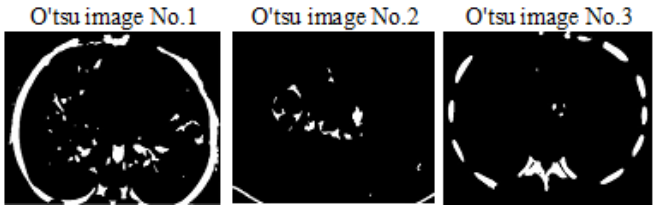

Figure 3: The Otsu thresholding for the Echincoccosis case images.

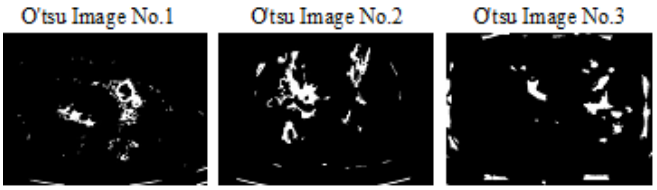

Figure 4: The Otsu thresholding for the Tuberculosis case images.

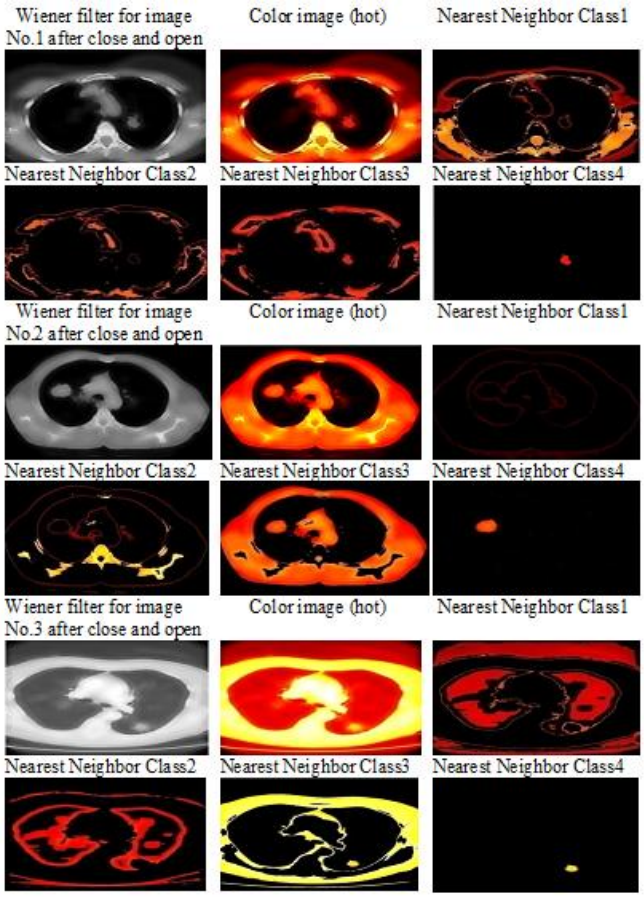

Figure 5: The Nearest Neighbor for the Cancer case images.

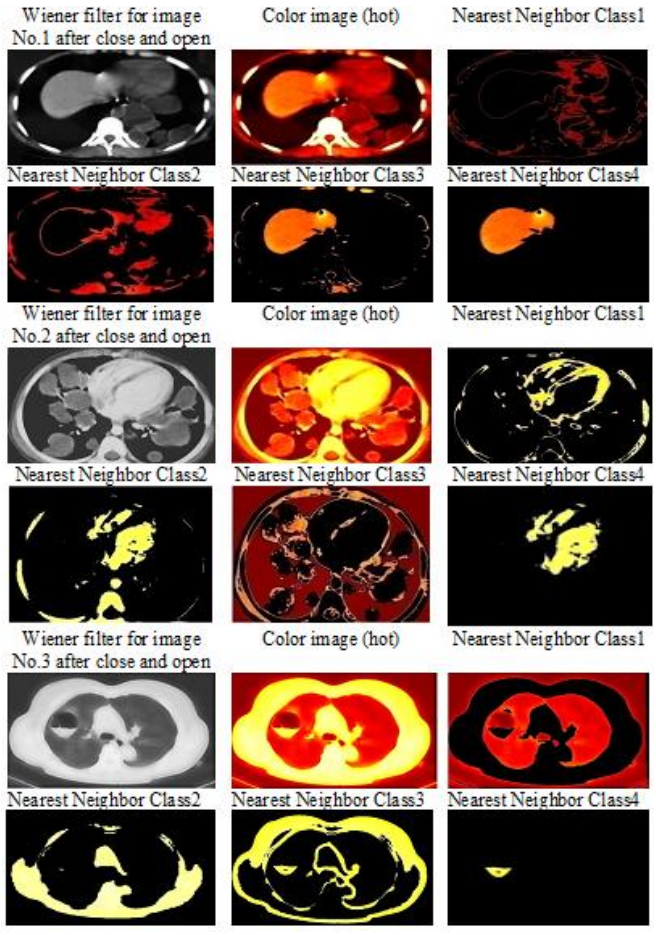

Figure 6: The Nearest Neighbor for the Echinococcosis case images.

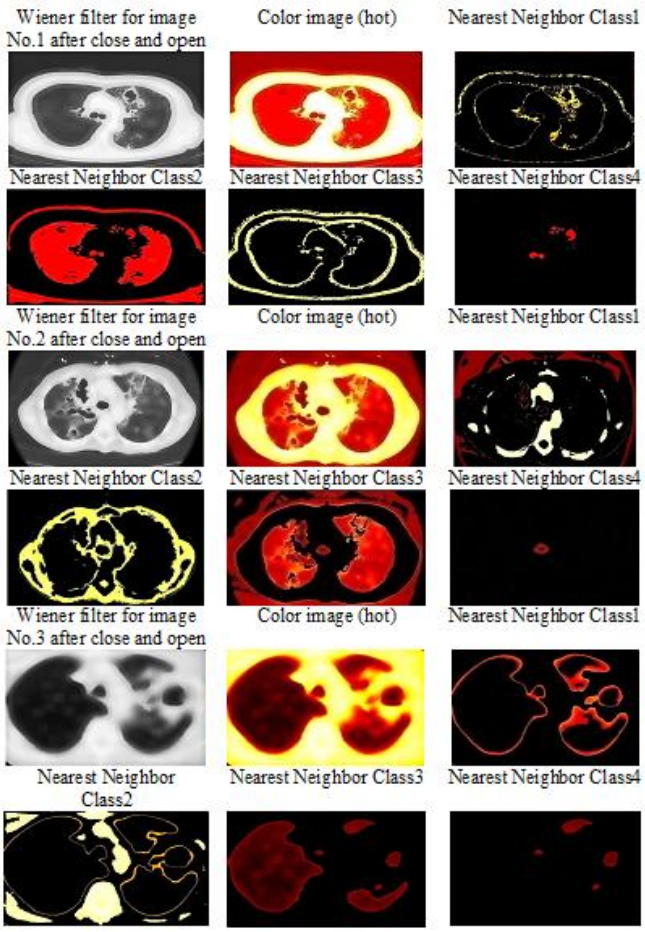

Figure 7: The Nearest Neighbor for the Tuberculosis case images.

Table 1: shows the first order statistical features for the cancer case

Image No.	First order for the Cancer case					
	Variance	Skewness	Kurtosis	Entropy	Energy	Mean
Image No.1	662.5091	1.6780e-004	2.5327e-007	0.2740	0.1146	42.321
Image No.2	2.5172e+003	1.2946e-005	5.1430e-009	-0.0142	0.0165	48.7222
Image No.3	6.8699e+003	1.8888e-006	2.7495e-010	0.0113	0.0139	149.2655
Average value	3349.8697	6.08783E-05	8.62293E-08	0.0903	0.0483333	80.102

Table 2: shows the first order statistical features for the Echinococcosis case

Image No.	First order for the Echinococcosis					
	Variance	Skewness	Kurtosis	Entropy	Energy	Mean
Image No.1	4.6772e+003	6.7568e-006	1.4446e-009	0.3193	0.2500	38.5882
Image No.2	2.8324e+003	2.4257e-005	8.5643e-009	0.0724	0.7175	16.5456
Image No.3	6.8513e+003	2.7224e-006	3.9736e-010	0.3889	0.1304	70.3973
Average value	4786.9666	1.12454E-05	3.46875E-09	0.2602	0.36596	41.8437

Table 3: shows the first order statistical features for the Tuberculosis case

Image No.	First order for Tuberculosis case.					
	Variance	skewness	Kurtosis	entropy	energy	Mean
Image No.1 right	282.2755	0.0023	8.1035e-006	-0.0182	0.8839	2.7046
Image No.1 left	211.7912	0.0040	1.8736e-005	-0.0267	0.9102	2.3031
Image No.22	806.9258	2.6378e-004	3.2690e-007	0.0421	0.6874	6.1747
Image No.23 right	20.5354	0.0835	0.0041	5.3184e-004	0.9030	1.6651
Image No.23 left	12.5031	1.1277	0.0902	-0.0322	0.9475	1.1845
Average value	266.8062	0.243552756	0.018865433	-0.006893632	0.8664	2.8064

12. Conclusion

For the Tuberculosis case the average value for the statistical features, Skewness, Kurtosis with higher value than the Echinococcosis and the cancer, the Skewness value for the normal distribution is between -1 and +1 any other value greater or less than these value can be consider to be shifted to the right or to the left, the Kurtosis value is high than the Cancer case and Echinococcosis this describe the behavior of the histogram, high value means the histogram is flat. So, the Cancer cell and Echinococcosis tend to be non-homogenous in its texture in the lung and this can obviously see from the value of the energy. The energy value for the cancer is very small compere to the Echinococcosis and this is very small compare to Tuberculosis case, this means the No. of gray level distribution is high. The cancer cell is non uniform distributed. The mean value is higher than that of Echinococcosis and Tuberculosis case this means that the cancer cell tend to be whiter than Echinococcosis and Tuberculosis. The variance of Tuberculosis, mean and entropy is lower than that for cancer and Echinococcosis it has less No. of gray level distribution and its cell is regular and homogenous in its texture. The conclusion is that the Tuberculosis cell is homogenous and regular than the cancer and Echinococcosis cell in its texture and the cancer is non uniform cell compare to Echinococcosis and Tuberculosis .

References

[1] M. C. Robert, Richard, E. M. Thomas, R. and Judith J. S., M.S.N, R.N., A.O.C.N., MaryW., Sc.D "LungDisease", M.P.H. Division of Respiratory Disease Studies National Institute for Occupational Safety and Health, Ph. D. Associate Director, Office of Cancer Content Management National Cancer Institute, 2010.

[2] K. Kanazawa, Y. Kawata, N. Niki, H. Satoh, H.Ohmatsu, R. Kakinuma, M. Kaneko, N. Moriyama and K. Eguchi, "Computer-aided diagnosis for pulmonary nodules based on helicalCT images", Compute. Med. Image Graph, vol. 22, no. 2, pp. 157-167, 1998.

[3] D. Lin and C. Yan, "Lung nodules identification rules extraction with neural fuzzynetwork", IEEE, Neural Information Processing, vol. 4, 2002.

[4] S. N. Mazhir, "Texture Analysis of smear of Leukemia Blood Cells after Exposing to Cold Plasma", Baghdad Science Journal, vol. 14, I:2, 2017.

[5] P. K. Sahoo, S. Soltani, A. K. C. Wong, and Y. Chen, "A Survey of Thresholding Techniques," Computer Vision Graphics Image Processing, Vol. 41, pp. 233-260. 1988.

[6] Tseng, C.C. Spletters, "An Efficient Design of a Variable Fractional Delay Filter Using a First-Order Differentiator". 10(10):307–10, 2003.

[7] L. Dongju and Y. Jian, 2009, "Otsu Method and k-means," in Hybrid Intelligent Systems, 2009. HIS '09. Ninth International Conference on, vol. 1, pp. 344–349.

[8] Azeez, M. A., Mazhir, S. N., & Ali, A. H. "Detection and Segmentation of Lung Cancer using Statistical Features of X-Ray Images," International Journal of Computer Science and Mobile Computing, Vol.4 Iss.2, 307-313, 2015.

[9] S. Mazhir, "Studying the Effect of Cold Plasma on Living Tissues Using Images Texture analysis," Diyala Journal for Pure Science. vol. 13. No.2 pp184-202, 2017.

[10] M.Tuceryan and A.Jain. Texture Analysis. In Handbook of pattern Recognition and Computer Vision, Chapter 2, Pages 235-276. Word Scientific, 1998.

[11] M. A. Azeez,, Ali, A. H., and S. N. Mazhir, "Detection and segmentation of lung disease using Law Mask with Watershed on X-ray images," International Journal of Scientific & Engineering Research, Vol. 6, Issue 3, March-2015 .

[12] M. Abramowitz, I.Stegun, "Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables"9th printing. New York, Dover, PP. 928, 1972.

[13] J. F. Kenny, and Keeping, E.S. 1951, "Mathematics of Statistics", Pt.2, 2nd ed. Princeton, Nj:VanNostrand

[14] X. Q. Shi, P. Sallstrom, and U. Welander, "A Color Coding Method for Radiographic Images," Image and Vision Computing Vol. 20, pp.761-767, 2002.