Volume 4, No. 2, February 2013
Journal of Global Research in Computer Science
ISSN-2229-371X
RESEARCH PAPER
Available Online at www.jgrcs.info

STATISTICAL FEATURE EXTRACTION TO CLASSIFY ORAL CANCERS

Anuradha.K[*1], Dr. K. Sankaranarayanan[2]
[*1] Research Scholar, Karpagam University, Coimbatore, Tamilnadu, India
k_anur@yahoo.com
[2] Dean, Easa College of Engineering and Technology, Coimbatore, Tamilnadu, India
kkd_sankar@yahoo.com

Abstract: Oral Cancer is the most common cancer found in both men and women. The proposed system segments and classifies oral cancers at an earlier stage. The tumor is detected using Marker Controlled Watershed segmentation. The features extracted using Gray Level Co occurrence Matrix (GLCM) is Energy, Contrast, Entropy, Correlation, Homogeneity. The extracted features are fed into Support Vector Machine (SVM) Classifier to classify the tumor as benign or malignant. The accuracy obtained for the proposed system is 92.5%.

***Keywords*:** Marker Controlled Watershed Algorithm, GLCM, SVM

INTRODUCTION

Oral cancer refers to the cancer that occurs in the head and neck region [1]. India accounts for 86% of oral cancer cases [2]. Oral cancer is the most common cancer found in both men and women. Chewing or smoking tobacco is the main cause of oral cancer, a condition which claims the lives of 10,000 people each year, more than cervical cancer or malignant melanoma. Because of the difficulty in detecting oral cancer early, it has one of the worst survival rates of all cancers, less than 50% of patients survive more than 5 years after diagnosis [3]. Oral cancer starts in the cells of the mouth (oral cavity). The oral cavity is made up of many parts like lip, tongue, inside of the lip and cheeks, hard palate (roof of the mouth), floor of the mouth, gums and teeth. Oral Cavity cancers have been increasing in the recent years and each year more new cases of oral cancer are reported. "Ahmedabad is considered the capital of oral cancers with 40% of cancers recorded being cancers of the mouth mostly caused by tobacco and gutkha chewing" [4].

"Maharashtra has the highest incidence of mouth cancer in the world". The common oral precancerous lesions are leukoplakia, erythroplakia, and oral sub – mucous fibrosis (OSF). The diagnosis of Oral precancer and cancer remains a challenge to the dental profession, particularly in the detection, evaluation and management of early phase alterations or frank disease [5]. The symptoms of the early oral cancer include: Persistent red /white patch non – healing ulcer, progressive swelling, sudden tooth mobility without apparent cause, unusual oral bleeding. Though oral cancers are detected easily, identification becomes difficult in initial stages. Oral Cancer can save life if they are diagnosed earlier. This paper presents the classification of normal and abnormal sections from oral images. The proposed work is shown in Figure 1.

The input image obtained is digitized and preprocessed using Contrast Linear stretching. After image enhancement, the tumor part is segmented and the features of the tumor are extracted using Grey Level co – occurrence Matrix (GLCM). Performance measure is made to identify the abnormal portions in the image. Once an abnormal portion is detected, radiologist recommends for Biopsy. As biopsy in mouth cavity is a painful task, only patients who are detected with abnormal sections are recommended.

The remainder of the paper is organized as follows: Section 2 describes the previous works in this field. Section 3 describes the methodology for the proposed system. The experiments and results are presented in Section 4. Finally Section 5 describes the conclusion of proposed work.

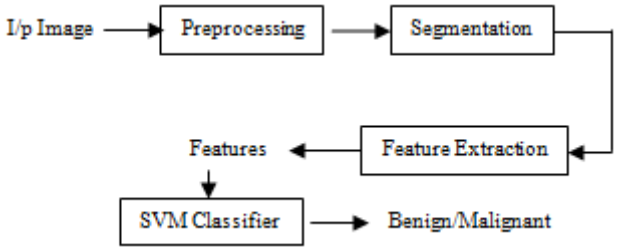

Figure 1. Proposed System

PREVIOUS WORK

In the literature various techniques are described to detect and classify the cancer in digital images. A lot of research has been done on Feature Extraction for classification of cancers.

Lalit Gupta et al [6] proposed a new method of Feature Selection using Mean – shift and Recursive Feature Elimination techniques to increase discrimination ability of the feature vectors. Performance of the algorithm is evaluated on a in-vivo recorded LIF data set consisting of spectra from normal, malignant and pre-malignant patients. Sensitivity of above 95% and specificity of above 99% towards malignancy are obtained using the proposed method.

Sebastian Steger et al [7] have proposed a method for novel image feature extraction approach that is used to predict oral cancer reoccurrence. Several numeric image features that characterize tumors and lymph nodes are also proposed. In order to automatically extract those features Registration and supervised segmentation of CT/MR images form the base of automated extraction of geometric and texture

features of tumor and lymph nodes. Higher accuracy and robustness is achieved compared to today's clinical practice. Micheletti A et al [8] classified tumor cells based on statistical shape analysis. Here the Theory of Size Functions is introduced and joined to some statistical techniques of discriminant analysis, to perform automatic classification of families of random shapes. The method is applied to the classification of normal and malignant tumor cell nuclei, described via their section profiles. The results here reported are compared with other techniques of shape analysis, already applied to the same data, showing some improvements.

M. Muthu Rama Krishnan et al [9] have proposed a wavelet based texture classification for oral histopathological sections. As the conventional method involves in stain intensity, inter and intra observer variations leading to higher misclassification error, a new method is proposed. The proposed method, involves feature extraction using wavelet transform, feature selection using Kullback – Leibler (KL).

G. Landini [10] analysed epithelial lining architecture in radicular cysts and odontogenic keratocysts applying image processing algorithms to follow a traditional cell isolation based approach. This formed the basis for later estimation of tissue layer level and architectural analysis of oral epithelia. Jadhav et al [11] carried out segmentation of the Histological OSF images using region growing and hybrid segmentation algorithm. Misclassification rate were calculated for both the algorithms. Finally, Hybrid Segmentation method found to be suitable for segmentation of cancers in OSF images.

K.V. Kulhalli et al [12] proposed a computer aided diagnostic system and ANN to detect and classify oral cancers present in Biopsy Image. The system was tested with many different types of images and found to be good.

METHODOLOGY

As shown in Figure 1, the proposed work is carried out in three stages. Dental X – rays are digitized and given as the input. The input image is preprocessed to remove the noise (Figure 3). Later, the enhanced image is segmented to detect tumor from image and the features are extracted to identify the tumor as benign or malignant (Figure 4).

Image Preprocessing:

The first stage is the Image Preprocessing. The input image which is obtained is preprocessed to remove noise from the image. In this paper, Linear Contrast enhancement is used which linearly expands the original digital values of the remotely sensed data into a new distribution. The enhanced image is shown in Figure 3.

Image Segmentation:

From the enhanced image, the tumor has to be detected using Image Segmentation algorithm. In [13], segmentation algorithms were compared and Marker Controlled Watershed Segmentation found to be suitable. The Marker Controlled Watershed Segmentation algorithm is used to segment unique boundaries from an image [13]. The segmented part is shown in Figure 4, in which the features are extracted from it.

Feature Extraction:

Feature extraction is a method of capturing visual content of images for indexing and retrieval. Feature extraction is used to denote a piece of information which is relevant for solving the computational task related to a certain application. There are two types of texture features measure. They are first order and second order. In the first order, texture measures are statistics calculated from an individual pixel and do not consider pixel neighbor relationships. The intensity histogram and intensity features are first order calculation. In the second order, measures consider the relationship between neighbor relationships. The GLCM is a second order texture calculation. In this work, GLCM texture features are extracted from the given input image.

GLCM:

A gray level co-occurrence matrix (GLCM) or co-occurrence distribution (less often co-occurrence matrix or co-occurrence distribution) is a matrix or distribution that is defined over an image to be the distribution of co-occurring values at a given offset. A GLCM is a matrix where the number of rows and columns is equal to the number of gray levels, G, in the image The use of statistical features is therefore one of the early methods proposed in the image processing literature. Haralick [14] suggested the use of co-occurrence matrix or gray level co-occurrence matrix. It considers the relationship between two neighboring pixels, the first pixel is known as a reference and the second is known as a neighbor pixel. Given an image I, of size N×N, the co-occurrence, matrix P can be defined as:

$$P(i,j) = \sum_{x=1}^{N} \sum_{y=1}^{N} \begin{cases} 1, \text{ if } I(x,y) = I \text{ and } I(x+\Delta x, y+\Delta y)=j \\ 0, \text{ otherwise} \end{cases} \quad \text{------ 1}$$

For the offset (Δx, Δy), is specifying distance between the pixel-of-interest and its neighbor.

Steps for GLCM:
Step 1: Read the command line from the users
Step 2: Read the content of the image from .bmp file
Step 3: Calculate the co-occurrence matrix
Step 4: Calculate Haralick texture features
Step 5: Save acquired information to a database file.

The Features that are extracted from the images are shown in Table I:

Table I. GLCM Features

Moment	Formulae
Energy	$\mu = (1/MN) * \sum_{i=1} \sum_{j=1} p(i,j)$
Contrast	$f_2 = \sum_{n=0}^{N_g-1} n^2 \left\{ \sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} p_{d,\theta}(i,j) \right\}$, where $n = \lvert i-j \rvert$
Entropy	$f_3 = \sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} p_{d,\theta}(i,j) log\,(p_{d,\theta}(i,j))$
Correlation	$f_5 = \sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} p_{d,\theta}(i,j) \frac{(i-\mu_x)(j-\mu_y)}{\sigma_x \sigma_y}$
Homogeneity	$\sum_{i,j=0}^{N-1} \frac{P_{ij}}{1+(i-j)^2}$

The values are obtained for various tumor cases are shown. The values obtained identify the classification of tumor as benign or malignant.

SVM Classifier:

Support Vector Machine (SVM) is supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. The original SVM algorithm was invented by Vladimir N. Vapnik and the current standard incarnation (soft margin) was proposed by Vapnik and Corinna Cortes in 1995. The basic SVM takes a set of input data and predicts, for each given input, which of two possible classes forms the output. The classification process is divided into the training phase and the testing phase. The known data is given in the training phase and unknown data is given in the testing phase. The accuracy depends on the efficiency of classification.

Implementation:

The Implementation for the proposed system is shown in Figure 2, 3, 4. The Home screen of the System is shown in Figure 2. Selecting the Image preprocessing button, the image is loaded which is then preprocessed (Figure 3). The preprocessed image is segmented and the features are obtained immediately. (Figure 4).

Figure 2. Initial Screen

Figure 3. Image Preprocessing

After a series of operations of the Marker Controlled Segmentation Algorithm, the segmented tumor is obtained in Figure 4.

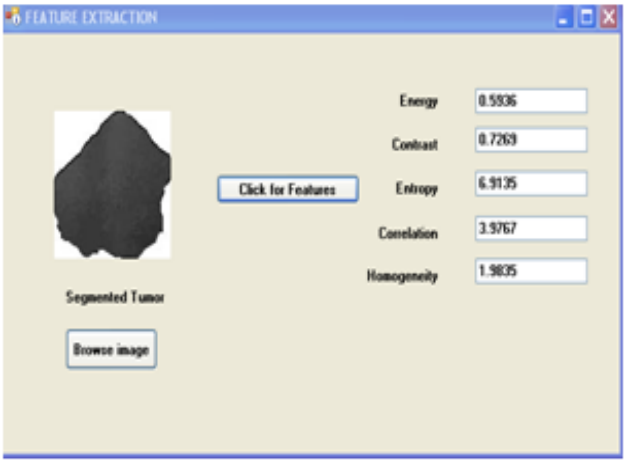

Figure 4. Image Feature Extraction

MEASURES OF PERFORMANCE EVALUATION

Different measures are used to evaluate the performance of the system. The measures used are Classification Accuracy (AC) and Mathews Correlation Coefficient (MCC). These values are calculated from the Confusion Matrix. A confusion matrix (Kohavi and Provost, 1998) contains information about actual and predicted classifications done by a classification system. Performance of such systems is commonly evaluated using the data in the matrix. The following table shows the confusion matrix for a two class classifier.

Table: 2 Confusion Matrix

		Predicted	
		Negative	***Positive***
Actual	Negative	TN	FN
	Positive	FP	TP

TN (True Negative) – Correct Prediction as normal
FN (False Negative) – Incorrect prediction of normal
FP (False Positive) – Incorrect prediction of abnormal
TP (True Positive) – Correct prediction of abnormal.
From the confusion matrix, accuracy (AC) can be obtained:

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+FP+TN+FN)} \quad \text{---------- 2}$$

The Matthews correlation coefficient (MCC)is used in machine learning as a measure of the quality of binary (two-class) classifications. The MCC is in essence a correlation coefficient between the observed and predicted binary classifications; it returns a value between −1 and +1. A coefficient of +1 represents a perfect prediction, 0 no better than random prediction and −1 indicates total disagreement between prediction and observation. The MCC is calculated using:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)\,(TP+FN)\,(TN+FP)\,(TN+FN)}} \quad \text{--- 3}$$

Sensitivity and specificity are terms used to evaluate a clinical test.

The sensitivity of a clinical test refers to the ability of the test to correctly identify those patients with the disease which is calculated from equation 4.

Sensitivity: TP / (TP+FN) ---------- 4

The specificity of a clinical test refers to the ability of the test to correctly identify those patients without the disease which is calculated from equation 5.

Specificity: TN / (TN + FP) ---------- 5

RESULTS AND DISCUSSION

For the proposed work 27 images were chosen randomly. Texture Features are obtained for the segmented part of the tumors (Figure 4). GLCM features are extracted and its classification was obtained. From Table III, we observe the feature values for the various sample images.

Table: 3 Feature Extraction

Feature	Img1 (normal)	Img2 (normal)	Img3 (abnormal)	Img4 (abnormal)
Energy	0.1453	0.1961	0.5936	0.7214
Contrast	0.1904	0.2661	0.7269	0.8175
Entropy	4.9486	5.0543	6.9135	7.4569
Correlation	2.2454	2.5357	3.9767	4.1253
Homogeneity	1.1227	1.2647	1.9835	2.0626

From Table III, the images are classified as normal and abnormal using SVM Classifier. Also the graph shown in Figure 5 represents the statistical feature values for benign and malignant lesions of oral cancer.

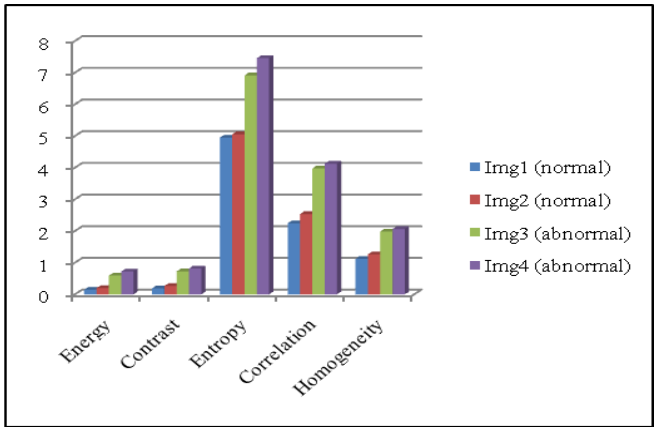

Figure 5 Performance Analysis

The confusion matrix is shown in Table IV.

Table: 4 Confusion Matrix for GLCM features

		Predicted	
		Negative	Positive
Actual	Negative	12 (TN)	1 (FN)
	Positive	1 (FP)	13 (TP)

Accuracy, Mathews Correlation Coefficient, Sensitivity and Specificity are calculated using the values from Table IV and equations 2,3,4 and 5. The evaluation results are obtained as follows:

Table 5: Evaluation Results

Feature	AC (%)	MCC (-1 to +1)	Sensitivity (%)	Specificity (%)
GLCM	92.5	0.85	92.85	92.30

From Table V, it is observed that the accuracy obtained is 92.5% and Mathews correlation coefficient (between -1 and 1) as 0.85. It is also noted that the Sensitivity and Specificity obtained is 92.85% and 92.30%.

CONCLUSION AND FUTURE WORK

In this work, the images are captured and the series of operations are performed to identify the classification as normal or abnormal. The tumor is segmented using Marker Controlled Watershed segmentation and features are extracted using GLCM. Further SVM classifier is used to identify the classification. Accuracy obtained for GLCM feature extraction is 92.5% and MCC is 0.85. In future, the classification performance of several classifiers will also be compared to find the best classifier.

REFERENCES

[1] Konstantinos P. Exarchos1, Yorgos Goletsis, Dimitrios I. Fotiadis1, "Unification of heterogeneous data in the prediction of oral cancer reoccurrence" in the AIAI 2009 Workshop proceedings, pp 24 – 35.

[2] National Institute of Public Health, February, 2011.

[3] Petra Wilder Smith, "Early Detection of Oral Cancer", TRDRP Research for a Healthier California, www.trdrp.org.

[4] Radha Sharma, "Oral Cancer goes viral", Times of India, 27th November 2012, http://articles.timesofindia.indiatimes.com/keyword/oral-cancer.

[5] "Oral Cancer understanding your diagnosis", Canadian Cancer Society.

[6] Lalit Gupta, Sarif Kumar Naik, Srinivasan Balakrishnan, "A new feature selection and classification scheme for screening of oral cancer using laser induced fluorescence", Proceedings of the First International Conference on Biometrics (ICMB'08), pp 1-8.

[7] Sebastian Steger, Marius Erdt, Gianfranco Chiari and Georgios Sakas, "Feature Extraction from Medical Images for an oral cancer reoccurrence prediction environment", World Congress on Medical Physics and Biomedical Engineering, September 7 - 12, 2009, Munich, Germany.

[8] Micheletti A, G. Landini, "Size functions applied to the statistical shape analysis and classification of tumor cells", In: Proceedings of ECMI2006 Springer, 2007.

[9] M. Muthu Rama Krishnan, Chandran Chakraborthy, Ajoy Kumar Ray, "Wavelet based texture classification of oral histopathological sections", International Journal of Microscopy, Science, Technology, Applications and Education, pp 897-906.

[10] G. Landini. "Quantitative analysis of the epithelial lining architecture in radicular cysts and odontogenic keratocysts." Head & Face Medicine 2, 2006.

[11] Jadhav. A.S, S.Banerjee, P.K.Dutta, R.R. Paul, M. Pal, P. Banerjee, K. Chaudhuri, J. Chatterjee "Quantitative analysis of histopathological features of precancerous lesion and condition using Image Processing Techniques", Proceedings of the IEEE Symposium on Computer-Based Medical Systems 02/2006.

[12] K.V.Kulhalli, V.T.Patil, V.R.Udupi, "Image Processing for Computer Aided Diagnosis of Cancer", International Conference on Advances in Computing and Management 2012 (ICACM 2012) 297 – 301.

[13] K. Anuradha, Dr.K. Sankaranarayanan, "Detection of Oral Tumors using Marker Controlled Segmentation", International Journal of Computer Applications, Vol. 52, No.2, August 2012. pp 15 -18.

[14] K. Shanmugam R. M. Haralick and I. H. Dinstein, "Textural features for image classification" IEEE Transactions on Systems, Man and Cybernetics 3 (1973), 610 - 621.

Short Bio Data of the Authors

K. Anuradha completed B.Sc and MCA from Bharathiar University, Coimbatore. Presently pursuing Ph.D (Computer Science). Her areas of Interest are Image Processing and Computer Graphics.

Dr.K. Sankaranarayanan completed his B.E. (Electronics and Communication Engineering) in 1975 and M.E. (Applied Electronics) in 1978 from P.S.G. College of Tech., Coimbatore under University of Madras. He did his Ph.D. (Biomedical Digital Signal Processing and Medical Expert System) in 1996 from P.S.G. College of Technology, Coimbatore under Bharathiar University. His areas of interest include Digital Signal Processing, Computer Networking, Network Security, Biomedical Electronics, Neural Networks and their applications and Opto Electronics.

