International Journal of Application or Innovation in Engineering & Management (IJAIEM)

1 author:

Anuradha .K

Karpagam Academy of Higher Education

7 PUBLICATIONS   **0** CITATIONS

Comparison of Feature Extraction Techniques to classify Oral Cancers using Image Processing

K. Anuradha[1], Dr. K. Sankaranarayanan [2]

[1]Research Scholar, Karpagam University, Coimbatore, Tamilnadu
[2] Dean, Sri Ramakrishna Institute of Technology, Coimbatore, Tamilnadu, India

Abstract
Oral cancer is the sixth most vulnerable cancer in the world, which accounts for over thirty percent of all cancers reported in the country and oral cancer control is quickly becoming a global health priority. In this work, a system is developed to segment, extract features and classify cancers. Later, a comparison is made. The proposed system consists of five steps. First, the images are enhanced and the Region of Interest (ROI) is segmented using Marker Controlled Watershed Segmentation. Feature Extraction methods like Gray Level Co-occurrence Matrix (GLCM), Intensity Histogram and Gray Level Run Length Matrix (GLRLM) are used to extract features from ROI. Next, classification is made using Support Vector Machine (SVM) classifier to classify the tumor as benign or malignant mass and a comparative study is performed to identify the best feature extraction technique.

Keywords: Marker Controlled Watershed Segmentation (MCWS), Gray Level Co-occurrence Matrix (GLCM), Gray Level Run Length Matrix (GLRLM), Support Vector Machine (SVM), Feature Extraction

1. INTRODUCTION

Cancer has become one of the leading causes of death in India. Oral cancer has high mortality ratio among all malignancies. It constitutes 17% of all cancers in males and 10.5% of all cancers in females making it the commonest cancers in males and the third commonest cancer among females [1]. In 96% of oral cancer cases, the tumor starts as small lesions in the mouth. These lesions are often visible, but only some actually become cancer. At present, the only way for a dentist to know whether an oral lesion is cancerous or not is to perform a biopsy. That however poses its own problems. When a painful biopsy turns out to be non cancerous, patients are less likely to cooperate with future biopsies. Highest rates are reported in South Asian Countries such as India and Sri Lanka [2]. "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" [3]. Early detection of Oral cancer improves survival rate to 80% as against late detection where in the survival rate drops to 30%. Overall the 5% year survival rate is only 50% [4]. There are many suspicious lesions; among these are Oral Lichenoid, Oral Sub mucous Fibrosis (OSF) and Oral Leukoplakia. Out of these Oral lichenoid are harmless while oral leukoplakia and OSF has a potential to transform into cancer. OSF is an insidious chronic, progressive, precancerous condition with a high degree of malignant potentiality [5]. Actually, a precancerous state generally depicts mixed features of normalcy as well as pro or pre malignancy. These cancers look visually similar. It is difficult for the oral oncologists to differentiate them. The only way to differentiate them is to diagnose through biopsy and histopathological examination. This procedure is expensive and cause inconvenience to the patients. Visual Inspection of oral cancer is highly subjective based on clinical features and it is often goes unnoticed at early stage. Though oral tumors are detected easily, identification becomes difficult at initial stages. In this study, an approach for classification of cancers using feature extraction techniques is proposed and a comparison is made to identify the best technique. The proposed system is shown in Figure.1.

The paper is structured as follows: Related Works are discussed in Section 2. Section 3 describes the methodology of the proposed system. Section 4 shows the Implementation of the system. Section 5 presents the performance evaluation. Section 6 presents the experimental results. Finally, section 7 describes the conclusion of the proposed system.
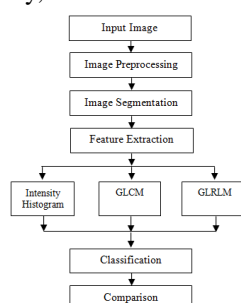

Figure 1 Proposed System

2. RELATED WORK

In 2006, G. Landini et al. [6] analysed epithelial lining architecture in radicular cysts and odontogenic keratocysts applying image processing algorithms to follow 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 [7] carried out segmentation of the Histological OSF images using region growing and hybrid segmentation algorithm. Misclassification rate was calculated for both the algorithms. Finally, Hybrid Segmentation method found to be suitable for segmentation of cancers in OSF images.

In 2007, Naik et al [8] performed a comparative study of different feature extraction methods for the classification of LIF spectra recorded from oral tissue. Features used were Principal Component Analysis, Linear Discriminant Analysis and Discrete Wavelet Transform. Among these Discrete Wavelet Transform features gives better classification accuracy than other features. Support Vector Machine classifier was used for classification.

In 2008, Tathagata Ray et al [5] segmented OSF into layers. Hybrid segmentation algorithm (HSA) yields better results of segmentation when compared to Region Growing Algorithm (RGA). The HSA provides good accuracy. But the algorithm is not fully automatic due to the choice of two thresholds.

In 2009, Sebastian Steger et al [9] 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.

A. Chodorowski et al. [10] classified oral lesions using true color images. Five different color representations and four common classifiers are used to classify oral lichenoid and oral leukoplakia. Best classification accuracy was achieved in HSI color system and Linear discriminant function.

M. Muthu Rama Krishnan et al [11] 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 with feature selection using Kullback – Leibler (KL) method.

In 2012, 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.

3. METHODOLOGY

Referring Figure 1, the proposed work is carried out in various stages. Dental X – rays are digitized and given as the input. The First step is Image Preprocessing which is used to remove the noise from the image. The enhanced image is segmented to detect tumor. Intensity Histogram, GLCM and GLRLM methods are used to extract features from the image. Next, Support Vector Machine (SVM) is used to classify the tumor as benign and malignant. Later, a comparative study is performed to identify the suitable feature extraction.

3.1 Image Preprocessing

The first step is the Image Preprocessing. The input image (Figure 2) 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 preprocessed image is shown in Figure 3.
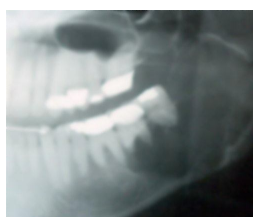

Figure 2 Input Image   **Figure 3**. Enhanced

International Journal of Application or Innovation in Engineering & Management (IJAIEM)
Web Site: www.ijaiem.org Email: editor@ijaiem.org, editorijaiem@gmail.com
Volume 2, Issue 6, June 2013   **ISSN 2319 - 4847**

3.2 Image Segmentation

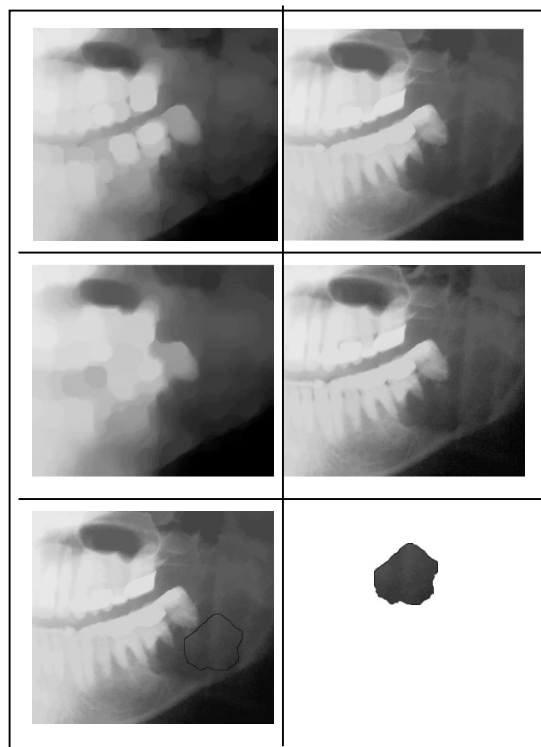From the enhanced image, the tumor is detected using Image Segmentation algorithm.

Figure 4: Segmentation Results

In [13], segmentation algorithms were compared and Marker Controlled Watershed Segmentation (MCWS) was found to be suitable. The MCWS algorithm is used to segment unique boundaries from an image [13]. Segmentation Results are shown in Figure 4

3.3 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 certain application system. There are two types of texture feature measures. They are given as first order and second order measures. 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 calculations. In the second order, measures consider the relationship between neighbors The GLCM is a second order texture calculation. In this work, Intensity Histogram, GLCM and GLRLM texture features are extracted from the given input image.

3.3.1 Intensity Histogram

A frequently used approach for texture analysis is based on statistical properties of Intensity Histogram. A histogram is a statistical graph that allows the intensity distribution of the pixels of an image, i.e. the number of pixels for each luminous intensity, to be represented. By convention, a histogram represents the intensity level using X-coordinates going from the darkest (on the left) to lightest (on the right). Thus, the histogram of an image with 256 levels of grey will be represented by a graph having 256 values on the X-axis and the number of image pixels on the Y-axis. The histogram graph is constructed by counting the number of pixels at each intensity value.

3.3.2 Gray Level Co – occurrence Matrix

A gray level co-occurrence matrix (GLCM) 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. The features extracted from GLCM are Energy, Contrast, Entropy, Correlation and Homogeneity [15].

International Journal of Application or Innovation in Engineering & Management (IJAIEM)
Web Site: www.ijaiem.org Email: editor@ijaiem.org, editorijaiem@gmail.com
Volume 2, Issue 6, June 2013   **ISSN 2319 - 4847**

3.3.3 Gray Level Run Length Matrix

A gray level run –length matrix (GLRLM) method is a way of extracting higher order statistical texture measures. A set of consecutive pixels with the same gray level, collinear in a given direction, constitute the gray level run. The run length is the number of pixels in the run and the run length value is the number of times such a run occurs in an image. The GLRLM is a two dimensional matrix in which each element $p(i, j | \theta)$ gives the total number of occurrences of runs of length "j" at gray level "i" in a given direction θ [16].

3.4 Classification

The final step of the proposed system is classification. SVM classifier is used to classify the tumor as benign or malignant.

3.4.1 SVM Classifier

Support Vector Machine (SVM) is a supervised learning model 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, the best 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.

4. IMPLEMENTATION

The implementation of the proposed system is shown in Figure 5, 6 and 7. The Home screen of the system is shown in Figure 5. Selecting the Image preprocessing button, the image is loaded which is then preprocessed (Figure 6). The preprocessed image is segmented and the segmentation results obtained shown in Figure 4
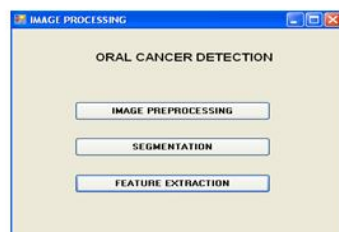

Figure 5 Initial Screen   **Figure 6** Image Preprocessing
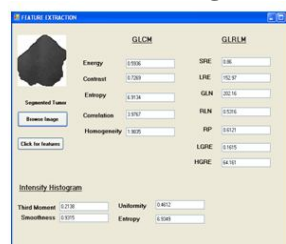

Figure 7 Feature Extraction

5. MEASURES OF PERFORMANCE EVALUATION

Different measures are used to evaluate the performance of the system. The measures used are Classification Accuracy (AC), Mathews Correlation Coefficient (MCC), Sensitivity (SN), Specificity (SP) and Positive Predictive Value (PPV). These values are calculated from the Confusion Matrix. A confusion matrix (Kohavi and Provost, 1998) [17] 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 (Table 1) shows the confusion matrix for a two class classifier.

Table 1: 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 shown in Table 1, Accuracy (AC) can be obtained as:

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+FP+TN+FN)} \tag{1}$$

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:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{2}$$

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 3.

$$\text{Sensitivity} = TP / (TP+FN) \tag{3}$$

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 4.

$$\text{Specificity} = TN / (TN + FP) \tag{4}$$

The Positive Predictive Value (PPV) or precision rate is the proportion of positive test results that are the true positives (such as correct diagnosis). The PPV is given by,

$$PPV = TP / (TP + FP) \tag{e5}$$

The performance of the system is examined by demonstrating correct and incorrect patterns. They are defined in Table 2. The higher value of both sensitivity and specificity shows better performance.

6. RESULTS AND DISCUSSION

For the proposed work, 50 images were chosen randomly. Features are extracted and its classification was obtained. The constructed feature sets are separately tested using the SVM classifier. The computational results are presented.

Table 2: Matrix for all three techniques

Value	Intensity Histogram	GLCM	GLRLM
TP	24	24	24
FP	2	0	1
FN	4	2	3
TN	20	24	22

Table 3: Evaluation Results

Measure	Intensity Histogram	GLCM	GLRLM
AC	88%	96%	92%
MCC	0.76	0.92	0.84
SN	85%	92.71%	88%
SP	90%	100%	95.45%
PPV	92.31%	100%	96%

Accuracy, Mathews Correlation Coefficient, Sensitivity, Specificity and Precision Rate are calculated using the values from Table 2 and equations 1, 2, 3, 4 and 5. The evaluation results are obtained as follows:

From Table 3, it is observed that the texture feature extracted using different models shows different accuracy. *The individual performance measures are shown in Figure 8, 9, 10 and the overall performance measure is shown in Figure 11*
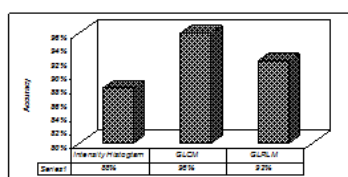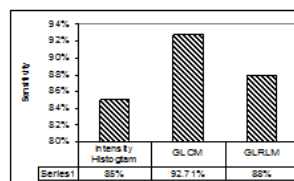

Figure 8 Accuracy
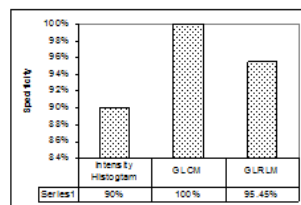

Figure 9 Sensitivity
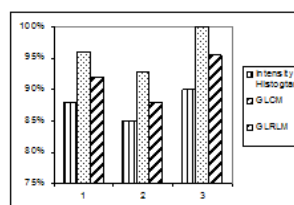

Figure 10 Specificity

Figure 11 Performance Measures of all methods

The results from Table 3 shows that the Intensity Histogram gives a classification rate of 88%, statistical based features such as GLRLM and GLCM shows a classification rate of 92% and 96% respectively.
From Figure 8, 9, 10 and 11, it is observed that GLCM technique gives a good result of classification and hence GLCM is the best Feature extraction technique.

7. CONCLUSION

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, GLRLM and Intensity Histogram. Further SVM classifier is used for classification. Accuracy obtained for GLCM feature extraction is 96% and MCC is 0.92. GLCM gives a better performance when compared with other techniques

References

[1] Iype EM, Pandey.M, Mathew.A, Thomas.G, Sebastian.P, Nair.MK, "Oral cancer among patients under the age of 35 years", Journal of Postgrad. Medicine, 2001; 47: 171.
[2] www.mouthcancerfoundation.org

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

[4] V.K.Joshi, "Oral Cancer: a growing concern", Preventive Dentistry, (1), 2006.

[5] Tathagata Ray, D.Shivashanker Reddy, Anirban Mukherjee, Jyotirmoy Chatterjee, Ranjan R.Paul, Pranab K.Dutta, "Detection of constituent layers of histological oral sub – mucous fibrosis: Images using the hybrid segmentation algorithm", Journal of Oral Oncology, No.44, pp 1167 – 1171, 2008.

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

[7] 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.

[8] S.K. Naik, L.Gupta, C. Mittal, and et al, "Optical Screening of oral cancer: Technology and emerging markets", In proc. of 29th Annual Intl.Conf. of the IEEE in Engg. In Medicine and Bio. Soc., 2007.

[9] 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.

[10] A.Chodorowski, U. Mattson, T.Gustavasson, "Oral lesion classification using true color images", Department of Signals and Systems, Chalmers University of Tech., Sweden.

[11] 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.

[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 Tumor based on Marker Controlled Watershed 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.

[15] K. Anuradha, Dr.K. Sankaranarayanan, "Statistical Feature Extraction to classify oral cancers", Journal of Global research in Computer Science", Vol. 4, No.2, February 2013, pp 8 -12.

[16] Fritz Albregtsen, "Statistical Texture Measures computed from Gray Level Run Length Matrices", Image Processing Laboratory, Department of Informatics, University of Oslo, Nov.1995.

[17] R. Kohavi and F. Provost. Glossary of terms, Special Issue on "Applications, of Machine Learning and the Knowledge Discovery Process", Journal of Machine Learning, 30(2–3):271–274, 1998.

AUTHORS

Mrs. K. Anuradha received B.Sc and MCA from Bharathiar University, Coimbatore in the year 2000 and 2003.. Presently pursuing her Ph.D (Computer Science) in Karpagam University. She has published 5 papers in International Journals. She has 9 years of Teaching Experience and currently working as Assistant Professor, Department of Computer Applications, Karpagam College of Engineering, Coimbatore. Her areas of Interest are Image Processing and Computer Graphics. She is a member in various bodies like ISTE, IAENG and ACM.

Dr. K. Sankaranarayanan received his B.E. (ECE) in 1975 and M.E. (Applied Electronics) in 1978 from P.S.G. College of Technology, 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. He has more than 35 years of Teaching Experience. His areas of interest include Digital Signal Processing, Computer Networking, Network Security, Biomedical Electronics, Neural Networks and their applications and Opto Electronics.