ISSN: 0976-9102 (ONLINE)
DOI: 10.21917/ijivp.2016.0170

A COMPARATIVE ANALYSIS OF SINGLE AND COMBINATION FEATURE EXTRACTION TECHNIQUES FOR DETECTING CERVICAL CANCER LESIONS

S. Pradeep Kumar Kenny[1] and S.P. Victor[2]
[1]*Centre for Information Technology & Engineering, Manonmaniam Sundaranar University, India*
E-mail: kenny_isles@yahoo.co.in
[2]*Department of Computer Science, St. Xavier's College, Palayamkottai, India*
E-mail: drspvictor@gmail.com

Abstract

Cervical cancer is the third most common form of cancer affecting women especially in third world countries. The predominant reason for such alarming rate of death is primarily due to lack of awareness and proper health care. As they say, prevention is better than cure, a better strategy has to be put in place to screen a large number of women so that an early diagnosis can help in saving their lives. One such strategy is to implement an automated system. For an automated system to function properly a proper set of features have to be extracted so that the cancer cell can be detected efficiently. In this paper we compare the performances of detecting a cancer cell using a single feature versus a combination feature set technique to see which will suit the automated system in terms of higher detection rate. For this each cell is segmented using multiscale morphological watershed segmentation technique and a series of features are extracted. This process is performed on 967 images and the data extracted is subjected to data mining techniques to determine which feature is best for which stage of cancer. The results thus obtained clearly show a higher percentage of success for combination feature set with 100% accurate detection rate.

Keywords:

Cervical Cancer, Feature Extraction, Texture Features, Content Based Image Retrieval

1. INTRODUCTION

In the modern world human beings are affected by various diseases and disorders. Some of these diseases or disorders result in a very high fatality rate in a particular race, gender, culture, or region. Cancer is one such disease that affects many people worldwide. Cancer is formed when the mitosis process goes berserk by splitting cells even though there is no room for the cell hence it progresses to the nucleoplasm stage. The cells keep accumulating on top of each other and form tumours which are called pre-cancers which eventually form the real cancer. This process can form in any part or organ of the body. The name of the cancer often reflects the place where the cancer has originated [1][2].

In this paper we are dealing with a type of cancer that affects women predominately in third world countries and is the third deadly cancer in women. It forms in the cervix of a women's conception channel and hence it is named cervical cancer. According to estimates there are about 12.7 million cancer cases around the world out of which 530,000 are in developing countries. Around 85% of this estimate represents cases from third world countries [3].

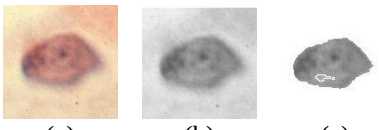
The reason for such high numbers from developing countries or so called third world countries can mean only one thing, people do not have medical awareness or they don't have the access to proper medical care. Another contributing factor to such high number is the lack of pathologists to scan each and every woman. To overcome such problems an automated approach could help not only to speed up but also help in detecting people with suspect of cancer accurately.

A pathologist normally swipes and obtains a sample of the cervix, examines it under a microscope and tells the verdict. This same process can be replicated by an automated system. A lot of work has been carried out in this front and is discussed in the literature [4]-[10].

In this work we have set out to compare two feature extraction techniques in tackling this problem so that the many solutions available to this problem can be refined. Section 2 gives a description of the two techniques taken for discussion. Section 3 gives the comparative results of the two techniques and section 4 concludes the comparative work been done here.

2. COMPARATIVE ANALYSIS

Any feature extraction procedure carried out follows three basic steps viz Image acquisition, image segmentation and feature extraction. Image acquisition is done by digitally converting the pap smear slide sample into a digital image. Once the image is acquired the next basic step is to segment the image. Both the techniques that are to be compared here use a Multiscale Morphological Watershed Segmentation using Gradient and Marker extraction [11] technique to segment the samples. The technique employed here helps to preserve the edge by implementing a gradient operation rather than an edge operation. A gradient is the difference of an image being dilated and eroded and hence helps in preserving the edge of the image. Further a technique called marker extraction puts markers so that when we apply watershed segmentation over segmentation is avoided. The resultant segmented image has all the original information of the cell intact. The output of this technique is shown in Fig.1.



(a) (b) (c)

Fig.1. (a) Original Image (b) Grayscale Image (c) Segmented Image

2.1 ANALYSIS FOR TEXTURAL FEATURES IN NUCLEI OF CERVICAL CYTO IMAGES

This technique extracts four features from the segmented image namely mean, standard deviation, skewness, and kurtosis. A feature is the contrast of an image surface with which the brain can differentiate between rough and a smooth surface. Hence

when a pathologist scans the cell with his eyes he is just looking for patterns in the texture composition that aren't in place. Here in this technique four features are extracted and compared with a manual diagnosis for the various stages of cancer [12]. This technique provides a basic insight of how an automated system can be created. This work is based on the concept of content based image retrieval where image are retrieved by the contents they posses and not just the name of the file in which they are stored. This concept when used with medical image processing is really a captivating idea.

This technique also concludes with a possibility of adding more features so that this technique can be further strengthened [12].

2.2 COMBINATION FEATURE SET FOR THE DETECTION OF CERVICAL CYTO IMAGES

This technique improves accuracy by working on the suggestions provided by the previous technique and also by removing a critical drawback in the previous technique. The reason for such a low percentages in the previous technique is that each image has a different composition and is affected by lighting effects and how it is acquired. So when you have to detect the cancer stage using a single feature technique it is going to heavily depend on the image. If the image is not segmented properly or if it has a smear which shows as a blob then your result is going to be wrong. Also the previous technique shows that more than one feature can be employed in detecting the stage of cancer [13].

Using these information's this technique has used the wisdom of the crowd. Instead of judging the cancer stages using a single feature a group of features are computed and their collective knowledge is harnessed. The concept is similar to our body's neural receptors. The intensity of pain is directly proportional to the number of neural inputs. This same concept is used here. A combination of features are calculated for a single stage of cancer and depending upon the number of positives we can predict for certain whether the cancer has progressed to a particular stage or not [13].

To derive the combination of features this technique has implemented an array of data analysis concepts on the 967 images from the database. From each image 34 features are extracted. Then a ranking system is used to compute the prominent features both in terms of features and in terms of stages. Outlier detection helps in removing features that are not suited for that particular image and a hierarchical clustering technique is used to find the best features for a particular stage. By intersecting these two data analysis we can obtain a refined combination set of features for a particular stage of cancer as shown in Table.1 [13].

Table.1. Combination Set of Features

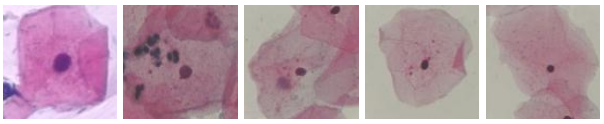
Normal Cells
Superficial Squamous
Entropy, Entropy of GLCM, Sum of Variance, Measurement of Correlation 1
Intermediate Squamous
Sum, Entropy of GLCM, Homogeneity, Sum of entropy, Measurement of Correlation 1
Columnar
Sum, Max, Homogeneity, Sum of square variance, Difference of variance, Directionality
Abnormal Cells
Mild Dysplasia
Nucleocytoplasmic Ratio (NCR), Max, Correlation, Homogeneity, Maximum probability, Information Measurement of Correlation 1
Moderate Dysplasia
Hyperchromasia, Standard Deviation, Max, Entropy, Autocorrelation, Contrast, Sum of Variance, Sum of average, Sum of square variance, Difference variance
Severe Dysplasia
Standard Deviation, Entropy, Cluster Prominence, Sum of Variance, Sum of average, Directionality Tamura feature Contrast
Carcinoma in Situ
Sum, Standard Deviation, Entropy of GLCM, Sum of average, Sum of square variance, Difference entropy, Inverse difference normalized, Directionality

2.3 DATASET

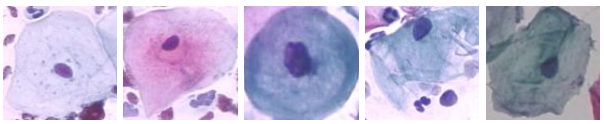
The first technique has taken images from a normal image atlas and has proved its results. However the second technique has taken its images from the Herlev Dataset [14]. This dataset contains a total of 917 images which is been pre-classified broadly into two subsets namely Normal and Abnormal and in each subset it has been further sub divided into 3 and 5 subsets making a total of 7 sub sets or classes. A sample of a few images of various subsets is shown in Fig.2 [14].

Normal Cells

Superficial Squamous



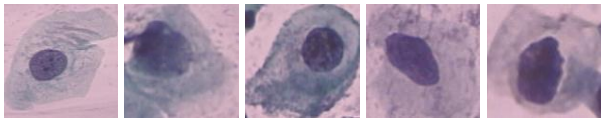
Intermediate Squamous



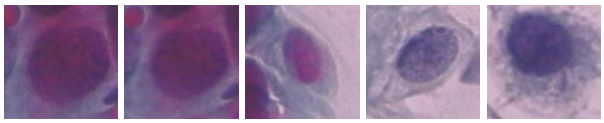
Columnar

Abnormal Cells

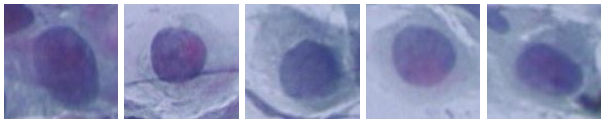
Mild Dysplasia



Moderate Dysplasia



Severe Dysplasia



Carcinoma in Situ

Fig.2. Sample images of the Herlev Dataset

This dataset has manually classified the images and hence it would be easier for us to find the perfect technique for an automated system. Also the large size of the dataset gives us an ideal platform to accurately test these two techniques. In this work for testing the effectiveness of these two techniques we have tested them with these images.

3. EXPERIMENTAL RESULTS

For our analysis we first need a base threshold so that we can see if the features obtained are true or not. Although the two techniques discussed here have proposed feature sets that help in detecting the cancer none have given a threshold value for the feature. Hence to find the threshold we have borrowed a technique from the combination feature technique. The combination analysis technique used an outlier removal technique to remove features that won't help in detecting features of a particular class [13]. We have used the output of this technique and obtained the min and max of the available data which will be used as the threshold. This is given in the Eq.(1) and Eq.(2).

$$\text{Min\_Thres} = \min(\text{OR}(f(t))) \quad (1)$$

$$\text{Max\_Thres} = \max(\text{OR}(f(t))) \quad (2)$$

where,

Min_thres = Minimum Threshold for a feature.

Max_thres = Maximum Threshold for a feature

OR = OR represents the data set with outliers removed

$f(t)$ = represents the features computed for an image class

By computing Eq.(1) and Eq.(2) we get the min and max threshold has shown in Table.2.

Table.2. Threshold of Features

Sl. No	Normal Cells		
	Superficial Squamous		
	Feature	**Min**	**Max**
1	Entropy	1	200
2	Entropy of GLCM	0.01	0.10
3	Sum of Variance	0.974	1.109
4	Measurement of Correlation 1	-0.84	-0.74
Sl. No	**Intermediate Squamous**		
	Feature	**Min**	**Max**
1	Sum	24	71
2	Entropy of GLCM	0.04	0.21
3	Homogeneity	0.994	0.999
4	Sum of entropy	0.04	0.20
5	Measurement of Correlation 1	-0.8	-0.7
Sl. No	**Columnar**		
	Feature	**Min**	**Max**
1	Sum	23	93
2	Max	38	200
3	Homogeneity	0.92	0.99
4	Sum of square variance	4.84	23.47
5	Difference of variance	0.04	0.67
6	Directionality	2.4	4.0
Sl. No	**Abnormal Cells**		
	Mild Dysplasia		
	Feature	**Min**	**Max**
1	Nucleocytoplasmic Ratio (NCR)	0.04	0.6
2	Max	79	659
3	Correlation	0.86	0.95
4	Homogeneity	0.94	0.99
5	Maximum probability	0.60	0.95
6	Information Measurement of Correlation 1	-0.8	-0.5
Sl. No	**Moderate Dysplasia**		
	Feature	**Min**	**Max**
1	Hyperchromasia	42.72	270.28
2	Standard Deviation	24.11	156.99
3	Max	88	541
4	Entropy	1	167
5	Autocorrelation	1.77	7.19
6	Contrast	0.04	0.49
7	Sum of Variance	1.76	7.29
8	Sum of average	2.3	4.5
9	Sum of square variance	4.75	22.01
10	Difference variance	0.1	0.9
Sl. No	**Severe Dysplasia**		
	Feature	**Min**	**Max**

1	Standard Deviation	14.44	132.99
2	Entropy	1	75
3	Cluster Prominence	7.29	752.84
4	Sum of Variance	1.66	10.33
5	Sum of average	2.39	5.87
6	Directionality	2.82	4.95
7	Tamura feature Contrast	0.0035	0.0597
Sl. No	**Carcinoma in Situ**		
	Feature	**Min**	**Max**
1	Sum	21	90
2	Standard Deviation	16.57	131.37
3	Entropy of GLCM	0.52	1.85
4	Sum of average	2.48	6.50
5	Sum of square variance	5.65	40.0
6	Difference entropy	0.108	0.62
7	Inverse difference normalized	0.88	0.98
8	Directionality	3.37	5.05

A cancer progresses from nucleoplasm stage to precancerous stage and then to the stage of actual cancer. The single feature detection technique [12] follows this premise. The combination feature technique [13] follows a more refined version of this premise. However both follow the same principle. From the analysis of the combination feature [13] technique, it's clear that skewness and kurtosis analysed in the single [12] feature technique is less significant and hence doesn't appear in the combination feature technique [13]. Hence we have analysed the remaining features namely mean (NCR) and standard deviation. To showcase our testing method a small sample of 5 images from each abnormal class of the Herlev dataset is chosen [15] and compared with the thresholds shown in Table.2. The results for these samples are shown in Table.3 to Table.6.

Table.3. Detection Analysis of Light Dysplasia Class

S	1	2	3	4	5	6
	0.12	397	0.95	0.98	0.88	-0.8
	ND	D	D	D	D	D
	0.09	229	0.91	0.98	0.90	-0.78
	D	D	D	D	D	D
	0.25	130	0.89	0.96	0.78	-0.73
	ND	D	D	D	D	D
	0.16	289	0.93	0.98	0.85	-0.78
	ND	D	D	D	D	D
	0.12	200	0.92	0.98	0.88	-0.8
	ND	D	D	D	D	D

where,
S- Light Dysplasia
1-Nucleocytoplasmic Ratio (NCR)
2-Max
3-Correlation
4-Homogeneity,
5-Maximum probability,
6-Information Measurement of Correlation 1
D - Detected
ND - Not Detected

Table.4. Detection Analysis of Moderate Dysplasia Class

S	1	2	3	4	5	6	7	8	9	10
	225	180.09	677	1	3.90	0.94	3.90	3.48	11.004	0.10
	D	ND	ND	D	D	ND	D	D	D	D
	76	40.46	142	62	3.20	0.22	3.26	3.04	8.53	0.22
	D	D	D	D	D	D	D	D	D	D
	270.28	171.67	652	1	3.73	0.10	3.73	3.38	10.46	0.10
	D	ND	ND	D	D	D	D	D	D	D
	90	37.92	127	68	2.44	0.11	2.45	2.62	7.46	0.11

	D	D	D	D	D	D	D	D	D	ND
	44.77	28.90	101	81	3.14	0.22	3.20	2.92	9.43	0.22
	D	D	D	D	D	D	D	D	D	ND

where,
S-Moderate Dysplasia 4-Entropy 8-Sum of average ND-Not Detected
1-Hyperchromasia 5-Autocorrelation 9-Sum of square variance
2-Standard Deviation 6-Contrast 10-Difference variance
3-Max 7-Sum of Variance D-Detected

Table.5. Detection Analysis of Severe Dysplasia Class

S	1	2	3	4	5	6	7
	131.67	1	21.73	2.90	2.967	3.33	0.0111
	D	D	D	D	D	D	D
	61.54	1	20.81	4.67	3.80	4.39	0.0216
	D	D	D	D	D	D	D
	36.79	23	24.93	3.20	3.089	3.96	0.0123
	D	D	D	D	D	D	D
	68.71	50	18.94	3.24	3.148	4.034	0.015
	D	D	D	D	D	D	D
	66.56	1	29.70	3.45	3.19	4.03	0.0175
	D	D	D	D	D	D	D

where,
S- Severe Dysplasia 3-Cluster Prominence 6-Directionality ND-Not Detected
1-Standard Deviation 4-Sum of Variance 7-Tamura feature Contrast
2-Entropy 5-Sum of average D-Detected

Table.6. Detection Analysis of Carcinoma in situ Class

S	1	2	3	4	5	6	7	8
	41	185.49	1.28	4.31	16.57	0.39	0.98	3.86
	D	ND	D	D	D	D	D	D
	34	91.25	0.92	3.17	9.34	0.27	0.98	3.79
	D	D	D	D	D	D	D	D
	37	165.94	1.13	4.15	13.76	0.38	0.98	4.06
	D	ND	D	D	D	D	D	D
	38	52.25	1.33	3.92	12.71	0.50	0.97	4.54
	D	D	D	D	D	D	D	D
	48	70.35	1.18	6.35	41.66	0.31	0.98	4.73
	D	D	D	D	ND	D	D	D

where,
S-Carcinoma in situ 4-Sum of average 8-Directionality
1-Sum 5-Sum of square variance D-Detected
2-Standard Deviation 6-Difference entropy ND-Not Detected
3-Entropy of GLCM 7-Inverse difference normalized

Those that are within the threshold are marked as 'D' meaning detected and those that are not are marked 'ND' meaning not detected. The summary of detection for the above computed samples is shown in Table.7 and a graphical representation of it is shown in Fig.3.

Clearly as you can see single detection technique fails most of the time and hence brings down the detection rate. On the other hand if you could combine this with the other features as suggested by the combination feature technique you drastically improve the detection rate. This same procedure was performed for all the 917 images in the dataset and the overall detection rate is shown in Table.8 and graphically shown in Fig.4.

Table.7. Detection Rate for the samples

Classes	Single Feature (%)	Combination Feature (%)
Light Dyplasia	20	100
Moderate Dyplasia	60	100
Severe Dyplasia	100	100
Carcinoma in situ	60	100

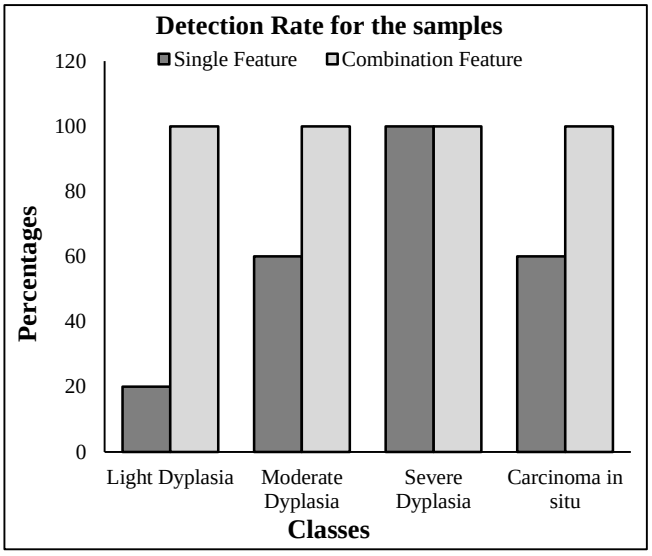


Fig.3. Graphical representation for the samples

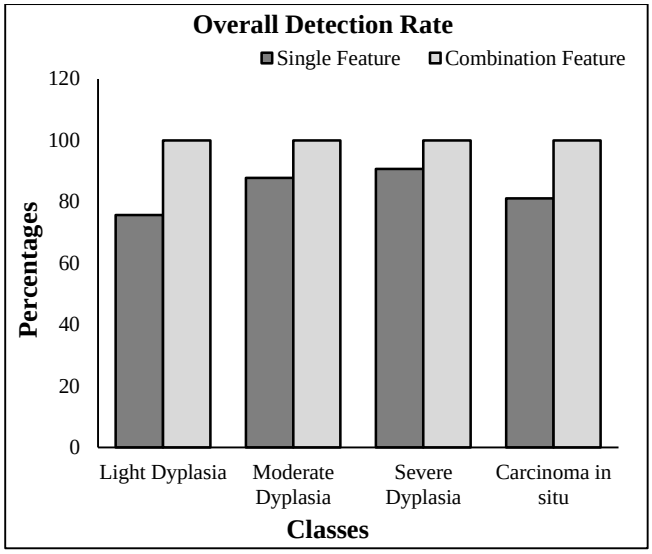


Fig.4. Graphical Representation of the Overall Detection Rate

Table.8. Overall Detection Rate

Classes	Single Feature (%)	Combination Feature (%)
Light Dyplasia	75.71	100
Moderate Dyplasia	87.76	100
Severe Dyplasia	90.66	100
Carcinoma in situ	81.08	100

4. CONCLUSION

The two techniques that are analyzed here are an innovation towards the novel cause of finding an automated system for detecting cervical cancer lesions by themselves. But arming feature extraction with the combined knowledge of group drastically improves the detection rate. Hence we conclude that the combination set of feature technique is far better than the single feature detection technique as shown in the results. In future, as system processing power increases, more features like gabor, wavelet etc., can be incorporated into the combination feature set. Also the robustness towards false detection should also be considered in detail.

REFERENCES

[1] Malcom R. Alison, "*Cancer*", eLS, John Wiley & Sons Ltd., 2001.

[2] Mitosis and Meiosis [Online] http://biocominstitute.org

[3] GLOBOCAN 2008 database (version 1.2), Available at: http://globocan.iarc.fr.

[4] R.C. Bostrom, H.S. Sawyer, and W.E. Tolles, "Instrumentation for Automatically Prescreening Cytological Smears", *Proceedings of the IRE*, Vol. 47, No. 11, pp. 1895-1900, 1959.

[5] P.H. Bartels, and G.L. Wied, "Computer Analysis and Biomedical Interpretation of Microscopic Images: Current

Problems and Future Directions", *Proceedings of the IEEE*, Vol. 65, No. 2, pp. 252-261, 1977.

[6] Y. Srinivasan, E. Corona, B. Nutter, S. Mitra and S. Bhattacharya, "A Unified Model-Based Image Analysis Framework for Automated Detection of Precancerous Lesions in Digitized Uterine Cervix Images", *IEEE Journal of Selected Topics in Signal Processing*, Vol. 3, No. 1, pp. 101-111, 2009.

[7] Amir Alush, Hayit Greenspan, and Jacob Goldberger, "Automated and Interactive Lesion Detection and Segmentation in Uterine Cervix Images", *IEEE Transactions on Medical Imaging*, Vol. 29, No. 2, pp. 488-501, 2010.

[8] P. Mitra, S. Mitra and S.K. Pal, "Staging of Cervical Cancer with Soft Computing", *IEEE Transactions on Biomedical Engineering*, Vol. 47, No. 7, pp. 934-940, 2010.

[9] C. Balas, "A Novel Optical Imaging Method for the Early Detection, Quantitative Grading, and Mapping of Cancerous and Precancerous Lesions of Cervix", *IEEE Transactions on Biomedical Engineering*, Vol. 48, No. 1, pp. 96-104, 2001.

[10] A.N. Esgiar, R.N.G. Naquib, B.S. Sharif, M.K. Bennett and A. Murray, "Microscopic Image Analysis for Quantitative Measurement and Feature Identification of Normal and Cancerous Colonic Mucosa", *IEEE Transactions on Information Technology in Biomedicine*, Vol. 2 , No. 3, pp. 197-203, 1998.

[11] K. Nallaperumal et al., "An Efficient Multiscale Morphological Watershed Segmentation using Gradient and Marker Extraction", *Proceedings of Annual IEEE India Conference*, pp. 1-6, 2006.

[12] K. Krishnaveni, S. Allwin, S. Pradeep Kumar Kenny and G. Mariappan, "Analysis for Textural Features in Nuclei of Cervical Cyto Images", *Proceedings of IEEE International Conference on Computational Intelligence and Computing Research*, pp. 1-4, 2010.

[13] S. Pradeep Kumar Kenny and S. Allwin, "Combination Feature Set for the Detection of Cervical Cyto Images", Sylwan Journal, Vol. 158, No. 6, pp. 423-429, 2014.

[14] MDE Lab, Available at: http://labs.fme.aegean.gr/decision/downloads