IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 11, Issue 1 (May. - Jun. 2013), PP 01-07
www.iosrjournals.org

Detection of Cancer in Pap smear Cytological Images Using Bag of Texture Features

Edwin Jayasingh.M[1], Allwin.S[2]
[1]*Research Scholar,Manonmaniam Sundaranar University,Tirunelveli,India*
[2]*Professor, Infant Jesus College of Engineering,Thoothukudi,India*

Abstract: *We present a visual dictionary based method for content based image retrieval in cervical microscopy images using texture features. The nucleus region in each image is identified by a simple and reliable segmentation algorithm and texture features are extracted from blocks of the region. These features from the entire database are clustered to build a visual dictionary. The histogram of the visual words present in an image is used as the representation of the image. Histogram intersection serves as the distance measure to do content based image retrieval. Experiments were conducted for various block sizes and number of clusters and the results are presented. The task was to identify images of cancerous cells from normal ones. The method offers encouraging results to utmost 90% accuracy. A brief discussion of the results and possible future directions are given*

Keywords - *Cervical Images, Visual Dictionary, Bag of Features, Tamura Features*

I. Introduction

Cervical cancer accounts for approximately 500,000 new cases and 274,000 deaths every year worldwide[1]. It is the second leading cause of death from cancer among women. However significant progress in reducing these deaths have occurred in the last few decades and it can be attributed to increasingly accurate early screening test. The screening tests consist of Papanicolaou test which involves microscopic examination of exfoliated cells from the transformation zone of the cervix. The screening tests still need to be made easy and prevalent to prevent cancer deaths especially in the developing world. Around 80% of all cervical cancer deaths occur in the developing world[2]. In this regard an medical image processing approach towards automatic detection of the presence and level of cancer in test samples becomes highly desirable.

Automatic diagnosis from medical images is done by extracting a certain set of relevant features from the images and comparing them with the features from the images of known cases. Usually a set of images from known cases called the training set is formed and its features are extracted. Normally an image has features of color, texture and shape. Several such features are proposed for cytological images in the past. Human eyes are more sensitive to color variation than the variation of gray levels [3]. Therefore, color images are widely being used in many applications of image processing but in medical imaging, gray level images are still popular. Color features may not be very helpful in this case because most of the color information in the cytology images pertain to the dye present in the preparation rather than the cells themselves. Shape and size features are useful but extracting them depends on accurate segmentation of the cells and nuclei. Segmentation methods abound in medical image processing literature but they suffer from several problems. Accuracy of these algorithms usually depends on the value of several parameters and may not generalize to all types of images. Even a small inaccuracy in segmentation may distort the shape feature resulting in less accuracy. Texture features on the other hand can be extracted easily and accurately. They are highly relevant to the microscopic structures in the cell that are telltale signs of malignant cancer.

In this paper we present a simple and efficient method based on texture features for detection of cancer in cervical images. We use the bag of features approach to build a visual dictionary from the entire corpus of images and then represent each image as a histogram of visual words. Bag of features methods are based on orderless collections of quantized local image descriptors; they discard spatial information and are therefore conceptually and computationally simpler than many alternative methods[4]. In [5],Cai *et al.* presented a visual word weighting factor learning approach for image classification and retrieval. This approach is an evolution of texton based representations and is also influenced by the bag-of-words representation for text classification and retrieval [6]. This methodology avoids the tough problem of segmentation of different objects in the image by representing the complex image contents using characteristic image regions. The features are first extracted from all the images and are clustered to get the centroids which represent the different classes of image contents that are possible. Then every image can be summarized as a histogram by the number of image regions belonging to each cluster. This approach ignores the spatial location information of the image contents which are not relevant in our case. To do classification, a knowledge-based method such as Support Vector Machines (SVM) or a neural network may be trained on the available training set. For image retrieval tasks a suitable distance measure can be used to find the most relevant images from the database for the given query image

The paper is organized as follows: In Section 2 we present the details of the proposed method, results and analysis are presented in Section 3 and Section 4 ends the paper with concluding remarks

II. RELATED WORK

Medical Image processing has given more importance in the area of microscopic images since from 1980's when many analysis and decisions has been done through microscopic image results. The medical cytologists and doctors are greatly in need of good quality images even though the test images were of acceptable range. One of the main problems in analysis was found to be the separation of single cell from the whole cytology image. Identification of this cell and finding their boundaries in an image is found to be more subjective than quantitative. Taking this into account, if the measurements of a single cell properties are defined, then it is easy for the experts to reproduces this in the other cells which paved easy way for identification. Then the automatic grading approach was introduced. Automatic grading is approached as an image segmentation problem to identify those regions that correspond to each grade [7]. When identifying stages of cancer, different rules were followed and everyone followed the tissue architectural features. In [8], the author used neuro blastoma histological slides for processing and identifying pathological regions and which are identified by multi resolution framework.

Automatic image classification, annotation and retrieval are the other popular areas in cytology images. Many transformation features are extracted and then the training classifiers decide in which category it may match. This was evaluated by Orlov et al[9]. Tang et al[10] and Naik et al[11]worked on the design of image similarity measures and makes the system to use the information obtained through semantics. The present work put the challenges of data mining and visualization, comparison and analysis of all features related to cytology images inside the large database, which is referred as bio-image informatics[12]. It is very much complicated in mining of visual patterns from the microscopic images. The bag of features approach is used to learn biased models for automatic image annotation, analyzing relationships between local visual patterns and image categories from a wider perspective, adding an interpretation layer that aims to explain image collection structures and that supports high-level decision making in histology.

III. APPROACHES AND METHODOLOGY

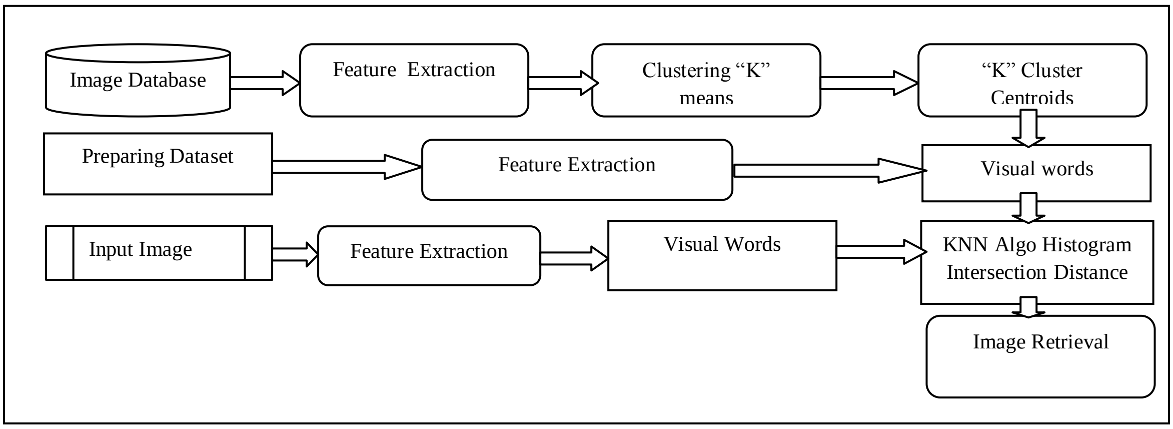

Figure 1 : Overall Architecture

All traditional methods of detection of cancer using medical images uses a segmentation process. But most of these segmentation methods suffer from distinctive disadvantages. Poor accuracy and the need for manual adjustment of parameters makes these methods unfit for automatic detection of cancer. In this work however we rely on a Bag of features approach for automatic detection of cancer in Pap smear images.

3.1 Image Acquisition and Pre-processing

The pap smear samples are collected and are used as the dataset for the experiments. The images are subject to a preprocessing step. The purpose of the pre-processing stage is to remove the noise found in the image samples and makes them well suited for further processing. Normally the nucleus of the cell is found to be darker than other region which could be inverted. The input image is binarized and the morphological closing operation with structuring element size 5 is performed followed by the morphological filling operation. This procedure produces the rough segmentation of cell nuclei

3.2 Extraction of Features

Feature selection is the process of selecting an optimum subset of features from the enormous set of potentially useful features which may be available in a given problem domain [13]. In medical image analysis, selection of precise features among various image modalities plays toughest and challenging task. Some features are predominant in some category of medical images and others do not show that. So finding out the best features for unique images are the very crucial task in medical image processing. By comparing the various features of cancerous images with the normal images one can spot out the severity and the stage of cancer in Pap smear image.

A feature is a characteristic that can capture a certain visual property of an image either globally for the whole image, or locally for objects or regions. Features are numerical values computed from each image properties which are present or absent. We divide the image cell region into the blocks of various sizes of 8,16,24,32 square regions. The features extracted are Tamura's features and gray level co-occurrence matrix (GLCM) based features. Tamura et al took the approach of devising texture features that correspond to human visual perception [14].The Tamura features include coarseness, coarseness histogram 1, coarseness histogram 2, coarseness histogram 3 , directionality and contrast. Energy, Homogeneity and correlation cover the GLCM features. We briefly define the features which are as follows.

An image will contain textures at several scales; *Coarseness* aims to identify the largest size at which a texture exists, even where a smaller micro texture exists [15]. It is calculated depending on the variance of the image It is related to the distances of notable spatial variations of grey levels, that is, implicitly, to the size of the primitive elements (texels) forming the texture. In the first step n averages of $2^k \times 2^k$ image sub windows are calculated around the central pixel

$$A_k(x,y) = \sum_{i=x-2^{k-1}}^{x+2^{k-1}-1} \sum_{j=x-2^{k-1}}^{x+2^{k-1}-1} {}^{I(i,j)}\!/_{2k} \qquad (3.1)$$

Then the absolute differences between pairs of non-overlapping averages in opposing sides, both in the horizontal and vertical directions, are calculated and from that coarseness of the entire image is calculated as

$$F_{crs} = \frac{1}{n \times m} \sum_i^n \sum_j^m S_{max}\,(i,j) \qquad (3.2)$$

Directionality is a global property over a region. It does not differentiate between orientations or patterns but measures the total degree of directionality. In the first step image is convoluted into smaller arrays and the gradient vector at each pixel is computed. By quantizing and counting the pixels with the corresponding magnitude, a histogram can be constructed. Summarization of entire histogram leads to the overall directionality.

$$F_{dir} = \sum_p^{n_p} \sum_{\emptyset \in w_p} \left(\emptyset - \emptyset_p\right)^2 H_D(\emptyset) \qquad (3.3)$$

Contrast aims to capture the dynamic range of grey levels in an image, together with the polarisation of the distribution of black and white. The first is measured using the standard deviation of grey levels and the second the kurtosis $\propto_4$.

$$F_{con} = {}^{\sigma}\!/_{(\propto_4)^n} \ where \ \propto_4 = \ \mu_4/\sigma^4 \qquad (3.4)$$

where μ_4 is the fourth moment about the mean and σ^2 is the variance.

Homogeneity is largely related to the local information extracted from an image and reflects how uniform a region is [16]. The value of the homogeneity at each location of an image has a range from 0 to 1. The more uniform the local region surrounding a pixel is, the larger the homogeneity value the pixel has. Furthermore, using a larger window in the computation of the homogeneity increases smoothing effect, and makes the derivative operations less sensitive to noise [16].

$$F_{homo} = \sum_{i=1}^{L} \sum_{j=1}^{L} \frac{C_d(i,j)}{1+\,|i-j|} \qquad (3.5)$$

where C_d is the co-occurrence matrices.

Energy, also called Angular Second Moment [17] and Uniformity in [16], is a measure of textural uniformity of an image. Energy reaches its highest value when gray level distribution has either a constant or a periodic form. A homogenous image contains very few dominant gray tone transitions have large value for the energy feature. The energy is the sum squared element in the normalized GLCM.

$$F_{energy} = \sum_i \sum_j P_d^2(i,j) \tag{3.6}$$

Correlation is a measure of how connected pixel is to its neighbor over the whole image. It calculates the correlation of each attribute with the label attribute and returns the absolute or squared value as its weight. Suen and Healy[18]used correlation functions across multiple color bands to determine basis textures for each texture class.

$$F_{corr} = \frac{\sum_i \sum_j (i-\mu_x)(j-\mu_y)P_d(i,j)}{\sigma_x \sigma_y} \tag{3.7}$$

The visual dictionary is built for the features of the entire database. All the features are taken and clustered with K-means algorithm to obtain K-cluster centroids. These centroids form the visual words. Features of each block is compared with the cluster centroids using Euclidean distance and assigned to that corresponding cluster. Each image then is expressed as a histogram of visual words. This representation is calculated for all images in the data set.

3.3 Retrieval of Images

When extracting the visual patterns from microscopic images of large collections, image processing techniques are facing two problems. First is unstructured image representations and the other is the selection of precise machine learning tools which could extract most meaningful visual patterns.

The bag of features (BOF) approach is an evolution of texton based representations and is also influenced by the bag-of-words representation for text classification and retrieval [19]. Image segmentation tries to find out the whole objects inside the background image, here in bag of features looks only a small image regions without explicitly model the objects. The smaller regions are extracted from all input images. These extracted regions are represented by the name called feature vectors. Now a visual dictionary with the set of code words is built by these collected feature vectors. Then each image in that collection is represented by a codeword histogram. Many used this approach in content based image processing. Tomassi et al. [20] adapted the BOF representation to effectively classify radiological images in an automatic image annotation task.

The input query image is preprocessed and its visual word representation is calculated using the same visual dictionary. Using the histogram intersection as the distance measure the image from the dataset is returned which have the most histogram intersection with the query.

$$H1(h_1, h_2) = \sum_{\mathrm{i}=1}^{\mathrm{N}} \min(h_1 i, h_2 i) \tag{3.8}$$

where N is the number of bins in the histogram.

IV. Expriments And Results

The cytological images of the cells in the cervix region are collected to form a dataset containing around 30 normal and malicious cases. An example image is shown in Figure 1.

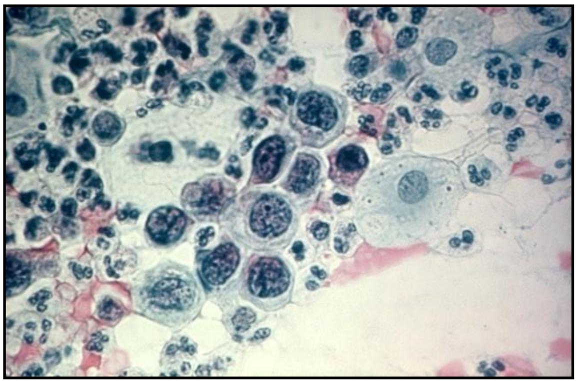

Figure 1. A Pap Smear Cervical Cytology Image

The features of a particular image are shown in Figure 2.

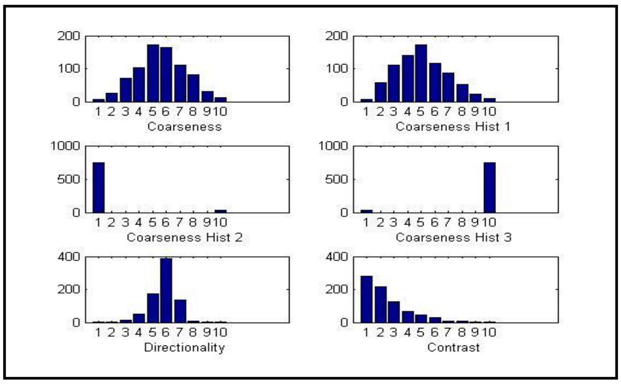

Figure 2. Texture Features of an example Image

We build the histogram with 30 images. We had conducted the experiments with varying block sizes of 8×8, 16×16 and 32×32 and different values of K, the number of clusters as 5, 10, 20, 40 and 80. Corresponding histograms each for malicious and normal images are shown in Figure 3, 4, 5 and 6 for different parameters.

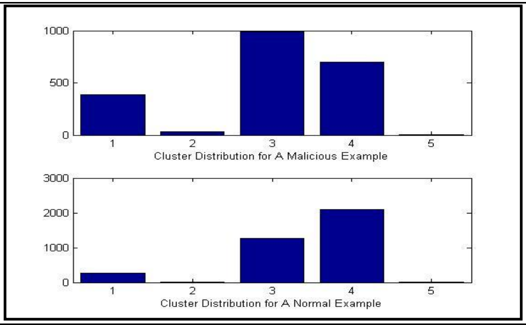

Figure 3.Histogram for Number of Clusters = 5

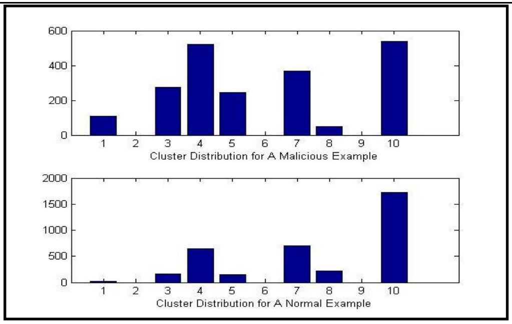

Figure 4.Histogram for Number of Clusters = 10

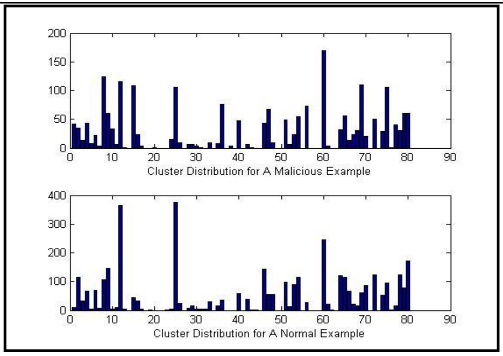

Figure 3.Histogram for Number of Clusters = 80

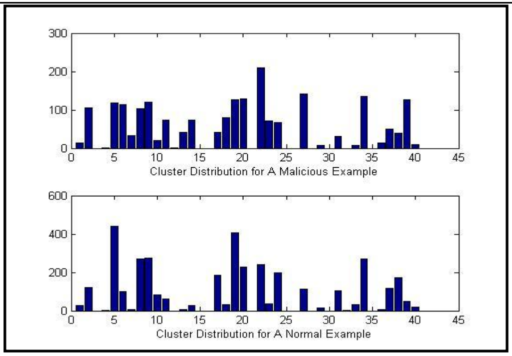

Figure 4.Histogram for Number of Clusters = 40

We have calculated the accuracy by the percentage of images returned which share the same characteristics of being normal or malicious with the query image and we present the result below for varying block size and parameter K.

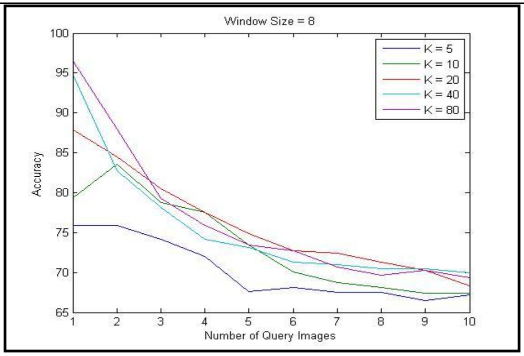

Figure 5. Accuracy for Block Size = 8

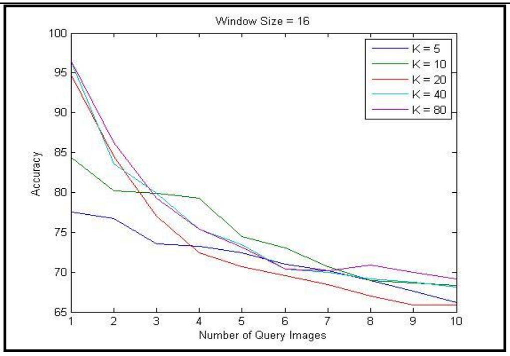

Figure 6. Accuracy for Block Size = 16

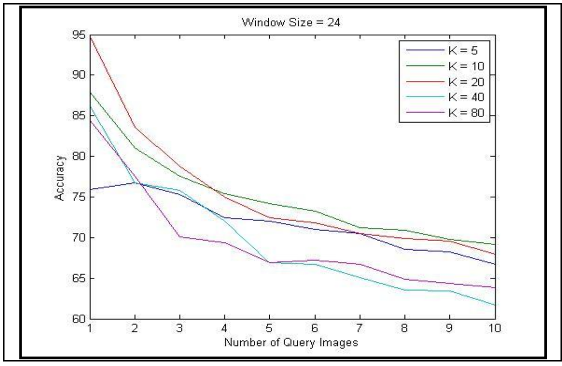

Figure 7. Accuracy for Block Size = 24

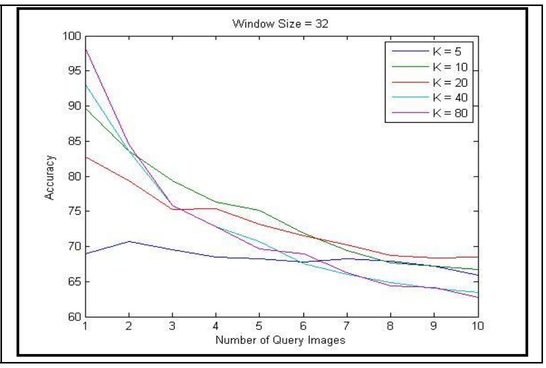

Figure 8. Accuracy for Block Size = 32

Using a visual dictionary enables us to avoid the problems associated with segmentation. From the results we note that the method can identify certain signatures of malignancy from the images automatically. The texture features from the signatures of these situations. For example a tell-tale feature of the images of malignant cells is that the nucleus is more enlarged and occupies larger portions of the cell's cytoplasm as seen in Figure 11. In the segmented nuclei region the malignant cell's image contains regions of dark and dense texture while the normal contains light and sparse structures. These correspond to different values of the texture features. The normal cells are marked by lower coarseness, correlation, homogeneity and energy whereas the malignant cells display lower contrast. The directionality is found not to vary a lot among the classes of images

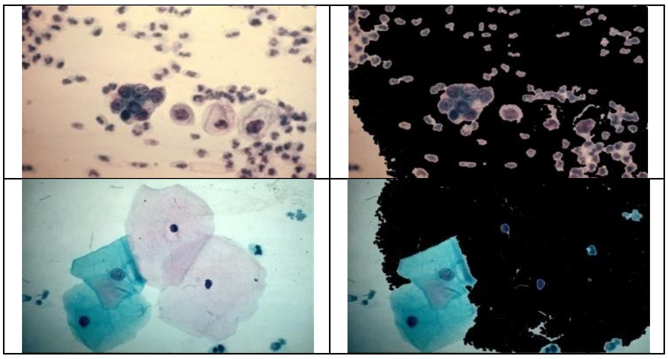

Figure 11. A Malignant Cells (top) and Normal Cells (bottom)

The cluster numbers for different blocks are shown in Figure 12. This clearly shows the different signatures of the classes of images. The cluster indices appearing on the inter-cellular space is to be ignored and does not have much discriminatory information. The cluster indices of blocks close to the nuclei region have more discriminatory value.

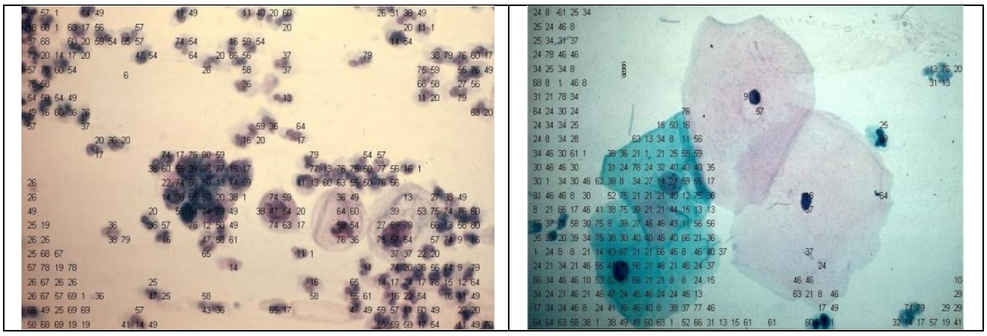

Figure 12. The Cluster Indices of Malignant (left) and Normal (right) Cervical Images

V. Conclusion

In this paper we have presented a method for content based image retrieval and diagnosis of cervical cancer cytological images using bag of features approach. We used texture features and created a visual dictionary through which we formed an image representation method using visual words. By experiments we demonstrated the effectiveness of the approach with good classification accuracies. Future work remains in improving accuracy as well as the ability to classify the cases based on the severity of occurrence of malignancy.

REFERENCES

[1] Hamakim AA, Lin PS, Wilczynski S, Nguyen K, Lynes B, Wakabayashi MT, Indications and Efficiency of the Human Papillomavirus Vaccine

[2] http://www.cervicalcanceraction.org/whynow/about.php, Accessed on 12th March 2013.

[3] H. R. Myler, and A. R. Weeks, "Computer Imaging Recipes in C.," 1993: Prentice Hall.

[4] Stephen O Hara, Bruce A. Draper, Introduction to the bag of features paradigm for Image Classification and Retrieval, Jan 2011

[5] H. P. Cai, F. Yan, and K. Mikolajczyk, "Learning weights for codebook in image classification and retrieval," in *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 2320– 2327.

[6] Lewis DD. Naive (bayes) at forty: The independence assumption in information retrieval. In: Nédellec C, Rouveirol C, editors. Proceedings of ECML-98, 10th European conference on machine learning, No. 1398. 1998. p. 4–15.

[7] Diamond J, Anderson NH, Bartels PH, Montironi R, Hamilton PW. The use of morphological characteristics and texture analysis in the identification tissue composition in prostatic neoplasia. Human Pathology 2004;35(9):1121–31.

[8] Kong J, Sertel O, Shimada H, Boyer KL, Saltz JH, Gurcan MN.Computer-aided evaluation of neuroblastoma on whole-slide histology images: classifying grade of neuroblastic differentiation. Pattern Recognition 2009;42(6):1080–92

[9] Orlov N, Shamir L, Macura T, Johnston J, Eckley DM, Goldberg IG. Wnd-charm:multi-purpose image classification using compound image transforms.Pattern Recognition Letters 2008;29(11):1684–93.

[10] Tang H, Hanka R, Ip H. Histological image retrieval based on semantic content analysis. IEEE Transactions on Information Technology in Biomedicine 2003;7(1):26–36.

[11] Naik J, Doyle S, Basavanally A, Ganesan S, Feldman MD, Tomaszewski JE, et al. A boosted distance metric: application to content based image retrieval and classification of digitized histopathology. SPIE Medical Imaging: Computer-Aided Diagnosis 2009;7260, 72603F1-12.

[12] Peng H. Bioimage informatics: a new area of engineering biology. Bioinformatics 2008;24(17):1827–36.

[13] Gose, Johnsonbough & Jost, 1996, Pattern Recognition and Image Analysis

[14] Tamura, H., Mori, S., Yamawaki, T.Textural features corresponding to visual perception. IEEE Trans on Systems, Man and Cybernetics 8 (1978) 460–472

[15] Peter Howarth and Stefan R¨uger, Evaluation of Texture Features for Content-Based Image Retrieval

[16] R. C. Gonzalez and P.Wintz, Digital Image Processing. Reading, MA: Addison-Wesley, 1987.

[17] R.M. Haralick, K.Shanmugam, and I. Dinstein, "Textural Features for Image Classification", IEEE Trans. on Systems, Man and Cybernetics, vol. SMC-3, no. 6, November 1973, pp. 610-621.

[18] Suen, P. and G. Healey: 2000, `The analysis and reconstruction of real-world textures in three dimensions'. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(5), 491-503.

[19] Lewis DD. Naive (bayes) at forty: The independence assumption in information retrieval. In: Nédellec C, Rouveirol C, editors. Proceedings of ECML-98, 10th European conference on machine learning, No. 1398. 1998. p. 4–15.

[20] Tommasi T, Orabona F, Caputo B. Clef2007 image annotation task: an svmbased cue integration approach. In: Nardi A, Peters C, editors. Working notes for the cross-language retrieval in Image Collections 2007 Workshop. 2007