# Knowledge Extraction from HEp2 Dataset

Ligendra Kumar Verma
Raipur Institute of Technology Raipur (C.G.)
Email: ligendra@rediffmail.com

Priyanka Tripathi and Kesari Verma
National Institute of Technology Raipur
Email: priyanka_tripathi@hotmail.com, keshriverma@gmail.com

***Abstract*—Pattern recognition is the process of taking the raw data, extract the features and categorize the patterns based on the features. Pattern recognition has been widely used in the area of diseases diagnosis. Autoimmune diseases are proven to be connected with the occurrence of autoantibodies in patient serum. Antinuclear autoantibodies (ANAs) identification can be accomplished in a laboratory using indirect immunofluorescence (IIF) imaging. In this paper we present the results of image analysis image analysis, feature extraction and classification of HEp-2 cell images. The experimental studies are performed on images of HEp-2 cells [1].**

***Index Terms*—image mining, decision tree induction, pattern recognition, HEp-2 classification**

## I. Introduction

Pattern recognition is one of the important problem in the field of medical imagine. Indirect Immuno Fluorescence (IIF) is considered a powerful, sensitive, and comprehensive test for antinuclear autoantibodies (ANAs) analysis. The IIF method has some disadvantages [1]. The major ones are: the low level of standardization, the inter observer variability, which limits the reproducibility of IIF readings; the lack of resources and adequately trained personnel; the photo bleaching effect, which bleaches significantly the tissues in a few seconds. Such drawbacks affect the diagnosis repeatability, therefore limiting the procedure reliability. Indeed, humans are limited in their ability to detect and diagnose a disease during image interpretation due to their non-systematic search patterns and to the presence of noise. In addition, the vast amount of image data that is generated makes the detection of potential disease a burdensome task and may cause oversight errors. Another problem is that similar characteristics of some abnormal and normal structures may cause interpretational errors. These problems result in an intra-laboratory variability estimated in the literature equal to 7-10%.ersight errors. The performance of the human expert with an error rate of 23.6% is very poor that motivates us to create a automated system to classify the cell.

The cells we consider in this paper are HEp-2 cell. The HEp-2 cells, which are used for the identification of antinuclear auto antibodies (ANA) [2], [3]. HEp-2 cells allow for recognition of over 30 different nuclear and cytoplasmic patterns, which are given by upwards of 100 different autoantibodies. We use the data set from the HEp-2 Cells Classification contest [1] organized by the International Conference on Pattern Recognition 2012. The training data set consists of 721 segmented and classified images. The six classes of the data set are:

1) Homogeneous
2) Coarse-speckled
3) Fine-speckled
4) Nucleolar
5) Centromere
6) Cytoplasmatic

The dataset consist of Images of Hp-2 cells. Each image contains multiple cultured HEp-2 cells. HEp-2 cells are cells of a human larynz epithelioma cancer cell line. These are highly specific to most human auto-antibodies [4].

In section 2 we present the preprocessing and noise removal method for image of cell. Section 3 focus on experimental work and important features. The paper is concluded in section 4.

### *A. Preprocessing and Noise Removal*

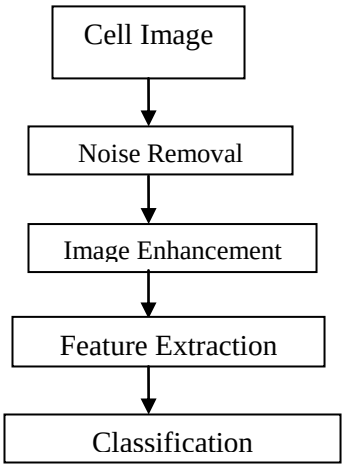


Figure 1. The block diagram for classification process

The image scanned from camera are converted from colored images to black & white image using following formula, where R denotes the Red component, G

Manuscript received February 11, 2013; revised April 13, 2013.


doi: 10.12720/joig.1.3.117-119

indicates the green components and B for B blue component of colored images.

$$gray = .2989 * R + .5870 * G + .1140 * B$$

If any noise in the image remove the noise from the image using noise removal method. The block diagram for image classification is shown in Fig. 1.

### B. Image Enhancement

In order to improve the contrast of image histogram [6] equalization is used.

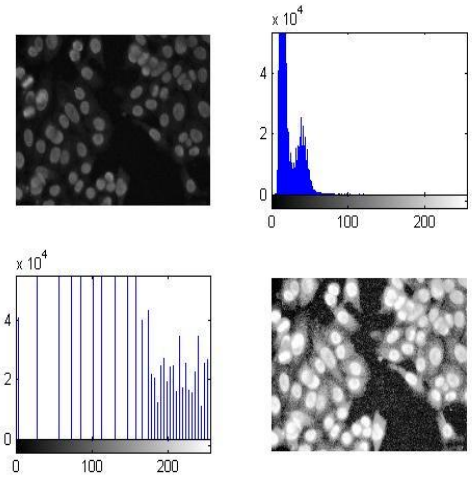


Figure 2. (a) Image before preprocessing; (b) Histogram of original images; (c) Histogram after image equalization; (d) Image after preprocessing

### C. Noise Removal

The image may contain noise during the enhancement. If the feature extracted from the noisy image the resultant may be the garbage result, in order to increase the accuracy of classifier the noise removal methods are applied on images.

*1) Median Filter - A median filter is an example of a non-linear it preserves the details of image.*

Algorithm -

*a) It consider each pixel in the image*

*b) Sort the neighbouring pixels into order based upon their intensities*

*c) Replace the original value of the pixel with the median value from the list*

A median filter is a rank-selection (RS) filter, a particularly harsh member of the family of rank-conditioned rank-selection (RCRS) filters.

*2) Wiener2 lowpass-filters*

wiener2 lowpass-filters a grayscale image that has been degraded by constant power additive noise.wiener2 uses a pixelwise adaptive Wiener method based on statistics estimated from a local neighborhood of each pixel.

*3) Ord Filter It is a nonlinear filter that comprising the following steps:*

*a) Define the neighbourhood of the target pixel(NxN)*

*b) Rank them in ascending order lowest to highest*

*c) Choose the order of the filter (from 1 to N)*

*d) Set the filtered value to be equal to the value of the chosen rank pixel.*

The experimental results after applying different filtering are shown in Fig 3.

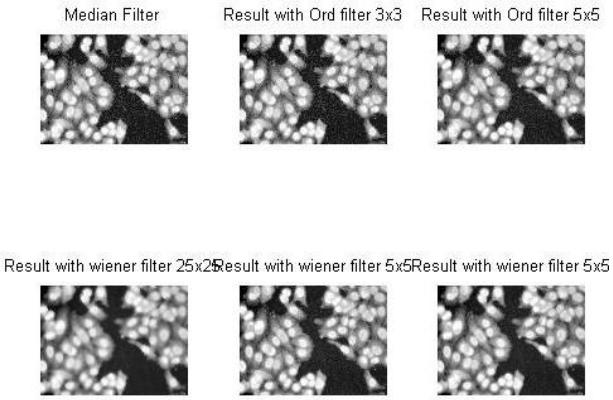


Figure 3. Result after applying different filters

## II. Important Features

The Main Geometrical and Statistical features of cell images are as follows.

*1) Diameter: The diameter is defined as the longest distance between any two points on the margin of the leaf. It is denoted as D.*

*2) Length: The length of the cell can be calculated automatically by the position of highest length as shown in Fig. 1*

*3) Physiological Width: Drawing a line passing through the two terminals of the main cell, one can plot infinite lines orthogonal to that line. The number of intersection pairs between those lines and the cell image.*

The relationship between physiological length and physiological width is illustrated in Fig. 4.

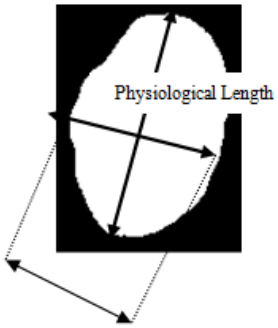


Figure 4. The relationship between physiological length and width.

*4) Variance [5] – Variance is used to determines the variation of pixel values from mean pixel intensity. This features gives the possibility of finding the amount of structure within a cell.*

*5) Mean – Mean of the intensity of the image is one of the important feature for clustering the images of cell.*

## III. Experimental Work

The experimental works is performed in Pentium IV machine with 800 MMz processor and Weka [7] software is used to perform the studies. The features we used are height, width, mean and variance. The experiment was SVM Classifier, Naïve Bayes classifier, Bayes Net Classifier, BF tree and Classification via regression. The accuracy of the algorithm is shown in Table I.

TABLE I. Experimental results using various classifiers

| Classifier | Correctly Classified | Incorrect Classified | Mean Square Error |
|---|---|---|---|
| SVM Classifier | 37.31% | 62.69% | .3465 |
| Naïve Bayes | 37.73% | 62.27% | 0.3478 |
| Bayes Net Classifier | 38.28% | 61.72% | 0.3462 |
| BF Tree | 38.28% | 61.72% | 0.3575 |
| Classification via Regression | 37.45% | 62.55% | 0.3482 |
| Human Expert | 23.6% | | |

## IV. Conclusion

The goal of this paper was to research the possibility to build a pattern recognition system suited for the classification of HEp-2 cell types. We used length, width, mean and standard deviation of the cell as major feature for classification of the HEP-2 cells. We applied it to SVM Classifier, Naïve Bayes, Bays Net classifier, BF tree and Classification via regression in weka [7] software. The average accuracy 37.5% for HEp-2 cell, which is very less. The texture feature , GLCM feature will be used for classification and recognition in future.

## Acknowledgment

The authors wish to thank University of Salerno Laboratory of Machine Intelligence for Video, Image and Audio Processing MIVIA Lab for providing dataset.

## References

[1] HEp-2 Cells Classification. [Online]. Available: http://mivia.unisa.it/hep2contest
[2] K. Conrad, R.-L. Humbel, M. Meurer, Y. Shoenfeld, and E. M. Tan, "Autoantigens and autoantibodies: Diagnostic tools and clues to understanding autoimmunity," K. Conrad, R.-L. Humbel, M. Meurer, Y. Shoenfeld, and E. M. Tan, eds, Pabst Science Publisher, Lengerich, Berlin, Riga, Rom, Wien, Zagreb, 2000.
[3] T. Y. Hsieh, Y. C. Huang, C. W. Chung, and Y-L. Huang, "HEp-2 Cell Classification in Indirect Immunofluorescence Images," in *Proc. 7th International Conference on Information, Communications and Signal Processing*, 2009, pp. 1-4.
[4] A. Bradwell and R. Sokes, "Atlas of HEp2 patterns," University of Birminghamm, 1995.
[5] D. V. Steirtenthn, S. Geerts, and F. V. Schueren, "Implementation of a semi exhaustive classification framework for the classification of HEp-2 cell types," Papers of the E-lab Master's Theses
[6] N. Otsu, "A threshold selection method from gray level histograms," *IEEE Trans*, vol. SMC-9, pp. 38-52, Jan 1979.
[7] Data Mining Software in Java. [Online]. Available: www.cs.waikato.ac.nz/ml/weka/

**Kesari Verma** Post graduated in 1998 and got doctorate in 2007. She is current Assistant professor in National Institute of Technology, Raipur. She has 12 year R & D and Teaching experience. She has more than 30 publications in journals and conferences.

**Ligendra Verma** completed his postgraduate in 1998 from Govt. Engineering College Raipur. He has 10 year of teaching and 10 year of industries experience. He is pursing his Ph.D. in Image Mining.