HUMAN EPITHELIAL TYPE 2 CELL CLASSIFICATION BASED ON CONCATENATED FEATURES AND MACHINE LEARNING ALGORITHMS

MAHSA NASSERI

UNIVERSITI TEKNOLOGI MALAYSIA

HUMAN EPITHELIAL TYPE 2 CELL CLASSIFICATION BASED ON CONCATENATED FEATURES AND MACHINE LEARNING ALGORITHMS

MAHSA NASSERI

A project report submitted in partial fulfilment of the
requirements for the award of the degree of
Master of Engineering (Electrical - Computer & Microelectronics System)

Faculty of Engineering
Universiti Teknologi Malaysia

JUNE 2015

Dedicated to my adorable parents Farah and Elyas, beloved husband Shahab and gorgeous daughter Elsa.

ACKNOWLEDGEMENT

I would like to express my deepest gratitude to my supervisor, Prof. Dr. Syed Abdul Rahman bin Syed Abu Bakar, for his excellent guidance, caring, patience, and providing me with an excellent atmosphere for doing research.

I would also like to thank my parents. They were always supporting me and encouraging me with their best wishes.

Finally, I would like to thank my husband, Shahab Ensafi. He was always there cheering me up and stood by me through the good times and bad.

I would also like to thank the developers of the utmthesis LaTeX project for making the thesis writing process a lot easier for me. Thanks to them, I could focus on the content of the thesis, and not waste time with formatting issues. Those guys are awesome.

Mahsa Nasseri

ABSTRACT

Medical researches show that Autoimmune Diseases (AD) are among the top ten leading causes of death among women in all age groups. The detection of AD in human body is performed by testing a type of antinuclear antibody, Human Epithelial Type 2 (HEp-2). The indirect immunofluorescence (IIF) imaging technique is used to capture the HEp-2 images to be used by physicians for detection. However, the detection of AD by IIF technique and analysis of them depends heavily on the experience of the physicians and may consume a long time. An accurate and automatic Computer Aided Diagnosis system will help greatly for the classification of the patterns in HEp-2 test. In this project an automatic HEp-2 cell image classification technique is proposed that exploits different feature types such as Rotation Invariant Co-Occurrence Local Binary Patterns (RICLBP), Scale Invariant Feature Transform and Speeded-Up Robust Feature. Additionally, a three layer of feature concatenation technique is proposed to extract spatial information of the images rather than local features. Finally, for the classification model, variety of classifiers are exploited to evaluate the performance of the system. These classifiers are Support Vector Machines (SVM), Random Forest (RF) and AdaBoost. As the results show, the performance of the system by using the RICLBP features with SVM and RF classifier is improved for ICPR2012 and ICIP2013 datasets, respectively. The performance on ICPR2012 is almost reached to the human expert accuracy and in image level classification it goes beyond the expert physicians' accuracy and reaches to 86%. In ICIP2013 dataset, specially for intermediate intensity level, almost 8% improvement is achieved by comparing with other methods.

ABSTRAK

Kajian perubatan menunjukkan bahawa Penyakit Autoimun atau (AD) merupakan salah satu di antara sepuluh penyebab utama kematian di kalangan wanita pada semua peringkat umur. AD dapat dikenalpasti pada tubuh manusia dengan menjalankan ujian antinuklear antibodi, Human Epithelial Type 2 (HEp-2). Teknik pengimejan immunofluorescence tidak langsung (IIF) digunakan oleh pakar perubatan untuk mengenalpasti imej-imej HEp-2. Walaubagaimanapun, teknik IFF dan analisisnya banyak bergantung kepada pengalaman doktor dan memakan masa yang lama. Satu sistem diagnosis bantuan komputer yang tepat dan automatik akan banyak membantu pengkelasan jenis sel gambar HEp-2. Dalam projek ini, satu teknik pengelasan automatik sel imej HEp-2 adalah dicadangkan yang mengeksploitasi ciri berbeza seperti ciri putaran berubah sesama kejadian corak binari tempatan (RICLBP), Skala Ciri berubah dan ciri teguh yang dipercepatkan. Selain itu, untuk klasifikasi model, teknik dinamika di cadangkan untuk mengesktrak informasi spatial bagi imej. Seterusnya bagi pengelasan, kepelbagaiannya dieksploitasi untuk menilai prestasi sistem. Pengelasan-pengelasan ini ialah mesin sokongan vektor (SVM), hutan rawak (RF) dan AdaBoost. hasilnya, keputusan menunjukkan prestasi sistem dengan menggunakan ciri-ciri pengelasan RICLBP dengan SVM dan RF meningkat untuk set data ICPR2012 dan ICIP2013. Prestasi pada ICPR2012 hampir sama dengan keputusan kepakaran manusia dan pada tahap klasifikasi imej, ianya jauh lebih dari itu dan ketepatan 86% diperolehi. Dalam set data ICIP2013 khasnya, untuk tahap intensiti pertengahan, peningkatan sebanyak 8% dicapai bila dibandingkan dengan kaedah lain.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ANA	-	Antinuclear Antibody
IIF	-	Indirect Immunofluorescence
HEp-2	-	Human Epithelial Type 2
LDA	-	Linear Discriminant Analysis
PCA	-	Principal Component Analysis
ICA	-	Independent Component Analysis
SIFT	-	Scaled Invariant Feature Transform
SURF	-	Speeded Up Robust Features
LBP	-	Local Binary Patterns
RICLBP	-	Rotation Invariant Co-occurrence Local Binary Patterns
SVM	-	Support Vector Machine
AdaBoost	-	Adaptive Boost
CART	-	Classification And Regression Trees
RF	-	Random Forest

CHAPTER 1

INTRODUCTION

1.1 Introduction

Abnormal immune response of the body against substances and tissues, which are normally present in the body, is called Autoimmune Diseases (AD). This may be restricted to certain organs, for instance in autoimmune thyroiditis or involve a particular tissue in different places. For example, Goodpasture's disease which may affect the basement membrane in both the lung and the kidney [1].

The immune system of body uses antibody, which is a large Y-shape protein produced by plasma cells, to identify and neutralize foreign objects such as bacteria and viruses (antigens). Additionally, autoantibody is produced by the immune system, which is an antibody that is directed against one or more of the individual's own proteins. Moreover, Antinuclear Antibodies (ANAs) are type of autoantibodies that bind to contents of the cell nucleus. The immune system of normal individuals produces antibodies to diminish antigens but not to human proteins (autoantigens). However, the disordered immune system produces antibodies to remove autoantigens.

For detecting the autoantibodies presence in an individual's blood serum, an ANA test is performed. Indirect Immunofluorescence (IIF) is the common test, which is used for detecting and quantifying ANAs. This test is very important as the American College of Rheumatology recommends that immunofluorescence test should remain the gold standard for ANA testing [2].

To visualize the antibodies, a fluorescent tagged autoantigens binds to them. This fluorescent is usually fluorescein isothiocyanate (FITC) or rhodopsin B [3]. To see the molecule under the microscope, a specific wavelength of light shines on it and the fluorescent reacts to it and makes the molecule visible. Depending on the antibody

present in the human serum and the localization of the antigen in the cell, the patterns of fluorescence will be seen on the HEp-2 cells [4].

1.2 Problem Statement

In the IIF test, the slides of the affected tissue are captured and the physicians need to diagnose the AD by categorizing the underlying patterns of the images. The most frequent and clinically useful patterns are as follows;

- Centromere: characterized by several discrete speckles ($\approx 40 - 60$) distributed throughout the interphase nuclei and characteristically found in the condensed nuclear chromatin during mitosis as a bar of closely associated speckles.
- Nucleolar: characterized by clustered large granules in the nucleoli of interphase cells which tend towards homogeneity, with less than six granules per cell.
- Homogeneous: characterized by a diffuse staining of the interphase nuclei and staining of the chromatin of mitotic cells.
- Fine Speckled: characterized by a fine granular nuclear staining of the interphase cell nuclei.
- Coarse Speckled: characterized by a coarse granular nu- clear staining of the interphase cell nuclei [5].

Although IIF is an excellent screening test, it contains many technical difficulties such as;

- Large number of IIF slides to be processed. By increasing the number of AD tests, the number of IIF slides are increased, which results in long diagnosis time.
- Diagnosing manually is very time consuming. As physicians need to manually identify the IIF patterns and there are a lot of cells in each specimen image, the procedure is very time consuming.
- Experienced physicians are needed for diagnosis. Specially in intermediate level of images, where the intensity of pixels are very low and the underlying pattern is not clearly visible, the experience of the physicians plays an important role for diagnosing ADs.

- Technologists should be trained to extract cells in the tissue. Segmentation of cells in the images is another issue, which makes the diagnosing procedure more challenging. The cells are sometimes overlapped and connected to each other or the images themselves are not captured well and may have some artefacts. Therefore, high level of technology for Ad diagnosis is needed.
- Need to use dark room for assessment. Physicians need to work in dark room in order to categorize the patterns ore efficiently. Due to the large number of IIF slides, physicians need to spend a lot of time in darkroom which is a hard labour.
- Repeatability of the test is costly too. Due to the time consuming and costly procedure of screening test, repeating the test doubles the difficulties.

1.3 Objectives

The objective of this thesis is to design an algorithm that will reduce the time for AD diagnosis in one hand and improve the efficiency and accuracy of screening test simultaneously. The proposed algorithm increases the diagnosis speed, which was done manually by the expert physicians and can perform the proof reading of the AD test. Additionally, it can be used for training the physicians to double check their decisions about specific tests. Moreover, the need of working in a dark room for diagnosis is solved.

1.4 Scopes

The proposed method is evaluated on two publicly available datasets, ICPR2012 and ICIP2013 and obtained classification accuracies on two tasks of cell and image level classification of HEp-2 images. These datasets are produced by the MIVIA (Labratorio di Macchine Intelligenti per il riconoscimento di Video, Immagini e Audio[1]) lab, which is a research laboratory of the University of Salerno in Italy. This laboratory is active in the fields of Pattern Recognition and Computer Vision.

To provide the ground truth, specialists manually segmented and annotated each cell. In particular, a biomedical engineer manually segmented the cells using a tablet PC. Subsequently, each image was verified by a medical doctor specialized

[1] http://mivia.unisa.it/

in Immunology and with 11 years experience, who annotated information at both image and cell levels. At the image level, the specialist annotated if there are enough mitotic cells to ensure, as previously stated, the correct preparation of the sample, the fluorescence intensity, and the staining pattern.

1.5 Chapter Organization

The rest of this thesis is organized as follows. In Chapter 2 the literature about the problem is reviewed. An algorithm is proposed in chapter 3. The experiment and results are discussed in chapter 4, which evaluated on two publicly available datasets introduced in Section 4.1. Finally, the conclusion and future works are presented in chapter 5.

REFERENCES

1. Cotsapas, C. and Hafler, D. A. Immune-mediated disease genetics: the shared basis of pathogenesis. *Trends in immunology*, 2013. 34(1): 22–26.

2. Meroni, P. L. and Schur, P. H. ANA screening: an old test with new recommendations. *Annals of the rheumatic diseases*, 2010: annrheumdis127100.

3. Storch, W. B. *Immunofluorescence in clinical immunology: a primer and atlas*. Springer. 2000.

4. González-Buitrago, J. M. and González, C. Present and future of the autoimmunity laboratory. *Clinica chimica acta*, 2006. 365(1): 50–57.

5. Foggia, P., Percannella, G., Soda, P. and Vento, M. Benchmarking HEp-2 Cells Classification Methods. *Medical Imaging, IEEE Transactions on*, 2013. 32(10): 1878–1889. ISSN 0278-0062.

6. Ludovic, R., Daniel, R., Nicolas, L., Maria, K., Humayun, I., Jacques, K., Frédérique, C., Catherine, G., Gilles, L., Metin, N. *et al.* Mitosis detection in breast cancer histological images An ICPR 2012 contest. *Journal of Pathology Informatics*, 2013. 4(1): 8.

7. Irshad, H., Jalali, S., Roux, L., Racoceanu, D., Hwee, L. J., Le Naour, G. and Capron, F. Automated mitosis detection using texture, SIFT features and HMAX biologically inspired approach. *Journal of pathology informatics*, 2013. 4(Suppl).

8. Hiremath, P., Bannigidad, P. and Geeta, S. Automated identification and classification of white blood cells (leukocytes) in digital microscopic images. *IJCA special issue on "recent trends in image processing and pattern recognition" RTIPPR*, 2010: 59–63.

9. Lee, L. H., Mansoor, A., Wood, B., Nelson, H., Higa, D. and Naugler, C. Performance of CellaVision DM96 in leukocyte classification. *Journal of pathology informatics*, 2013. 4.

10. Chan, J. W., Lieu, D. K., Huser, T. and Li, R. A. Label-free separation of human embryonic stem cells and their cardiac derivatives using Raman

spectroscopy. *Analytical chemistry*, 2009. 81(4): 1324–1331.

11. Plissiti, M. E. and Nikou, C. Cervical cell classification based exclusively on nucleus features. In: *Image Analysis and Recognition*. Springer. 483–490. 2012.

12. Yang, Y., Wiliem, A., Alavi, A. and Hobson, P. Classification of human epithelial type 2 cell images using independent component analysis. *ICIP*. 2013. 733–737.

13. Roullier, V., Lézoray, O., Ta, V.-T. and Elmoataz, A. Multi-resolution graph-based analysis of histopathological whole slide images: Application to mitotic cell extraction and visualization. *Computerized Medical Imaging and Graphics*, 2011. 35(7): 603–615.

14. Kiyan, T. Breast cancer diagnosis using statistical neural networks. *IU-Journal of Electrical & Electronics Engineering*, 2011. 4(2).

15. Malon, C., Cosatto, E. *et al.* Classification of mitotic figures with convolutional neural networks and seeded blob features. *Journal of Pathology Informatics*, 2013. 4(1): 9.

16. Nosaka, R. and Fukui, K. Hep-2 cell classification using rotation invariant co-occurrence among local binary patterns. *Pattern Recognition*, 2014. 47(7): 2428–2436.

17. Watanabe, T., Ito, S. and Yokoi, K. Co-occurrence histograms of oriented gradients for pedestrian detection. In: *Advances in Image and Video Technology*. Springer. 37–47. 2009.

18. Cristianini, N. and Shawe-Taylor, J. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press. 2000.

19. Di Cataldo, S., Bottino, A., Islam, I. U., Vieira, T. F. and Ficarra, E. Subclass Discriminant Analysis of morphological and textural features for HEp-2 staining pattern classification. *Pattern Recognition*, 2014. 47(7): 2389–2399.

20. Ghosh, S. and Chaudhary, V. Feature analysis for automatic classification of HEp-2 florescence patterns: Computer-Aided Diagnosis of Auto-immune diseases. *Pattern Recognition (ICPR), 2012 21st International Conference on*. IEEE. 2012. 174–177.

21. Ersoy, I., Bunyak, F., Peng, J. and Palaniappan, K. HEp-2 cell classification in IIF images using Shareboost. *Pattern Recognition (ICPR), 2012 21st*

International Conference on. IEEE. 2012. 3362–3365.

22. Ponomarev, G. V., Arlazarov, V. L., Gelfand, M. S. and Kazanov, M. D. ANA HEp-2 cells image classification using number, size, shape and localization of targeted cell regions. *Pattern Recognition*, 2014. 47(7): 2360–2366.

23. Theodorakopoulos, I., Kastaniotis, D., Economou, G. and Fotopoulos, S. Hep-2 cells classification via fusion of morphological and textural features. *Bioinformatics & Bioengineering (BIBE), 2012 IEEE 12th International Conference on*. IEEE. 2012. 689–694.

24. Ensafi, S., Lu, S., Kassim, A. A. and Tan, C. L. A Bag of Words Based Approach for Classification of HEp-2 Cell Images. *Pattern Recognition Techniques for Indirect Immunofluorescence Images (I3A), 2014 1st Workshop on*. IEEE. 2014. 29–32.

25. Kong, X., Li, K., Cao, J., Yang, Q. and Wenyin, L. HEp-2 cell pattern classification with discriminative dictionary learning. *Pattern Recognition*, 2014. 47(7): 2379 – 2388. ISSN 0031-3203. doi:http://dx.doi.org/10.1016/j.patcog.2013.09.025. URL `http://www.sciencedirect.com/science/article/pii/S003132031300397X`.

26. Lowe, D. G. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 2004. 60(2): 91–110.

27. Bay, H., Ess, A., Tuytelaars, T. and Van Gool, L. Speeded-up robust features (SURF). *Computer vision and image understanding*, 2008. 110(3): 346–359.

28. Vapnik, V., Golowich, S. E. and Smola, A. Support vector method for function approximation, regression estimation, and signal processing. *Advances in Neural Information Processing Systems 9*. Citeseer. 1996.

29. Freund, Y., Schapire, R. E. *et al.* Experiments with a new boosting algorithm. *ICML*. 1996, vol. 96. 148–156.

30. Kittler, J., Hatef, M., Duin, R. P. and Matas, J. On combining classifiers. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 1998. 20(3): 226–239.

31. Breiman, L. Random forests. *Machine learning*, 2001. 45(1): 5–32.

32. Bosch, A., Zisserman, A. and Munoz, X. Image classification using random forests and ferns. *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*. IEEE. 2007. 1–8.

33. Ensafi, S., Lu, S., Kassim, A. A. and Tan, C. L. Automatic cad system for hep-2 cell image classification. *Pattern Recognition (ICPR), 2014 22nd*

International Conference on. IEEE. 2014. 3321–3326.

34. Han, X.-H., Wang, J., Xu, G. and Chen, Y.-W. High-Order Statistics of Microtexton for HEp-2 Staining Pattern Classification. *Biomedical Engineering, IEEE Transactions on*, 2014. 61(8): 2223–2234. ISSN 0018-9294.