Artificial Intelligence in Medicine 65 (2015) 239–250

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

journal homepage: www.elsevier.com/locate/aiim

Benchmarking human epithelial type 2 interphase cells classification methods on a very large dataset

Peter Hobson[a], Brian C. Lovell[b], Gennaro Percannella[c,*], Mario Vento[c], Arnold Wiliem[b]

[a] Sullivan Nicolaides Pathology, 134 Whitmore street, Taringa, Queensland 4068, Australia
[b] School of Information Technology and Electrical Engineering, The University of Queensland, St Lucia, Queensland 4072, Australia
[c] Department of Information Engineering, Electrical Engineering and Applied Mathematics, University of Salerno, Via Giovanni Paolo II, 132, Fisciano, SA I-84084, Italy

ARTICLE INFO

Article history:
Received 12 September 2014
Received in revised form 31 July 2015
Accepted 3 August 2015

Keywords:
Large-scale benchmarking
Computer-aided diagnosis systems
Indirect immunofluorescence
Hep-2 cell classification

ABSTRACT

Objective: This paper presents benchmarking results of human epithelial type 2 (HEp-2) interphase cell image classification methods on a very large dataset. The indirect immunofluorescence method applied on HEp-2 cells has been the gold standard to identify connective tissue diseases such as systemic lupus erythematosus and Sjögren's syndrome. However, the method suffers from numerous issues such as being subjective, time consuming and labor intensive. This has been the main motivation for the development of various computer-aided diagnosis systems whose main task is to automatically classify a given cell image into one of the predefined classes.
Methods and material: The benchmarking was performed in the form of an international competition held in conjunction with the International Conference of Image Processing in 2013: fourteen teams, composed of practitioners and researchers in this area, took part in the initiative. The system developed by each team was trained and tested on a very large HEp-2 cell dataset comprising over 68,000 images of HEp-2 cell. The dataset contains cells with six different staining patterns and two levels of fluorescence intensity. For each method we provide a brief description highlighting the design choices and an in-depth analysis on the benchmarking results.
Results: The staining pattern recognition accuracy attained by the methods varies between 47.91% and slightly above 83.65%. However, the difference between the top performing method and the seventh ranked method is only 5%. In the paper, we also study the performance achieved by fusing the best methods, finding that a recognition rate of 85.60% is reached when the top seven methods are employed.
Conclusions: We found that highest performance is obtained when using a strong classifier (typically a kernelised support vector machine) in conjunction with features extracted from local statistics. Furthermore, the misclassification profiles of the different methods highlight that some staining patterns are intrinsically more difficult to recognize. We also noted that performance is strongly affected by the fluorescence intensity level. Thus, low accuracy is to be expected when analyzing low contrasted images.

1. Introduction

Recently there has been a growing interest in introducing automated pattern classification systems for microscopy images [1–5]. The results from these systems may offer a more objective classification which would improve result consistency and resolve any discrepancies in the subjective analyses.

The anti-nuclear antibodies (ANA) test is commonly used to diagnose connective tissue diseases (CTD) such as systemic lupus erythematosus (SLE) and Sjögren's Syndrome [6]. The gold standard for performing this test is the indirect immunofluorescence (IIF) protocol using human epithelial type 2 (HEp-2) cells [6,7] due to the expression of a wide range of antigens on HEp-2 cells. Nevertheless, the protocol is time and labor intensive [8,9]. In addition, there is high intra- and inter-laboratory variation of the test [8,10,11].

One way to address these issues is by applying computer-aided diagnosis systems. These provide a more objective analysis which could be incorporated into the overall test results. In recent years, we have seen significantly growing interest in developing such systems [2,10–20]. Nevertheless, the use of private datasets with

* Corresponding author. Tel.: +39 089 964253; fax: +39 089 964218.
E-mail addresses: peter_hobson@snp.com.au (P. Hobson), lovell@itee.uq.edu.au (B.C. Lovell), pergen@unisa.it (G. Percannella), mvento@unisa.it (M. Vento), a.wiliem@uq.edu.au (A. Wiliem).

http://dx.doi.org/10.1016/j.artmed.2015.08.001
0933-3657/

non-standard evaluation protocols makes it difficult to draw meaningful conclusions from the existing works. Therefore, it is critical to develop a standard evaluation platform in order to advance the domain [2]. One notable example is the first contest initiative held in conjunction with the International Conference on Pattern Recognition (ICPR) 2012, here denoted ICPR2012Contest [2], which is then followed by publications of a Pattern Recognition journal special issue on the same theme [21].

Despite the merit of being the first initiative in this research area and the attention received from the scientific community, there were some shortcomings in the benchmarking platform introduced through the ICPR2012Contest. Among such issues, the most relevant were:

- Small size of the dataset: the dataset provided in ICPR2012Contest has six classes: centromere, coarse speckled, cytoplasmic, fine speckled, homogeneous and nucleolar. It has a total of 1457 cell images extracted from 28 specimen images. It is assumed that each specimen image comes from a unique patient serum and a specimen image contains a distribution of HEp-2 cells. The specimen images are equally divided for training and testing. Although at first glance the number of cell images may appear significant, larger numbers of images are required to draw more meaningful conclusions [2]. In fact, the overall analysis is mainly affected by the number of specimen images, as the cell images from the same specimen are similar. More specifically, the classes in both training and test sets only have two or three specimen images, thus, the evaluation protocol is limited to the variation generated from two specimen images. This also renders a biased view during the cross validation training process which may have misled participants in designing their systems.
- Focusing only on common patterns: whilst in general there are four ANA patterns commonly found in day-to-day operation – homogeneous, speckled, centromere and nucleolar – correctness in identifying less common patterns is equally significant as they may have clinical significance. Unfortunately, the ICPR2012Contest dataset did not include these less common patterns.

In the present work, we address the above two issues by constructing a very large dataset consisting of 68,429 cell images extracted from 419 patient sera. In particular, there are now six classes: homogeneous, speckled, centromere, nucleolar, nuclear membrane and Golgi. Nuclear membrane and Golgi patterns are less common than the other four patterns. This not only offers a more realistic evaluation protocol, but also, more flexibility for doing cross validation. These factors allow the present work to offer a more realistic benchmarking of systems in this domain.

We note that, unlike ICPR2012Contest that considers the cytoplasmic pattern, we exclude the cytoplasmic pattern from our current benchmarking platform as it is not considered an ANA pattern [7]. In addition, our benchmarking platform also does not differentiate between the fine and coarse speckled classes for two reasons. Firstly, the speckled pattern subdivision is generally more complex than simply dividing it into fine and coarse speckled groups. In general, the subdivision is done by relating each individual sub-group with specific antibodies [7]. For instance, fine speckled could be further divided into several sub-groups with distinct characteristics such as fine speckled patterns caused by SSA(Ro)/SSB(La) and DFS-70 [22]. Secondly, given the above fact, a better analysis would be to consider the fine-grained classification scheme [23,24] on the sub-groups of the speckled patterns once a specimen is identified as speckled.

Our benchmarking platform is not aimed to evaluate the performance of CAD systems in the fine-grained speckled classification problem. Thus, using only one speckled class gives us an advantage to avoid confusion in analyzing the evaluation results (e.g. whether the classification mistakes are due to the inability of a method in addressing the fine-grained speckled classification problem or the general ANA HEp-2 cell classification problem).

Finally, it is worth to highlight that the benchmarking platform presented here refers to the classification of the HEp-2 cells in the interphase, as the ICPR2012Contest, and does not consider the issue of the recognition of the cells in the mitotic stage.

The paper is organized as follows: Section 2 provides a brief description on methods to perform the ANA test; in Section 3, we describe our dataset that has been used for the benchmarking; in Section 4, we first define formally the pattern recognition task that was proposed to the participants in the initiative and then provide a short summary of each method. The results and analysis of the benchmarking work are presented in Section 5. Finally, we draw conclusions and delineate future work in Section 6.

2. The ANA test

The ANA test is used for screening a wide range of CTDs [6,7]. Methods to detect ANA include indirect immunofluorescence using HEp-2 cells, enzyme immunosorbent assay (EIA)/enzyme-linked immunosorbent assay (ELISA), farr assay, multiplex immunoassay (MIA) and western blot [25].

Amongst these methods, the IIF using HEp-2 cell method is considered the gold standard as the method has high sensitivity due to the expression of wide range of antigens on HEp-2 cells [6]. Generally, other techniques are used as secondary/confirmatory tests. For instance, EIA/ELISA are specifically designed to target single autoantigens (e.g. dsDNA and SSA-A/Ro). The Farr assay is a radio-labeled assay for quantifying anti-dsDNA [25]. In western blot, antigens are separated according to their molecular weight and then transferred onto strips or a membrane [25]. The strips are then incubated with the patient serum. Positive reactions are compared to a positive control strip. For MIA, serum is incubated with a suspension of multi-colored polystyrene micro-spheres coated with a range of antigens. The binding, determining the test result, is then quantified using a specific instrument platform.

For the IIF method, the slides are examined under a fluorescent microscope by two scientists. The analysis starts by determining the specimen positivity from the observed fluorescent signal. The guidelines established by the Center of Disease Control and Prevention, Atlanta, Georgia (CDC) suggest the use of a scoring system ranging from 0 to 4+ wherein 0 represents negative (no fluorescent signal observed), and 4+ represents the strongest positive (very bright fluorescent signal observed) [26]. As this process is subjective, it is possible to reduce the scoring system into merely determining whether the fluorescence intensity level of the sample is positive, intermediate or negative [12]. Positive ANA patterns are then titred by serial dilution to obtained a more objective fluorescence intensity level [26]. Finally, the last step in the analysis is to determine the visual pattern appearing in the positive and intermediate specimens.

Generally, scientists consider at least three visual cues when examining positive and intermediate specimens: (1) at least one or two mitotic cells can be found in the specimen [26]; (2) the visual features of the mitotic cells and (3) the visual features of the interphase cells.

Unlike the interphase cells, the amount of cell chromatin in mitotic cells is doubled. The cells undergoing the mitosis stage may express different antigens or antigens in different concentrations to those in the interphase stage [27,28]. Thus, in some cases, scientists need to consider the mitotic cell visual cues before correctly