Benchmarking Human Epithelial Type 2 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, Australia*
[b]*School of Information Technology and Electrical Engineering, The University of Queensland, Australia*
[c]*Department of Information Engineering, Electrical Engineering and Applied Mathematics, University of Salerno, Italy*

Abstract

This paper presents benchmarking results of Human Epithelial type 2 (HEp-2) 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 (SLE) and Sjögren's Syndrome. However, the method suffers from numerous issues such as being subjective, time and labour intensive. This has been the main motivation for the development of various Computer Aided Diagnosis (CAD) systems whose main task is to automatically classify a given cell image into one of the predefined classes. The benchmarking was done 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 cell images. In this paper, we briefly describe all the submissions and provide an in-depth analysis on the benchmarking results and their design choices.

Keywords: Computer-aided diagnosis (CAD), HEp-2 cells classification, indirect immunofluorescence

Email addresses: peter_hobson@snp.com.au (Peter Hobson), lovell@itee.uq.edu.au (Brian C. Lovell), pergen@unisa.it (Gennaro Percannella), mvento@unisa.it (Mario Vento), a.wiliem@uq.edu.au (Arnold Wiliem)

1. Introduction

Recently there has been a growing interest in introducing automated pattern classification systems to microscopy images [1, 2, 3, 4, 5]. The results from these systems may offer a more objective classification which would resolve any discrepancies in the subjective analyses and improve result consistency.

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 labour intensive [8, 9]. In addition, there is high intra- and inter- laboratory variation of the test [10, 11, 8].

One way to address these issues is by applying Computer Aided Diagnosis (CAD) 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 [12, 13, 10, 14, 15, 16, 17, 2, 18, 11, 19]. Nevertheless, the use of private datasets with non-standard evaluation protocols made 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 was then followed by a special issue of the Pattern Recognition journal on the same theme [20].

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: (1) the small size of the dataset causing a limited realistic evaluation of the considered methods and (2) focusing only on common HEp-2 cell patterns.

The dataset provided in ICPR2012Contest has six classes: centromere, coarse speckled, cytoplasmic, fine speckled, homogeneous and nucleolar. It has a total of 1,457 cell images extracted from 28 specimen images. It is

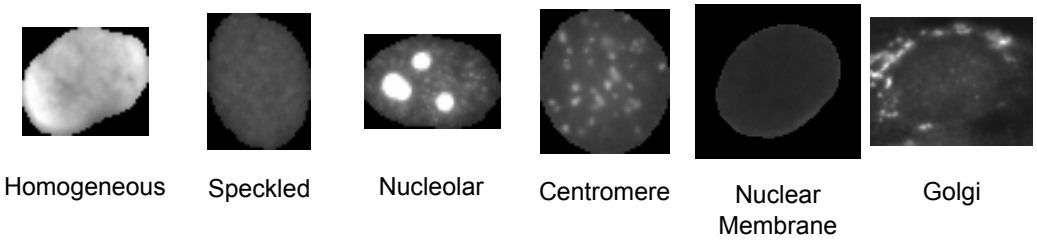


Figure 1: Sample cell images from ICIP2013 dataset.

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.

Small size of the dataset. 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 at least 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 [21]. Secondly, given the above fact, a better analysis would be to consider the fine-grained classification scheme [22, 23] 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 also gives us an advantage to avoid confusion in analysing 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).

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 to the initiative and then provide a short summary of each method; The results and analysis of the benchmarking work is 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 (IIF) using HEp-2 cells, Enzyme Immunosorbent Assay (EIA)/Enzyme-Linked Immunosorbent Assay (ELISA), Farr Assay, Multiplex Immunoassay (MIA) and Western Blot [24].

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 [24]. In western blot, antigens are separated according to their molecular weight and then transferred onto a membrane or strips [24]. 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-coloured 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) [25]. 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 [25]. 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 [25]; (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 [26, 27]. Thus, in some cases, scientists need to consider the mitotic cell visual cues before correctly identifying an ANA pattern. For instance, the mitotic spindle pattern, where the mitotic spindle is positively stained, can only be observed in mitotic cells.

While it is important to study the automated mitotic pattern classification which is shown in a number of recent works [28, 19], in this work, we primarily focus on the interphase cell classification problems. Addressing the interphase cell classification problems is one of the early steps to develop a CAD system for the ANA IIF HEp-2 test.

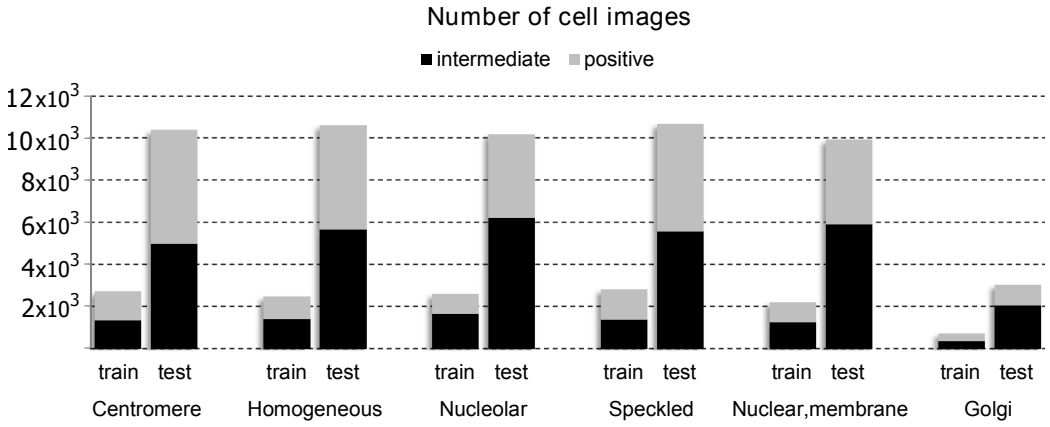


Figure 2: The number of cell images for each pattern contained in train and test sets.

3. Dataset description

The dataset was obtained between 2011 and 2013 at Sullivan Nicolaides Pathology laboratory, Australia[1]. The dataset contains the following six classes [7] (see Figure 1 for some examples):

- *homogeneous*: a uniform diffuse fluorescence covering the entire nucleoplasm sometimes accentuated in the nuclear periphery;
- *speckled*: this pattern is generally divided into two groups[2]:
  - *coarse speckled*: densely distributed, variously sized speckles, generally associated with larger speckles, throughout the nucleoplasm of interphase cells; nucleoli are negative;
  - *fine speckled*: fine speckled staining in a uniform distribution, sometimes very dense so that an almost homogeneous pattern is attained; nucleoli may be positive or negative;
- *nucleolar*: brightly clustered large granules corresponding to decoration of the fibrillar centers of the nucleoli as well as the coiled bodies;

[1] We name this dataset as ICIP2013 dataset.

[2] In this dataset, we consider these two sub-categories as one category, while in the ICPR2012Contest they were kept distinct.

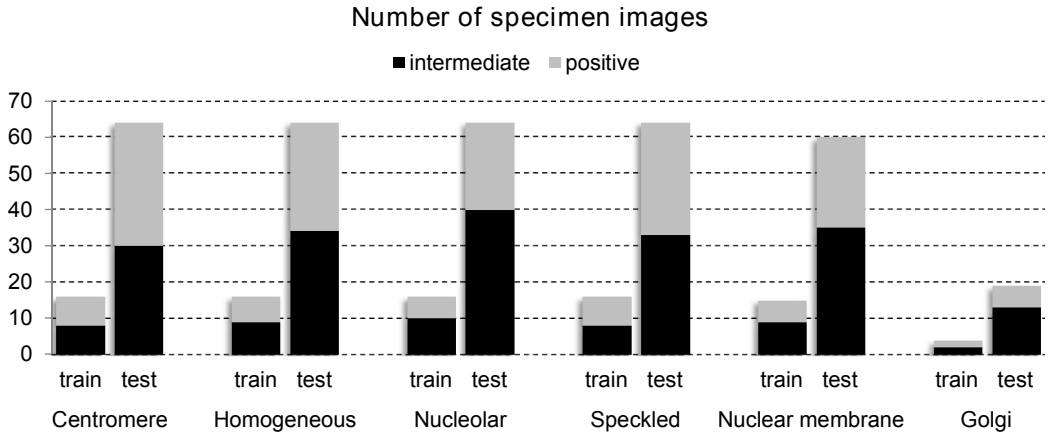


Figure 3: The number of specimen images for each pattern contained in train and test sets.

- *centromere*: rather uniform discrete speckles located throughout the entire nucleus;

- *Golgi*: staining of a polar organelle adjacent to and partially surrounding the nucleus, composed of irregular large granules. Nuclei and nucleoli are negative. Diffuse staining of the cytoplasm of dividing cells sometimes with accentuation around chromosomal material;

- *nuclear membrane*: a smooth homogeneous ring-like fluorescence of the nuclear membrane in interphase cells.

The dataset utilises 419 unique positive sera extracted from 419 different patients which were prepared on 18-well slides of HEP-2000 IIF assay from Immuno Concepts N.A. Ltd. using a screening dilution 1:80. As per the manufacturer's description, the assay contains at least one or two mitotic cells. The specimens, one for each patient serum, were then automatically photographed using a monochrome high dynamic range cooled microscopy camera which was fitted on a microscope with a plan-Apochromat 20x/0.8 objective lens and an LED illumination source. Approximately 100-200 cell images were extracted from each patient serum. In total there were 68,429 cell images extracted. We divided these into 13,596 images for training and 54,833 for testing. The division was deliberately made so that the test set only contained cells from patients who were not included in the training set.

Specifically, the training set contains the specimen images of 83 patients, while the remaining 336 were reserved for the test set.

The training set is publicly available and in particular was used by the ICIP 2013 competition for HEp-2 cells classification, while the test set is undisclosed[3].

The labelling process involved microscopic reading by two scientists. A third opinion was sought to adjudicate any discrepancies. We used each specimen label as the groundtruth of cells extracted from it. Furthermore, the labels were investigated further using secondary tests such as ENA, and anti-ds-DNA to confirm specificity of the ANA pattern.

Figures 2 and 3 present the number of exemplars for each pattern class in both training and test sets. The more common patterns such as centromere, homogeneous, nucleolar and speckled have a similar number of exemplars. However, the less common patterns such as the nuclear membrane and Golgi have fewer exemplars. In particular, Golgi has significantly fewer exemplars than the other patterns. This depicts a more realistic condition where the system needs to perform reasonably well on both common patterns and significantly less common patterns.

We note that the creation of this benchmarking platform is possible due to the recent advancements that sufficiently address several practical problems in the automated acquisition of HEp-2 images; allowing us to capture high quality images of a patient specimen in approximately 20 seconds [27]. In particular, the acquisition system uses two channels: (1) the Fluorescein isothiocyanate (FITC) channel that is normally used for ANA test and (2) the 4',6-diamidino-2-phenylindole (DAPI) channel that is used in the cell image segmentation. DAPI which is a fluorescent stain that binds strongly to cell DNA [29], specifically delineates the HEp-2 cell nuclei (i.e. the area of interest in the ANA test). Therefore, this may be used to perform high precision HEp-2 cell segmentation regardless of the patient pattern exhibited in the FITC channel [27, 26] (refer to Fig. 4 for some challenging examples where the HEp-2 cells are successfully segmented in high precision). This

[3]The test set is intentionally kept private with the aim of avoiding the problem that after consecutive loops of train-test sessions researchers may continuously improve performance over the test set through specialization to that specific set of data. However, the interested researchers and practitioners can evaluate performance of their method on the test set by submitting the executable of their algorithm. Submission is allowed once per method in order to avoid the specialization problem described above.

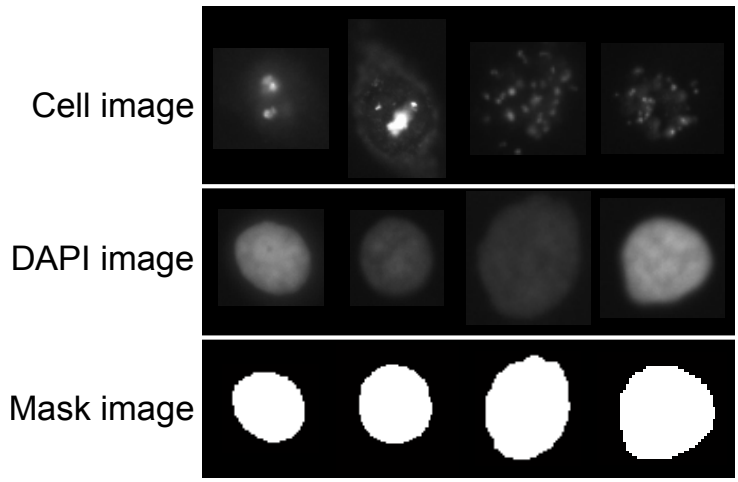


Figure 4: Examples of challenging HEp-2 cell images with their corresponding FITC, DAPI and mask images. The high-precision segmentation masks in the third row are obtained by applying a simple image thresholding approach as suggested in [27].

approach addresses issues such as misclassifications due to poor segmentation, stemming from imperfections in the manual segmentation process in the previous benchmarking set, ICPR2012Contest.

4. Classification methods

We now describe the methods which participated in the benchmarking activity held at ICIP 2013. For the sake of brevity, the description is intentionally short so as to focus on the most relevant aspects of each method. However, interested readers may find more details about each individual method in the ICIP 2013 competition report[4] where each participant provides an extended description of their methods. In the following, each method is reported using the first three letters of the surname of its first author.

Before going into the description of the methods it is worth recalling the classification goal proposed for each team that participated in the competition. The task was to develop a classifier φ which classifies a set of HEp-2 cell images. Each image is represented by the three-tuple $(\boldsymbol{I}, \boldsymbol{M}, \delta)$ [18]: (1) $\boldsymbol{I}$ represents the cell fluorescence image in FITC channel; (2) $\boldsymbol{M}$ is the cell

[4]The report is available at http://nerone.diiie.unisa.it/contest-icip-2013/ICIP2013_report.pdf

mask which is automatically extracted from the DAPI channel and (3) δ represents the cell positivity strength which has two values weak/borderline (*intermediate*) or strong (*positive*). Let $\boldsymbol{Y}$ be a test image, ℓ be its true class label and $\mathcal{G} = \{(\boldsymbol{I}, \boldsymbol{M}, \delta)_1, ...(\boldsymbol{I}, \boldsymbol{M}, \delta)_n\}$ be a given gallery set. The classifier task is to predict the test label, $\hat{\ell}$. In other words, $\varphi : \boldsymbol{Y} \times \mathcal{G} \mapsto \hat{\ell}$, where ideally $\hat{\ell} = \ell$.

CHA - The rationale of the method is to selectively exploit texture information from different regions of an image. To this end each cell is divided into six partially overlapped regions which extend from the cell boundary to the inner circle area. In total 18 features are calculated from each region: region brightness, contrast and 16 one-dimensional bispectral invariants [30]. These are successively concatenated to form a vector of 108 features which were used to train a set of classifiers (each HEp-2 cell class has one corresponding classifier). For each pattern class, Adaboost [31] is used to generate a 10-stage binary classifier, combined with a hand-crafted decision tree. In particular, the authors evaluated the performance of each binary classifier and constructed a decision tree that placed them in the order of performance, highest first. If all the binary classifiers reject a query image, it is then assigned to a default class.

HAN - The proposed method uses the distribution of local pixel neighborhoods (denoted micro texton) with Gaussian mixture model as its histogram encoding method. As for the image representation, they compute and concatenate the gradient with respect to the model parameters. The final representation can be considered as Fisher Vector. A random forest classifier is adopted as the classifier.

LAR - In the preprocessing stage each image is augmented with its logarithmic representation [32]. Then, each representation is mapped linearly to $[0, 1]$ such that their minimum attains a value of zero and their maximum a value of one. The features are extracted from both representations of each image. For each cell a feature vector is built consisting of the intensity information, morphological features extracted from the provided mask (including area, eccentricity, major and minor axis length, perimeter), and the "annulus" shape index histogram feature. The latter is the most significant descriptor and consists of weighted histograms of second order image features derived form the local Hessian eigenvalues [33] over a number K of band-shaped regions. Each region is defined by its distance to the center pixel of the image, while the weight for each pixel is assigned based on a Gaussian distribution centered on the radial band. Classification is performed through

a multi-class SVM with RBF kernel using a one-vs-one scheme.

LIU - The proposed method initially normalizes the brightness of the input image. Then, local patches of size 9×9 pixels are extracted on a dense sampling grid. In the training phase, these patches are projected through PCA and a codebook with N codewords is created, as described in [34]. This codebook is used to partition all the local patches into N groups. Then, discriminative projections are obtained for each group by a partial least square analysis in order to re-project the image patches to low dimensional vectors. According to the BoW pipeline, the final image representation is obtained by concatenation of the histograms from different groups. A linear SVM is used for the classification stage.

MAR - The proposed approach [35] builds upon the use of square subwindows randomly extracted from the original image with respect to the position, the rotation angle and the size. The subwindows are resized to 16×16 pixels and encoded in normalized RGB color space. A very large set of visual features is generated using randomized trees. In particular, an ensemble of 50 trees is built according to [36] and then is used to generate an image-level signature inspired by the bags of visual words [37] or textons [38]. Each terminal node of individual tree is a real-valued feature that corresponds to the number of subwindows that reach the terminal node divided by the number of subwindows extracted in the image. Each cell image is represented by these sparse and high-dimensional signatures. Finally, a linear SVM, adopting a one vs one multi-class strategy, is used for the final class prediction.

NAN - The method is based on the combination of three texture descriptors: the multiscale Pyramid Local Binary Pattern (PLBP) [39], which is based on the LBP operator applied to each of the $l = (0, ..., L)$ levels of the gaussian pyramid built from the original image; the Strandmark morphological features (STR), which are a reduced version of the features in [40] and the canonical Haralick features (HAR) defined in [41]. The classification is performed using a multiclass SVM with RBF kernel, according to the one-vs-all approach. The SVMs are trained for each of the three sets of features and the results are combined according to the sum rule.

PAI - After a preprocessing phase that includes denoising (median filtering) and normalization (histogram equalization), the proposed approach relies on three different sets of features in addition to the information regarding the image intensity level: (1) region covariance of image statistics, as the intensity value, the first and second order derivative in the vertical and

horizontal directions, the magnitude of the gradients; (2) CoALBP features, the extension of LBP [42] and (3) the Strandmark morphological features (STR) [40]. Finally, the classification is carried out using a multi-class boosting algorithm that can adaptively select the most discriminative feature in each boosting iteration and combine them into an effective classifier.

POM - In the preprocessing stage the cell image is binarized and resized to a canonical size. The employed features are based on the Complete LBP (CLBP) approach [43]. The CLBP approach is based on the assumption that the local appearance and textural structure can be defined by the histogram of the local sign, magnitude and central pixel defined on a dense grid. The CLBP histograms of the sign magnitude and central pixel combine structural and statistical information, and capture the distribution of the classified structures. Classification is performed using $k-$NN.

PON - The proposed method relies on the characterization of the morphological properties of the stained regions of the HEp-2 cells such as nucleoli, nucleous and chromosomes. The authors suggest two different preprocessing steps depending on the type of the descriptor: (1) the image is thresholded using Otsu binarization and (2) the image is normalized in the range $[0, 255]$. Twenty one features belonging to the following logical groups are used: number of stained regions (also called objects by the authors), object size, holes inside objects, holes intensity depth, foreground/background intensity properties, normalized image intensity properties, object localization and object shape. Final image classification is carried out through a kernel SVM and includes two independently trained classification models, one for the positive level of the image intensity and the other for the intermediate level.

SAR - The method first applies histogram equalization on the foreground part of the image, then resizes it to the size 100×100 pixels. Then the following features are extracted: statistical features as average intensity, average contrast, smoothness, skewness, uniformity and entropy [44]; invariant moments [44], Haralick features [41], discrete wavelet frame texture descriptors [45] with three resolution levels. The classification is performed using the maximum probability normal classifier.

SHE - In the proposed approach [46], each cell image is represented through the combination of two rotationally invariant descriptors based on SIFT [47] and Co-occurrence LBP [48]. In particular, for SIFT approach, a large number of SIFT features are clustered to form a dictionary, which is then used for cell representation. For pairwise LBP, the uniform pattern LBP operator was applied to two neighboring points for feature extraction.

Finally, the two features are fused and input to a multi-class SVM with linear kernel trained with one vs one strategy.

STO - In the preprocessing stage image denoising, normalization and enhancement are performed. For each type of adopted image descriptor a separate feature space with its proper metric is employed and specifically: the LBP descriptor, the Haralick features, the color structure, the surface descriptor and the radial cell structure descriptor are used with L1 metric, while for the granulometry descriptor, the author defines a specific distance function. The final classification is obtained through a custom combination of $k-$NN searches over the considered feature spaces.

THE - The proposed method preprocesses the input image by denoising and normalization. Then, a set of binary images are constructed by thresholding the image using a set of 14 equally spaced threshold values in the range $[0, 1]$. Connected component analysis is performed in each binary image, and the following set of morphological features is computed for each of them: number of detected objects, density in binary image and mean objects solidity. Objects of size less than 1% of the mean objects size of each binary image, are considered as noise and ignored during the calculation of the above features. Finally, the complexity of the cells contour is considered and modeled as the difference between the cells contour and the perimeter of the equivalent circle. The resulting feature vector is normalized by subtracting the mean vector and dividing each feature by the standard deviation of the corresponding values of the training set. The final classification is performed using the $k-$NN classification rule.

THI - The proposed method is based on two different statistical descriptors. The Fuzzy Size Zone Matrix, a fuzzy version of Gray Level Size Zone Matrix [49], which provides a statistical representation by the estimation of a bivariate conditional probability density function of the image distribution values. Features are calculated as moments of order -2 and 2 computed on such a matrix. The Multi-resolution Local Binary Patterns is a rotation and gray level invariant descriptor which attributes to each pixel a unique code according to its neighborhood. The feature vector is derived from the histogram of the codes. Classification is done according to one-vs-all scheme. In particular, for each class two classifiers (one per descriptor) are built with Random Forests [50] and the probabilities provided by the model are averaged in order to provide a final probability for the cell under study to belong to the class. The cell is then labelled with the class corresponding to the highest probability.

5. Experiments

In this section, we analyze the results of the benchmarking activity that we conducted within the ICIP 2013 competition on IIF images. The analysis is focused on the following aspects: first, we briefly review the adopted experimental protocol; then, we consider the classification results at the cellular and specimen levels[5] of the fourteen submitted methods over the test set of the ICIP2013 dataset by analyzing in detail the behavior of the methods with respect to the different cellular staining patterns; finally, we compare the results with those presented in [2] with the aim of drawing several general conclusions.

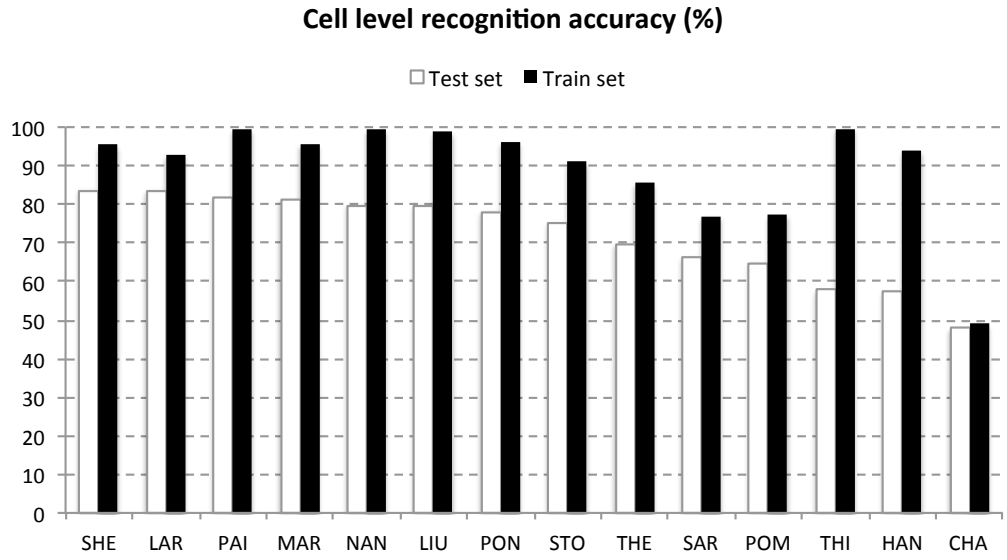


Figure 5: The cell recognition accuracy obtained by the considered methods over the train and the test set. The performance is sorted in descending order left to right with respect to the test set (i.e.the best performing method, SHE, is listed first).

[5] As in [2], we use here the expression *at the cell level* when we refer to a single cell without the surrounding part of the specimen image, while we use the expression *at the specimen level* as a synonym of the expression *at the image level*, used in [2], when we refer to the specimen image as a whole.

5.1. Experimental protocol

We briefly summarize the protocol as follows. Each participant receives the train set with the original images of the automatically segmented cells. In particular, for each cell image we provide the bounding box and the foreground mask. The cells are provided along with the information about the intensity pattern and the ID of the specimen image they belong to. This information is critical for creating unbiased cross-validation splits during training. More specifically, in order to construct the unbiased cross-validation splits during training, one needs to ensure that cell images extracted from the same patient are not in both training and testing sets. We note that this information is not available in the ICPRContest2012, thus, severely disadvantaging the contest participants in training their systems. Therefore, the adopted experimental protocol in this work is not identical to the ICPRContest2012 widely described in [2].

The participants use the train set provided to tune their systems and then release an executable file for independent evaluation on the test set. Finally, submitted executables are run on the train and test sets. The results are discussed below.

5.2. Performance analysis

Each team performance is evaluated based on the Correct Classification Rate (CCR), $CCR = \frac{TP+TN}{P+N}$, where TP, TN, P and N are the number of True Positive, True Negative, Positive and Negative samples, respectively. We note that this performance evaluation might be biased towards the method performance on common classes. To that end, we also analyse the accuracy for each class.

Figure 5 shows the cell recognition accuracy attained by each method on both the train and test sets. We firstly observed that performance varied in a wide range, from 47.19% to a maximum value at 83.65%. However, it is interesting to note that the top seven performing methods are contained within a much smaller range (approximately 6 percentage points). In fact, the performance gap between the best and the second best methods are markedly similar (i.e. 83.65% vs 83.54%). We further confirm that the top two methods are similar by applying the Cochran's Q test [51] where the null hypothesis cannot be rejected at 5% significance level (i.e. $p = 0.5065$).

Generally, the best performing methods make use of two ingredients: (1) features extracted from the local statistics of an image and (2) using a strong classification framework. For instance, the contest winner, SHE, employs two

local feature descriptors based on SIFT and Co-occurence LBP. The method uses the SVM training framework; a strong classification maximising margin between the classes. Furthermore, for the case of linear SVM, the weights on the SVM model indicate how important a particular feature to the classifier output. This could provide an implicit feature selection. Another example is the second best method, LAR, which uses a novel descriptor namely "annulus" shape index histogram features which introduces a rotation-invariant spatial pooling scheme over the shape index histograms. Again, they use a strong classifier such as a multi-class one-vs-one SVM in conjunction with the RBF Kernel. From this observation, we conjecture that the combination of the local descriptor and a strong classifier has significant relevance. The SVM seems to be an effective classifier for this problem as it offers good generalisation error as well as an effective feature selection process.

We observe that for the second tier methods, the above two ingredients either only appear individually or not in the right balance. For example, although STO does use local feature descriptors such as LBP, it does not employ a strong classifier. Instead, it uses k-NN method as the classifier. Another example is the THI method. In this method both ingredients are present. The feature selection is done via a probabilistic framework which may be prone to over training (Refer to Figure 5) when the cross-validation training protocol is not carefully constructed.

Results reported in Figure 5 highlight that in a large number of cases there is a very high discrepancy between the cell level recognition accuracy attained by each method in both the train and test sets. In fact, such a difference is generally around or above 10%, reaching the maximum values in the cases of THI and HAN (41.97% and 36.52%, respectively). The unique exception is represented by CHA: in this case the difference is only 1.21%. This can be motivated by the design of the method that was kept simple (e.g. the fact that the binary classifiers adopted in the classification stage are combined in hand-crafted decision tree); this choice could potentially avoid overfitting, thus achieving by far the highest generalization level with respect to the other submissions, despite its low overall accuracy on the test set.

In order to focus the attention on the recognition capabilities of the methods with respect to six staining patterns of the cells, we report in Figure 6 the confusion matrices of all the methods. A first observation is that the method by SHE, while obtaining the highest global accuracy, it never achieves the highest class accuracy; in fact, the method by LAR (the 2nd ranked) is the

	Centromere	Homogeneous	Nucleolar	Speckled	Nuclear Memb.	Golgi
Centromere	93.83% (1.51)	0.24% (0.18)	2.29% (0.74)	2.95% (0.92)	0.35% (0.09)	0.34% (0.17)
Homogeneous	0.46% (0.18)	71.93% (4.41)	5.67% (1.15)	10.55% (1.87)	10.43% (2.69)	0.95% (0.58)
Nucleolar	1.50% (0.71)	1.45% (0.43)	90.94% (1.58)	2.58% (0.83)	2.52% (1.09)	1.01% (0.64)
Speckled	14.13% (3.40)	12.56% (2.48)	3.25% (0.60))	66.08% (3.69)	3.33% (0.86)	0.65% (0.37)
Nuclear Memb.	0.28% (0.11)	4.54% (1.72)	1.95% (0.88)	2.24% (0.49)	89.4% (2.01)	1.59% (1.21)
Golgi	0.77% (0.36)	6.34% (2.53)	10.95% (2.41)	2.09% (1.03)	18.58% (5.31)	61.27% (5.43)

Table 1: Each entry of the table contains the average value and the standard deviation (in parentheses) calculated over the homologous entries of the cell classification accuracy confusion matrices of the best seven methods that participated to the ICIP 2013 contest.

	Centromere	Homogeneous	Nucleolar	Speckled	Nuclear memb.	Golgi
Centromere	98.44% (1.28)	0.22% (0.59)	0.22% (0.59)	1.12% (1.18)	0.00% (0.00)	0.00% (0.00)
Homogeneous	0.00% (0.00)	81.25% (3.93)	4.46% (1.67)	6.92% (2.83)	7.14% (2.36)	0.22% (0.59)
Nucleolar	0.00% (0.00)	1.12% (1.18)	97.10% (1.41)	0.45% (0.76)	1.34% (1.08)	0.00% (0.00)
Speckled	13.17% (4.85)	7.81% (4.42)	0.45% (0.76)	77.01% (5.83)	1.56% (1.80)	0.00% (0.00)
Nuclear Memb.	0.00% (0.00)	3.10% (2.24)	0.00% (0.00)	0.48% (0.81)	96.19% (2.09)	0.24% (0.63)
Golgi	0.00% (0.00)	2.26% (5.97)	4.51% (3.63)	0.00% (0.00)	13.53% (10.01)	79.70% (7.70)

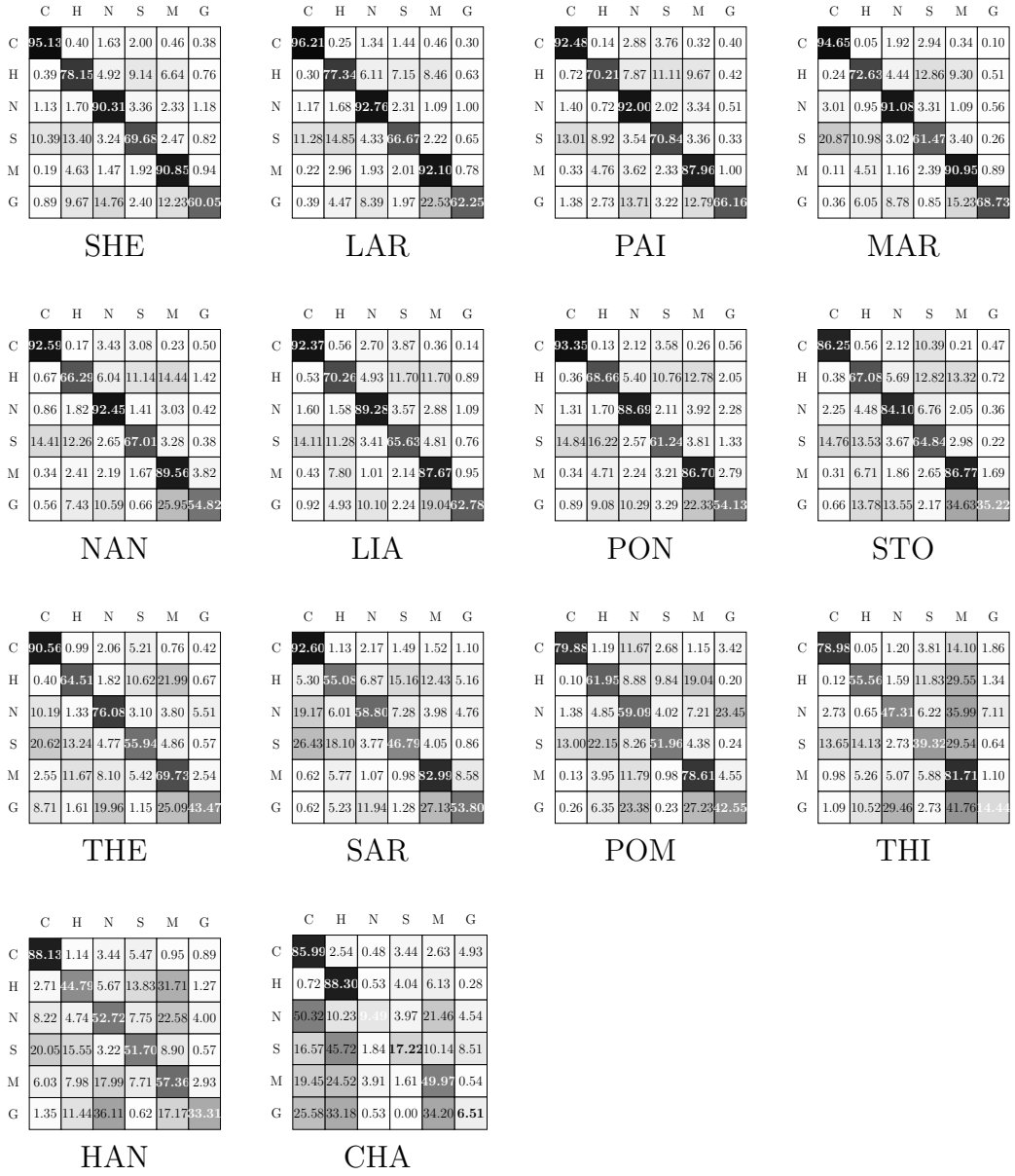
Table 2: Each entry of the table contains the average value and the standard deviation (in parentheses) calculated over the homologous entries of the specimen classification accuracy confusion matrices of the best seven methods that participated to the ICIP 2013 contest.

best performing over the centromere, nucleolar and nuclear membrane patterns, the methods by PAI (the 3rd ranked), by MAR (the 4th ranked) and by CHA (the 14th ranked) have the highest value of accuracy on the speckled, Golgi and homogeneous patterns, respectively.

We note that the best two methods show similar behaviors with respect to all the classes (the difference of the class accuracies of the two methods is always below 3%); further confirming the statistical significance test performed previously. When focusing again on the best seven approaches we notice that all share a common behavior over the six classes. This observation is supported by the data reported in Table 1 where each entry contains the average value of the cell classification performance and, in parentheses, the standard deviation calculated over the values of the homologous entries of the confusion matrices of the best seven methods. It is noteworthy to mention that in most cases the standard deviation is below 3% confirming that at least for the best seven methods there is a common trend in the recognition capabilities. The best seven approaches recognize the centromere, the nucleolar and the nuclear membrane patterns in around 9 cases out of 10, performing worst on the remaining classes with a class accuracy ranging within 60% and 70%.

From the data in Table 1 we notice that the homogeneous pattern is often confused with the speckled and the nuclear membrane and partially with

Figure 6: Confusion matrices of all the fourteen methods that participated to the ICIP 2013 competition. The label of the row/column is the true/guessed class name, with the following meanings: C = centromere, H = homogeneous, N = nucleolar, S = speckled, M = nuclear membrane, G = Golgi.



the nucleolar; such behavior is not surprising because the homogeneous pattern visually appears as a smoothed version of the speckled (both coarse and fine); a similar consideration stands for the nuclear membrane pattern that visually resembles the homogeneous pattern (notice as a large fraction of the errors on the nuclear membrane pattern is due to misclassification with the homogeneous). Furthermore, the speckled pattern is almost equally confused with the homogeneous and centromere patterns; the first type of error is coherent with the previous finding regarding the errors with the homogeneous patterns; the confusion with the centromere can be justified through the observation that both patterns are characterized by the presence of small dots on a quite homogeneous cellular background, with the only difference that the dots are darker than the background in the case of the speckled while are lighter in the case of the centromere. The confusion between speckled and centromere patterns seems to have originated from the use of features which do not explicitly encode the gradient directionality. Thus, future research efforts should also be directed toward the development of appropriate descriptors for addressing this issue. Finally, in the case of the Golgi pattern we deem that the low recognition rate is due to a combination of three factors: (1) the skewness of the dataset (the number of images belonging to Golgi pattern is significantly less than the other patterns); (2) the visual similarity with other patterns (in particular with nuclear membrane) and (3) the fact that the discriminating visual cues for Golgi pattern occurring at the periphery of the cell are outside the provided binary masks. This may need a dedicated directive in the future benchmarking campaigns.

5.2.1. Further performance analysis

From previous discussions, we observed that the top seven methods were similar in terms of their performance (i.e. space ranging from 78%-83%). However it is not clear whether their performance is indicative of a similar error profile. In other words, we cannot draw any conclusions whether these methods generally have mistakes on the same images. Again, we performed the Cochran's Q test against the top seven methods. The test concluded that the null hypothesis can be rejected at the 5% significance level; indicating that the these methods have different error profiles. We note that unlike in the previous section where the Cochran's Q test was performed against the top two methods, for this time, we performed the test against the top seven methods. To further confirm this result, we introduced a fusion rule whose input is the output of selected participant classifiers.

Formally, we define the fusion rule φ_f as follows. Let $\boldsymbol{X} \in \mathbb{R}^k$ be a vector whose the i-th component be the output of the i-th participant method φ_i; $k = \{1, 3, 5, 7, 9, 11, 13\}$ be the number of selected participant methods. The fusion rule is described as:

$$\varphi_f(\boldsymbol{x}) = \underset{c \in \mathcal{C}}{\operatorname{argmax}} \ \ \xi(\boldsymbol{X}, c) \tag{1}$$

where $\xi(\boldsymbol{X}, c)$ is the function that counts the number of c in $\boldsymbol{X}$; $\mathcal{C}$ is the set of pattern classes $\mathcal{C} \in \{$centromere, homogeneous, nucleolar, speckled, nuclear membrane, Golgi$\}$. For instance, $\xi([\text{homogeneous}, \text{homogeneous}, \text{speckled}], \text{homogeneous})$ equals 2 since there are two homogeneous patterns in $\boldsymbol{X}$.

We first ranked the participant methods according to their recognition performance as presented in Fig. 5. Then, we evaluated the fusion rule performance by progressively selecting the methods ordered by their performance in descending order. Figure 7 reports the results of the study. We found that it was possible to improve the performance of the ICIP2013 winner by 2% point. The fusion rule reached its optimal performance (85.60%) when the top seven methods were employed, thus, corroborating our previous observation that these methods are heading in the right direction to solve the classification problem. This also indicates that each participant does not have the same classification errors suggesting that there could be more room to improve performance.

We note that the performance improvement resulted from the fusion rule was significant due to the significantly large number of test data (i.e.2% means there were additional 1,097 test data points correctly classified).

The Cochran's Q test was also performed to further analyse whether the error profile of the fusion rule differed to the top seven methods. More precisely, the test was applied on paired output of the fusion rule and one of the top seven methods. This resulted in seven tests. All tests rejected the null hypothesis indicating that the error profile and the performance gains of the fusion rule over the top seven methods was statistically significant.

Figure 8 shows the performance of each method in positive and intermediate fluorescence intensity images. All methods consistently have lower performance on intermediate fluorescence intensity images than the positive fluorescence intensity images. This indicates that the classification problem in intermediate fluorescence intensity images is much more complex than in positive ones.

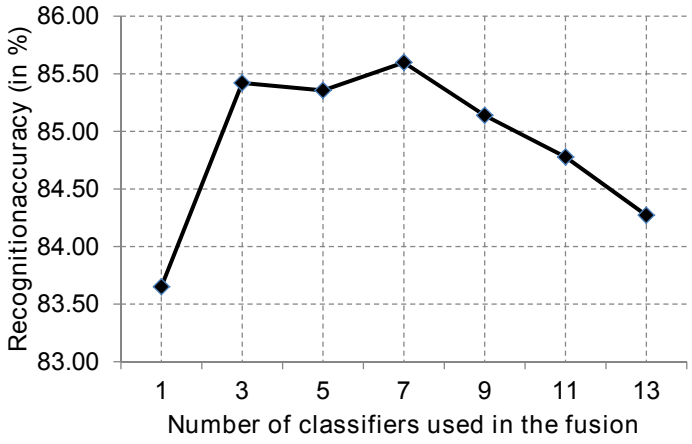


Figure 7: The recognition accuracy at the cell level over the test set obtained by the fusion rule for different number of classifiers in the fusion.

Figure 9 reports the specimen-level performance of the considered methods. The label of a specimen image is simply determined by the most dominant cell pattern in the image. It is noteworthy to mention that the top performing methods could not achieve 100% performance despite their excellent performance in the cell image classification. Upon closer examination of the top seven methods, we found that most error was associated with the speckled patterns. This can be observed in Table 2 that reports the specimen classification performance of the top seven methods. In addition, this is consistent with the previous finding that considers speckled as the second most difficult pattern to classify due to misclassification with other patterns such as centromere, nuclear membrane, homogeneous and nucleolar.

5.3. Lessons from the past

Although a direct comparison of the classification performance of the methods that were analyzed at the ICIP 2013 and at the ICPR 2012 contests [2] is not immediate as the datasets are different and the cellular staining patterns under consideration only partially overlap, indeed some general considerations can be drawn.

A first observation is that we immediately notice a general increase of the recognition rates and a reduction of the performance variability for the newly considered methods. Both phenomena can be explained by the availability of a larger dataset which allowed more reliable training and testing of the

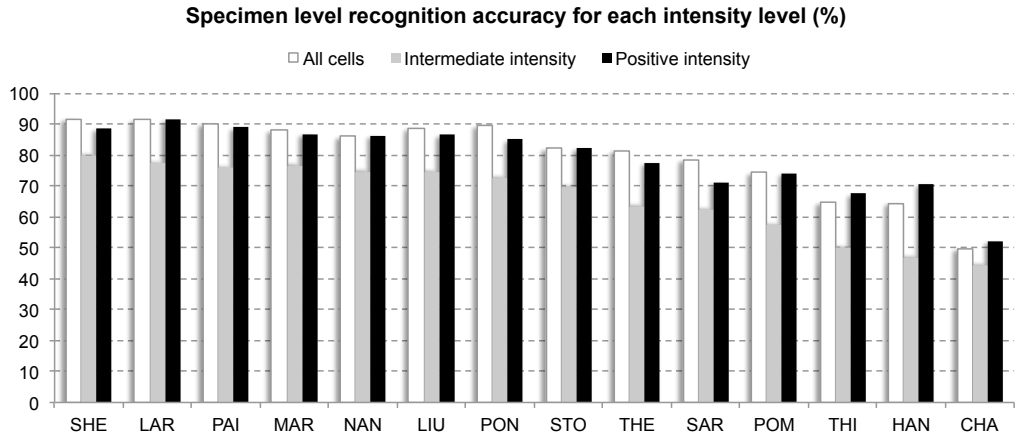


Figure 8: The recognition accuracy at the cell level obtained by considered methods over the test set for intermediate and positive fluorescence intensities.

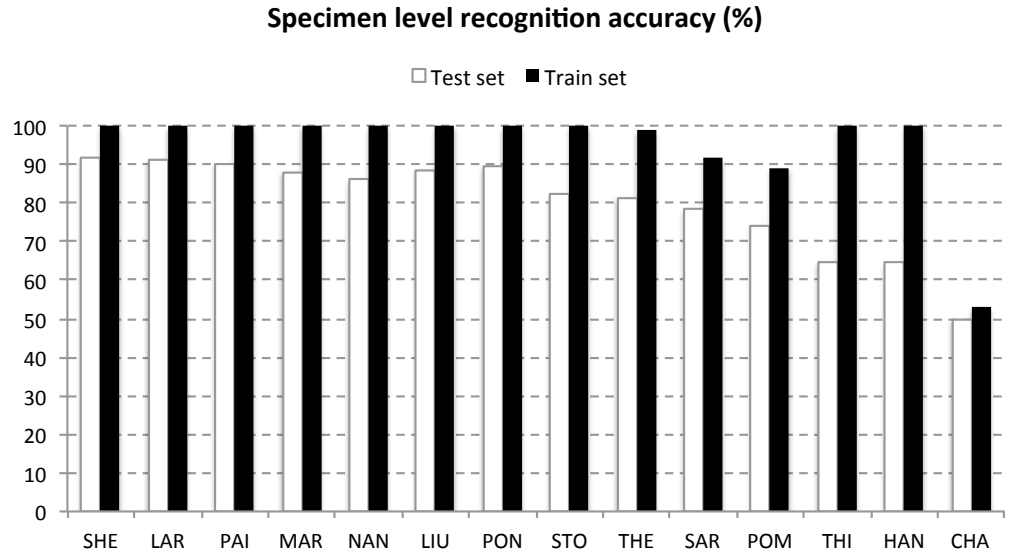


Figure 9: The recognition accuracy at the specimen level obtained by the considered methods over the train and the test set.

approaches and by the experience gained through the ICPR2012Contest that provided important cues for designing more effective methods.

We note that most methods can reliably classify centromere pattern. In fact cells belonging to this class were recognized with the highest accuracy also in the ICPR2012Contest. In both current benchmarking work and ICPR2012Contest, the centromere cells were correctly recognized in more than 9 cases out of 10. The class accuracy of the nucleolar pattern increases to almost 10% (90.3% of the best method of ICIP 2013 competition, SHE, versus 80.6% of the best method of ICPR2012Contest). The homogeneous and the speckled patterns still remain more difficult to recognize confirming the outcomes of the previous competition although significant improvements can be observed: the class accuracy of the best method of the ICPR2012Contest with respect to the above two patterns was slightly over 50% while the method by SHE in 2013 improved to around 15% (speckled) and 25% (homogeneous).

6. Main findings and future direction

In this paper we have presented an in-depth analysis of the benchmarking results from the international competition on HEp-2 cell classification held in conjunction with the International Conference of Image Processing 2013. The benchmarking platform described here overcomes some limitations of the previous benchmarking platform (ICPR2012Contest) through the adoption of a brand new dataset consisting of a significantly larger number of images and offering a more realistic evaluation by providing more than one specimen image for each pattern and by introducing less common pattern images such as nuclear membrane and Golgi. We have performed an evaluation on 14 methods on the same test set that has not been released to the contest participants.

From this evaluation, we found several significant findings:

- The first seven top performing methods such as: SHE, LAR, PAI, MAR, NAN, LIU and PON are closely matched in terms of performance.

- We found that top performing methods normally employ two ingredients: (1) features extracted from local statistics of an image (e.g. bag of words approach) and (2) a strong classifier. We conjecture that

the most effective solution for the latter point is to employ an SVM classifier.

- We observed that there is a significant discrepancy amongst most participant's performance on train and test sets. This may suggest that in most cases participants overtrained their system.

- On closer observation we found that although the seven top performing methods have similar performance, they do not have the same error profile. This was shown in further evaluation of the fusion classifier which was able to achieve better performance than the competition winner (85.36% vs 83.65%).

We also had some confirmations from the experience of the ICPR2012Contest.

- The classification problem on intermediate fluorescence intensity images is still considered more difficult as the performance of all participants on this set of images is lower than their performance on the positive fluorescence intensity images. We envision that the future methods will give different treatment for images with different fluorescence intensity.

- We verified that some staining patterns are simpler to recognize. This is the case for centromere and for nucleolar patterns that can recognized in more than 9 cases out of 10.

- We found the speckled pattern to be a source of confusion for the participant methods when classifying specimen images. The speckled pattern has been noted as the second most confused pattern class in cell image classification problem.

The substantial and systematic effort to solve HEp-2 image classification problems started since the previous benchmarking work at the ICPR 2012 Cell Classification Competition has led to a series of high quality publications in numerous venues. The present work provides insight into how far we are from the prescribed goal: to develop a reliable and robust HEp-2 ANA test CAD system that can be used for routine operation in pathology laboratories. From the present analysis, we found several important ingredients to make CAD systems successful in performing the classification task. We also found several potential issues wherein solving these could significantly

improve the classification success rate of CAD systems. Some of the issues are also linked to the previous benchmarking work which then should receive more attention from the community. Among these issues, we deem that the most noteworthy ones that should be considered in the future are related to the classification problems on intermediate fluorescence intensity images, possibly investigating also on the impact that the explicit use of the fluorescence intensity information might have over the achievable accuracy.

There are also several other questions worth exploring that could significantly advance the field such as how we can effectively classify HEp-2 specimen images. Currently, the specimen image classification is merely carried out by using the dominant cell pattern presence in the image. Despite the high performance exhibited in the present evaluation, this approach may not be appropriate for other ANA patterns that do not have a dominant pattern such as mitotic spindle pattern and cell cycle dependent patterns. This warrants further investigation of the mitotic cell pattern classification problem. It is also not clear how current CAD systems deal with cases where multiple ANA patterns exist within a patient serum.

We also believe that future benchmarking initiatives in this area should take into account issue related to the reproducibility of the research, to allow validation of methodologies on other datasets, to assess the robustness with respect to the choice of input parameters, to analyse the computational burden, and so on.

Given the steady advancement witnessed in the present work, we are confident that despite the long road to accomplishing the goal, we are getting closer to solving the problems posed in this challenging area.

Acknowledgment

The authors would like to thank the teams that participated to the competition held at the International Conference on Image Processing in 2013, whose members are reported below.
CHA: V. Chandran, J. Banks, B. Chen, I. Toneo-Reyes, *Queensland University of Technology, Australia.* **HAN**: X. Han, J. Wang, Y. Chen, *Ritsumeikan University, Japan.* **LAR**: A.B.L. Larsen, J.S. Vestergaard, R. Larsen, *Technical University of Denmark, Denmark.* **LIU**: L. Liu[1], J. Zhang[2], L. Wang[2], [1] *Australian National University, Australia*, [2] *University of Wollongong, Australia.* **MAR**: R. Marée, *University of Liège, Belgium.* **NAN**: L. Nanni[1], M. Paci[2,3], J. Hyttinen[2,3], S. Severi[4], [1] *University of Padua, Italy*, [2] *Tampere*

University of Technology, Finland, [3]*BioMediTech, Finland*, [4]*University of Bologna, Italy*. **PAI**: S. Paisitkriangkrai, R. Hill, C. Shen, A. den Hengel, *University of Adelaide, Australia*. **POM**: V. Pomponiu, H. Hariharan, *University of Pittsburgh, USA*. **PON**: G.V. Ponomarev, M.S. Gelfand, M.D. Kazanov, *Institute for Information Transmission Problems, Russia*. **SAR**: O. Sarrafzadeh, H. Rabbani, *Isfahan University of Medical Sciences, Iran*. **SHE**: L. Shen, J. Lin, S. Yu, *Shenzhen University, China*. **STO**: R. Stoklasa, *Masaryk University, Czech Republic*. **THE**: I. Theodorakopoulos, D. Kastaniotis, *University of Patras, Greece*. **THI**: G. Thibault, *Oregon Health & Science University, USA*.

This work has been partly funded by Sullivan Nicolaides Pathology, Australia and the Australian Research Council (ARC) Linkage Projects Grant LP130100230.

[1] P. J. Tadrous, Computer-assisted screening of ziehl-neelsen-stained tissue for mycobacteria. algorithm design and preliminary studies on 2,000 images, American Journal of Clinical Pathology 133 (6) (2010) 849–858.

[2] P. Foggia, G. Percannella, P. Soda, M. Vento, Benchmarking HEp-2 cells classification methods, Medical Imaging, IEEE Transactions on 32 (10) (2013) 1878–1889. doi:10.1109/TMI.2013.2268163.

[3] R. D. Labati, V. Piuri, F. Scotti, All-IDB: the acute lymphoblastic leukemia image database for image processing, in: 2011 18th IEEE International Conference on Image Processing (ICIP), IEEE, 2011, pp. 2045–2048.

[4] D. C. Wilbur, Digital cytology: current state of the art and prospects for the future, Acta Cytologica 55 (3) (2011) 227–238.

[5] M. N. Gurcan, L. E. Boucheron, A. Can, A. Madabhushi, N. M. Rajpoot, B. Yener, Histopathological image analysis: A review, Biomedical Engineering, IEEE Reviews in 2 (2009) 147–171.

[6] P. L. Meroni, P. H. Schur, ANA screening: an old test with new recommendations, Annals of the Rheumatic Diseases 69 (8) (2010) 1420 –1422.

[7] A. S. Wiik, M. Hier-Madsen, J. Forslid, P. Charles, J. Meyrowitsch, Antinuclear antibodies: A contemporary nomenclature using HEp-2 cells, Journal of Autoimmunity In Press, Corrected Proof.

[8] N. Bizzaro, R. Tozzoli, E. Tonutti, A. Piazza, F. Manoni, A. Ghirardello, D. Bassetti, D. Villalta, M. Pradella, P. Rizzotti, Variability between methods to determine ANA, anti-dsDNA and anti-ENA autoantibodies: a collaborative study with the biomedical industry, Journal of Immunological Methods 219 (1-2) (1998) 99–107.

[9] B. Pham, S. Albarede, A. Guyard, E. Burg, P. Maisonneuve, Impact of external quality assessment on antinuclear antibody detection performance, Lupus 14 (2) (2005) 113–119.

[10] R. Hiemann, T. Bttner, T. Krieger, D. Roggenbuck, U. Sack, K. Conrad, Challenges of automated screening and differentiation of non-organ specific autoantibodies on HEp-2 cells, Autoimmunity Reviews 9 (1) (2009) 17–22.

[11] P. Foggia, G. Percannella, A. Saggese, M. Vento, Pattern recognition in stained HEp-2 cells: Where are we now?, Pattern Recognition 47 (7) (2014) 2305 – 2314.

[12] A. Rigon, P. Soda, D. Zennaro, G. Iannello, A. Afeltra, Indirect immunofluorescence in autoimmune diseases: Assessment of digital images for diagnostic purpose, Cytometry Part B - Clinical Cytometry 72 (6) (2007) 472–477.

[13] P. Perner, H. Perner, B. Mller, Mining knowledge for HEp-2 cell image classification, Artificial Intelligence in Medicine 26 (2002) 161–173.

[14] P. Elbischger, S. Geerts, K. Sander, G. Ziervogel-Lukas, P. Sinah, Algorithmic framework for HEp-2 fluorescence pattern classification to aid auto-immune diseases diagnosis, in: IEEE International Symposium on Biomedical Imaging: From Nano to Macro, 2009, pp. 562–565.

[15] T. Hsieh, Y. Huang, C. Chung, Y. Huang, HEp-2 cell classification in indirect immunofluorescence images, in: Int. Conf. Information, Communications and Signal Processing, 2009, pp. 1–4.

[16] P. Foggia, G. Percannella, P. Soda, M. Vento, Early experiences in mitotic cells recognition on HEp-2 slides, in: Computer-Based Medical Systems (CBMS), 2010 IEEE 23rd International Symposium on, 2010, pp. 38–43. doi:10.1109/CBMS.2010.6042611.

[17] A. Wiliem, P. Hobson, R. Minchin, B. Lovell, An automatic image based single dilution method for end point titre quantitation of antinuclear antibodies tests using HEp-2 cells, in: Digital Image Computing: Techniques and Applications, Noosa, Australia, 2011.

[18] A. Wiliem, Y. Wong, C. Sanderson, P. Hobson, S. Chen, B. C. Lovell, Classification of human epithelial type 2 cell indirect immunofluoresence images via codebook based descriptors, in: IEEE Workshop on Applications of Computer Vision (WACV), 2013.

[19] G. Iannello, G. Percannella, P. Soda, M. Vento, Mitotic cells recognition in HEp-2 images, Pattern Recognition Letters 45 (1) (2014) 136–144.

[20] P. Foggia, G. Percannella, P. Soda, M. Vento, Special issue on the analysis and recognition of indirect immuno-fluorescence images, Pattern Recognition 47 (7) (2014) 2303 – 2304.

[21] A. Watanabe, M. Kodera, K. Sugiura, T. Usuda, E. M. Tan, Y. Takasaki, Y. Tomita, Y. Muro, Anti-dfs70 antibodies in 597 healthy hospital workers, Arthritis and Rheumatism 50 (3) (2004) 892–900.

[22] Y. Chai, V. Lempitsky, A. Zisserman, Symbiotic segmentation and part localization for fine-grained categorization, in: International Conference on Computer Vision, 2013.

[23] P.-H. Gosselin, N. Murray, H. Jègou, F. Perronnin, Revisiting the fisher vector for fine-grained classification, Pattern Recognition Letters 49 (1) (2014) 92–98.

[24] Y. Kumar, A. Bhatia, R. Minz, Antinuclear antibodies and their detection methods in diagnosis of connective tissue diseases: a journey revisited, Diagnostic Pathology 4 (1) (2009) 1.

[25] Quality assurance for the indirect immunofluorescence test for autoantibodies to nuclear antigen (IF-ANA): Approved guideline, Vol. 16, NCCLS I/LA2-A. Wayne, PA, 1996.

[26] P. Hobson, B. C. Lovell, G. Percannella, M. Vento, A. Wiliem, Classifying HEp-2 specimen images: A benchmarking platform, in: International Conference on Pattern Recognition (ICPR), 2014.

[27] A. Wiliem, P. Hobson, R. F. Minching, B. C. Lovell, A bag of cells approach for antinuclear antibodies hep-2 image classification, Cytometry Part A. In press.

[28] G. Percannella, P. Soda, M. Vento, Mitotic HEp-2 cells recognition under class skew, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 6979 LNCS (PART 2) (2011) 353–362.

[29] J. Kapuscinski, Dapi: A dna-specific fluorescent probe, Biotech Histochem 70 (5) (1995) 220–233.

[30] V. Chandran, S. Elgar, Pattern recognition using invariants defined from higher order spectra- one dimensional inputs, Signal Processing, IEEE Transactions on 41 (1) (1993) 205–212.

[31] Y. Freund, R. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, Journal of Computer and System Sciences 55 (1) (1997) 119–139.

[32] A. Larsen, J. Vestergaard, R. Larsen, HEp-2 cell classification using shape index histograms with donut-shaped spatial pooling, Medical Imaging, IEEE Transactions on 33 (7) (2014) 1573–1580.

[33] J. J. Koenderink, A. J. van Doorn, Surface shape and curvature scales, Image and Vision Computing 10 (8) (1992) 557 – 564.

[34] L. Liu, L. Wang, HEp-2 cell image classification with multiple linear descriptors, Pattern Recognition 47 (7) (2014) 2400 – 2408. doi:http://dx.doi.org/10.1016/j.patcog.2013.09.022.

[35] R. Marée, P. Geurts, L. Wehenkel, Towards generic image classification: an extensive empirical study, Tech. rep., University of Liege (2014). URL http://orbi.ulg.ac.be/retrieve/217317/270648/maree-Towards-TR2014-full.pdf

[36] P. Geurts, D. Ernst, L. Wehenkel, Extremely randomized trees, Machine Learning 63 (1) (2006) 3–42.

[37] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, C. Bray, Visual categorization with bags of keypoints, in: In Workshop on Statistical Learning in Computer Vision, ECCV, 2004, pp. 1–22.

[38] F. Moosmann, E. Nowak, F. Jurie, Randomized clustering forests for image classification, Pattern Analysis and Machine Intelligence, IEEE Transactions on 30 (9) (2008) 1632–1646. doi:10.1109/TPAMI.2007.70822.

[39] X. Qian, X.-S. Hua, P. Chen, L. Ke, Plbp: An effective local binary patterns texture descriptor with pyramid representation, Pattern Recogn. 44 (10-11) (2011) 2502–2515.

[40] P. Strandmark, J. Ulén, F. Kahl, HEp-2 staining pattern classification, in: Int. Conf. Pattern Recognition, 2012.

[41] R. Haralick, K. Shanmugam, I. Dinstein, Textural features for image classification, Systems, Man and Cybernetics, IEEE Transactions on SMC-3 (6) (1973) 610–621.

[42] R. Nosaka, Y. Ohkawa, K. Fukui, Feature extraction based on co-occurrence of adjacent local binary patterns, in: Y.-S. Ho (Ed.), Advances in Image and Video Technology, Vol. 7088 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2012, pp. 82–91.

[43] F. Liu, Z. Tang, J. Tang, Wlbp: Weber local binary pattern for local image description, Neurocomputing 120 (2013) 325 – 335.

[44] R. C. Gonzalez, R. E. Woods, Digital Image Processing (3rd Edition), 3rd Edition, Prentice Hall, 2007.

[45] M. Sonka, V. Hlavac, R. Boyle, Image Processing, Analysis, and Machine Vision, Thomson-Engineering, 2007.

[46] HEp-2 image classification using intensity order pooling based features and bag of words, Pattern Recognition 47 (7) (2014) 2419 – 2427. doi:http://dx.doi.org/10.1016/j.patcog.2013.09.020.

[47] D. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision 60 (2) (2004) 91–110.

[48] X. Qi, R. Xiao, J. Guo, L. Zhang, Pairwise rotation invariant co-occurrence local binary pattern, in: Proceedings of the 12th European Conference on Computer Vision - Volume Part VI, ECCV'12, Springer-Verlag, Berlin, Heidelberg, 2012, pp. 158–171.

[49] G. Thibault, J. Angulo, F. Meyer, Advanced statistical matrices for texture characterization: Application to dna chromatin and microtubule network classification, in: Image Processing (ICIP), 2011 18th IEEE International Conference on, 2011, pp. 53–56.

[50] L. Breiman, Random forests, Machine Learning 45 (1) (2001) 5–32.

[51] W. G. Cochran, Biometrika 37 (3/4) (1950) 256–266.