Benchmarking HEp-2 Cells Classification Methods

Pasquale Foggia, Gennaro Percannella*, Paolo Soda, and Mario Vento

***Abstract*—In this paper, we report on the first edition of the HEp-2 Cells Classification contest, held at the 2012 edition of the International Conference on Pattern Recognition, and focused on indirect immunofluorescence (IIF) image analysis. The IIF methodology is used to detect autoimmune diseases by searching for antibodies in the patient serum but, unfortunately, it is still a subjective method that depends too heavily on the experience and expertise of the physician. This has been the motivation behind the recent initial developments of computer aided diagnosis systems in this field. The contest aimed to bring together researchers interested in the performance evaluation of algorithms for IIF image analysis: 28 different recognition systems able to automatically recognize the staining pattern of cells within IIF images were tested on the same undisclosed dataset. In particular, the dataset takes into account the six staining patterns that occur most frequently in the daily diagnostic practice: centromere, nucleolar, homogeneous, fine speckled, coarse speckled, and cytoplasmic. In the paper, we briefly describe all the submitted methods, analyze the obtained results, and discuss the design choices conditioning the performance of each method.**

***Index Terms*—Computer-aided diagnosis (CAD), HEp-2 cells classification, indirect immunofluorescence.**

I. Introduction

PATTERN recognition techniques are widely used in the field of medicine for the development of computer-aided diagnosis (CAD) systems. Such systems may support the physician in many ways: they can be adopted as a second reader, thus augmenting the physician's capabilities and reducing errors; they make it possible to perform a preselection of the cases to be examined, enabling the physician to focus the attention only on the most relevant cases and hence facilitating mass screening campaigns; they may aid the physician in carrying out the diagnosis; finally, they can be used as a tool for the instruction and training of specialized medical personnel.

These techniques have been applied to several fields of medical science, both for research purposes and for actual clinical practice: for instance, the detection and the classification of breast microcalcifications in mammographic images [1], [2], the differential count of white cells from bone marrow preparations [3], cell classification for the diagnosis of reactive histiocytic hyperplasia [4], the high-throughput topological analysis of lymphocytes in tissue sections [5], the shape and color classification of haematic cells [6], and many others.

Among such applications, over the few last years there has been a certain interest in the realization of CAD systems for the analysis of indirect immunofluorescence (IIF) images. IIF is a diagnostic methodology based on image analysis that reveals the presence of autoimmune diseases by searching for antibodies in the patient serum. As a result of its effectiveness, we have witnessed to a growing demand for diagnostic tests for systemic autoimmune diseases. Unfortunately, however, IIF as yet remains a subjective method that depends too heavily on the experience and expertise of the physician.

Indeed, it has been found that the inter-laboratories agreement is 92.6% for the simple task of positive/negative intensity classification, while it drops to 76.0% for the recognition of staining patterns, which is required for a more detailed diagnosis [7]. The main reasons for this variability are: 1) the lack of quantitative information supplied to physicians, 2) varieties of reading systems and optics, 3) the photo-bleaching effect caused by a light source irradiating the cells over a short period of time [8], 4) the low reproducibility of the diagnostic protocol.

This limitation has motivated the recent developments of tools and systems supporting the diagnostic procedure. The investigated topics have covered several aspects of this procedure, such as image acquisition [9], image segmentation [10]–[13], mitotic cell recognition [14], fluorescence intensity classification [15], and staining pattern recognition [12], [16]–[18].

However, research in the field of IIF image analysis is still in its early stages and has great potential for further growth. A review of the literature reveals that a comprehensive CAD system for IIF is not yet available, although different algorithms and methods exist. Furthermore, all existing approaches have been developed and tested on private datasets with different characteristics (e.g., they contain images acquired from wells prepared according to different criteria, they consider different classes of fluorescence intensity and of staining patterns, etc.), thus making it impossible to reproduce the results and to compare the different approaches.

The aim of this paper is to describe recent advances in this field. In particular we report on the outcomes of the first edition of the HEp-2 cells classification contest hosted by the 21st International Conference on Pattern Recognition (ICPR 2012). The contest focused on the comparison of systems able to automatically recognize the staining pattern of cells within IIF images. The initiative was well received by the scientific community, with 28 methods submitted for benchmarking. This paper describes all the methods analyzed within the competition and attempts to draw conclusions on key aspects in the design of the methods, achieving a high correlation with the obtained results.

Manuscript received May 15, 2013; revised May 28, 2013; accepted June 03, 2013. Date of publication June 18, 2013; date of current version October 02, 2013. *Asterisk indicates corresponding author.*

P. Foggia and M. Vento are with the Department of Information Engineering, Electrical Engineering and Applied Mathematics, University of Salerno, 84084 Fisciano, Italy (e-mail: pfoggia@unisa.it; mvento@unisa.it).

*G. Percannella is with the Department of Information Engineering, Electrical Engineering and Applied Mathematics, University of Salerno, 84084 Fisciano, Italy (e-mail: pergen@unisa.it).

P. Soda is with the Computer Science and Bioinformatics Laboratory, Integrated Research Centre, University Campus Bio-Medico of Rome, 00128 Roma, Italy (e-mail: p.soda@unicampus.it).

Digital Object Identifier 10.1109/TMI.2013.2268163

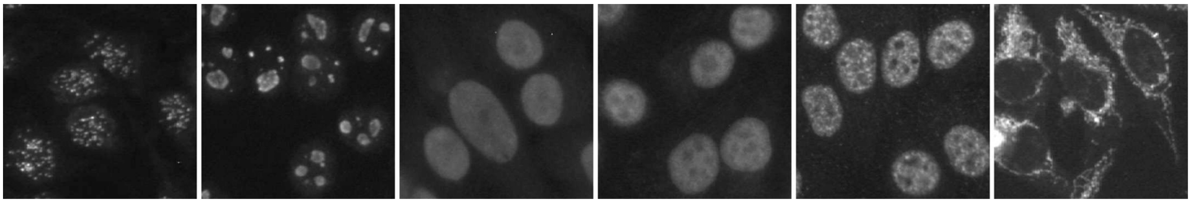


Fig. 1. Examples of HEp-2 cells with different staining patterns (from left to right): centromere, nucleolar, homogeneous, fine speckled, coarse speckled, cytoplasmic.

The present paper is organized as follows. First, we briefly describe the medical context and the procedure adopted by medical doctors to formulate a diagnosis, focusing on the crucial role played by pattern recognition techniques in order to set up a computer aided system in this applicative area. We then describe the dataset used for the benchmarking and the main characteristics of the tested methods. Finally, we analyze the results and draw some conclusions.

II. Medical Context

Antinuclear autoantibodies (ANAs) play a pivotal role in the serological diagnosis of autoimmune diseases, which are connective tissue disorders characterized by a chronic inflammatory process involving different organs. ANAs are directed against a variety of nuclear antigens and can be detected in patient serum through laboratory tests. In this respect, IIF is considered the "gold standard" for ANA testing [19]. IIF uses the human larynx carcinoma (HEp-2) substrate, which bonds with serum antibodies forming a molecular complex. This complex then reacts with human immunoglobulin conjugated with a fluorochrome and becomes visible under a fluorescence microscope which reveals the antigen-antibody reaction. During the reading phase under the microscope, medical doctors classify the fluorescence intensity, recognize mitotic cells[1] and classify the staining patterns for each well.[2]

In order to classify the fluorescence intensity, the guidelines established by the Center for Disease Control and Prevention (CDC), Atlanta, GA, USA [20] suggest scoring it semi-quantitatively and independently by two physicians who are experts in IIF. The score ranges from 0 up to 4+ as follows:

- 4+: brilliant green (maximal fluorescence);
- 3+: less brilliant green fluorescence;
- 2+: defined pattern but diminished fluorescence;
- 1+: very subdued fluorescence;
- 0 negative.

Since technical problems can affect test sensitivity and specificity, the same guidelines suggest comparing the sample with a positive and a negative control. The former allows the physician to check the correctness of the preparation process, whereas the latter represents the auto-fluorescence level of the slide under examination. Recently, Rigon *et al.* [21] statistically analyzed the variability between a set of physician's fluorescence intensity classifications and then proposed to classify the fluorescence intensity into three classes, namely negative, intermediate, and positive. Although the detailed description of these classes lies outside the scope of this paper, it is useful to report that they maintain the clinical significance of IIF testing and establish a more robust ground truth. Therefore, we have adopted this classification protocol to build the dataset used in this contest.

It is worth observing that fluorescence intensity affects the performance of the subsequent phases (mitotic cell recognition and staining pattern classification) since intermediate intensity samples usually exhibit low contrast, making the classification task harder even for experts.

With regard to mitotic cells, their presence assures medical doctors that the well has been correctly produced. Thus, the well is discarded if the number of observed mitotic cells is below a certain threshold (usually 1 or 2).

The last step in the reading phase consists of staining pattern recognition. This is an extremely challenging task as several patterns, corresponding to different autoimmune diseases, may be observed [22]. The most frequent and clinically useful ANA patterns are as follows.

- *Centromere*: characterized by several discrete speckles (≈40–60) distributed throughout the interphase[3] 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 nuclear staining of the interphase cell nuclei.

Besides ANA patterns, HEp-2 slides may also reveal the *cytoplasmic* staining pattern, which is relevant to diagnostic purposes since it can be associated with specific and heterogeneous autoantibodies, e.g., anti-Jo1, anti-M2. Indeed, many speckled and fiber cytoplasmic staining patterns are described [23]. We include all these staining patterns in the cytoplasmic class. Currently, such an assumption is reasonable and is adopted in all

[1] A mitotic cell is a cell in mitosis. Mitosis is the process by which a eukaryotic cell separates the chromosomes in its cell nucleus into two identical sets in two nuclei.

[2] A well is a portion of the slide where the patient serum is dispensed, diluted, and incubated to react with the substrate.

[3] The interphase is the nonmitotic phase of the cell cycle in which the cell spends the majority of its time and performs the majority of its purposes.

works dealing with IIF automation. Fig. 1 shows examples of the aforementioned staining patterns. In most cases, all the cells in the image show the same staining pattern; thus the image is labeled with the common pattern. In the remaining few cases where the cells of the image show more than one pattern, the image is said to have a *mixed* pattern. It is worth noting that mitotic cell observation also confirms the classification of some staining patterns when the interphase cells stainings are similar, e.g., fine speckled versus homogeneous. The interested reader may refer to [7] and [24] to further delve into IIF methods and related issues.

III. Dataset Description

In this section, we provide some information on the dataset used for the benchmarking activity, which is publicly available for research purposes and can be obtained from the web page in [25]. In particular, we describe the procedures and the equipment used for acquiring and labeling the pattern of the images and of the cells, and we go on to present its composition. For the sake of brevity, hereafter we will use the expressions *at the cell level* to mean "considering a single cell without the surrounding part of the image," and *at the image level* to mean "considering the image as a whole;" for instance, *classification at the cell level* will be used to mean "classification performed looking only at a single cell."

HEp-2 images were acquired by means of a fluorescence microscope (40-fold magnification) coupled with a 50W mercury vapor lamp and with a digital camera. The camera has a charge-coupled device with square pixels of 6.45 $\mu\mathrm{m}$. The images have a resolution of 1388 × 1038 pixels, a color depth of 24 bits and they are stored in an uncompressed format. Please note that the adopted acquisition system provides RGB color images for the benefit of the specialist, who is used to looking at the cells under the microscope and thus would not be comfortable with a gray level image, even though a single channel is sufficient to convey all the information. We have decided to provide the contest participants with the original images and leave them the choice of the preprocessing steps needed before features extraction, as would be the case if their classifiers interface directly with the acquisition system.

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 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.

The attribution of a fluorescence intensity level to an image, chosen from five intensity scores (as defined by [20]), proves to be a difficult task even for experts in the field. This is confirmed by the variability of the intensity labels on the same images produced by different medical doctors, in spite of their comparable experience. In order to take this variability into account, we asked a second expert from a different diagnostic center to independently evaluate the fluorescence intensity. According to the methodology proposed in [21], three different labels have been assigned to the images:

- *negative*, if both experts classified the image as negative;
- *positive*, if both experts attributed a score of 2+ or more;
- *intermediate*, otherwise.

At the cell level, the specialist annotated whether or not the cell is in mitosis and, in the latter case, also the staining pattern. In dubious cases, the specialist asked the second expert to independently label the cell; in case of disagreement, they determined the final label through mutual discussion.

Once the database had been constructed, the ground truth was validated with the assistance of another specialist from a different laboratory (an immunologist with 10 years experience) who independently labeled the cells.

At the image level, this validation provided the same classes as the ground truth for all the images. At the cell level, the validation expert disagreed on 11 cells, which the first expert had excluded considering them as artifacts. Since those cells were not included in the training and test sets used for the contest, this disagreement does not influence the reported results of the participating methods. It is worth noting that both the experts used for the ground truth creation and the one performing the validation had the opportunity to examine the whole image, thus also exploiting information on surrounding cells in the uncertain cases. Since all the used images contain cells with the same staining pattern (there are no *mixed pattern* images), this explains the high degree of agreement between the experts (together with the fact that all the experts viewed the very same images, not just images from the same patient acquired with different reading devices as in the study cited in [7]). Note that this is not always the case: experts start looking at the whole image, but then consider individual cells to confirm whether or not they have the same pattern as the whole image.

The dataset, whose composition is outlined in Table I, contains 28 images almost equally distributed with respect to the different patterns (four images belonging to the cytoplasmic, fine speckled and nucleolar patterns, five homogeneous and coarse speckled images, and six centromere ones) and intensities (from two to three images per intensity and pattern). The fourth column of the table also reports the number of cells in each image.

For the experimental analysis, the image dataset was divided into a training set and a test set. The composition of the two sets is described in Table II, where it is possible to note that the subdivision was performed while maintaining approximately the same image pattern distribution over the two sets. Note also that in our dataset the cells have the same pattern as the image to which they belong, i.e., there is no image with a mixed staining pattern. Furthermore, the number of cells per image varies significantly due to the cell culture process within companies, influencing the *a priori* distribution of cell patterns in the dataset which turns out to be somewhat imbalanced. For instance, the number of centromere cells is about three times higher than the number of cytoplasmic cells.

IV. Classification Methods

In this section, we provide a brief description of each of the 28 methods that were submitted to the contest, focusing on some

TABLE I
COMPOSITION OF THE HEp-2 IMAGES DATASET USED FOR BENCHMARKING. EACH ROW REPORTS DATA RELATED TO AN IMAGE: THE UNIQUE ID ASSIGNED TO THE IMAGE, THE PATTERN, THE INTENSITY, AND THE NUMBER OF THE CELLS WITHIN THE IMAGE

Image ID	Pattern	Intensity	Number of cells
1	Homogeneous	Positive	61
2	Fine speckled	Intermediate	48
3	Centromere	Positive	89
4	Nucleolar	Intermediate	66
5	Homogeneous	Intermediate	47
6	Coarse speckled	Positive	68
7	Centromere	Intermediate	56
8	Nucleolar	Positive	56
9	Fine speckled	Positive	46
10	Coarse speckled	Intermediate	33
11	Coarse speckled	Intermediate	41
12	Coarse speckled	Positive	49
13	Centromere	Positive	46
14	Centromere	Intermediate	63
15	Fine speckled	Intermediate	63
16	Centromere	Positive	38
17	Coarse speckled	Positive	19
18	Homogeneous	Positive	42
19	Centromere	Intermediate	65
20	Nucleolar	Intermediate	46
21	Homogeneous	Intermediate	61
22	Homogeneous	Positive	119
23	Fine speckled	Positive	51
24	Nucleolar	Positive	73
25	Cytoplasmic	Intermediate	24
26	Cytoplasmic	Positive	36
27	Cytoplasmic	Positive	38
28	Cytoplasmic	Intermediate	13

relevant aspects, such as the adopted preprocessing procedure, the features, the classification paradigm, etc. Please note that on the contest website [26] it is possible to download the technical report of the contest, containing a more extended description of each method written by its authors. Before presenting the methods, in the following table we describe the task they were required to perform, specifying the input and the output:

INPUT :	• a single cell image • the foreground mask of the cell • the fluorescence intensity level of the image the cell belongs to
OUTPUT :	• the guessed staining pattern of the cell

The methods participating in the contest[4] are:

CHEPLYGINA—The features adopted by the system are the size and perimeter of the foreground, variance, and covariance of the intensity values, and the histograms of the red and green channels, each with 100 bins. The classifier is k-NN, where k is set using a leave-one-out procedure on the training set.

DI CATALDO—The system preprocesses the image by creating rotated versions at multiples of 45°, rescaling to 64×64 pixels, and normalizing the contrast and the intensity. It uses texture features as gray level co-occurrence matrix (GLCM) [27] and discrete cosine transform (DCT) coefficients. The most significant features are selected by the minimum redundancy maximum relevance algorithm [28] and forward selection, obtaining a set of 21 features. Classification is performed by an SVM with a radial basis function (RBF) kernel.

ERSOY—The features adopted in the proposed method [29] are the intensity values, ARST-HOG, texture features as the local binary patterns (LBP) [30], and several local features derived from the eigenvalues of the Hessian matrix. ARST-HOG are an extension proposed by the authors to the histograms of oriented gradients (HOG) [31], which combines the description power of HOG with their earlier work [32] on adaptive robust structure tensors (ARST) for improved orientation estimation. The classification is based on ShareBoost, a variant of AdaBoost proposed by the same authors in [33]. Since the ShareBoost algorithm is for binary classification, they apply error correcting output codes decoding with 6-bits codes.

FIASCHI—The proposed approach exploits information extracted from the green channel. The system uses filter banks comprising Gaussian gradient magnitude, eigenvalues of the structure tensor and eigenvalues of the Hessian matrix at several scales, as well as the raw values of the image in the selected channel. The data obtained by these filters over several patches of the image are divided into clusters during the training phase. Finally, a radial histogram of the cluster labels is calculated for each cell. An SVM with a χ^2 kernel is adopted as a classifier.

GHOSH—The proposed method [34] first converts the images to grayscale. Then the system relies on shape features extracted from the binary mask of the cell, such as area, eccentricity, major and minor axis length and perimeter of the region, the standard deviation, the 30th (P_{30}) and 60th (P_{60}) percentiles of the gray values, the percentile range $(P_{range} = P_{60} - P_{30})$ and the roundness of the region; texture features, such as contrast, homogeneity, correlation, and energy from the GLCM, entropy; normalized HOG. The final classification is obtained through a multiclass SVM with linear kernel.

GILBERT—The 22 used features are calculated on the green channel of the image. In particular, the author describes the pattern of the cells through features related to image intensity, size, statistics on pixel intensity, texture information obtained from Prewitt filtering, Otsu class distribution, statistics on the size of bright and dark blobs (detected with the maximally stable extreme regions technique in [35]). The classifier used is an SVM with an RBF kernel, with parameters obtained by grid search.

HASSAINE—The proposed method uses features which are derived from both soundtrack restoration [36] and writer identification [37]. Each of the RGB, $\mathrm{L}^*\mathrm{a}^*\mathrm{b}^*$ and HLS channels of every image is considered separately for feature extraction. Several morphological features are extracted from the resulting images as well as many geometrical features including number of holes of the segmented images, moments, projections, distributions, position of barycenter, number of branches in the skeleton, Fourier descriptors, tortuosities, directions, curvatures, and chain codes. All these features are combined using a logistic regression in order to perform the classification.

KASTANIOTIS—The proposed method [38] initially converts the image to grayscale. Then, a set of binary images is produced via a thresholding operation, using multiple threshold

[4] In the paper, we refer to each method using the name of the corresponding author of the software submission. The names of the teams members with their affiliation are provided in the acknowledgment section.

TABLE II
SUBDIVISION OF THE DATASET IN TRAINING AND TEST SETS CONSIDERING THE FLUORESCENCE INTENSITY AND STAINING PATTERNS. EACH TABLE ITEM REPORTS THE NUMBER OF IMAGES AND, IN PARENTHESES, THE NUMBER OF CELLS

	Training set		**Test set**		**Total**	
	Interm.	**Posit.**	**Interm.**	**Posit.**	**Interm.**	**Posit.**
Centromere	2 (*119*)	1 (*89*)	1 (*65*)	2 (*84*)	**3 (*184*)**	**3 (*173*)**
Coarse speckled	1 (*41*)	1 (*68*)	1 (*33*)	2 (*68*)	**2 (*74*)**	**3 (*136*)**
Cytoplasmic	1 (*24*)	1 (*36*)	1 (*13*)	1 (*38*)	**2 (*37*)**	**2 (*74*)**
Fine speckled	1 (*48*)	1 (*46*)	1 (*63*)	1 (*51*)	**2 (*111*)**	**2 (*97*)**
Homogeneous	1 (*47*)	2 (*103*)	1 (*61*)	1 (*119*)	**2 (*108*)**	**3 (*222*)**
Nucleolar	1 (*46*)	1 (*56*)	1 (*66*)	1 (*73*)	**2 (*112*)**	**2 (*129*)**
Total	**7 (*325*)**	**7 (*398*)**	**6 (*301*)**	**8 (*433*)**	**13 (*626*)**	**15 (*831*)**

values derived from statistical analysis of the image. On each binary image the authors perform connected components analysis in order to extract a set of morphological features. Additionally, in order to capture genuinely textural characteristics of the image, the LBP descriptor is incorporated, appropriately modified in order to extract rotation-invariant features. The computed morphological features from all the binary images, along with the LBP histogram are concatenated to form the image's feature vector. The classification is performed using a nonlinear SVM classifier with RBF kernel.

KAZANOV—Features are calculated over the original and the sharpened images. In addition, Otsu thresholding is applied to each version of the image. The same set of features is computed twice: once on the original image, and once on a sharpened version of it. In both cases, Otsu thresholding is applied to obtain a binarized image. Then several descriptors are determined over the binarized and the nonbinarized images. The features calculated on the binarized image are the number of connected components (objects), total object area, number of holes and total holes area, average distance of the object's pixels from the border, length of the longest concave arc; the features calculated on the nonbinarized image are related to the statistics of the pixels intensity inside and outside the objects area. The most significant features are chosen through forward selection. The classifier is a naive Bayes.

KOVACS—The proposed method uses a statistical approach for classification. First, an optimal chain of image transforms is determined by best-first search through a large set of chains consisting of commonly used intensity transforms (filters, morphological operators, etc.). Then, four sets of features are extracted: the set named R1 contains conventional region based descriptors such as area, range of intensities, mean/min/max/variance of the intensities covered by the cell region, etc.; the set R2 is the same as R1, but the extraction of the features is preceded by the application of the best transform chain; the set H1 contains the normalized 8-bin histogram of the intensities of the cells. Finally, the set H2 contains the normalized 8-bin histogram extracted after the application of the best transform chain. For each set of features (R1, R2, H1, H2), five classifiers are trained by applying a grid search for the best parameters. The adopted classifiers are ϵ-SVM, ν-SVM, k-NN, naive Bayes, and averaged one-dependence estimator. The 20 trained models are combined using the majority voting decision rule.

KUAN—The method [39] uses the following four texture descriptors: a rotation invariant form of LBP with multi-scale analysis, DCT, the mean values and standard variances of 2-D Gabor wavelets, and some global appearance based statistical features. A multiclass posterior probability SVM is utilized on each of the four feature sets. The four SVMs are then merged into an integrated classifier. Furthermore, the AdaBoost.M1 algorithm is modified to embed the integrated classifier in the boosting procedure.

MALON—Only the green channel is analyzed. For each cell a 100×100 pixels area centered on the largest component of the cell mask is taken. The resulting frame is normalized by mapping the first and 99th percentile values to 0 and 1. Each cell image is classified using a convolutional neural network (CNN) starting from the raw pixel values of the normalized frame, applying tricks such as absolute value rectification, subtractive spatial normalization, and max pooling.

MAREE—The proposed method [40] extracts from each cell thousands of square subwindows at random positions, using random sizes varying from 15% to 45% of the original image size. Each subwindow is then resized using bilinear interpolation to a fixed-size patch of 16×16 pixels, and encoded in normalized RGB color space. The subwindows are indexed by extremely randomized trees and the subwindow frequencies at terminal nodes are then used as features by a linear SVM classifier.

MATEOS-GARCIA—The system resizes all the images to 100×100 pixels. The value of the pixels are used as basic features. Then, the authors employ a combination of correlation feature selection (CFS) method and of the genetic algorithms for feature selection. The final classification is obtained by C4.5 boosted by AdaBoost whose parameters were optimized through genetic algorithms.

NANNI—The proposed method is based on multithreshold nonbinary coded texture descriptors, derived from LBP, combined with the Haralick features (GLCM). The chosen nonbinary coded texture descriptors are local quinary patterns and local phase quantization with ternary coding. The authors chose different thresholds for the ternary coding (or a set of different pairs of thresholds for the quinary coding), in order to extract a different features set according to each threshold (or pair of thresholds), using each features set to train a different SVM and fusing all these results according to the sum fusion rule.

NOSAKA—The algorithm uses only the green channel, which is then filtered by a Gaussian function for noise reduction. The adopted descriptor is CoALBP, an extension of LBP defined in [41], which can describe complex textures by observing not only each local LBP but also the spatial relations among adjacent LBP. The classifier is a linear SVM trained with a learning set including the various rotated patterns of the original images.

PERSSON—The system uses a manually designed decision tree with four decision nodes. Each node consists of a nearest neighbor classifier that adopts a weighted Euclidean distance. The weights have been defined ad hoc by the author for each decision node. The system uses nine scalar features related to texture, size, and intensity distribution.

REZVANI—Eight sets of features are extracted after performing background subtraction and converting the images to grayscale: geometrical features, pixel correlation features, connected components, statistical texture features, consecutive Fourier transformations, Radon transform of the histograms, first-order histogram, pixel intensities of a rescaled image. Of these, the first-order histogram proved best both for its performance and for its low complexity. Then PCA was used to select the best features. Finally, an expert system was designed: it was made up of five RBF networks each designed with optimized parameters and trained on overlapping sections of the dataset.

RUUSUVUORI—There are 470 extracted features overall, including quantities related to size and shape (e.g., area, perimeter, major axis length), statistical quantities (e.g., grayscale variance and mean), LBP, HOG, and the nominal image intensity (positive or intermediate). After extraction, each feature (except the nominal intensity feature) is normalized to have zero mean and unit variance. For classification, the authors used an SVM with an RBF kernel.

SHEN—The histograms of R, G, and B channels are calculated independently and then concatenated into a feature vector. The classifier is an SVM with an RBF kernel.

SNELL—The system [42] converts the image to grayscale. Shape information for separating the cytoplasmic pattern is included in the feature vector alongside texture, which is represented by DCT coefficients and difference statistics computed at different scales. The binary intensity label supplied for each image is also included. Classification is accomplished through a multiclass SVM with an RBF kernel. The authors also introduce a second stage expert classifier to deal with the case of the images of intermediate intensity from homogeneous or fine speckled classes. This stage uses a subset of features found to have the best discriminative power for this particular task and a binary SVM. Full details can be found in [42].

STOKLASA—Only the green channel is used. After a preprocessing stage that implements denoising and contrast enhancement, the method removes impulse noise, i.e., small groups of pixels with remarkably high intensity, by analyzing the histogram. Then, the method calculates several global image descriptors, namely LBP, statistics of the GLCM, color structure (from the MPEG-7 descriptors) [43], granulometry-based [44] and surface descriptors (cumulative derivative of neighboring pixels), SIFT local descriptor [45]. For each set of descriptors, a separate features space with a suitable metric is created. The classifier is a k-NN (with $k = 16$) using a custom aggregated distance function, which combines the descriptors selected from the list above with different weights.

STRANDMARK—This algorithm [46] projects the RGB color of each pixel onto the principal component of all pixel colors in the training set. Then it preprocesses the image by removing very bright pixels. The image is thresholded at 20 intensities equally spaced from its minimum to its maximum intensity. The used features take into account geometrical aspects, statistics about pixel intensity, texture (modeled through GLCM). Such features are calculated on the original and smoothed versions of the images after thresholding. The classifier is a random decision forest.

THIBAULT—The approach [47] is based on two different families of descriptors and three different classifiers. The first descriptor is the gray level size zone matrix (GLSZM), which provides a statistical representation by the estimation of a bivariate conditional probability density function of the image distribution values. The second descriptor is the pattern spectrum, a study of the size distribution of the objects of an image [44]. Classification is accomplished using a one per class classification approach, where the base classifiers are a logistic regression, a random forest and a neural network, whose outputs are combined using a weighted average.

WAFA—The proposed method [48] adopts bio-inspired features relying on the distribution of contrast information, modeled through a pyramid of centered bidimensional differences of Gaussians. Then, a supervised learning process selects prototypical samples which are used in a leveraged k-NN framework to predict the class of unlabeled cells.

WANG—The proposed solution is based on an improvement of the standard bag of words. Images are first normalized with respect to the intensity. Then, the algorithm adopts a scheme that automatically learns the discriminative descriptors from the raw image pixels. Final classification is performed by a linear SVM.

WILIEM—The proposed approach [49] divides each HEp-2 cell image into two regions: inner and outer. The inner region includes only the cell content, while the outer region contains cell edges. The regions are represented using a variant of probabilistic histograms of visual words [50]. In particular, each region is represented by two histograms, derived from the two scales at which the image is analyzed, so that each image is represented by four histograms. Two cell images are compared by computing an L_1-based distance between the corresponding histograms. The multi cue kernel [51] is used to fuse the four distances. Finally, a nearest neighbor classifier is applied to obtain the classification label.

XIANGFEI—The proposed system adopts Varmas' MR8 method [52] to extract statistical intensity features. Before calculating filter responses, the local regions where the filter is convolved are normalized. A global texton dictionary is trained using K-means clustering. Then each image is represented by the frequency histogram of the textons. The adopted classifier is a k-NN, with χ^2 distance.

A summary of all the considered approaches is reported in Table III, where we analyze the main design choices made by the contest participants. In particular, we consider the projection of the pixel onto specific color channels, the adopted preprocessing procedures, the descriptors, the features selection procedure, the number of features, and the classification architectures.

V. Experimental Results

In this section, we evaluate the recognition accuracy at the cell level of each considered method, as required by the contest rules. In addition, in order to provide the reader with a rough estimate of the achievable performance at the image level we

TABLE III

MAIN DESIGN CHOICES FOR EACH METHOD THAT PARTICIPATED IN THE CONTEST: THE PROJECTION OF THE PIXEL ONTO SPECIFIC COLOR CHANNELS, THE ADOPTED PREPROCESSING PROCEDURES, THE ADOPTION OF A MULTI-RESOLUTION ANALYSIS, THE DESCRIPTORS, THE FEATURES SELECTION PROCEDURE, THE SIZE OF THE FEATURE VECTOR, THE CLASSIFICATION ARCHITECTURES. EMPTY CELLS ARE DUE TO THE FACT THAT THE CONTEST PARTICIPANTS WERE NOT REQUIRED TO DISCLOSE ALL THE DETAILS OF THEIR METHOD, AND CHOSE NOT TO REVEAL THE CORRESPONDING INFORMATION

Submission	Color channel	Preprocessing	Multiresolution analysis	Descriptors	Features selection	# of features	Classifier
Cheplygina	Green-Red			Intensity, size		205	k-NN
Di Cataldo		Resizing 64×64, contrast and intensity normalization		GLCM, DCT	minimum Redundancy Maximum Relevance (mRMR)	21	SVM
Ersoy				Intensity, ARST-HOG, LPB, eigenvalues of the Hessian		130	ShareBoost with error correcting code
Fiaschi	Green			Gaussian gradient, eigenvalues of the tensor of the structure and of the Hessian (Bag of words)		42	SVM with kernel χ^2
Ghosh	Grayscale			Area, Eccentricity, GLCM, HOG		1346	Multiclass SVM with linear kernel
Gilbert	Green			Image intensity, pixels intensity, Prewitt filtering, number and size of blobs		22	SVM with RBF kernel
Hassaine	RGB, HLS and L*a*b*			Morphological, FFT			Logistic regression
Kastaniotis	Grayscale			Morphological, LBP rotation invariant		86	SVM with RBF kernel
Kazanov		Sharpening		Morphological, pixels intensity	Forward selection	11	Naive Bayes
Kovacs				Pixels intensity, size (area)			Combination of ϵ-SVM, ν-SVM, k-NN, Naive Bayes, Averaged one-dependence estimator
Kuan			Yes	LBP rot. inv., DCT, GLCM, Wavelet		180	SVM + AdaBoost.M1
Malon	Green	Contrast stretching		Pixels intensity		160	Convolutional neural networks
Maree	Normalized RGB		Yes	Raw image pixels (random trees indexing)		352842	SVM
Mateos-Garcia		Resizing 100×100		Pixels intensity	Correlation based Feature Selection (CFS) + Genetic Algorithm (GA)	246	C4.5+AdaBoost
Nanni				LQP, LPQ, GLCM		7972	Combination of SVM
Nosaka	Green	Gaussian smoothing		CoALBP		5120	Linear SVM
Persson				Intensity, Size, Texture		9	Decision tree with NN nodes
Rezvani	Grayscale			Eccentricity, compactness, circularity, pixels intensity, raw pixels values, FFT	PCA	25	Combination of five RBF networks
Ruusuvuori				Intensity, pixels intensity (mean and variance), LBP, HOG, size (area, perimeter, major axis length)		470	SVM with RBF kernel
Shen				Histograms of R,G,B channels		765	SVM with RBF kernel
Snell	Grayscale		Yes	Shape, DCT, difference statistics		40	SVM with RBF kernel
Stoklasa	Green	Impulse noise removal, contrast enhancement		LBP, GLCM, color structure, granularity, surface, SIFT		628	k-NN
Strandmark	Projection onto the principal component of the training set	Impulse noise removal	Two scales (original image and Gaussian smoothed)	Geometrical, pixels intensity, GLCM		966	random decision forest
Thibault				Pattern Spectrum, GLSZM			Combination of logistic regression, random forest, neural network
Wafa			Yes	Pyramid of DoG			Leveraged k-NN
Wang		Intensity normalization		Raw image pixels (Bag of words)			Linear SVM
Wiliem			Two scales	Probabilistic histograms of visual words		4096	NN, Multi-Cue kernel
Xiangfei		Local normalization		Frequency histograms of textons		85	k-NN with χ^2 distance

also present the results obtained by applying a simple plurality voting scheme [53] to each classifier result; while this analysis should only be considered as preliminary, and does not represent the best attainable result at the image level, it is significant because it shows that, even with the limitations of this scheme, some of the classifiers were able to reach an accuracy comparable to a human expert.

The adopted experimental protocol was the following.

1) Each participant received the training set with the original images of the cells already segmented by specialists. In particular, for each cell we provided the bounding box and the foreground mask (see Fig. 2). The cells were provided along with the information on the intensity pattern of the image they belong to.
2) The participants used the training set to tune their HEp-2 cells classification systems and then they released the executable for independent evaluation on the test set.
3) We ran all the submitted executables on the test set, collecting the results that are reported in this section.

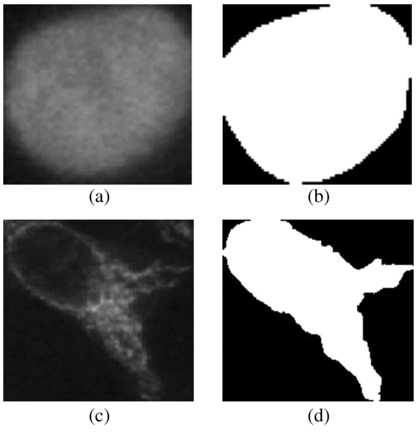
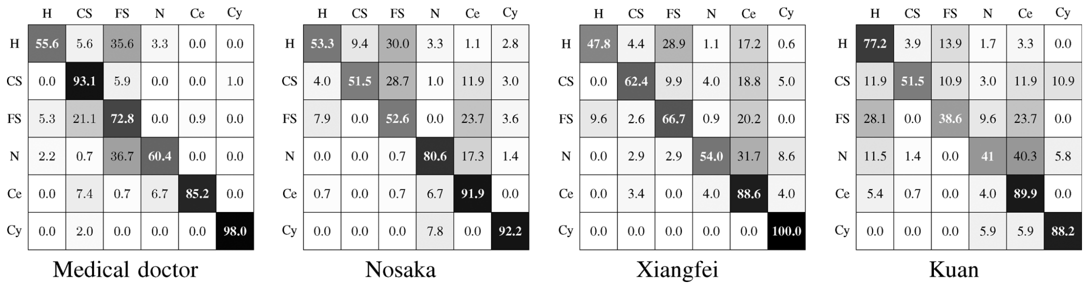


Fig. 2. Examples of the two images associated to a cell of the training set. (a) and (c) The cell bounding box cropped from the original image. (b) and (d) The binary mask used to discriminate the pixels inside or outside the cell.

In order to establish a level of accuracy to be used as a reference when analyzing the results, a medical doctor manually classified each cell. This doctor was a specialist in Immunology with 12 years experience, was different from the ones who provided the ground truth, and worked in a different research center. In particular, this specialist was asked to annotate each cell by looking only at its bounding box extracted from the whole image (as shown in Fig. 2), without observing the other surrounding cells. We were interested in assessing the performance of the specialist working in exactly the same conditions as the automatic methods. The specialist achieved a recognition accuracy equal to 73.3%. It is worth noting that the high level of discordance between this doctor and the other ones is due to the fact that they operated under different conditions, as the first experts were also allowed to exploit information on the whole image to which the cells belong.

When analyzing the performance of the considered methods at the cell level, we focused on two issues: we first tried to find some relations between the design choices of each approach and the corresponding performance, then we examined the behavior



Medical doctor

	H	CS	FS	N	Ce	Cy
H	55.6	5.6	35.6	3.3	0.0	0.0
CS	0.0	93.1	5.9	0.0	0.0	1.0
FS	5.3	21.1	72.8	0.0	0.9	0.0
N	2.2	0.7	36.7	60.4	0.0	0.0
Ce	0.0	7.4	0.7	6.7	85.2	0.0
Cy	0.0	2.0	0.0	0.0	0.0	98.0

Nosaka

	H	CS	FS	N	Ce	Cy
H	53.3	9.4	30.0	3.3	1.1	2.8
CS	4.0	51.5	28.7	1.0	11.9	3.0
FS	7.9	0.0	52.6	0.0	23.7	3.6
N	0.0	0.0	0.7	80.6	17.3	1.4
Ce	0.7	0.0	0.7	6.7	91.9	0.0
Cy	0.0	0.0	0.0	7.8	0.0	92.2

Xiangfei

	H	CS	FS	N	Ce	Cy
H	47.8	4.4	28.9	1.1	17.2	0.6
CS	0.0	62.4	9.9	4.0	18.8	5.0
FS	9.6	2.6	66.7	0.9	20.2	0.0
N	0.0	2.9	2.9	54.0	31.7	8.6
Ce	0.0	3.4	0.0	4.0	88.6	4.0
Cy	0.0	0.0	0.0	0.0	0.0	100.0

Kuan

	H	CS	FS	N	Ce	Cy
H	77.2	3.9	13.9	1.7	3.3	0.0
CS	11.9	51.5	10.9	3.0	11.9	10.9
FS	28.1	0.0	38.6	9.6	23.7	0.0
N	11.5	1.4	0.0	41	40.3	5.8
Ce	5.4	0.7	0.0	4.0	89.9	0.0
Cy	0.0	0.0	0.0	5.9	5.9	88.2

Fig. 3. Confusion matrices of the medical doctor asked to perform staining pattern recognition at the cell level and of the three algorithms that achieved the highest accuracy at the cell level (Nosaka, Xiangfei, and Kuan, respectively). The label of the row/column is the true/guessed class name, with the following meanings: H = homogeneous, CS = coarse speckled, FS= fine speckled, N = nucleolar, Ce = centromere, Cy = cytoplasmic.

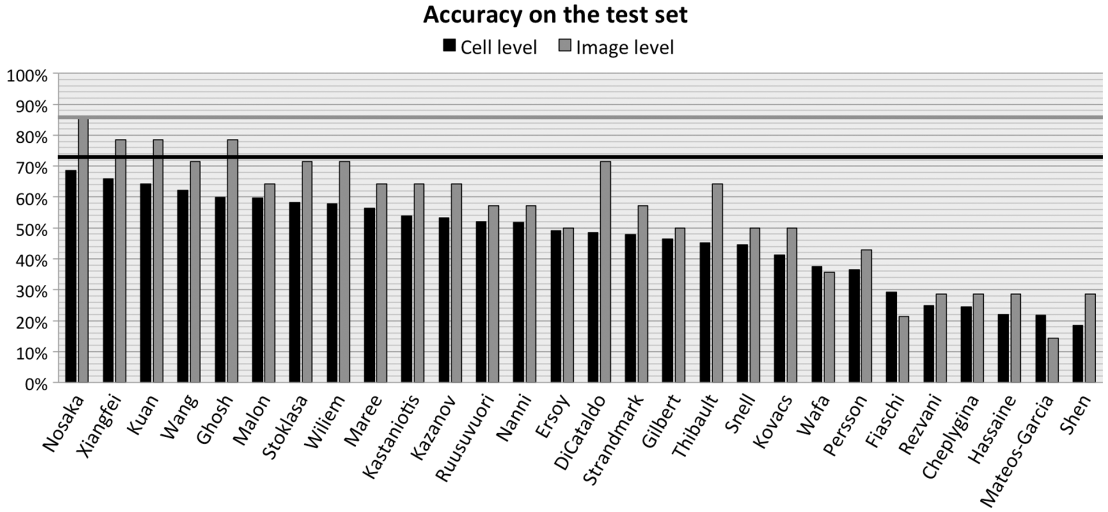


Fig. 4. Recognition accuracy obtained by the considered methods over the test set at the cell and the image levels. Two horizontal lines represent the accuracy obtained by a specialist at the cell level (black line) and at the image level (gray line).

of the best performing methods with respect to the different cell classes.

Fig. 4 shows the staining pattern recognition accuracy at the cell level achieved by all the considered methods. We notice that the accuracy values of the considered methods are almost uniformly distributed between the minimum and the maximum values. Fig. 4 also shows the accuracy at image level, obtained by applying the plurality voting to the results of each classifier in order to determine the staining pattern of the whole image. In Fig. 5, we provide an indication of the agreement level reached by the plurality rule for each image of the test set.

It is useful to look in greater detail at the distribution of the errors over the six staining patterns. Unfortunately, for reasons of space here we cannot present this information for all 28 methods. So we have decided to make detailed data available on the web site of the contest [26]; here we will show only the three methods achieving the highest performance: Nosaka, Xiangfei, and Kuan. Fig. 3 graphically shows the confusion matrices for these methods.

As a final step of our study, we consider the dependence of the classification performance at the cell level on the fluorescence intensity. Fig. 6 shows the accuracy that each method obtained on the intermediate and on the positive cells of the test set. For the sake of comparison, in the figure we once again show the accuracy over the whole test set.

A. Analysis of the Results

The accuracy values at the cell level (see Fig. 4) range over a fairly wide interval, starting from about 20% and reaching the maximum at 68.7%. About half of the methods are able to correctly recognize more than 50% of the samples, while no method was able to perform better than the medical doctor who manually annotated each cell. However, the absolute difference between the accuracy of the best method and that of the specialist is very small, as it is only 4.6%.

Focusing on the tail of the plot we can observe that it is populated by methods which do not use texture-based features; on the contrary, the large majority of the remaining methods adopt features that model the texture (through LBP and its derivations, GLCM, etc.).

Analogously, a consideration on the features can also be made for the best performing methods. Indeed, it can be observed that most of these approaches adopt local descriptors: examples are the Nosaka's method, which (through CoALBP) also models the spatial relations among the local binary patterns; the Xiangfei's one, which creates a dictionary of textons used to model the textures in local regions; the Wang's method, which uses the raw image pixels according to the bag of words approach; or the Malon's method, which uses the convolutional neural networks, which also operate by analyzing small patches extracted from the image.

	21 (H)	22 (H)	10 (CS)	12 (CS)	17 (CS)	15 (FS)	23 (FS)	4 (N)	24 (N)	13 (Ce)	16 (Ce)	19 (Ce)	27 (Cy)	28 (Cy)
Medical doctor	77.0 (FS)	83.2	97.0	95.9	78.9	63.5	84.3	71.2 (FS)	89.0	100.0	97.4	67.7	97.4	100.0
Nosaka	55.7	52.1	54.5	69.4	94.7 (FS)	69.8 (Ce)	90.2	62.1	97.3	93.5	100.0	86.2	100.0	69.2
Xiangfei	37.7 (Ce)	67.2	75.8	69.4	52.6 (FS)	49.2	88.2	62.1 (Ce)	89.0	80.4	94.7	90.8	100.0	100.0
Kuan	68.9	81.5	45.5	69.4	52.6 (FS)	42.9 (Ce)	76.5	84.8 (Ce)	78.1	80.4	100.0	90.8	100.0	53.8

Fig. 5. Each entry of the table represents the percentage of cells attributed to the staining pattern selected by the plurality rule for each image. Entries with the gray background correspond to misclassification of the plurality rule. For such cases the attributed class is shown in parentheses. Classes are indicated with the same abbreviations used in Fig. 3.

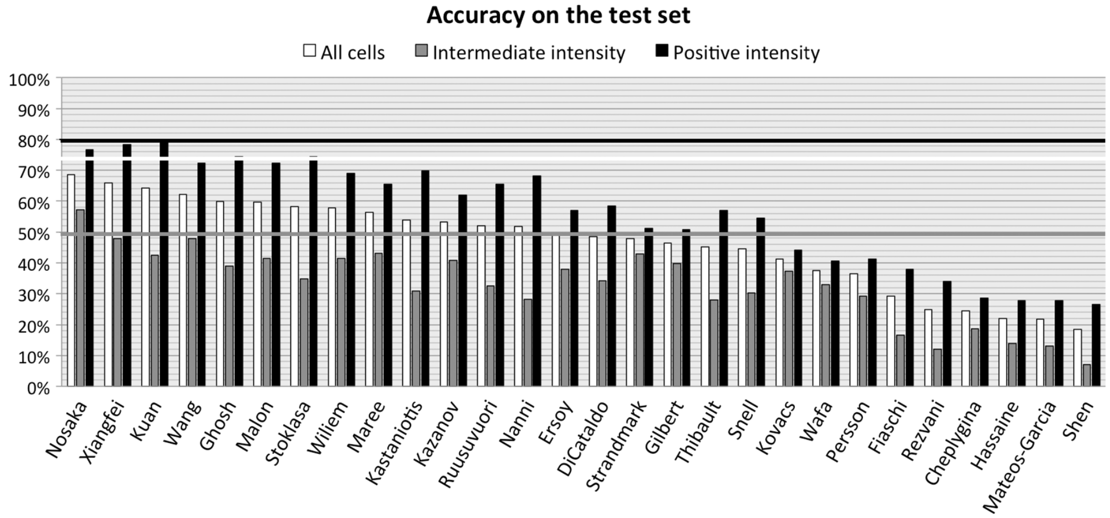


Fig. 6. Recognition accuracy obtained by the considered methods over the test set at the cell level for the intermediate and the positive fluorescence intensities. Three horizontal lines represent the accuracy obtained by a specialist on the whole test set (white line), on the intermediate images (gray line), and on the positive ones (black line).

While the use of local, texture-based descriptors seems a key factor for obtaining a good performance, we notice that the adoption of a specific classification paradigm does not significantly influence the performance.

We note that the centromere and the cytoplasmic patterns are detected much more accurately than the remaining four classes. For the cytoplasmic pattern, whose accuracy ranges between 88.2%–100.0%, the result is not very surprising as cells belonging to this class can be easily distinguished from the others by using simple features related to shape. The centromere pattern is also recognized quite reliably (88.6%–91.9%); moreover, we note that the three considered techniques tend to confuse the remaining classes with the centromere pattern. This phenomenon is probably due to the partially unbalanced distribution of the samples in the training set, where the centromere pattern is more represented than the others. It is also interesting to note that most of the centromere misclassified samples were attributed to the nucleolar class and vice versa. This behavior might be explained by considering that the nucleolar and the centromere patterns are both characterized by a body with highly fluorescent dots surrounded by an area with a homogeneous fluorescence intensity. The main difference between the two patterns lies in the size of the dots: centromere cells have many small dots, while nucleolar only have very few large dots.

The recognition rates over the remaining four types of pattern is around or below 50%, with only few exceptions: the nucleolar pattern that is classified by the Nosaka's method with an accuracy of 80.6%; the two speckled classes that the Xiangfei's method recognizes in 62.4% and 66.7% of the cases; the homogeneous pattern that is correctly classified by the Kuan's method in 77.2% cases.

As regards the homogeneous and the two speckled classes, we notice that most misclassification errors remain confined within these three classes, with the exception of those errors due to the training set bias toward the centromere pattern. This result must be expected: in fact, the fine speckled and the homogeneous classes appear as increasingly smoothed versions of the coarse speckled pattern. In order to improve the recognition ability between staining patterns with similar interphase cells stainings we should look at mitotic cells. For instance, consider the homogeneous and fine speckled staining patterns: the former has positive mitotic cells,[5] whereas the latter has negative mitotic cells.[6]

A final, interesting consideration on the misclassification matrices can be derived from a comparison of the three best methods and the medical doctor. While the overall accuracy is similar, the human expert tends to make different kinds of errors. For example, the doctor performs quite well at classifying the coarse speckled pattern (93% accuracy), while the three best methods only achieve a performance between 51%

[5] Positive mitosis is characterized by a cell body which is weakly fluorescent or nonfluorescent, while the chromosomes mass is fluorescent.

[6] Negative mitosis is characterized by a fluorescent cell body, while the collapsed chromosomes mass located in the middle part of the cell does not exhibit a fluorescent pattern or has a weak fluorescence.

and 62%. On the other hand, the doctor displays a significant confusion between the nucleolar and the fine speckled patterns, which is instead fairly infrequent in the automated methods. This difference can be explained by the fact that the automated methods base their decision mainly on quantitative measures related to the texture, while the doctor tries to recognize shapes and structures within the image. The results show that there is some complementarity between the two approaches, leaving room for an improved method that attempts to combine both of them.

Turning our attention to the accuracy at the image level, we note that the Nosaka's method is again confirmed as the best performing method and is able to attribute the correct class to 12 out of 14 images. In particular, in one case the method misclassifies an intermediate fine speckled image (the image in Table I with $ID = 15$) as a centromere; in the second case it attributes a positive coarse speckled ($ID = 17$) to the fine speckled class.[7] An important consideration derives from the observation that at the image level the Nosaka's method performs as well as the medical doctor. The latter is able to recognize 12 images out of 14, but it is worth noting that the two errors are different from those of the algorithm: the doctor attributed the images with $ID = 4$ and with $ID = 21$ to the fine speckled class. It is also interesting to note that the Nosaka's method is the only one able to recognize the pattern of the image with $ID = 4$, which is a challenging image with a nucleolar pattern over an intermediate fluorescence intensity, that makes the pattern very weak. Even the doctor was unable to recognize the staining pattern of the image, by looking at each cell separately. Also the image with $ID = 17$ is particularly interesting as only the human expert was able to correctly classify it.

The Xiangfei's and Kuan's methods correctly classify 11 images out of 14. They both misclassify the image with $ID = 4$ attributing it to the centromere pattern, and the image with $ID = 17$, classified as fine speckled. Moreover, Xiangfei misclassifies the image with $ID = 21$ as centromere, which is the same class erroneously attributed to the image with $ID = 15$ by the Kuan's method. Note that these two methods have the same performance at image level even though Kuan has a slightly lower accuracy at the cell level. Fig. 5 shows that this is due to the fact that, roughly, the errors are uniformly distributed among the different images, and thus the plurality rule is able to guess the right class in the same number of instances.

It is worth noting that, while most methods improve their performance at the image level with respect to the cell level, the Wafa's, Fiaschi's, and Mateos–Garcia's methods do the opposite. This is explained by the combination of two factors: first, those methods have a recognition rate below 50%, and thus most cells are misclassified; second, the misclassified cells are often assigned to the same wrong class, and not distributed across several classes, and thus the wrong class is able to reach the majority over the whole image. The same kind of consideration explains why the Ghosh's method outperforms the Wang's method at the image level, even though it is slightly worse at the cell level.

Finally, focusing on the plot in Fig. 6 that shows the recognition accuracy at the cell level separately for the intermediate and the positive fluorescence intensities obtained by the considered methods over the test set, we note that, as might be expected, most classification errors are made on the intermediate images. For all the methods, including the human expert, the recognition performance over the intermediate dataset is much lower than that obtained on the positive one. It is interesting to observe that the Kuan's and Xiangfei's methods provide a performance similar to that of the specialist, while on the intermediate images the Nosaka's method performs better than the doctor. As a final remark, focusing on the best three performing methods, we note that the Nosaka's method is much more robust with respect to the fluorescence intensity variations.

VI. Conclusion

In this paper, we have reported the results of the first international contest on HEp-2 cells classification hosted by the 21st edition of the International Conference on Pattern Recognition (ICPR 2012). We have analyzed 28 methods on the same private dataset that had not been released to the contest participants.

The comparative analysis of the considered methods, integrated with the performance achieved by a specialist doctor invited to perform the same classification task, highlights some important issues.

- The problem of automatic HEp-2 cells classification is somewhat complex, which is demonstrated by the fact that even a human expert recognizes on average only three cells out of four, thus performing only slightly better than the best automatic approach.
- The fluorescence intensity significantly affects the recognition performance of the classifiers, so that, as could be intuitively expected, the positive cells are recognized much more reliably than the intermediate ones.
- The key design choice for obtaining a high performance in this task is the adoption of a local description that explicitly models the textures, while the use of specific classification paradigms does not significantly influence the recognition rates; if we consider only the approaches for which it is known what kind of preprocessing steps has been adopted, the choice of a particular preprocessing does not appear to alter significantly the results.
- The experiments confirm that the obtained recognition rate for each of the considered patterns is, as expected, highly correlated with the degree of visual similarity to other patterns. In fact, cytoplasmic cells appear very different from the cells belonging to the remaining patterns and they are easily recognized. The nucleolar and the centromere patterns share a similar visual appearance and are often confused with each other. The same observation can be made for the remaining three patterns (homogeneous, fine speckled, and coarse speckled).
- The complementarity of the errors made at the image level by the best method and the medical doctor confirms the initial statement about the importance of automatic methods in supporting the human expert.

[7] The mentioned images are not shown in the paper because their low contrast would make them almost illegible; the interested reader is encouraged to obtain the image files using the web page given in [25].

We also notice that the skew of the dataset toward the centromere class had some impact during the training of the classifiers: consequently, in several cases the recognition systems were biased toward this class.

The competition had the merit of raising considerable interest within the pattern recognition researchers community. However, as any first attempt in any field, it is not immune to some limitations that we want to acknowledge here. A first problem is the limited number of images with respect to the number of classes: as shown in Table I each class is represented by four to six images which are too few to draw definitive conclusions on classification at the image level.

Another issue that has to be taken into account in future is the fact that all images used in this work were acquired using the same equipment. Consequently, the methods were only tested for these specific image types and might not perform as well on new images acquired with different settings and/or equipment. Future initiatives should take this issue into account by adopting a larger dataset including images obtained by acquisition devices from different diagnostic centers. This would make it possible to test the robustness of the method against this source of variability.

A third important issue regards the choice according to which the submitted methods were required to classify each cell without using any information on the surrounding cells. This choice, motivated by the narrowness of the dataset, had a major impact on the design of the submitted methods and on their performance, as they were required to work under more difficult conditions.

In the future, we will address three main issues. First, we intend to test the best approaches on a larger dataset. Second, we will explore the possibility of boosting performance by combining the best methods that have already shown to have quite complementary behaviors, at least on some classes. Third, we will enrich the dataset with mitotic cells to help to distinguish between patterns with similar interphase cells.

Acknowledgment

The authors would like to thank all the teams participating in the first edition of the HEp-2 Cells Classification contest. The team members and their relative affiliations are reported in the following: **CHEPLYGINA**: V. Cheplygina, A. Ibba, D. M. J. Tax, *Delft University of Technology, The Netherlands*. **DI CATALDO**: S. Di Cataldo, A. Bottino, E. Ficarra, E. Macii, *Politecnico di Torino, Italy*. **ERSOY**: I. Ersoy[1], F. Bunyak[1], J. Peng[2], K. Palaniappan[1], [1]*University of Missouri-Columbia, USA*, [2]*Montclair State University, USA*. **FIASCHI**: L. Fiaschi, F. A. Hamprecht, *IWR/HCI Heidelberg, Germany*. **GHOSH**: S. Ghosh, *State University of New York, USA*. **GILBERT**: G. Lim Yong San, *National University of Singapore, Singapore*. **HASSAINE**: A. Hassaïne, *Qatar University, Qatar*. **KASTANIOTIS**: I. Theodorakopoulos, D. Kastaniotis, *University of Patras, Greece*. **KAZANOV**: G. V. Ponomarev, V. L. Arlazarov, M. D. Kazanov, *Institute for Information Transmission Problems, Russia*. **KOVACS**: G. Kovacs, *University of Debrecen, Hungary*. **KUAN**: K. Li[1], J. Yin[1], Z. Lu[2], X. Kong[2], R. Zhang[2], W. Liu[2], [1]*National University of Defense Technology, China*, [2]*City University of Hong Kong, China*. **MALON**: P. F. Laquerre, C. Malon, *NEC Laboratories America, USA*. **MAREE**: R. Marée, *University of Liège, Belgium*. **MATEOS-GARCIA**: D. Mateos-García, J. García-Gutiérrez, J. C. Riquelme-Santos, *University of Seville, Spain*. **NANNI**: M. Paci[1], L. Nanni[2], S. Severi[1], [1]*University of Bologna, Italy*, [2]*University of Padua, Italy*. **NOSAKA**: R. Nosaka, K. Fukui, *University of Tsukuba, Japan*. **PERSSON**: B. Persson, *FOSS Analytical AB, Denmark*. **REZVANI**: M. Naghshnejad, Z. Rezvani, *Shahid Beheshti University, Iran*. **RUUSUVUORI**: T. Manninen, H. Huttunen, P. Ruusuvuori, *Tampere University of Technology, Finland*. **SHEN**: L. Shen, J. Lin, *Shenzhen University, China*. **SNELL**: V. Snell, W. Christmas, J. Kittler, *University of Surrey, UK*. **STOKLASA**: R. Stoklasa, T. Majtner, D. Svoboda, M. Batko, *Masaryk University, Czech Republic*. **STRANDMARK**: P. Strandmark, J. Ulen, F. Kahl, *Lund University, Sweden*. **THIBAULT**: G. Thibault, J. Angulo, *MINES-ParisTech, France*. **WAFA**: W. Bel Haj Ali, M. Barlaud, *University of Nice Sophia Antipolis, France*. **WANG**: L. Liu[1], L. Wang[2], [1]*Australian National University, Australia*, [2]*University of Wollongong, Australia*. **WILIEM**: A. Wiliem[1], Y. Wong[1,3], C. Sanderson[1,3], P. Hobson[2], S. Chen[1,3], B. C. Lovell[1,3], [1]*NICTA, Australia*, [2]*Sullivan Nicolaides Pathology, Australia*, [3]*University of Queensland, Australia*. **XIANGFEI**: X. Kong, K. Li, Z. Lu, L. Wenyin, *City University of Hong Kong, China*.

References

[1] A. Jalalian, S. B. Mashohor, H. R. Mahmud, M. I. B. Saripan, A. R. B. Ramli, and B. Karasfi, "Computer-aided detection/diagnosis of breast cancer in mammography and ultrasound: A review," *Clin. Imag.* 2012.

[2] "Approaches for automated detection and classification of masses in mammograms," *Pattern Recognit.*, vol. 39, no. 4, pp. 646–668, 2006.

[3] J. Keller, P. Gader, S. Sohn, and C. Caldwell, "Soft counting networks for bone marrow differentials," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, 2001, vol. 5, pp. 3425–3428.

[4] D. Milenovic, L. Stoiljkovic, and N. Stojanovic, "Cell classification for diagnostic of reactive histocytic hyperplasia using neural networks," in *Proc. 9th Mediterranean Electrotechn. Conf.*, May 1998, vol. 2, pp. 1466–1470.

[5] T. Nattkemper, H. Ritter, and W. Schubert, "A neural classifier enabling high-throughput topological analysis of lymphocytes in tissue sections," *IEEE Trans. Inf. Technol. Biomed.*, vol. 5, no. 2, pp. 138–149, Jun. 2001.

[6] W. Lin, J. Xiao, and E. Micheli-Tzanakou, "A computational intelligence system for cell classification," in *Proc. IEEE Int. Conf. Inf. Technol. Appl. Biomed.*, May 1998, pp. 105–109.

[7] N. Bizzaro, R. Tozzoli, E. Tonutti, A. Piazza, F. Manoni, A. Ghirardello, D. Bassetti, D. Villalta, M. Pradella, and P. Rizzotti, "Variability between methods to determine ANA, anti-dsDNA and anti-ENA autoantibodies: A collaborative study with the biomedical industry," *J. Immunol. Methods*, vol. 219, pp. 99–107, Aug. 1998.

[8] L. Song, E. J. Hennink, I. T. Young, and H. J. Tanke, "Photobleaching kinetics of fluorescein in quantitative fluorescence microscopy," *Biophys. J.*, vol. 68, no. 6, pp. 2588–2600, June 1995.

[9] R. Hiemann, N. Hilger, U. Sack, and M. Weigert, "Objective quality evaluation of fluorescence images to optimize automatic image acquisition," *Cytometry Part A*, vol. 69, pp. 182–184, 2006.

[10] Y. L. Huang, Y. L. Jao, T. Y. Hsieh, and C. W. Chung, "Adaptive automatic segmentation of HEp-2 cells in indirect immunofluorescence images," in *Proc. IEEE Int. Conf. Sensor Netw., Ubiquitous Trustworthy Comput.*, 2008, pp. 418–422.

[11] Y. L. Huang, C. W. Chung, T. Y. Hsieh, and Y. L. Jao, "Outline detection for the HEp-2 cells in indirect immunofluorescence images using watershed segmentation," in *Proc. IEEE Int. Conf. Sensor Netw., Ubiquitous Trustworthy Comput.*, 2008, pp. 423–427.

[12] P. Perner, H. Perner, and B. Muller, "Mining knowledge for HEp-2 cell image classification," *J. Artif. Intell. Med.*, vol. 26, pp. 161–173, 2002.

[13] G. Percannella, P. Soda, and M. Vento, "A classification-based approach to segment HEp-2 cells," in *Computer Based Medical Systems*. Los Alamitos, CA: IEEE Comput. Soc., 2012, pp. 1–5.

[14] P. Foggia, G. Percannella, P. Soda, and M. Vento, "Early experiences in mitotic cells recognition on HEp-2 slides," in *Proc. 23rd IEEE Int. Symp. Computer-Based Med. Syst.*, 2010, pp. 38–43.

[15] P. Soda, G. Iannello, and M. Vento, "A multiple experts system for classifying fluorescence intensity in antinuclear autoantibodies analysis," *Pattern Anal. Appl.*, vol. 12, no. 3, pp. 215–226, Sep. 2009.

[16] U. Sack, S. Knoechner, H. Warschkau, U. Pigla, F. Emmerich, and M. Kamprad, "Computer-assisted classification of HEp-2 immunofluorescence patterns in autoimmune diagnostics," *Autoimmun. Rev.*, vol. 2, pp. 298–304, 2003.

[17] R. Hiemann, T. Büttner, T. Krieger, D. Roggenbuck, U. Sack, and K. Conrad, "Challenges of automated screening and differentiation of non-organ specific autoantibodies on HEp-2 cells," *Autoimmun. Rev.*, vol. 9, no. 1, pp. 17–22, 2009.

[18] R. Hiemann, N. Hilger, J. Michel, J. Nitscke, A. Bohm, U. Anderer, M. Weigert, and U. Sack, "Automatic analysis of immunofluorescence patterns of HEp-2 cells," *Ann. NY Acad. Sci.*, vol. 1109, no. 1, pp. 358–371, 2007.

[19] P. Meroni and P. Schur, "ANA screening: An old test with new recommendations," *Ann. Rheumatic Diseases*, vol. 69, no. 8, pp. 1420–1422, 2010.

[20] Center for Disease Control, "Quality assurance for the indirect immunofluorescence test for autoantibodies to nuclear antigen (IF-ANA): Approved guideline," *NCCLS I/LA2-A*, vol. 16, no. 11, Dec. 1996.

[21] A. Rigon, P. Soda, D. Zennaro, G. Iannello, and A. Afeltra, "Indirect immunofluorescence in autoimmune diseases: Assessment of digital images for diagnostic purpose," *Cytometry B (Clin. Cytometry)*, vol. 72, pp. 472–477, 2007.

[22] A. Kavanaugh, R. Tomar, J. Reveille, D. H. Solomon, and H. A. Homburger, "Guidelines for clinical use of the antinuclear antibody test and tests for specific autoantibodies to nuclear antigens," *Am. College Pathol., Arch. Pathol. Lab. Med.*, vol. 124, no. 1, pp. 71–81, 2000.

[23] A. Bradwell and R. Hughes, *Atlas of HEp-2 Patterns: And Laboratory Techniques*. Birmingham, U.K.: Binding Site, 2007.

[24] D. H. Solomon, A. J. Kavanaugh, and P. H. Schur, "Evidence-based guidelines for the use of immunologic tests: Antinuclear antibody testing," *Arthritis Care Res.*, vol. 47, no. 4, pp. 434–444, 2002.

[25] MIVIA HEp-2 images dataset 2010 [Online]. Available: http://mivia.unisa.it/datasets/biomedical-image-datasets/hep2-image-dat

[26] HEp-2 Cells Classification contest website [Online]. Available: http://mivia.unisa.it/hep2contest/

[27] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Trans. Syst., Man Cybern.*, vol. 3, no. 6, pp. 610–621, Nov. 1973.

[28] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

[29] I. Ersoy, F. Bunyak, J. Peng, and K. Palaniappan, "Hep-2 cell classification in IIF images using shareboost," in *Proc. 21st Int. Conf. Pattern Recognit.*, Nov. 2012, pp. 3362–3365.

[30] T. Ojala, M. Pietikainen, and D. Harwood, "Performance evaluation of texture measures with classification based on Kullback discrimination of distributions," in *Proc. 12th IAPR Int. Conf. Pattern Recognit. Conf. A: Comput. Vis. Image Process.*, Oct. 1994, vol. 1, pp. 582–585.

[31] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, vol. 1, pp. 886–893.

[32] S. K. Nath and K. Palaniappan, "Adaptive robust structure tensors for orientation estimation and image segmentation," in *Proceedings of the First International Conference on Advances in Visual Computing*. Berlin, Germany: Springer-Verlag, 2005, pp. 445–453.

[33] J. Peng, C. Barbu, G. Seetharaman, W. Fan, X. Wu, and K. Palaniappan, "Shareboost: Boosting for multi-view learning with performance guarantees," in *Machine Learning and Knowledge Discovery in Databases*, ser. Lecture Notes in Computer Science, D. Gunopulos, T. Hofmann, D. Malerba, and M. Vazirgiannis, Eds. Berlin, Germany: Springer, 2011, vol. 6912, pp. 597–612.

[34] S. Ghosh and V. Chaudhary, "Feature analysis for automatic classification of hep-2 florescence patterns: Computer-aided diagnosis of auto-immune diseases," in *Proc. 21st Int. Conf. Pattern Recognit.*, Nov. 2012, pp. 174–177.

[35] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide baseline stereo from maximally stable extremal regions," in *Proceedings of the British Machine Vision Conference*. Cambridge, U.K.: BMVA Press, 2002, pp. 36.1–36.10.

[36] A. Hassaïne, E. Decenciére, and B. Besserer, "Efficient restoration of variable area soundtracks," *Image Anal. Stereol.* vol. 28, no. 2, 2011.

[37] A. Hassaïne, S. Al-Maadeed, J. M. Alja'am, A. Jaoua, and A. Bouridane, "The ICDAR2011 Arabic writer identification contest," in *Proc. Int. Conf. Document Anal. Recognit.*, 2011, pp. 1470–1474.

[38] I. Theodorakopoulos, D. Kastaniotis, G. Economou, and S. Fotopoulos, "Hep-2 cells classification via fusion of morphological and textural features," in *Proc. IEEE 12th Int. Conf. Bioinformat. Bioeng.*, Nov. , pp. 689–694.

[39] K. Li, J. Yin, Z. Lu, X. Kong, R. Zhang, and W. Liu, "Multiclass boosting SVM using different texture features in hep-2 cell staining pattern classification," in *Proc. 21st Int. Conf. Pattern Recognit.*, Nov. 2012, pp. 170–173.

[40] R. Marée and L. Wehenkel, "Extremely randomized trees and random subwindows for image classification, annotation, retrieval," in *Decision Forests for Computer Vision and Medical Image Analysis*, ser. Adv. Comput. Vis. Pattern Recognit., A. Criminisi and J. Shotton, Eds. London, U.K.: Springer, 2013, pp. 125–141.

[41] R. Nosaka, Y. Ohkawa, and K. Fukui, "Feature extraction based on co-occurrence of adjacent local binary patterns," in *Advances in Image and Video Technology*, ser. Lecture Notes Computer Science, Y.-S. Ho, Ed. Berlin, Germany: Springer, 2011, vol. 7088, pp. 82–91.

[42] V. Snell, W. Christmas, and J. Kittler, "Texture and shape in fluorescence pattern identification for auto-immune disease diagnosis," in *Proc. 21th IEEE Int. Conf. Pattern Recognit.*, Nov. 2012, pp. 3750–3753.

[43] P. Salembier and T. Sikora, *Introduction to MPEG-7: Multimedia Content Description Interface*, B. Manjunath, Ed. New York: Wiley, 2002.

[44] J. Serra, *Image Analysis and Mathematical Morphology*. Orlando, FL: Academic, 1983.

[45] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.* vol. 60, no. 2, pp. 91–110, Nov. 2004.

[46] P. Strandmark, J. Ulen, and F. Kahl, "Hep-2 staining pattern classification," in *Proc. 21st Int. Conf. Pattern Recognit.*, Nov. 2012, pp. 33–36.

[47] G. Thibault and J. Angulo, "Efficient statistical/morphological cell texture characterization and classification," in *Proc. 21st Int. Conf. Pattern Recognit.*, Nov. 2012, pp. 2440–2443.

[48] W. Bel haj ali, D. Giampaglia, M. Barlaud, P. Piro, R. Nock, and T. Pourcher, "Classification of biological cells using bio-inspired descriptors," in *Proc. 21st Int. Conf. Pattern Recognit.*, Nov. 2012, pp. 3353–3357.

[49] A. Wiliem, Y. Wong, C. Sanderson, P. Hobson, S. Chen, and B. Lovell, "Classification of Human Epithelial type 2 cell indirect immunofluoresence images via codebook based descriptors," in *IEEE Workshop Appl. Comput. Vis.*, Jan. 2013, pp. 95–102.

[50] C. Sanderson and B. C. Lovell, "Multi-region probabilistic histograms for robust and scalable identity inference," in *Proceedings of the Third International Conference on Advances in Biometrics*. Berlin, Germany: Springer-Verlag, 2009, pp. 199–208.

[51] T. Tommasi, F. Orabona, and B. Caputo, "Discriminative cue integration for medical image annotation," *Pattern Recognit. Lett.* vol. 29, no. 15, pp. 1996–2002, Nov. 2008.

[52] M. Varma and A. Zisserman, "A statistical approach to texture classification from single images," *Int. J. Comput. Vis.* vol. 62, no. 1–2, pp. 61–81, Apr. 2005.

[53] L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*. New York: Wiley-Interscience, 2004.