SIPAKMED: A NEW DATASET FOR FEATURE AND IMAGE BASED CLASSIFICATION OF NORMAL AND PATHOLOGICAL CERVICAL CELLS IN PAP SMEAR IMAGES

Marina E. Plissiti[1], P. Dimitrakopoulos[1], G. Sfikas[1,2], Christophoros Nikou[1], O. Krikoni[3], A. Charchanti[3]

[1]Dept. of Computer Science & Engineering, University of Ioannina, Greece
[2] CIL/IIT, NCSR "Demokritos", Athens, Greece
[3] Dept. of Anatomy-Histology and Embryology, Faculty of Medicine, University of Ioannina, Greece

ABSTRACT

Classification of cervical cells in Pap smear images is a challenging task due to the limitations these images exhibit and the complexity of the morphological changes in the structural parts of the cells. This procedure is very important as it provides fundamental information for the detection of cancerous or precancerous lesions. For this reason several algorithms have been proposed in order to classify normal and abnormal cells in such images. However, it is a common phenomenon that each research group usually creates its own dataset of images, as well-established datasets are not publicly available. To overcome this obstacle and to assist the research progress in this field, we present an annotated image database of Pap smear images, in which the cells are categorized in five different classes, depending on their cytomorphological features. The area of the cytoplasm and the nucleus in each image is manually defined by experts and salient features of intensity, texture and shape are calculated for each region of interest. Several experiments have been performed for the classification of these images and they include feature and image based classification schemes. In this direction, methods based on support vector machines and deep neural networks are tested and the performance of each classifier is presented in order to constitute a reference point for the evaluation of future classification techniques.

Index Terms— Pap smear images, cervical cell classification, cell image database, cell features, convolutional neural network.

1. INTRODUCTION

The automated interpretation of Pap smear images is one of the most interesting fields in cytological image analysis. This is a crucial problem and it combines several aspects of digital image processing, such as image enhancement, limitation of artifacts, object segmentation, delineation of overlapping cells etc. Many efforts have been made for the automated detection of the regions of interests in these images and they include several techniques [1, 2, 3].

Additionally, the integrated Pap smear image analysis includes classification of the image, based on its features. The cytomorphologic classification of cervical squamous cells in Pap smear images is important for the accurate diagnosis and the detection of cancerous or precancerous lesions. In general, the methods proposed for the automated classification of these images require images of single cells, which are cropped from cell clusters and further analyzed

[4, 5]. To the best of our knowledge, the only available dataset containing images of single cells is the Harlev dataset [6], which consists of a limited number of images (917). Thus, several researchers create their own annotated image dataset to evaluate the performance of their method. However, a major drawback of the existence of non-public specific datasets is that it is difficult to compare the efficiency of different classification techniques, as they are evaluated in individual datasets.

In this paper, we introduce the novel publicly available image dataset SIPaKMeD, which consists of 4049 annotated cells images. The cells are classified by expert cytopathologists into five different classes, depending on their cellural appearance and morphology. More specifically, normal cells are divided into two categories (superficial-intermediate, parabasal), abnormal but not malignant cells are divided into two categories (koilocytes and dyskeratotic) and there is also one category of benign (metaplastic) cells. In each image of our database, the area of the cytoplasm and the nucleus of the cells, is manually defined. In every region of interest, 26 features are computed, characterizing the intensity, the texture and the shape of the region of interest. Finally, we provide evaluation results using feature and image based classification schemes and some remarks on the discriminative ability of each classifier are presented.

2. SIPAKMED DATABASE

SIPaKMeD Database consists of 4049 images of isolated cells (Fig. 1) that have been manually cropped from 966 cluster cell images of Pap smear slides, which are also included. These images were acquired through a CCD camera (Infinity 1 Lumenera) adapted to an optical microscope (OLYMPUS BX53F). The distribution of the cells in classes is depicted in Table 1. In the following paragraphs, a brief description of each class is provided.

Table 1. SIPaKMeD Database

Category	Num of Images	Num of Cells
Superficial/Intermediate	126	813
Parabasal	108	787
Koilocytotic	238	825
Metaplastic	271	793
Dyskeratotic	223	813
Total	966	4049
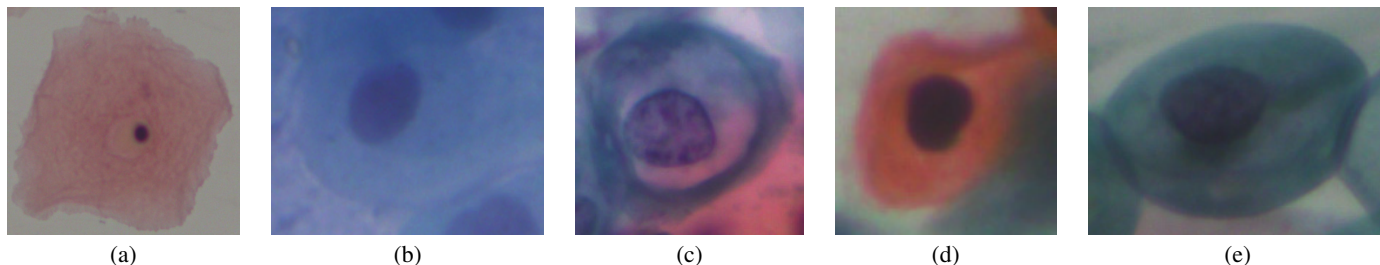
Fig. 1. Cell images of five categories: (a) Superficial-Intermediate, (b) Parabasal, (c) Koilocytotic, (d) Dyskeratotic, (e) Metaplastic.

2.1. Normal cells

These are squamous epithelial cells and their type is defined according to their position at epithelium layers and their degree of maturation.

2.1.1. Superficial-Intermediate cells

They constitute the majority of the cells found in a Pap test. Usually they are flat with round, oval or polygonal shape (Fig. 1(a)). The cytoplasm stains mostly eosinophilic or cyanophilic. They contain a central pycnotic nucleus. They have well defined, large polygonal cytoplasm and easily recognized nuclear limits (small pycnotic in the superficial and vesicular nuclei in intermediate cells). These type of cells show the characteristics morphological changes (koilocytic atypia) due to more severe lessions.

2.1.2. Parabasal cells

Parabasal cells are immature squamous cells and they are the smallest epithelial cells seen on a typical vaginal smear (Fig. 1(b)). The cytoplasm is generally cyanophilic and they usually contain a large vesicular nucleus. It must be noted that parabasal cells have similar morphological characteristic with the cells identified as metaplastic cells and it is difficult to be distinguished from them.

2.2. Abnormal cells

Abnormal cells are characterized by morphological changes in their structural parts and they indicate the existence of pathological situations . Human papillomavirus (HPV) is the cause of almost all cases of cervical cancers, manifests itself by characteristic changes of the squamous cells, two of which are pathognomonic: koilocytosis and dyskeratosis.

2.2.1. Koilocytotic cells

Koilocytotic cells correspond most commonly in mature squamous cells (intermediate and superficial) and some times in metaplastic type koilocytic cells (Fig. 1(c)). They appear most often cyanophilic, very lightly stained and they are characterized by a large perinuclear cavity. The periphery of the cytoplasm is very dense stained. The nuclei of koilocytes are usually enlarged, eccentrically located, hyperchromatic and exhibit irregularity of the nuclear membrane contour. In many cases there are binucleated and/or multinucleated cells. Koilocytic cells are pathognomonic cells for HPV infection and the nucleus of koilocytes usually displays various degrees of degeneration, depending at the different stage of infection and also of the different virus type infection.

2.2.2. Dyskeratotic cells

Dyskeratotic cells are squamous cells which undergone premature abnormal keratinization within individual cells or more often in three-dimensional clusters (Fig. 1(d)). They exhibit a brilliant orangeophilic cytoplasm. They are characterized by the presence of vesicular nuclei, identical to the nuclei of koilcytotic cells. They constitute a prominent characteristic of HPV infection, and sometimes even in the total absence of koilocytes, can be a pathognomonic evidence. They are usually in thick, three-dimensional clusters and it is difficult to distinguish either the nucleus or the cytoplasm margins.

2.3. Benign cells

These cells represent the transformation zone, the area in which almost all cervical precancerous and cancerous conditions develop.

2.3.1. Metaplastic cells

Metaplastic cells are in essence small or large parabasal-type cells with prominent cellular borders, often exhibiting eccentric nuclei and sometimes containing a large intracellular vacuole (Fig. 1(e)). The staining in the center portion is usually light brown and it often differs from that in the marginal portion. Also, there is essentially a darker-stained cytoplasm and they exhibit great uniformity of size and shape compared to the parabasal cells, as their characteristic is the well defined, almost round shape of cytoplasm. Their presence in the Pap test is associated with higher detection rates of pre-cancerous lesions (HSIL).
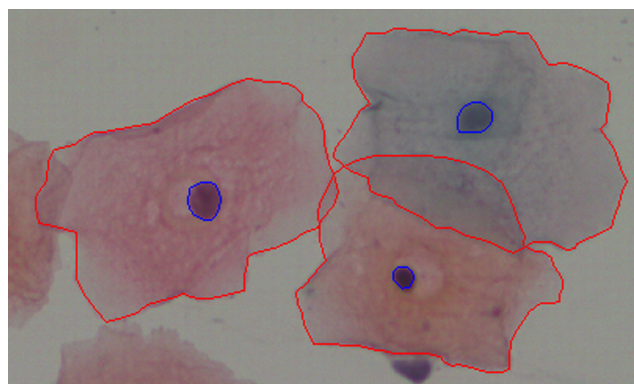
Fig. 2. The boundaries of the cytoplasm and the nucleus of each cell in images of cell clusters.

3. EVALUATION ON SIPAKMED

We have tested several classification schemes on the SIPaKMeD Database in order to evaluate their performance on the discrimination of the various cell types. Furthermore, we have used three different feature sets, namely cell features, image pixel features and deep features. We have followed a 5-fold cross validation scheme and the same training and test sets were used in all the experiments for the consistency of the results. In the following paragraphs a detailed description of each classification scheme is provided.

3.1. Cell Features

In each image of our database, the boundaries of the regions of interest, i.e. the area of the cytoplasm and the nucleus of the cells, are manually defined by expert observers. The coordinates of the contours of each area are provided for both the isolated cell images and the images of cell clusters (fig. 2). In every region of interest, we calculate 26 features concerning the intensity (average intensity, average contrast) and the texture (smoothness, uniformity, third moment, entropy) in all three color channels. Also, some shape features for each area were calculated (area, major and minor axis length, eccentricity, orientation, equivalent diameter, solidity and extent). These features were calculated for both the region of the nucleus and the cytoplasm of each cell and they are stored in 5 tables (one for each class) of 28 columns (the two added fields denote the number of the image and the cell correspondingly). These features were used for the classification of the cells using Support Vector Machines (SVM) and standard Multi-layer Perceptrons (MLP), as described in the following paragraphs.

3.1.1. Support Vector Machine (SVM)

In our experiments, the SVM classifier with Radial Basis Function (RBF) kernel was used. We have tested several values for the parameters C (0.05, 0.1, 0.5, 1, 2, 5, 10, 15, 20, 40) and γ (1, 2, 4, 8, 10, 15, 20, 30, 40, 50) and we used a 5-fold cross validation scheme for the selection of the optimal parameters. Furthermore, we performed two experiments using as input the features extracted from the nucleus and the features extracted from the cytoplasm. Since the SVM originally separates two categories, we have followed the one-versus-one approach and we have decomposed the 5-class problem into 10 pairwise problems. Applying each classifier to a test pattern would give one vote to the winning class. A test pattern is then labeled to the class with the most votes.

3.1.2. Multi-layer perceptron (MLP)

A multi-layer perceptron was employed using the same training and test set configuration as described above. MLPs are the simplest class of neural networks in terms of layer variability, comprising fully-connected hidden layers exclusively. We have test several different network architectures, and picked the best one by cross-validating on the architecture parameters. In particular, we have tested different numbers of hidden layers (1-10) and several number of hidden neurons (10, 20, 30, 40, 50). We have set all hidden layers to the same number of neurons in all cases. The hyperbolic tangent sigmoid function was selected as the activation function for all layers, save for the last one which uses a 5-class softmax function. The last, softmax-activated layer is of size 5 in order to suit the current problem requirements for 5 possible cell pathologies/types. The (softmax) output is a probability vector, denoting the probability that the input cell is of a given cell type. Regarding our architecture

cross-validation setup, we have used 10% of the training set for the construction of the validation set. The scaled conjugate gradient method was used to train the model, with a cross-entropy classification loss. Training is terminated after 30 epochs of increasing validation error.

3.2. Image Features

3.2.1. Convolutional Neural Network (CNN)

We have run classification tests using a convolutional neural network (CNN). The input to the network is cropped cell images, resized to a fixed size (in our case, 80×80 pixels each). Hence, no feature extraction step is required – the CNN is to learn a suitable internal representation automatically, given raw input RGB values. We have adapted the "very deep" architecture (VGG-19) of Krizhevsky et al. [7] for the current problem. In the employed architecture, the input is transformed by a series of stacks of convolutional layers topped by max-pooling layers. All convolutional filters are of size 3×3, and all max-pooling layers are of size 2×2. There are in total 5 convolutional stacks. The 2 first convolutional stacks are comprised of 2 convolution layers each, and other 3 stacks are comprised of 4 convolutional layers each. The depth (i.e., number of filters) of the convolutional layers in each stack is 64, 128, 256, 512, 512 respectively. The convolutional stacks are followed by a series of fully-connected layers, 3 in total. The size of the fully-connected layers are 4096, 4096, 5 respectively. In order to improve network generalization, dropout is employed on the first two fully-connected layers, with a keep probability of 50%. All layers, save for the last one, use ReLU activations. The last layer uses a 5-class softmax activation, as in the case of the standard MLP architecture (sec. 3.1.2). Also following standard practice, we train the model using a cross-entropy loss.

As the training set size is relatively small, especially for the needs of a deep neural network, we employ the technique of data augmentation to artificially create a larger training set. Hence, we augment our training set by creating 3 additional cell images for each existing training cell image. We do this by flipping the input image either horizontally, vertically, or in both directions. In this manner, we effectively quadruple the size of our training set.

A stochastic gradient descent (SGD) optimizer was used to train the model (batch size 50, learning rate 10^{-4}). Each training batch is selected from the training set at random. Training is terminated after $200,000$ iterations.

3.3. Deep features

We also use our convolutional network in an alternative way to its standard use as a softmax classifier. Following the trend of a number of recent works, we employ the network as a feature extractor [8, 9, 10, 11, 12, 13]. A neural network can be used as a feature extractor by feeding it an input image and using the intermediate layer activations or pre-activations to construct a feature vector. Features produced in this manner are usually referred to as deep features. The rationale behind this technique is that the neural network can be seen a machine that automatically learns the most suitable data representation [14]. Furthermore, deep features are known to have a more abstract character, lending to more transferable features. This especially holds for deep features based on layers closer to the input [13]. Deep features have been produced either given standard, fully-connected layers [8] or convolutional layers [10, 13]. In the case of fully-connected layers, the feature is an M-sized vector, where M is the number of neurons in the specified layer. Concerning convolutional layers, feature construction is a two-step process [10, 11],

Table 2. Comparison of classification accuracies (%) using the presented methodologies.

Features	MLP	SVM	CNN
Nuclei	78.81 ± 1.83	83.45 ± 1.53	−
Cytoplasm	88.54 ± 5.60	91.68 ± 0.98	−
Colour (RGB)	−	−	95.35 ± 0.42
Deep (convolutional)	−	93.35 ± 0.62	−
Deep (fully-connected)	−	94.44 ± 1.21	−

as layer activation is in general a $H \times W \times D$ tensor, where H, W, D are the height, width and depth of the current convolutiona layer. In general, H and W are different from the original image height and width, as max-pooling or strided convolutional layers (or layers equivalent to this effect) may lie between the input and the current layer, producing subsampled, filtered versions of the original input. The $H \times W \times D$ layer output can be seen as a set of $H \times W$ vectors of size D. These vectors can then be aggregated to form a single vector. Aggregation by simple sum-pooling has been shown to produce very useful features [10].

In the current context, we use the trained models to produce deep features by using the pre-activations of the last convolutional layer (layer $conv5$), as suggested in [10]. These are then aggregated by sum-pooling to form a feature vector of size 512 to describe each cell image. We also use the pre-activation of the first fully-connected layer (layer $fc6$), following [8]. These features are of size equal to the number of neurons in the corresponding layer, i.e. 4096. In both cases, we compress the extracted deep features to size 256 using PCA. Features are then fed into SVMs and class estimates are computed using the previously described voting scheme (sec. 3.1.1).

4. EXPERIMENTAL RESULTS

We have tested the presented classification methodologies on our database and the classfication results are included in Table 2. Mean/standard deviation accuracy values over 5 training-test folds are presented. Also, in figures 3, 4, 5 we present confusion matrices for all the employed classifiers. As we can see, in all cases the koilocytotic cells are the most challenging cells to be distinguished correctly. With respect to methods that employ the hand-crafted cytoplasm and nuclei features, it can be observed that the SVM classifier is in general more effective than the MLP classifier. We can also observe that cytoplasm features exhibit higher discriminative ability than the nuclei features. Methods based on hand-crafted features are however outperformed by convolutional neural network-based methods. The standard CNN setup gives the best average performance, with performance computed using deep features following closely behind.

5. CONCLUSION

In this paper, the publicly available SIPaKMeD cell image database is introduced. It contains both images of isolated cells and images of cell clusters, which were acquired from Pap smear slides. The images are divided into five categories of superficial-intermediate, parabasal, koilocytotic, dyskeratotic and metaplastic cells. The area of the cytoplasm and the nucleus of each annotated cell is manually identified by expert observers and the coordinates of their boundaries are also included. Three different types of features are provided for each pattern of our database. These are the hand-crafted

(a)   (b)
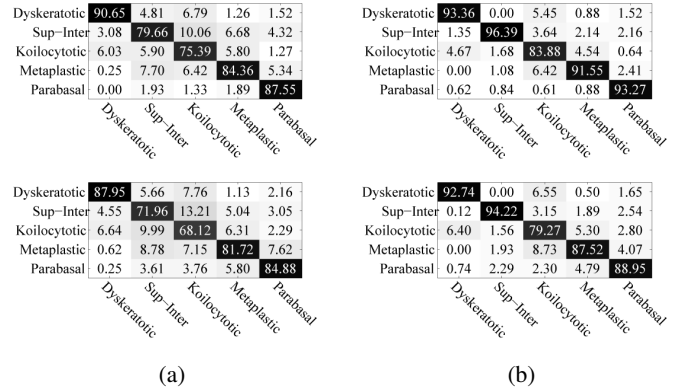
Fig. 3. Confusion matrices for classification using SVM (upper) and MLP (bottom) with (a) nuclei and (b) cytoplasm features.
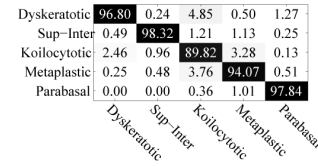
Fig. 4. Confusion matrix for classification using CNN.

(a)   (b)
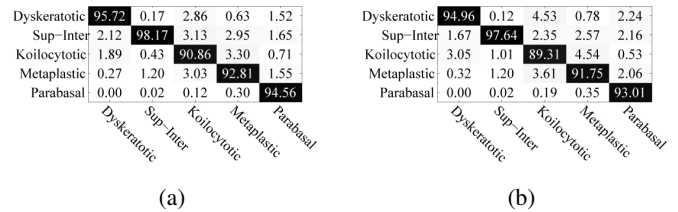
Fig. 5. Confusion matrices for classification using deep features based on (a) convolutional layer activations (b) fully-connected layer activations.

cell features, the image features and the deep features. The results of the classification schemes that were used in our experiments provide a reference point for the evaluation of future techniques for cell image classification. The database can be used not only for classification purposes but also for the evaluation of image segmentation techniques for isolated cells (cropped images) or overlapping cells (cell cluster images), since the ground truth for the regions of interest in each image is defined. Thus, the SIPaKMeD database provides new challenges and it constitutes a solid basis for competitive evaluations for the cell image analysis community.

6. REFERENCES

[1] M. H. Tsai, Y. K. Chan, Z. Z. Lin, S. F. Yang-Mao, and P. C. Huang, "Nucleus and cytoplasm contour detector of cervical smear image," *Pattern Recognition Letters*, vol. 29, pp. 1441–1453, 2008.

[2] K. Li, Z. Lu, W. Liu, and J. Yin, "Cytoplasm and nucleus segmentation in cervical smear images using radiating gvf snake," *Pattern Recognition*, vol. 45, no. 4, pp. 1255–1264, 2012.

[3] M. E. Plissiti and C. Nikou, "Overlapping cell nuclei segmentation using a spatially adaptive active physical model," *IEEE Transactions on Image Processing*, vol. 21, no. 11, pp. 4568–4580, 2012.

[4] Y. F. Chen, P. C. Huang, K. C. Lin, H. H. Lin, L. E. Wang, C.C. Cheng, T. P. Chen, Y. K. Chan, and J. Y. Chiang, "Semi-automatic segmentation and classification of pap smear cells," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 1, pp. 94–108, 2014.

[5] B. Sokouti, S. Haghipour, and A. D. Tabrizi, "A framework for diagnosing cervical cancer disease based on feedforward mlp neural network and thinprep histopathological cell image features," *Neural Computing and Applications*, vol. 24, no. 1, pp. 221–232, 2014.

[6] J. Jantzen, J. Norup, G. Dounias, and B. Bjerregaard, "Pap-smear benchmark data for pattern classification," in *Proceedings of Nature inspired Smart Information Systems (NiSIS)*, 2005, pp. 1–9.

[7] A. Krizhevsky, I. Sutskeverband, and G. E. Geoffrey, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems (NIPS)*, 2012, pp. 1097–1105.

[8] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky, "Neural codes for image retrieval," in *IEEE European Conference in Computer Vision (ECCV)*. Springer, 2014, pp. 584–599.

[9] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Hypercolumns for object segmentation and fine-grained localization," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 447–456.

[10] A. Babenko and V. Lempitsky, "Aggregating local deep features for image retrieval," in *IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[11] Y. Kalantidis, C. Mellina, and S. Osindero, "Cross-dimensional weighting for aggregated deep convolutional features," in *IEEE European Conference in Computer Vision (ECCV)*. Springer, 2016, pp. 685–701.

[12] G. Sfikas, G. Retsinas, and B. Gatos, "Zoning aggregated hypercolumns for keyword spotting," in *Proceedings of the 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 2016, pp. 283–288.

[13] G. Retsinas, G. Sfikas, and B. Gatos, "Transferable deep features for keyword spotting," in *IWCIM, in conjuction with EU-SIPCO*, 2017.

[14] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*, MIT press Cambridge, 2016.