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

OVERVIEW

Motivation: Classification of cervical cells in Pap smear images is a challenging task due to the limitations these images exhibit and well-established datasets are not publicly available.

Objective:
- We introduce the novel publicly available image dataset SIPAKMED .
- We demonstrate several classification schemes on the database.

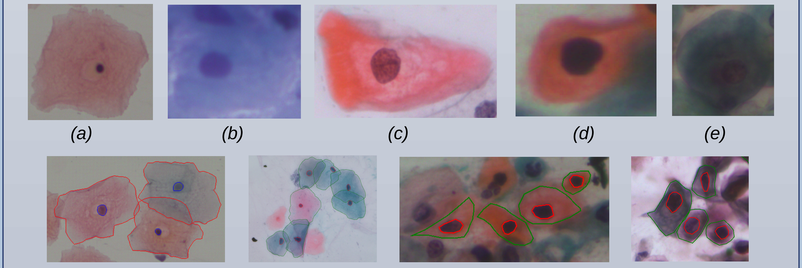
SIPAKMED DATABASE

- It consists of 4049 annotated images of isolated cells that have been manually cropped from 966 cell cluster images of **Pap smear slides**.
- The cells are classified into **five different classes**.
- The area of the **cytoplasm** and the **nucleus** of the cells is manually defined by expert cytopathologists.

Distribution of the cells in classes

Category	Num of Images	Num of Cells
Superficial/Intermediate	126	831
Parabasal	108	787
Koilocytotic	238	825
Metaplastic	271	793
Dyskeratotic	223	813
Total	966	4049

Categories of cell images
Normal cells: *(a)* Superficial-Intermediate, *(b)* Parabasal
Abnormal cells: *(c)* Koilocytotic, *(d)* Dyskeratotic
Bening cells: *(e)* Metaplastic



(a) *(b)* *(c)* *(d)* *(e)*

The boundaries of the cytoplasm and the nucleus of each cell in images of cell clusters.

EVALUATION ON SIPAKMED

We have tested the following classification schemes using 5-fold cross validation.
- Support Vector Machines (SVM) and Multi Layer Perceptron (MLP) based on features extracted from cytoplasm and nucleus.
- Convolutional Neural Network (CNN) based on RGB cropped cell images.
- SVM based on features extracted from the CNN.

Cell Features

In each image, for both the region of the nucleus and the cytoplasm of each cell we calculate 26 features concerning:

Intensity (average intensity, average contrast)
Texture (smoothness, uniformity, third moment,entropy)
Shape (area, major and minor axis length, eccentricity, orientation, equivalent diameter, solidity and extent)

Cell features were divided to **nuclei** and **cytoplasm** features.
These features were used for the classification of the cells using SVM and MLP.

Support Vector Machines (SVM)

Kernel: Radial Basis Function (RBF).
Parameters (C and γ): The optimal parameters were selected using 5-fold cross validation.
Training: One vs One approach (10 classifiers).

Multi Layer Perceptron (MLP)

Network Architecture: The optimal architecture was selected by cross validating on the architecture parameters.
Activation Functions: Last layer: 5-class softmax.
All the other layers: Hyperbolic tangent Sigmoid.
Training: Scaled conjugate gradient method terminated after 30 epochs of increasing validation error.
Loss: Cross-Entropy classification loss.

Image features

Convolutional Neural Network (CNN)

Input: Cropped cell images (80x 80 pixels, Raw RGB values).
Architecture: Vgg-19 [1].
Data Augmentation: 3 additional images for each image (horizontal, vertical and both flips).
Activation Functions: ReLU except the last one 5-class softmax.
Training: Stochastic Gradient Descent (batch size=50, lr=10^{-4}) with dropout, terminated on 200000 iterations.

Deep Features

We also use our convolutional network as a feature extractor [1].
- We feed our CNN an input image (cropped cell image).
- We use the pre-activations of the **last convolutional layer** aggregated by sum pooling [2] and the **first fully connected layer** [3].
- We construct two feature vectors (512,4096 in size both compressed to 256 using PCA).
- We finally feed these features to SVMs.

EXPERIMENTAL RESULTS

Comparison of classification techniques

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

Indicative Confusion Matrices

Cytoplasm SVM

	Dyskeratotic	Sup-Inter	Koilocytotic	Metaplastic	Parabasal
Dyskeratotic	93.36	0.00	5.45	0.88	1.52
Sup-Inter	1.35	96.39	3.64	2.14	2.16
Koilocytotic	4.67	1.68	83.88	4.54	0.64
Metaplastic	0.00	1.08	6.42	91.55	2.41
Parabasal	0.62	0.84	0.61	0.88	93.27

RGB CNN

	Dyskeratotic	Sup-Inter	Koilocytotic	Metaplastic	Parabasal
Dyskeratotic	96.80	0.24	4.85	0.50	1.27
Sup-Inter	0.49	98.32	1.21	1.13	0.25
Koilocytotic	2.46	0.96	89.82	3.28	0.13
Metaplastic	0.25	0.48	3.76	94.07	0.51
Parabasal	0.00	0.00	0.36	1.01	97.84

Observations:

- CNN setup gives the best average performance with deep features following.
- Koilocytotic cells are the most challenging to be distinguished.
- With respect to methods based on cell features SVM classifier is in general more effective than MLP.

CONCLUSION

- We introduce the publicly available SIPAKMED cell image database.
- It contains both images of isolated cells and images of cell clusters, which are divided into five categories.
- Three different types of features are provided.
- The results of the classification schemes provide a reference point for the evaluation of future methodologies.
- The database can be also used for evaluation of image segmentation methods for isolated cells (cropped images) or overlapping cells (cell clusters).

REFERENCES

[1] A. Krizhevsky, I. Sutskeverband and G. E. Geoffrey, "Imagenet classification with deep convolutional neural networks", in Proceedings of Advances in Neural Information Processing Systems (NIPS), pp. 1097-1105, 2012.
[2] A. Babenko and V. Lempitsky, "Aggregating local deep features for image retrieval", in Proceedings of IEEE International Conference on Computer Vision (ICCV), December 2015.
[3] A. Babenko, A. Slesarev, A. Chigorin and V. Lempitsky, "Neural codes for image retrieval", in Proceedings of IEEE European Conference on Computer Vision (ECCV), Springer, pp. 584–599, 2014.

ACKNOWLEDGMENT

This work was co-financed by the European Union (European Regional Development Fund-ERDF) and Greek national funds through the Operational Program THESSALY- MAINLAND GREECE AND EPIRUS-2007-2013 of the National Strategic Reference Framework (NSRF 2007-2013). Also has been co-financed by the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH–CREATE–INNOVATE (project code:T1EDK04517).

The SIPAKMED database is available on ***www.cse.uoi.gr/~marina.***