Cervical Cell Classification Based Exclusively on Nucleus Features

Marina E. Plissiti and Christophoros Nikou*

Department of Computer Science, University of Ioannina, Greece
{marina,cnikou}@cs.uoi.gr

Abstract. In this work, we present a framework for the efficient classification of cervical cells in normal and abnormal categories, based on features extracted exclusively from the nucleus area and ignoring the contingent cytoplasm features. This task is very important, since the nuclei are the only distinguishable areas in complex Pap smear images, as these images present a high degree of cell overlapping and the exact borders of the cytoplasm areas are ambiguous. We have examined the ability of non-linear dimensionality reduction schemes to produce accurate representation of the features manifold, along with the definition of an efficient feature subset, and their influence on the classification performance. Two unsupervised classifiers were used and the results indicate that we can achieve high classification performance when only the nuclei features are used.

Keywords: Pap smear images, abnormal cell classification, non-linear dimensionality reduction, spectral clustering, fuzzy C-means.

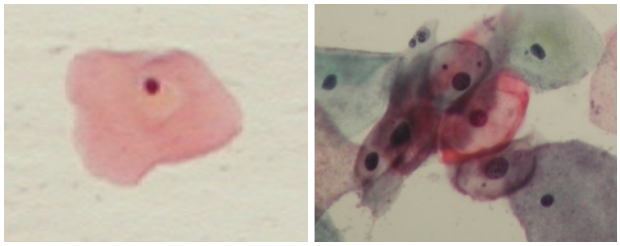
1 Introduction

Cervical smear screening is the most popular method used for the detection of cervical cancer in its early stages. The most eminent screening test is the Pap smear, which is based on the staining of cervical cells, using the technique that was first introduced by George Papanicolaou [1]. With this screening technique, the sample of exfoliated epithelial cells is stained and smeared onto a glass slide. After the careful examination of the slide by an expert cytologist, precancerous conditions and abnormal changes in cells that may develop into cancer are recognized. The widespread use of this test in developed countries has significantly reduced the incidence and mortality of invasive cervical cancer.

The interpretation of these images relies basically on the visual recognition of the changes of the structural parts of the cells (nucleus and cytoplasm). However, this process is a tedious, time-consuming and in many cases error-prone procedure due to the high degree of complexity that these images exhibit. Several approaches have been proposed for the classification of cells in Pap smear images and they concern techniques

* This work is 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).

A. Campilho and M. Kamel (Eds.): ICIAR 2012, Part II, LNCS 7325, pp. 483–490, 2012.



(a) (b)

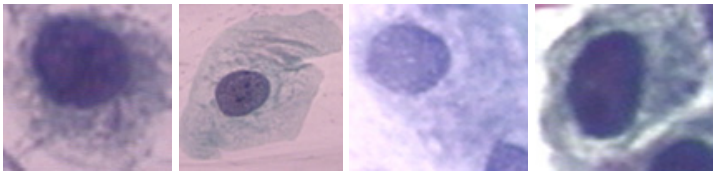
Fig. 1. (a) An isolated cell and (b) overlapping cells. Notice that the cytoplasm area is clearly recognized in (a) while in (b) its determination is very ambiguous for each cell.

such as Bayesian classifiers [2], artificial neural networks [3], support vector machines (SVM) [4] and nearest neighbor based classifiers [5]. It must be noted that most of these methods use presegmented images which contain only one cell, so the correct segmentation of the nucleus and the cytoplasm is feasible (Fig. 1(a)). In images containing cell clusters (Fig. 1(b)), the detection of the cytoplasm boundary is a difficult problem and until now, there is not any method in the literature that results in the automated delineation of the cytoplasm areas in cell clusters. However, the detection and segmentation of the nuclei in images containing cell overlapping and cell clusters has been successfully addressed by several studies [6], [7].

The methods which deal with the classification of Pap smear images are based on the calculation of features extracted from the areas of the nucleus and the cytoplasm [5,8]. These features are usually based on shape and intensity characteristics of the objects of interest. However, the calculated features do not exhibit the same discriminative ability. For the determination of the most efficient feature set which will be used as input in a classifier, some feature selection schemes have been proposed, and they concern genetic algorithms [5] and particle swarm optimization [8].

Based on the aforementioned facts, we can conclude that there are two open problems in the automated classification of a Pap smear image acquired directly from an optical microscope: a) the limitation to use only the features extracted from the nuclei areas, as these are the areas that can be automatically segmented, and b) the determination of the most efficient feature subset, which will provide the best discriminative ability.

In this work, we evaluate the classification of cervical cells, based exclusively on nucleus features and ignoring the features extracted from the cytoplasm area. This is a crucial step for the correct characterization of Pap smear images acquired directly from an optical microscope, where the cell overlapping is an often found phenomenon and the delineation of the cytoplasm area can not be obtained automatically. In this direction, we investigate the representation of the features in low dimensional spaces using non linear dimensionality reduction methods. These techniques are advantageous in comparison with their linear counterparts, because they can properly handle complex nonlinear data, as they better describe their manifold structure.



(a) (b) (c) (d)

Fig. 2. Types of cells included in the Pap smear benchmark [11]. (a)-(b) Abnormal cells and (c)-(d) normal cells.

The low dimensional feature subsets serve as input in two unsupervised classifiers (Spectral Clustering [9] and fuzzy C-means [10]). As it was verified by the results, the non-linear dimensionality reduction techniques lead to a construction of nucleus-only feature subsets which can be successfully used for the separation of normal and abnormal cells by the classifiers, presenting high performance.

2 Materials and Methods

2.1 Study Group

Our experiments are based on the Pap-smear benchmark database presented in [11] . The database consists of 917 images containing a single cell each (Fig. 2), and the samples are distributed unevenly in seven classes. Three of them are considered as normal and four of them are considered as abnormal types of cell. The detailed description of the database is depicted in Table 1.

Table 1. Distribution of cells in the Pap-smear benchmark database [11]

NORMAL	#cells
Superficial squamous epithelial	74
Intermediate squamous epithelial	70
Columnar epithelial	98
TOTAL	**242**
ABNORMAL	#cells
Mild squamous non-keratinizing dysplasia	182
Moderate squamous non-keratinizing dysplasia	146
Severe squamous non-keratinizing dysplasia	197
Squamous cell carcinoma in situ intermediate	150
TOTAL	**675**

2.2 Feature Generation and Dimensionality Reduction

The images of the database have been manually segmented by experts and the areas of the nucleus and the cytoplasm are accurately defined. From these areas, twenty features

concerning the intensity and the shape characteristics of the specific area are determined (Table 2). Nine out of twenty features concern the nucleus area and they can be calculated independently.

The techniques that we have used for the construction of the new feature sets concern non-linear dimensionality reduction schemes. In our study we have investigated the performance of four nonlinear techniques: Kernel-PCA [12], Isomap [13], Locally Linear Embedding [14] and Laplacian Eigenmaps [15]. A brief description of these techniques is presented in the following paragraphs.

Table 2. Features extracted from each image in the database

Cytoplasm Features	Nuclei Features
1. Area	1. Area
2. Brightness	2. Brightness
3. Short Diameter	3. Short Diameter
4. Longest Diameter	4. Longest Diameter
5. Elongation	5. Elongation
6. Roundness	6. Roundness
7. Perimeter	7. Perimeter
8. Maxima[1]	8. Maxima[1]
9. Minima[1]	9. Minima[1]
10. Nucleus Position	
11. Nucleus/Cytoplasm (size)	

[1] The number of pixels with the maximum/minimum intensity value in a 3×3 neighborhood of the specific area.

Kernel Principal Component Analysis (K-PCA) [12] is actually an extension of the conventional PCA in a high-dimensional space, which is obtained with the use of a kernel function. The main difference in comparison with the standard PCA is that the eigenproblem is solved for the "kernelized" covariance matrix. If $X = \{x_1, x_2, ..., x_N\}$ is the original data set, the elements of this $N \times N$ matrix are defined as $k_{ij} = K(x_i, x_j)$, where K is the kernel function and x_i, x_j are D-dimensional feature vectors of X. In our implementation, we have used the polynomial and the Gaussian kernels. The kernel matrix is centered, in order the features in the high dimensional space to be defined by a kernel function with zero mean and the eigenvectors α_i are then calculated. The projection of a datum y_i in the low dimensional space is defined as $y_i = \left\{ \sum_{j=1}^{N} a_1^j K(x_j, x_i), ..., \sum_{j=1}^{N} a_d^j K(x_j, x_i) \right\}$, where a_i^j denote the j-th component of the i-th vector and $d < D$ is the number of the retained eigenvectors.

Isomap [13] is a variant of multidimensional scaling (MDS) [16], in which the distances between the datapoints in the high dimensional space are also retained in the low dimensional space. In MDS, this is accomplished by the eigendecomposition of a pairwise distance matrix (instead of the covariance matrix which is involved in PCA). In Isomap, the Euclidean distance between the points is substituted by their geodesic distance. Thus, the pairwise geodesic distance between the datapoints in the high dimensional space is preserved in the low dimensional space, by the construction of a

neighborhood graph *G*, in which each datapoint is connected with its *k* nearest neighbors. The geodesic distance of two points may be approximated with the shortest path in the graph *G* between these points, using, for instance, Dijkstra's algorithm [17]. Having estimated the geodesic distances for all the points in the data set, the representation of the datapoints in the low dimensional space are computed by applying MDS on the resulting distance matrix.

Locally Linear Embedding (LLE) [14] is similar in spirit to Isomap, as it is also based on the construction of a distance graph *G*. However, in LLE each datapoint is described as a linear combination of its *k* nearest neighbors, thereby assuming that the manifold is locally linear. The weights w_{ij} describe the contribution of the j-th point to the reconstruction of the i-th point and are computed by minimizing the cost:

$$\arg\min_{W} E(W) = \sum_{i=1}^{N} ||x_i - \sum_{j=1}^{k} w_{ij} x_{i_j}||^2,$$

where x_{i_j} is the j-th nearest neighbor of the i-th point. Thus, the weights w_{ij} that best reconstruct each point x_i from its neighbors are used to compute the corresponding points y_i in the low dimensional space by minimizing the following cost function with respect to $Y = (y_1, y_2, ..., y_N)^T$:

$$\arg\min_{Y} \varphi(Y) = \sum_{i}^{N} ||y_i - \sum_{j=1}^{k} w_{ij} y_{i_j}||^2.$$

This minimization problem is equivalent to the calculation the eigenvectors corresponding to the smallest eigenvalues of the matrix $(I - W)^T(I - W)$, where *I* is the identity matrix and W is a matrix with elements w_{ij}. The above minimization is performed in two steps with the additional constraint $\sum_j w_{ij} = 1$ to make the representation translation invariant.

Laplacian Eigenmaps method [15] has as its main philosophy to calculate the low dimensional representation of the data in such a way that the local neighborhood information is optimally preserved. For this reason, the distance graph *G* is computed, in a way similar with the methods described above. Each edge of the graph is associated with a weight, which is a measure of closeness of the respective neighbors. The weights are attributed by the Gaussian kernel function $w_{ij} = e^{-\frac{||x_i - x_j||^2}{2\sigma^2}}$, where σ^2 indicates the variance of the Gaussian. Thus, the weights exhibit high values for nearest neighbors and small values for distant datapoints. Next, a diagonal matrix *A* is constructed, with elements $A_{ii} = \sum_j w_{ij}, i = 1, .., n$ and the generalized eigendecomposition $Lu = \lambda Au$ is performed, where $L = A - W$. The low dimensional representation is obtained using the *d* eigenvectors corresponding to the smallest nonzero eigenvalues.

3 Results and Discussion

In order to investigate the effectiveness of the above dimensionality reduction schemes, we have used two unsupervised classifiers and two datasets of patterns. More

specifically, spectral clustering and Fuzzy C-means are tested using patterns from two different feature sets (Table 1): one containing both cytoplasm and nucleus features (20 features) and the other containing only nucleus features (9 features). Several experiments were performed and the performance of the classification techniques was measured using patterns of increasing dimension varying from 1 to 20 features for the first subset and from 1 to 9 for the second subset. Furthermore, different values for the kernel width of spectral clustering have been tested $(10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 2, 5)$. In Isomap, LLE and Laplacian Eigenmaps, different numbers of nearest neighbors ranging from 4 to 20 were also tested for the construction of the distance graph G. The best results for each classifier are presented in this work.

For comparison purposes, PCA was also implemented. For the evaluation of the classification performance, the harmonic mean (H-mean) of the sensitivity and the specificity indices was calculated. The sensitivity measures the proportion of abnormal cells which are correctly identified as such by the classification algorithm, and the specificity measures the proportion of the normal cells that are correctly characterized as such.

The classification results of spectral clustering and fuzzy C-means are depicted in Table 3. For each feature subset, the H-mean and the number of features retained by the dimensionality reduction techniques are presented. As we can observe, the initial features without the use of dimensionality reduction schemes, lead to the weakest classification performance. The use of either linear or non-linear dimensionality reduction schemes results in a significant improvement of the classification.

More specifically, regarding the linear dimensionality reduction technique (PCA), we can conclude that there is a small improvement in the classification results, compared to the case where no dimensional reduction technique is used. Furthermore, in fuzzy C-means we observe a significant reduction in the retained features. Only 3 out of 20 dimensions are retained in first set of features and only 4 out of 9 for the nuclei feature subset. Finally, in spectral clustering, better classification results are produced when only the nuclei features are used, in comparison with the use of both nuclei and cytoplasm features.

In non-linear dimensionality reduction schemes, we can notice that the performance of the classifiers is clearly better when they are based only on nuclei features (except in the case of LLE with spectral clustering, where the results are approximately similar). In spectral clustering, an improvement of 12.16% in the classification is observed in Isomap, where the best value of H-mean (88.77%) using only the nuclei features is reached. Furthermore, in fuzzy C-means, the corresponding highest difference in classification rates is 11.17% and it is observed using K-PCA (polynomial kernel). Nevertheless, the best classification result using only the nuclei features is 90.58% with K-PCA (Gaussian kernel). It must be noted that this result is obtained using only seven features, while for different number of features the H-mean value is smaller (Fig. 3).

The obtained results clarify that the use of non-linear dimensionality reduction schemes, not only improves the classification performance of spectral clustering and fuzzy C-means, but they also allow the successful separation of normal and abnormal cercival cells, based exclusively on nuclei features.

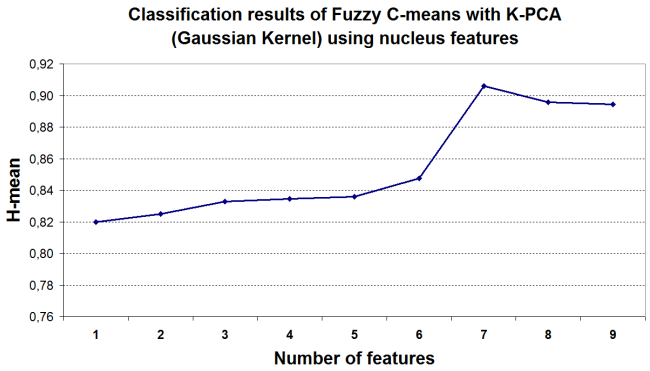


Fig. 3. Results obtained in terms of H-mean for Fuzzy C-means classification. Notice that the H-mean reaches its highest value for seven features.

Table 3. Performance of classification in terms of H-mean and the number of retained features

	Spectral Clustering				Fuzzy C-means			
	All Features		Nuclei Features		All Features		Nuclei Features	
	#feat	HM(%)	#feat	HM(%)	#feat	HM(%)	#feat	HM(%)
No dimensionality reduction	20	74.21	9	73.59	20	72.89	9	71.98
PCA	6	74.25	7	83.38	3	74.23	4	71.99
K-PCA (polynomial)	2	85.78	9	88.53	3	74.24	3	85.41
K-PCA (Gaussian)	16	84.44	7	87.52	9	**90.42**	7	**90.58**
Isomap	1	76.61	9	**88.77**	1	75.02	3	75.08
LLE	17	**86.97**	9	86.45	15	81.69	6	87.17
Laplacian Eigenmaps	20	80.84	3	87.52	11	85.31	1	87.20

4 Conclusion

The correct characterization of the cell nuclei in Pap smear images is a prerequisite for the derivation of accurate diagnostic decisions. Since in cell clusters presented in Pap smear images the automated cytoplasm segmentation is not feasible, in contrast to the automated nuclei segmentation [6], [7], we have investigated the case of the successful classification of cells with exclusively nuclei features using two unsupervised classifiers. In this direction, non-linear dimensionality reduction techniques were also used, for the more accurate representation of the features manifold. As it was verified by our experiments, the obtained results using only the nuclei features are better than the results obtained using all the extracted features (from the areas of nucleus and the cytoplasm). This implies that the characterization of a Pap smear image as normal or abnormal is feasible with the use of the nuclei features alone. This may contribute in the development of a fully automated method for the classification of microscopic cervical cell images, which embodies automated nuclei segmentation, nuclei feature extraction

and finally classification. At a next step, we intend to investigate the application of supervised techniques, such as support vector machines, to the correct classification of normal and abnormal cervical cells, based exclusively on nuclei features lying in low dimensional manifolds.

References

1. Papanicolaou, G.N.: A new procedure for staining vaginal smears. Science 95(2469), 438–439 (1942)
2. Riana, D., Murni, A.: Performance evaluation of Pap smear cell image classification using quantitative and qualitative features based on multiple classifiers. In: Proceedings of the International Conference on Advanced Computer Science and Information Systems, ACSIS 2009 (2009)
3. Mat Isa, N.A., Mashor, M.Y., Othman, N.H.: An automated cervical pre-cancerous diagnostic system. Artificial Intelligence in Medicine 42, 1–11 (2008)
4. Huang, P.-C., Chan, Y.-K., Chan, P.-C., Chen, Y.-F., Chen, R.-C., Huang, Y.-R.: Quantitative Assessment of Pap Smear Cells by PC-Based Cytopathologic Image Analysis System and Support Vector Machine. In: Zhang, D. (ed.) ICMB 2008. LNCS, vol. 4901, pp. 192–199. Springer, Heidelberg (2007)
5. Marinakis, Y., Dounias, G., Jantzen, J.: Pap smear diagnosis using a hybrid intelligent scheme focusing on genetic algorithm based feature selection and nearest neighbour classification. Computers in Biology and Medicine 39, 69–78 (2009)
6. Plissiti, M.E., Nikou, C., Charchanti, A.: Automated detection of cell nuclei in Pap smear images using morphological reconstruction and clustering. IEEE Transactions on Information Technology in Biomedicine 15(2), 233–241 (2011)
7. Plissiti, M.E., Nikou, C., Charchanti, A.: Combining shape, texture and intensity features for cell nuclei extraction in Pap smear images. Pattern Recognition Letters 32(6), 838–853 (2011)
8. Marinakis, Y., Marinaki, M., Dounias, G.: Particle swarm optimizaton for Pap-smear diagnosis. Expert Systems with Applications 35, 1645–1656 (2008)
9. Ng, A.Y., Jordan, M.I., Weiss, Y.: On spectral clustering: Analysis and an algorithm. Advances in Neural Information Processing Systems 14, 849–856 (2002)
10. Bishop, C.: Pattern recognition and machine learning. Springer (2006)
11. Jantzen, J., Norup, J., Dounias, G., Bjerregaard, B.: Pap-smear benchmark data for pattern classification. In: Proceedings of Nature inspired Smart Information Systems (NiSIS), pp. 1–9 (2005)
12. Scholkopf, B., Smola, A., Muller, K.R.: Nonlinear component analysis as a kernel eigenvalue problem. Neural Computation 10, 1299–1319 (1998)
13. Langford, J.C., Tenenbaum, J.B., De Silva, V.: A global geometric framework for nonlinear dimensionality reduction. Science 290, 2319–2323 (2000)
14. Roweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. Science 290(5500), 2323–2326 (2000)
15. Belikn, M., Niyogi, P.: Laplacian eigenmaps for dimensionality reduction and data representation. Neural Computation 15(6), 1373–1396 (2003)
16. Cox, T.F., Cox, M.A.A.: Multidimensional Scaling, 2nd edn. Chapman and Hall, CRC (2001)
17. Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C.: Introduction to algorithms, 2nd edn. MIT Press and McGraw-Hill (2001)