PAPDIA FINAL REPORT

I. PROJECT OBJECTIVES (*"expected" in gray*) AND OUTCOMES (*"realized" in black*)

The main goal of this project was to generate novel intelligent automated classification algorithms for the cervical Pap-smear sample sets for the implementation of a pilot intelligent system for automated diagnosis aid, which will be used in medical clinics and laboratories. This goal was achieved by checking through the following specific objectives:

1. Obtaining a marked image database and a test database

The main goal of this project can only be achieved after obtaining and processing an image database which is to be used in the following segmentation and classification stages. This database is the starting point for all the research activity. An important issue here is the the quality of images which will have a high impact on the other stages. We will have a strict set of guidelines for obtaining the images. The second challenge related to this database is the number of the analyzed medical cases. We estimate at least 2000 cases will be required.

Our (digital) images of Pap-smear stains are obtained with a scanner by VENTANA iScan Coreo, at an optic zoom of 40.

Cells from five subjects were collected and classified according to the Bethesda system (Table 1). A total of 9405 cells were collected, thus exceeding our estimate of 2000 by more than 4 times. Our collection contains two types of pathologic cells and three types of healthy cells. The two pathologic types are ASCUS and LSIL, which correspond to light and moderate progress in cancer development. The three healthy types are the superficial, intermediate and inflamated.

Table 1. Cell count per type.

Type	Grade	Count	# Subjects
ASCUS	light	1154	5
LSIL	moderate	637	5
superficial	normal	2693	4
intermediar	normal	3648	5
inflamated	normal	1273	5
Total Cells		9405	
non-nuclei	not appl.	10743	5
Total Patches		20148	

The collected patches are of size 301x301, but because we often focus on the nucleus only, the examples depicted in Figure 1 show a zoom-in. A significant proportion of collected nuclei are not isolated, but occur as part of overlapping cell agglomerates or other structures, meaning the collection is not selected for optimal cases, but is representative for the cases as found in an image.

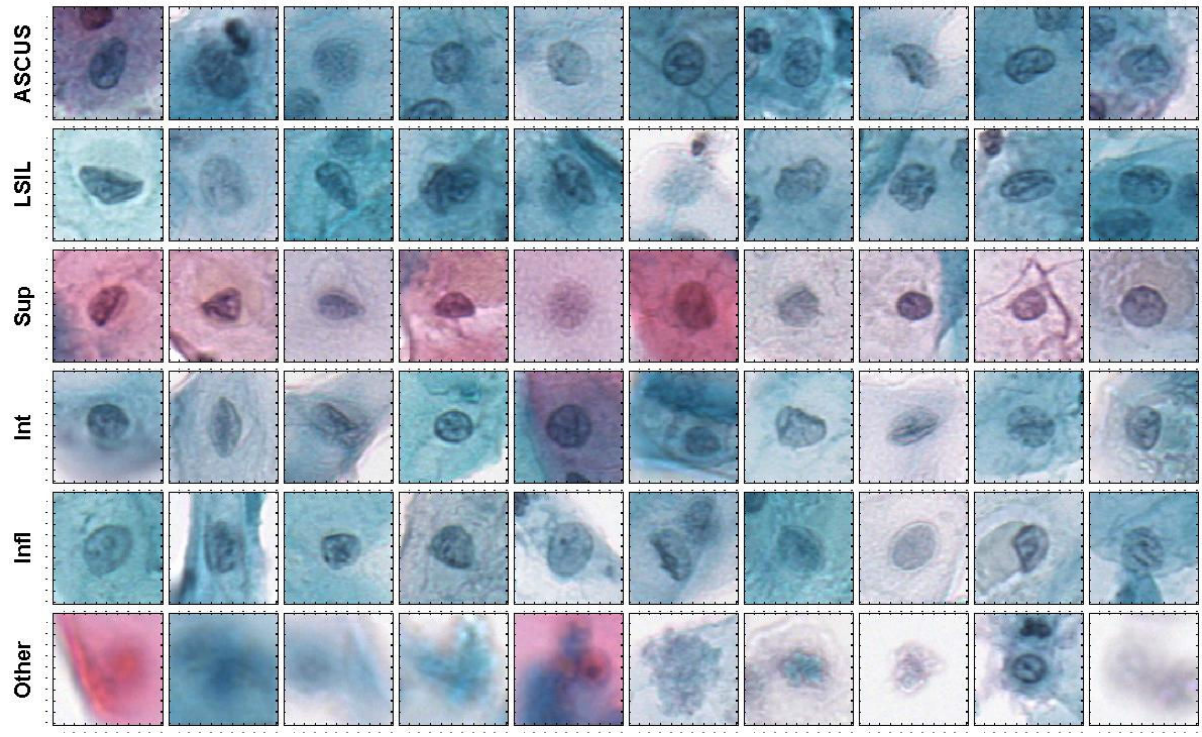


Figure 1. Examples of collected nuclei. The first two rows are representatives of pathologic cases (ASCUS, LSIL); rows three to five are examples of healthy cells (superficial, intermediar, inflamated). The bottom row (other) shows structures similar to those cell nuclei or other nuclei types. The images show 101x101-pixel patches, but optimal classification with a Deep Belief Network can be achieved with patches of size 61x61 approximately.

In contrast to the popular Herlev database (Marinakis2009), our set of images does not contain pathologic cells that are in severe or even carcinogen stage. Those two cell types are however easier to discriminate from healthy cells than the cell types ASCUS and LSIL. Our database is therefore suitable for trying to detect the early stages of cancer development, which is the more pressing issue.

The degree of achievement: We have fully completed this objective. We have collected many more cases than expected.

2. Developing and implementing new segmentation and feature extraction algorithms

This is also a very important step in achieving this project's ultimate goal. A possible issue here is identifying the relevant features as close as they are found in the manual diagnosis process. A thorough and accurate feature extraction is required.

We experimented with two novel approaches for segmentation and feature extraction. One is based on a polar description of the cell (Section 2.1). And another one based on a description by iso-contours (Section 2.2).

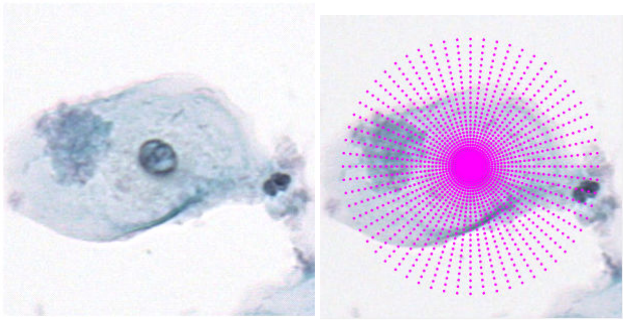
2.1. Polar Description

The polar-description method we developed consisted of the following three steps:

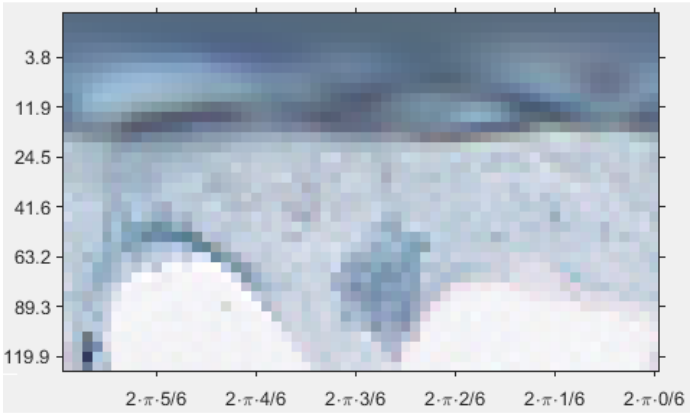
1. An image Cartesian-to-Polar transformation, re-dimensioning and the addition of a fourth pixel dimension.
2. Use of the k-means clustering method as the primary segmentation.

3. Post-processing, including merging of clusters.

We explain those three steps in some more detail:



a. Candidate image **b**. Sampling points



c. Polar transform of the re-sampled candidate image

Figure 2. Polar description of a nucleus. **a**: is a candiate image of a nucleus. **b**. Shows examples of sampling points - from the center outward in all directions. **c**. Shows the polar image with y-axis corresponding to radius (top-down) and x-axis corresponding to angle.

1. Image transformation The candidate images contain, in the ideal case, concentric layers (representing nucleus, cytoplasm and background), therefore the polar transformation comes as a natural first step. The transformation itself does not introduce or uncover new information, but it arranges the information in a way that makes further operations easier. This step is also used here for re-dimensioning, keeping, for each image, only pixels at certain angles theta and certain distances d from the center of the image, marked with magenta in **Fig. 2b**. Using a set of 60 equally spaced angles and 36 distances, the number of pixels can be reduced from 90601 to just 2160 while at the same retaining the most significant elements of the image. The number of angles and distances are adjustable parameters, but the chosen values seemed to be sufficient for the current resolution. Since the most significant information is concentrated around the nucleus (and the nucleus is assumed to contain the seed), the selected distances become more sparse with distance from seed.

In order to improve the k-means segmentation in step B, a fourth component is added pre-emptively to each pixel of the transformed image. The reason for the fourth value is to convey the information to the k-means clustering that, ideally, the top pixels form the nucleus, the middle pixels form the cytoplasm and the lower pixels form the background of the cell. Looking at the values v along any column of the transform image, the fourth dimension encourages pixels at very small or very large distances from the seed to classify naturally,

while also gently increasing the Euclidean distance between the top and bottom pixels in the transformed image.

2. K-means clustering The 4D pixel array is clustered using the k-means for three values of k (k=3, k=4 and k=5) and for determining the optimal number of clusters, we have considered the silhouette method (**Fig. 3c**).

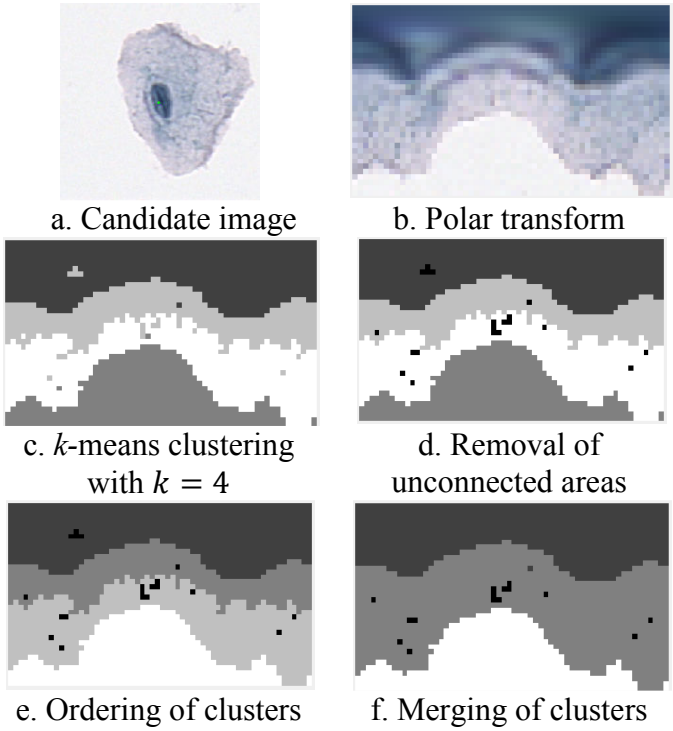


Fig. 3. Processing chain for a candidate image.

3. Post-processing Post-processing consists of various morphological operations that further improve the clustering results:

- *Removal of unconnected areas*: Considering that the cell should have the nucleus and cytoplasm in one connected area each, this step keeps for each cluster only the largest connected area and re-assigns any other group of pixels to label black (unused), represented as black in further images (**Fig. 3d**). For this step, the first and last columns of the transformed image are considered connected (since they represent consecutive sampling angles).

- *Merging of nucleus parts*: If the centroids of the uppermost two clusters are close enough, they should merge. Also as part of this step, black areas engulfed by the nucleus are assigned the same label, the nucleus label.

- *Ordering of clusters*: This step orders the clusters so that the uppermost cluster (usually the nucleus) is the darkest and the lowermost cluster (usually the whitish background) is the lightest (**Fig. 3e**). As decided at the previous step, the nucleus is the cluster with the lowest mean vertical position of the pixels (closest to the uppermost row). Cytoplasm is the cluster with the greatest contact to the nucleus. The other clusters are then sorted by the mean vertical position of the pixels.

- *Merging of cytoplasm parts*: For the cytoplasm cluster determined at the previous step and each of the other clusters except the nucleus, the statistical mean and standard deviation of the red layer or the blue layer are computed (depending on the dominant color of

the nucleus) (**Fig. 3f**). If the normalized distributions of the cytoplasm cluster and another cluster overlap significantly, the two clusters are merged.

Feature Extraction (for Polar Description) Once the segmentation has been completed, parameters useful to the decision making module can be extracted. The aim of the decision module is thus the identification of non-cells with the reasoning that anything that is not weird enough should be treated as an actual cell.
There are two types of parameters: strong indicators and weak indicators that the candidate is a non-cell.

The strong indicators are:

I_1: Whiteness of the nucleus class
I_2: Whiteness of the cytoplasm class
I_3: Contrast between the nucleus class and the cytoplasm class

The weak indicators that the candidate is a non-cell are:

I_4: A radius estimation for the nucleus cluster
I_5: A measure of nucleus cluster roundness
I_6: A contact index between the nucleus and the cytoplasm cluster
I_7: The area ratio between the nucleus and the cytoplasm cluster
I_8: A nucleus colour index

For each candidate, we compute the membership values to class non-cell $\boldsymbol{\mu_i}$ according to the strong indicators I_1, I_2 and I_3 and the weak indicators I_4 – I_8. The aggregate membership $\boldsymbol{\mu}$ is the weighted sum of the fuzzy memberships for all indicators. A candidate is deemed non-cell if the aggregate membership $\boldsymbol{\mu}$ exceeds threshold $\boldsymbol{T = 1}$.

The segmentation method has a number of adjustable parameters (the number of angles and distances used in re-dimensioning, the scaling factor for the fourth dimension of pixels, the type of distance for the adaptive k-means algorithm, merging rules, etc.) and one perspective would be to study the influence of these parameters on the overall classification. Furthermore, the number of indicators may be increased and finer classifications may be attempted. Thus, the automatic segmentation on the redimensioned polar transform shows potential to decrease the computational effort and outperform the approached based on Cartesian images.

2.2. Iso-Contour Description

Our cytologist Ciprian Tiganesteanu told us that when evaluating a cell, he would study carefully the 'granularity' of nuclei. We therefore decided to focus on iso-contours for nuclei description (**Fig. 4**), because they naturally outline the individual dark patches within a nucleus. It is pointed out, that this iso-contour description has nothing to do with active segmentation methods that merely start with an iso-contour: here iso-contours at different levels are taken without any further modification of them.

The first step toward that iso-contour description is to find the nucleus silhouette. For a patch containing a nucleus, the largest iso-contour of that patch is taken as the nucleus silhouette. In **Fig. 4**, the most outer iso-contour corresponds to that determined nucleus silhouette. As can be seen, this type of silhouette selection is not perfect: some silhouettes contain one or several protrusions. Three types of information are taken from such a localized nucleus: simple appearance, the structure of the silhouette and the structure of the inside.

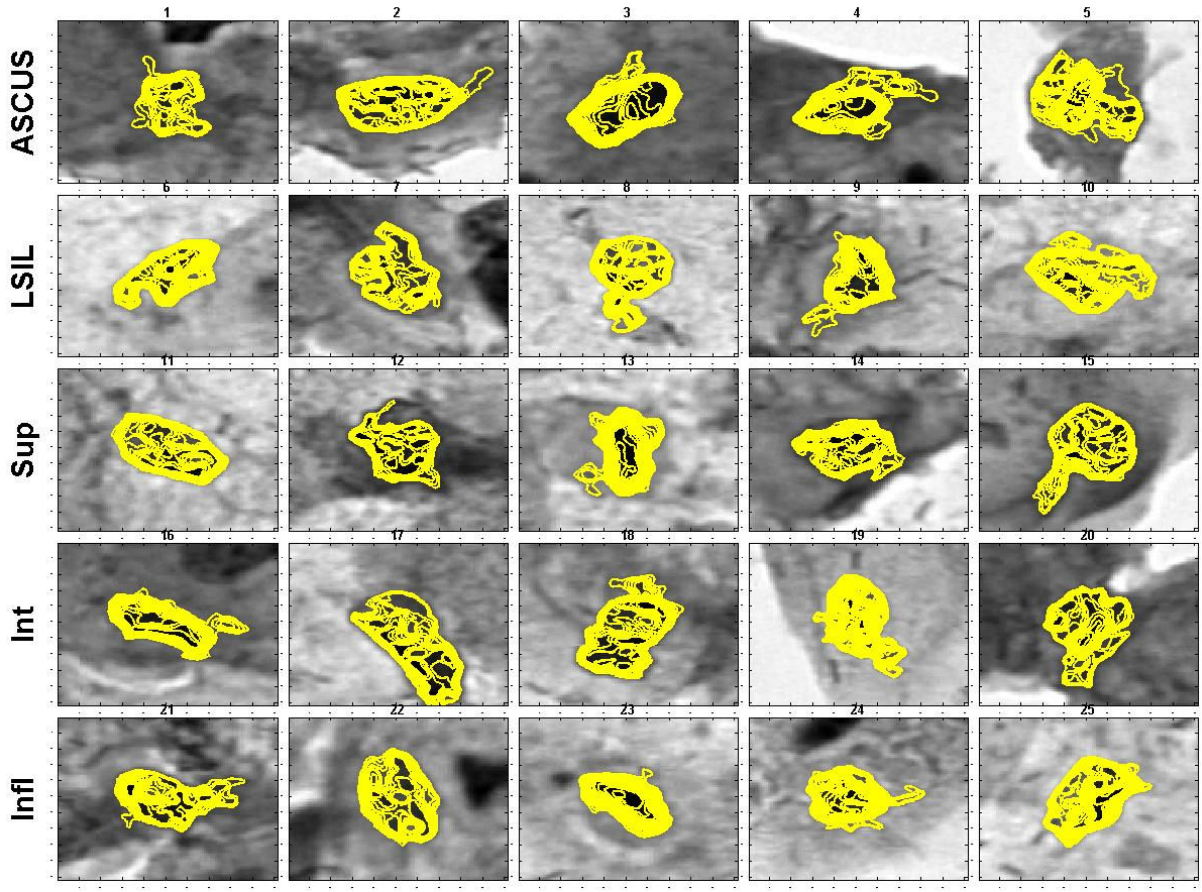


Figure 4. Examples of iso-contours for nuclei. The silhouette iso-contour is selected by observing a number of geometric conditions. With the iso-contours inside that selected silhouette, a feature description is generated that analyses the sizes and relations between those iso-contours.

Simple Appearance: Four attributes are determined: the nucleus area a, its mean intensity i_{mean}, the standard deviation of its pixel intensity values i_{std}, as well as the contrast, the range of intensity values i_{rng}.

Silhouette Structure: To describe the silhouette, the Fourier transform is applied to its radial signature. Studies on shape retrieval showed that this is the most efficient shape description with regard to the spatial and temporal complexity: shape descriptions that show a higher retrieval performance use excessive spatial and temporal complexity in comparison. More concretely, for a boundary $B(s)$ with arc length variable s, its radial signature $R(s)$ is determined, of which the first four (fast) Fourier descriptors are determined, f_1, f_2, f_3, f_4. More Fourier descriptors did not improve performance signicantly.

Inside Structure: The iso-contours inside the nucleus silhouette are determined, namely at a spacing value equal four, out of an intensity range of [0, 255]. That 'inside' count niso can vary between a few - typical for the small healthy nuclei - to several tens - typical for the bloated nuclei, see **Fig 5**; that count is therefore a parameter as well. For each inside iso-contour, its (average) radius R_i is calculated ($i = 1..n_{iso}$). To capture some structural aspects, two distributions are analyzed (right column in **Fig. 5**). One is the radius distribution, which is the sorted list of iso-contour radii R_i - in decreasing order - whereby the first value corresponds to the average radius of the nucleus silhouette, the last value corresponds to the

smallest iso-contour. The other distribution is the luminance level of the iso-contours Ii, whereby the ordering is in accordance with the radius distribution. Those two distributions can be characteristic for different nuclei types (compare right columns in **Fig 5**). The goal is therefore to characterize those distributions with a few parameters. This is done with help of the linear decay between the first and last point of the distribution, for which we measure the amount of 'deection' into the positive and negative range: if the decay occurs 'slower' than the linear decay, then it is a positive deection; if the decay occurs 'faster' then it is a negative deection. The amount of detection corresponds to the maximal distance between the linear decay and the distribution. For the luminance distribution I_i only a negative detection is possible, because the distribution is ordered according to the radius distribution R_i and the luminance decreases linearly due to sampling. For the radius distribution, both the positive and the negative deection is determined and normalized by the range of radii. In addition, we determine the minimum radius rmin = mini R(i) and the mean radius r_{mean}. In total there are 14 attributes and the following vector $\boldsymbol{f}$ is formed:

$$\boldsymbol{f} = [a,\ i_{mean},\ i_{std},\ i_{rng},\ f_1, f_2, f_3, f_4,\ r_{min},\ r_{mean},\ d_{R+},\ d_{R-},\ d_{L-},\ n_{iso}]$$

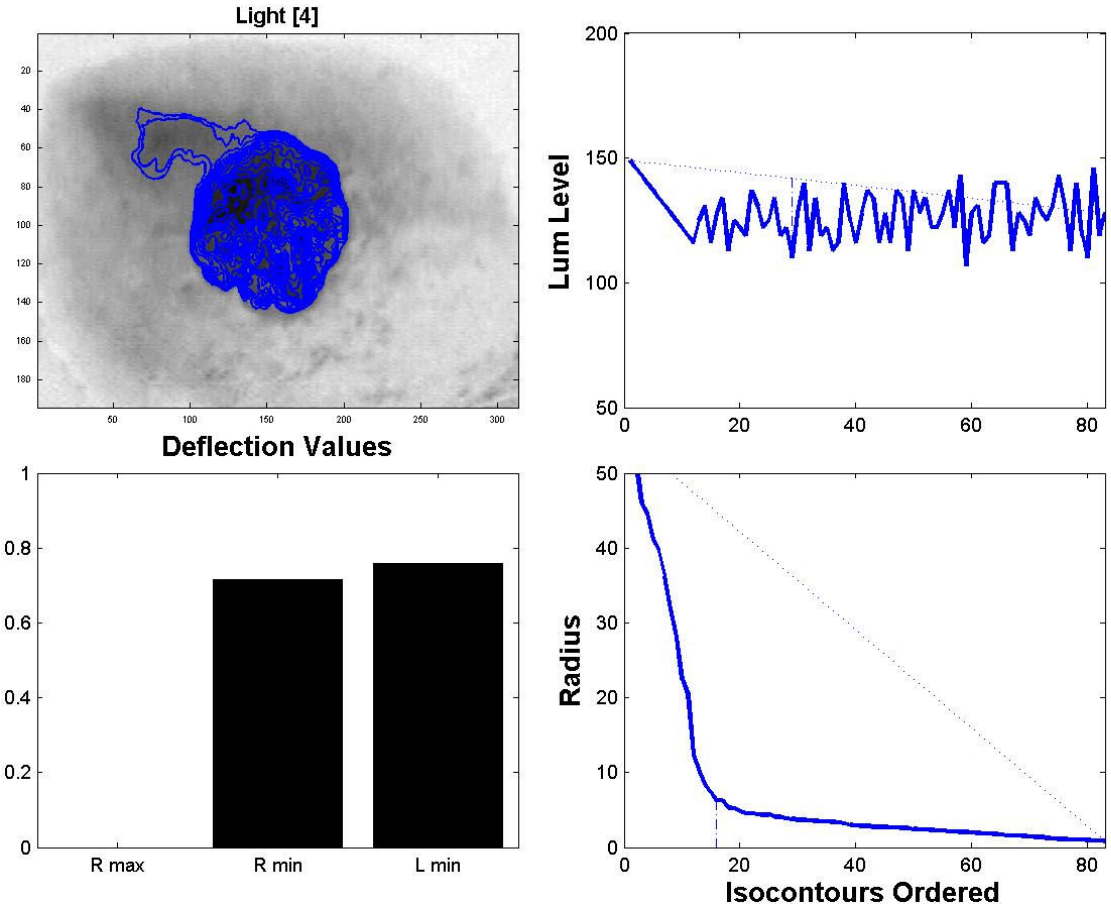


Figure 5. Iso-contours for a normal nucleus (from the Herlev database). **Upper right**: Luminance level (intensity) against iso-contours ordered by *R(i)*. The dotted line oblique connects the first and last value linearly. The vertical dashed line indicates the linear decay; the negative deflection is marked by a dashed-dotted line. **Lower right**: Radii of iso-contours, sorted in decreasing order. Only the positive deflection is clearly visible. **Lower left**: deflection values.

The degree of achievement: We have fully completed this objective as well. We have tried more approaches than any other group doing research on this subject.

3. Implementing appropriate cytological images classification methods

This is an important stage in obtaining the final integrated system. A possible challenge is dealing with real-life use-cases and obtaining good detection rates and low false positive rates. Novel approaches will be researched, so that error rates will be minimum.

We have tried three approaches: **a)** one based on the iso-contour description as developed under 2 (new feature extraction); **b)** one based on a polar-based image representation; **c)** and a third one that uses Deep Belief Networks (DBN), which is the most novel approach in the ComputerVision/MachineLearning community. Figure 6 summarizes most classification results.

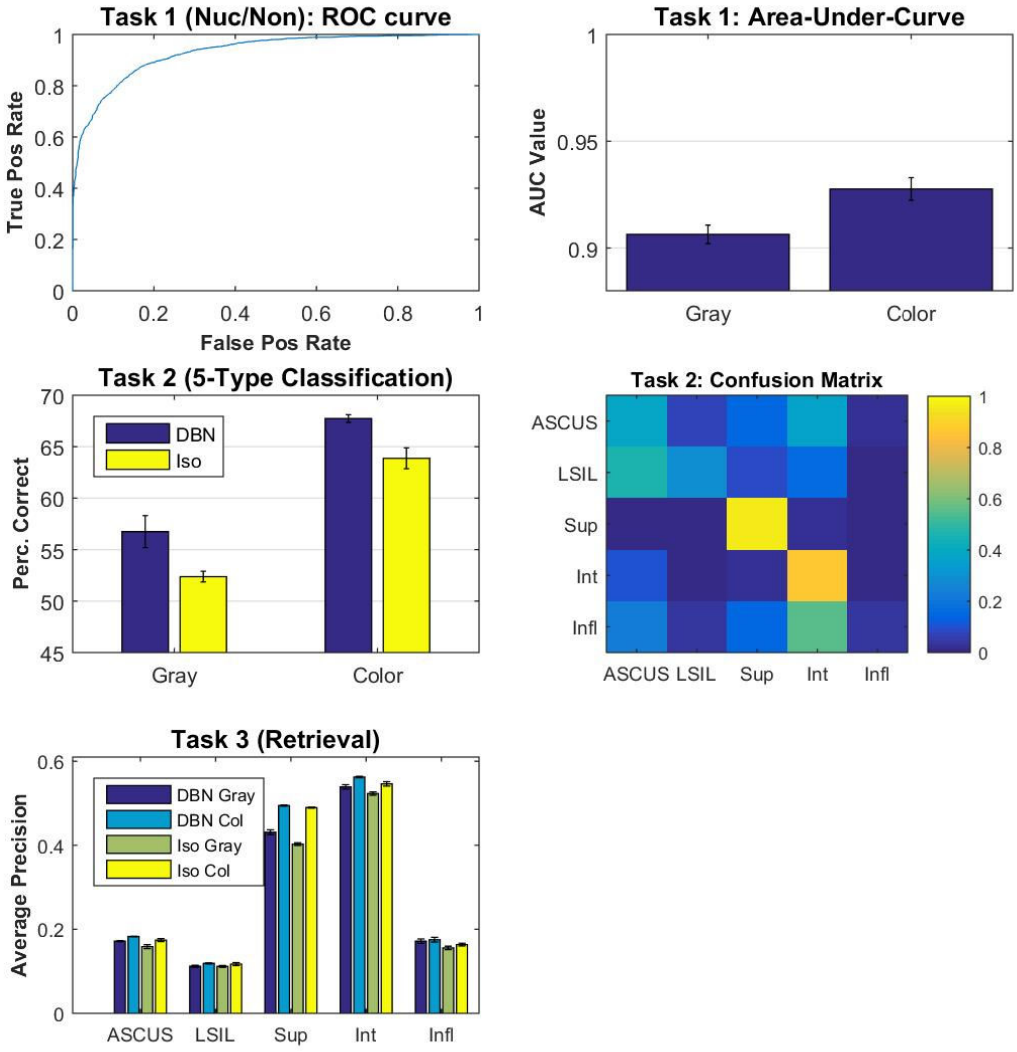


Figure 6: Classification and retrieval performance for the two tasks. **Pix**: pixels classified with a Belief Network; **Iso**: feature description based on iso-contours, classified with a Support Vector Machine. For retrieval, the posterior values of the classifiers are exploited - no actual classification decision is made; average precision is the area under the precision-recall curve.

We have performed the following three tasks:

- *Nucleus/Non-Nucleus Discrimination*: a binary classification between nucleus and non-nucleus, as just mentioned. For that purpose ca. 10'000 image patches were collected that contained structures easily mistaken as nuclei.

- *Five-Type Classification*: a classification into five different types of nuclei: we have collected three types of benign and two types of malign nuclei, as will be elaborated below.

- *Retrieval*: a retrieval task of the individual cell types: that retrieval is supposed to return the most likely affected cases to the doctor (or cytologist), because it cannot be expected that the five-type prediction accuracy (of task no. 2) will be perfect.

a) Iso-contour description: The feature vector $\boldsymbol{f}$ as introduced under 2.2 was classified with a Support-Vector-Machine.

b) Polar-based representation: The feature vector containing the 'indicators' as introduced under 2.1 was classified with various linear classifiers, but was not quite as competitive as our other two approaches (a and c) and we therefore do not mention it further.

c) Deep Belief Networks (DBN):

We use a four-layer Deep Belief Network (DBN) as described in [5], written in Matlab: an input layer that takes the raw image - the pixel values; two hidden layers; and an output layer, the number of classes. This type of network has been used to classify a number databases and achieved competing prediction accuracies amongst Deep Network approaches, often with only two hidden layers. The unit count of the input layer is now denoted as n_{inp} and will be 61 x 61 = 30721 pixels for instance if it is a gray scale image. We made the general observation that the nuclei classification accuracy is maximal, if the first hidden layer has a unit count that is three to four times as large as the input layer ninp; the second hidden layer has a unit count that is approximately one tenth of the second layer: n_{inp} x 10*n_{inp} x 4*n_{inp} x $n_{classes}$. Choosing more layers did not result in any performance gain. The batch size is 50; the learning rate is 0.2; the number of iterations 100 to 300; the other parameters are set as pointed out in. Interestingly however, the method of drop out did not improve but rather deteriorate performance and was therefore never used.

Evaluation Classification Accuracy:

- *Task 1: Nucleus/Non-Nucleus Discrimination*: In this task one discriminates between the 9405 nucleus patches and the 10743 non-nucleus patches. **Fig. 6** upper left displays the ROC curve for one of the five folds as obtained with a DBN; the upper right displays the average area-under-curve (AUC) value, for gray-scale and color patches. Classification accuracies for iso-features are not shown because they are significantly lower. Those AUC values are obtained with a patch size of 61 x 61 pixels, for which the network's unit count is 3'721 x 12k x 2k x 2 for gray-scale patches, and 11'163 x 30k x 3k x 2 for color patches. For smaller and larger patch sizes, 51 x 51 pixels or 81 x 81 pixels, the AUC values were significantly lower (not shown).

- *Task 2: Five-Type Classification*: Now the five cell types are discriminated (ASCUS, LSIL, etc.). For the DBN, the architecture and its parameters remain the same as for task no. 1. To classify the 14-dimensional feature description vectors *f*, a one-versus-all Support Vector Machine (SVM) is employed; the kernel function is a Gaussian function. Again, the DBN achieves higher prediction accuracy, but the feature classification also shows respectable prediction values (center left in **Fig. 6**). Color information again improves

accuracy and does so more than in task no. 1. The confusion matrix for both classifiers looks very similar and only the one for the DBN is shown (**Fig. 6**, center right).

- *Task 3: Retrieval*: A retrieval procedure is typically carried out with some sort of similarity measurement between pairs of items. If we intended to apply this principle to our case, this meant to either find a similarity measure for two pixel patches, or for its two vectors. As this is unlikely to produce best results, we directly exploit the power of classification algorithms: we use the exact same classifiers as above, but use the posterior values for retrieval, instead of choosing the class with maximal posterior value. More explicitly, instead of making a classification decision with the five posterior values per sample, the posterior values for *all* samples of *one* nucleus type are sorted in decreasing order; from that sorting one creates a precision-recall curve. As a performance measure, we take the commonly used area-under-curve (AUC) value, see lower left in **Fig. 6**.

The average precision value does however not tell us, if there are any hits amongst the first few retrievals. The retrieval of affected cases among the first few selected items is however important, because otherwise the doctor might as well analyze the entire image directly. We therefore verified for cancer types ASCUS and LSIL, that under the first 10 retrieval items at least one was present. This was always the case and is our **most significant result.**

The degree of achievement: We have fully completed this objective. We lack however a comparison to other approaches as we have used a much larger dataset than any other research group. But by using a DeepNeuralNetwork we have used the state-of-the-art classification approach and it is unlikely that any other method will perform better. We have shown that the use of iso-contours can be competitive, justifying our search for descriptors that correspond to the features that a cytologist looks for.

4. Testing and validation, optimizing the real-time diagnosis tools

The main challenge related to this specific objective is the optimisation of all the components so that the samples can be classified in real-time. The balance between harnessing the available processing power and the classification accuracy will be very important.

The most time consuming part of classification is the detection of potential nuclei sites. We have developed a method that is capable of processing a full microscopic image, five billion pixels, within a day - other methods required several days - if not even weeks. The majority of segmentation methods presently used in Pap smear images are very time consuming, as they often utilize computationally intensive propagation methods. A popular example are the level-set methods, which start with an iso-contour and some other image aspects and then gradually evolve to an optimal status. Apart from their issue of long computation duration, it is not clear how they perform on affected, large, bright nuclei - many studies have focused on healthy, small, dark nuclei. We therefore developed novel algorithms, namely the analysis of iso-contours. To obtain seed points, we perform blob-detection first (Section 4.1). Subsequently a local iso-contour analysis is performed at each seed point, which serves two purposes: one is to further eliminate seed points and to arrive at a final set of nuclei candidates (Section 4.2).

4.1 Seed Points (Blob Detection)

The process of seed detection is divided into the following three phases, which are illustrated in Fig.7:

- *Band-Pass Filtering and Thresholding*: The original image I_{orig} is band-pass filtered with a difference-of-Gaussian (DoG) and the output I_{dog} is thresholded to obtain a black-white image B_{sink} with regions corresponding to potential sites of nuclei, called sink image now. In the upper right in Fig. 7, the on-pixels in B_{sink} are replaced by the corresponding gray-scale values of I_{orig} for purpose of illustration.

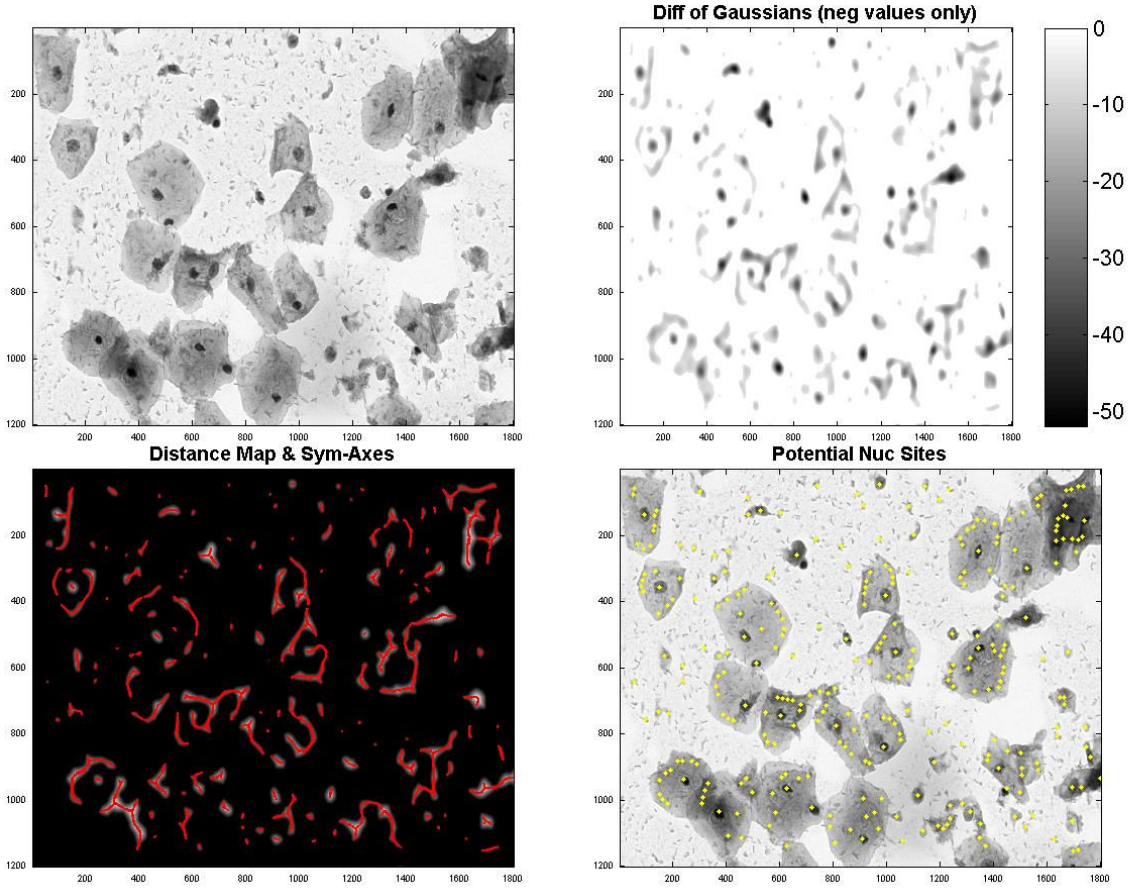


Figure 7. Detecting potential nuclei sites. **Upper left**: a typical Pap smear image (original image I_{orig}). **Upper right**: sink image B_{sink} with regions values taken from I_{dog} (negative values of the DoG-filtered I_{orig}; positive values set to 0). **Lower left**: distance map and symmetric axes (red in pdf file) of B_{sink}. **Lower right**: seed points, the minima of the luminance along symmetric axes (yellow in pdf file).

- *Distance Transform and Symmetric Axes*: The distance transform is applied to the sink image B_{sink}, followed by determining its symmetric axes (sym-axes) using a simple ridge detection algorithm, see lower left in **Fig. 7**. This phase results in a list S_i of sym-axes segments, each one containing a list of x- and y-coordinates and a symmetric distance value.

- *Luminance-Minima Detection Along Symmetric Axes*: The maxima in the symmetric axes S_i are determined. Their corresponding gray-scale value in I_{orig} may however not correspond to the local minimum in I_{orig}. Thus we determine the minimum intensity in a local neighborhood around those maxima, which then are taken as seed points. Seed points are shown in the lower right graph of **Fig. 7**.

4.2 Iso-contours and Candidate Selection

The iso-contour analysis at seed points occurs identically to our previously introduced analysis (see Section 2.2). More explicitly, at each seed point a small patch is extracted and the iso-contour analysis is applied and further candidates are discarded resulting in a final set of candidate patches. Those candidates - and their associated local iso-contours - are shown in Fig. 4. For each such 'cluster' of iso-contours, the one with the smallest elongation is selected as the nucleus silhouette. The selected nucleus silhouette and its inside is then described as in the section above.

Our nuclei recall values exceed those of other studies proving that our method of nuclei detection with isocontours is powerful even in the presence of strongly overlapping cells as demonstrated on the Plissiti database and on our own. Our nuclei precision values are rather low however - it requires further effort to minimize the number of false alarms.

The degree of achievement: We have fully completed this objective. A real-time diagnosis as originally envisioned is not quite possible with current technology, but we have developed the fastest nuclei-detection algorithm: it can process an image within hours. There is no other method to date, that can analyse a 4-billion pixel image in reasonable time.

5. Publishing results

The work carried out during the project has made the object of the following publications:

- Christoph Rasche, Ciprian Ţigăneşteanu, Mihai Neghină, Alina Sultana: Cervical Nuclei Classification: Feature Engineering Versus Deep Belief Network, Annual Conference on Medical Image Understanding and Analysis (MIUA2017), pp. 874-885, Edinburgh, Scotland.

- Neghina M., Rasche C., Ciuc M., Sultana A., Automatic detection of cervical cells in Pap-smear images using polar transform and k-means segmentation, The sixth international conference on image processing theory, tools and applications, IPTA 2016, Oulu, Finland

- C. Toca, C. Patrascu, M. Ciuc, AutoMarkov DNNs for Object Classification, Proc. of 23rd International Conference on Pattern Recognition, Cancun, Maxic, Dec. 2016

- M. Ivanovici, N. Richard, ENTROPY VERSUS FRACTAL COMPLEXITY FOR COMPUTERGENERATED COLOUR FRACTAL IMAGES, PROCEEDINGS of the 4th CIE Expert Symposium on Colour and Visual Appearance, 6 – 7 September 2016, Prague, pp. 432-437, 2016

- Automatic Pap Smear Nuclei Detection Using Mean-Shift and Region Growing, S. Oprisescu, T. Radulescu, A. Sultana, C. Rasche, M. Ciuc, 12-th International Symposium on Signals, Circuits and Systems (ISSCS), Iasi, Romania, 2015.

- Analysis of Pap Smear Images with Iso-, Edge-Contours, C. Rasche, S. Oprisescu, A. Sultana, T. Radulescu, International Conference on Intelligent Computer Communication and Processing (ICCP), Cluj, Romania, 2015.

- Fast Probabilistic Pseudo-Morphology for Noise Reduction in Color Images, R.-M. Coliban, M. Ivanovici, I. Szekely, 9th International Conference Interdisciplinarity in Engineering, INTER-ENG, Tirgu-Mures, Romania, 2015.

Also, a paper which sums up is currently under preparation, and will be submitted shortly to a journal on medical imaging.

The impact of the obtained results, emphasizing the most significant result obtained.

We present the first working system, that fully automatically analyzes entire microscopic images, within reasonable time and without the assistance of a doctor. Existing studies on Pap-smear diagnosis have typically focused on individual tasks of the diagnosis, namely detection or classification separately, thus not demonstrating their use in a fully automatic system.

Our most significant result is that our retrieval process always finds malign cases within the first 10 items, second-last paragraph under Section 3. That means, a doctor can be aided in his analysis. A Pap-smear image can contain up to hundreds of thousands of cells - of which only few are affected; it can take a cytologist an hour to spot any of those affected cells. In that real-case scenario, our system's retrieval success might be ten times worse, meaning only one affected cell would be found under the first one hundred retrieved items: but even that would still help the cytologist to find malign cases much faster than by visual search of the entire image. This potential speed-up needs to be confirmed on novel cases and not on those that were used for training the system. In order to avoid the loss of nuclei during the nucleus/non-nucleus discrimination stage, one can bias the decision to show a higher recall at the cost of lower precision, a bias shift which should not deteriorate the retrieval performance substantially.