Removing Artefacts from Microscopic Images of Cytological Smears. A Shape-Based Approach.

Dariusz Pietka, Annamonika Dulewicz, Pawel Jaszczak

Image Processing Systems Laboratory
Biomedical Information Processing Methods Department
Institute of Biocybernetics and Biomedical Engineering, PAS
darek@ibib.waw.pl

Summary. Most of reports on computer supported cytological investigations focus on searching for objective, quantitative descriptors enabling an automated system to distinguish between "normal" and "pathological" objects, usually cells or their organelles. A great number of sophisticated tools have been developed and reported. However, few reports may be found concerning the problem of detecting artefacts in cytological smears and reducing their influence on the overall system performance. On the other hand, the problem is crucial for the whole system setup and if not properly solved may spoil any attempts to implement the system in practice. The paper addresses this neglected problem trying to point out some general rules and procedures that should be followed to reject artefacts from automatic cytological analysis.

1. Objective of the work

Most projects related to clinical implementation of computerised image processing in cytology face the common problem of distinguishing between artefacts and objects of interest which should be measured and analysed. Let us imagine automated detection of early cancer cells in a microscopic smear. Since most algorithms of cancer cell identification rely on some kind of abnormality detection, artefacts left in a sample would generate too many undesired, false-positive alarms, making such a system impractical. Thus, efficient detection and removal of artefacts from later analysis is crucial for the entire system setup and its overall performance [1]. Regardless of the microscopic enlargement used, considerable number of artefacts will always be present in collected images (Fig.1). In the next section we shall explain in more detail what is understood by the term "artefact" in our work. **Generally, these are undesired objects or phenomena influencing the appearance of a smear and obstructing or even preventing proper analysis of important factors of the cytological sample.** As such, artefacts or their influence on the experiment should be avoided. Objective of this work was to

find and to point out some general rules and methods allowing efficient rejection of artefacts in the process of automatic analysis of cytological material. This is a trial to make a step towards breaking through one of the main obstacles in the practical implementation of computer-aided cytological screening. To make our deliberations more useful we illustrate them with the clinical material of Feulgen-stained epithelial cells from urinary bladder obtained by means of *bladder washing* technique in Medical University of Nijmegen, Holland. The material is used in our collective investigations on non-invasive, computer-aided detection of bladder cancer.

2. Artefacts in cytological smears

Let us define, for the purpose of this work, the objective of cytological investigations as measuring different morphological parameters of isolated nuclei found in a Feulgen-stained smear. The results of the measuring stage are usually used for successive statistical analysis and discrimination, but this is out of the scope of the paper.

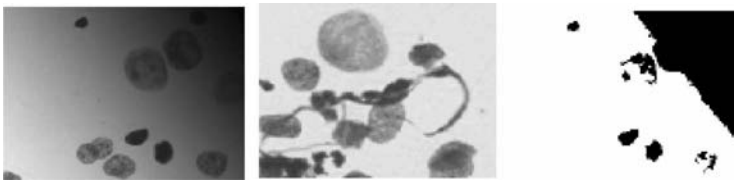

Fig. 1. Artefacts caused by different factors: non-uniform illumination of the scene, non-interesting objects appearing in a smear and an inadequate thresholding applied to an image.

(I) To get reliable measuring results it is expected that the visual appearance of nuclei are not affected by changeable physical factors of a system but only reflect their important biomedical features. To fulfil this requirement appropriate image correction and normalization algorithms should be applied. These are not covered in this paper but were the subject of our earlier works [5].
(II) It is necessary for correctness of the results that measured objects are isolated nuclei and only nuclei, not granulocytes or other biological or artificial objects encountered in a sample. Our method of shape-based discrimination between these interesting and non-interesting objects will be actually one of the main topics of this paper.
(III) As we are going to utilise shape features it is obvious that the objects should be properly extracted from the background. Otherwise, image processing algorithms used for objects extraction may themselves become a

source of artefacts. An efficient adaptive thresholding procedure built in the course of our investigations will be presented as it is crucial for preserving shapes of extracted objects.

To get an impression of the presence of artefacts in a smear a test was performed. 60 randomly chosen images from three smears of different persons were visually inspected by an expert who outlined each nucleus manually. Isolated nuclei size distribution was used to set up lower and upper limits for possible sizes of a single nucleus. Then, a rule for a human expert was established to classify artefacts in images. The order of artefacts identification steps is important as it simulates the way this algorithm is going to be implemented in a computer system. The script for checking off artefacts in a smear image is as follows:
1) mark all objects on the edge of an image as artefacts, then for the rest of objects
2) mark all objects smaller than nucleus lower size limit as artefacts, then for the rest of objects
3) mark all objects larger than nucleus upper size limit as artefacts, then for the rest of objects
4) mark all overlapping objects as artefacts, then for the rest of objects
5) mark each object that is not nucleus as artefact

By applying this artefact identification rule to each object or aggregate of objects, the isolated nuclei are found as those objects which are left unmarked. The overall results of this initial, interactively conducted experiment are very interesting. They are presented in the table (Fig.2). Let us summarize:
- 57% from all of 472 objects encountered in this particular material proved to be artefacts from the point of view of our study (not isolated nuclei),
- most of them (~41 %), three upper classes in the table, may be detected easily by means of simple and fast computer algorithms,
- the actual problem are overlapping objects (~16%) of overall area not exceeding acceptable sizes of a single nucleus.

Conclusions are straightforward. When implementing an automated cytological screening system, most efforts should be directed to detect overlapping objects and eliminate them from successive morphological and statistical analysis.

3. Adaptive thresholding of a smear image

As was stated earlier, to utilise shape features for discrimination it is a necessary condition that the objects were properly extracted from the background. Many of our images (dark objects on bright background) are characterised by a simple, bi-modal histograms. It is relatively easy to design an algorithm for

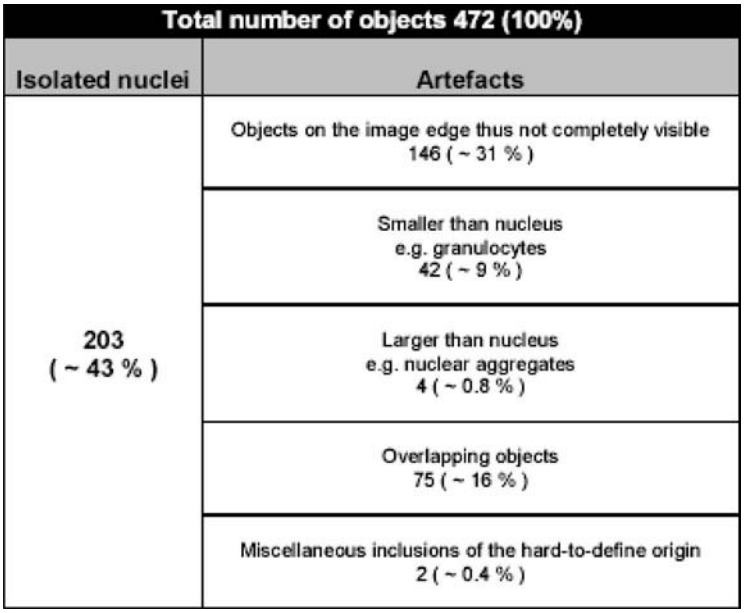

Fig. 2. Statistical summary of objects found in the smear.

finding proper global threshold for such cases. Unfortunately, real-life smears are not always so easy to analyze and global thresholding may lead to critical processing errors (see Fig.2.3). On the contrary, adaptive thresholding selects an individual threshold for each pixel based on the range of intensity values in its local neighborhood, allowing for thresholding of an image whose histogram doesn't contain distinctive peaks.

A general definition of a dynamic threshold $\mathbf{t_{xy}}$ that we are going to use can be written in as follows:

$$\mathbf{t_{xy}} = \mathrm{T}\ [\ \mathrm{x},\ \mathrm{y},\ \mathrm{f(x,y)},\ \mathrm{p(x,y)}\ ]$$

where $f(x,y)$ is the light intensity of point (x,y) in the original image, and $p(x,y)$ is some local property of this point. Several adaptive thresholding algorithms have been tested and the best one, adopted from Intel's Picture Processing Library (*Open Source licence*) [2], finally chosen. Let *f(x,y)* be the input image. For every pixel *(x,y)* the mean $\mathbf{m_{xy}}$ and a measure of intensity variations in its neighborhood $\mathbf{v_{xy}}$ are calculated as follows:

where $\mathbf{p}$ is the half-size of pixel neighborhood. Local threshold for pixel *(x,y)* is computed as follows:

where $\mathbf{v_{min}}$ is some application specific minimum variance value.

To find optimal parameters for the algorithm an experiment was performed. Its idea is illustrated in Fig.3. An original and artificially degraded

$$m_{xy} = 1/(2p+1)^2 \sum_{s=-p}^{p} \sum_{t=-p}^{p} f(x+s,y+t)$$

$$v_{xy} = 1/(2p+1)^2 \sum_{s=-p}^{p} \sum_{t=-p}^{p} | f(x+s,y+t) - m_{xy} |$$

$$t_{xy} = m_{xy} + v_{xy} \quad \text{for } v_{xy} > v_{min}$$

and

$$t_{xy} = t_{xy-1} \quad \text{for } v_{xy} \leq v_{min}$$

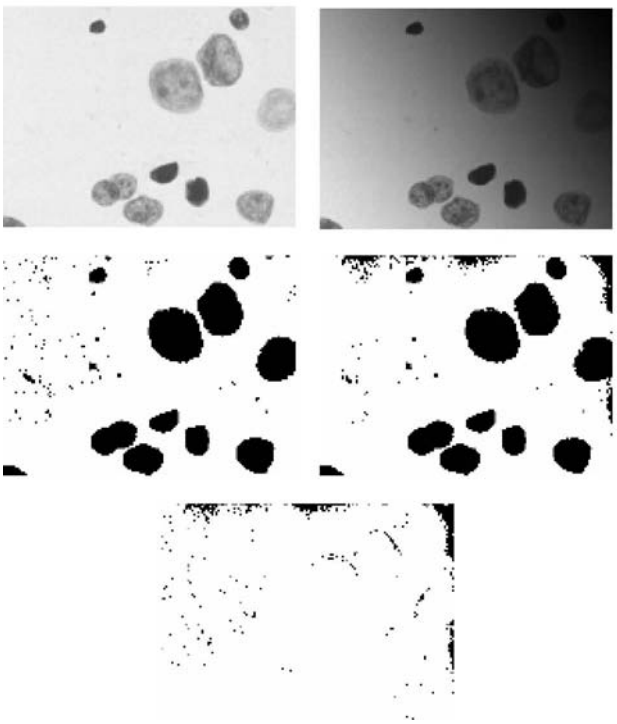

Fig. 3. Input images, thresholded binary images and their difference.

images were thresholded and subtracted to evaluate differences between them, especially differences between extracted regions of objects. Applying different values of parameters **p** and $\mathbf{v_{min}}$ for 30 randomly chosen images the best pair was found: **p=13,** $\mathbf{v_{min}}$**=4.** The range of measured differences between extracted areas of the same objects not exceeded 4% of their sizes. Although indirect, it is a rather strong confirmation that our adaptive segmentation algorithm is strongly independent of background variations and preserves the shape of objects. The results of object extraction by means of adaptive, local thresholding give us a good base for successive shape-based analysis.

4. Shape-based artefacts detection

In this section we put a short description of the methods used to cope with those 16% of artefacts which can't be detected by means of simple methods based on object's position (on the edge) and size. As we have demonstrated experimentally, most of those "hard cases" are overlapping objects. After the im-

Fig. 4. Processing steps of the image with overlapping objects.

age processing steps and adaptive thresholding procedure we get the extracted contours in Fig.4 (object on the edge was also automatically removed).
An initial attempts using simple scalar shape descriptors (e.g. eccentricity, elongatedness, rectangularity, compactness [6]), although supported by multivariate discriminant analysis, were not promising, so abandoned. Nevertheless, visual inspection of objects in many thresholded images suggested applying of some advanced shape descriptors. Experience of the laboratory staff in two-dimensional spectral analysis directed us to *Elliptic Fourier Descriptors (EFD)* for shape-based discrimination between objects. Basically, our method does not make an *a priori* choice of the relevant features; it rather tries to automatically associate an importance degree. For that purpose the *Principal Component Analysis (PCA)* is used. The Fourier descriptors were defined in such a way, that they remain translation, rotation and scaling invariant [3,4]. They identify a shape, independent of its position, orientation or size. After Fourier transformation of chain-coded contours we get the feature vectors consisting of Fourier coefficients (20 harmonics are used). After that, the

PCA was applied to capture most of the essential relations from the data, creating new factors *(Principal Components)*. Five of them were supplied for final discriminant analysis in order to obtain satisfactory shapes separation results.

5. Experiments and results

The biological material used in this work consisted of urinary bladder epithelial cells. It was obtained by means of *bladder washing*. Only nuclei were visible in a smear. Image acquisition and processing were performed in a computer system equipped with a frame-grabber, CCD camera and moving stage optical microscope. For proper shape description of objects relatively large optical magnification had to be used (60x). Although results of this stage of the work were not intended to be implemented in a real-life screening system, the same material, consisting of 60 images (472 objects), as described earlier in section 2, was used to verify overall effects of the work. It may be very informative for final conclusions to compare an expert screening data in the table from section 2 (Fig.2) with the results of fully automated processing, equipped with advanced thresholding and shape-based object identification (Fig.5). What is really interesting and worth discussion it is ability of our shape-based analysis to identify artefacts in the form of overlapping objects. Assuming that the real number of overlaps in examined material was 75 (expert evaluation) the total *sensitivity* of advanced processing tools that were applied in the work may be reported to be high and equals to 92%. In other words, only 8% of overlaps were not identified as artefacts and were left as isolated nuclei for successive diagnostic steps. Examples of missed occlusions are shown in Fig.6. They are usually two or three clustered nuclei or granulocytes forming quite regular, nuclear shapes. Since we have concentrated on detecting and removing artefacts to reduce false-positive signals, *sensitivity* of the method was the most important parameter. Nevertheless, a question about nuclei misclassified as artefacts must not be left without a short discussion. Pathology, to be diagnosed, must be evident at some level. Therefore, it does not seem critical if some insignificant amount of nuclei are classified as artefacts and removed from analysis. However, to confirm that misclassification actually concerns inessential number of nuclei, *specificity* of the method has to be evaluated. 53, from the total number of 203 isolated nuclei were chosen randomly and went through the shape-based artefacts identification procedure. Three of them were identified as artefacts (Fig.7). An evaluation of specificity yields some 94%, which may be accepted as it implies insignificant number of misclassifications.

Total number of objects 472 (100%)	
Isolated nuclei	**Artefacts**
203 (~ 43 %) 209 (~ 44 %)	Objects on the image edge thus not completely visible 146 (~ 31 %) 146 (~ 31 %)
	Smaller than nucleus e.g. granulocytes 42 (~ 9 %) 42 (~ 9 %)
	Larger than nucleus e.g. nuclear aggregates 4 (~ 0.8 %) 4 (~ 0.8 %)
	Overlapping objects 75 (~ 16 %) 69 (~ 15 %)
	Miscellaneous inclusions of the hard-to-define origin 2 (~ 0.4 %) 2 (~ 0.4 %)

Fig. 5. Overall results of the automatic, computer artefacts identification (larger numbers) and comparison with the interactive, human-expert results (smaller numbers).

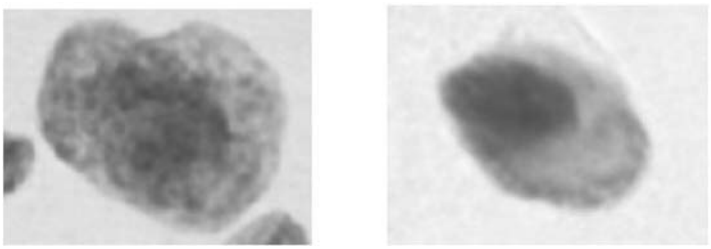

Fig. 6. Occlusions of two cytological objects forming regular, nuclear shapes.

Fig. 7. Nuclei misclassified as artefacts by the shape-based analysis.

6. Conclusions

It was demonstrated that it is possible to efficiently detect artefacts in cytological smears by means of advanced shape analysis of properly extracted objects. Well known and powerful tools of *Fourier Shape Descriptors* and *Principal Component Analysis* have shown to be very useful in solving the problem. Sensitivity of our artefacts detection method yields some 92% and it seems we have reached, or are very close to, limits of shape-based approach to artefacts detection. The cases likely to be misclassified are usually clustered nuclei or granulocytes forming quite regular shapes. It is easy for the human visual system to recognize most of such occlusions, but it is extremely difficult to translate physiological algorithms into machine procedures. Fortunately, "difficult" does not imply "impossible". Detection of those "hardest cases" may be a challenge and direction for future work.

References

1. I. Al and J.S. Ploem, 1979, ŞDetection of suspicious cells and rejection of artefacts in cervical cytology using the Leyden Television Analysis SystemŤ, Journal of Histochemistry and Cytochemistry, Volume 27, Issue 1, pp. 629-634
2. INTEL Corporation, 2004, ŞOpen Source Computer VisionLibrary - OpenCVŤ, http://www.intel.com/research/mrl/research/opencv/
3. Iwata, H. and Y. Ukai, 2004, ŞSHAPE: A computer program package for quantitative evaluation of biological shapes based on elliptic Fourier descriptorsŤ, Journal of Heredity, in press
4. Wallace, T. P. and Wintz, P. A., 1980, An Efficient Three-Dimensional Aircraft Recognition Algorithm Using Normalised Fourier DescriptorsŤ, Computer Graphics and Image Processing, Vol. 13, 99-106
5. Dulewicz A., Pietka D., Jaszczak P., Nechay A., Sawicki W., Ko?mi?ska E., Borkowski A., 1998, Computer Analysis of Epithelium Cell Nuclei of Urinary Bladder for Cancer DetectionŤ, VIII Mediterranean Conference on Medical and Biological Engineering & Computing. MEDICON Ś98, Proceedings of the conference, Limassol, 14-17, June
6. M.Sonka, V.Hlavac, R.Boyle, 1998, ŞImage Processing, Analysis and Machine VisionŤ, PWS Brooks & Cole Publishing