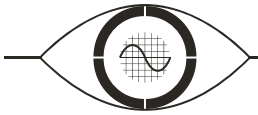


Lehrstuhl für Bildverarbeitung
Institute of Imaging & Computer Vision

Automated Classification of Analysis and Reference Cells in Microscopic Images for Cancer Diagnostics

Timna Esther Schneider
Institute of Imaging and Computer Vision
RWTH Aachen University, 52056 Aachen, Germany
tel: +49 241 80 27860, fax: +49 241 80 22200
web: www.lfb.rwth-aachen.de

in: Proceedings of the 11th International Student Conference on Electrical Engineering POSTER 2007. See also BibTeX entry below.

BibTeX:

```
@inproceedings{Sch07c,
  author = {Timna Esther Schneider},
  title = {Automated Classification of Analysis and Reference Cells
in Microscopic Images for Cancer Diagnostics},
  booktitle = {Proceedings of the 11th International Student Conference
on Electrical Engineering POSTER 2007},
  address = {Prague},
  year = {2007},
}
```


document created on: May 24, 2007
created from file: poster2.tex
cover page automatically created with CoverPage.sty
(available at your favourite CTAN mirror)

RWTH AACHEN
RHEINISCH-WESTFÄLISCHE TECHNISCHE HOCHSCHULE AACHEN

Automated Classification of Analysis and Reference Cells in Microscopic Images for Cancer Diagnostics

Timna Esther SCHNEIDER[1]

[1]Institute of Imaging and Computer Vision, RWTH Aachen University, Sommerfeldstrasse 24, 52074 Aachen, Germany

Timna.Schneider@lfb.rwth-aachen.de

Abstract. *To get the best possible chance of healing, cancer has to be detected as early as possible. As cancer starts within a single cell, cytopathological methods offer the chance of early detection. One such method is the standardized DNA image cytometry, for which the diagnostically relevant cells have to be searched for manually within a specimen. As this is a time-consuming process, a preselection of diagnostically relvant cells has to be performed automatically. For specimens of oral mucosa this means to distinguish between truly healthy epithelial cells and possibly cancerous epithelial cells.*

Based on cell images from a brightfield light microscope a set of morphological and textural features were implemented. To identify highly distinctive feature subsets the sequential forward floating search method is used. For these feature sets k-nearest neighbor and fuzzy k-nearest neighbor classifiers as well as support vector machines were trained. On a validation set of 400 *cells it could be shown that healthy and possibly cancerous cells can be distinguished at overall rates above 95.4% for different classifiers, enabling us to choose the support vector machine with a set of two features only as the classifier with the lowest computational costs.*

Keywords

cells, cytopathology, feature extraction, classification, Fuzzy k-nearest neighbor, support vector machine

1. INTRODUCTION

One of the most often reasons of death within industrial countries is cancer. So as to have the best possible chance of healing, cancer has to be detected and treated as early as possible. As cancer starts within a single cell, many types of cancer can be detected very early and from already marginal changes within single cells using cytopathological methods. These cell specimens can be obtained easily and painlessly, e.g. with tiny brushes for smears of the oral mucosa.

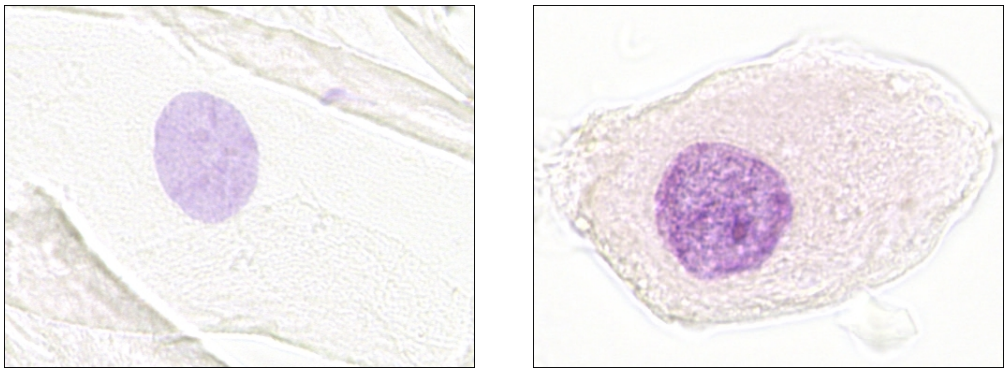
One cytopathological diagnostic method is DNA image cytometry (DNA-ICM). For this the cells are stained stoichometrically according to Feulgen to visualize the DNA content within the nuclei. Images of the nuclei are captured with a camera mounted on a brightfield light microscope.

Since DNA-ICM needs the DNA value of each cell, the integral optical densities of the nuclei are computed. To get the DNA value from these integral optical densities, a set of about 30 healthy, non-proliferating cells, called *reference cells*, has to be selected. Then the DNA values of nuclei are computed from the ratio of their integral optical densities to the mean density of the reference cells. Subsequently the diagnosis is performed based on the histogram of the DNA values of diagnostically relevant *analysis cells*. For oral mucosa these are epithelial cells with noticeable changed morphology or texture of their nuclei, cells that therefore are suspicious of cancer.

According to the guidelines for the DNA-ICM [1] a cytopathologist has to find and select manually about 30 reference cells and about 300 analysis cells. This is a time-consuming task, wherein the expert should therefore be assisted by an automated preselection of cells. A machine for fully automated screening of cervical smears using DNA-ICM is already available. This machine searches for cells with an abnormal high DNA content, by measuring the DNA value of all existing cells [2]. But as these cells are rare and cancer starts already with small changes of DNA content, the same high sensitivity and specificity as with the standardized interactive DNA-ICM [1] can not be achieved.

The approach in this paper therefore aims at an algorithmic implementation of such expert knowledge for specimens of the oral mucosa, i.e., an automatic discrimination of healthy epithelial nuclei as reference cells and non healthy epithelial nuclei as analysis cells.

The paper is organized as follows: In section 2 the database of cells from oral smears and the imaging modality is described. Features characterizing properties of epithelial reference and analysis cells, the process of reducing the whole feature set to a subset of features with optimal discrimating power and different classification algorithms to classify the cells automatically is presented in section 3. The results of these algorithms computed on the dataset are pre-



(a) (b)

Fig.1. Two examples of Feulgen stained epithelial cells. On the left a healthy cell is shown, on the right an analysis cell. The DNA within the nuclei is stained with Feulgen dye, whereas the cytoplasm surrounding is not stained.

sented in section 4, showing that epithelial reference and analysis cells from 15 specimens of the oral mucosa can be discrimated with different classification methods at total classification rates of above 95.4%. The paper ends with analyzing the classification results and an outlook.

2. MATERIAL

Our dataset is based on 15 Feulgen stained specimens from the oral mucosa, with seven specimens without cancer, three specimens with inflammation but without cancer, and five specimens containing cancer cells. From these specimens images have been acquired with a brightfield light microscope and a $63\times$ oil immersion objective (NA 1.32) and a three chip CCD camera with a resulting resolution of $\Delta x \approx 0.1\mu m$.

Within these specimens an experienced cytopathologist classified 950 reference cells from the specimens without cancer cells, and 748 analysis cells from specimens with cancer cells. For these cells the contours of the nuclei are given as chaincodes. Figure 1 shows two example cells.

3. METHODS

To be able to distinguish automatically between analysis and reference epithelial cells, a set of potentially relevant features has been implemented. To miminize the risk of overfitting during training of a classifier, a feature selection method is performend that can be combined with different objective functions to select good discriminative feature subsets. To solve the classification task, different classification algorithms can be chosen. Each of them has to be trained on a training set of cells, whereas the classification rate is calculated on an independent cell set (validation set). The process of feature selection and classifier generation is scetched in figure 2, the algorithms used within each step are described in the following.

3.1. FEATURES

Performing a manual discrimination between analysis and reference cells the cytopathologists consider geometrical criteria like area of the nuclei, shape, as well as textural characteristics of the chromatin pattern and the overall amount of stain related to the size of a nucleus.

To cover these criteria, several morphological features are used, as they are described in [3] and [4]. These comprise area, perimeter, form factor, Fourier descriptor energies, and, further on, translation, rotation and scale invariant features computed as combinations of the central 2D polynomial moments for the nuclear mask.

As textural features these moment based features are calculated for a density estimation of the chromatin (extincion image) and an edge image of the chromatin distribution. The extinction image is calculated on the green channel of the original RGB image, and the edge image is the difference of the extinction image and its median filtered version. These features are supplemented by histogram features of the topological gradient [3] and particle oriented features homogeneity, granularity and distribution of the chromatin [5]. See figure 3 for examples of the image transformations. This sums up to 203 features describing morphology of the nuclei and chromatin texture.

3.2. FEATURE SELECTION

Testing the separability performance of each possible feature subset is computationally too expensive for large feature sets. So the parameter-free search method sequential forward floating selection [6] is implemented to identify fea-

basic feature set	classification (a;r;o)	objective function	number of features	number of neighbors
morphology	75.5; 92.3; 85.2	B	14	3
chromatin green	92.5; 95.2; 94.1	B	14	1
chromatin extinction	90.2; 96.5; 93.8	B	19	3
chromatin+moments	90.7; 98.3; 95.1	m	2	4
all features	90.7; 98.3; 95.1	m	2	4

Tab.1. Results of the feature selection for the best overall classification result of each basic feature set. The classification results (a;r;o) in percentages for the (a)nalysis, the (r)eference cells, and the (o)verall rate are given for each of the basic sets. Additionally the used objective function ((m)utual information, (F)ishers' criterion, or (B)hattacharyya distance), the number of features and the number of neighbors of the kNN are noted. The computed optimal feature sets of chromatin+moments and all features are identical.

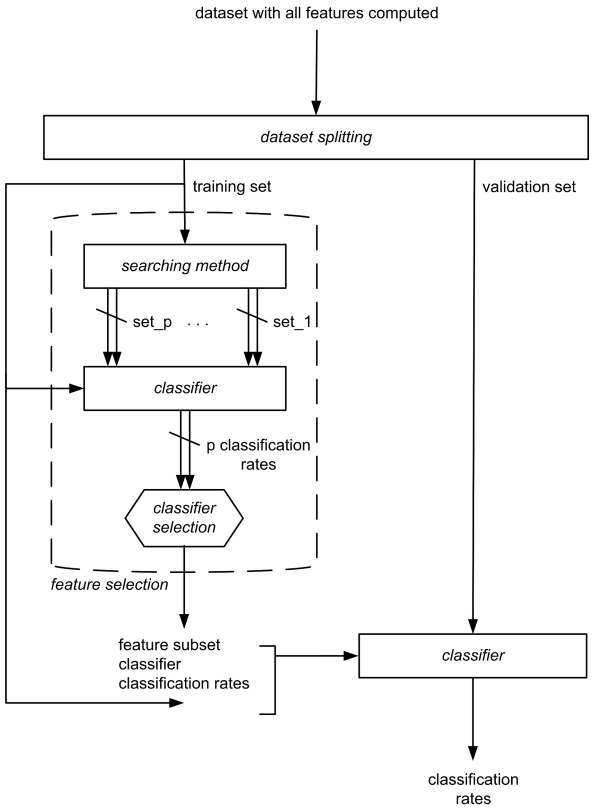


Fig. 2. Algorithmic workflow of feature subset selection and classifier generation. The different sized feature subsets resulting from the searching method are denoted as *set_1* to *set_p*, the output of classifier training and training set classification as *p classification rates*.

ture subsets with a good separability performance for our classification task. To rate the separability performance three different objective functions can be chosen. Based on the assumption of normally distributed data these are Fisher's criterion [7], and the Bhattacharyya distance [7], which is adapted to the special case of two classes only. Using the likelihood of feature vectors instead of a density model, Mutual information [8] can be selected as the third objective function.

3.3. CLASSIFIER

To carry out the classification task three classification algorithms are applicable. These have been selected due to the reproducibility of the classification results, as well as the possibly changing behavior of the distribution of the data during the feature selection. For the different subsets of features during this selection the distribution of the data may fit different distribution models, including non-gaussian, multimodal ones. Furthermore for higher dimensions of the feature space in generally the data is sparsely distributed and less compact, which leads to the use of non-parametric classifiers to be more general. As non-parametric classifiers that provide comprehensible decisions the kNN algorithm and the Fuzzy-kNN are chosen and additionally the support vector machines (SVM), that are known to provide a good generalization capability.

As a kNN a version is trained that makes its decision as soon as k neighbors belong to the same class. The Fuzzy-kNN is implemented according to [9]. For this the membership to each class is computed, taking the incfluence of a nearest neighbor into account using the distance to the sample. This distance can moreover be weighted. For this the proposed version of assigning complete membership of the neighbors into their own class and nonmembership in all other classes is being used. The second version of weighting the neighbors' influence based on their distances to the class means assumes a unimodal distribution of the classes with equal variances, which turned out to be too restrictive ([10], [11]). Since the third proposed weighting increases computational costs through a search for nearest neighbors for each nearest neighbor of a sample, it is excluded.

For the SVM algorithm ([12]) the implementation of the Spider toolbox for Matlab is used, interfacing the libSVM, with using gridsearch for the parameter search, the rbf kernel and cross-validation to train classifiers. All methods use the Euclidean distance.

4. EXPERIMENTS AND RESULTS

The cell set from section 2 is split into 750 reference cells and 550 analysis cells for the training set and a valida-

basic feature set	kNN (a;r;o)	F-kNN-a (a;r;o)
morphology	75.3; 92.5; 85.4	77.7; 88.5; 83.2
chromatin green	91.9; 97.5; 94.7	91.9; 97.5; 94.7
chromatin extinction	93.9; 95.5; 94.7	91.9; 94.5; 93.2
chromatin+moments/ all features	92.4; 98.5; 95.5	93.9; 97.5; 95.7

Tab.2. Classification results of the validation set, computed for the feature sets and the number of neighbors of the classifiers within table 1. The classification results (a;r;o) are computed for the classifiers kNN and F-kNN and are given as percentages for the (a)nalysis, the (r)eference cells, and the (o)verall rate.

tion set of 200 reference cells, and 198 analysis cells. On the basis of the training set all features are normalized to the range $[0, 1]$.

Firstly a feature selection has been done. To rate a wide range of feature combinations during the feature selection process, the features are grouped into five basic feature sets. These are *morphology*, *chromatin green* and *chromatin extinction* as the chromatin features computed for the green channel and the extinction image respectively. Additionally the combination of the two former chromatin sets is extended through moment based features computed for the extinction and the edge texture image to *chromatin+moments*. And, finally, within the last basic feature set *all features* are included.

On each of these basic feature sets a feature selection has been performed up to feature set sizes of 50 features for each objective function. To determine the best feature set size of each objective function, kNN classifiers have been trained for k from 1 to 10, using the overall classification rates from leave-one-out cross-validation on the training set to rate the feature subsets. Table 1 shows the results of the best sets for each basic feature set.

For these five best classifiers the classification rates on the validation set has been computed with kNN and Fuzzy-kNN (table 2). It turned out that the best distinction between the two classes needs two features only. These are a moment based feature (IMTOTE in [3]) as an estimate of the integral optical density, and, as an inhomogeneity measure of the chromatin distribution, the median of the topological gradient image (RG in [3]) for the green channel.

As SVMs are known to provide good generalization results, two SVMs were trained on the training set. One on the whole set of 203 features (SVM203) and one for the two features from the feature selection process, that have shown a good discriminative power with the kNN classifiers (SVM2). Both SVMs are trained with 13-fold cross-validation and an iteratively refined grid-search for the SVM parameter standard deviation of the rbf kernel and starting with a logarithmic scale. The classification results on the validation set are comparable for both classifiers as well as to the results of the kNN and the Fuzzy kNN, which is shown in table 3.

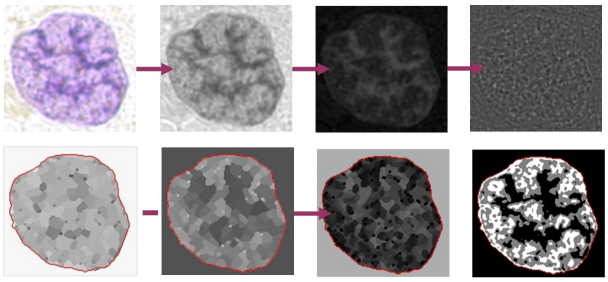


Fig. 3. Image transformations needed to compute the features. From left to right: Within the upper row the original RGB image, its green channel, the extinction image computed from the green channel and the edge image are displayed. Based on the extinction image the lower row contains the watershed regions computed for local maxima and for local minima, each filled with the gray value of the local optima, the topological gradient as the difference of the watershed regions and a three color image to partition the chromatin into darker and brighter particles.

5. DISCUSSION

For the different basic feature sets the classification results of the kNN for the best feature sets vary slightly. It can be seen that chromatin features provide a good separability performance, whereas only the geometrical features within the basic set morphology do not distinguish the two classes sufficiently.

To achieve the best overall classification result, only two features are needed, which provides low computational costs. These features also are due to visual criteria of the cytopathologists (Section 3), which simplifies an understanding of the computed decisions.

Validating these two features on the validation, set using kNN and F-kNN, as well as SVM2, results in comparable classification rates between the classifiers, and with slightly better classification results than those achieved with the kNN classifier after the feature selection on the training set. This is due to the distribution of the validation set within the feature space. Training a SMV on all features to possibly detect another relation between the features than has been tested during feature selection did not result in better classification results, but these still are comparable to the classifiers using two features only. So for this the number of training data might yet be too low. In consequence, to be sure not to have

classifiers	analysis cells	reference cells	overall rate
kNN	92.4	98.5	95.5
F-kNN	93.9	97.5	95.7
SVM2	92.9	98.5	95.7
SVM203	93.9	97.0	95.5

Tab. 3. The detection rates of analysis cells, reference cells, and the overall rates of the validation set are shown in percentages for the different classifiers. kNN, F-kNN and SVM2 use two features only, SVM203 classifies according to all features.

performed an overfitting or having used a training set that is not representative, these classifiers have to be tested with cells from specimens that are different from the specimens of the training set.

Furthermore a persistent difference between the classification results of reference cells and analysis cells for the kNN methods as well as for the SVMs is noticeable. The major reason for this might be the different number of training cells in the classes. Therefore the analysis cells should be supplemented with new prototypes.

But, overall it can be stated, that the classification results on the 398 validation cells result in overall classification results between 95.5% and 95.7% for the different classifiers, thus providing a good separability between the analysis cells and the reference cells. While the classification results of the different classification algorithms are comparable, one can choose the final classifier according to the best classifications results of either analysis cells or reference cells, or subject to computational costs.

Acknowledgements

I would like to thank Prof. Dr-Ing. T. Aach and Prof. emer. Dr-Ing. D. Meyer-Ebrecht, Institute of Imaging and Computer Vision, RWTH Aachen University, Germany, for many insightful discussions and I would like to thank Prof. Dr. med. Böcking, Institute for Cytopathology, Heinrich-Heine-University, Düsseldorf, Germany, for providing the dataset and his teaching about the medical background.

The project is supported by the Viktor-and-Mirka-Pollak-Fund for Biomedical Engineering.

References

[1] BÖCKING, A.B., et al. Consensus report of the ESACP task force on standardization of diagnostic DNA image cytometry. *Anal Cell Pathol*, 1995, vol. 8, no. 1, p. 67 - 74.

[2] SUN, X.R., WANQ, J., GARNER, D., PALCIC, B. Detection of cervical cancer and high grade neoplastic lesions by a combination of liquid-based sampling preparation and DNA measurements using automated image cytometry. *Cellular Oncology*, 2005, vol. 27, no. 8, p. 33 - 41.

[3] RODENACKER, K., BENGTSSON, E. A feature set for cytometry on digitized microscopic images. *Anal Cell Pathol*, 2003, vol. 25, no. 1, p. 1 - 36.

[4] SUK, T., FLUSSER, J. Graph Method for Generating Affine Moment Invariants. *Procs ICPR*, 2004, vol. 2, p. 192 - 195

[5] YOUNG, I. T., VERBEEK, P. W., MAYALL, B. H. Characterization of Chromatin Distribution in Cell Nuclei. *Cytometry*, 1986, vol. 7, no. 5, p. 467 - 474

[6] PUDIL, P., NOVOVIČOVÁ, J., KITTLER, J. Floating search methods in feature selection. *Pattern Recognition Letters*, 1994, vol. 15, no. 11, p. 1119 - 1125

[7] FUKUNAGA, K. *Statistical Pattern Recognition*. Academic Press, 1990

[8] PLUIM, J. P. W., MAINTZ, J. B. A. , Viergever, M. A. Mutual information based registration of medical images: a survey. *IEEE Trans. Med Imaging*, 2003, vol. 22, no. 8, p. 986 - 1004.

[9] KELLER, J. M., GRAY, M. R., GIVENS, J. A. A Fuzzy K-Nearest Neighbor algorithm. *IEEE Trans. System, Man, and Cybernetics*, 1985, vol. SMC-15, no. 4, p. 580 -585

[10] SCHNEIDER, T. E., BELL, A. A., MEYER-EBRECHT, D., BÖCKING, A., AACH, T. Computer aided cytological cancer diagnosis: Cell type classification as a step towards fully automatic cancer diagnostics on cytopathological specimens of serous effusions. *SPIE Medical Imaging 2007, Computer–Aided Diagnosis*. 2007, in print

[11] SCHNEIDER, T. E., BELL, A. A., GERBERICH, G., MEYER-EBRECHT, D., BÖCKING, A., AACH, T. Chromatinmuster-basierte Zellklassifizierung für die DNS-Bildzytometrie an Mundschleimhaut-Abstrichen. *Bildverarbeitung für die Medizin 2007*, 2007, to appear

[12] VAPNIK, V. N. *The Nature of Statistical Learning Theory*. Springer, 1995

About Authors...

Timna Esther SCHNEIDER received her diploma degree in computer science from the University of Kaiserslautern, Germany, in 2000. Since 2001, she is with the Institute of Imaging and Computer Vision, RWTH Aachen University as a Ph.D. student. Her current research interests and projects are focussed on feature extraction, feature selection, and classifiers with an emphasis on computerized cytopathological cancer diagnostics.