

# Automated Classification of Analysis- and Reference Cells for Cancer Diagnostics in Microscopic Images of Epithelial Cells from the Oral Mucosa

T. E. Schneider

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 possibility of early detection. One such method is standardized DNA image cytometry. For this, the diagnostically relevant cells have to be found within each specimen, which is currently performed manually. Since this is a time-consuming process, a preselection of diagnostically relevant cells has to be performed automatically. For specimens of the oral mucosa this involves distinguishing between undoubted healthy epithelial cells and possibly cancerous epithelial cells.

Based on cell images from a brightfield light microscope, a set of morphological and textural features was 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 cells it could be shown that normal and possibly cancerous cells can be distinguished at overall rates above 95.5 % 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

Cancer is the second leading cause of death in industrial countries. So as to have the best possible chance of cure, 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 from already marginal changes within single cells using cytopathological methods. These cell specimens can be obtained easily and painlessly, e.g. with tiny brushes, from the oral mucosa.

One cytopathological diagnostic method is DNA image cytometry (DNA-ICM). For this the cells are stained stoichiometrically 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 obtain the DNA value from these integral optical densities, a set of about normal, non-proliferating cells, called *reference cells*, has to be selected. Then the DNA content of each nucleus is computed from the ratio of its integral optical density to the mean integral optical density of the reference cells. Subsequently the diagnosis is performed based on the histogram of the DNA values of diagnostically relevant, suspicious *analysis cells*. For oral mucosa these are squamous epithelial cells with noticeably altered morphology and chromatin texture, i.e., amount and distribution of DNA, of their nuclei, cells that therefore are suspicious of cancer.

According to the European guidelines for DNA-ICM [1], a cytopathologist has to find and select manually about 30 ref-

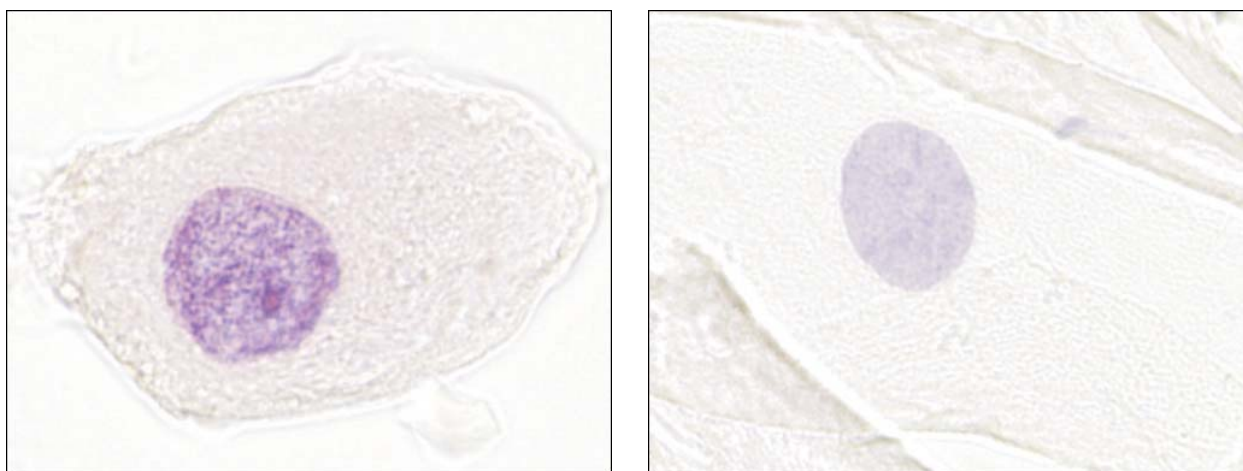


Fig. 1: Two examples of Feulgen stained squamous epithelial cells. On the left a normal cell is shown, on the right a morphologically abnormal analysis cell. Since Feulgen stains the DNA stoichiometrically, the DNA within the nuclei is stained, whereas the surrounding cytoplasm is not.

reference cells and about 300 analysis cells. This is a time-consuming task, which the expert should therefore be assisted by 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 abnormally high DNA content, by measuring the DNA content 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] cannot 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., automatic discrimination of normal 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 the properties of epithelial reference and analysis cells, the process of reducing the whole feature set to a subset of features with optimal discriminating 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 presented in section 4, showing that epithelial reference and analysis cells

from 15 specimens of the oral mucosa can be discriminated with different classification methods at total classification rates of above 95.5 %. The paper ends by analyzing the classification results and suggesting topics for future study.

## 2 Material

Our dataset is based on 15 Feulgen stained smears from the oral mucosa. Seven of these specimens are without cancer cells, including three with inflammation. Five specimens contain 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  $\Delta x \approx 0.1 \mu\text{m}$ .

An experienced cytopathologist classified 950 reference cells within the specimens without cancer cells, and 748 analysis cells within the specimens with cancer cells. Fig. 1 shows two example cells.

For these cells, the contours of the nuclei are given as chaincodes.

## 3 Methods

To be able to distinguish automatically between analysis and reference squamous epithelial cells, a set of potentially relevant features has been implemented. To minimize the risk of overfitting during training of a classifier, a feature selection method is performed 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 these 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 sketched in Fig. 2, and the algorithms used within each step are described below. The distance measure used here within each step is the Euclidean distance.

### 3.1 Features

When performing a manual discrimination between analysis and reference cells, 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 described in [3] and [4]. These comprise area, perimeter, form factor, Fourier descriptor energies, and, further on, translation, rotation and scale invariant features. The latter are 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 (extinction 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 Fig. 3 for examples of the image transformations. This sums

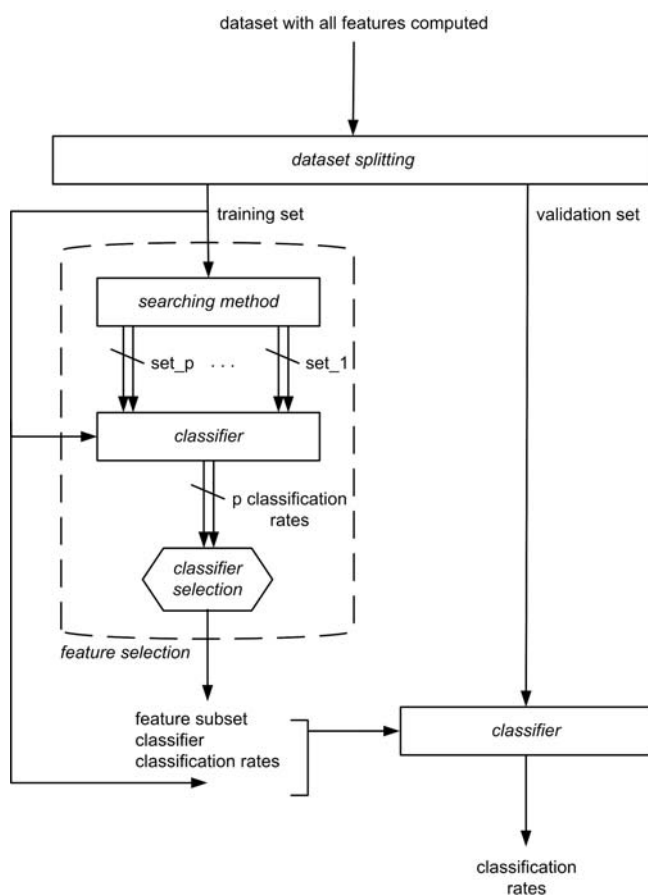


Fig. 2: Algorithmic workflow of feature subset selection and classifier generation. The different sized feature subsets resulting from the searching method are denoted as  $set_1$  to  $set_p$ , the output of classifier training and training set classification as  $p$  classification rates.

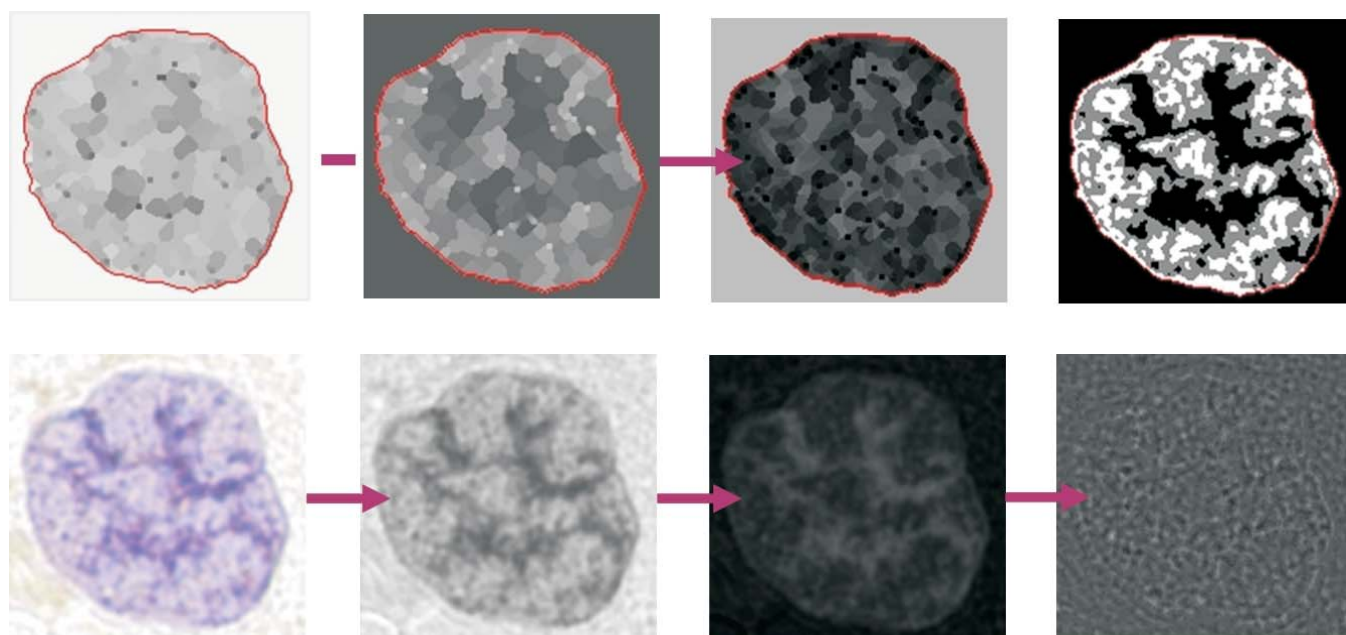


Fig. 3: Image transformations needed to compute the features. From left to right: Within the upper row the original RGB image, its green channel, the extinction image computed from the green channel and the edge image are displayed. Based on the extinction image the lower row contains the watershed regions 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

up to 203 features describing the morphology of the nuclei and their chromatin texture.

### 3.2 Feature selection

Testing the separability performance of each possible feature subset is computationally too expensive for large feature sets. Therefore the parameter-free sequential forward floating selection search method [6] is implemented to identify feature 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 feature selection. For the different subsets of features during this selection the distribution of the data may fit different distribution models, including non-Gaussian, multi-modal ones. Furthermore, for higher dimensions of the feature space, the data is in general 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, support vector machines (SVM), which are known to provide a good generalization capability.

A kNN version is trained that makes its decision as soon as  $k$  neighbors belong to the same class  $i$ . The Fuzzy-kNN is

Table 1: Results of the feature selection for each basic feature set, selected due to the best overall classification result. 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.

| 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; 95.5; 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                   |



implemented according to [9]. For this the membership to each class is computed, taking the influence of a nearest neighbor into account using the distance to the sample. This distance can moreover be weighted. For this we used the proposed version of assigning complete membership of the neighbors into their own class and nonmembership in all other classes. The second version of weighting the influence of the neighbours 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 the computational costs through a search for nearest neighbors for each nearest neighbor of a sample, it is excluded.

For the SVM algorithm [12] the Spider toolbox for Matlab is applied, interfacing the libSVM. The rbf kernel is chosen as kernel function, and the classifiers are trained using grid-search for the parameter search, and cross-validation on the training data.

## 4 Experiments and results

The cell set from section 2 is split into reference cells and analysis cells for the training set and a validation set of reference cells, and analysis cells. On the basis of the training set all features are normalized to the range [0, 1].

Firstly, feature selection was performed. 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.

Feature selection was performed on each of these basic feature sets up to feature set sizes of 50 features for each objective function (searching method in Fig. 2). To determine the best feature set size of each objective function, kNN classifiers were trained for  $k$  from 1 to 10 (classifier within the block feature selection in Fig. 2). These sets are chosen 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 were computed with kNN and Fuzzy-kNN

Table 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)analysis, the (r)eference cells, and the (o)verall rate.

| 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/<br>all features | 92.4; 98.5; 95.5 | 93.9; 97.5; 95.7  |

(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 showed 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 width of the rbf kernel and the penalty term. The grid search started at logarithmic scale for both parameters. The classification results on the validation set are comparable for both classifiers and also with the results of the kNN and the Fuzzy kNN, as shown in Table 3.

A comparison of the two classifiers with the best overall classification rate, F-kNN and SVM2, shows a higher detection rate of the diagnostically relevant analysis cells for F-kNN, but with a lower predictive value and higher computational costs compared to SVM2. Table 4 lists the number of misclassified cells and the predictive values in comparison for these two classifiers. The predictive value of each class is calculated as the ratio of the number of correctly classified cells of the class to the sum of the number of correctly classified cells of this class and the number of cells falsely classified into this class.

Table 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.

| classifiers | analysis cells | reference cells | overall rate |
|-------------|----------------|-----------------|--------------|
| kNN         | 92.4           | 98.5            | 95.5         |
| F-kNN       | 93.9           | 97.5            | 95.7         |
| SVM 2       | 92.9           | 98.5            | 95.7         |
| SVM203      | 93.9           | 97.0            | 95.5         |

Table 4: Absolute number of correctly classified cells within each class and the predictive values in percentages for F-kNN and SVM2. The total number of cells in each class is noted in brackets.

| classifier      | F-kNN     |        | SVM2      |        |
|-----------------|-----------|--------|-----------|--------|
| analysis cells  | 186 (198) | 97.4 % | 184 (198) | 98.4 % |
| reference cells | 195 (200) | 94.2 % | 197 (200) | 93.4 % |

## 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 rate, only two features are needed, which results in low computational costs. These features also are consistent with the visual criteria used by 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. These rates on the validation set are slightly better than those classification results achieved with the kNN classifier after feature selection on the training set. This is due to the distribution of the validation set within the feature space.

Training a SVM on all features for the possibility of detecting another relation between the features than was tested during feature selection did not result in better classification results. However, these results are still comparable to the classifiers using two features only. The quantity of training data may still be too low for this. In consequence, to be sure not to have performed an overfitting or to have 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, there is a persistent difference between the classification results of the reference cells and analysis cells for the kNN methods, and also for the SVMs. The major reason for this may be the different number of training cells in the classes. Therefore the analysis cells should be supplemented with new prototypes.

In conclusion, 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 good separability between the analysis cells and the reference cells. While the classification results of the different classification algorithms are comparable, we can choose the final classifier according to the best classification results of either analysis cells or reference cells, or on the basis of computational costs.

## Acknowledgments

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. A. Böcking, Institute for Cytopathology, Heinrich-Heine-University, Düsseldorf, Germany, for providing the dataset and for his teaching about the medical background.

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

## References

[1] Böcking, A., et al.: Consensus Report of the ESACP Task Force on Standardization of Diagnostic DNA Im-

age Cytometry. *Anal. Cell Pathol.*, Vol. **8** (1995), 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*, Vol. **27** (2005), No. 8, p. 33–41.
- [3] Rodenacker, K., Bengtsson, E.: A Feature Set for Cytometry on Digitized Microscopic Images. *Anal Cell Pathol*, Vol. **25** (2003), No. 1, p. 1–36.
- [4] Suk, T., Flusser, J.: Graph Method for Generating Affine Moment Invariants. *Procs ICPR*, Vol. **2** (2004), p. 192–195.
- [5] Young, I. T., Verbeek, P. W., Mayall, B. H.: Characterization of Chromatin Distribution in Cell Nuclei. *Cytometry*, Vol. **7** (1986), No. 5, p. 467–474.
- [6] Pudil, P., Novovičová, J., Kittler, J.: Floating Search Methods in Feature Selection. *Pattern Recognition Letters*, Vol. **15** (1994), 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*, Vol. **22** (2003), 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*, Vol. **SMC-15** (1985), 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*, Vol. **6514** (2007), p. 6514OG-1–10.
- [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, p. 257–261.
- [12] Vapnik, V. N.: *The Nature of Statistical Learning Theory*. Springer, 1995.

Timna Esther Schneider  
e-mail: Timna.Schneider@lfb.rwth-aachen.de

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