

Reduction of stain variability in bone marrow microscopy images

Influence of augmentation and normalization methods on detection and classification of hematopoietic cells

Philipp Gräbel¹, Martina Crysandt², Reinhild Herwartz², Melanie Baumann²,
Barbara M. Klinkhammer³, Peter Boor³, Tim H. Brümmendorf²,
Dorit Merhof¹

¹Institute of Imaging and Computer Vision, RWTH Aachen University, Germany

²Department of Hematology, Oncology, Hemostaseology and Stem Cell
Transplantation, University Hospital RWTH Aachen University, Germany

³Institute of Pathology, University Hospital RWTH Aachen University, Germany
`graebel@lfb.rwth-aachen.de`

Abstract. The analysis of cells in bone marrow microscopy images is essential for the diagnosis of many hematopoietic diseases such as leukemia. Automating detection, classification and quantification of different types of leukocytes in whole slide images could improve throughput and reliability. However, variations in the staining agent used to highlight cell features can reduce the accuracy of these methods. In histopathology, data augmentation and normalization techniques are used to make neural networks more robust but their application to hematological image data needs to be investigated. In this paper, we compare six promising approaches on three image sets with different staining characteristics in terms of detection and classification.

1 Introduction

Automated analysis of hematological image data is an emerging field in computer vision. Recent technological advances, particularly deep learning, made the classification of hematopoietic cells in bone marrow microscopy images possible. The core steps of an automated system include detection of individual cells, classification of cell type and quantification of entire Whole-Slide Images (WSI). In clinical practice, the analysis of bone marrow samples includes manual screening of selected regions by highly trained medical experts. Deep learning offers, for the first time, the potential to automate this procedure and thereby increase throughput and objectivity.

Bone marrow samples use staining agents to highlight specific properties of the cell. This results in a clearly distinguishable cell nucleus and cytoplasm that is colored according to cell type. While this staining process is standardized

and performed following a strictly defined workflow, a certain variability cannot be avoided. In extreme cases, this might lead to slides that are uninterpretable even by human experts, but also slight variations can degrade performance of automated systems.

In the case of histological image analysis, various effects of staining have been extensively researched. In terms of variability of a single stain type, two different kinds of approaches seem prevalent: stain normalization and stain augmentation [1]. The first approach aims at minimizing variance in both training and test images, resulting in simplified network training. Stain augmentation aims at augmenting the training data such that it exhibits a similar distribution (i.e., staining variability) as the test data. Classical approaches for stain normalization are conversion to grayscale, a color-based transform [2] and color deconvolution [3]. Stain augmentation can be performed by altering the different color channels in a transformed image, e.g. using Principal Component Analysis (PCA), the HSV color space or color deconvolution.

Hematology focusses on object (cell) detection and classification. While classification relies on the color of a cell (e.g. basophilic cytoplasm is rather blue while eosinophilic cytoplasm is red) as much as on structure, detection heavily relies on the shape of a cell. This leads to the assumption that classification and detection are affected differently by the aforementioned methods, which can significantly alter the color properties of an image. Consequently, we perform an analysis of several normalization and augmentation techniques on hematological image data with respect to detection and classification separately. To this end, we employ different hematological datasets with varying degrees of staining severity.

2 Materials and methods

2.1 Image data

We use a whole slide microscopy scanner with $63\times$ magnification and automated immersion oiling to obtain high resolution images from bone marrow samples. Each sample is stained in a standardized procedure with Pappenheim staining. Nevertheless, variabilities in the scanned images of stained bone marrow samples occur – both due to effects of the staining but also the subsequent digitalization with automatic focus and illumination setting. We divide the data into three subsets: normal images (NI) in which the staining appears as expected, images that tend towards pink over-staining (OS1) and images that tend towards darker blue over-staining with starker contrast (OS2). Examples are shown in Fig. 1. It needs to be mentioned that not necessarily all extracted patches of a slide with a particular stain exhibit the same stain characteristics as there is also some variation within extracted regions. All three sets have manual annotations from hematological experts.

For the detection task, we divide each set into slightly overlapping patches of size $1600 \times 1600 \text{ px}^2$, resulting in 913 images in NI, 268 images in OS1 and 215

images in OS2. For the classification task, we extract patches of size 244×244 px² around individual cells, resulting in 4560 patches in NI, 1385 patches in OS1 and 1075 patches in OS2.

2.2 Normalization Methods

The normalization methods are applied to images both in training and in testing and aim at minimizing the variance (caused among others by stain inhomogeneities) that a network is required to learn. Fig. 2 shows some examples for normalization as well as augmentation methods.

Macenko Based on the color deconvolution method by Macenko et al. [3], two vectors corresponding to the the two different staining components can be extracted. Using these, the image can be decomposed into a stain matrix and a concentration matrix. Images are normalized by replacing the stain matrix of an individual image with a target stain matrix, which can be extracted from a large number of normally stained images.

Grayscale The image is converted into gray-scale (y) from RGB using the conversion $y = 0.299R + 0.587G + 0.114B$.

LAB Reinhard et al. [2] proposed a normalization method using image statistics in LAB space. Images in LAB space are adapted so that mean and standard deviation of each of the three channels is the same as in a large number of normally stained images.

2.3 Augmentation Methods

Augmentation methods are only applied to images in the training process in order to increase the variance of normally stained images. In this work, we use a parameter x to determine the augmentation strength – higher values for x result in a more perturbed images. Nevertheless, each augmentation has a random element to it, such that the whole spectrum from maximum perturbation to no perturbation is utilized with a given probability. Fig. 2 shows some examples for normalization as well as augmentation methods.

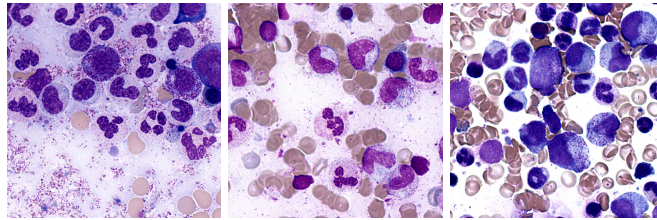


Fig. 1. Three patches from the detection dataset of (from left to right) NI, OS1 and OS2. For classification, patches around each individual cell are extracted.

Macenko Stain augmentation is performed by perturbing the elements of the concentration matrix extracted based on Macenko et al.’s color deconvolution method. [3]. Preliminary visual analysis showed that a multiplication with a factor from a Gaussian distribution with $\mu = 1$ and $\sigma^2 = \frac{x}{30}$ (with x being the augmentation strength) yields realistic results.

HSV The augmentation is performed by adding a value drawn from a Gaussian distribution with zero mean and variance x (augmentation strength) to the hue component. The hue component is chosen as most changes in differently stained images show a shift towards pink or blue, which can be represented by a shift in hue.

PCA Using images from all three datasets, a PCA is performed on the RGB pixel values. Preliminary analysis shows that the first component of PCA corresponds mostly to changes in brightness, while the second component corresponds to the change in color. Consequently, augmentation can be performed by perturbing the second component of an image in PCA space. Preliminary visual comparison of transformed images showed that a multiplication with a factor $f \in [1 - 0.01x, 1 + 0.01x]$, with x being the augmentation strength results in realistically augmented images.

2.4 Experimental setup

For the evaluation, the set of normally stained images (NI) is divided into six subsets. In six-fold cross-validation, we first train a network on four training and one validation set. Evaluation is performed separately on the held out test set as well as on OS1 and OS2. For augmentation, we evaluate augmentation strengths from $x \in [1, 9]$.

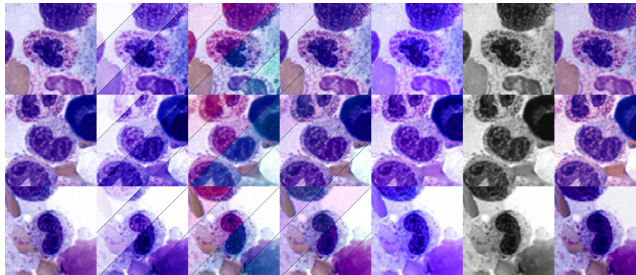
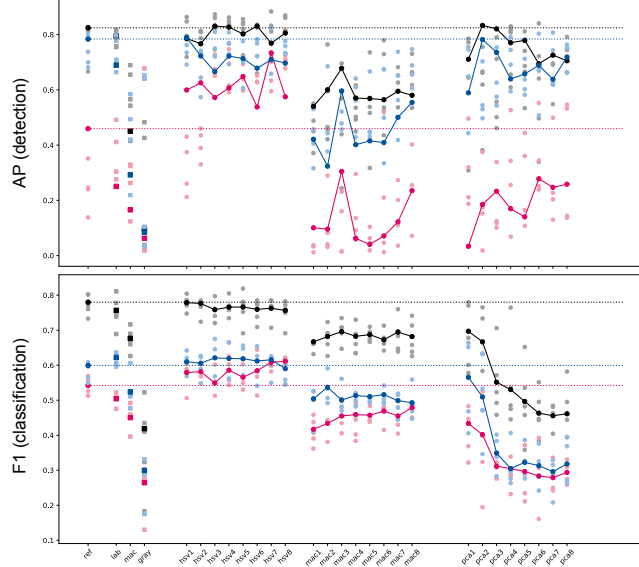


Fig. 2. Augmentation and normalization for three sample cells (normal stain classification dataset). From left to right: original, Macenko augmentation, HSV augmentation, PCA augmentation, Macenko normalization, grayscale normalization, LAB normalization. For the augmentation methods, the largest change in each direction is shown (top left and bottom right of each image, both augmentation strength 9), with a medium augmentation in the middle (augmentation strength 4).

Fig. 3. Top: results for detection in terms of Average Precision (AP). Bottom: results for classification in terms of macro f1-Score. Horizontal lines denote the reference results, colors denote the test set (black for NI, red for OS1, blue for OS2). Each column shows results from all six folds (brighter markers) as well as the result over all folds.



For detection, we employ a U-Net [4] of depth five on random patches of size $512 \times 512 \text{px}^2$ followed by Watershed [5] segmentation. This model showed excellent results in previous comparisons to other state-of-the-art architectures. Training takes at most 100 epochs with early stop after ten epochs without improvement. The matching of prediction and ground truth is performed using centroid distance. For classification, we employ a Dense-121 [6] architecture with cross entropy loss. Training takes at most 300 epochs with early stop after 50 epochs without improvement. In both cases, we additionally perform data augmentation with random rotation, mirroring and random crop.

3 Results

Fig. 3 shows the results for detection and classification. In general, augmentation methods yield better results than normalization methods. However, in both cases results frequently decrease compared to the baseline without normalization or augmentation. Notable exception is the HSV augmentation method: while the absolute score stays mostly unchanged on NI with a slight increase in variance between folds, results for the stained images of OS1 (for both detection and classification) and OS2 (only for detection) improve. It is further noteworthy that for classification, the PCA augmentation method has decreasing scores with increasing augmentation strength, which is not the case for detection.

4 Discussion

The results show that most standard augmentation and normalization methods commonly used in digital histopathology do not yield improved results on heavily stained hematology data and even decrease classification and detection scores in most cases. Only the HSV augmentation methods mostly improves results: particularly the dataset OS1 benefits from the additional variance in the training data. While methods based on color deconvolution yield good results on histopathological image data, no improvement is observed for hematological images. The varying density of a large number of very different objects (cells) might make computation of accurate stain and concentration matrices more challenging. The poor results of the PCA augmentation methods are unexpected as their color transformation is based on a large data basis containing patches from all three staining types.

It stands to reason that these common data-augmentation methods are not able to sufficiently capture the color distribution of the staining variabilities in hematological image data. In future work, we will further look into generative models that are able to learn more accurately from training data.

It needs to be further noted that the datasets OS1 and OS2 in this work constitute extreme cases, which are already difficult to analyze for human experts. While it might be an option to exclude such images from training or in application, researching the limitations of a neural network is important for translation into clinical practice.

Acknowledgement. The authors would like thank Xinyi Gao for his support in improving the data transformation pipeline. This study was supported by the following grants: DFG: SFB/TRR57, SFB/TRR219, BO3755/6-1, BMBF: STOP-FSGS-01GM1901A, BMWi: EMPAIA project to PB.

References

1. Tellez D, Litjens G, Bandi P, et al. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. arXiv preprint arXiv:190206543. 2019;.
2. Reinhard E, Adhikhmin M, Gooch B, et al. Color transfer between images. IEEE Comput Graph Appl. 2001;21(5):34–41.
3. Macenko M, Niethammer M, Marron JS, et al.; IEEE. A method for normalizing histology slides for quantitative analysis. 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro. 2009; p. 1107–1110.
4. Ronneberger O, Fischer P, Brox T; Springer. U-net: Convolutional networks for biomedical image segmentation. Medical Image Computing and Computer-Assisted Intervention (MICCAI), International Conference on. 2015; p. 234–241.
5. Beucher S, Meyer F. The morphological approach to segmentation: the watershed transformation. Mathematical morphology in image processing. 1993;34:433–481.
6. Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017; p. 4700–4708.