Robust slide cartography in colon cancer histology

Evaluation on a multi-scanner database

Petr Kuritcyn[1], Carol I. Geppert[2], Markus Eckstein[2], Arndt Hartmann[2],
Thomas Wittenberg[1], Jakob Dexl[1], Serop Baghdadlian[1], David Hartmann[1],
Dominik Perrin[1], Volker Bruns[1], Michaela Benz[1]

[1]Fraunhofer Institute for Integrated Circuits IIS
[2]Institute of Pathology, University Hospital Erlangen,
Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU)
petr.kuritcyn@iis.fraunhofer.de

Abstract. Robustness against variations in color and resolution of digitized whole-slide images (WSIs) is an essential requirement for any computer-aided analysis in digital pathology. One common approach to encounter a lack of heterogeneity in the training data is data augmentation. We investigate the impact of different augmentation techniques for whole-slide cartography in colon cancer histology using a newly created multi-scanner database of 39 slides each digitized with six different scanners. A state of the art convolutional neural network (CNN) is trained to differentiate seven tissue classes. Applying a model trained on one scanner to WSIs acquired with a different scanner results in a significant decrease in classification accuracy. Our results show that the impact of resolution variations is less than of color variations: the accuracy of the baseline model trained without any augmentation at all is 73% for WSIs with similar color but different resolution against 35% for WSIs with similar resolution but color deviations. The grayscale model shows comparatively robust results and evades the problem of color variation. A combination of multiple color augmentations methods lead to a significant overall improvement (between 33 and 54 percentage points). Moreover, fine-tuning a pre-trained network using a small amount of annotated data from new scanners benefits the performance for these particular scanners, but this effect does not generalize to other unseen scanners.

1 Introduction

Histopathology involves the microscopic examination of tissue sections. The sections undergo various preparation steps before analysis: fixation, embedding, sectioning, staining and digitization [1]. Each step introduces some form of variation into the resulting whole-slide image. A trained pathologist can cope with these variances; however, a human analysis is prone to subjectivity and interobserver variance. Computational pathology aims to support pathologists in their decision-making process by automating in a very objective and validated fashion the calculation of scores or extraction of parameters hidden to the human eye. Deep learning approaches achieve very good results in many applications [2]. However, training robust models for the analysis of histopathological slides is still a challenge. The algorithm's accuracy suffers from the high variability encountered in slides in the field [2]. Sources of heterogeneity are different assays, unstandardized preparation protocols, different characteristics of the scanner components - most relevantly it's camera and microscope objective – and finally color post-processing steps. There are two main approaches to counter the problem of variances in WSIs: (i) normalization of images during runtime and (ii) representing the variance already in the training database.

The first approach aims to provide a standardized input to the classifier by adjusting the slide's color and scale to match the properties of the reference slides the network was trained on. Earlier solutions rely on a normalization based on color deconvolution of the underlying staining components [3]. The disadvantage of these solutions is that they frequently produce unrealistic color alterations and are not robust against severe stain variations. More recent techniques use machine learning algorithms to improve the normalization quality by taking into account morphological properties in addition to the color [1,4]. Two prominently employed network architectures for transferring the reference slides' style are sparse auto-encoders or generative adversarial networks (GANs) [5]. These methods produce more reliable results, however, they are computationally expensive and prone to false color estimations in unseen regions.

The second approach requires a sufficiently heterogeneous multi-centric dataset. When this is not available, a viable workaround is to introduce variance synthetically using domain-specific data augmentation. Native image patches can be duplicated and altered in terms of their geometry and color with the goal of increasing the network's capability to generalize to unseen data. Geometric transformations leave the color information intact and modify only morphological information. Patches can be rotated, flipped, scaled or images can be artificially blurred to simulate out-of-focus scans. Color augmentations, on the other hand, include variations of the color's hue, saturation, gamma, etc. This can help to mimic different stain protocols or color alterations of WSI scanners. A color augmentation tool specific to the domain of histopathology is a stain variation [6], where the hematoxylin and eosin components are separated using a color deconvolution in order to vary their color properties independently. Training and evaluating classifiers on slides that stem from one and the same laboratory and scanner can result in overly optimistic accuracy estimations. On the contrary, evaluating on a multi-centric dataset frequently shows poor performance [7]. Telez et al. compared for different applications the performance gain from stain normalization and stain augmentation and conclude that the latter is more beneficial [2].

Previous research has largely focused on the variance introduced by variations in the staining protocol. In this work, we investigate the effects of variations that stem from the use of different slide-scanners.

Table 1. WSI resolution, size of test database and number of patches.

Scanner	Resolution in µm per pixel	number of test patches	number of patches for fine-tuning
3DHISTECH MIDI	0.22	1,381,316	40,230
Fraunhofer iSTIX	0.17	2,123,364	49,005
Fraunhofer SCube	0.27	857,511	38,528
PreciPoint M8	0.35	514,397	35,524
Hamamatsu Nanozoomer S210	0.22	1,424,716	-
Hamamatsu Nanozoomer S360	0.23	1,298,056	-

2 Materials and Methods

2.1 Materials

The dataset used for the baseline cartography network comprises 161 hematoxylin and eosin (HE) stained colon sections from the Institute of Pathology at the University Hospital Erlangen. First, all samples were digitized with a 3DHISTECH MIDI scanner (20x magnification, 0.22 µm per pixel) and annotated manually by accurately outlining the contours of seven different tissue classes: tumor, necrosis, inflammation, connective combined with adipose tissue, muscle tissue, mucosa and mucus. Based on the annotated WSIs, labelled non-overlapping patches of pixel size 224x224 are generated. Patches that do not intersect with a manual annotation or contain no or only little foreground are discarded. The number of patches per class and slide is limited to 10,000 (using random sampling) in order to limit the overall dataset size while ensuring that information from all available slides is used. The training database comprises 2,173,515 patches from 92 slides. The validation set contains 719,010 patches from a disjoint set of 30 slides. These two datasets were used to train the CNN, while the remaining 39 glass slides were additionally digitized with four other automated scanners and with a manual microscope using the real-time stitching software iSTIX. The resolution as well as the color varies significantly between the different scanners (see Tab. 1 and Fig. 1).

For each slide, the annotations are transferred from the original scan to the new scans by co-aligning the WSIs. The main steps for the registration are the adjustment of resolution, calculation of features, brute-force-matching of the slides' feature points and finally the estimation of a global transformation (translation and rotation). Afterwards, for each scanner a test database with labelled patches (224x224) from 30 slides is generated without limiting the number of patches per class/slide. Due to the different resolution and background detection, the amount of image patches varies among the scanner datasets. A set of nine slides is excluded from these datasets and reserved for additional fine-tuning experiments for three scanners (M8, SCube, iSTIX). Moreover, tiles from the original scanner are included in a mixed database for fine-tuning comprising 163,287 patches.

2.2 Methods

CNNs are a common choice to solve image classification tasks. One popular CNN architecture is Xception, which was introduced by Chollet [8]. Its main characteristics are a depthwise separable convolution with residual connections. We slightly adapted the architecture by introducing two dropout layers between the fully connected layers at the top and replacing the logistic regression with softmax. Moreover, our input image size is 224x224 (instead of 299x299) in order to obtain more image patches that lie entirely within the bounds of the manual annotations. All experiments were carried out using the TensorFlow framework (version 2.2). For training, cross entropy loss and Adam optimizer with a learning rate of 0.001 and an exponential decay was applied. Image patches are zero-centered and the batches are generated with respect to class labels ensuring that each class is equally represented on average in a batch. Class imbalances are compensated by oversampling of underrepresented classes. A dropout rate of 0.5 was chosen.

First, a baseline experiment is carried out, where no data augmentation is employed. Afterwards several data augmentation techniques are applied during training and the robustness of the resulting model is evaluated on our six per-scanner test databases. Based on the observation that the slides vary mainly in color and in scale, we focus on color transformations and zoom variations. For each augmentation type, a probability and a valid parameter range is defined. We investigate variations of saturation, hue and contrast and apply gamma correction to the image patches. Additionally, the hematoxylin and eosin components are separated [6] and manipulated independently (HE augmentation). All experiments that employ data augmentation start with the weights of the baseline training. Moreover, we investigate the robustness of a model trained only on grayscale image tiles. Finally, we combine the most promising data augmentation types in a single run in order to evaluate if their individually observed benefits add up. In addition, the impact of fine-tuning with a small number of

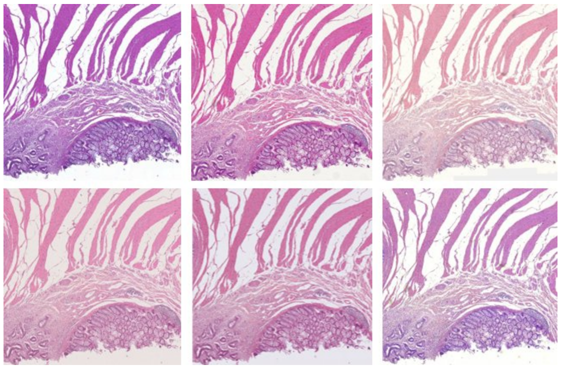

Fig. 1. Digitized slide scanned with MIDI, M8, iSTIX (upper row, from left to right), S210, S360, SCube (lower row, from left to right).

scanner-specific images is evaluated for three of the scanners individually and in combination (see Tab. 1).

3 Results

Results obtained on the scanner-specific test sets are listed in Tab. 2. The baseline and grayscale models are trained without any data augmentation. Starting point for the fine-tuning is the model trained on the MIDI scanner with HE and gamma augmentation.

Table 2. Classification accuracies (number of true positive classifications divided by the total number of classified patches) for different models on each scanner test set.

	Classification accuracy on					
Model	3DHISTECH	PreciPoint	Fraunhofer		Hamamatsu	
	MIDI	M8	iSTIX	SCube	S210	S360
Baseline	0.939	0.394	0.290	0.731	0.354	0.361
Grayscale	0.908	0.849	0.680	0.885	0.864	0.882
HE + Gamma	0.925	0.696	0.432	0.884	0.663	0.727
Hue + Sat + HE	0.918	0.891	0.621	0.896	0.880	0.901
Hue + Sat + Bright + Cont	0.933	0.894	0.493	0.858	0.863	0.917
M8 fine-tuning	0.610	0.939	0.519	0.759	0.668	0.780
iSTIX fine-tuning	0.566	0.642	0.821	0.614	0.576	0.533
SCube fine-tuning	0.894	0.592	0.386	0.936	0.489	0.536
Mixed fine-tuning	0.902	0.917	0.852	0.904	0.778	0.799
Zoom	0.931	0.418	0.310	0.750	0.343	0.344
Gamma	0.921	0.398	0.370	0.848	0.305	0.331
HE	0.932	0.679	0.401	0.831	0.637	0.680
Hue	0.927	0.861	0.410	0.852	0.821	0.896
Saturation	0.934	0.452	0.365	0.856	0.422	0.459
Brightness	0.918	0.494	0.371	0.815	0.372	0.397
Contrast	0.927	0.455	0.336	0.796	0.368	0.380

4 Discussion

A baseline experiment confirms earlier observations that a trained CNN shows poor performance on unseen data from another scanner. In our experiments, the influence of changes in resolution is less critical than the deviation in color: the accuracy of the baseline model on SCube scans (73%), which are similar in color but have a different resolution, is significantly higher than that on Hamamatsu scans (35%), which share the same resolution with the original MIDI scans. The highest impact on the results is gained by employing the HE and hue augmentations. On the contrary, zoom augmentation yields only little benefit.

Surprisingly, the grayscale model shows comparatively robust results and evades the problem of color variation - the highest burden for robust models. A combination of multiple methods (hue, saturation and HE augmentations) lead to a significant overall improvement. By fine-tuning the model using a small set of patches from the newly targeted scanner, the overall accuracy could be raised to a level close to that obtained on the native dataset for all new scanners except iSTIX. A likely explanation is that the manual scanning concept inherently suffers from stronger variances and the overall quality is inferior to that of high-end automated scanners (more out-of-focus areas, stitching artefacts).

Fine-tuning the network with additional patches from four scanner datasets ("mixed") yields a solid performance on all scanners. However, this model does not generalize as well as the model trained on a data augmented (hue, saturation and HE color augmentation together) database and shows worse results on the unseen Hamamatsu datasets: 78-80% against of 88-90%. In future work we will focus on extending the augmentation range and developing an automated approach for increasing the robustness of a pre-trained CNN.

Acknowledgement. This work was supported by the Bavarian Ministry of Economic Affairs, Regional Development and Energy through the Center for Analytics – Data – Applications (ADA-Center) within the framework of „BAYERN DIGITAL II" (20-3410-2-9-8). A part of the employed scanners and the deep learning cluster were funded by the Federal Ministry of Education and Research under the project reference numbers 16FMD01K, 16FMD02 and 16FMD03. Hamamatsu scans courtesy of Hamamatsu Photonics Deutschland GmbH.

References

1. Bejnordi BE, Litjens G, et al. Stain Specific Standardization of Whole-Slide Histopathological Images. IEEE Trans Med Imaging. 2016;35(2):404–415.
2. Tellez D, Litjens G, Bándi P, et al. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. Med Image Anal. 2019;58:516–24.
3. Reinhard E, Adhikhmin M, Gooch B, et al. Color transfer between images. IEEE Comput Graph Appl. 2001;21(5):34–41.
4. Khan AM, Rajpoot N, Treanor D, et al. A Nonlinear Mapping Approach to Stain Normalization in Digital Histopathology Images Using Image-Specific Color Deconvolution. IEEE Trans Biomed Eng. 2014;61(6):1729–1738.
5. Zanjani FG, et al. Stain normalization of histopathology images using generative adversarial networks. Proc IEEE Int Symp Biomed Imaging. 2018; p. 573–577.
6. Tellez D, Balkenhol M, Otte-Höller I, et al. Whole-Slide Mitosis Detection in H E Breast Histology Using PHH3 as a Reference to Train Distilled Stain-Invariant Convolutional Networks. IEEE Trans Med Imaging. 2018;37(9):2126–2136.
7. Leo P, Lee G, Shih NNC, et al. Evaluating stability of histomorphometric features across scanner and staining variations: prostate cancer diagnosis from whole slide images. J Med Imaging (Bellingham). 2016;3(4):1–11.
8. Chollet F. Xception: Deep Learning with Depthwise Separable Convolutions. Conf Comput Vis Pattern Recognit. 2017; p. 1800–1807.