Deep learning and color variability in breast cancer histopathological images: a preliminary study

Gobert Lee, Mariusz Bajger, Kevin Clark

Gobert Lee, Mariusz Bajger, Kevin Clark, "Deep learning and color variability in breast cancer histopathological images: a preliminary study," Proc. SPIE 10718, 14th International Workshop on Breast Imaging (IWBI 2018), 107181E (6 July 2018); doi: 10.1117/12.2316613

SPIE.

Event: The Fourteenth International Workshop on Breast Imaging, 2018, Atlanta, Georgia, United States

Deep Learning and Color Variability in Breast Cancer Histopathological Images - a Preliminary Study

Gobert Lee[a,b], Mariusz Bajger[a,b], and Kevin Clark[b]

[a]Medical Devices Research Institute, Flinders University, Adelaide, Australia
[b]College of Science and Engineering, Flinders University, Adelaide, Australia

ABSTRACT

Variability in the color appearance in H&E stained histopathological images are typically observed. Color normalization has been found useful in standardizing the color appearance of H&E stained histopathological images prior to quantitative analysis with machine learning (using handcrafted features). However, its usefulness has not been previously studied when deep convolutional neural networks (CNNs) are used in classifying H&E stained breast cancer histopathological images. In this paper, we have adopted a representative CNN for classifying breast cancer histopathological images and evaluated the benefit/necessity of color normalisation using the commonly used Macenko, Khan and Reinhard color normalization methods. The representative CNN was implemented in-house and was verified. The BreaKHis dataset was used to train and test the CNN model. The preliminary results did not show significant superiority in the CNN performance when color normalization was used to standardize the color appearance of histopathological image. Furthermore, the classification performance of a magnification-independent CNN is comparable to that of magnification-specific CNNs with an additional benefit of a simpler classification scheme and training for only one CNN models (rather than multiple magnification-specific models). It may also have an advantage in clinical practice when the magnification factor of a histopathological image is not known.

Keywords: Deep learning, convolutional neural networks, H&E stains, color normalization, breast cancer histopathological images

1. INTRODUCTION

Histopathological analysis is the corner stone in many disease diagnosis, including breast cancer. In a breast cancer histopathological analysis, biopsied breast tissues are prepared, sectioned, mounted on slides and stained with hematoxylin and eosin (H&E). The H&E stains enhance the tissue structures for visual examination by pathologists using a light microscope. Histopathological analysis based on visual examination by pathologists is a time-consuming and laborious task. The subjectivity nature of the examination and the inter- and intra-observer variability raise the question of quantitative reproducibility and consistency in assessments. Computer-aided histopathological image analysis is a key element in digital pathology. However, the variability in color appearance in H&E stained histopathological images is problematic for automated image analysis.

Pathologists can easily adapt to the variability in color appearance without hindering their capability in histopathological diagnosis. However, color variability can adversely affect the performance of computer based analysis. Color normalization (or color standardization) has become a useful tool in quantitative analysis of color stained histopathological imagesA number of color normalization methods,[1,2,3,4] can be found in the literature. Among them, the methods proposed by Macenko et al.,[2] Khan et al.[3] and Reinhard et al.[4] are widely used. Running codes of these methods are available from the University of Warwick.[5]

Recent years has seen deep learning and deep convolutional neural networks (CNNs) drawing a lot of attentions and becoming increasingly popular among researchers. CNN is a type of deep learning or deep neural network particularly for image recognition. Neural network has been studied for many years. The recent sudden surge of interest in the topic

Further author information:
G.L.: gobert.lee@flinders.edu.au
M.B.: mariusz.bajger@flinders.edu.au
K.C.: clar0737@flinders.edu.au

14th International Workshop on Breast Imaging (IWBI 2018), edited by Elizabeth A. Krupinski, Proc. of SPIE Vol. 10718, 107181E · © 2018 SPIE
CCC code: 1605-7422/18/$18 · doi: 10.1117/12.2316613

is not unrelated to the AlexNet CNN model[6] winning the 2012 ImageNet competition. CNN model are computationally expensive to train due to the very large number of parameters (the AlexNet ImageNet model[6] has 60 million parameters and 650,000 neurons). However, the key attraction of CNN is that it does not required field-knowledge, making it particularly useful when indepth knowledge of the field is lacking. As such, it is conjectured that the color normalization process that has been found useful with handcrafted feature based machine learning algorithms may not be having the same effect when CNN is subsequently use. This has not been previously studied.

In this paper, we investigate the benefit/necessity of color normalization when CNN is used to classify H&E stained breast cancer histopathological images. The AlexNet[6] is widely used in research areas and has been adapted to many different applications including medical and histopatholgoical image analysis. As the investigation is on the usefulness of color normalization when CNN is subsequently used, and there are many CNN models, we first need to decide on a representative CNN model. After the successful implementation of a representative CNN model, the benefit/necessity of color normalization will be evaluated using three widely used color normalization methods. In the following sections, Section 2 describe experimental setup including the BreaKHis[7] breast cancer histopathological image dataset, the convulotional neural networks, color normalization methods as well as the training protocol. Section 3 presents the findings and discuss the results. Concluding remarks will be presented in Section 4.

2. METHODS AND MATERIALS

In investigating the benefit/necessity of color normalization when CNN is used to classify H&E stained breast cancer histopathological images, a representative CNN is required. The AlexNet[6] is first published for the application of scene images in the 2012 ImageNet competition. It is widely used in research areas and has been adapted to many different applications, including medical and digital pathological areas. Thus, an AlexNet-like CNN could be considered as a representative CNN. In particular, this study is concerned with the classification of H&E stained breast cancer histopathological images, a CNN model with successful application in this particular type of images is desirable. Two such CNN models were found in the literature, the CNN models proposed seperately by Spanhol et al.[8] and Bayramoglu et al.[9] In addition, the above CNNs are both trained and tested using the same BreaKHis[7] dataset and the same cross-validation training and testing protocol.[7] This facilitates comparison among CNNs. Judging by the performance and other attributes (see Section 3) the Bayramoglu CNN was adopted as the representative CNN to be employed in the investigation of the color normalization effect. A representative CNN was then implemented and its performance was verified. In the following Section 2.1 describes the BreaKHis[7] dataset and Section 2.2 describes the architecture of the representative CNN and its implementation. Color normalisation methods and the training and testing protocol are described in Section 2.3 and Section 2.4, respectively.

2.1 BreaKHis Dataset

The publicly available BreaKHis[7] breast cancer histopathology image dataset was used in this study. Details of the dataset can be found elsewhere.[7] A brief summary is included as follow. The histopathology images in the BreaKHis dataset were collected from the P&D Laboratory, Brazil, in a clinical study from January 2014 to December 2014. The dataset contains a total of 7909 histopathology images over four magnification factors ($40\times, 100\times, 200\times$, and $400\times$) from 82 patients with surgical open biopsy. Of the 82 patients, 24 have benign breast lesions and 58 have malignant breast cancer. Sections of breast tissues were mounted on slides and stained with hematoxyline and eosin (H&E). The tumor area in each slide was identified manually. On average, about 24 images of the lowest magnification ($40\times$) are required to cover the tumor area (region of interest, ROI). On these 24 $40\times$ images, a similar number of 24 areas were manually selected for further magnification ($100\times$). This procedure is repeated to obtain images of $200\times$ and $400\times$. The image distribution over magnification factor and disease status is shown in Table 1. Final diagnosis of each case was made by experienced pathologists and confirmed by complementary exams such as immunohistochemistry analysis.

Digital histopathology images were obtained with a Olympus BX-50 system microscope coupled by a relay lens (of magnification $3.3\times$) to a Samsung digital color camera SCC-131AN with a 1/3" Sony Super-HAD (Hole-Accumulation Diode) interline transfer charge-coupled device (CCD). The CCD sensor has a specification of a unit cell size of 6.5 μm $\times 6.25$ μm and an image size of 752×582 pixels. Images were acquired in a three-channel RGB TrueColor color space ($24-$bit color depth, 8 bits per color channel). The camera is set for automatic exposure and focusing was done manually. The images produced with the above setup have black borders on both left and right side and text annotations in the upper

left corner and were subsequently cropped, resulting in the image size of 700×460 pixels in the BreaKHis[7] dataset. Out-of-focus images are removed from the dataset in a final visual inspection.

Table 1: BreaKHis[7] Dataset: Image distribution by magnification factor and disease status. (Table reproduced from Spanhol et al.[7])

Magnification	Benign	Malignant	Total
40×	625	1370	1995
100×	644	1437	2081
200×	623	1390	2013
400×	588	1232	1820
Total	2480	5429	7909
No. of Patients	24	58	82

2.2 The Representative CNN

Figure 1 shows the network architecture used in our experiments. It comprises of three convolutional layers and two fully-connected layers. The first convolutional layer has 96 filters of size $3 \times 7 \times 7$. The second contains 256 filters of size $3 \times 5 \times 5$ and the last layer uses 384 filters of size $3 \times 3 \times 3$. Each convolutional layer is followed by ReLU, a max pooling layer and a local response normalization layer. Each fully-connected layer contains 512 neurons and is followed by a ReLU and a dropout layer except the last softmax layer which maps to the final class label for cancer type. The class with maximum probability for the given test image is taken as the predicted class. The drop out ratio was set to 0.5. The max pooling layers take maximum values of 3×3 regions with two-pixel strides.

The Caffe[10] framework was used for the implementation of the above CNN architecture. The framework is a popular choice for researchers implementing CNN as it supports multiple APIs and programming languages. Caffe provides reference models for visual tasks, including the landmark AlexNet ImageNet Model.[6] For this study, a variation of this model (as shown in Figure 1) was implemented.

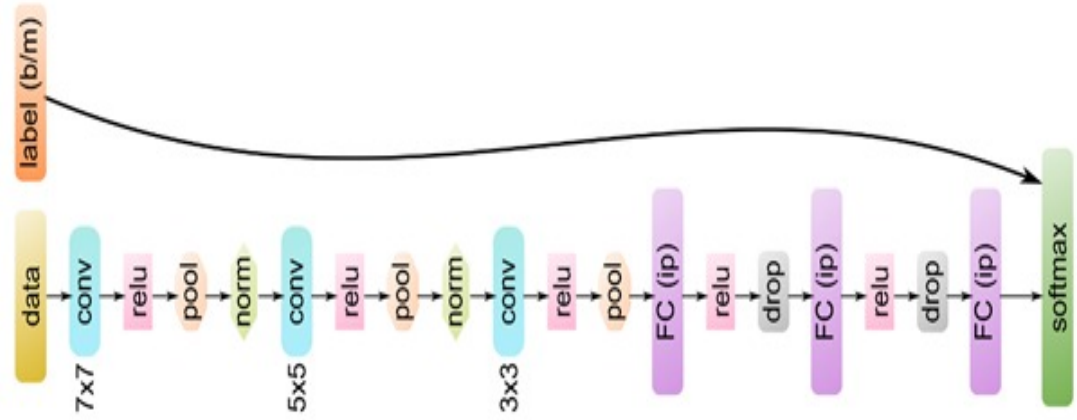

Figure 1: The CNN architecture proposed by Bayramoglu et al.[9] (Schematic diagram extracted from Bayramoglu et al.[9])

2.3 Color Normalization

Hematoxylin and eosin stains (H&E stains) are applied to histopathological samples in order to highlight the structures in the tissue. The hematoxyline stain binds to DNA, giving the blue/purple color to the nucleus. The eosin binds to proteins,

giving the pink color to such as cytoplasm and fibrous stroma. Variability of the color intensity and color appearance is observed. Different staining protocol found in inter-laboratories, stain manufacturing process across vendors, stain storage and other factors contribute to the variability in the color appearance of H&E stained histopathological slides. While this seldom affect the diagnostic ability of pathologists, it is problematic for quantitative analysis using computer algorithms and machine learning (using handcrafted features). Color normalization has been found useful in standardize the color appearance of stained histopathological images. Macenko,[2] Khan[3] and Reinhard[4] are three methods that have been widely used and have achieved good results. Figure 2 illustrate the three color normalization methods.

In this study, we employed these three color normalization methods in investigating if color normalization benefit CNN in classifying H&E stained histopathological images. The stain normalization toolbox[5] available from the University of Warwick was used in performing the color normalization.

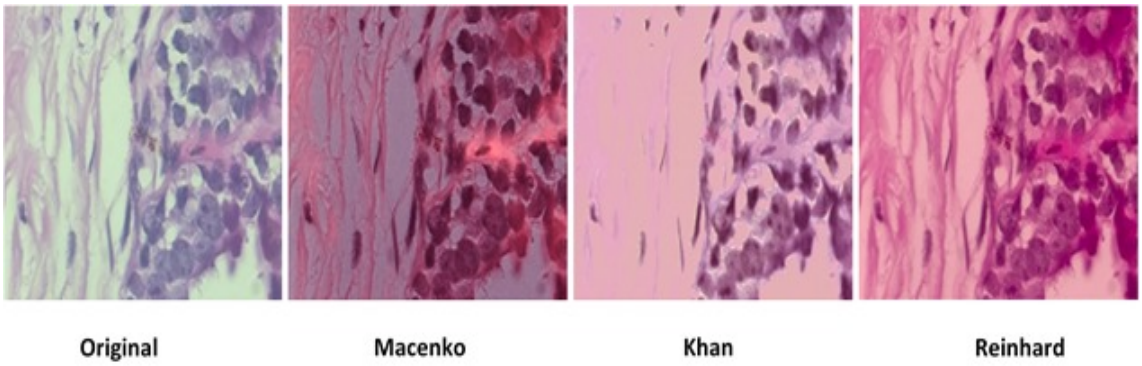

Figure 2: Illustration of the three color normalization methods. A sample histopathological image from the BreaKHis[7] dataset was used showing (left to right) original, Macenko,[2] Khan,[3] Reinhard[4]

.

2.4 Training and Testing Protocol

Along with the BreaKHis dataset, Spanhol et al,[7] published their cross-validation training and testing protocol. In their protocol, the BreaKHis dataset was randomly divided into a training set (70%) and a test set (30%). Efforts were made to guarantee that images from the same patient did not span over both the training set and the test set. Five trials were conducted. Performance of a classification scheme were averaged over the five trials. The grouping of images into the training and test sets has been made available. The CNNs in,[8,9] both employed this training and testing protocal. To allow comparison with these results, the BreaKHis training and testing protocol as well as the performance measures were adopted in this study. In this baseline experiment, the performance of our implemented CNN was established. Color normalization and other preprocessing was not performed on the input images except for the image mean subtraction which is routinely performed to input images as described in the AlexNet.[6] The Spanhol CNN[8] and the Bayramoglu CNN[9] was trained and tested without and with data augmentation, respectively. Data augmentation was not used in this study.

3. RESULTS AND DISCUSSIONS

Experimental results are presented in this section. Section 3.1 compares the magnification-specific and magnification-independent CNNs and verifies our implementation. Section 3.2 reports the results in investigating the benefit/necessity of color normalization of H&E stained histopathological images when classification is performed with CNNs.

3.1 Magnification-Specific and Magnification-independent CNN

Table 2 shows the performance of two CNN models reported in the literature. The two CNNs are both representative CNNs adapted from the AlexNet.[6] The first one (Spanhol's) is magnification-specific. That is, the CNN model is trained on images of a specific magnification and tested on that magnification. Whereas the second one (Bayramoglu's) is magnification-independent. That is, the CNN model is trained on images of all four magnifications ($40\times, 100\times, 200\times, 400\times$).

Table 2: Performance comparison with other CNN-based classification studies (malignant/benign) on BreaKHis database. Table entries are accuracies in percentage.

		Magnification Factors				
		40×	100×	200×	400×	Magnification independent
Spanhol[8]	Patient Level	90.0%	88.4%	84.6%	86.1%	–
	Image Level	85.6%	83.5%	83.1%	80.8%	–
Bayramoglu[9]	Patient Level	83.1%	83.2%	84.6%	82.1%	83.3%
	Image Level	–	–	–	–	–

When comparing Spanhol's and Bayramoglu's results, the classification performances are similar except Spanhol's seems to perform better in lower magnification while Bayramoglu's has a steady performance across all magnification factors. The magnification-independence in Bayramoglu's has an advantage of simplicity having only one CNN for all magnification factors and will reduce training time and complexity of a scheme. For example, in the second stage of this study, a representative CNN is retrained when input images are normalized with three different color normalization methods. Using Spanhol's CNNs, a total of 12 (4 magnification factors $\times 3$) CNNs need to be trained while Bayramoglu's has only 3 CNNs to be trained. Based on this, and that the Bayramoglu's and the Spanhol's CNNs have comparable performance, the Bayramoglu's magnification-independent CNN is preferred over Spanhol's magnification-specific CNNs in deciding a representative CNN to be implemented for the second stage (color normalization) in this study. In clinical practice, the magnification-independent CNN may also has an advantage when the magnification factor of the image is not known.

Verification of our in-house implementation of Bayramoglu's[9]is performed by comparing the classification performance of Bayramoglu's and the in-house implementation. In the in-house CNN implementation, the Bayramoglu's ([9]) CNN architecture (Figure 1) was used. However, data augmentation was performed in [9] but not in our study. On the other hand, [9] subsampled all images to the size of 100×100 while we retained the original image resolution. Overall, the average accuracy of our CNN implementation is 80% which is comparable to that of Bayramoglu's.

3.2 Necessity of color normalization with CNN

Table 3 shows the effect of color normalization on the performance of a representative CNN classifying H&E stained breast cancer histopathological images as benign or malignant. The use of color normalization methods Macenko[2] and Khan[3] produced results marginally better, if at all, than no color normalization. On the other hand, the use of Reinhard[4] produced results that were significantly worse than no color normalization which is unexpected.

Color normalization have, so far, been found useful in analysing color stained histopathological images using handcrafted features based computer algorithms.? From Table 3, the superiority in employing Macenko and Khan (two of the widely used color normalization techniques) is not observed. This may suggest that the CNNs were effective in learning the task in the presence of color variation. Or, that the color variability in the dataset is smaller than expected. Section 2.1 described that the BreaKHis dataset was collected from a single clinical facility with a single microscope-camera setup. This well-controlled data collection environment could have resulted in smaller variabilty in the color appearance when compare to data collected from multi-centre, with different equipment and protocol from laboratory to laboratory. The latter can be investigated by using dataset collected from multi-centre or employing more than one dataset.

Existing color normalization methods were mostly developed for machine learning schemes employing handcrafted features (as opposed to deep learning/CNN). These schemes may or may not perform other prepossessing processes on the input image. CNN, on the other hand, typically perform (channel-wise) mean subtraction on the images as described in AlexNet.[6] This may have an interaction with specific color normalization method and need to be investigated.

Table 3: Effect of color normalization on CNN. Classification (benign/malignant) performance on the BreaKHis[7] dataset with and without color normalization (Macenko,[2] Khan[3] or Reinhard[4]). Table entries are the 5-fold averages.

	No color normalize	Macenko[2]	Khan[3]	Reinhard[4]
Sensitivity	90.9%	89.2%	88.4%	88.1%
Specificity	56.4%	62.6%	63.6%	44.9%
Accuracy	79.2%	80.2%	80.0%	73.5%

4. CONCLUSION

We have presented a preliminary study on the benefit/necessity of color normalization when classifying color stained breast cancer histopathological images using deep convolutional neural network. We have investigated and implemented a representative CNN with the use of the BreaKHis breast cancer histopathological image dataset. The results show, at best, a marginal benefit with Macenko's and Khan's color normalisation methods and an adverse effect with Reinhard's. The adverse results observed with Reinhard's color normalisation method warrants further investigation of interaction, if any, between color normalisation and the routine mean subtraction preprocessing used in AlexNet-like CNN. The marginal or no superiority when Macenko's and Khan's color normalisation methods are used in correcting the variability in the color appearances of stained histopathological images in preparation for CNN classification could indicate that CNN is capable of learning in the presence of unwanted variability in color appearances. However, due to the well-controlled environment (e.g. single-centre and single microscope and camera setup in digital image acquisition) that the BreaKHis dataset was collected from, the variability in color appearances may be less than that observed in images collected from multi-centre. This should be verified using another dataset collected from multi-centre, or combining several datasets together.

REFERENCES

[1] Vahadane, A., Peng, T., Sethi, A., Albarqouni, S., Wang, L., Baust, M., Steiger, K., Schlitter, A. M., Esposito, I., and Navab, N., "Structure-preserving color normalization and sparse stain separation for histological images," *IEEE Transactions on Medical Imaging* **35(8)**, 1962–1971 (2016).

[2] Macenko, M., Niethammer, M., Marron, J. S., Borland, D., Woosley, J. T., Guan, X., Schmitt, X. C., and Thomas, N. E., "A method for normalizing histology slides for quantitative analysis," in [*Biomedical Imaging: From Nano to Macro. ISBI 2009. IEEE Interantional Symposium on*], 1107–1110 (2009).

[3] Khan, A. M., Raipoot, N., Treanor, D., and Magee, D., "A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution," *IEEE Transactions on Biomedical Engineering* **61(6)**, 1729–1738 (2014).

[4] Reinhard, E., Ashikhmin, M., Gooch, B., and Shirley, P., "Color transfer between images," *IEEE Computer Graphics and Applications* **21(5)**, 34–41 (2001).

[5] Trahearn, N. and Khan, A., "Stain normalization toolbox for MATLAB." Last Assessed: 17-08-2017.

[6] Krizhevsky, A., Sutskever, I., and Hinton, G. E., "Imagenet classification with deep convolutional neural networks," in [*Advances in neural information processing systems*], 1097–1105 (2012).

[7] Spanhol, F. A., Oliveira, L. S., and Heutte, L., "A dataset for breast cancer histopathological image classification," *IEEE Transactions on Biomedical Engineering* **63(7)**, 1455–1462 (2016).

[8] Spanhol, F. A., Oliveira, L. S., and Heutte, L., "Breast cancer histopathological image classification using convolutional neural networks," in [*International Joint Conference on Neural Networks (IJCNN)*], 2560–2567 (2016).

[9] Bayramoglu, N., Kannala, J., and Heikkila, J., "Deep learning for magnification independent breast cancer histopathology image classification," in [*23rd International Conference on Pattern Recognition (ICPR)*], 2440–2445 (2016).

[10] Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Grishick, R., Guadarrama, S., and Darrell, T., "Caffe: Convolutional architecture for fast feature embedding," in [*Proceedings of the 22nd ACM International conference on Multimedia*], *ACM*, 675–678 (2014).