Colorimetrical evaluation of color normalization methods for H&E-stained images

Liu, Jocelyn, Lam, Samuel, Lemaillet, Paul, Cheng, Wei-Chung

Jocelyn Liu, Samuel Lam, Paul Lemaillet, Wei-Chung Cheng, "Colorimetrical evaluation of color normalization methods for H&E-stained images," Proc. SPIE 11603, Medical Imaging 2021: Digital Pathology, 116030U (15 February 2021); doi: 10.1117/12.2582281

SPIE.

Event: SPIE Medical Imaging, 2021, Online Only

Colorimetrical Evaluation of Color Normalization Methods for H&E-Stained Images

Jocelyn Liu[1,2], Samuel Lam[1,3], Paul Lemaillet[1], Wei-Chung Cheng[1]

[1]Division of Imaging, Diagnostics, and Software Reliability, Office of Science and Engineering Laboratories, Center for Devices and Radiological Health, U.S. Food and Drug Administration, Silver Spring, MD 20993, USA
[2]University of Southern California, Los Angeles, CA 90089, USA
[3]University of Maryland, College Park, College Park, MD 20742, USA

ABSTRACT

Color normalization is one of the pre-processing steps employed by many deep learning-based algorithms used for aiding pathology diagnoses with whole-slide images. Due to variability in tissue type, specimen preparation, staining protocol, and scanner performance, whole-slide images acquired from different sources may exhibit pronounced color variability that hinders algorithms from executing effectively. In the literature, numerous methods have been proposed to color-normalize hematoxylin and eosin (H&E)-stained images. However, the objective of color normalization has not been colorimetrically defined or evaluated beyond visual comparison. In this study, a quantitative metric, *color normality*, was defined to evaluate the degree of color similarity between images involved in a color normalization process. The pixel-wise spectral data of eight H&E-stained tissue slides were optically measured as the ground truth to test the Reinhard, Macenko, and Vahadane methods. Principal component analysis was conducted on the spectral data to derive a new color normalization method as the reference. Experiment results show that the H&E color gamut needs to be expressed with three components, but the widely used Macenko and Vahadane methods compressed the three-dimensional color gamut volume into a two-dimensional surface and reduced color gamut volumes by 40% or more. None of the color normalization methods could achieve a color normality of greater than 0.6174 when the image was not self-normalized.

Keywords: color normalization, digital pathology, hyperspectral imaging, whole-slide images

1. INTRODUCTION

The color information presented in digital pathology is the result of an involved process of realization from a colorless tissue slice to the final digital color image. The color of a tissue slide is determined by factors in a three-stage pipeline: the tissue sample, histological process, and image acquisition. The tissue sample determines how and which cell types will react with chemicals in the histological stage to exhibit different colors. The histological process consists of dozens of steps in both the fixation and staining protocols; any slight variation in each of the steps or chemicals can pronouncedly shift the resultant color. Lastly, the image acquisition stage is a complicated optical-to-digital conversion process which depends on both the hardware and software components of the scanner. As a result, even for the same tissue sample, the color can vary greatly across different histology labs and scanners.

Thanks to the unparalleled adaptation capability of the human color vision system, color variation does not usually trouble pathologists when the glass slides are directly viewed under a light microscope. However, color variation presents a challenge in digital pathology when artificial intelligence/machine learning (AI/ML) algorithms— where the mechanism of usage of color information in discriminative tasks is not fully understood— fail to deliver equivalent diagnostic performance as the image color deviates from the training sets. A common solution is to introduce an additional pre-processing step, known as color normalization, to transform the color for each pixel.

Medical Imaging 2021: Digital Pathology, edited by John E. Tomaszewski, Aaron D. Ward, Proc. of SPIE Vol. 11603, 116030U · © 2021 SPIE
CCC code: 1605-7422/21/$21 · doi: 10.1117/12.2582281

In the literature, numerous methods have been proposed to color-normalize H&E-stained images. A comprehensive survey of previous color normalization methods is omitted, as it is beyond the scope of this paper. Three frequently used color normalization methods, Reinhard [1], Macenko [2], and Vahadane [3], that were included in this study are described in Section 2.

The efficacy of such color normalization techniques has also been studied and classified in the literature, but a general consensus within the discourse has not been reached. For example, Sethi et al. reported that applying the Vahadane and Khan algorithms incrementally increased the pixel classification accuracy of two comparative models, including a convolutional neural network (CNN) [4]. Furthermore, Khan et al. presented a color normalization algorithm and demonstrated the nontrivial importance of colorimetrical consistency in both pixel-based color segmentation and texture-based tumor segmentation of histopathological images as well [5]. Other studies, however, suggested that color normalization algorithms, when applied to highly variable stained datasets, could modify features recognized by CNN-based algorithms and negatively affect nuclei segmentation [6]. Magee et al. used Hotelling's T-square statistic as a basis of three-channel (RGB) comparisons between the normalized and target images. However, this metric is unintuitive to interpret, and the proposed benchmark value is solely statistical in nature, which does not translate easily into the visual domain [7][8]. Ziaei et al. used a pixel-to-pixel color difference value averaged over a large dataset in order to evaluate the color differences before and after an applied normalization algorithm. Experiment results showed that the application of a color normalization algorithm, among which they assessed Reinhard, Macenko, and Vahadane, in fact decreased the average color difference between their paired datasets [8]. Roy et al. proposed a new methodology based on Reinhard's implementation, and compared against others using locally defined metrics with the Pearson correlation coefficient, absolute mean color error (AMCE), and a defined contrast difference metric, respectively [9]. However, beyond comparative statistical and empirical analyses, a rigorous, colorimetrically defined framework for color normalization is still lacking.

In this study, a quantitative metric, *color normality*, was defined to evaluate the degree of color-wise similarity between images involved in a color normalization process. The pixel-wise spectral data of eight H&E-stained tissue slides were optically measured as the input data to test the Reinhard, Macenko, and Vahadane methods. Principal component analysis was conducted on the spectral data to derive a new color normalization method as the reference.

The specific aim of this study is to answer the following questions: (1) What is the colorimetrical definition of color normalization? (2) How can we measure the degree of color normalization? (3) Can the color normalization methods used in digital pathology be applied to general color images?

The approach used in this study is different from the previous ones in the following ways: (1) the test images were obtained by optical measurement of actual glass slides of various organs; (2) the proposed evaluation method was based on a perceptually linear color space; (3) the reference color normalization method was based on the spectral data.

2. METHODS AND MATERIALS

The study design is illustrated in Figure 1 to lay out the materials, instrumentation, input data, test subjects, reference, and evaluation method.

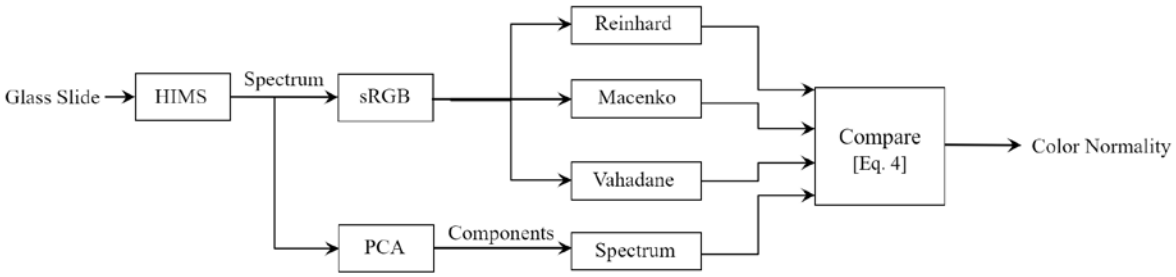

Figure 1. Overview of the study design, where the comparative metric *color normality* derives from color gamut volume analysis; the delineative Eq. 4 is defined in section 2.6.

2.1 Tissue samples

Eight formalin-fixed, paraffin-embedded (FFPE), H&E-stained tissue samples were selected from eight tissue microarray slides (US Biomax, Inc., MD, USA) to represent a spectrum of human organs and associated diseases. Appropriate, histologically meaningful regions of interest (ROI) for each tissue type were selected for various organ-specific features as listed in Table 1.

Table 1. Tissue types and tissue-specific characteristics used in the study.

Id#	Organ	Selection of region of interest
1	bladder	Prominently stained urothelium, or transitional epithelium, lining.
2	brain	A diversity of glial cells (namely, astrocytes and oligodendrocytes) in the proximity of blood vessels.
3	breast	Juxtaposition of stroma and lobules/ducts.
4	colon	Clear stratification of the lamina propria, muscularis mucosa, and muscularis propria, respectively
5	kidney	Well-preserved renal corpuscles and a prominent proximal convoluted tubule.
6	liver	Healthy hepatocytes and a well-preserved portal tract.
7	lung	Prominent smooth muscle lining the bronchioles and well-preserved alveoli.
8	uterine	Clear stratification of the squamous mucosa and stroma.

2.2 Measurement of color truth

The spectral transmittance of each sample was measured at the pixel level using a hyperspectral imaging microscopy system (HIMS) [10][11][12], which was built upon a conventional upright light microscope (AxioPhot 2, Carl Zeiss Microscopy, NY, USA) equipped with a 20X objective (Plan-Apochromat 20x NA=0.8, Carl Zeiss Microscopy). The slide was set on a motorized stage system (MAC 6000, Ludl Electronic Products Ltd, NY, USA) and was illuminated by a tunable light source (OL490 Agile light source, Gooch and Housego, TX, USA), with λ = 380 nm to 780 nm in steps of 10 nm. A calibrated scientific monochrome CCD camera (Grasshopper3 9.1 MP Mono USB3 Vision, Point Grey Research Inc., BC, Canada) with a linear response was used as a light detector to measure the luminance of each pixel. A series of 41 images each with a resolution of 844×676 were acquired. The set of 41 images was compared with a flat-field image to calculate the spectral transmittance, which would be used by the proposed spectrum-based color normalization method. The spectral power distribution of the CIE D65 illuminant was mathematically applied to the spectral transmittance to obtain the CIELAB coordinates, from which the sRGB images were derived to test the color normalization methods. The spectral data are publicly available on GitHub [12].

2.3 Previous color normalization methods

Three frequently used color normalization methods were tested in this study. A well-referenced method proposed by Reinhard et al. used simple linear statistical analysis to impose the color characteristics of a target image on a source image in the decorrelated $l\alpha\beta$ color space, which is based on cone cell responses [1]. The standard deviation and mean of a source image were scaled to that of a target image; in other words, the size of the color gamut of the source image was equalized to that of its respective target. In a histopathological context, this global color transfer can negatively affect the contrast characteristics and background luminance of the source image on which it is applied [2][3][9]. This method was proposed for general color images and not optimized for histology images.

Macenko et al. implemented a color deconvolution algorithm which first converted an H&E-stained image into a two-dimensional optical density (OD) space [2]. A key assumption was that two extreme stain vectors, spanning the planar OD space, were sufficient to describe the total color gamut of a dual-stained histopathological image, such as H&E. The optimal linear combination was found by using the principle of singular value decomposition, and the robust extremes were found and converted back into OD space as the optimal and independent hematoxylin and eosin stain vectors. As pointed out in the literature, the foundation of using optical density, which must be non-negative, may limit performance and produce abnormal results [3].

Vahadane et al. extended Macenko's work in stain separation by introducing non-negativity and sparsity constraints in the matrix factorization technique, based on a biological interpretation of the stain matrix. Structure-preserving normalization was performed by combining stain density maps with a two-stain color basis extracted from the target image [3].

2.4 Spectrum-based color normalization method

Using the spectral measurements of the transmittance of the reference tissue slides, we designed an approach using principal component analysis (PCA) to conduct the color normalization. For a source image *S* and a target image *T* of identical dimensions (*l, w)* comprising $k = 1 \times w$ pixels, the color normalization process converted *S* into the result image *R* that minimized the color gamut difference between *T* and *R*. A PCA software tool (MATLAB, MathWorks) was used to process the spectral transmittance data, a $k \times 41$ matrix, to generate a 41×41 eigenvector matrix (i.e., principal components) and the weighting coefficients of *S* (or *T*). To perform the color normalization, the weighting coefficients of *S* were plotted in the principal component space of *T* to generate the resultant color-normalized image *R*. Only the first three principal components were used for reconstruction in this study, but the process is generalizable for up to all 41 components. The final spectral transmittance data was mathematically combined with the CIE standard illuminant D65 to obtain the spectral power distribution, which was then converted into the CIEXYZ and CIELAB color spaces. Finally, the CIELAB data was converted to the sRGB color space as a PNG image for visualization.

2.5 Color gamut volume arithmetic

The color gamut volume of each image was evaluated in the CIELAB color space by counting the occurrences of all pixels. The CIELAB color space consists of three perceptually linear dimensions: L^*, a^*, and b^*, where L^* indicates the lightness attribute, from dark to bright, a^* the chromaticity attribute, from green to red, and b^* the chromaticity attribute, from blue to yellow. We divided the CIELAB color space into unit cells (i.e., $1\ L^* \times 1\ a^* \times 1\ b^*$) for binning all pixels *p* belonging to image *i*. Equation 1 defines the cardinality of the set that contains all pixels within the unit cell located at (L,a,b). $L^*(p)$, $a^*(p)$, and $b^*(p)$ are transfer functions that convert a pixel *p* from the sRGB color space to the CIELAB color space. The color gamut volume size of image *i*, $CGV_{Size}(i)$, is defined by the number of non-empty bins as formulated in Eq. 2. The intersection between two color gamut volumes of images *i* and *j*, $CGV_{Int}(i,j)$, is defined in Eq. 3 as the count of bins that are non-empty for both images.

$$B_i(L,a,b) = |\{\quad p:\quad p \in i \; and \; L \le L^*(p) < L+1 \; and \; a \le a^*(p) < a+1 \; and \; b \le b^*(p) < b+1\}| \tag{1}$$

$$CGV_{Size}(i) = \sum_{L,a,b} \begin{cases} 1 & if\ B_i(L,a,b) > 0 \\ 0 & otherwise \end{cases} \tag{2}$$

$$CGV_{Int}(i,j) = \sum_{L,a,b} \begin{cases} 1 & if\ B_i(L,a,b) > 0 \; and \; B_j(L,a,b) > 0 \\ 0 & otherwise \end{cases} \tag{3}$$

2.6 Color normality

The mathematical objectives of existing color normalization methods have not been described in the literature, but are understood to be as matching the color gamut volumes between two images. Thus, we define a metric, *color normality*, to measure the degree of color normalization between images *i* and *j* by calculating the mutual inclusion of two color gamut volumes as formulated in Eq. 4. Equation 5 defines a ratio to indicate whether the color gamut volume of image *i* expanded ($CGV_{Scale} > 1$) or shrank ($CGV_{Scale} < 1$) after normalization with image *j*.

$$ColorNormality(i,j) = \sqrt{\frac{CGV_{Int(i,j)}}{CGV_{Size(i)}} * \frac{CGV_{Int(j,i)}}{CGV_{Size(j)}}} \tag{4}$$

$$CGV_{Scale}(i,j) = \frac{CGV_{Size}(j)}{CGV_{Size}(i)} \tag{5}$$

2.7 Implementation

All of the previous color normalization methods were implemented in Python 3 based on the open source code StainTools published in GitHub by Peter554/StainTools and wanghao14/Stain_Normalization under the MIT license. The programs ran in the Anaconda JupyterLab environment in an Ubuntu system. The programs relied on libraries including OpenCV (4.3.0.6) for color space conversion and SPAMS (2.6.1) for sparse matrix operations.

3. EXPERIMENT RESULTS

Figure 2 shows the RGB input images for the color normalization methods. These RGB images were converted from the spectral transmittance data measured from the eight samples and considered as the color truth of the glass slides. The color gamut size is noted in the bottom left in each image. Notice the visual variation in color gamut across the tissue types.

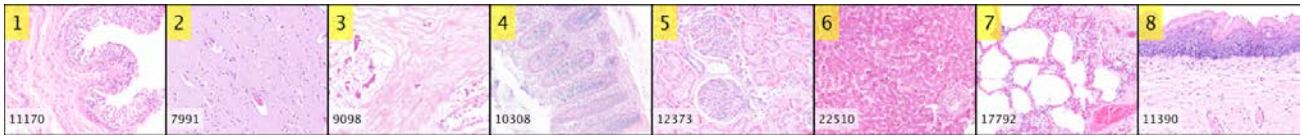

Figure 2. Images reconstructed from spectral transmittance data. The number in the white box is the color gamut size.

After removing the white background (i.e., empty areas), principal component analysis was applied to the spectral transmittance dataset. Figure 3 shows the first three principal components plotted in the spectral space. It was found that across the set of eight images, three components described an average of 98.68% explained variance in each dataset. In each image case, the fourth component increased less than 1% of explained variance, so three principal components were determined to adequately represent all eight H&E-stained images. Notice the correlation, visually presented as the similarity between the component curves, across the eight images for each of the three components.

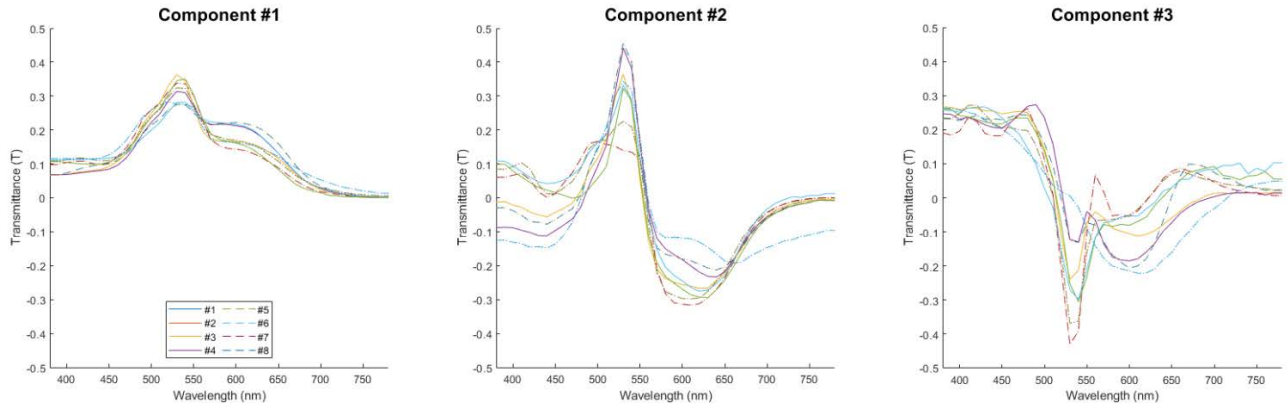

Figure 3. The first (left), second (center), and third (right) principal components of the eight images, corresponding to tissues #1 through #8 as defined in Table 1, in the spectral space.

3.1 Color normalization results

Figure 4 shows the results of each source-target pair processed by the four color normalization methods in an 8×8 matrix. Each of the eight images was divided into five regions. The four quadrants were processed by the four color normalization methods – Spectrum-based (upper left), Reinhard (upper right), Macenko (lower left), and Vahadane (lower right), while the center region is occupied by the source image without any processing. The cell in the *i*-th row, *j*-th column indicates the *i*-th source image color-normalized against the *j*-th target image. Thus, the center region is constant across each row, while the four quadrants vary with the corresponding target image. Qualitatively, for example, we can convince ourselves that column four from the left, corresponding to source images #1 through #8 color-normalized against target image #4, yields results tinted a deeper purple shade (quadrants) compared with the original image (center). This is consistent with the color gamut spanned by target image #4 and the deeper purple staining of the mucosa.

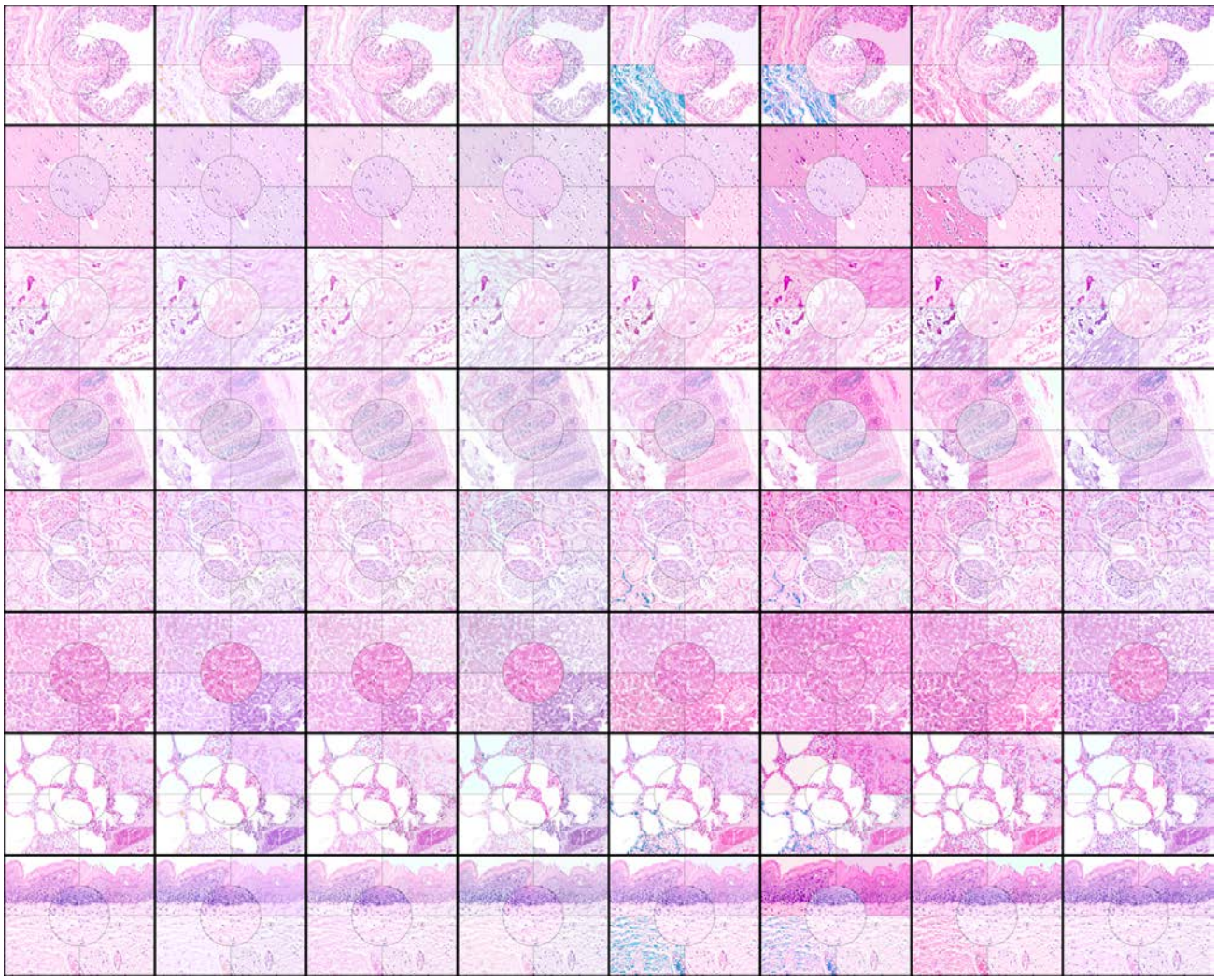

Figure 4. The 8×8×5 image matrix showing the results of color normalization of 8 source images with respect to 8 target images by using 4 different color normalization methods. Each square is occupied by the results from the Spectrum-based (upper left), Reinhard (upper right), Macenko (lower left), and Vahadane methods (lower right) surrounding the unchanged source image (center). Of the source-image pair, the source image is constant across the row, and the target image iteratively varies across the corresponding columns of images #1, #2, …, #8. For example, in the first row only, the leftmost square contains the results of source image #1 color-normalized against itself as the target, the second square contains the results of source image #1 color-normalized against target image #2, the third contains the results of source image #1 color-normalized against target image #3, and so on.

It should be noted that an abnormal color transfer is observed in instances of the Macenko transformation, as shown in the unexpected bright blues apparent in the bottom left portions of columns 4 and 5 in Figure 4. This strong deviation was not reported in previous literature. We anticipate that the algorithm or implementation failed to cope with true colors measured directly from glass slides and generated negative OD values. Regardless, this does not affect the dimensionality-reducing nature of the algorithm, which is the primary feature addressed in the following sections.

3.2 Color gamut distributions

Figure 5 shows the visualizations of the color gamut volumes of the source (#1), target (#8), and color normalization results in the 3D CIELAB space. Notice that while the input source and target images constitute 3D point clouds in the color space, the Macenko and Vahadane methods generated post-processed color gamut volumes that were curved 2D surfaces in 3D LAB space. This dimensionality reduction is directly a result of the assumption that H&E-stained images could be expressed by exactly two stain vectors, such that all reconstructed colors would be constrained to the plane defined by the extracted hematoxylin and eosin stain vectors' spanning space. Consequently, the resultant color gamut only encapsulates a small fraction of the original image gamut, and 3D color information is inevitably lost in the transformation process.

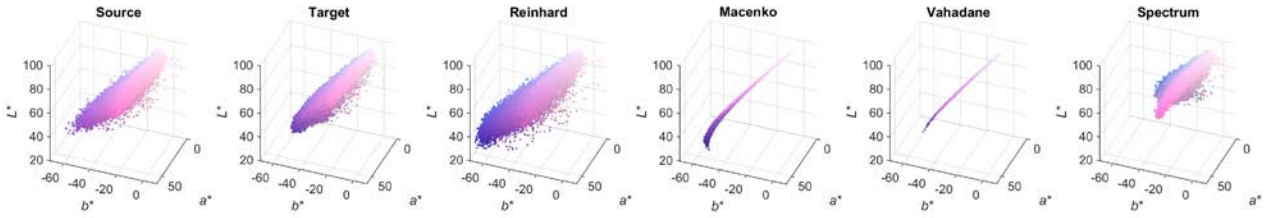

Figure 5. Color distributions in the CIELAB color space; each data point is plotted in its original color. From left to right: Source image #1, Target image #8, then the results of Reinhard, Macenko, Vahadane, and Spectrum-based methods, respectively.

3.3 Color gamut size changes

Figure 6 shows the results calculated to express the changes in color gamut size, relative to the original image, after normalization using Eq. 5. The graphs clearly show that the Reinhard and Spectrum-based methods generated much larger resultant color gamut volumes than their counterparts, sometimes augmenting the target gamut and exceeding 100% coverage, most likely translating to enhanced color contrast in the visual domain. The other two OD-based methods, Macenko and Vahadane, shrank the color gamut to at most 53.77% coverage for the former and 33.18% for the latter, and to as low as 4.08% and 5.71%, respectively, suggesting a significant reduction of color gamut volume as a result of the dimensionality reduction.

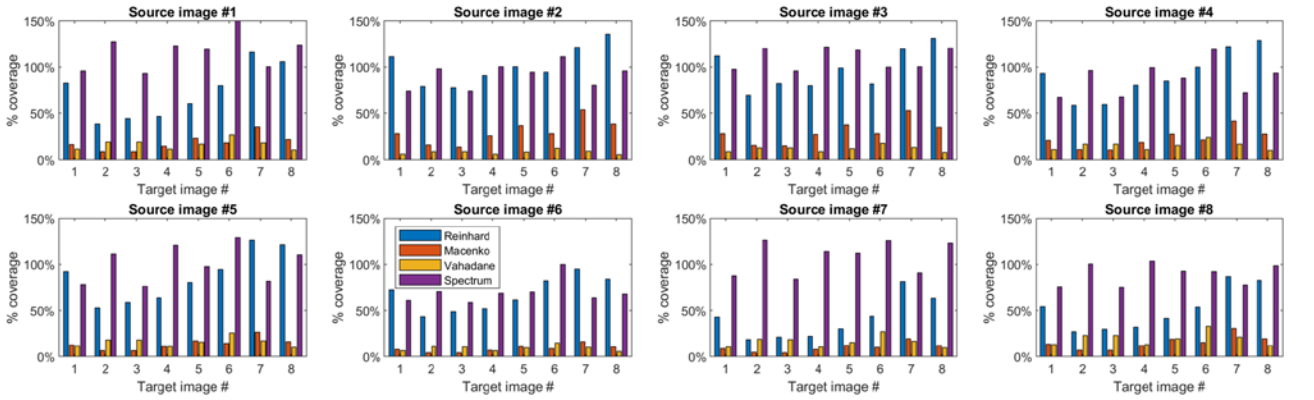

Figure 6. Color gamut volume after color normalization applied to source images #1, #2, ...#8 for all possible target images #1, #2, ...#8, expressed as a percentage of the original gamut.

3.4 Color normality

Figure 7 shows the color normality measures of four methods applied to all 8x8 combinations. No method could achieve a score of higher than 0.6174 when the image was not color-normalized to itself, the highest being achieved at the instance of source and target image pairs #4 and #3 with the Spectrum-based method. Moreover, no method could achieve the ideal color normality of 1.0, even when the image was self-normalized. The maximum color normality was 0.9167 when image #4 was normalized to itself, suggesting that these color normalization methods are irreversible and color gamut information is lost even when the source and target images for normalization are the same. Additionally, the

data reveal that the operative normalization process is not necessarily commutative; for example, the Spectrum-based method performs the best for source image #5 normalized to target image #7, but the Reinhard method performs the best for source image #7 normalized to target image #5. In general, the Spectrum-based method achieved the best results for all target images, and the Macenko and Vahadane method scores were consistently low, never exceeding 0.4.

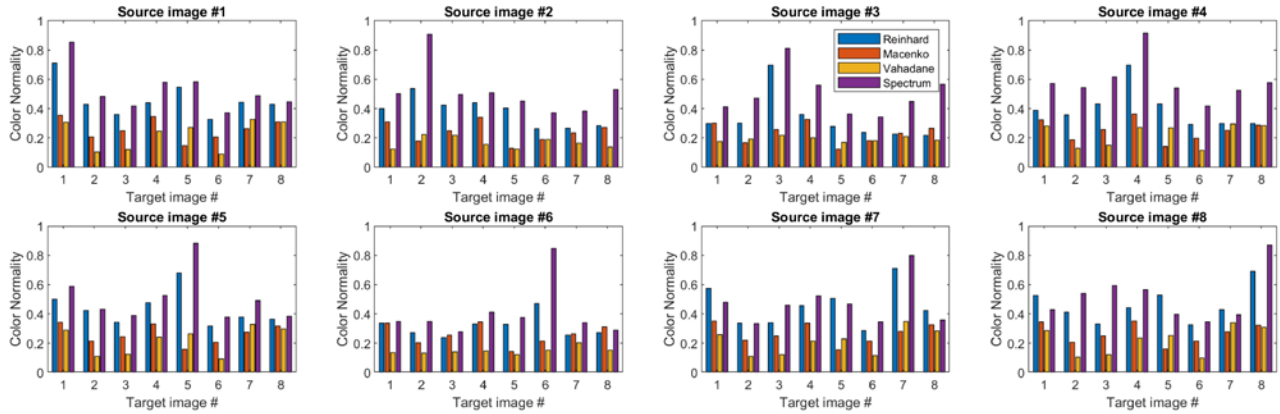

Figure 7. Color normality of four methods applied to source images #1, #2, ...#8 for all possible target images #1, #2, ...#8, respectively.

4. DISCUSSION AND CONCLUSIONS

Color normalization is a common image pre-processing method used by deep learning-based algorithms to improve discriminative performance. The term color normalization was coined by pioneering researchers without a clear colorimetrical definition. In this work, a colorimetrical metric, color normality, was proposed to measure the degree of color similarity between two images involved in a color normalization process. The proposed metric was used to test three color normalization methods, Reinhard, Macenko, and Vahadane. The test images were obtained by a hyperspectral imaging microscopy system to represent the true color variety of eight tissue samples prepared and stained separately. Experiment results show that the two optical density-based methods, Macenko and Vahadane, compressed the originally three-dimensional color gamut volumes into two-dimensional surfaces and reduced the color gamut size significantly. Although the Reinhard method achieved higher color normality, it was nevertheless unable to fully preserve the color information. In other words, these color normalization methods may improve deep learning-based algorithms by pre-processing the images, but they may not be adequate color transforms in the other applications when subtle color information needs to be preserved. Further research needs to be conducted to understand the colorimetrical behavior of more modern color normalization methods, especially those used for deep learning techniques such as artificial dataset augmentation. The proposed color normality can be an objective metric in such comparative studies.

ACKNOWLEDGEMENTS

The authors thank Drs. Mary Barcus, Anant Agrawal, Andrew Lamont, Weijie Chen, Brandon Gallas, Frank Samuelson, and Aldo Badano for their helpful comments and discussions. This study was supported in part by a grant from the Critical Path Initiative at the Food and Drug Administration and by an appointment to the Research Participation Program at the CDRH administered by the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between DOE and FDA. The mention of commercial products herein is not to be construed as either an actual or implied endorsement of such products by the Department of Health and Human Services.

REFERENCES

[1] Reinhard, E., et al., *Color transfer between images.* Ieee Computer Graphics and Applications, 2001. **21**(5): p. 34-41.

[2] Macenko, M., et al., *A method for normalizing histology slides for quantitative analysis*. 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, p. 1107-1110.

[3] Vahadane, A., et al.,*Structure-Preserving Color Normalization and Sparse Stain Separation for Histological Images.* IEEE Trans Med Imaging, 2016. **35**(8): p. 1962-71.

[4] Sethi, A., Sha, L., Vahadane, A., Deaton, R.J., Kumar, N., Macias, V., & Gann, P.H. *Empirical comparison of color normalization methods for epithelial-stromal classification in H and E images*. Journal of pathology informatics, 7, 17, (2016).

[5] Khan, A.M., Rajpoot, N.M., Treanor, D., & Magee, D.R. A., *Nonlinear Mapping Approach to Stain Normalization in Digital Histopathology Images Using Image-Specific Color Deconvolution*. IEEE Transactions on Biomedical Engineering, 61, 1729-1738, (2014).

[6] Pontalba, JT., Gwynne-Timothy, T., David, E., Jakate., K, Androutsos, D., and Khademi, A., (2019) *Assessing the Impact of Color Normalization in Convolutional Neural Network-Based Nuclei Segmentation Frameworks. Front. Bioeng. Biotechnol.* 7:300. doi: 10.3389/fbioe.2019.00300

[7] Magee, D., Treanor, D., Crellin, D., Shires, M., Smith, K., Mohee, K. and Quirke. P., *Color normalization in digital histopathology images*. in Proc. Opt. Tissue Image Anal. Microsc., Histopathol. Endosc., pages 100–111, (2009).

[8] Ziaei, D., Li, W., Lam, S., Cheng, W.C., and Chen, W., *Characterization of color normalization methods in digital pathology whole slide images*, Proc. SPIE 11320, Medical Imaging 2020: Digital Pathology, 1132017 (16 March 2020); https://doi.org/10.1117/12.2550585

[9] Roy, S., Lal, S., and Kini, J. R. (2019). *Novel color normalization method for hematoxylin eosin stained histopathology images*. IEEE Access 7, 28982–28998. doi: 10.1109/ACCESS.2019.2894791

[10] Cheng, W.C., F. Saleheen, and A. Badano, *Assessing color performance of whole-slide imaging scanners for digital pathology.* Color Research and Application, 2019. **44**(3): p. 322-334.

[11] Lemaillet, P. and Cheng, W.C.**,** *Colorimetrical uncertainty of a hyperspectral imaging microscopy system for assessing whole-slide imaging devices.* Biomed Opt Express, 2020. **11**(3): p. 1449-1461.

[12] Lemaillet, P., *Hyperspectral Imaging Microscopy System (HIMS)*, (2020), Github repository; https://github.com/DIDSR/HIMSPEC