EXPLORING CNN-BASED SELF-SUPERVISED ILLUMINATION INHOMOGENEITY COMPENSATION FOR SERIAL OPTICAL COHERENCE TOMOGRAPHY

Joël Lefebvre

Département d'informatique
Université du Québec à Montréal (UQÀM)
Montréal, Qc, Canada

ABSTRACT

Serial blockface histology is a 3D imaging modality that combines a vibratome with a microscope. Whole samples are acquired by sequentially removing small tissue layers with the vibrating blade and by generating a mosaic of several images of the revealed tissue which can be assembled to obtain a 3D representation of the sample at a high resolution. Due to many factors, the acquired mosaic tiles can be affected by complex illumination inhomogeneity that negatively affects the data reconstruction and analysis. Here, we propose a convolutional neural network approach to estimate and compensate the illumination inhomogeneity. The model is trained with simulated vignettes without using illumination ground truth, which is many times harder or even impossible to obtain. Using a small multiresolution dataset consisting in serial OCT images from whole mouse brains, we show that our proposed approach has many advantages compared to an unsupervised *a posteriori* illumination compensation method.

***Index Terms*—** Serial blockface histology, Optical Coherence Tomography, Illumination Inhomogeneity, Convolutional Neural Network, Data augmentation

1. INTRODUCTION

Serial blockface histology (SBH) is a 3D imaging modality that combines a vibratome with a microscope. Whole 3D samples are embedded within an agarose matrix and acquired by sequentially removing small tissue layers with the vibrating blade, and by generating a mosaic of the revealed tissue containing multiple image tiles. The several thousands of images acquired with this procedure can be assembled to obtain a 3D representation of the whole sample at a high resolution [1, 2]. When coupled with optical coherence tomography (OCT), this imaging modality can reveal the 3D distribution of white matter in whole mouse brains without necessitating complex tissue labeling or cleaning as with other optical whole brain 3d imaging modalities such as light-sheet microscopy.

Due to many factors, the acquired tiles within the mosaic can be affected by complex illumination inhomogeneities (also called vignetting), which can negatively affect the data reconstruction and analysis. When degraded by a vignette, the assembled mosaics will contain seam artifacts at locations where neighboring tiles overlap. Thus, it is common practice when generating image mosaics to compensate the vignetting effect before assembling the data. Two general approaches exist: *a priori* methods that will acquire the vignette before an acquisition and use these calibration images to fix the illumination during reconstruction, and an *a posteriori* approach that will estimate the vignette from already acquired image tiles and use the extracted illumination bias to reduce its effect prior to reconstruction. One such *a posteriori* method is the BaSiC Background and Shading Correction algorithm [3]. It uses a low-rank and sparse decomposition to estimate multiplicative flatfield (shading) and additive darkfield (background) vignettes for a sequence of existing tiles.

One drawback of the BaSiC method is that it supposes that at most $< 50\%$ of the tiles contain foreground, otherwise the estimated vignettes converge towards the foreground instead of the background. We also observed that the BaSiC algorithm will not converge correctly when the number of tiles is not sufficient (e.g., when applied to a small 5×4 tiles mosaic as illustrated in Fig.3). On the contrary, when using the algorithm with a large number of tiles convergence can require considerable amount of time and computational resources, which limits the applicability of this technique when working with the large mosaics often required for high magnification serial blockface histology. To address these issues, we propose a Convolutional Neural Network (CNN) based approach inspired by VoxelMorphCNN [4] to estimate the illumination inhomogeneity affecting serial OCT 2D tiles. Due to the lack of illumination ground truth, we propose to generate synthetic vignettes to train the network in a self-supervised way. These experiments have shown promising results compared to the BaSiC method, which we will investigate more thoroughly in future works.

The rest of this paper is organized as follows. Section 2 describes the experimental data, introduces two synthetic vignette generation methods and summarizes the illumination inhomogeneity extraction network architecture. Section

3 presents the results, including examples of synthetic vignettes and a comparison between our method and the BaSiC method. Section 3 also discusses about the performance of the self-supervised strategy and states a few potential extensions of the proposed approach.

2. MATERIALS AND METHODS

2.1. Experimental Data

The data used for this paper was acquired in whole mouse brains with a custom serial OCT system [1, 2]. These datasets were generated by other previous and ongoing research projects in accordance with the animal ethics committee of the Montreal Heart Institute. They consist in 3D mosaics acquired with a serial Fourier-Domain OCT (FD-OCT) at 3 different resolutions using 3x, 10x and 25x magnification objectives. With FD-OCT, each mosaic tile is a 3D volume describing the tissue's optical backscattering at multiple depth within the brain. For the purpose of this paper, the volumetric tiles were converted into 2D tiles by computing average intensity projections along the tissue depth. Each mosaic for every tissue slice was intensity normalized with 0.1% clipping, and the 512×512 tiles were saved as separate jpeg files. In total, 121 mosaics were used (110 mosaics at 3x magnification, 1 mosaic at 10x magnification, and 10 mosaics at 25x magnification) resulting in 6572 tiles for the training dataset and 40 tiles for the validation dataset (2 mosaics out of 110 mosaics at 3x magnification were used for the validation dataset).

2.2. Synthetic vignette generation

For this study, it was assumed that the illumination inhomogeneity consisted only in a multiplicative vignette $V(x, y)$, often called a *flatfield*. This vignette is assumed to be shared for all the image tiles I_k within a same mosaic. The simplified image formation model can be expressed as

$$I'_k(x, y) = V(x, y) \cdot I_k(x, y), \tag{1}$$

where $I'_k(x, y)$ is the observed tile. The goal of the proposed method is to estimate this vignette $\hat{V}$ and use it to compensate the illumination inhomogeneity

$$\hat{I}_k(x, y) = I'_k / \hat{V}(x, y). \tag{2}$$

In many cases, the real flatfield can be hard or impossible to estimate. For this reason, we propose to use synthetic vignettes to train the illumination inhomogeneity extraction network. Two vignette generation methods are proposed: Gaussian vignettes and Zernike vignettes. Both of these models were chosen to create vignettes with some degree of symmetry and smoothness. A Gaussian vignette consists in

$$V_G(x, y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}, \tag{3}$$

where $\sigma = 1$ was used for the Gaussian function standard deviation. A more complex model using Zernike polynomials [5] was also considered for the vignette generation. Zernike polynomials are often used to represent the wavefront distortion in optics, and it forms an orthonormal basis on the unit circle. For this paper, a 5th order Zernike polynomial was used. Its coefficients were generated using uniform sampling between -1 and 1. The vignette was synthesized with

$$W(r, \theta) = \sum_{n,m} C_n^m Z_n^m(r, \theta), \tag{4}$$

where $C_n^m \sim \mathcal{U}(-1.0, 1.0)$ are the Zernike coefficients, and Z_n^m are the n^{th} order polynomials

$$Z_n^m(r, \theta) = \begin{cases} R_n^m(r) \cos m\theta & \text{for } m \geq 0, \\ R_n^m(r) \sin m\theta & \text{for } m < 0 \end{cases} \tag{5}$$

with the radial function $R_n^m(r)$

$$R_n^m(r) = \sum_{l=0}^{(n-m)/2} \frac{(-1)^l (n-l)! r^{n-2l}}{l! \left(\frac{n+m}{2} - l\right)! \left(\frac{n-m}{2} - l\right)!} \tag{6}$$

To generate the vignettes with both models, the 2D simulation domain was first initialized in the range $[-1, 1]$ in the XY directions. Then to represent random illumination centering, scaling and rotations, a random geometry transform was generated. This was performed by using uniform sampling for the rotation angle ($\theta \sim \mathcal{U}(-\pi/2, \pi/2)$) and for the scale ($s_{x,y} \sim \mathcal{U}(1/2, 2)$), and Gaussian sampling for the center translation ($t_{x,y} \sim \mathcal{N}(\mu = 0.0, \sigma = 1/4)$). The spatial transforms were represented as 3×3 matrices and combined to obtain the global spatial transform $\mathbf{M}$. This transformation is applied to the vignettes simulation domain prior to vignettes intensity computation.

$$\mathbf{M} = \mathbf{TS} \cdot \mathbf{R}, \tag{7}$$

with

$$\mathbf{TS} = \begin{bmatrix} s_x & 0 & t_x \\ 0 & s_y & t_y \\ 0 & 0 & 1 \end{bmatrix}, \mathbf{R} = \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{8}$$

the combined translation/scale ($\mathbf{TS}$) and rotation ($\mathbf{R}$) matrices, respectively. Figure 1 show some generated vignettes using the Gaussian model (top row) and the Zernike model (bottom row).

2.3. Vignette Extraction Network

For the vignette extraction network, we used a model inspired by the VoxelMorphCNN [4] network for 3D deformable, pairwise medical image registration. We adapted the encoder/decoder architecture to receive 2D images as input and

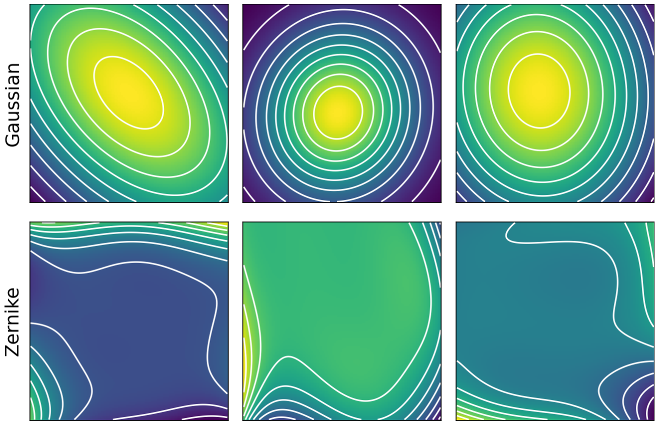

Fig. 1. Example of synthetic vignettes generated with the Gaussian (top) and Zernike (bottom) models, respectively. White contour lines are overlaid to help visualize the generated illumination inhomogeneity.

to predict only 1 scalar image as output instead of a 3D deformation vector field. To summarize, the network consists of 4 parts: (1) an input block, (2) an encoder and (3) decoder linked with skip connection, (4) and a prediction head. The input of the model is a single image formed by concatenating k random tiles extracted for a mosaic grid and affected by the same synthetic illumination inhomogeneity $V(x, y)$. The tiles are concatenated over the channel dimension. The input block performs an early image fusion of the input tiles using large (7×7) 2D convolutions followed by a Leaky ReLU activation, resulting in 48 channels. This is followed by a UNet encoder/decoder network [6, 7] with skip connection over 4 levels. Each layer consists of two convolution blocks (3×3 2D convolution + LeakyReLU), and by a 2D max pooling. The number of features is doubled for each layer. For the decoder part, nearest neighbor interpolation is used for the upsampling. The prediction head consists of 4 convolution blocks (3×3 2D convolution + Leaky ReLU) that progressively reduce the number of features, to finally predict a single feature per pixel. This output is the predicted illumination bias $\hat{V}(x, y)$.

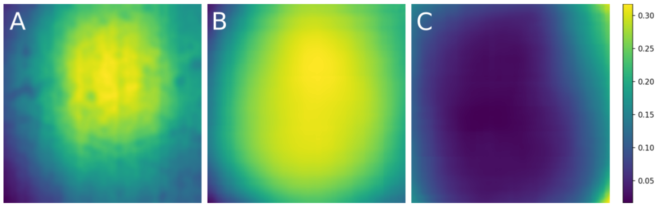

Fig. 2. Extracted vignette using (A) the BaSiC method, and (B) our proposed method with $k = 5$ input tiles. (C) represents the relative vignette standard deviation for our method

2.4. Data augmentation and training

Due to the lack of vignette ground truth and the small dataset size, data augmentation was used to train the vignette extraction network. For every iteration, k tiles were selected randomly within the training dataset. Data augmentation (random scaling, cropping and flipping) was applied to each tile separately to obtain tiles of shape 128×128. Then a synthetic vignette $V(x, y)$ was generated using either the Gaussian or Zernike model (both the model and their parameters are sampled randomly) and all k tiles are multiplied with it. Finally, the tiles were concatenated over the channel dimension to form a single input image for the network. Using the input of shape $(k, 128, 128)$, the network outputs a vignette estimation $\hat{V}$ of shape $(1, 128, 128)$. Similar to VoxelMorphCNN [4], the loss function used for gradient descent was

$$\mathcal{L}(V, \hat{V}) = \mathcal{L}_{sim}(V, \hat{V}) + \lambda \mathcal{L}_{smooth}(\hat{V}), \tag{9}$$

where $\mathcal{L}_{sim}(V, \hat{V}) = \|V, \hat{V}\|_2$ is the $L2$ norm measuring the similarity between the generated and the estimated vignettes, $\mathcal{L}_{smooth} = \|\nabla \hat{V}\|_2$ is a smoothness regularization using the spatial gradient, and $\lambda = 1$ controls the regularization. The training was performed for 625,000 random synthetic samples, separated in 250 epochs of 2,500 samples with a batch size of 8. An Adam optimizer with learning rate $l_r = 0.4$ was used. The model was implemented with PyTorch [8] and trained with Pytorch-Lighting [9] on a NVidia GeForce RTX 3080 GPU. To compare the performance of the proposed method with the BaSiC method, we used our custom Python implementation of the BaSiC algorithm [10].

3. RESULTS AND DISCUSSION

3.1. Model Training

Using simulated vignettes and data augmentation, a modified U-Net encoder-decoder network was trained to estimate the multiplicative illumination bias $V(x, y)$ given k input tiles affected by the same illumination bias. Training the network with this approach takes approximately 40 minutes on a single GPU. During training, we logged the training loss every 15 steps and validation loss every epoch. Both were consistently decreasing when considering their moving average. In addition to training/validation losses, vignettes of 8 random samples were visualized every epoch to assess the training progression. Qualitatively, the network quickly learned to extract Gaussian vignettes (similar to the top-middle vignette in Fig. 1) but required additional training iterations for more complex vignettes (ex. Zernike). Furthermore, initial vignette estimation was affected by many rectilinear artifacts. Many additional training iterations were necessary to obtain spatially smooth vignettes. One possible explanation for these artifacts is with the choice of the nearest neighbor upsampling method in the network's decoder. Adopting another in-

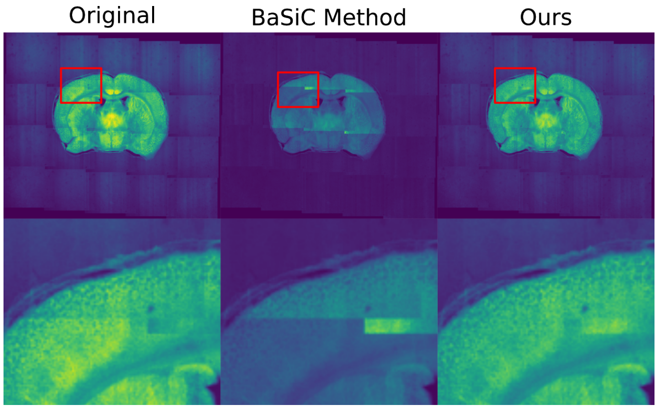

Fig. 3. Comparison between the vignette compensation with the BaSiC method and our proposed method using $k = 5$ input tiles.

terpolation method, or replacing by transposed convolutions or dilated kernels could help with reducing the smoothness artifacts. To evaluate the influence of network architecture on the vignette extraction performance, we trained the network with different number of input tiles ($k \in \{1, 2, 5\}$). These resulted in a final validation loss of $\mathcal{L}_{k=1} = 9.4 \times 10^{-3}$, $\mathcal{L}_{k=2} = 4.5 \times 10^{-3}$ and $\mathcal{L}_{k=5} = 2.0 \times 10^{-3}$, which shows that the synthetic vignette extraction improves as the number of input images increases. This was an expected result, as at least two tiles would be necessary to distinguish between the vignette illumination bias and the underlying real image intensity. The network trained with a single input still is able to get a good estimated of the vignette, especially in cases where the vignette greatly affects the tile appearance (ex: Gaussian vignette with small covariance).

3.2. Inference and illumination compensation

After the training, we compared the extracted vignette predicted by our network with the vignette obtained using the unsupervised BaSiC method (Fig. 2). For inference, we used $N = 25$ permutations of k unaugmented input vignettes originating from the same mosaic. The N extracted vignettes were then averaged together pixel-by-pixel. We also computed the relative standard intensity deviation ($\sigma(x, y)/\mu(x, y)$) on a per pixel basis across the $N = 25$ vignettes (Fig. 2c). Compared to the vignette extracted with the BaSiC method (Fig. 2a), the illumination bias estimated by the CNN model is smoother (Fig. 2b). The BaSiC vignette was computed with only 20 tiles (see Fig. 3), with almost half of them not containing tissues. Thus the extracted vignette for this method is biased toward background tiles. It also tries to capture finer details such as vertical stripes which is not the case for the network derived vignette. In fact, due to our choice of synthetic vignettes (Gaussian and 5th order Zernike), our proposed method estimates vignettes containing mostly low spatial frequencies and is biased toward angular symmetry. We performed some inference tests with tiles affected by rectilinear vignettes (e.g., mirror reflections) and the trained method was not able to accurately estimate those inhomogeneities (data not shown). This indicates that additional synthetic vignette models would be necessary to capture other types of illumination biases. Regarding the vignette extraction repeatability, Fig. 2c shows that our method exhibits more variability in the corners and tile boundary ($\approx 22.2\%$ for $k = 5$) compared to the middle of the tile ($< 2.5\%$). This was expected, as the vignetting effect increases with distance from the image center. This variability could indicate that the chosen vignette synthesis models cannot accurately represent the real illumination biases in the OCT tile corners, or that the chosen network architecture is not well adapted for encoding and decoding near edges. Additionally, the maximum relative STD decreases with an increasing number of input tiles ($\approx 34.6\%$ for $k = 1$, $\approx 22.6\%$ for $k = 2$) and when using smoothness regularization ($\approx 27.0\%$ for $k = 5$ without regularization), which support the early fusion approach and spatial smoothness regularization benefit, although a thorough ablation study would be necessary to confirm these observations.

Finally, we tested the vignette correction capacity of our method (Fig. 3). We compensated the illumination bias by dividing each tile pixel-by-pixel with the estimated vignette obtained with the BaSiC method (Fig. 3, middle) and with our method (Fig. 3, right). Qualitatively, we can see that our method is able to reduce the vignetting effect in the assembled mosaic which results in less visible seams between neighboring tiles. Compared with the BaSiC method, our method performs better in tiles containing tissue and worse in background tiles. In particular, for this example only 20 tiles are available for the BaSiC method, and half of them contain only background, thus it is biased toward fixing the illumination in the background tiles to the detriment of illumination compensation in tissue slices. Despite this advantage of our CNN approach, we can still see some areas where the illumination compensation was too strong in tissues and not strong enough in the background. To address this, the training strategy could be modified to consider the similarity between neighboring tile overlapping areas, or by also predicting an additive darkfield in addition to a multiplicative flatfield.

4. CONCLUSION

We explored CNN-based self-supervised illumination inhomogeneity compensation for serial OCT data. It showed better vignetting compensation in tissue areas compared to the BaSiC method, resulting in better performance when few tiles are available to apply the algorithm. Future work will investigate in more details our CNN approach, considering for example neighboring tiles information during training.

5. COMPLIANCE WITH ETHICAL STANDARDS

The animal tissue data used in this study were generated by other previous and ongoing research projects in accordance with the animal ethics committee of the Montreal Heart Institute.

6. ACKNOWLEDGMENTS

This work was supported by a *FRQNT* grant (2021-PR-282231) and a *NSERC Discovery* grant (RGPIN-2020-06109). Special thanks to Patrick Delafontaine-Martel for some the data acquisition and optics work, and to Pr. Frederic Lesage for providing access to the imaging facilities.

7. REFERENCES

[1] Joël Lefebvre, Alexandre Castonguay, Philippe Pouliot, Maxime Descoteaux, and Frédéric Lesage, "Whole mouse brain imaging using optical coherence tomography: reconstruction, normalization, segmentation, and comparison with diffusion MRI," *Neurophotonics*, vol. 4, no. 4, pp. 041501–041501, 2017.

[2] Joël Lefebvre, Patrick Delafontaine-Martel, Philippe Pouliot, Hélène Girouard, Maxime Descoteaux, and Frédéric Lesage, "Fully automated dual-resolution serial optical coherence tomography aimed at diffusion MRI validation in whole mouse brains," *Neurophotonics*, vol. 5, no. 4, 2018.

[3] Tingying Peng, Kurt Thorn, Timm Schroeder, Lichao Wang, Fabian J. Theis, Carsten Marr, and Nassir Navab, "A BaSiC tool for background and shading correction of optical microscopy images," *Nature Communications*, vol. 8, no. 1, pp. 14836, 2017.

[4] Guha Balakrishnan, Amy Zhao, Mert R. Sabuncu, John Guttag, and Adrian V. Dalca, "VoxelMorph: A Learning Framework for Deformable Medical Image Registration," *IEEE Transactions on Medical Imaging*, vol. 38, no. 8, pp. 1788–1800, 2019.

[5] Vasudevan Lakshminarayanan and Andre Fleck, "Zernike polynomials: a guide," *Journal of Modern Optics*, vol. 58, no. 7, pp. 545–561, 2011.

[6] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *arXiv*, 2015.

[7] Thorsten Falk, Dominic Mai, Robert Bensch, Özgün Çiçek, Ahmed Abdulkadir, Yassine Marrakchi, Anton Böhm, Jan Deubner, Zoe Jäckel, Katharina Seiwald, Alexander Dovzhenko, Olaf Tietz, Cristina Dal Bosco, Sean Walsh, Deniz Saltukoglu, Tuan Leng Tay, Marco Prinz, Klaus Palme, Matias Simons, Ilka Diester, Thomas Brox, and Olaf Ronneberger, "U-Net: deep learning for cell counting, detection, and morphometry," *Nature Methods*, vol. 16, no. 1, pp. 67–70, 2019.

[8] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019.

[9] William Falcon and The PyTorch Lightning team, "PyTorch Lightning," 3 2019.

[10] Joël Lefebvre, "linum-uqam/PyBaSiC: v1.0.0," in *GitHub*. 11 2022.