scientific reports

Check for updates

OPEN

Rapid artefact removal and H&E-stained tissue segmentation

B. A. Schreiber[1,2]✉, J. Denholm[1,2,3], F. Jaeckle[1,3], M. J. Arends[4], K. M. Branson[5], C.-B. Schönlieb[2,3] & E. J. Soilleux[1,3]✉

We present an innovative method for rapidly segmenting haematoxylin and eosin (H&E)-stained tissue in whole-slide images (WSIs) that eliminates a wide range of undesirable artefacts such as pen marks and scanning artefacts. Our method involves taking a single-channel representation of a low-magnification RGB overview of the WSI in which the pixel values are bimodally distributed such that H&E-stained tissue is easily distinguished from both background and a wide variety of artefacts. We demonstrate our method on 30 WSIs prepared from a wide range of institutions and WSI digital scanners, each containing substantial artefacts, and compare it to segmentations provided by Otsu thresholding and Histolab tissue segmentation and pen filtering tools. We found that our method segmented the tissue and fully removed all artefacts in 29 out of 30 WSIs, whereas Otsu thresholding failed to remove any artefacts, and the Histolab pen filtering tools only partially removed the pen marks. The beauty of our approach lies in its simplicity: manipulating RGB colour space and using Otsu thresholding allows for the segmentation of H&E-stained tissue and the rapid removal of artefacts without the need for machine learning or parameter tuning.

Otsu thresholding[1] is often applied to the luminance of whole-slide images (WSI) of haematoxylin and eosin (H&E)-stained tissue for the purposes of segmentation[2–11] (see Fig. 1), including in popular histopathological analysis tools Histolab[12] and PyHist[13]. However, Otsu thresholding only successfully segments the tissue from the background when the tissue and background pixels are well-separated in a greyscale representation of the WSI. While this is often the case in artefact-free WSIs, WSIs often contain artefacts such as pen marks and dark scanning artefacts, which cause this assumption to fail, thus resulting in artefacts wrongly identified as tissue, tissue rejected as background, or both (see Fig. 1). While there are a large and diverse range of artefacts that can occur on a WSI, in the context of this paper artefacts will refer only to pen marks (see Figs. 3a–d, g), bounding boxes added by the scanners (see Fig. 3e–g) scanning artefacts such as dark blobs or text (see Fig. 3e, f).

The exclusion of pen marks in particular is a crucial first step for any machine learning-based automated WSI analysis pipeline; pathologists often use pen marks to highlight areas of interest which, if observed by a machine learning algorithm, could result in deleterious bias, spurious classifications or even data leakage, thus reducing confidence in the performance metrics and generalizability of the algorithm[14].

In this paper, we propose a new tissue segmenting algorithm for H&E-stained tissue which can segment tissue in the presence of artefacts. We tested our method on WSIs of H&E-stained duodenal biopsies prepared at multiple different institutions, scanned using multiple different scanners, and containing a large range of artefacts of different types, shapes and colours.

Method

Our method improves on Otsu thresholding by selecting a representation of the WSI data that better separates H&E-stained tissue from background and artefacts than luminance. Given a three channel image $I = [I_R, I_B, I_G]$, the channels are normalized so that the channels of each pixel are represented by floats ranging from 0 to 1. Then, the following representation of the data is calculated:

[1]Department of Pathology, University of Cambridge, Tennis Court Road, Cambridge CB2 1QP, Cambridgeshire, UK. [2]Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Wilberforce Road, Cambridge CB3 0WA, Cambridgeshire, UK. [3]Lyzeum Ltd., Cambridge CB1 2LA, Cambridgeshire, UK. [4]Edinburgh Pathology, Institute of Genetics and Cancer, University of Edinburgh, Crewe Road, Edinburgh EH4 2XR, UK. [5]Artificial Intelligence and Machine Learning, GSK plc., Great West Road, Brentford TW8 9GS, Middlesex, UK. ✉email: bas43@cam.ac.uk; ejs17@cam.ac.uk

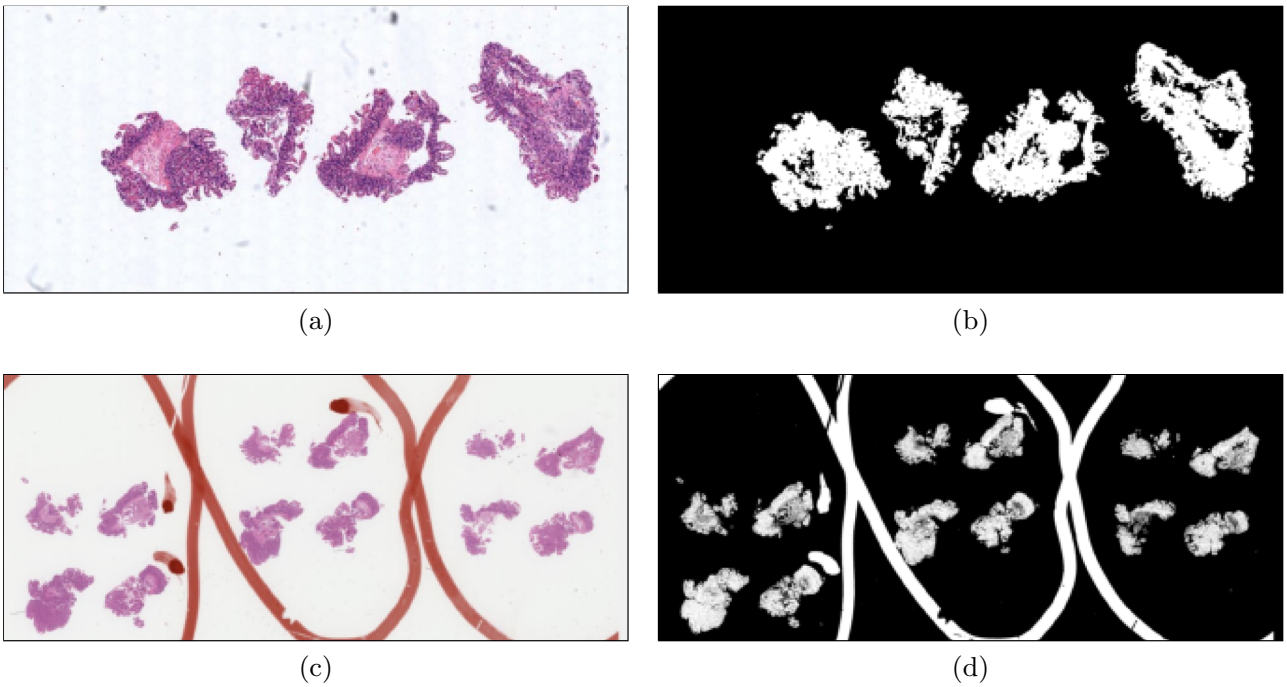

Figure 1. (**a**) A WSI of a H&E-stained biopsy containing only minor, non-interfering artefacts. (**b**) The tissue segmentation provided by applying Otsu thresholding to the luminance of the WSI. The clear distinction between the intensities of the tissue and the rest of the WSI caused the Otsu threshold to lie between the maximum intensity of the tissue and the minimum intensity of background, allowing for a successful tissue segmentation. (**c**) A WSI of an H&E-stained biopsy containing orange pen marks splitting the tissue of different levels and used to identify three features of interest. (**d**) The tissue segmentation provided by applying Otsu thresholding to the luminance of the WSI. The pen marks interfered with the Otsu threshold calculation, resulting in a tissue segmentation that contains tissue and pen marks.

$$T = \mathrm{ReLU}(I_R - I_G) \odot \mathrm{ReLU}(I_B - I_G) \tag{1}$$

where $\mathrm{ReLU}(x) = \max(x, 0)$ is the rectifier linear unit and $\odot$ is the Hadamard product, both of which act element-wise. Otsu thresholding is then used to separate tissue and non-tissue pixels[1]. Note that this calculation requires no parameter training or tuning. A Python implementation of this previously unreported algorithm can be found here https://gitlab.developers.cam.ac.uk/bas43/h_and_e_otsu_thresholding in accordance with the Guidelines for Authors Submitting Code & Software presented in Nature Research https://www.nature.com/nature-portfolio/editorial-policies/reporting-standard#reporting-requirements. All relevant guidelines were followed in the development and testing of this algorithm.

Algorithm 1. Our method for segmenting H&E stained tissue

RGB Image: $[I_R, I_G, I_B]$
Normalize:
$I_R, I_G, I_B \leftarrow I_R/255, I_G/255, I_B/255$
R - G Representation:
$I_{R-G} \leftarrow \mathrm{ReLU}[I_R - I_G]$
B - G Representation:
$I_{B-G} \leftarrow \mathrm{ReLU}[I_B - I_G]$
Tissue Representation: $T \leftarrow I_{R-G} \odot I_{B-G}$
Otsu threshold: $\gamma \leftarrow \mathrm{Otsu}[T]$
if $T[p] > \gamma$ **then**
  Pixel p is segmented as tissue
else
  Pixel p is rejected
end

Figure 2. Left: Two 24-bit colour cubes, one with the white corner at the origin and one with the black corner at the origin. Middle: The values of the pixles in the representation specified in Eq. 1 and Alg. 1. Right: The pixels that have values greater than 0 in this representation.

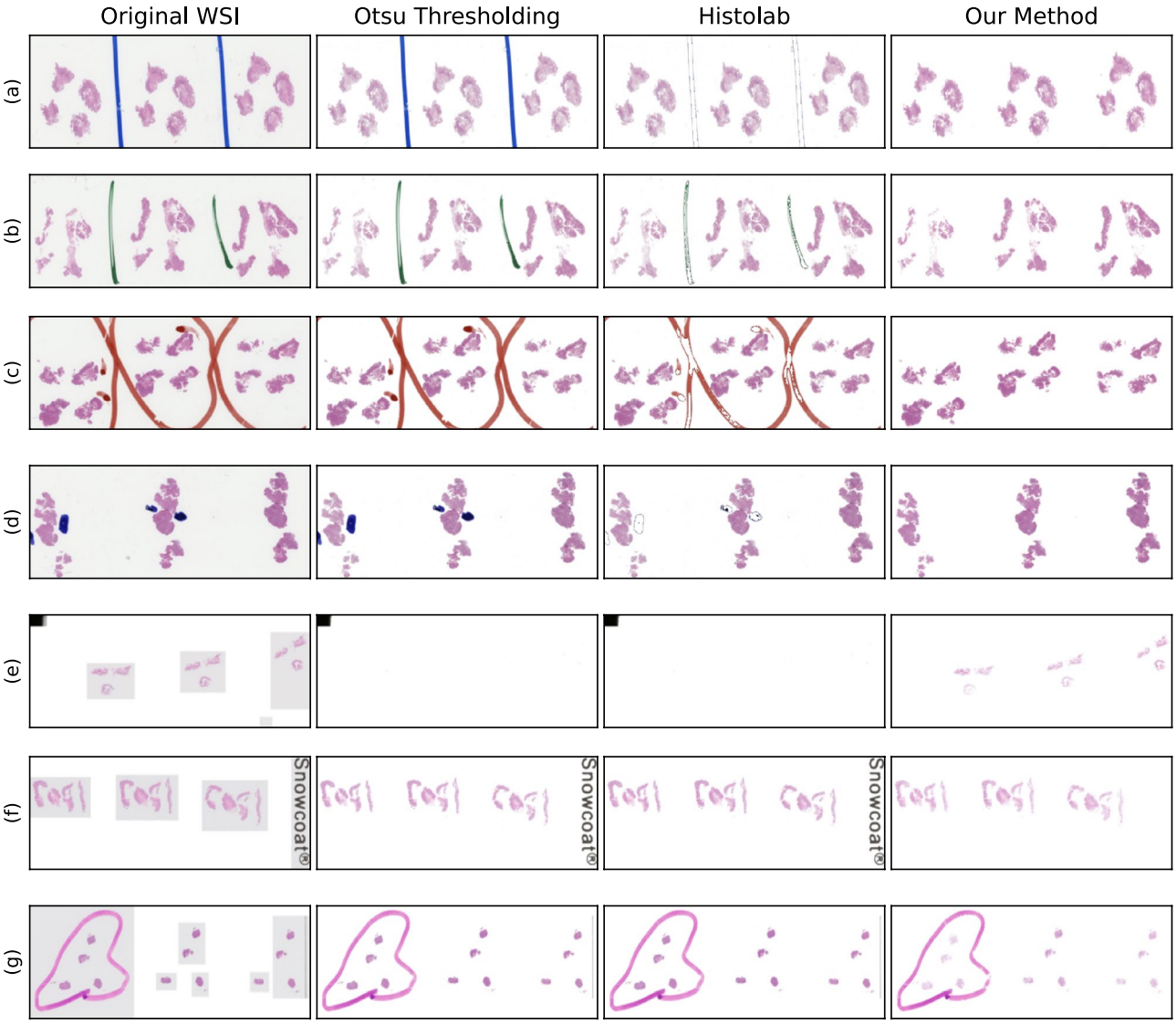

Figure 3. Seven WSIs of H&E-stained biopsies containing artefacts of a wide range of types and colours. The aim was to segment the tissue without including background and artefacts. First Column: The original WSI. Second Column: The tissue segmentation provided by applying Otsu thresholding to the luminance of the WSIs placed on a white background. Otsu thresholding failed to reject a single artefact and failed to segment the tissue in (**e**). Third column: the tissue segmentation from Histolab tissue thresholding and pen filters. While there was partial pen mark removal in (**a–d**), the pen marks were not fully removed in any image, no tissue was segmented in (**e**), no pen marks were removed in (**g**). Forth column: the tissue segmentation from our method placed on a white background. Our method successfully segmented all tissue and rejected all background and artefacts except the pen marks in (**g**). Our method failed to reject the pen marks in (**g**) because the pen is the same colour as the eosin.

The assumption made by Otsu thresholding is that tissue and non-tissue pixels can be separated through luminance, which is not the case when artefacts are present. However, our method, which is described in Eq. 1 and Algorithm 1, is based on the assumption that the tissue pixels can be identified by being *both* more blue than green and more red than green as compared to non-tissue pixels. The advantage of our method is that all shades of grey have approximately the same value in the red channel as the green channel, so their difference is 0, while pixels of H&E-stained tissue have higher values in both the blue and red channels than the green. Setting all negative values in both representations to zero ensures that artefacts with high green channels compared to blue or red channels do not adversely influence the threshold calculation, and are thus considered background. Thus, this representation results in a bimodal distribution that separates pixels that are the most "purple-pink" from others, so pen marks (which are often black, blue, green or red) are also excluded, independent of the pixel's light intensity. Pixels on an RGB colour cube that have a non-zero value in this representation are shown in Fig. 2 and comparisons between Otsu thresholding and our method on an RGB colour cube can be seen in the Supplementary Material.

We compared the performance of our method against Otsu thresholding and Histolab's pen filtering tools by applying these methods to a dataset of WSIs and assessing the resulting tissue segmentations qualitatively.

Data

To compare the performances of the Otsu thresholding, Histolab and our method, we applied both methods to a selection of 60 WSIs of H&E stained duodenal biopsies. Of the 60 WSIs selected:

- 15 contained pen marks
- 15 contained scanning artefacts
- 30 contained no significant artefacts

The WSIs were hand-picked so that they contained a wide range of artefacts of different types, shapes and colours. The WSIs were scanned with a wide range of digital scanners (Ventana, Aperio, Hamamatsu and Philips), and the 30 WSIs with no significant artefacts were selected at random and matched for scanner type of the 30 WSIs with pen marks or artefact.

Ethical statement

All fully anonymized slide scans (and patient data) were obtained with full ethical approval from the Oxfordshire Research Ethics Committee A (IRAS: 162057; PI: Prof. E. Soilleux), and the method was performed in accordance with their guidelines and regulations. Informed consent was obtained from all subjects and/or their legal guardian(s).

Results

Otsu thresholding, Histolab and our method were used to segment the tissue from the 60 WSIs described above. Examples of the WSIs selected and the tissue segmentation of these methods can be viewed in Fig. 3. Examples of the tissue segmentation masks provided by Otsu thresholding and our method, and the Sørensen–Dice coefficient's between the segmentations and a manually segmented tissue mask are displayed in the supplementary material. The tissue segmentations were assessed by a single observer, and considered "successful" if all the following were true:

- All tissue was segmented
- All background was rejected from the segmentation
- All bounding boxes were rejected from the segmentation
- All artefacts were rejected from the segmentation

Otsu thresholding rejected pen and scanning artefacts from the tissue segmentation in 0/30 WSIs containing artefacts. In 2/30 WSIs containing artefacts, the influence the artefacts had on the threshold was so great that the tissue was not segmented as tissue (see Fig. 3e).

The Histolab pen filtering tool only partially removed pen marks in Fig. 3a–d, and removed no pen marks in Fig. 3g. Other artefacts such as scanning artefacts were not effected by the Histolab tools.

Our method segmented the tissue in all 60/60 WSIs and rejected all artefacts in 29/30 WSIs containing pen and scanning artefacts. The only WSI where pen marks were included in the tissue segmentation can be seen in Fig. 3g. Here our method failed to reject the pen marks because tissue and non-tissue pixels could not be separated through their "pinkness", when the pen marks were also pink.

Discussion

While Otsu thresholding segmented the tissue in all artefact-free WSIs and most WSIs with artefacts, it identified all artefacts as tissue as well. In 2 out of 30 WSIs with artefacts, the presence of artefacts caused the threshold to ignore tissue and include background in the tissue segmentation as seen in Fig. 3e.

The Histolab pen filtering tools were applied to the tissue segmentations in order to remove the remaining pen marks. The filtering tools performed best on blue pen marks, as seen in Fig. 3a, d. However, the tools did not remove the edges of pen marks of all colours, and failed to detect the majority of all green and orange pen marks, as seen in Fig. 3b, c respectively. The pink pen marks presented in Fig. 3g remained untouched. Additionally, the Histolab tools were not designed to remove scanning artefacts and bounding boxes so these features remained.

Our method, on the other hand, segmented the tissue in all WSIs and rejected artefacts in all WSIs containing artefacts but one. The only exception can be seen in Fig. 3g, which contained pink pen marks that caused all methods to fail.

The thresholding algorithm presented here is a rapid, reliable and easily implementable tissue segmentation and artefact removal tool for WSIs of H&E-stained tissue. In machine learning tasks especially, this tool can be used as a preprocessing step that ensures artefacts do not cause the machine learning algorithm to train on irrelevant patches or patches that contain data leaking pen marks.

It should be noted that this method is built to segment H&E-stained tissue only, and will not perform as intended on tissue which has been stained with stains that do not appear pink/purple. However, this method should be relatively simple to generalize to other stains by using representations of the WSI data that uniquely differentiate the stained tissue from background and artefacts, and will be studied in future research.

Data availability

The datasets of WSIs analysed during this current study have not publicly available due to the large size of the WSIs and legal considerations. However, low-level representations of the WSIs used in the study have been made available at https://gitlab.developers.cam.ac.uk/bas43/h_and_e_otsu_thresholding.

Received: 28 October 2023; Accepted: 16 December 2023
Published online: 03 January 2024

References

1. Otsu, N. A threshold selection method from gray-level histograms. *IEEE Trans. Syst. Man Cybern.* **9**, 62–66. https://doi.org/10.1109/TSMC.1979.4310076 (1979).
2. Wang, X. *et al.* Weakly supervised deep learning for whole slide lung cancer image analysis. *IEEE Trans. Cybern.* **50**, 3950–3962. https://doi.org/10.1109/TCYB.2019.2935141 (2020).
3. Denholm, J. *et al.* Multiple-instance-learning-based detection of coeliac disease in histological whole-slide images. *J. Pathol. Inform.* **13**, 100151. https://doi.org/10.1016/j.jpi.2022.100151 (2022).
4. Campanella, G. *et al.* Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat. Med.* **25**, 1301–1309. https://doi.org/10.1038/s41591-019-0508-1 (2019).
5. Khened, M., Kori, A., Rajkumar, H., Krishnamurthi, G. & Srinivasan, B. A generalized deep learning framework for whole-slide image segmentation and analysis. *Sci. Rep. (Nature Publishing Group)* **11**, 11579. https://doi.org/10.1038/s41598-021-90444-8 (2021).
6. Anghel, A. *et al.* A high-performance system for robust stain normalization of whole-slide images in histopathology. *Front. Med.* **6**, 193. https://doi.org/10.3389/fmed.2019.00193 (2019).
7. Haghighat, M. *et al.* Automated quality assessment of large digitised histology cohorts by artificial intelligence. *Sci. Rep. (Nature Publishing Group.)* **12**, 5002. https://doi.org/10.1038/s41598-022-08351-5 (2022).
8. Smith, B., Hermsen, M., Lesser, E., Ravichandar, D. & Kremers, W. Developing image analysis pipelines of whole-slide images: Pre- and post-processing. *J. Clin. Transl. Sci. (Cambridge University Press)* **5**, e38. https://doi.org/10.1017/cts.2020.531 (2021).
9. ...Veta, M. *et al.* Predicting breast tumor proliferation from whole-slide images: The TUPAC16 challenge. *Med. Image Anal.* **54**, 111–121. https://doi.org/10.1016/j.media.2019.02.012 (2019).
10. Schmauch, B. *et al.* Transcriptomic learning for digital pathology. *BioRxiv*https://doi.org/10.1101/760173 *(2019).*
11. Zhang, H. *et al.* DTFD-MIL: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 2022. 18780–18790. ISSN: 2575-7075 https://doi.org/10.1109/CVPR52688.2022.01824 (2022).
12. Marcolini, A. *et al.* histolab: A python library for reproducible digital pathology preprocessing with automated testing. *SoftwareX* **20**, 101237 https://doi.org/10.1016/j.softx.2022.101237.https://www.sciencedirect.com/science/article/pii/S2352711022001558 (2022).
13. Muñoz-Aguirre, M., Ntasis, V. F., Rojas, S. & Guigó, R. PyHIST: A histological image segmentation tool. *PLOS Comput. Biol.* **16**, e1008349. https://doi.org/10.1371/journal.pcbi.1008349 (2020).
14. Kaufman, S., Rosset, S., Perlich, C. & Stitelman, O. Leakage in data mining: Formulation, detection, and avoidance. *ACM Trans. Knowl. Discov. Data* **6**, 151–1521. https://doi.org/10.1145/2382577.2382579 (2023).

Author contributions

B.A.S. devised the thresholding algorithm and wrote the manuscript. J.D. and F.J. independently tested and compared Otsu thresholding, Histolab, and the thresholding algorithm presented here. Histological expertise was provided by M.J.A. and E.J.S. The project was initialized by E.J.S. and supervised by K.M.B., C.-B.S. and E.J.S. All authors were given the opportunity to review and comment on the manuscript.

Funding

This work was supported by the Pathological Society [PKAG/924] and GlaxoSmithKline [LEAG/781].

Competing interests

The following authors are shareholders in Lyzeum ltd.: Elizabeth Jane Soilleux and Carola-Bibiane Schönlieb.

Additional information

Correspondence and requests for materials should be addressed to B.A.S. or E.J.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.