DOI:10.34918/87336

Patryk DĄBKOWSKI[1], Joanna POLAŃSKA[1,*]

Chapter 10. INTELLIGENT SELF-ADAPTIVE H&E-STAINED TISSUE SCAN QUANTISATION

10.1. Introduction

Tumour-infiltrating lymphocytes (TILs) can effectively indicate the immune risk [1] as they provide information about the organism's response to the cancer cells, which play a significant role in therapy or treatment selection. The score representing TIL's concentration relates to patient survival [2]. To obtain necessary information about tissue structure and cell spatial distribution in the collected specimen, samples are stained with hematoxylin and eosin (HE) to selectively colour cancer cells, blood cells, and fibres in a specific way that enables their identification. Analysing such preparations by doctors or specialists is highly time-consuming and challenging, and obtaining precise results is difficult. Unsupervised machine learning methods must be adopted to extract the crucial diagnostic information from H&E-stained tissue scans efficiently. We propose a three-stage modular pipeline for automated tissue detection on HE-stained scans, staining quantisation and image segmentation and demonstrate its properties using the exemplary breast cancer tissue scans from the TCGA database.

[1] Department of Data Science and Engineering, Faculty of Automatic Control, Silesian University of Technology, Gliwice, Poland.
[*] Corresponding author: joanna.polanska@polsl.pl.

10.2. Materials and Methods

10.2.1. Data

Images are scans of breasts tissues cancer, as the example in Fig.1, saved in TIFF format and of a substantial spatial resolution, containing millions of pixels and 400 + GB volume. They are considered Big Data and need to develop suited and intelligent solutions that avoid increasing computational costs. Data come from TCGA (The Cancer Genome Atlas) database and contain over thousand images of breast cancer H&E stained scans.

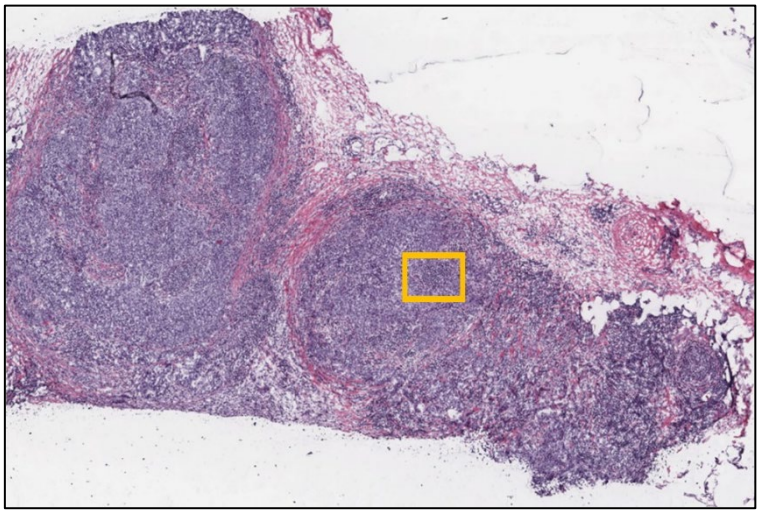

Fig. 1. Exemplary tissue scan with the zoom region marked as yellow box
Rys. 1. Przykładowy skan wycinka tkankowego z obszarem zainteresowania oznaczonym ramką w kolorze żółtym

10.2.2. Tissue mask detection

Among the various image segmentation methods available, the aim was to choose an approach that would effectively separate the tissue region from a relatively homogeneous background. The most popular watershed segmentation algorithm was discarded through experimental trials as it led to over-segmentation for complex images. The reasons for rejecting the watershed segmentation algorithm were also low computational efficiency, the need for calculations performed on each pixel, and the large number of hyperparameters that impede code automation and task universality.

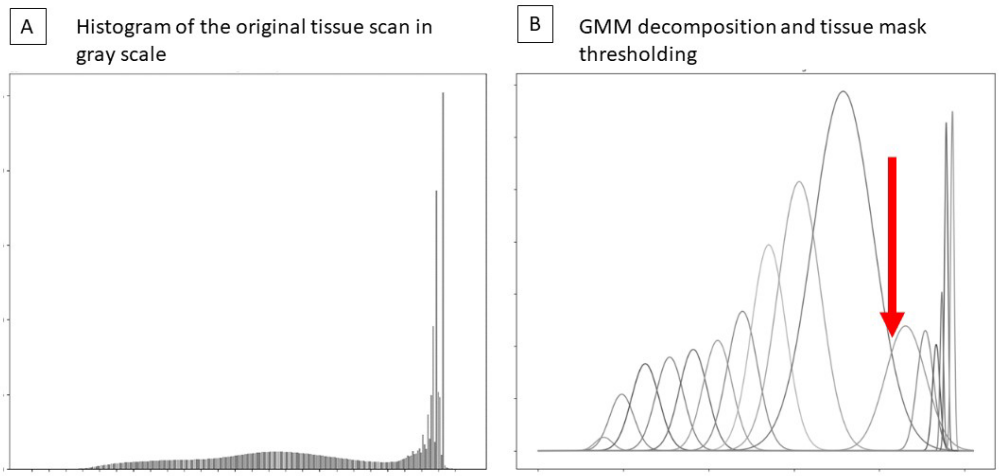

Fig. 2. Grayscale histogram (A) and its GMM decomposition (B) for the exemplary tissue scan
Rys. 2. Histogram w kolorach szarości (A) i jego dekompozycja do mieszaniny rozkładów normalnych (B) dla przykładowego skanu wycinka tkankowego

Segmentation based a threshold usage obtained from the analysis of the grayscale image's histogram's envelope allowed us to reduce computationally expensive iterative operations on each pixel. The selection of the threshold point was performed through histogram analysis using the Gaussian Mixture Model (GMM) decomposition (Fig. 2) [3]. The GMM-type algorithm can be applied to segment the image histogram and determine the threshold for thresholding mask segmentation. The signal frequency distribution is modelled as a Gaussian mixture following the formula:

$$f(x_n) = \sum_{k=1}^{K} \alpha_k f_k(x_n, \mu_k, \sigma_k) \quad (1)$$

where K stands for the number of Gaussian components in the model, the coefficients α_k that sum up to 1 are the weights of the individual component with means μ_k and standard deviations σ_k. Dempster's expectation-maximisation algorithm is used to estimate the model parameters. Let us define the term *envelope,* which refers to a curve resulting from the sum of Gaussian components in the obtained model (Fig. 3). Additionally, we define a set of potential thresholds similarly as it is done for GMM-based classification problems. The crosspoints between consecutive GMM components are considered candidate greyscale image segmentation thresholds, and the first one after the dominating Gaussian component is chosen (as value 211 in the exemplary envelope in Fig. 2B – red arrow – and Fig. 3).

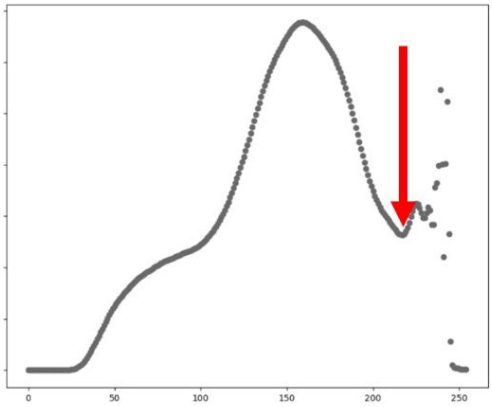

Fig. 3. GMM envelope for the exemplary scan
Rys. 3. Obwiednia modelu mieszanin rozkładów normalnych dla przykładowego skanu wycinka tkankowego

The number of Gaussian components is determined by the Bayesian Information Criterion (BIC). The criterion is conceptually defined as the model probability expressed through Bayes' theorem for model M and dataset y, where P(M|y) represents the marginal likelihood function:

$$P(M_i|y_1,\cdots,y_n) = \frac{P(y_1,\cdots,y_n|M_i)}{P(y_1,\cdots,y_n)} P(M_i) \quad (2)$$

Pixels with colour values above the chosen threshold are considered background, while the remaining ones constitute the tissue mask. To guarantee mask spatial homogeneity, the standard morphological operations as dilation and erosion are applied.

10.2.3. Staining quantisation

Quantisation is the process of reducing the number of signal or data values. This reduction can be helpful in data compression, denoising, standardisation, and extracting crucial and relevant information. Colour images can contain millions of unique colours, making it difficult for algorithms to effectively select features and make predictions or classifications with the desired accuracy.

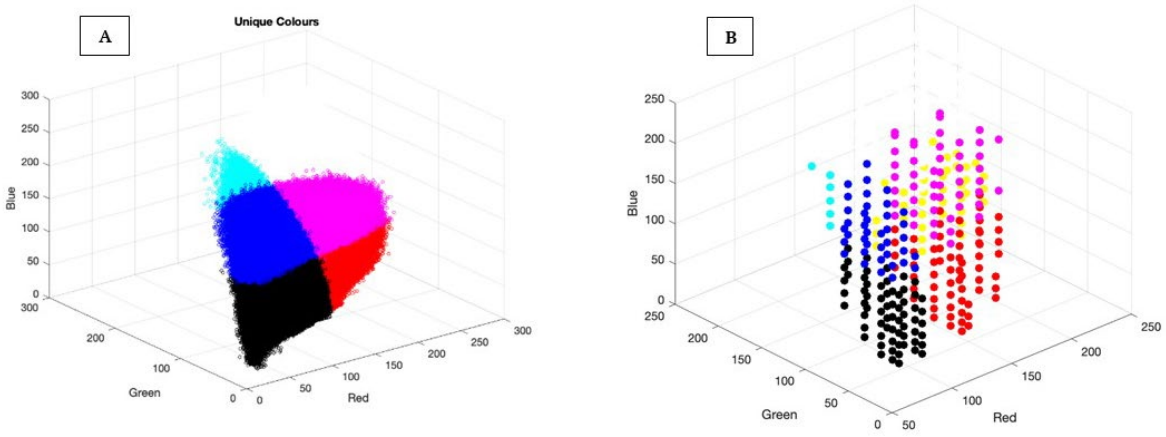

Fig. 4. (A) Unique colours of the original image (B) Unique colours of the quantised image in expanded RGB space (eRGB)
Rys. 4. (A) Unikatowe kolory oryginalnego skanu (B) Unikatowe kolory dla obrazu po wstępnej kwantyzacji barw w rozszerzonej przestrzeni RGB (eRGB)

Therefore, minimising the loss of information and the number of unique colours through colour reduction can be essential for identifying tumour-infiltrating lymphocytes. In our method, quantisation was achieved by utilising Gaussian Mixture Models for each RGB channel's histogram separately. Gaussian distributions are calculated based on the analysis of these histograms and with the Bayesian Information Criterion (BIC) to determine the number of components. The new colour values for each pixel are defined by their relation to RGB-specific Gaussian components. We propose to use a colour projection method that translates H&E stained colour images into new expanded colour space. New colour space is dynamic and defined by GMM components identified per each RGB channel. The final colour vector is of Mx1 dimension, where M denotes the total of GMM components from all RGB channels. The new colour feature value is 1 if GMM-based classifier assigns the pixel to the respective RGB channel specific GMM component, otherwise is equal to 0. Instead of 3D colour space we get the expanded M-dimensional space with 0/1 values only. That dimensionality expanding and colour coding let us distinguish colours and features that could not be distinguished in RGB space.

10.2.4. Image segmentation

The final step involves image segmentation performed in the expanded colour domain via pixel grouping according to their colour in the expanded colour domain. Clustering is done by the K-means++ algorithm, first introduced in [4]. The method is based on iterative fitting data into a static and predefined number of groups represented

by the mean of each group, aiming to minimise the error of data fitting. The K-Means++ variant ensures more evenly distributed and distant initial centroids. As a result, the algorithm is less prone to convergence to local minima, which helps achieve better cluster quality. The properly chosen initial in K-means++ results in faster convergence and a reduction of the number of iterations required to achieve convergence. Dimensionality expanding and colour quantisation let us effectively spread the colour values into distinguishable groups. The number of K-means clusters is defined by the Calinski and Harabasz score. Finally, the obtained colour clusters are represented by a static colour palette, which aims to standardise the obtained results, enabling more effective analysis of the scans by doctors. The kNN algorithm is applied to identify palette colours for each cluster. The reference palette is represented in Fig. 5.

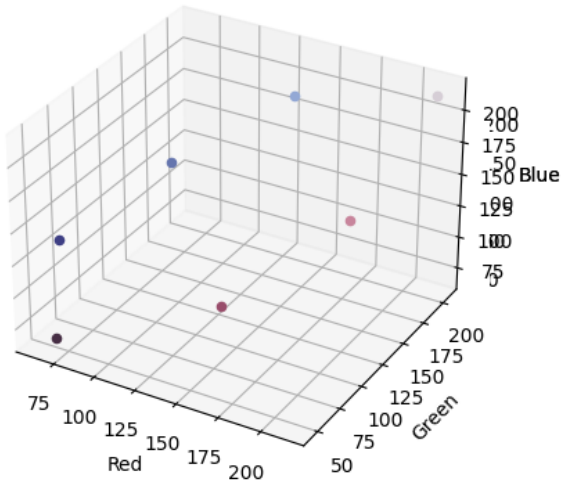

Fig. 5. The final reference colour palette in RGB space
Rys. 5. Ostateczna paleta pseudobarw klastrów przedstawiona w przestrzeni RGB

The resulting pipeline is presented in Fig. 6. It effectively utilises machine learning techniques for the segmentation, quantisation, and clustering of histopathological images stained with the H&E method.

Fig. 6. The image processing pipeline
Rys. 6. Proponowany potok przetwarzania obrazów

10.3. Results

The developed pipeline has been applied to a diverse set of images from the TCGA dataset, and the proposed method proves to be successful in extracting principal features important for TIL scoring. The developed tissue region segmentation method provides an accurate and precise binary mask, as presented in Fig. 7A.

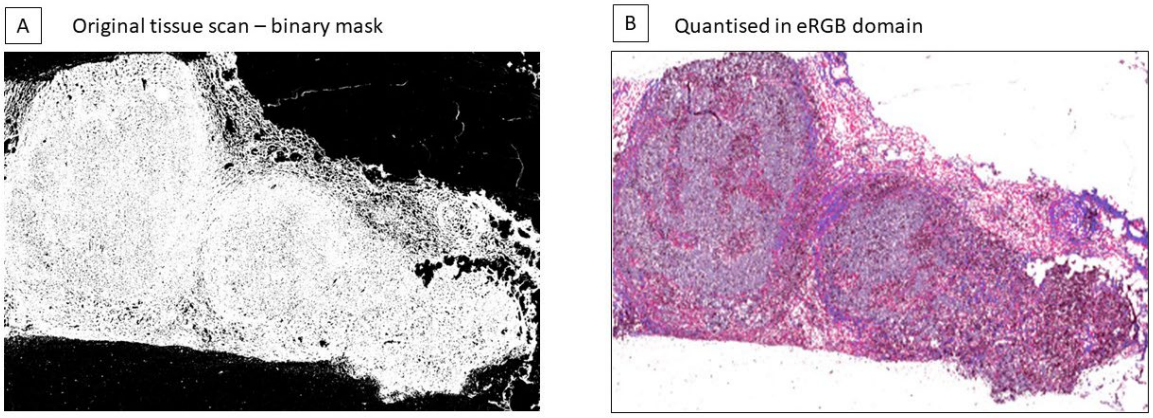

Fig. 7. Segmentation mask (panel A) and quantisation result (panel B) for the exemplary tissue scan from Fig. 1
Rys. 7. Binarna maska wycinka tkankowego (panel A) oraz wyniki kwantyzacji barw (panel B) dla przykładowego skanu wycinka tkankowego z Rys. 1

The number of unique colours in the exemplary image from Fig. 1 was 1,152,408. The applied colour quantisation method resulted in the reduction to 730 unique colours, and K-means clustering ended up with seven tissue segments (as shown in Fig. 7B and Fig. 8). The details of the cell neighbourhood are presented in Fig. 9, where one can easily distinguish the cell nucleus and cytoplasm. The preservation of this information in the process of image segmentation is essential from the point of view of the TILS index value assessment.

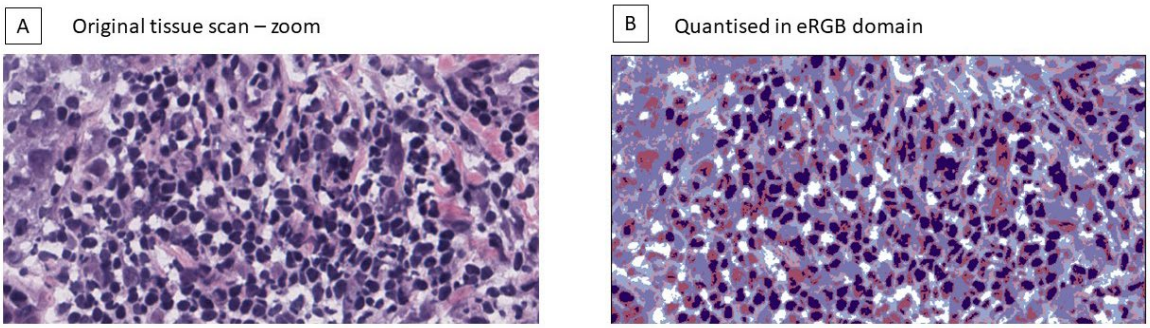

Fig. 8. Image segmentation for the zoom of the exemplary tissue scan from Fig. 1
Rys. 8. Wynik segmentacji obrazu dla wybranego regionu zainteresowania z przykładowego skanu wycinka tkankowego przedstawinego na Rys. 1

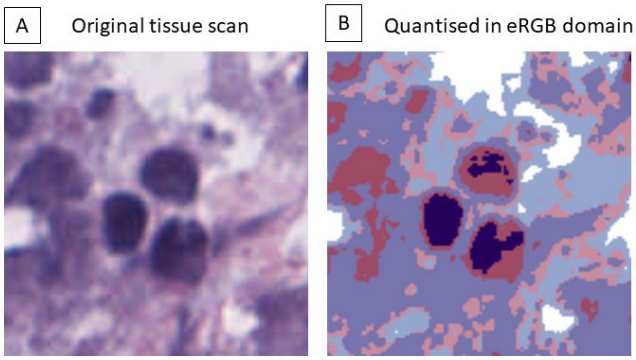

Fig. 9. Details of the original (panel A) and segmented (panel B) tissue region
Rys. 9. Szczegóły wybranego fragmentu tkanki: oryginalny obraz (panel A) i po segmentacji (panel B)

To investigate the importance of the RGB colour domain expansion, a similar image segmentation pipeline using the K-means clustering algorithm and the number of clusters set based on the analysis performed in expanded RGB (eRGB) was applied to the image represented in RGB colour space. Figure 10 presents the results of tissue region segmentation without (panel B) and with colour domain expansion (panel C). One can notice that quantisation without domain expansion provides a lower quality of the compressed images, where cells are inseparable. That phenomenon is not observed in the second case (panel C).

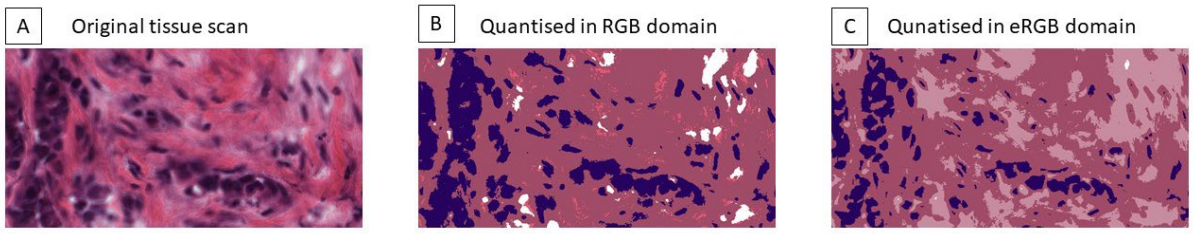

Fig. 10. Zoom of the original (panel A) and segmented tissue regions done without (panel B) and with (panel C) expansion of the colour space
Rys. 10. Szczegóły wybranego fragmentu oryginalnego skanu (panel A) i po segmentacji wykonanej bez (panel B) oraz z (panel C) proponowanym rozszerzeniem przestrzeni barw

10.4. Conclusions

We have shown that it is possible to perform fast and effective H&E-stained tissue scan segmentation by using the expanded RGB colour domain combined with the classical K-means approach. As the staining quantisation and standardisation are crucial either for cell identification as well as for further Deep Learning TIL score estimation, the above-developed image preprocessing pipeline should be a required step

in data analysis schema. For a more comprehensive exploration, we defer further discussion on the topic of effectively determining the number of clusters and estimation of TILs score to future research endeavours.

Bibliography

1. H.J. Wanebo, P.P. Rosen, T. Thaler, J.A. Urban, H.F. Oettgen. Immunobiology of operable breast cancer: an assessment of biologic risk by immunoparameters. Ann Surg. 1976 Sep;184(3):258–67. DOI: 10.1097/00000658-197609000-00003. PMID: 962394; PMCID: PMC1344377.
2. A. Suwalska, L. Zientek, J. Polanska, M. Marczyk. Quantifying Spatial Heterogeneity of Tumor-Infiltrating Lymphocytes to Predict Survival of Individual Cancer Patients. Journal of Personalized Medicine, 12(7):1113, 2022.
3. A. Polanski, M. Marczyk, M. Pietrowska, P. Widlak, J. Polanska. Signal Partitioning Algorithm for Highly Efficient Gaussian Mixture Modeling in Mass Spectrometry. PLOS ONE, 10(7), 2015.
4. R. Ostrovsky, Y. Rabani, L.J. Schulman, C. Swamy, “The Effectiveness of Lloyd-Type Methods for the k-Means Problem”, 2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06), Berkeley, CA, USA, 2006, pp. 165–176, DOI: 10.1109/FOCS.2006.75.

INTELLIGENT SELF-ADAPTIVE H&E-STAINED TISSUE SCAN QUANTISATION

Abstract

H&E-stained tissue scans are widely used to estimate the tumour-infiltrating lymphocyte score, serving as an essential prediction factor in cancer treatment. The proportion and location of the lymphocytes can determine the organism's response. However, this analysis is a demanding and time-consuming task. The support of machine learning approaches is crucial for achieving complete and precise estimates. Existing computer-aided methods allow image segmentation and lymphocyte identification by supervised and unsupervised learning methods. Although still, the obtained results are unsatisfactory and sensitive to different stain concentrations.We present a three-stage, modular self-adaptive image quantisation pipeline which uses

statistical modelling and clustering techniques. DL-based TILS regression models can then efficiently analyse the quantified images. At first, RGB colour channels are analysed independently. The channel-specific colour quantisation is based on the Gaussian Mixture Model of its intensity histogram. Then, the k-means pixel grouping is done, with the variance ratio criterion applied to set the number of clusters. The proposed quantisation reduces the number of unique colours from millions to 330,000 and afterwards to ~700 in the exemplary image without any apparent loss of quality or detail. The K-means grouping detected seven main tissue subtypes, which concord with the biological structure of a given tissue. The apriori set of pseudocolours is assigned to each cluster.

Keywords: H&E tissue stained scans, image processing, segmentation, colour quantisation, clustering, unsupervised machine learning