Automated adenocarcinoma lung cancer tissue images segmentation based on clustering

Segmentación automatizada de imágenes de tejido de cáncer de pulmón de adenocarcinoma basado en agrupamiento

Bryan Cervantes-Ramirez[1], Francisco Siles[2]

Cervantes-Ramirez, B.; Siles, F. Automated adenocarcinoma lung cancer tissue images segmentation based on clustering. *Tecnología en marcha*. Vol. 35, special issue. November, 2022. International Work Conference on Bioinspired Intelligence. Pág. 16-23.

https://doi.org/10.18845/tm.v35i8.6442

[1] Pattern Recognition and Intelligent Systems Laboratory (PRIS-Lab) and Cancer and Surgery Research Center (CICICA). Universidad de Costa Rica. Costa Rica. E-mail: bryan.cervantesramirez@ucr.ac.cr

[2] Pattern Recognition and Intelligent Systems Laboratory (PRIS-Lab) and Cancer and Surgery Research Center (CICICA). Universidad de Costa Rica. Costa Rica. E-mail: francisco.siles@ucr.ac.cr
https://orcid.org/ 0000-0002-6704-0600

Keywords

Digital pathology; pattern recognition; lung cancer.

Abstract

Cancer is one of the main dead causes worldwide. It is re- sponsible for an approximate of 1 out of 6 deaths globally and lung cancer is along breast cancer, the most common types of cancer in the population, which confirms the importance of studies associated with it. This work presents an approach toward lung cancer histological tissue images segmentation based on colour. The proposed method for the segmentation is K-means clustering, providing promising results that may become as an assistance for pathologists, as it can help them reduce the time consumed reviewing the slides and giving a more objective perspective in order to provide a diagnose and specific treatment.

Palabras clave

Patología digital; reconocimiento de patrones; cáncer de pulmón.

Resumen

El cáncer es una de las principales causas de muerte en todo el mundo. Es responsable de aproximadamente 1 de cada 6 muertes a nivel mundial y el cáncer de pulmón, junto al cáncer de mama, es el tipo de cáncer más común en la población, lo que confirma la importancia de los estudios asociados a él. Este trabajo presenta un enfoque hacia la segmentación de imágenes de tejido histológico de cáncer de pulmón en función del color. El método propuesto para la segmentación es el agrupamiento de K-medias, brindando resultados prometedores que pueden convertirse en una ayuda para los patólogos, ya que puede ayudarlos a reducir el tiempo consumido revisando las diapositivas y dando una perspectiva más objetiva para brindar un diagnóstico y tratamiento específico.

Introduction

According to World Health Organization (WHO) data, cancer is the second cause of death worldwide and is responsible for an estimated 9.6 million deaths in 2018 [1] . Globally, about 1 in 6 deaths is due to cancer. In addition, for the particular case of lung cancer, it is estimated that it corresponds to a total of 2.09 million deaths of the aforementioned total, which makes it, together with breast cancer, the most common types of cancer in the population [1]. There are different types of tumors associated with lung cancer, such as non-small cell lung cancer (NSCLC) which is the case of the adenocarcinomas, small cell lung cancer (SCLC), carcinoid tumors, among others [2].An accurate diagnosis is essential for lung cancer, since the spread of the disease and the response to treatment varies greatly between patients.

In general, when there is suspicion of the presence of some kind of cancer, the diagnosis is confirmed with a biopsy. A biopsy is a medical procedure that involves taking a small sample of tissue or cells to be examined in detail through a microscope [3]. Pathologists, based on the biopsy study, either diagnose cancer or, if it is already confirmed, evaluate its severity or current stage. They examine biopsy tissues through naked eye, looking for any visible abnormalities, and selecting certain regions of interest (ROIs) for further detailed analysis under the microscope. These small sections of tissue are stained with different types of chemicals that allow different structures of the cells to be observed, in case they need to be evaluated in more detail. Histology is the study of the microscopic structure of tissues [4], and one of the most popular staining techniques used in histology is Haematoxylin and Eosin (H & E).

Microscopic examination of tissue slides is a crucial stage in the diagnosis of cancer and, simultaneously, the process is time-consuming, subjective and generates considerable interobserver and intraobserver variability. The interobserver variability occurs when different people evaluate the same case obtaining different results, while the intraobserver variability occurs when the same person evaluates a case in different moments in time, also differing in the results. Due to the subjectiveness and time-consuming nature of slide analysis by pathologists, the idea of automating the process at least to some extend sounds appealing, since having an image previously classified will benefit the pathologist, making it's labor faster and more objective. Of course, this tools would be produced as a complement to the work of the pathologist, and not as a substitution for it. Some approaches to automate the analysis of images from cell populations or histology slides, related to cancer cell tracking and classification, uses for example Hidden Markov Models [5], or classifies mitosis stages for phenotype classification [6]. Some other approaches more focused in the efficiency rather than the accuracy are based on convolutional neural networks (CNN) to achieve this classification, such as [7] and [8]. The approach showed in [9], which compares two CNN performance, reports an accuracy of 75%, and as it has been described previously, high accuracy is still crucial for diagnose. In addition, just for the adenocarcinoma cases, it is estimated that in 80% of cases there is presence of a wide mixture of histological patterns that must be qualitatively classified by a pathologist [10], which clearly shows the complexity of the task. Methods based on CNN be- have themselves as a 'black box', which creates a challenge in the extraction of relevant biological information and obtain meaningful insights from the trained models. There is still a need to clearly identify ROIs, in order to make a deeper analysis in specific areas of the issue and have an improvement in diagnose and treatment.

In the present work, the results of an algorithm for segmentation of bright- field microscopy lung cancer tissue images H&E-stained, based on clustering is presented. The clustering algorithm selected is K-means, which identifies different ROIs on the images associated to the dominant colors detected. In order to start working just with relevant areas of the image, a threshold was set in order to remove the background. Then, the histogram analysis will give a significant insight about the behavior of the color distribution in the image, which helps to get an idea about the potentially dominant colors.

Methods

In Figure 1, a simple diagram with the followed process is presented. Initially, the images are converted from SVS format to TIFF format, then is converted to HSV colour space in order to extract the ROI and finally apply K-Means algorithm.

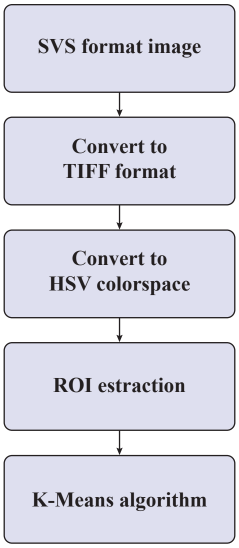


Figure 1. Overview of general process.

Dataset

All the images used for this work are adenocarcinoma images, which were obtained from the Clinical Proteomic Tumor Analysis Consortium Lung Adenocar-cinoma (CPTAC-LUAD) group, which is part of the National Cancer Institute of the United States [11].

The images used for this work are ScanScope Virtual Slide (SVS) format. This format is proprietary from Aperio and it is specifically used for brightfield scanning. The slides are stored in a single file TIFF format, the first image in an SVS file is always the baseline image (full resolution). This image is always tiled, usually with a tile size of 240 x 240 pixels. The second image is always a thumbnail, typically with dimensions of about 1024x768 pixels. Unlike the other slide images, the thumbnail image is always stripped [12].

As these images are in a proprietary format, the first step is to convert them in TIFF format images so they can be processed with tools such as OpenCV [13]. In order to achieve this, the library ImageMagick was used.

It is important to mention that due to the conversion from SVS to TIFF format, in order to use OpenCV the maximum TIFF image size allowed is 1 GB. Nowadays scanned images in SVS format can easily go over 30-40 GB of storage, which brings a challenge in order to process any kind of algorithms on them.

ROI Segmentation

All of the scanned images have a white-tone background which is not relevant for the classification (see Figures 3a and 4a), so, in order to remove it, a threshold was defined as it is clearly showed that the background is not fully white. This threshold was selected experimentally, indicating in the mask, zeros in the pixels within the 'white' range.

The first step was to convert the original RGB image into HSV colour space, this was done because with classic RGB colour space it is more difficult to separate color information from elements such as lighting or intensity. So, after this con- version, the relevant channel is the hue one (H), as it has the color information. Then, using the threshold range previously explained, a mask was applied using an AND bitwise operation between the original HSV image and this mask.

Histogram analysis

The distribution of each of the pixels that make up the image was analyzed via histogram. As it was mentioned before, the relevant channel is the hue channel, as the desired result is to be able to discriminate each group of colors within the image, as one of each colors represent a different structure of the tissue that may be relevant for the pathologist analysis.

After visualizing the histogram for the hue channel, a series of considerations were made in order to select the steps to follow.

Based on the Central Limit Theorem, if each colour is considered as an independent random variable X_1, ..., X_N, with overall mean μ, finite variance σ^2 and $\bar{x}$ is the sample mean. Then the distribution of

$$Z = \frac{\bar{X}_{N - \mu}}{(\sigma\sqrt{N})} \qquad (1)$$

as N → ∞, is the standard normal distribution [14].

After considering this, it is expected in the H channel histogram to get a mixture of gaussian distributions, made up of independent single distributions for each one of the dominant colors in the picture. In the histogram, this can be identified if multiple 'peaks' are being showed.

Segmentation

For the segmentation, a k-means clustering algorithm is used. The k-means is an unsupervised algorithm that it is used to assign to each pixel of the image one of the k-available tags with the degree of belonging to that particular cluster. One of the other parameters selected for the clustering was the amount of iterations, a total of 10 iterations was selected experimentally as a standard value for each test.

Results

The following results correspond to testing using K=3 and K=4 clusters. The amount of clusters selected was chosen based on naked-eye examination and identification on potentially amount of colors within each image. Also, as showed in Figure 2, the Elbow Method was used in order to get a first approach of which of the values may fit with the needs.

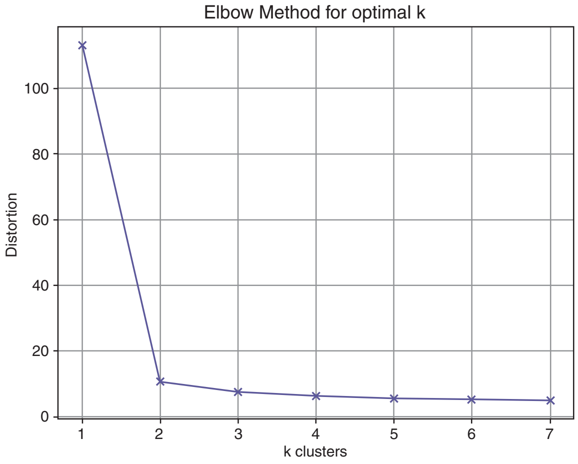
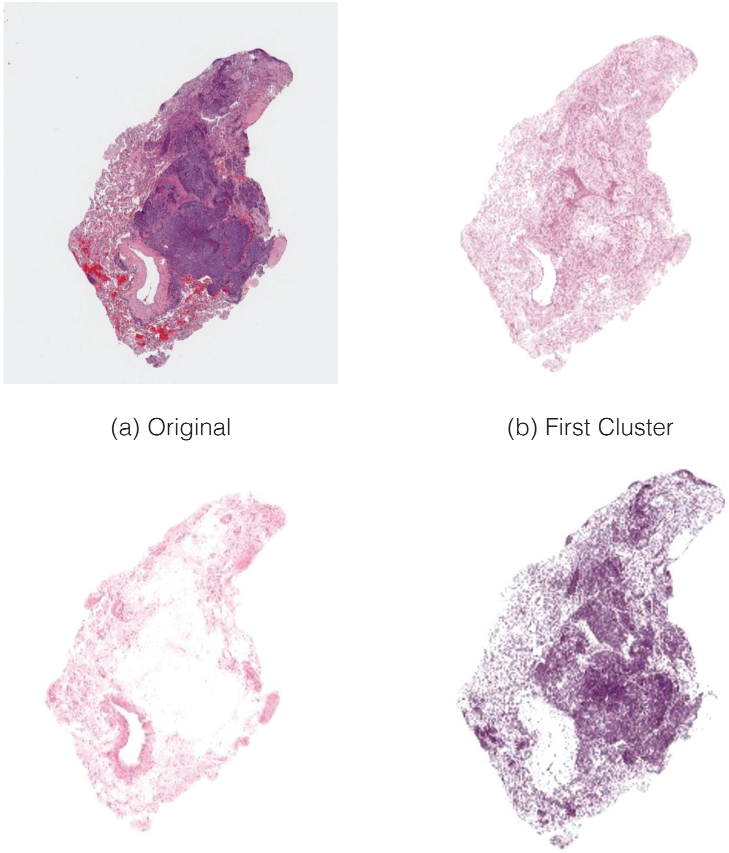


Figure 2. Elbow method to get optimal k.

Discussion and conclusions

The results show an appropriate segmentation within different structures of the tissue images. There is a clear challenge in the areas where the colour is very similar, and also whether to know if the number of clusters that are being suggested by the 'Elbow Method' is precise, since there might be no need to consider some of the areas of the slide that are being detected by the algorithm. The regions of interest for the pathologists may be just a small part of the complete tissue image that it is being processed, so as future work, after having previously annotated images by expert pathologists, then a complete validation of the algorithm can be performed, in order to have a more complete understanding on which parts are completely irrelevant and which ones are not.



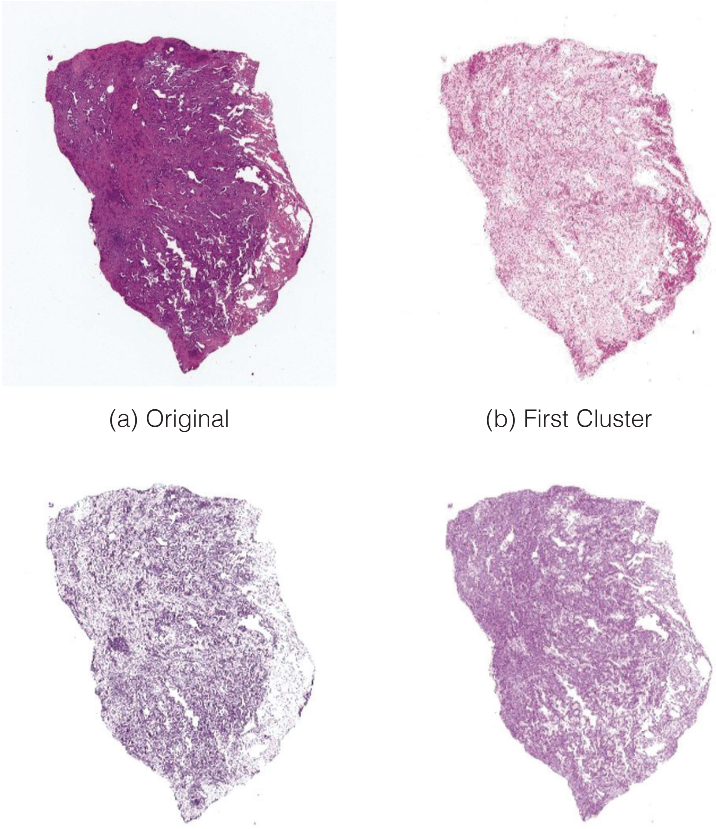
(a) Original

(b) First Cluster

(c) Second Cluster

(d) Third Cluster

Figure 3. Original image and resulting clusters of the first slide.



(a) Original

(b) First Cluster

(c) Second Cluster

(d) Third Cluster

Figure 4. Original image and resulting clusters of the second slide.

In addition, this will create a better understanding of which other methods can be explored not just for the general area segmentation, but then into specific and deeper parts of it.

References

[1] World Health Organization. Cancer key facts. https://www.who.int/news-room/ fact-sheets/detail/cancer, Sep 2018.

[2] American Cancer Society. What is lung cancer? https://www.cancer.org/cancer/ lung-cancer/about/what-is.html, 2019.

[3] The Royal College of Pathologists. What is a biopsy? https://www.rcpath.org/ discover-pathology/news/fact-sheets/what-is-a-biopsy.html, 2018.

[4] Brian Nation and Guy Orchard. Histopathology. Oxford University Press, 2 edition, 2012.

[5] P. Quinde and F. Siles. Diseño, implementación y validación un algoritmo de ras- treo de células cancerígenas basado en hmm's a partir de imágenes de microscopía de campo claro. Universidad de Costa Rica, 2018.

[6] A. Mora and F. Siles. Cell phenotype classification using m-phase features in live-cell bright field time-lapse microscopy. Universidad de Costa Rica, 2018.

[7] B. Peng, L. Chen, M. Shang, and J. Xu. Fully convolutional neural networks for tissue histopathology image classification and segmentation. In 2018 25th IEEE International Conference on Image Processing (ICIP), October 2018.

[8] Y. Xu, Z. Jia, and L. Wang. Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features. BMC Bioinformatics, 2017.

[9] M. Saric, M. Russo, M. Stella, and M. Sikora. CNN-based method for lung cancer detection in whole slide histopathology images. In 2019 4th International Conference on Smart and Sustainable Technologies (SpliTech), pages 1–4, 2019.

[10] W.D. Travis, E. Brambilla, A.P. Burke, A. Marx, and A. G Nicholson. Who classification of tumours of the lung, pleura, thymus and heart, 2015.

[11] National Cancer Institute. Clinical proteomic tumor analysis consortium lung adenocarcinoma. https://wiki.cancerimagingarchive.net/display/Public/ CPTAC-LUAD, 2018.

[12] OpenSlide. Aperio format. https://openslide.org/formats/aperio/.

[13] G. Bradski. The OpenCV Library. Dr. Dobb's Journal of Software Tools, 2000.

[14] G. Cooper and C. McGillen. Probabilistic Methods of Signal and System Analysis. Oxford University Press, NJ, USA, 3rd edition, 1998.