Encuentro Internacional de Educación en Ingeniería ACOFI
2º Congreso Latinoamericano de Ingeniería
2019 10 al 13 de septiembre - Cartagena de Indias, Colombia
RETOS EN LA FORMACIÓN DE INGENIEROS EN LA ERA DIGITAL

SEGMENTACIÓN DE NÚCLEOS EN IMÁGENES HISTOLÓGICAS

Allison Yineth Rodriguez Martínez, Sandra Liliana Cancino Suárez

Escuela Colombiana de Ingeniería Julio Garavito
Bogotá, Colombia

Resumen

Reconociendo que el cuerpo humano está formado por millones de células, y que gracias a la proliferación de estas es que existen tejidos, órganos y sistemas, además, que son las células en caso de enfermedad las que responden a la agresión o lesión; se presenta relevante el hacer investigación a nivel celular con el objetivo de estudiar características propias de las células de cada órgano, como estructura y función, al reaccionar a diversos tratamientos; permitiéndole al investigador comprender los procesos biológicos subyacentes en el trabajo. En función de lo anterior, la base de datos usada en este proyecto es un conjunto de datos obtenidos de la primera fase del concurso Data Science Bowl 2018 (DSB) que consta de 670 imágenes histológicas con sus respectivas máscaras de núcleos previamente segmentados. Dichas imágenes varían entre sí, en el tipo de célula, el tamaño y la modalidad de estás (campo claro frente a fluorescencia). El proyecto tiene como objetivo principal la segmentación de los núcleos en cada uno de los diferentes tipos de imágenes histológicas, a través del uso de técnicas de procesamiento de imágenes. La metodología empleada abarca tres fases, pre-procesamiento, procesamiento y post-procesamiento.

La primera, consta de procesamiento de color en función de las características del histograma para tomar el plano con la mayor cantidad de información posible. La segunda, implica segmentación mediante contornos activos y crecimiento de regiones utilizando múltiples pixeles semilla en las zonas de interés. Finalmente, se realiza un post-procesamiento de los núcleos celulares segmentados por medio de operaciones morfológicas y el empleo de la técnica de Watershed. Los resultados fueron obtenidos por medio de la validación con las máscaras de referencia del total de imágenes de la base de datos. Se usó como métrica de desempeño la correlación entre la imagen de los núcleos segmentados utilizando el método propuesto, y los núcleos segmentados en las máscaras de referencias disponibles. Obteniendo entonces, una correlación promedio de 0.9468 ± 0.1792. Además, en función de las 670 imágenes se determinó tanto la cantidad de falsos

positivos o núcleos no identificables en cada una como la cantidad de falsos negativos o zonas identificadas erróneamente como núcleos.

Palabras clave: imágenes histológicas; núcleos celulares; procesamiento de imágenes

Abstract

We know that the human body is made up of millions of cells, and tissues, organs and systems exist for the proliferation; also, the cells, in case of illness, are those that respond to the aggression or injury. It is relevant to do research at the cellular level with the objective of studying the characteristics of the cells of each organ, such as structure and function, by reacting to different treatments; allowing the researcher to understand the biological processes. For the above, the database used in this project is a set of data obtained from the first phase of the Data Science Bowl 2018 (DSB). This consists of 670 histological images with their respective masks of previously segmented nuclei. These images vary among themselves, in the type of cell, the size and the modality (clear field versus fluorescence). The main objective of the project is the segmentation of the nuclei in each of the different types of histological images, through the use of image processing techniques. The methodology used covers three phases, pre-processing, processing and post-processing. The first consists of color processing based on the characteristics of the histogram to take the plane with as much information as possible. The second involves segmentation through active contours and growth of regions using multiple seed pixels in the areas of interest. Finally, a post-processing of the segmented cell nuclei is performed through morphological operations and the Watershed technique. The results were obtained by validating with the reference masks of the total images of the database. The correlation between the image of the segmented nuclei using the proposed method and the segmented nuclei in the available reference masks was used as the performance metric. Obtaining then, an average correlation of 0.9468 ± 0.1792. In addition, depending on the 670 images, the number of false positives or non-identifiable nuclei in each was determined, as well as the number of false negatives or zones erroneously identified as nuclei.

Keywords*: histological images; nuclei, image processing*

1. Introducción

La importancia de hacer investigación a nivel celular es que permite estudiar características propias de las células de cada órgano, como estructura y función para más adelante entender la patología y otras disciplinas.

Por lo tanto, la identificación de los núcleos de las células es el punto de partida para la mayoría de los análisis a nivel celular, ya que gran parte de los 30 billones de células del cuerpo humano contienen un núcleo. La identificación de núcleos permite a los investigadores identificar cada célula individual en una muestra, y al medir cómo reaccionan las células a diversos tratamientos, el investigador puede comprender los procesos biológicos subyacentes **(**2018 Data Science Bowl | Kaggle, 2018).

Ahora bien, el procesamiento de imágenes ha sido utilizado recientemente para el desarrollo de aplicaciones que permiten automatizar este tipo de procesos, permitiendo obtener resultados rápidos, reproducibles y no sesgados (Khan, et. al., 2018), (Brugger, et. al., 2012). La aplicación de metodologías de procesamiento de imágenes puede ser una forma útil para aumentar la eficiencia en la identificación de células y por ende en los diagnósticos emitidos por profesionales de la salud, permitiendo disminuir el tiempo invertido en el análisis de los datos y a su vez en la entrega de resultados.

Con este propósito, en este artículo se presenta un método para automatizar la identificación de núcleos mediante la creación de un algoritmo generalizado teniendo en cuenta variaciones que presentan las imágenes histológicas.

2. Métodos y Materiales

a. Materiales

670 imágenes histológicas obtenidas de la primera fase del DSB 2018, con sus respectivas máscaras de núcleos previamente segmentados fueron empleadas. Siendo las máscaras la referencia para verificar la calidad de los resultados del método propuesto para la segmentación de núcleos. El algoritmo fue desarrollado en el lenguaje de programación propio de MatLab.

b. Metodología

La metodología empleada para las 670 imágenes abarcó distintas técnicas tal y como se muestra en la figura 1. A continuación, se explicará el alcance de cada fase.

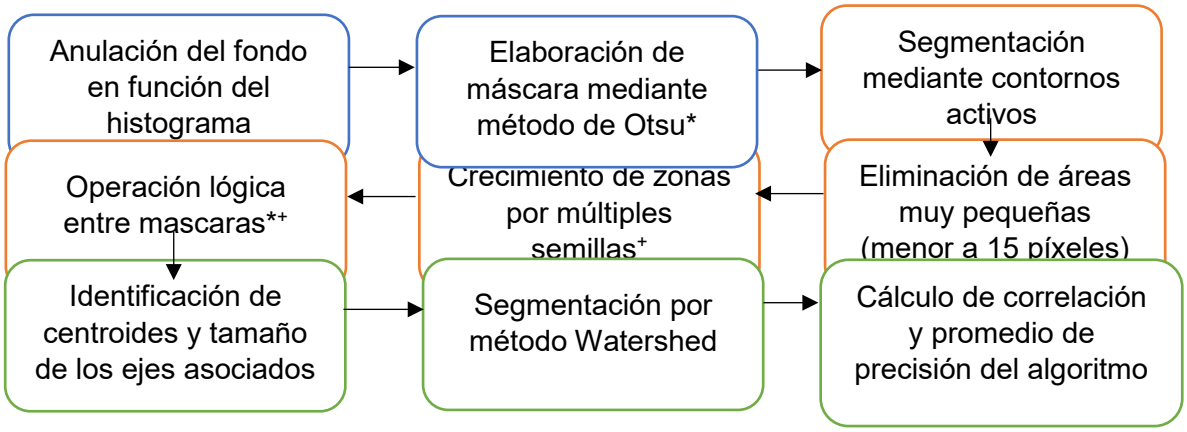

Figura 1. Método propuesto para la identificación de núcleos en imágenes histológicas. Recuadros azules asociados a etapa de pre-procesamiento. Recuadros naranjas asociados a etapa de procesamiento. Recuadros verdes asociados a etapa de post-procesamiento.

- **Etapa de pre-procesamiento**

Esta etapa implica la anulación del fondo en función del histograma debido a que una umbralización fija por color no permitiría que funcionara para todas las imágenes posibles.
Ahora bien, reconociendo que el histograma es una representación gráfica de la cantidad de píxeles en una intensidad de grises y que la estadística permite medir la forma de una distribución por medio de la asimetría, pues esta permite identificar y describir la manera como los datos tienen

a reunirse de acuerdo a la frecuencia; fue con los postulados previos que se realizó la segmentación, pues se sabe que una imagen con fondo claro se caracteriza porque sus datos son más cercanos a uno y una con fondo oscuro es porque sus datos tienden a cero, por lo tanto, si se hace una equivalencia con la asimetría estadística de una distribución, se hablaría de una asimetría negativa y positiva, respectivamente. Así que, si se compara la media, la mediana y la moda de la distribución sabríamos a qué tipo de asimetría corresponde. En este sentido, si la media es menor a la moda se necesitaría invertir la imagen para que el fondo sea oscuro y las zonas de interés blancas. Para el caso de imágenes RGB, fue necesario seleccionar solo un plano de trabajo y esto se empleó al comparar los histogramas de cada componente y establecer el que menor cantidad de píxeles tuviera cercanos a 0 y a este realizarle el procedimiento previamente descrito.

Finalmente, para esta etapa, se generó una máscara mediante la binarización de la imagen por el método de Otsu, el cuál calcula el valor umbral de forma que la dispersión dentro de cada clase de dato sea lo más pequeña posible, pero al mismo tiempo que la dispersión sea lo más alta posible entre segmentos diferentes. Para ello, se calcula el cociente entre ambas varianzas con la característica de que el valor umbral se establece cuando el cociente es máximo.

- **Etapa de procesamiento**

Para esta etapa se considera de manera general la segmentación mediante contornos activos y crecimiento de zonas por múltiples semillas.
Respecto a la segmentación mediante contornos activos, también llamado segmentación con serpientes, se especifican curvas en la imagen que se mueven para encontrar límites de objetos. El objetivo es minimizar la energía controlando la deformación de la curva y ajuste del contorno de la imagen.

Luego, se empleó el crecimiento de zonas por múltiples semillas, el cuál consta que un conjunto inicial de áreas pequeñas se fusiona de manera iterativa de acuerdo con las restricciones de similitud. De tal forma que la región se cultiva a partir del píxel semilla añadiendo píxeles vecinos que son similares, lo que aumenta el tamaño de la región (Gonzalez, et. al., 2004).

- **Etapa de post-procesamiento**

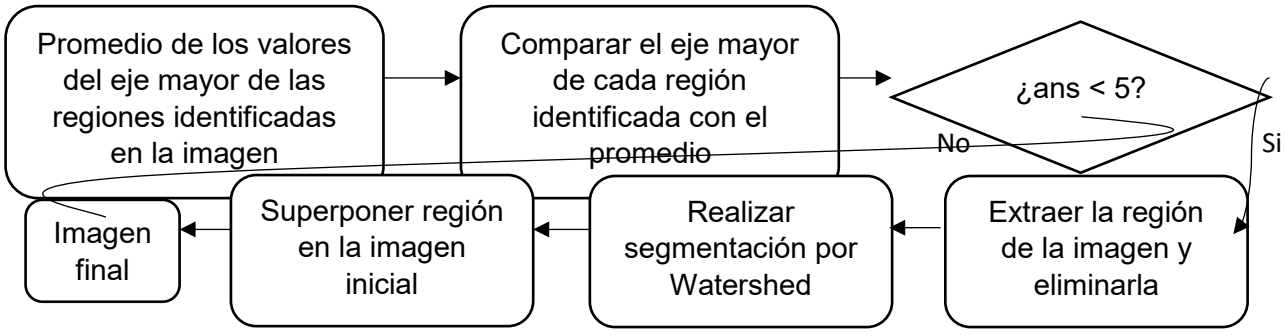

Figura 2. Método empleado para realizar o no segmentación por Watershed

Se generó la segmentación por Watershed, que análogamente es la transformación de la cuenca hidrográfica. La cual genera una transformación de distancia donde los píxeles claros representan

elevaciones altas y los píxeles oscuros elevaciones bajas. Esta se realiza siempre y cuando la zona a segmentar tenga en su eje mayor una diferencia superior a 5 unidades respecto al promedio de los ejes mayores de los otros núcleos, tal y como se muestra en la figura 2.
Por último, la comparación entre la máscara obtenida con respecto a la máscara proporcionada por DSB se estableció en función de la correlación y la cantidad de núcleos detectados en función de la cantidad de centroides identificados y los que se debieron identificar, cantidad de máscaras proporcionadas por cada imagen por DSB.

3. Resultados

A continuación, se mostrarán dos imágenes con fondo distinto, es decir, una con fondo oscuro y otra con fondo claro, y los resultados al aplicar las etapas propuestas en la metodología. Adicionalmente, se presentan los resultados de validación del algoritmo con las 670 imágenes analizadas.

- **Imagen con fondo oscuro**

Respecto a la imagen de fondo oscuro, ver figura 3a, DSB proporcionó su máscara tal y como se muestra en la figura 3b.

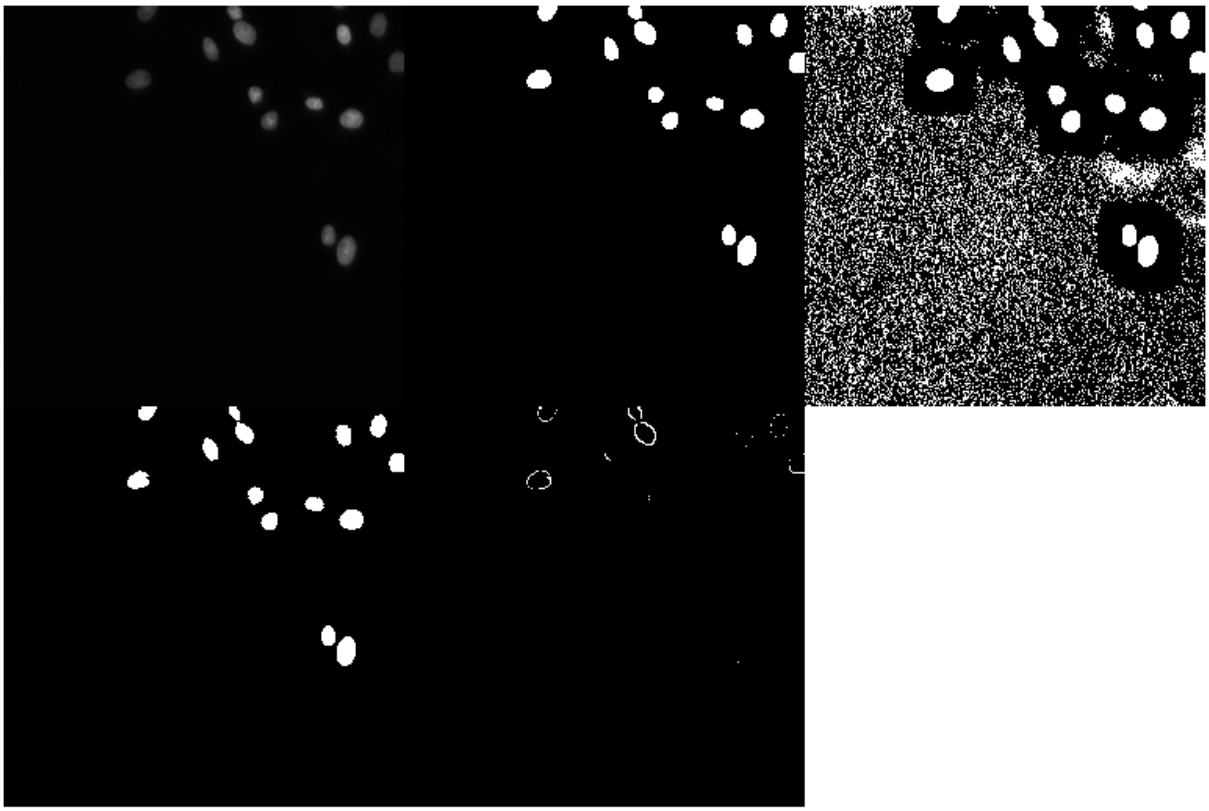

Figura 3. a) Imagen original con fondo oscuro a procesar. b) Mascara original proporcionada por DSB. c) Mascara obtenida en la etapa de pre-procesamiento. d) Mascara final del algoritmo propuesto. e) Diferencias entre la máscara original proporcionada por DSB y la máscara final del algoritmo propuesto.

Para la etapa de pre-procesamiento al correlacionar la máscara, figura 3c, respecto a la proporcionada por DSB, figura 3b, se obtuvo una correlación del 29.10%. Y al aplicar la metodología restante a dicha mascara se obtuvo como mascara final la mostrada en la figura 3d. Al compararla con la proporcionada por DSB, se obtuvieron diferencias en tamaño de algunos núcleos, donde la máscara obtenida en el algoritmo propuesto lo maneja de un tamaño inferior, ver figura 3e. La correlación entre estas máscaras es del 97.18, y de los 14 núcleos a identificar se obtuvieron todos y no hubo identificaciones erróneas, es decir, no se asumieron zonas que no correspondían a núcleos como uno de estos.

- **Imagen con fondo blanco**

Ahora bien, se mostrarán los resultados asociados del procesamiento de una imagen de fondo claro y de tamaño diferente al de la imagen anterior, ver figura 4a. La máscara proporcionada por DSB se encuentra en la figura 4b.

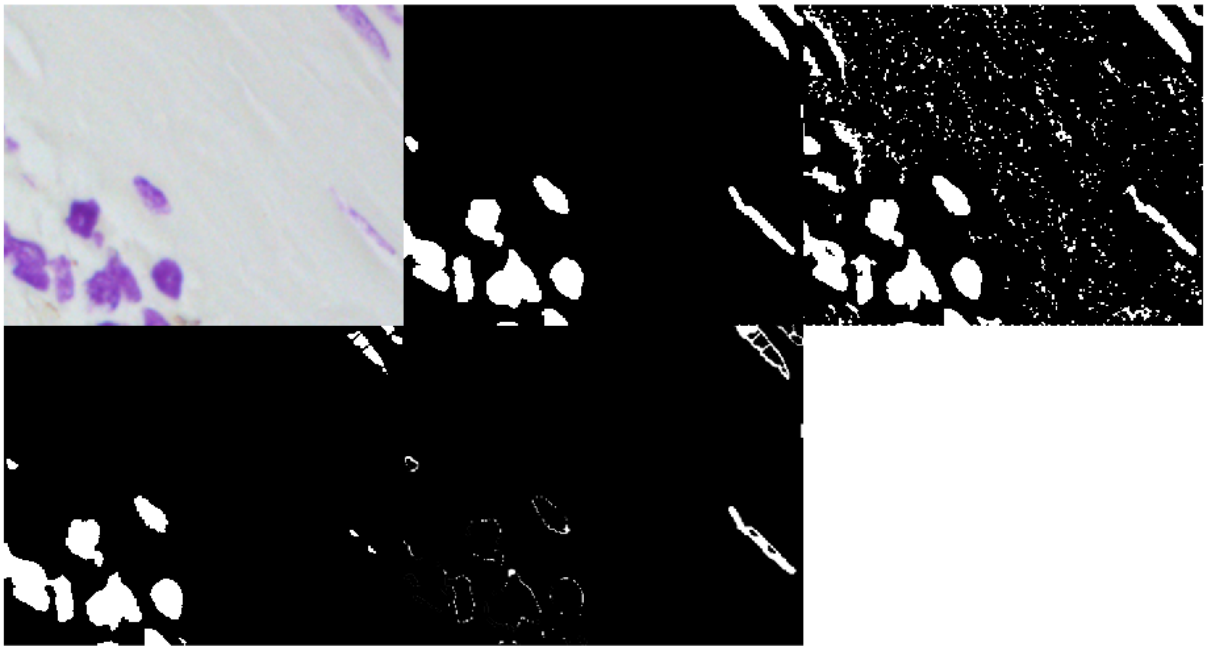

Figura 4. a) Imagen original con fondo claro a procesar. b) Mascara original proporcionada por DSB. c) Mascara obtenida en la etapa de pre-procesamiento. d) Mascara final del algoritmo propuesto. e) Diferencias entre la máscara original proporcionada por DSB y la máscara final del algoritmo propuesto.

La máscara asociada a la etapa de pre-procesamiento de esta imagen, figura 4c, tiene una correlación con la proporcionada por el DSB, figura 4b, del 69.97% y se observan la mayor cantidad de diferencias en zonas que no son núcleos, pero que hasta el momento se asumieron que sí.

La máscara final, ver figura 4d, se comparó con la proporcionada por el DSB, obteniendo las diferencias visualizadas en la figura 4e. La correlación entre estas máscaras es del 97.48%, y de los 19 núcleos a identificar se obtuvieron 17 regiones identificadas de los cuales 5 deberían ser 3 pues se dividieron incorrectamente en la segmentación por Watershed y otros no fueron divididos por lo que se asumen como uno, cuando deberían ser varios.

Finalmente, al validar el algoritmo, en función de las 670 imágenes, mediante la correlación se obtuvieron valores en la sección de pre-procesamiento de 0.6433 ± 0.2356 y en la de procesamiento de 0.9468 ± 0.1792, además, de 30801 núcleos a identificar en las 670 imágenes, se identificaron 21419 regiones como núcleos, de los cuales 1082 fueron falsos positivos. Lo anterior, se encuentra condensado en la tabla 1.

Tabla 1. Matriz de confusión al identificar regiones como núcleo o no.

	Clasificación de regiones según las mascaras proporcionadas por el DSB	
Resultado de la identificación de núcleos por el algoritmo propuesto	**Es núcleo**	**No es núcleo**
Es núcleo	20337	1082
No es núcleo	10464	

Se reconoce que el procesamiento no exige muchos recursos computacionales aun cuando si automatiza de manera adecuada la mayoría de los núcleos de las células en las imágenes. Además, los errores presentados en la identificación de núcleos pueden reducirse introduciendo mejoras al algoritmo, basadas en la observación del tipo de imágenes que presentan mayor dificultad en la identificación correcta de los núcleos, instancia actual del proyecto; y el empleo de inteligencia artificial para generar resultados más robustos.

4. Conclusiones

Una metodología alterna para la identificación de núcleos en imágenes histológicas ha sido presentada. Apoyando y sustentando aportes en la investigación celular, por parte de ingeniería biomédica con áreas disciplinarias siendo estas: biología celular y procesamiento de imágenes, permitiendo modificar los tiempos que un investigador invierte en la identificación de zonas de interés y en el análisis de datos.

5. Referencias

Libros

- Gonzalez R., Woods R., Eddins S. (2003). Digital Image Processing Using MATLAB. Pearson. pp. 408 – 410.

Artículos de revistas

- Brugger S.D., Baumberger C., Jost M., Jenni W., Brugger U., Mühlemann K. (2012). Automated Counting of Bacterial Colony Forming Units on Agar Plates, PLoS ONE Vol. 7, Issue 3, e33695.
- Khan, AU, Torelli, A, Wolf, I, Gretz, N, AF Khan, Arif Ul Maula, Torelli, Angelo, Wolf, Ivo, Gretz, Norbert TI. (2018). AutoCellSeg: robust automatic colony forming unit (CFU)/cell analysis using adaptive image segmentation and easy-to-use post-editing techniques. SCIENTIFIC REPORTS, Vol. 8. 7302.

Fuentes electrónicas

- 2018 Data Science Bowl | Kaggle. (2018). Spot Nuclei. Speed Cures. Consultado el 12 de marzo de 2019 en https://www.kaggle.com/c/data-science-bowl-2018/overview/description

Sobre los Autores

- **Allison Yineth Rodriguez Martinez:** Estudiante del Programa de Ingeniería Biomédica, Miembro del Semillero de Investigación Procesamiento de Imágenes y Señales PROMISE. Escuela Colombiana de Ingeniería Julio Garavito – Universidad del Rosario. allison.rodriguez@mail.escuelaing.edu.co
- **Sandra Liliana Cancino Suarez:** Profesora Asociada del Programa de Ingeniería Biomédica en el área de procesamiento de señales e imágenes médicas. Ingeniera Electrónica de la Escuela Colombiana de Ingeniería y Magister en Ingeniería de la Universidad Los Andes. Escuela Colombiana de Ingeniería Julio Garavito. sandra.cancino@escuelaing.edu.co