

# Segmentación de núcleos celulares mediante métodos de aprendizaje de máquina

Eduard Leonardo Sierra Ballén \*

\*Maestría en Ingeniería - Ingeniería de Sistemas y Computación

Departamento de Ingeniería de Sistemas e Industrial

Universidad Nacional de Colombia

Email: elsierrab@unal.edu.co

**Abstract**—El principal objetivo de este trabajo es desarrollar métodos de medida para dar soporte al patólogo en el diagnóstico de lesiones escamosas cervicales. Para esto, una de las primeras labores se concentra en la segmentación de las células obteniendo una representación más simple de la imagen para poder establecer medidas que sean útiles en el diagnóstico de dichas lesiones. El presente escrito muestra algunos avances iniciales en la segmentación de núcleos celulares.

**Index Terms**—Cancer cervical, lesiones escamosas, segmentación de imágenes en color, segmentación de células, Maquinas de vectores de soporte (Support Vector Machine SVM), perceptrones multicapa.

## I. INTRODUCCIÓN

EL cáncer cuello uterino (cáncer cervical) es hoy el segundo más común en mujeres en todo el mundo [1], el diagnóstico temprano de esta enfermedad en una etapa precancerosa es muy importante para la efectividad de los tratamientos. Existen diversos exámenes para su diagnóstico como la citología vaginal, la colposcopia, la inspección visual directa, las espectroscopías fluorescente y reflectiva, entre otras; la mayoría de países del tercer mundo usan estudios de citología vaginal o inspección visual directa ya que son de muy bajo costo. En los sitios donde se ha implementado la citología vaginal se ha logrado bajar la incidencia de cáncer en un 70% pero esta prueba tiene inconvenientes ya que la sensibilidad de la misma es sólo del 51%, y la interpretación para dar un diagnóstico es altamente subjetiva [1]. Se han desarrollado diferentes proyectos que permitan ayudar a controlar esta enfermedad.

Como por ejemplo el desarrollo de redes de investigación para estudiar la incidencia del virus del papiloma humano en el desarrollo de este cancer como se ve en [1].

Este artículo trata de la segmentación de núcleos celulares, esta tarea es un paso inicial en el desarrollo de un proyecto en el cual se pretende determinar medidas morfométricas sobre imágenes de cervix. Dichas medidas se toman con el fin de ayudar a los expertos patólogos en el diagnóstico de lesiones escamosas intraepiteliales, las cuales son un indicador de la presencia de tejido canceroso en el cuello uterino. Para esto, el proyecto se apoya en el conocimiento a priori acerca de los criterios empleados por los patólogos para dicho diagnóstico, algunos de estos son:

- Pérdida de la polaridad de las células.

- Mitosis anormales.
- Alteración de la relación núcleo-citoplasma.
- Hiperchromasia nuclear.
- Alteración de la maduración.

El artículo se divide de la siguiente manera la sección III presenta el conjunto de imágenes y algunos métodos empleados, en la sección IV se revisan los experimentos realizados y resultados obtenidos y finalmente en la sección V se dan algunas conclusiones y perspectivas de trabajo a desarrollar.

## II. ANTECEDENTES

### A. Diagnóstico de cáncer asistido por computador

Recientemente se han reportado diferentes proyectos que buscan introducir el diagnóstico automático de cáncer basado en imágenes histopatológicas. En el reporte técnico de Demir y Yener [2] se encuentran documentados algunos de los pasos del diagnóstico de cáncer, ilustrados en la figura 1, estos pasos incluyen: preprocesamiento para eliminar ruido y segmentar regiones de interés en imágenes microscópicas, selección y extracción de características. Para el caso del cáncer de origen cervical entre las características mencionadas se encuentra la textura, basada en fractales y la topología. Finalmente en el diagnóstico se emplean algoritmos de aprendizaje con clasificadores, métodos de evaluación y pruebas estadísticas. Otro ejemplo es del de Price et. al. quien presenta el desarrollo de un sistema de decisión basado en redes de decisión bayesianas (bayesian belief networks) para dar soporte más objetivo al diagnóstico asociado a los exámenes de patología; que permitan a patólogos dar mejor sus diagnósticos, reproducirlos y que puedan ayudar a patólogos jóvenes en el aprendizaje [3].

Por otro lado, Loukas y Linney [4], hicieron un análisis de imágenes histológicas basado en tres factores biológicos que influyen en la respuesta de los tumores a la radiación: la proliferación, la hipoxia y la vasculatura. Al igual que la revisión de Demir y Yener [2] presentan diferentes pasos donde se destaca el problema de segmentar las células.

### B. Segmentación de células

La segmentación de células se ha tratado desde diversos enfoques que incluyen técnicas tradicionales de imágenes usando umbralización, detección de bordes con máscaras de gradiente, hasta técnicas donde se aprovecha la información

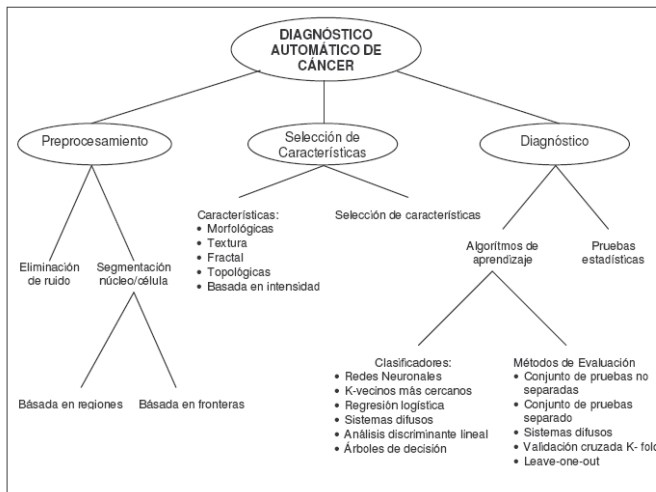


Fig. 1. Revisión de pasos en diagnóstico automático de cáncer. Adaptado de Demir y Yener [2]

de color incluyendo métodos usando modelos donde se separa la componente de iluminación como el HSI (Hue Saturation Intensity) y el HSL (Hue Saturation Luminance). También uso de algoritmos de clasificación de píxeles en el espacio de color RGB, usando análisis de componentes principales (ACP), fusión de regiones, emparejamiento de plantillas (contornos activos, transformada de Hough), análisis de texturas y clasificación (Redes neuronales, lógica difusa) [4].

Lezoray [5] en su tesis doctoral hace segmentación de células en imágenes en color con morfología matemática aplicada a imágenes de redes neuronales. Un trabajo anterior hecho por Walker et al. en 1994 [6], aplicado a células cervicales usa también morfología y extracción de características de textura.

Otros enfoques tratan de solucionar problemas que se presentan con segmentación de células usando representaciones paramétricas para hacer emparejamiento de plantillas como contornos activos o transformada Hough [4]. Entre estas también hay técnicas que plantean el problema de segmentar las células como un problema de optimización de una representación paramétrica de la forma celular. Segmentación satisfactoria de células aisladas, múltiples células y células traslapadas se logra con la técnica de Wu, Barba y Gil [7]. Este método hace minimización sobre una representación paramétrica pero el costo computacional es prohibitivo para imágenes grandes donde hay muchas células. Otro esquema basado también en optimización usa un método evolutivo basado en búsqueda tabú de Jiang y Yang [8]. Este método usa memoria de corto plazo y tiene una lista para diversificar el espacio de búsqueda pero el problema está precisamente en la determinación del tamaño de esta lista. El número de familias es importante, si es pequeño se acerca a la búsqueda tabú, de otra manera la velocidad de convergencia empeora. Además el determinar al número de hijos también afecta la velocidad de convergencia para ese caso usaron valores entre 2 y 8. Además los autores encontraron que este método tenía casos donde no se obtenían segmentaciones óptimas.

Entre las técnicas de segmentación de células con aplicaciones en cancerología hay una comparación de métodos supervisados, basada en píxeles hecho por Meurie et al. [9], con imágenes de cáncer bronquial, en donde usa un esquema de segmentación en cinco etapas: simplificación de imágenes usando pre-tratamiento con suavizado, eliminación de ruido; seguido de clasificación de píxeles usando Bayes, kNN, SVM; extracción de marcadores basados en morfología matemática; luego watersheds para hacer crecimiento usando la información de color de la imagen obteniendo las regiones correspondientes al núcleo y al citoplasma. En este trabajo se concluyó que es de importancia la escogencia del binomio compuesto por el modelo de color y del clasificador para obtener mejores resultados, hay diferentes combinaciones que dan mejores resultados ellos encontraron que para Bayes los mejores espacios de color son YCh1Ch2, RGB y XYZ, para SVM son YCh1Ch2, HSL y RGB, y para kNN los mejores son HSL, XYZ y RGB [9].

### III. MATERIALES Y MÉTODOS

#### A. Material Histológico

El proyecto se desarrolla con imágenes digitales en color tomadas de muestras de cérvix con método de tinción (Hematoxilina-Eosina).

#### B. Análisis de imágenes

El estudio para determinar la medidas se ha centrado en la segmentación de los núcleos celulares. Para esto se realizaron algunas pruebas donde se analizó el uso de diversas técnicas de las disponibles en la literatura. Para las imágenes disponibles se identificaron varios casos de presencia de núcleos en los que se pueden identificar los núcleos celulares. Estas son núcleos aislados, en mitosis, traslapados con otras células y traslapados con otros núcleos celulares, además hay casos donde en este tipo de tejidos se presentan células que tienen más de un núcleo, la figura 2 presenta algunos de estos casos.

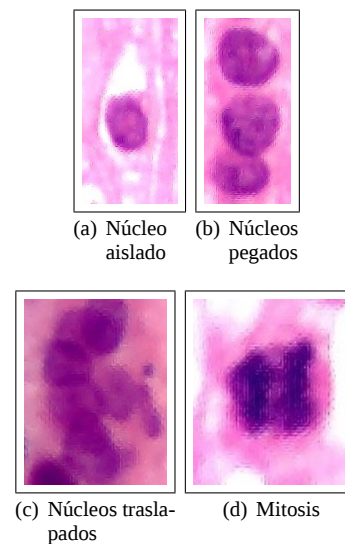
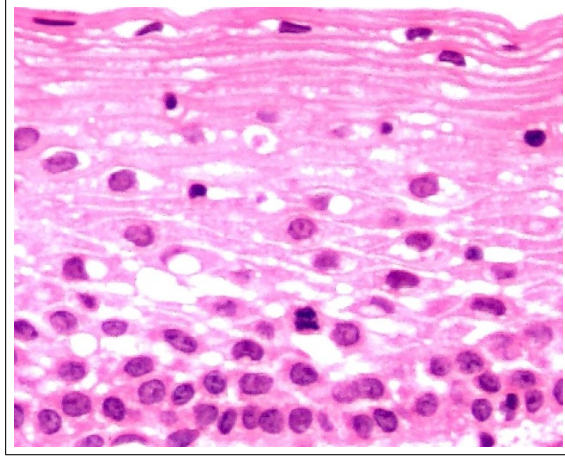
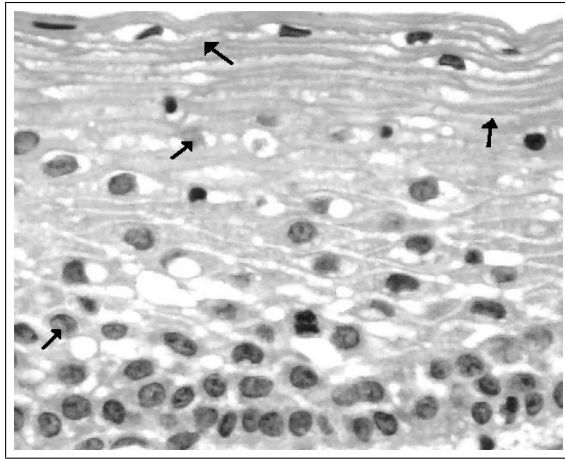


Fig. 2. Algunos casos que se presentan para la segmentación de núcleos celulares.

Hasta el momento se ha trabajado específicamente en la determinación de núcleos aislados. Inicialmente el problema de segmentación se trabajó sobre las imágenes en escalas de grises, tomando la componente de iluminación  $Y$  del modelo de color  $YIQ$  ( $0.299R+0.587G+0.114B$ ), la figura 3 presenta un ejemplo donde se nota cómo al emplear la imagen en escala de grises se pueden confundir algunos núcleos esto debido al fenómeno de metamerismo [10]



(a) Imagen Original



(b) Imagen en escala de grises

Fig. 3. Imagen original en color y escala de grises usando la componente  $Y$  del modelo de color  $YIQ$ , las flechas indican algunos sitios donde se pueden confundir núcleo y citoplasma celular

La complejidad de estas imágenes se hace manifiesta debido al método de captura empleado ya que el método de tinción afecta bastante la calidad de las imágenes, lo cual obliga a buscar alternativas de segmentación empleando la información de color.

### C. Segmentación de núcleos celulares

Debido a que la coloración de los objetos en cada imagen puede variar bastante, pero que también se encuentra que en general la coloración se encuentra separada en cuanto a su tonalidad, los núcleos son de tonalidad violeta casi azul, y el citoplasma de coloración rosa. Se plantea un esquema de segmentación de imágenes que haga una separación en el espacio de color RGB, esta segmentación se hace mediante

una metodología experimental usando métodos de aprendizaje de máquina, específicamente se está trabajando con máquinas de vectores de soporte.

1) *Máquinas de Vectores de Soporte (Support Vector Machine -SVM-)*: Este es un método de aprendizaje de máquina que hace una clasificación que se aplica a patrones de entrenamiento de la forma  $\{y_i, \mathbf{x}_i\}$ , donde  $y_i \in \{1, -1\}$ , y  $\mathbf{x}_i \in \mathbf{R}^n$ . El método da por resultado un hiperplano de separación determinado por la ecuación:

$$\mathbf{w}^T \mathbf{x} + w_0 = 0 \quad (1)$$

Donde la clasificación estaría dada por:

$$\mathbf{w}^T \mathbf{x}_i + w_0 \geq 1, \text{ para } y_i = 1 \quad (2)$$

$$\mathbf{w}^T \mathbf{x}_i + w_0 \leq -1, \text{ para } y_i = -1 \quad (3)$$

Se plantea un esquema donde se busca maximizar la separación entre las clases usando unos puntos de entrenamiento denominados vectores de soporte los cuales van a soportar el hiperplano. Esto es equivalente a la minimización de  $\frac{1}{2} \|\mathbf{w}\|^2 = \frac{1}{2} \mathbf{w}^T \mathbf{w}$ . Para obtener el vector de pesos  $\mathbf{w}$  y el sesgo  $w_0$ . Así el problema se plantea cómo:

Minimizar :

$$J(\mathbf{w}) = \frac{1}{2} \mathbf{w}^T \mathbf{w} \quad (4)$$

Sujeto a:

$$y_i(\mathbf{w}^T \mathbf{x}_i + w_0) - 1 \geq 0 \quad i = 1, \dots, N \quad (5)$$

Usando el método de multiplicadores de Lagrange esto es:

Minimizar :

$$J(\mathbf{w}, w_0, \alpha) = \frac{1}{2} \mathbf{w}^T \mathbf{w} - \sum_{i=1}^N \alpha_i [y_i(\mathbf{w}^T \mathbf{x}_i + w_0) - 1] \quad (6)$$

Se encuentra que el dual de este problema, se puede expresar en términos de los datos de entrenamiento de manera que se obtiene:

Maximizar :

$$Q(\alpha) = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j \quad (7)$$

sujeto a  $0 \leq \alpha_i \leq C$  y  $\sum_i \alpha_i y_i = 0$  y  $C$  es un parámetro a seleccionar que se utiliza cuando las clases no son linealmente separables.

De la solución del anterior problema se puede obtener los valores de  $\mathbf{w}$  y  $w_0$  usando las expresiones:

$$\mathbf{w} = \sum_{i=1}^N \alpha_i y_i \mathbf{x}_i \quad (8)$$

$$w_0 = 1 - \mathbf{w}^T \mathbf{x}^{(s)} \quad (9)$$

donde  $\mathbf{x}^{(s)}$  es un vector de soporte, es decir un vector donde  $y_i = \pm 1$  y  $\alpha_i \neq 0$ . Detalles más a fondo acerca de este método se pueden ver en [11], [12].

#### IV. EXPERIMENTACIÓN Y RESULTADOS

##### A. Conjunto de datos

Dado que se emplearon técnicas basadas en aprendizaje de máquina se requería tomar un conjunto de entrenamiento para esto se tomó sobre una imagen un conjunto de 29781 puntos marcados por un patólogo experto, de estos puntos se rein su coordenada  $x, y$  en la imagen y su color en modelo RGB. Donde los puntos que el patólogo determinó corresponden a las fronteras de núcleo y citoplasma celulares en un conjunto de 35 células.

De él conjunto de datos se tomó solamente la información de color en modelo RGB, con este se entrenó un SVM aplicando el algoritmo SMO de la librería Weka [13]. La figura 4 muestra la imagen original, los contornos marcados por el experto y la segmentación de núcleos empleando SVM. Se puede observar como existen casos donde los núcleos no

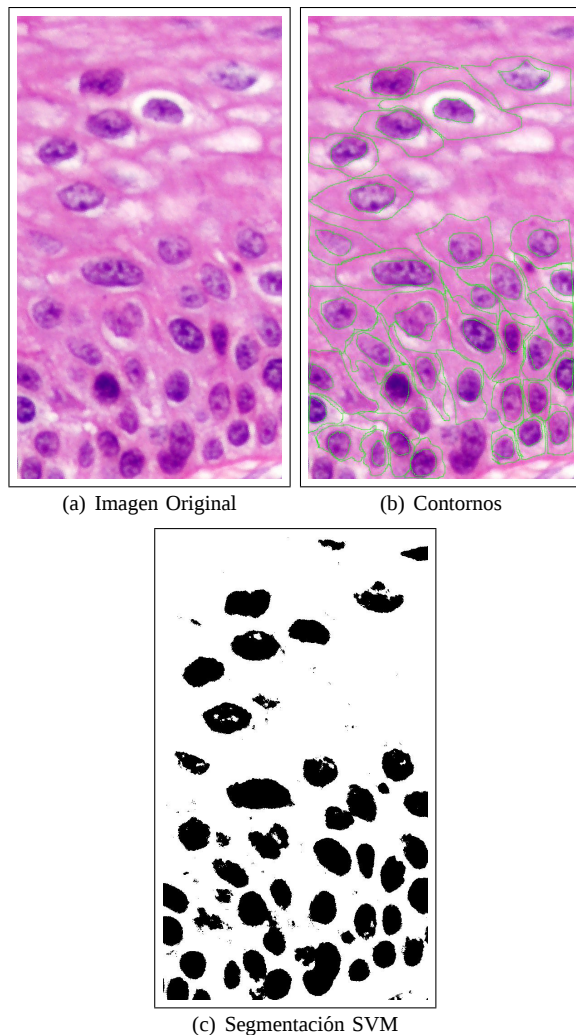


Fig. 4. Imagen original, contornos delineados por un experto y resultados de segmentación usando SVM algoritmo SMO

quedan totalmente separados especialmente en las regiones donde hay alta proliferación de células, además de huecos dentro de algunos núcleos producto de tonalidades similares a las del citoplasma, además de regiones del citoplasma cuya

tonalidad la parecer es similar a la del núcleo y se marcan como núcleo (ver figuras 4(b) y 4(c) izquierda abajo).

##### V. CONCLUSIONES Y TRABAJO FUTURO

Se presentó el primer paso para la construcción de una morfometría en imágenes de cérvix empleando la información de color en el modelo de color RGB para hacer una segmentación de núcleos celulares.

Las imágenes de cérvix debido al método de tinción usado y a fenómenos como el metamerismo no son buenas candidatas para ser procesadas en escala de grises.

La aproximación presentada es un paso inicial, se plantea la necesidad de buscar un modelo de segmentación donde los núcleos aislados queden bien segmentados ya que como se ve en la figura 4 este método requiere afinarse porque aún da regiones que no corresponden.

Se debe revisar la segmentación afinando parámetros del SVM, además de explorar modelos de color diferentes la RGB como es el caso de modelos que trabajan el color desde el punto de vista de percepción humana como el CIELab o modelos donde se separe el color en componentes separadas para tonalidad, saturación e iluminación como HSV ó HSL.

##### REFERENCES

- [1] D. G. Ferris, "Analysis of digitized cervical images detect cervical neoplasia," in *Proceedings of SPIE*, ser. Medical Imaging 2004: Image Processing, J. M. Fitzpatrick and M. Sonka, Eds., vol. 5370, 2004, pp. 181–194.
- [2] C. Demir and B. Yener, "Automated cancer diagnosis based on histopathological images: a systematic survey," Rensselaer Polytechnic Institute, Department of Computer Science., Troy, NY 12180, US, Tech. Rep. TR-05-09, March 2005. [Online]. Available: <http://www.cs.rpi.edu/research/pdf/05-09.pdf>
- [3] G. J. Price, W. G. McCluggage, M. L. Morrison, G. McClean, L. Venkatraman, J. Diamond, H. Bharucha, R. Montironi, P. H. Bartels, D. Thompson, and P. W. Hamilton, "Computerized diagnostic decision support system for the classification of preinvasive cervical squamous lesions," *Human Pathology*, vol. 34, no. 11, pp. 1193–1203, 2003.
- [4] C. G. Loukas and A. Linney, "A survey on histological image analysis-based assessment of three major biological factors influencing radiotherapy: proliferation, hypoxia and vasculature," *Computer Methods and Programs in Biomedicine*, vol. 74, no. 03, pp. 183–199, June 2004.
- [5] O. M. Lezoray, "Segmentation d'images couleur par morphologie mathématique et classification de données par réseaux de neurones : application à la classification de cellules en cytologie des séreuses," Ph. D. These, Ecole Doctorale Structure, Information, Matière et Matériaux, U.F.R de Sciences, Université de Caen/Basse-Normandie, 2000. [Online]. Available: <http://users.info.unicaen.fr/lezoray/Publis/These.pdf>
- [6] R. Walker, P. Jackwa, B. Lovell, and D. Longstaff, "Classification of cervical cell nuclei using morphological segmentation and textural feature extraction," in *2nd ANZIS Australian and New Zealand Conference on Intelligent Information Systems*, 1994.
- [7] H.-S. Wu, J. Barba, and J. Gil, "A parametric fitting algorithm for segmentation of cell images," *IEEE Transactions on Biomedical Engineering*, vol. 45, no. 03, pp. 400–407, March 1998.
- [8] T. Jiang and F. Yang, "An evolutionary tabu search for cell image segmentation," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 32, no. 05, pp. 675–678, October 2002. [Online]. Available: <http://nlpr-web.ia.ac.cn/english/staff/jiangtz/Jiang.Yang.IEEE.TSMCB02.pdf>
- [9] C. Meurie, G. Lebrun, O. Lezoray, and A. Elmoataz, "A comparison of supervised pixel-based color image segmentation methods. application in cancerology," *WEAS Transactions on Computers, special issue on ICOSSIP'03*, vol. 02, no. 03, pp. 739–744, July 2003. [Online]. Available: <http://users.info.unicaen.fr/lezoray/Publis/meurie.icossip03.pdf>
- [10] T. Gevers and F. C. A. Groven, "Segmentation of color images," in *Proceedings of 7th Scandinavian Conference on Image Analysis*, 1991. [Online]. Available: <http://staff.science.uva.nl/gevers/>

- [11] S. Haykin, *Neural Networks: A Comprehensive Foundation*, 2nd ed. Prentice Hall, 1999, ch. Support Vector Machines.
- [12] C. J. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, no. 2, pp. 121–167, 1998.
- [13] I. H. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques*. San Francisco: Morgan Kaufmann, 2005.

PLACE  
PHOTO  
HERE

**Eduard Sierra** Ingeniero de Sistemas de la Universidad Nacional de Colombia en 2002. Docente del Programa de Ingeniería en Multimedia Universidad Militar "Nueva Granada" desde 2003. En 2005 inició sus estudios de Maestría en Ingeniería - Ingeniería de Sistemas y Computación en la Universidad Nacional de Colombia.