LIBRO DE ACTAS

INGENIERIA BIOMEDICA avanzando hacia el futuro

Ciudad Real, 21 - 23 noviembre de 2018

CASEIB 2018 XXXVI Congreso Anual de la Sociedad Española de Ingeniería Biomédica

Diseño y desarrollo de un sistema para la detección automática de sangre en imágenes de cápsula endoscópica

P. Pons Suñer[1], R. Alexander Noorda[1], V. Naranjo[1]. A. Nevárez Heredia[2], V. Pons Beltrán[2]

[1] Grupo de investigación CVBLab, Universitat Politècnica de València, Valencia, España,

pedropons96@gmail.com, vnaranjo@dcom.upv.es, reinoo@upv.es

[2] Unidad de Endoscopia Digestiva y el Grupo de Investigación de Endoscopia Digestiva, Hospital Universitari i Politècnic La Fe, Valencia, España, nevarezja78@gmail.com, pons_vicbel@gva.es

Resumen

La endoscopia por cápsula inalámbrica permite observar el tracto gastrointestinal completo de forma sencilla y no invasiva. Sin embargo, se genera una gran cantidad de imágenes por examen que los médicos tardan aproximadamente 2 horas en analizar. Esto no solo supone un elevado coste, sino que el diagnóstico puede ser erróneo debido a la fatiga y a la naturaleza variable de las lesiones, que exige una alta concentración.

En el presente trabajo se diseña y desarrolla un sistema capaz de detectar automáticamente aquellas imágenes que contienen sangre, siguiendo dos enfoques distintos. El primero consiste en escoger y extraer ciertas características de color de las imágenes con las que entrenar modelos de aprendizaje automático clásico (SVM y Random Forest) que permitan distinguir entre tejido sano y sangre. Además, se implementa la técnica de segmentación "waterpixels" para tratar de mejorar la clasificación. El segundo método consiste en utilizar técnicas de aprendizaje profundo (redes neuronales convolucionales), capaces de extraer las características relevantes de la imagen por sí solas. La configuración que ha obtenido los mejores resultados (95,7% de sensibilidad y 92,3% de especificidad) ha sido un modelo Random Forest entrenado con los histogramas de los canales del espacio de color HSV.

1. Motivación y objetivos

1.1. Introducción

Actualmente, el personal clínico dispone de múltiples técnicas e instrumentos que permiten diagnosticar las enfermedades que afectan al tracto gastrointestinal (tracto GI) humano. Las técnicas convencionales como la gastroscopia y la colonoscopia permiten acceder tanto con finalidad diagnóstica como terapéutica a distintos tramos del tracto GI. Sin embargo, comparten una misma limitación: no permiten la observación del intestino delgado completo, vital para el diagnóstico de numerosas patologías. Este tramo, de mayor longitud, es tan solo accesible por medio de otras técnicas más invasivas como la enteroscopia por empuje o la endoscopia intraoperatoria.

La endoscopía por cápsula endoscópica (en inglés, w*ireless capsule endoscopy* o WCE), introducida en el año 2000 por la empresa Given Imaging, permite observar el tracto GI completo de forma sencilla y no invasiva, aunque sin capacidad terapéutica. El paciente tan solo ha de ingerir una cápsula endoscópica (Figura 1), que va avanzando por el tubo digestivo gracias a los movimientos peristálticos. A medida que avanza, la cámara instalada en uno de sus extremos va tomando entre dos y tres imágenes por segundo del interior del tubo, iluminado con una linterna.

Figura 1. *Cápsula endoscópica PillCam™SB3*

Uno de los principales problemas de la WCE reside en la gran cantidad de imágenes que se generan por examen (más de 150.000). Aunque el *software* proporcionado por el fabricante, RAPID Reader, permite visualizarlas en forma de vídeo a velocidad rápida, los médicos experimentados tardan aproximadamente entre 1 y 2 horas en analizar el vídeo completo. Esto no solo supone un elevado coste, sino que además produce una importante fatiga en el médico, lo cual aumenta el riesgo de realizar un diagnóstico erróneo. Además, este riesgo se ve agravado por la naturaleza de las lesiones, de tamaño variable, distribuidas aleatoriamente por las imágenes y en ocasiones visibles en una o muy pocas imágenes.

En este trabajo se centra la atención en la detección automática de sangre en imágenes de WCE, ya que su presencia en el intestino es síntoma de numerosas patologías del tracto GI, como pueden ser pólipos, tumores, úlceras o la enfermedad de Crohn. Así pues, la detección de sangre suele considerarse un paso prioritario preliminar a la búsqueda de otras anormalidades más específicas.

Aunque RAPID Reader dispone de una herramienta de detección automática de sangre en las imágenes de WCE, llamada SBI, varios estudios afirman que este sistema presenta tanto una baja sensibilidad como una baja especificidad [1-3], por lo que no puede ser utilizado como una herramienta fiable con la que tomar un diagnóstico.

1.2. Revisión bibliográfica

Son muchos los esfuerzos realizados para tratar de desarrollar sistemas que permitan detectar de forma automática las anomalías en el tracto GI, siguiendo enfoques muy diversos. En cuanto al tipo de características de las imágenes que se utilizan, algunos autores optan por utilizar la información del color junto con información de

ISBN: 978-84-09-06253-9, pags. 257- 260

las texturas, aunque aquellas relacionadas con el color demuestran ser mucho más discriminativas [4]. La mayoría de los autores explora distintos espacios de color a fin de encontrar las características que mejor separen la sangre del tejido sano. Aunque la mayoría utiliza, entre otras opciones, el espacio RGB, su uso puede resultar problemático debido a que la iluminación en las imágenes de WCE no es uniforme, pues la linterna ilumina con mayor intensidad la zona del tejido más cercana, afectando drásticamente a los tres canales del espacio RGB [5]. Por otro lado, el espacio HSV constituye una alternativa más popular, pues este problema tan solo afecta al canal V.

Se encuentran también distintos enfoques en cuanto a la estrategia de procesamiento de las imágenes. Los métodos de análisis a nivel de pixel son los más sencillos y rápidos, pero ignoran la información espacial, proporcionando a menudo resultados poco satisfactorios e incoherentes. Los métodos a nivel de imagen aprovechan esta información, pero pierden la capacidad de detectar las lesiones más pequeñas, cuyas características quedan diluidas en las del resto de la imagen. Por último, los métodos a nivel de bloque son un paso intermedio entre los dos anteriores, donde los bloques más grandes incorporan más información espacial mientras que los más pequeños permiten obtener una mayor sensibilidad [6].

1.3. Objetivos

En el presente trabajo se trata de desarrollar un sistema de detección automática de sangre que permita reducir el tiempo de revisión de las imágenes de WCE comprometiendo en la menor medida posible la calidad del diagnóstico, siguiendo para ello dos enfoques distintos. El primero consiste en extraer características que permitan distinguir la sangre del tejido sano y entrenar con ellas algoritmos de aprendizaje automático. El segundo enfoque consiste en utilizar técnicas de aprendizaje profundo, concretamente redes neuronales convolucionales, capaces de extraer las características necesarias de las imágenes por sí solas. Por último, se compararán los resultados obtenidos con cada procedimiento.

2. Materiales

La base de datos disponible consiste en 75 imágenes de WCE obtenidas utilizando la cápsula endoscópica PillCam™SB3, proporcionadas por la Unidad de Endoscopia Digestiva y el Grupo de Investigación de Endoscopia Digestiva del Hospital Universitari i Politècnic La Fe (Valencia). De estas 75 imágenes, 40 contienen sangre y las 35 restantes son totalmente sanas.

3. Métodos

3.1. Preparación de la base de datos

En primer lugar, se genera manualmente el *groundtruth* de las imágenes, que es validado por un experto (Figura 2).

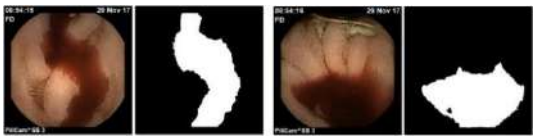

Figura 2. *Ejemplos de groundtruth de las imágenes.*

El siguiente paso consiste en extraer bloques cuadrados de distintos tamaños (32x32, 64x64 y 96x96 píxeles) de cada una de las imágenes mediante una ventana deslizante con un solapamiento del 50% tanto en la dirección vertical como en la horizontal. El etiquetado de los bloques se realiza al mismo tiempo, considerándose como sangre aquellos cuya área correspondiente en el *groundtruth* está conformada en más de un 10% por sangre.

Posteriormente, para garantizar la capacidad de generalización de los modelos se reparten los bloques en conjuntos de entrenamiento y de test, garantizando que todos los bloques de una misma imagen pertenezcan al mismo conjunto. Con el fin de reducir la dependencia de los resultados de la división realizada, se utiliza la técnica conocida como *nested cross-validation* [7]. En resumen, esta técnica consiste en entrenar varios modelos, cada uno con su correspondiente conjunto de test, asegurando que cada imagen participa una vez en un conjunto de test. Los parámetros de cada modelo son optimizados mediante una validación cruzada de 10 iteraciones en un bucle interno. Al final, se promedian los resultados sobre test de cada modelo para obtener un valor más general. Cabe destacar también que los grupos de entrenamiento son balanceados mediante la eliminación de bloques de la clase mayoritaria.

3.2. Selección de características

En primer lugar, se consideran las características relacionadas con las texturas. Sin embargo, guiándose por la literatura y tras comprobar que los histogramas LBP y HOG son prácticamente idénticos en ambas clases, se decide utilizar únicamente características de color. Concretamente, se utilizan los espacios de color RGB y HSV. Además, con el fin de corregir el problema que suponen las inhomogeneidades en la iluminación, se transforman las imágenes originales al espacio HSV, se aplica un filtro homomórfico sobre el canal V (que contiene toda la información de la iluminación) y se devuelven las imágenes al espacio RGB (Figura 3). Se entrenarán distintos modelos con cada uno de los siguientes conjuntos de características:

- Histogramas de los canales R, G y B.
- Histogramas de los canales H, S y V.
- Histogramas de los canales H y S.
- Histogramas de los canales R, G y B tras aplicar un filtro homomórfico sobre el canal V en HSV.

Figura 3. *Ejemplos de imágenes de WCE filtradas.*

3.3. Selección y entrenamiento de los algoritmos

Tras estudiar varios de los algoritmos de aprendizaje automático clásico más utilizados actualmente, se escogen los algoritmos Support Vector Machine (SVM) [8] y Random Forest [9], debido a su gran poder predictor, su relativa sencillez y su resistencia al sobreentrenamiento.

En el caso de los modelos SVM se decide comparar el rendimiento logrado con un kernel lineal, de gran sencillez, y con un kernel RBF, cuyo rendimiento suele sobrepasar al del resto y es también sencillo de implementar.

En cuanto a los modelos Random Forest, se estima el número óptimo de árboles observando que la curva del error de clasificación (concretamente, el *out-of-bag error*) se mantiene prácticamente constante a partir de 80 árboles.

3.4. Implementación del método waterpixels

Hasta ahora, el enfoque seguido se ha basado en dividir las imágenes en bloques cuadrados y extraer características de ellos. Sin embargo, la ventana deslizante utilizada divide la imagen en bloques sin tener en cuenta su contenido, mezclando ambas clases. Con el fin de superar este problema, se utiliza la técnica de segmentación en superpíxeles, regiones conexas cuyos píxeles comparten características similares y cuyas fronteras tienden a ajustarse a los contornos de los objetos de la imagen. Concretamente, se utiliza una variante de los métodos de generación de superpíxeles basada en la técnica *watershed*, denominada ***waterpixels*** [10], que ha demostrado proporcionar mejores resultados que otros métodos [9]. Uno de los pasos de la técnica *waterpixels* que presente la mayor oportunidad de mejora es la elección del marcador a partir del cual se genera cada uno de los waterpíxeles. En este trabajo se seleccionan como marcadores los puntos de mínimo gradiente. En caso de existir varios mínimos dentro de una misma celda (área reservada inicialmente para cada waterpíxel), se selecciona aquel que presenta el máximo coeficiente de extinción de superficie [11].

3.5. Entrenamiento de las redes neuronales convolucionales

Por último, se aborda el enfoque basado en las técnicas de aprendizaje profundo, capaces de extraer las características necesarias de las imágenes por sí solas. Se entrenan, de la misma forma que en los casos anteriores, cinco redes neuronales convolucionales (CNN) para cada tamaño de bloque. En este caso, el subconjunto de validación, constituido por el 20% del conjunto de entrenamiento, será fijo en el entrenamiento de cada modelo. Las CNN son entrenadas a partir de otra red pre-entrenada con cientos de miles de ejemplos, la red VGG19 [12], mediante el proceso conocido como "*fine tuning*". Esta técnica consiste en bloquear las primeras capas de la red ya entrenada, capaces de extraer los rasgos más básicos de las imágenes, y permitir únicamente cambios en los pesos de las últimas capas para adecuar la red al nuevo problema.

4. Resultados

Se presentan en este apartado los resultados obtenidos por los distintos modelos entrenados con cada tamaño de bloque, además de aquellos entrenados con superpíxeles.

En cuanto a los modelos SVM, se muestran únicamente los resultados obtenidos con el kernel RBF en la Tabla 1, pues ha demostrado ser muy superior al kernel lineal. Por otro lado, se muestran en la Tabla 2 los resultados obtenidos con los modelos Random Forest y en la Tabla 3 aquellos correspondientes a las redes neuronales convolucionales.

Regiones procesadas	Características	Exactitud	Sens.	Espec.	AUC
Bloq. 32x32	RGB	0,9221	0,9137	0,9279	0,9685
	HSV	0,9051	0,9090	0,9030	0,9589
	HS	0,9175	0,9260	0,9128	0,9718
	RGB filtrado	0,9236	0,9181	0,9274	0,9709
Bloq. 64x64	RGB	0,9045	0,8587	0,9353	0,9566
	HSV	0,8968	0,8931	0,8982	0,9571
	HS	0,9132	0,9162	0,9104	0,9694
	RGB filtrado	0,9148	0,8880	0,9326	0,9689
Bloq. 96x96	RGB	0,8497	0,7954	0,8906	0,9303
	HSV	0,8817	0,9026	0,8658	0,9549
	HS	0,8957	0,9034	0,8894	0,9637
	RGB filtrado	0,8991	0,8524	0,9339	0,9640
Superpíxeles	RGB	0,9257	0,9485	0,9132	0,9752
	HSV	0,9086	0,9526	0,8851	0,9734
	HS	0,9075	0,9479	0,8845	0,9668
	RGB filtrado	**0,9325**	**0,9500**	**0,9258**	**0,9769**

Tabla 1. *Resultados promediados obtenidos por los modelos SVM con un kernel RBF sobre los conjuntos de test.*

Regiones procesadas	Características	Exactitud	Sens.	Espec.	AUC
Bloq. 32x32	RGB	0,9266	0,9206	0,9313	0,9756
	HSV	0,9252	0,9363	0,9198	0,9806
	HS	0,9238	0,9372	0,9169	0,9802
	RGB corregido	0,9292	0,9342	0,9275	0,9787
Bloq. 64x64	RGB	0,9276	0,9189	0,9343	0,9769
	HSV	0,9318	0,9483	0,9210	0,9828
	HS	0,9316	0,9443	0,9231	0,9822
	RGB corregido	0,9365	0,9357	0,9386	0,9823
Bloq. 96x96	RGB	0,9273	0,9225	0,9317	0,9760
	HSV	**0,9378**	**0,9568**	**0,9233**	**0,9844**
	HS	0,9357	0,9549	0,9209	0,9841
	RGB corregido	0,9389	0,9364	0,9424	0,9849
Superpíxeles	RGB	0,9058	0,9448	0,8835	0,9701
	HSV	0,8987	0,9580	0,8672	0,9781
	HS	0,8966	0,9612	0,8621	0,9777
	RGB corregido	0,9073	0,9548	0,8873	0,9770

Tabla 2. *Resultados promediados obtenidos utilizando los modelos Random Forest sobre los conjuntos de test.*

Bloques	Exactitud	Sensibilidad	Especificidad	AUC
32x32 píxeles	**0,9496**	**0,9390**	**0,9560**	**0,9901**
64x64 píxeles	0,9494	0,9208	0,9688	0,9917
96x96 píxeles	0,8943	0,7545	0,9838	0,8920

Tabla 3. *Resultados sobre los conjuntos de test promediados obtenidos utilizando las redes neuronales convolucionales.*

5. Discusión de los resultados

Debido a la naturaleza del problema abordado, donde el error de clasificar como sana una imagen patológica tiene consecuencias más graves que el error contrario, que tan solo alargaría el tiempo de revisión, se le da una mayor importancia a la sensibilidad que a la especificidad.

De los resultados se extrae que, en general, los histogramas de los canales RGB extraídos directamente de la imagen original parecen proporcionar los peores resultados. Sin embargo, al extraerlos tras aplicar un filtro homomórfico los resultados se aproximan a los obtenidos con el espacio HSV. También se observa que los resultados de los modelos entrenados con los tres canales del espacio HSV son muy similares a los obtenidos al omitir el canal V. Esto podría deberse a que, aunque el canal V contiene

información útil, también incluye el problema de las diferencias en la iluminación, por lo que en ocasiones su uso podría llegar a empeorar el modelo.

En cuanto a los diferentes tamaños de bloque, se comprueba que generalmente el uso de bloques más grandes afecta negativamente a la sensibilidad. Sin embargo, los modelos Random Forest parecen ser más insensibles al tamaño de bloque, pues tan solo se aprecian cambios del orden de las centésimas, probablemente debidos a la distinta selección de imágenes en cada caso.

En cuanto al uso de la técnica *waterpixels*, se comprueba que mejora los resultados en el caso de los SVM, especialmente en términos de sensibilidad. Sin embargo, en Random Forest, los resultados obtenidos al utilizar bloques son ligeramente superiores.

En la Figura 4, donde se comparan los resultados de los mejores algoritmos de cada tipo, puede observarse que, aunque globalmente las CNN parecen mejores clasificadores, ofrecen la menor sensibilidad.

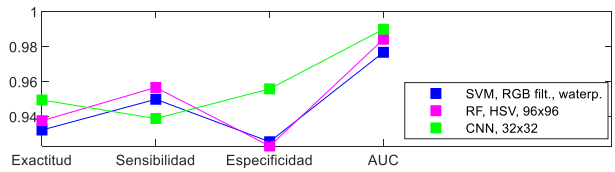

Figura 4. *Comparación de los mejores modelos de cada tipo.*

La Figura 5 muestra los resultados obtenidos con algunas imágenes de test, que no han participado en el proceso de aprendizaje. Cada imagen se divide en bloques (fila inferior) y waterpíxeles (fila superior), que se clasifican con los modelos correspondientes y se combinan para obtener una clasificación a nivel de pixel. En la misma figura puede comprobarse que, al utilizar *waterpixels*, la región marcada como sangre es mucho más precisa.

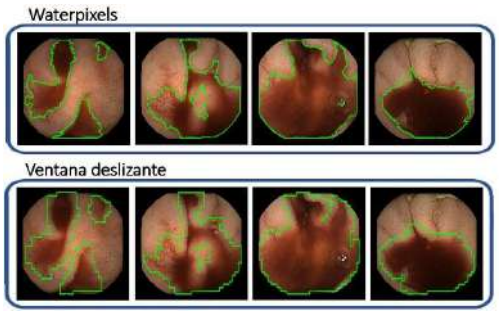

Figura 5. *Comparación de las regiones clasificadas como sangre utilizando waterpixels y bloques cuadrados.*

Una de las mayores limitaciones encontradas durante la realización del trabajo ha sido la falta de disponibilidad de una mayor base de datos. La segmentación a nivel de píxel realizada con *waterpixels* podría aprovecharse como herramienta de etiquetado supervisado con la que generar un mayor *groundtruth* para entrenar y probar los modelos.

6. Conclusiones

Se han entrenado modelos SVM, Random Forest y redes neuronales convolucionales capaces de detectar la sangre automáticamente en las imágenes de WCE, con el fin de desarrollar una herramienta que sirva de apoyo para el médico. Las características de color cuyo uso ha proporcionado los mejores resultados son los histogramas de los canales H y S del espacio de color HSV y los de los tres canales del espacio RGB tras haber aplicado un filtro homomórfico que corrija las variaciones en la iluminación.

El modelo que ha obtenido la mayor sensibilidad es un Random Forest entrenado a partir del espacio HSV, con un 95,68% de sensibilidad y un 92,33% de especificidad (sobre las imágenes de test). Por otro lado, los resultados al utilizar *waterpixels* en la clasificación de muestras de test son más precisos a nivel de pixel, siendo el mejor modelo obtenido que utiliza esta técnica un SVM entrenado a partir del espacio RGB tras haber aplicado el filtro corrector, con un 95% de sensibilidad y un 92,58% de especificidad.

Referencias

[1] Signorelli, C y col. (2005). “Sensitivity and specificity of the suspected blood identification system in video capsule enteroscopy”. En: *Endoscopy* 37.12, págs. 1170-1173 (vid. pág. 8).

[2] Buscaglia, Jonathan M y col. (2008). “Performance characteristics of the suspected blood indicator feature in capsule endoscopy according to indication for study”. En: *Clinical gastroenterology and hepatology* 6.3, págs. 298-301 (vid. pág. 8).

[3] Francis, RD y col. (2004). “Sensitivity and specificity of the red blood identification (RBIS) in video capsule endoscopy”. En: *3Rd, Int. Conf. on Capsule Endoscopy* (vid. pág. 8).

[4] Li, Baopu y Max Q-H Meng (2009a). “Computer-aided detection of bleeding regions for capsule endoscopy images”. En: *IEEE Transactions on biomedical engineering* 56.4, págs. 1032-1039 (vid. pág. 9).

[5] Berens, Jeff, Graham D Finlayson y Guoping Qiu (2000). “Image indexing using compressed color histograms”. En: *IEE Proceedings-Vision, Image and Signal Processing* 147.4, págs. 349-355 (vid. pág. 10).

[6] Novozámsk, Adam y col. (2016). “Automatic blood detection in capsule endoscopy video”. En: *Journal of biomedical optics* 21.12, pág. 126007 (vid. pág. 11).

[7] Varma, Sudhir y Richard Simon (2006). “Bias in error estimation when using cross-validation for model selection”. En:BMC bioinformatics7.1, pág. 91 (vid. pág. 20).

[8] Cortes, Corinna y Vladimir Vapnik (1995). “Support-vector networks”. En: Machine learning 20.3, págs. 273-297 (vid. pág. 34).

[9] Breiman, Leo (2001). “Random forests”. En: Machine learning 45.1, págs. 5-32 (vid. pág. 38).

[10] Machairas, Vaıa y col. (2015). “Waterpixels”. En: *IEEE Transactions on Image Processing* 24.11, págs. 3707-3716 (vid. págs. 28, 31).

[11] Vachier, Corinne y Fernand Meyer (1995). “Extinction value: a new measurement of persistence”. En: IEEE Workshop on nonlinear signal and image processing. Vol. 1, págs. 254-257 (vid. pág. 31).

[12] Simonyan, Karen y Andrew Zisserman (2014). “Very Deep convolutional networks for large-scale image recognition”. En: *arXiv preprint arXiv:1409.1556* (vid. pág. 46).