Detección del espacio glotal en imágenes laríngeas mediante transformada Watershed y Merging JND

Víctor Osma Ruiz, Nicolás Sáenz Lechón, Juan I. Godino Llorente, Rubén Fraile.

Dpt. De Ingeniería de Circuitos y Sistemas, Universidad Politécnica de Madrid, Ctra. Valencia, Km. 7, 28031, Madrid, España, (vosma@ics.upm.es, nicolas.saenz@upm.es, igodino@ics.upm.es, rfraile@ics.upm.es)

Resumen

El presente artículo describe un nuevo método para la detección del espacio glotal en imágenes laríngeas obtenidas de vídeos de alta o baja velocidad. El proceso de detección basa su eficacia en la combinación de varias técnicas de gran relevancia en el campo del tratamiento digital de imágenes. Una de estas técnicas es la transformada Watershed que junto con varios tipos de Merging y un proceso final de predicción lineal, hacen posible la detección automática en un 99% de las imágenes analizadas. La potencia del método se ve incrementada por la ausencia de cualquier tipo de inicialización y por no necesitar condiciones estrictas sobre las características de las imágenes a procesar. Evidentemente es importante que el algoritmo integre información a priori del espacio glotal, pero este conocimiento es bastante relajado comparado con las condiciones impuestas por otros trabajos que también intentan la segmentación.

1. Introducción

La forma de vida actual ha producido un aumento de las patologías vocales haciendo cada vez más necesario un buen diagnóstico de las mismas. Esto ha contribuido al desarrollo de mecanismos de visualización y herramientas software orientados a facilitar y hacer más preciso el trabajo del profesional médico. Uno de los métodos más útiles para la detección de patologías es la observación directa de la vibración de las cuerdas vocales durante la fonación.

Existen dos técnicas fundamentales que permiten registrar el movimiento vibratorio de las cuerdas vocales [1]: grabaciones de alta velocidad y grabaciones de baja velocidad con luz estroboscópica (en adelante grabaciones estroboscópicas o vídeos estroboscópicos).

Las primeras permiten almacenar fotogramas de las cuerdas vocales durante la fonación a gran velocidad (2000 fotogramas por segundo).

El segundo sistema usa velocidades de grabación mucho menores, entorno a 25 fotogramas por segundo, velocidad claramente insuficiente para registrar un ciclo de vibración ya que las cuerdas se mueven a un ritmo de 50 a 300 Hz dependiendo de la persona y su sexo. La solución pasa por realizar un submuestreo haciendo uso de cortos destellos luminosos. De esta forma se consigue una ilusión óptica que permite la visualización del ciclo vibratorio de las cuerdas [1]. No obstante, la difícil sincronización entre el destello luminoso, la frecuencia de vibración de las cuerdas y la velocidad de grabación del vídeo introduce errores en las imágenes capturadas mediante esta técnica de las que están exentas las imágenes tomadas con sistemas de alta velocidad. Los principales problemas son: cambios de iluminación, borrosidad y falta de continuidad en la secuencia de imágenes [1].

Existen otros problemas que afectan a ambos tipos de sistemas: rotación de la cámara, movimientos laterales de la cámara y/o del paciente que producen la deslocalización de las cuerdas vocales. Si bien, los problemas son más graves en las grabaciones estroboscópicas porque para registrar el mismo número de ciclos de vibración es necesario un tiempo de grabación mucho más grande.

Por si todo ello fuera poco, las fuentes de iluminación son diferentes de un sistema a otro lo que introduce una variabilidad de la iluminación inter-vídeo. En grabaciones estroboscópicas el sistema de iluminación produce además una gran variación intra-vídeo.

La detección del espacio glotal (o glotis [1]) en imágenes laríngeas no es por tanto una tarea sencilla y sin embargo resulta una operación fundamental para el cálculo de numerosos parámetros de fonación ya sea directamente o a través de alguna representación (forma de onda glotal [2], perfiles de vibración [3], quimogramas [4], relación de amplitudes de vibración, relación de periodos de vibración, medidas en fases de apertura y cierre, etc...[5]) y también supone, en ocasiones, un paso previo para la segmentación de las cuerdas vocales [3].

Todas estas utilidades han hecho necesario el desarrollo de numerosas técnicas de tratamiento digital de imagen orientadas a la segmentación de la glotis de forma más o menos automática, desde las técnicas basadas en procesado clásico de imágenes (umbralización, filtrado, operaciones morfológicas, etc...) [6] hasta los modernos contornos activos, *snakes* [5;7] y los *balloon models* [3] pasando por las conocidas técnicas de crecimiento de región [2;4].

El principal problema de todas estas técnicas es que resultan muy dependientes del punto de inicialización del proceso de segmentación, además de ser altamente sensibles al ruido. Existen trabajos como el de Palm [3] que con una serie de variaciones sobre los métodos de *snakes* consigue mejorar el comportamiento ante el ruido además de independizarse en cierta medida de la

inicialización, pero a cambio se ve obligado a establecer una señal de parada realmente fuerte al buscar la glotis como un objeto oscuro y centrado en la imagen, cosa que no ocurre en muchas imágenes estroboscópicas, debido a los movimientos y problemas de iluminación anteriormente descritos. En [2] se busca el punto inicial mediante una técnica de umbralización avanzada basada en el histograma de la imagen. Este método da muy buenos resultados con los vídeos de alta velocidad analizados en el artículo. Sin embargo cuando los vídeos son estroboscópicos es difícil distinguir mediante umbralización la glotis de otras zonas oscuras con un nivel de gris semejante. En [7] la inicialización se consigue analizando previamente las diferencias existentes entre fotogramas correlativos para detectar zonas de movimiento (*Motion Energy Images*). Como el anterior, el método es difícil de aplicar en vídeos estroboscópicos donde el movimiento no sólo depende de la vibración de las cuerdas vocales sino de los ya citados movimientos de la cámara y/o el paciente.

El nuevo método que se propone en este artículo no necesita ningún tipo de inicialización y el único conocimiento a priori que integra es realmente generalizable a cualquier imagen de las cuerdas vocales, al basarse en características que siempre debe cumplir la glotis para que un observador humano pudiera reconocerla. La señal de parada del proceso de detección aprovecha también las características de la visión humana para distinguir niveles de gris (JND – *Just Noticeable Difference*). Estos tres aspectos hacen del sistema una herramienta útil para la detección de la glotis en cualquier tipo de imágenes ya provengan de vídeos de alta velocidad con una buena calidad o de vídeos estroboscópicos de baja velocidad con los problemas inherentes a su técnica.

2. Transformada Watershed y Merging

La transformada *Watershed* es una de las herramientas más valoradas en el campo de la segmentación digital de imágenes [8]. Su éxito radica en las ventajas que introduce esta herramienta para dividir la imagen en regiones claramente delimitadas, frente a otros métodos que devuelven una segmentación de líneas inconexas que permiten la diferenciación de objetos a simple vista pero resultan difíciles de manejar para un sistema automático.

El concepto de *watersheds* proviene del mundo de la topografía y se refiere a la división de un terreno en sus cuencas de recepción de agua. Cuando llueve sobre ese terreno las gotas de agua irán descendiendo desde el punto donde caen hasta el lugar más profundo posible siguiendo el camino que tenga un mayor desnivel.

Para llevar este concepto al mundo del procesado digital de imágenes, se entienden las imágenes como superficies donde el nivel de gris representa la altura de cada píxel, considerándose el blanco (nivel 255) como la máxima altura y el negro (nivel 0) como la mínima. Después se simula una lluvia sobre la imagen de forma que cuando una gota de agua cae sobre un píxel esta bajaría buscando el camino de máximo desnivel hasta el punto mínimo más cercano. Todos los píxeles que dirigen sus gotas hacia el mismo mínimo formarán parte de la misma cuenca de recepción y serán etiquetados con el mismo identificador [8]. Una implementación muy eficiente de este método se presenta en [9].

El objetivo es que cada cuenca de recepción represente un objeto en la imagen, sin embargo la mayor parte de las veces el resultado es frustrante ya que aparecen miles de cuencas donde sólo se esperaban unas pocas. Este problema se denomina sobresegmentación y es debido principalmente al ruido en la imagen [8].

Una solución a este problema pasa por preprocesar la imagen. En este sentido una práctica muy extendida es la umbralización del gradiente [10]. En este caso la transformada *Watershed* se aplica sobre el gradiente de la imagen que es donde más sentido tiene ya que éste posee los máximos justo en los bordes de los objetos. Con esta técnica se consigue eliminar de la imagen a procesar bordes insignificantes carentes de información.

Sin embargo, esto por si solo no resuelve completamente el problema y hay que buscar un post-procesado del resultado. La mejor solución consiste en realizar una unión de las diferentes cuencas siguiendo distintos criterios descritos en la literatura [8;11]. Esta unión de cuencas es lo que denominamos *Merging*.

En líneas generales todos estos métodos se basan en la realización continua de iteraciones sobre la transformada *Watershed*. En cada iteración la herramienta calcula cuáles son las dos cuencas vecinas que pueden unirse con un menor coste y las fusiona. El proceso finalizará cuando sólo nos queden en la imagen el número de cuencas deseado (idealmente una por objeto) o cuando el menor coste de unión supere un umbral establecido. La definición de la función coste es lo que esencialmente diferencia unos métodos de otros.

En nuestro caso vamos a emplear un proceso de *Merging* como el definido teóricamente en [11] donde la función de coste se calcula según el JND de los distintos niveles de gris de la imagen. El JND representa la sensibilidad del sistema visual humano a los cambios de luminancia. El sistema visual humano no es capaz de diferenciar determinados cambios de luminancia, como ejemplo y representando la luminancia en términos de nivel de gris de una imagen no seríamos capaces de distinguir entre un nivel 80 y un nivel 85. Además, ésta insensibilidad no sigue una pauta lineal, siendo el ojo menos sensible a los cambios de luminancia en niveles oscuros que en claros.

La ecuación 1 permite calcular esta sensibilidad T(x,y) en función del nivel de gris I(x,y) considerado.

$$T(x,y)=\begin{cases} 17\cdot\left(1-\sqrt{\frac{I(x,y)}{127}}\right)+3 & \text{Si I(x, y)} \le 127 \\ \frac{3}{128}\cdot\left(I(x,y)-127\right)+3 & \text{Resto} \end{cases}$$

Ecuación 1. *Fórmula para el cálculo del JND.*

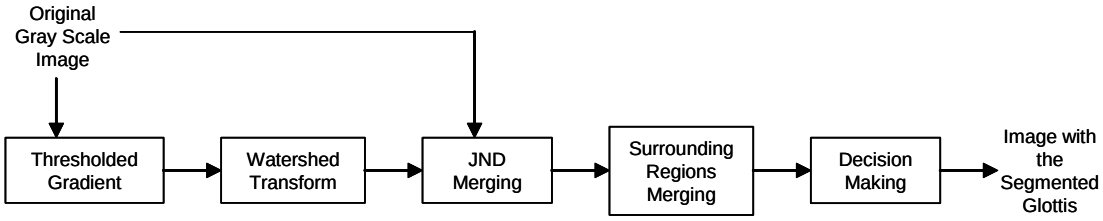

Figura 1. *Esquema del proceso seguido para la segmentación de la glotis.*

3. Método de segmentación y resultados

El método seguido para lograr la segmentación de la glotis es el que se presenta de forma esquematizada en la figura 1.

En el primer paso convertimos la imagen original en color (RGB) a escala de grises mediante una transformación al modelo YIQ del que tomamos la luminancia "Y".

Posteriormente obtenemos la transformada *Watershed* del gradiente de la imagen, previamente umbralizado con un valor de 2. Es decir, aquellos píxeles del gradiente que no superan el valor 2 son colocados a cero y por tanto son convertidos en mínimos que sólo podrán pertenecer al interior de una cuenca. De esta forma se consigue reducir el número inicial de estas en aproximadamente un 20% eliminando cuencas insignificantes debidas a la presencia de ruido en la imagen. El umbral de gradiente se ha obtenido empíricamente tras analizar el comportamiento de todas las imágenes empleadas en las pruebas.

El segundo paso consiste en una operación de *Merging* basado en JND. La función de coste para la unión de cuencas se calcula según indica la expresión de la ecuación 2, donde *mR*i representa el valor medio de gris en cada una de las cuencas *Ri* y *LimitArea* representa un valor límite de área (número de píxeles de las cuencas) por debajo del cual se facilita la unión y por encima del cual se dificulta.

$$F_c = \left[\left|mR_1 - mR_2\right| - MinJND(mR_1, mR_2) + 255\right] \cdot \frac{MinArea(R_1, R_2)}{LimitArea}$$

Ecuación 2. *Función coste para Merging.*

El área límite se establece también empíricamente en un 0,5% del área total de la imagen. Esto permite que las regiones muy pequeñas disminuyan su función coste a pesar de que en un principio pudieran resultar distintas al ojo. Este hecho no rompe la filosofía inicial ya que al comienzo las cuencas, debido a la sobresegmentación inherente a las *watersheds*, tendrán áreas muy pequeñas y por tanto pueden existir pequeñas regiones que perteneciendo a grandes zonas con nivel de gris similar estén por encima del umbral de insensibilidad del ojo. Con esta modificación de la función coste permitimos que estas áreas se unan con sus vecinas más similares. En definitiva lo que se consigue es aislar aún más el proceso de segmentación de los problemas introducidos por el ruido y la pobre iluminación de las imágenes estroboscópicas. La idea es permitir la unión de cuencas mientras la función de coste entregue resultados inferiores a 255, ya que por debajo de ese nivel el sistema visual humano considera que los niveles medios de gris de las cuencas a unir son idénticos (no es capaz de distinguirlos).

La función JND utilizada en la ecuación 2 sigue en esencia la expresión indicada en la ecuación 1, pero hemos introducido una modificación que mejora significativamente los resultados. El umbral de visualización se coloca a 255 para valores de gris superiores al 90% del máximo valor del histograma de la imagen. Esto es lo mismo que decir que todas las regiones claras de la imagen son indistinguibles entre si. Con esta condición el número de objetos resultantes después del proceso de merging se ve drásticamente disminuido pero seguimos detectando perfectamente la glotis al tratarse esta de un objeto oscuro.

Tras el paso 2 la imagen nos quedará segmentada como muestra el ejemplo de la figura 2. Puede observarse como la glotis aparece perfectamente segmentada, además de otras zonas de la imagen con nivel de gris homogéneo para el sistema visual humano. Las zonas claras de la imagen aparecen todas unidas como si de un único objeto se tratara.

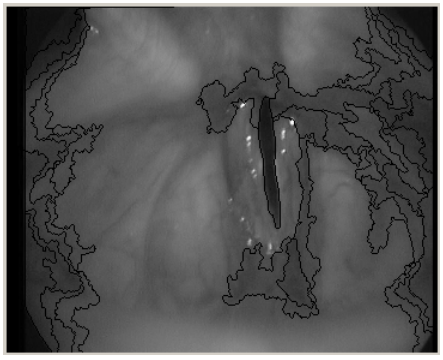

Figura 2. *Segmentación de la glotis tras el primer Merging.*

El tercer paso consiste en una segunda operación de *Merging* pero esta vez destinada a unir cuencas que posean alguna vecina con nivel medio de gris inferior al suyo propio. Ahora se trata de reducir el número de objetos segmentados uniendo regiones que no pueden ser glotis. Como siempre nos basamos en las decisiones que tomaría un observador humano y evidentemente este siempre esperaría que la glotis sea el objeto más oscuro de todos los que la rodean.

La división obtenida tras el tercer paso será como la mostrada en el ejemplo de la figura 3. El número de objetos presentes en la imagen es ya muy reducido.

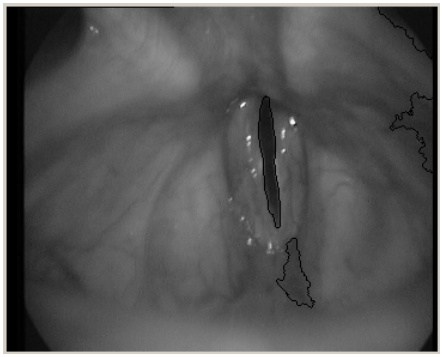

Figura 3. *Segmentación de la glotis tras el segundo Merging.*

El último paso está constituido por un proceso de clasificación mediante el que se pretende distinguir la glotis del resto de objetos presentes en la imagen. Los defectos laterales se eliminan por tener un nivel medio de gris demasiado bajo. El objeto de fondo se elimina porque es el que mayor área tiene de la imagen. Para distinguir entre el resto de objetos (sombras) y la glotis entrenamos un clasificador discriminante [12] basado en momentos invariantes. La idea es detectar la glotis principalmente por su forma. Para entrenar el predictor se han utilizado el 88% de las imágenes disponibles (98 fotogramas recogidos de 13 vídeos sobre un total de 15), lo que nos entrega 263 objetos sombra y 98 objetos glotis. El resto (13 fotogramas de los 2 vídeos restantes) se han dejado fuera para después validar los resultados.

La glotis se ha detectado con éxito en el 99% de las imágenes usadas para entrenamiento y en todas las imágenes de validación. Sólo en un 25% de las imágenes hubo que cambiar el umbral de coste, principalmente debido a grandes variaciones en el tamaño de la glotis.

La figura 4 recoge un ejemplo de la detección de la glotis incluso cuando ésta aparece deslocalizada y cortada.

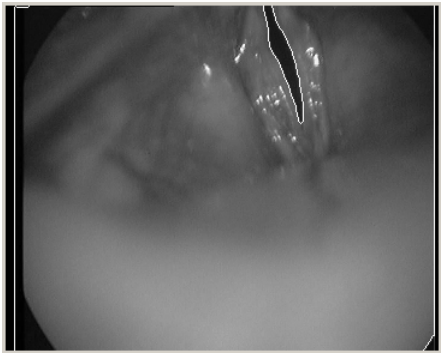

Figura 4. *Segmentación de una glotis deslocalizada y cortada.*

4. Conclusiones

Se ha presentado un método para la segmentación automática de la glotis que entrega buenos resultados incluso con imágenes estroboscópicas de iluminación pobre. El método se ha demostrado muy robusto ante los problemas de iluminación inter-vídeo e intra-vídeo.

El método no necesita ningún tipo de inicialización ni señal de parada estricta ya que se comporta como lo haría un observador humano por distinción de zonas con nivel de gris homogéneo.

Se ha logrado detectar la glotis en el 99% de las imágenes de entrenamiento y en todas las imágenes de validación.

Agradecimientos

Investigación realizada bajo el programa TEC 2006-12887-C02-02. Ministerio de Ciencia y Tecnología (España)

Referencias

[1] Baken, R. J. and Orlikoff, R. F., *Clinical measurement of speech and voice*, 2 ed., Singular, 2000.

[2] Yan, Y., Chen, X., and Bless, D., "Automatic tracing of vocal-fold motion from high-speed digital images," *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 7, pp. 1394-1400, July2006.

[3] Palm, C., Lehmann, T. M., Bredno, J., Neuschaefer-Rube, C., Klajman, S., and Spitzer, K., "Automated analysis of stroboscopic image sequences by vibration profiles," in *Proceedings of the 5th International Workshop on Advances in Quantitative Laryngology, Voice and Speech Research*, Groningen, Netherlands, Apr.2001.

[4] Wittenberg, T., Moser, M., Tigges, M., and Eysholdt, U., "Recording, processing, and analysis of digital high-speed sequences in glottography," *Machine Vision and Applications*, vol. 8, pp. 399-404, 1995.

[5] Manfredi, C., Bocchi, L., Bianchi, S., Migali, N., and Cantarella, G., "Objetive vocal fold vibration assessment from videokymographic images," *Biomedical signal processing and control*, vol. 1, no. 2, pp. 129-136, 2006.

[6] Sáenz-Lechón, N., Osma-Ruiz, V. J., and Godino-Llorente, J. I., "Kymogram synthesis from pre-recorded low speed video data," in *Proceedings of IEEE EMBS/BMES*, vol. 1, pp. 1088-1089, Oct.2002.

[7] Friedl, S. and Wittenberg, T., "Automatic segmentation of vocal folds using active shape models," in *Proceedings of the 6th International Workshop on Advances in Quantitative Laryngology, Voice and Speech Research*, Hamburg, Germany, Apr.2003.

[8] Bleau, A. and Leon, L. J., "Watershed-based segmentation and region merging," *Computer Vision and Image Understanding*, vol. 77, no. 3, pp. 317-370, Mar.2000.

[9] Osma-Ruiz, V. J., Godino-Llorente, J. I., Sáenz-Lechón, N., and Gómez-Vilda, P., "An improved watershed algorithm based on efficient computation of shortest paths," *Pattern Recognition*, vol. 40, no. 3, pp. 1078-1090, 2007.

[10] Hernandez, S. E. and Barner, K. E., "Joint region merging criteria for watershed-based image segmentation," in *Proceedings of IEEE ICIP 2000*, vol. 2, pp. 108-111, Sept.2000.

[11] Shen, D. F. and Huang, M. T., "A watershed-based image segmentation using JND property," in *Proceedings of IEEE ICASSP 2003*, vol. 3, pp. 377-380, Apr.2003.

[12] Duda, R. O., Hart, P. E., and Stork, D. G., *Pattern Classification*, 2 ed., Wiley-Interscience, 2001.