

HRIMAC - Una Herramienta de Recuperación de Imágenes Mamográficas por Análisis de Contenido para el Asesoramiento en el Diagnóstico del Cáncer de Mama.

J. Martí (a), J. Freixenet (a), D. Raba (a), A. Bosch (a), J. Pont (b), J. Español (b),
R. Bassaganyas (b), E. Golobardes (c), X. Canaleta (c).

(a) Instituto de Informática y Aplicaciones. Universidad de Girona.
Ed. P4 Campus de Montilivi, 17071 Girona.

(b) Hospital Universitario de Girona Doctor Josep Trueta.
Avda. de Francia s/n. 17007 Girona.

(c) Departamento de Informática. Universidad Ramon Llull.
Ingeniería y Arquitectura La Salle. Paseo Bonanova, 8 08022 Barcelona.

Resumen

En este trabajo presentamos HRIMAC, un proyecto concebido para funcionar como un sistema de recuperación de imágenes por contenido que permita acceder a una determinada tipología de imágenes de mamografías digitales almacenadas en las diversas bases de datos públicas, a partir del contenido de una imagen ejemplo, siguiendo determinados criterios de afinidad. Así, a partir de una imagen mamográfica sobre la cual se pretende emitir un diagnóstico, HRIMAC busca en la base de datos mamográfica las n mamografías más similares, de acuerdo con los criterios especificados en la búsqueda. De esta manera, cada búsqueda proporciona un conjunto limitado de casos (mamografías digitales) con ciertas características (la forma de los clusters de microcalcificaciones, la presencia de determinadas lesiones espiculares, la forma de las masas, etc.) muy similares a la mamografía que se somete a estudio. El análisis de estos casos, ya patológicos, puede sin lugar a dudas ayudar al radiólogo a diagnosticar con más garantías de éxito, y aumentar de esta forma el grado de eficacia en la interpretación.

Keywords: Sistema basados en Conocimiento, Análisis de imágenes

1. Introducción

La utilización de la técnica de la mamografía digital constituye uno de los métodos más efectivos para la detección precoz del cáncer de mama [4]. En los últimos años, se ha podido comprobar en la mayoría de países occidentales que la implantación de programas masivos de *screening* ha repercutido en un incremento de la tasa de supervivencia de las mujeres frente a esta enfermedad [15][18]. Sin embargo, el porcentaje de mujeres que sufren cáncer y que es detectable mediante métodos clínicos o mamográficos constituye únicamente un reducido porcentaje de las mujeres que tienen una mama con alguna abnormalidad [19] y que aún no es clínicamente visible [1][2]. Consecuentemente, se hace necesario el desarrollo de nuevas estrategias que permitan la identificación de forma prematura de aquellas pacientes sometidas a un riesgo elevado de padecer cáncer invasivo de mama, con las correspondientes ventajas terapéuticas y sociales que esto puede representar [7].

En este sentido, es de común acuerdo entre la comunidad científica, técnica y médica que la incorporación de sistemas de tipo CAD (*Computer-Aided Diagnosis*) en las etapas iniciales del diagnóstico y análisis de la enfermedad suelen traducirse en una mayor supervivencia de la población afectada por este tipo de neoplasia [12][13], de forma que la tendencia en la lucha contra el cáncer consiste en la paulatina incorporación de nuevas técnicas y métodos orientados a la detección precoz de los elementos propios de la sintomatología cancerígena.

1. Motivación

Una de las repercusiones más notorias de los últimos avances tecnológicos aplicados al ámbito médico ha sido sin duda alguna la implantación de sistemas informáticos en los Hospitales. Esta informatización ha comportado, entre otras consecuencias, la masiva digitalización de imágenes médicas hasta ahora registradas mediante el uso de metodologías basadas en film analógico. Con ello, desaparece paulatinamente el concepto clásico de archivo de casos (localizado en hospitales y centros médicos asociados a la procedencia de las pacientes) y se introduce la noción de base de datos de imágenes con las consiguientes ventajas asociadas: concepto de ubicuidad mediante telegestión de las bases, accesibilidad de los casos más relevantes para su aplicabilidad con fines pedagógicos y de diagnóstico, etc. Sin embargo, dado el ingente volumen de datos almacenado, la necesidad de disponer de herramientas eficientes de búsqueda de los casos de interés se ha convertido (o se convertirá) en esencial.

Un Sistema de Recuperación de Imágenes por Contenido (más conocido como CBIR, iniciales que corresponden a la definición inglesa *Content-Based Image Retrieval*) es un sistema que permite buscar imágenes de una base de datos a partir de información visual o de contenido [5]. Dada una imagen ejemplo (la imagen *query*), el objetivo del sistema CBIR consiste en buscar en la base de datos las n imágenes más similares a la imagen *query*. La propuesta más simple consiste en utilizar descriptores globales de la imagen, como por ejemplo caracterizar las imágenes a partir de calcular sus histogramas de nivel de gris [6]. A partir de ahí, el sistema CBIR recupera imágenes a partir de evaluar la similitud entre los histogramas de diferentes imágenes. Propuestas más elaboradas son aquellas que tienen en cuenta las características de color y textura de la imagen [17].

En lo que se refiere a imágenes médicas, estos tipos de caracterización no aportan hasta el momento resultados demasiado notables, ya que no tienen en cuenta la estructura de la imagen y en consecuencia, son incapaces de capturar la información relevante contenida en ella [14][16]. Como antecedentes más directos al proyecto HRIMAC, cabe destacar los sistemas ASSERT [3] y WebMIRS [8]. ASSERT es un CBIR para imágenes médicas desarrollado en la Purdue University (EE.UU.) que analiza imágenes tomográficas de alta resolución (*HRCT High Resolution Computed Tomography*) del pulmón, y que actualmente se está expandiendo para analizar imágenes CT del hígado y imágenes MRI de la rodilla. WebMIRS en cambio, es en realidad un sistema que recupera imágenes de rayos X de la columna vertebral a partir de descripciones textuales (queries basadas en texto), pero que indexa su base de datos (formada por 17.000 imágenes) a partir de la información del contenido de las imágenes (es decir, a partir del análisis mediante técnicas de Visión por Computador).

3. El proyecto HRIMAC

El diagnóstico médico suele llevarse a cabo mediante el reconocimiento visual de las anomalías que presenta externamente el paciente, o mediante el análisis de imágenes de rayos X, resonancia magnética, TAC, u otras técnicas. Partiendo de la hipótesis de que el tratamiento médico resulta más efectivo cuando se puede suministrar en etapas iniciales de la enfermedad, resulta de vital importancia disponer de herramientas efectivas que permitan buscar de forma precisa y rápida imágenes ejemplo que complementen el caso objeto de estudio, y facilitar así la emisión de un diagnóstico más fiable.

Con este objetivo hemos creado el proyecto HRIMAC, un sistema CBIR que permite la recuperación de imágenes de mamografías digitales contenidas en bases de datos, a partir del contenido de una imagen ejemplo, siguiendo determinados criterios de afinidad. Así, dada una imagen ejemplo (la de la mamografía a estudiar), HRIMAC busca en las diferentes bases de datos mamográficas las n mamografías más similares a la mamografía ejemplo, de acuerdo con una serie de criterios especificados en la búsqueda (Figura.1). De esta manera, cada búsqueda proporciona un conjunto limitado de casos (mamografías digitales) con ciertas características (la forma de los *clusters* de microcalcificaciones, la presencia de determinadas lesiones espiculares, la forma de las masas, etc.) muy similares a la mamografía que se somete a estudio, de la cual se pretende dictar un diagnóstico. Asimismo, en la búsqueda también intervienen rasgos destacables del propio paciente, la edad, antecedentes... El análisis de estos casos, ya patológicos, puede sin lugar a dudas ayudar al radiólogo a diagnosticar con más garantías de éxito, y aumentar de esta forma el grado de eficacia en la interpretación.

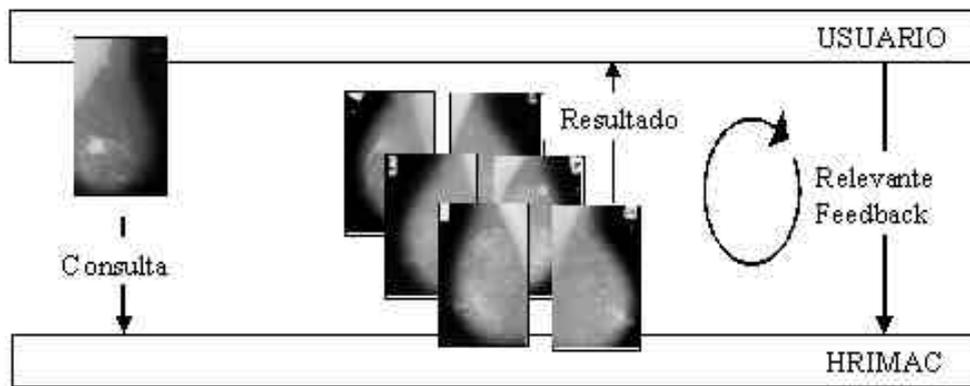


Figura 1: Proceso de consulta al sistema HRIMAC

El desarrollo de sistemas CAD (*Computer-Aided Diagnosis*) aplicados al ámbito médico ha sido un eje central en la investigación de los últimos años en disciplinas informáticas aplicadas, como son Visión por Computador o Inteligencia Artificial. No obstante, la mayoría de estos sistemas están diseñados para actuar como un segundo lector, con el recelo y las dudas que un sistema de estas características puede suscitar en los colectivos afectados (tanto médicos como pacientes). A diferencia de estos sistemas CAD mencionados, el proyecto HRIMAC no persigue como objetivo la emisión de un diagnóstico, sino que su principal utilidad para el especialista que lo use consiste en la obtención de un conjunto de casos muy parecidos al que en esos momentos intenta evaluar. HRIMAC constituirá, por tanto, una herramienta que pretende “ayudar al diagnóstico” y no diagnosticar.

Desde esta perspectiva, HRIMAC ofrece unas magníficas posibilidades de documentación de casos de especial relevancia para el caso que se está analizando en cada momento: el acceso inteligente a grandes bases de datos mamográficas de uso público (hoy en día: MIAS, Nijmegen, University of South Florida, -University of Washington, etc; pero en un futuro próximo la mayoría de hospitales-), en lugar de únicamente restringirse a la memoria de casos similares que el equipo de un determinado centro médico pueda gestionar de forma local. Obviamente, la búsqueda de información relevante en grandes bases de datos no es una tarea fácil, y muchas veces, si no se dispone de un buscador adecuado, la búsqueda se convierte en inviable. Es en este sentido que se dirige la propuesta de proyecto, consistente en la creación de una herramienta que permita buscar la información relevante de forma sencilla y eficiente

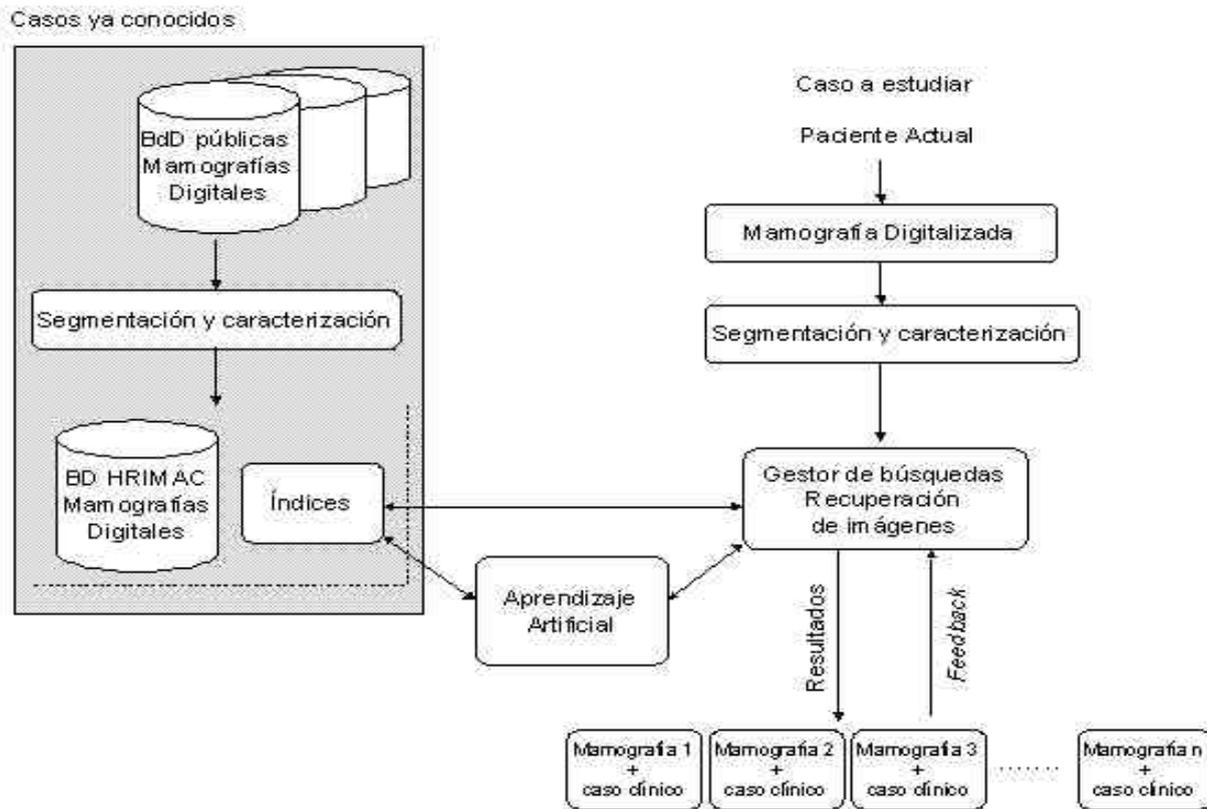


Figura 2: Esquema de la arquitectura de HRIMAC, un CBIR orientado a imágenes médicas.

Adicionalmente, HRIMAC también puede ser utilizado como una herramienta didáctica de ayuda a la formación de jóvenes radiólogos especialistas en la lectura de mamografías. HRIMAC ha de permitir la ejemplificación de casos prototipos de las diferentes sintomatologías registradas en el análisis de imágenes mamográficas. Además, en estos casos y cuando el usuario lo requiera, HRIMAC puede sugerir un diagnóstico, siguiendo la especificación BI-RADS^a, lo que completará sus utilidades como soporte de formación.

3.1. Arquitectura

La figura 2 muestra, a grandes rasgos, el diagrama de bloques de la arquitectura de HRIMAC. El proceso de búsqueda se inicia mediante una imagen de la mamografía a estudiar. Mediante un proceso de segmentación [11], se identificarán las zonas susceptibles a estudio y en una etapa de caracterización se clasificarán todas y cada una de las estructuras localizadas en la mamografía. La segmentación y caracterización [9] de esta imagen resultarán en un “vector de características” que describe la imagen que conforma el objeto de búsqueda. A partir de este punto debe llevarse a cabo un proceso de recuperación que consiste en buscar similitudes, utilizando las características de la imagen junto con los datos del paciente. Evidentemente, esta búsqueda no se puede realizar sobre todo el conjunto de imágenes de las bases de datos, ya que resultaría un proceso inviable en cuanto a tiempo de procesamiento; la búsqueda está dirigida mediante unos índices contruidos a partir del análisis (segmentación y caracterización [10]) de todas las imágenes de diferentes bases de datos públicas. Los índices agrupan las diferentes imágenes de las bases de datos según tipos, formas y clases de anomalías o patologías, de forma que permitan delimitar el área de búsqueda. Después de la consulta de índices, el gestor de la búsqueda iniciará el proceso de recuperación que consistirá en determinar el grado de similitud entre las características de la imagen *query* y las de las imágenes seleccionadas por los índices.

Este concepto de similitud es subjetivo para cada usuario, y es factible que ante una misma imagen *query* sometida por diferentes usuarios se busquen en realidad resultados distintos. Las técnicas de *relevance feedback* (o aprendizaje por reforzamiento) permitirán que el usuario guíe la búsqueda que se está realizando, indicando qué resultados han sido realmente relevantes. Ello permitirá a HRIMAC aprender (utilizando técnicas de inteligencia artificial) cuáles son las características importantes de la imagen *query* con la intención de refinar y particularizar los resultados de la búsqueda

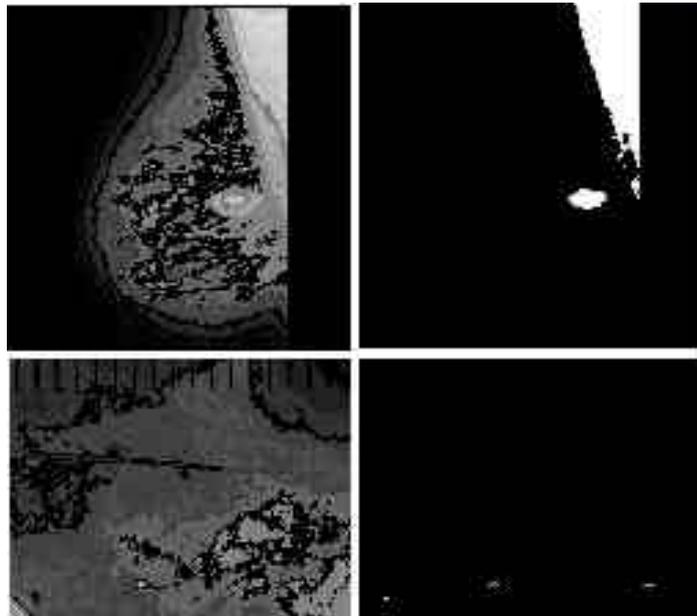


Figura 3: Proceso de segmentación de imagen. Localización de las áreas de interés i de las abnormalidades

En el momento de realizar la búsqueda, el funcionamiento interno del sistema es el siguiente: al proporcionar una imagen a analizar, el sistema realiza una primera búsqueda, mediante una función de similitud que tiene en cuenta datos de la imagen y del paciente. Para este proceso, el sistema desarrollado realiza una segmentación de la imagen (Figura.3) obteniendo así las posibles lesiones de esta. Cada lesión, es analizada extrayendo de ella características de forma y de textura. Junto con datos de carácter personal de los pacientes, nos permiten aplicar funciones de similitud entre los diferentes casos que conforman nuestra base de datos, y así poder determinar un grado de semejanza. Cabe destacar, que esta primera búsqueda es autónoma, no es necesaria la intervención del experto. Aun así, el conocimiento experto es útil en una segunda etapa de refinamiento, en la que se requiere de la actuación del usuario (radiólogo), marcando aquellas imágenes del conjunto resultado, que según su criterio, se adecuan a un resultado óptimo. Utilizando esta información, HRIMAC perfila la siguiente búsqueda, para proporcionar un mejor resultado.

Hay que remarcar que el proyecto se ha concebido dentro de la filosofía *Freeware*, utilizando software de libre distribución. La plataforma sobre la que actualmente el sistema trabaja consta de un servidor Linux. Se ha utilizado PostgreSQL para realizar el almacenamiento de los datos, debido a su robustez, y su compatibilidad con la gran mayoría de las herramientas de programación. En la fase actual de desarrollo, la base de datos almacena 30 casos de diferentes pacientes y un total de 330 imágenes

La GUI de HRIMAC se realiza via web, mediante cualquier navegador estándar existente, sea Explorer, Netscape, etc. aportando al sistema independencia al poder ser consultado desde cualquier punto de red en el mundo. Dicha ventaja constituye a su vez un requerimiento de seguridad que es tenido en consideración, mediante verificación de usuario y sistemas de encriptación de datos.

El entorno del proyecto, es una estructura web intuitiva, personalizada y de fácil uso (Figura.4 y 5). Permite una gestión de casos y gestión de bases de datos de imágenes. El proceso de búsqueda se ha concebido teniendo en cuenta el coste computacional de los procesos. Así pues, el usuario puede ordenar realizar búsquedas, y recibirá la notificación de final de proceso mediante la interfaz. Obviamente esto implica que los resultados se pueden consultar en otro momento.

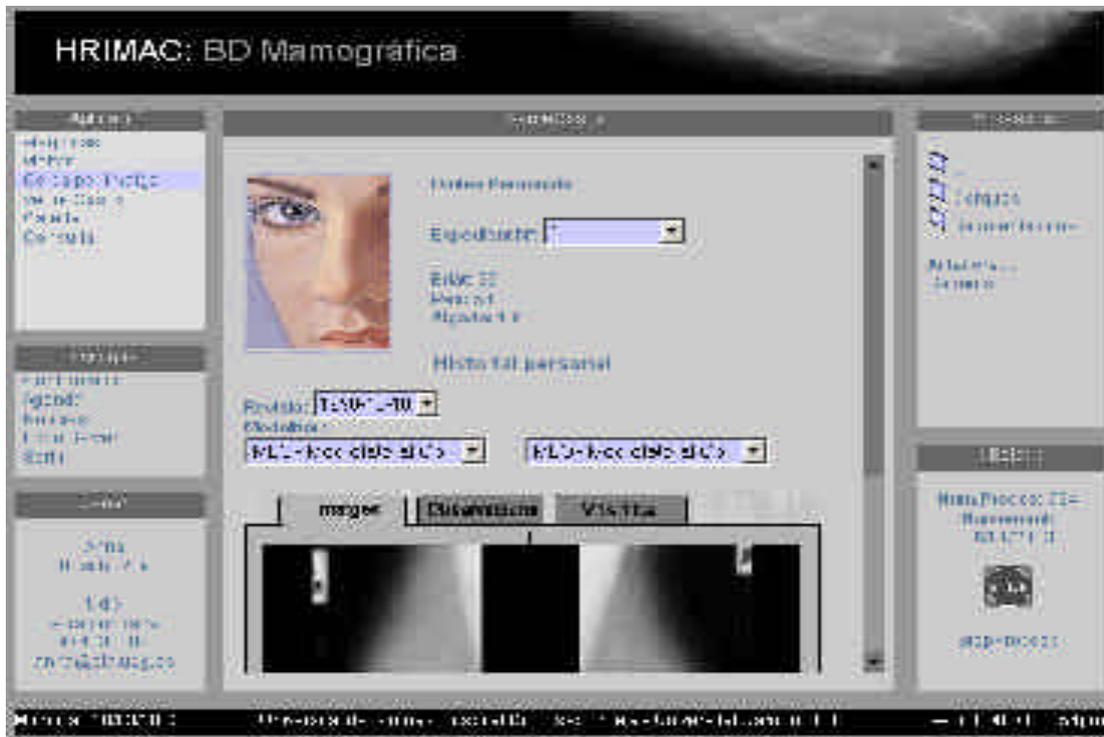


Figura 4: Interfaz web del sistema. Visualización del caso.

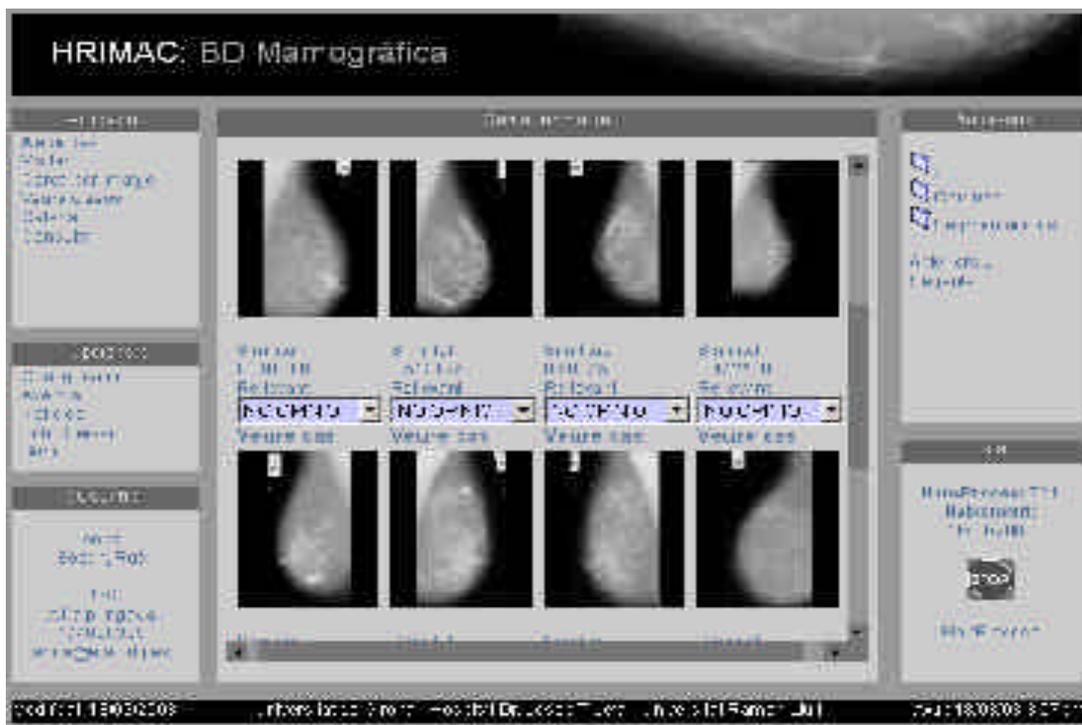


Figura 5: Interfaz web del sistema. Proceso de búsqueda.

5. Conclusiones y Trabajos futuros

En este artículo hemos presentado los primeros pasos realizados en el proyecto HRIMAC. La principal finalidad del proyecto es proporcionar un sistema informático sobre plataforma *web* que permita ayudar al colectivo de lectores mamográficos radiólogos a evaluar más eficazmente el diagnóstico de cáncer de mama a partir de imágenes de mamografías digitales.

En la actualidad se está trabajando en mejorar los algoritmos de segmentación no asistida, incorporando nuevas técnicas que aporten mejoras en los resultados.

Otro de los frentes de actuación, se centra en solventar el coste computacional de los procesos. Se ha diseñado el sistema para que pueda aprovechar el rendimiento que se obtiene mediante el multiproceso (clustering). En un futuro se planea instalar el sistema en un conjunto de máquinas que cooperen para realizar las búsquedas, y así incrementar la velocidad de proceso.

6. Agradecimientos

Este proyecto cuenta con el financiamiento del Ministerio de Ciencia y Tecnología (TIC 2002-04160-c02-01).

7. Bibliografía

- [1] Alberta Cancer Board, “Screen Test: Alberta Program for the Early Detection of Breast Cancer.” 1999/01 Biennial Report, 2001.
- [2] R. Highman, M. Brady, B. Shepstone. “A representation for mammographic image processing”, in *Medical Image Analysis*, Vol. 1, núm.1, pp.1-18, Marzo 1996.
- [3] C.Brodley, A.Kak, C.Shyu, J.Dy, L.Broderick, A.M.Aisen. “Content-Based Retrieval from Medical Image Databases: A Synergy of Human Interaction, Machine Learning and Computer Vision”.
- [4] C.J. D’Orsy. “Early detection of breast cancer: mammography”. *Breast-Cancer Res. Treat.* 18 Suppl. 1:S107-9. 1991.
- [5] E. Golobardes, X. Llorà, M. Salamó, J. Martí. Computed aided diagnosis with case-based reasoning and genetic algorithms. *Journal of Knowledge-Based Systems* 15, pp. 45-52, Elsevier Science, 2002.
- [6] P.M. Kelly, T.M.Cannon, D.R. Hush. “Query by image example: The CANDID approach”. In *SPIE Vol. 2420 Storage and retrieval for image and video databases III*, pp. 238-248, 1995.
- [7] J.L. Lamarque et al. “Digitalización en mamografía”. Edit. Díaz de Santos, pp. 201-208 (*Mastología Dinámica*). 1995.
- [8] Long LR, Pillemer SR, Lawrence RC, Goh G-H, Neve L, Thoma GR. “World Wide Web platform-independent access to biomedical text/image databases”. *Proc. SPIE Medical Imaging 1998: PACS Design and Evaluation: Engineering and Clinical Issues*, SPIE Vol. 3339, San Diego, CA, February 21-26, 1998; 52-63.
- [9] J. Martí, X. Cufí, J. Regincós, J. Español, J. Pont, C. Barceló. “Shape-based features characterization for microcalcification”, in *Proceedings of the Medical Imaging*, pp. 1215-1224, SPIE Vol. 3338, San Diego, California, 1998.

- [10] J. Martí, J. Español, E. Golobardes, J. Freixenet, R. García, M. Salamó. “Classification of microcalcifications in digital mammograms using Case-Based Reasoning”. 5th International Workshop on Digital Mammography, celebrado el Junio del 2000 en Toronto (Canadá).
- [11] J. Martí, P. Planiol, J. Freixenet, J. Español, E. Golobardes. Development of a Mammographic Analysis System Using Computer Vision Techniques. *Medical Data Analysis*, pp. 174-180, vol. 2199 Series Lecture Notes in Computer Science. Springer-Verlag Berlin Heidelberg, 2001. J.Crespo, V. Maojo, F. Martin Eds.
- [12] A.J.Méndez, P.G.Tahoces, et al. «Computer-aided diagnósticos: detection of masses on digital mammograms», in *Proceedings of the 3rd International Workshop on Image and Signal Processing*, pp. 465-468, Manchester, U.K., 1996.
- [13] R.M. Nishikawa, M.L. Giger, K. Doi, C.J. Vyborny, R.A. Schmidt. “Computer-aided detection and diagnósticos of masses and clustered microcalcifications from digital mammograms”, in A.G. Gale et al. editors, *Digital Mammography, Proceedings of the 2nd International Workshop on digital Mammography*, pp. 82-102, York, U.K. 1994.
- [14] Liang Shen, Rangaraj M. Rangayyan, and J.L. Leo Desautels. “Detection and classification of mammographic calcifications”, in K.W. Bowyer and S. Astley, editors, *State of the Art in Digital Mammographic Image Analysis*, pp. 198-212. World Scientific Publishing, 1994.
- [15] F. Shtern. “Digital mammography and related technologies: a perspective from the National Cancer Institute”. *Radiology*, 183 (3): pp. 629-630. 1992.
- [16] C.Shyu, C.Brodley, A.Kak, A.Kosaka, A.Aisen, L.Broderick. “ASSERT, a physician in the loop content-based image retrieval system for HRCT image databases”. In *Computer Vision and Image Understanding*, 2000.
- [17] N. Vujovic, D. Brzakovic, and K. Fogarty. “Detection of cancerous changes in mammograms using intensity and texture measures”, in *Proceedings of the Medical Imaging*, pp. 612-620, SPIE Vol. 2434, San Diego, California, 1995.
- [18] D. Winfields, M. Silbiger, G.S. Brown, L. Clarke, S. Dwyer, S. Yaffe, F. Shtern. “Technology transfer in digital mammography”. Report of the *Joint National Cancer Institute – National Aeronautics and Space Administration*. Workshop of May 19-20, Invest. Radiol. April, 29(4): pp. 507-515, 1994.
- [19] Ming Zhang and Maryellen L. Giger. “Automated detection of spiculated lesions and architectural distortions in digitized mamograms”, in *Proceedings of the Medical Imaging*, pp. 846-854, SPIE Vol. 2434, San Diego, California, 1995.