Congreso Internacional en Ingeniería Electrónica. Mem. Electro 2015, Vol. 37, pp 193-198, Chihuahua, Chih. México
http://depi.itchihuahua.edu.mx/display/memorias_electro/MemoriaElectro2015
ISSN 1405-2172

COMPARACIÓN Y PROPUESTA EN COMPRESIÓN HÍBRIDA EN MASTOGRAFÍAS DIGITALES "ELECTRO 2015"

Vilma Sánchez López, Inés Fernando Vega López,
Universidad Autónoma de Sinaloa (UAS)
Facultad de Informática Culiacán
Calle Josefa Ortiz de Domínguez s/n, Ciudad Universitaria, 80013 Culiacán Rosales, Culiacán SIN
Tel. 01 (667) 1 65 25 31
v.sanchez12@info.uas.edu.mx, ines.vegalopez@gmail.com

RESUMEN

El cáncer de mama es la principal causa de muerte en mujeres en todo el mundo, la detección temprana de este padecimiento aumenta la probabilidad de vida de quien la padece. La mastografía digital es a la fecha el estudio preventivo más confiable para la detección de dicho padecimiento, por esto la Asociación Americana del Cáncer recomienda realizar estudios de mastografías una vez por año en mujeres mayores de 40 años. Con el incremento de la realización de estudios de mastografías digitales la necesidad de trasmisión y almacenamiento de estas imágenes ha aumentado de igual manera. La compresión de imágenes es aplicada para reducir el tamaño de la representación de las imágenes y conservar la información relevante o necesaria. En este trabajo se propone un algoritmo de compresión para mastografías digitales. Este algoritmo segmenta la información de la mama y descarta el resto de la imagen. El segmento de interés (la mama) se comprime utilizando un algoritmo de compresión sin pérdida. Este tipo de compresión se le conoce como de tipo híbrida. Al comparar en promedio la escala de compresión de nuestra propuesta contra JPEG2000 y una propuesta algorítmica de tipo híbrida que utiliza el mismo método de segmentación, nuestra propuesta superó en escala de compresión a los demás algoritmos analizados.

Palabras Clave: mastografías, compresión sin pérdida, compresión con pérdida

ABSTRACT

Breast Cancer is the leading cause of death in women around the world. Early detection of this disease greatly improves life expectancy. Digital mammogram analysis is, at this time, the most effective preventive procedure for breast cancer detection. Because of this, the American Cancer Society recommends breast scanning once a year on woman of age 40 and older. With the rapid increment on the number of mammogram-based diagnosis, there has been a corresponding increase in the need for storing and transmitting digital mammogram files. Image compression is at the time, the best choice for reducing a digital image's file while preserving relevant information of the image, in this case, information required for a medical diagnostic. This work introduces a new algorithm for compressing digital mammograms' files. The proposed approach extracts a region of interest (ROI) and treats it separately from the background, which is treated as an information-less segment. This kind of compression is known as hybrid compression. Experimental evidence shows that our proposed algorithm is superior to JPEG2000, as well as to previously proposed hybrid algorithms that the use the same segmentation method.

Keywords: mammograms, lossless compression, lossy compression

1. INTRODUCCIÓN

El cáncer de mama es la principal causa de muerte en mujeres de todo el mundo. Según la Sociedad Americana del Cáncer, una de cada ocho mujeres tiene riesgo de padecer cáncer de mama a lo largo de su vida (12.2%) y una de cada 28 de morir por esta enfermedad [1]. Dicho padecimiento se origina con un tumor maligno en las células de la mama que luego invade a otras células sanas [2]. La posibilidad de una célula de crecer sin control e invadir otros tejidos, es lo que la hace cancerosa.

La mastografía digital es el estudio preventivo, no invasivo, más confiable para detectar cáncer de mama. Esta consiste en una imagen obtenida por un aparato especializado de rayos X. La Asociación Americana del Cáncer recomienda realizar mastografías anuales en las mujeres mayores de 40 años.

A medida que la cantidad de mastografías aumenta, el costo de almacenamiento y la necesidad de trasmisión de las imágenes aumenta de igual manera, sin embargo existe otro problema común al que se enfrentan las instituciones médicas, en ocasiones el médico oncólogo que realiza el diagnóstico no se encuentra en el mismo lugar donde se realizó el estudio. Esto implica trasmitir las imágenes a través de redes de computadora. El costo de esto, medido en tiempo, está en función del tamaño de las imágenes y de la velocidad de transmisión de la red utilizada. Conservar la información y reducir el tamaño de ésta, es el objetivo de los métodos de compresión de imágenes.

Los métodos de compresión de imágenes son a la fecha, la mejor opción para reducir el tamaño y conservar la información y por ello son ampliamente utilizados en las instituciones médicas. Existen dos tipos generales de métodos para la compresión de imágenes; Métodos de compresión con pérdida o irreversibles y métodos de compresión sin pérdida o reversibles.

Los métodos de compresión reversibles son los más utilizados en la comunidad médica y en la literatura científica son los mayormente estudiados. Sin embargo los métodos de compresión irreversibles proporcionan un grado más alto de

compresión en comparación de los métodos reversibles y no necesariamente se pierde información relevante para el diagnóstico [4].
En la búsqueda de una mayor compresión en las imágenes de mastografías digitales, en los últimos años han surgido propuestas de algoritmos híbridos. Estos métodos segmentan la imagen y comprimen la parte del fondo de forma diferente a la parte de la mama, que es la región de interés en este estudio. Segmentar significa identificar la parte en la mastografía que representa la mama del resto de la imagen. En la Sección 3 se describe el proceso de segmentación utilizado en el presente trabajo.

En el presente trabajo se propone un método que implica los dos tipos de compresión básicos antes mencionados. La parte que representa la mama es comprimida sin pérdida mientras el fondo de la mastografía se descarta. Además de realizar una propuesta algorítmica híbrida, dicha propuesta se compara contra los resultados de otro trabajo que utiliza el mismo método de segmentación [4]. Las dos propuestas híbridas son comparadas con el bien conocido algoritmo de compresión sin pérdida JPEG2000.

El resto de este trabajo se organiza de la siguiente forma. En el apartado 2 se da una introducción de los trabajos previos similares a nuestra propuesta. En el apartado 3 describimos nuestra propuesta algorítmica. Se describe detalladamente las dos piezas fundamentales del algoritmo (Compresor y descompresor). La explicación de la configuración experimental, los resultados y la comparación con los algoritmos es explicada en la sección 4, la sección de resultados. Por último se presentan las conclusiones del trabajo en la sección 5.

2. TRABAJO PREVIO

Han sido muchos los estudios que han experimentado con algoritmos de compresión con y sin pérdida en mastografías digitales. Otros estudios han utilizado combinaciones de estos. Recientemente ha surgido un nuevo paradigma que utiliza los dos tipos de compresión, a este tipo de algoritmos se les conoce como compresión híbrida.
En 2004, H. Chan et al. [5] realizaron un trabajo en compresión de mastografías digitales basado en contexto. Es decir se identificaban mediante segmentación fractal las estructuras relevantes para el diagnóstico y posteriormente se aplicaba el compresor JPEG2000 modificado. Dicha modificación fue realizada para aplicar compresión sin pérdida en las regiones de interés y compresión con pérdida en el resto de la imagen. Las regiones de interés son las estructuras significantes en la mama tales como; las masas, microcalcificaciones y ductos.

Otra propuesta en compresión por zonas de interés en donde se combinan los dos tipos principales de compresión, también conocida por compresión híbrida es la propuesta realizada por Tayer y Mohsen [3]. En este trabajo se realizó un preprocesamiento de la imagen. Primero se separó por medio de un umbral de intensidad la zona que pertenece a la mama del fondo de la imagen. Para determinar el umbral que delimita cuando el valor de un píxel pertenece a los valores habituales de la mama y no de la parte de fondo se utilizó el resultado de un trabajo estadístico implementado en las imágenes que fueron utilizadas como prueba [4]. Posteriormente se limpió la imagen de cualquier artefacto o valores que pudiera haber en el fondo, quedando solo la parte de la mama y el fondo como negro.
Posterior al preprocesamiento de la imagen se aplicó el operador lógico XOR a todos los valores. Como paso final se utilizó un algoritmo de compresión sin pérdida basado en diccionario.

Una propuesta similar a la de Tayer fue la de Xu et al. [7] en 2011, en donde utiliza compresión con pérdida en el área que no es relevante para el diagnóstico. Xu et al. comenzaron por dividir la imagen de mastografía entre zonas de interés con diferentes prioridades. Para esto se creó una máscara de intensidad de grises en donde la prioridad va en descenso del valor más alto (Valor más claro) al valor más bajo (valor más obscuro). La segmentación de la imagen está dividida en tres secciones, el área en la imagen que representa la mama, el área en la imagen que representa el pectoral y el fondo de la imagen. Posteriormente cada una de las regiones de interés fue transformada en regiones independientes a las cuales se les aplicó La Trasformada Wavelet entero-a-entero (del inglé integer-to-integer) adaptativa, de acuerdo al orden de prioridad etiquetado en la máscara, respectivamente a cada zona de interés. Por último cada región de interés con su respectiva prioridad fue codificada con el algoritmo SPIHT (Set Partitioning in Hierarchical Tres) modificado [6].

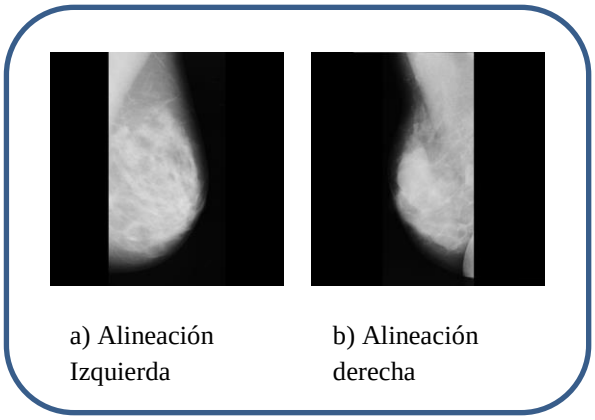
3. ESQUEMA PROPUESTO

En los últimos años han surgido propuestas algorítmicas en compresión, que aprovechan las características particulares de las mastografías. Los estudios de mastografías digitales son imágenes en escala de donde cada píxel toma una valor 0 a (2^b -1), donde *b* es la resolución de escala de grises. Los tonos de grises que obtiene cada píxel, construyen de forma abstracta la forma de la mama. Entre más denso es el tejido de la mama el tono de gris es más claro. En esta sección describimos nuestra propuesta en compresión algorítmica de tipo híbrida para mastografías digitales.
Dicho algoritmo separa la parte de la imagen que representa la mama de la parte del fondo. Esta separación se obtiene mediante la utilización de un umbral de intensidad.
La parte de la imagen que nos interesa, es conocida comúnmente como ROI por sus siglas en inglés (Region of Interest).
Posteriormente a la separación se comprime la ROI (los píxeles que representan la mama) con un algoritmo de

compresión sin pérdida basado en diccionario. El fondo es considerado como negro total por el algoritmo descompresor.

3.1. Compresor

Nuestra propuesta comienza por leer la imagen de mastografía y convertirla a una representación digital *F1(m,n)*. También se obtienen todos los parámetros de entrada del algoritmo (Alineación de la imagen, *m*, *n* y *F1(m,n))*.
Donde la "Alineación de la imagen" nos indica si la mastografía digital es la representación de la mama derecha o de la mama izquierda.
Posteriormente se crea *M(m,n)*, una máscara de referencia.



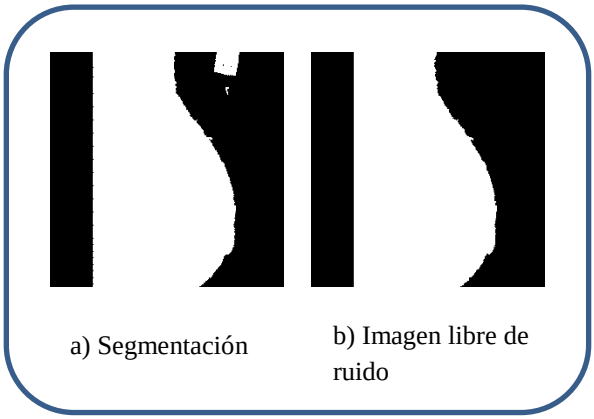
a) Alineación Izquierda b) Alineación derecha

Figura 2. Alineación de las imágenes

Para poder crear *M* se separa la parte de la imagen que representa la mama del fondo. Para esto se utiliza un umbral de intensidad *U*. El umbral utilizado en el presente trabajo fue obtenido del trabajo realizado por Tayel y Mohsen [3] en donde utilizaron propiedades estadísticas para definir el umbral de segmentación. El valor establecido fue de 32. La máscara fue construida como se muestra en la siguiente ecuación.

$$M(m,n) = \begin{cases} 1 \; F1(m,n) > U \\ 0 \; si \; no \end{cases}$$

La información segmentada puede contener no solo información del seno sino también puede contener ruido causado por artefactos que se encuentran alrededor de la mama o etiquetas que las instituciones les colocan a las imágenes para identificarlas y que también tienen valores altos. Por lo tanto, dichos elementos pasaron el umbral de segmentación. Para mitigar dicha situación y obtener realmente sola la zona de interés (la mama) se detectan los elementos que están conectados entre sí, es decir los objetos que no forman parte del fondo y, se eliminan todos aquellos que son diferentes al más grande. El elemento más grande es la representación de la mama. De esta forma nos quedamos con la máscara limpia ver Figura 1.



a) Segmentación b) Imagen libre de ruido

Figura 1. Segmentación

Después de crear la máscara se obtiene el desplazamiento de la imagen *D*. Este desplazamiento varía dependiendo la imagen.
El desplazamiento es un valor entero que nos indica que tan alejada está la imagen según su alineación del extremo vertical.

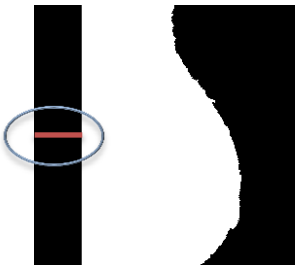


Figura 3. Desplazamiento

La alineación de la imagen pueden ser derecha o izquierda.
Como podemos observar en la Figura 3 existe una separación desde el borde de la imagen hasta donde comienza la representación de la mama. Este desplazamiento se obtiene dependiendo la alineación de la mama (la imagen 3 es alineación izquierda). Para el caso de la alineación izquierda el desplazamiento se obtiene desde el borde izquierdo hasta el comienzo de la mama. Para el caso de la alineación derecha el desplazamiento se obtiene desde el borde derecho de la imagen hasta el comienzo de la representación de la mama. D representa el valor desde el punto medio de la imagen (512) hasta encontrar el primer valor diferente de 0 en la máscara. Solo se requiere de un valor *D* debido a que la mama en las imágenes de la base de datos analizada se encuentra en posición

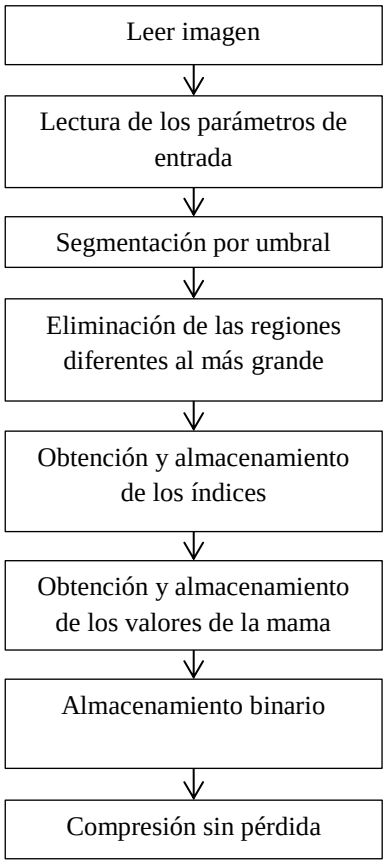


Figura 3. Diagrama de flujo del algoritmo Compresor

vertical y los valores tienden a ser el mismo o a variar insignificantemente.
Es por esta razón que solo nos quedamos con un valor y no con un arreglo de valores verticales.

De igual manera se obtienen los índices de la representación de la mama del lado contrario al desplazamiento. Los índices se almacenan en una estructura *I*(m). La estructura que guarda los índices de la mama tiene el valor de cada índice donde termina la mama. Para poder determinar estos índices se hace uso de *M*. Para obtener *I* se toma el índice de cada renglón en la máscara que cumpla con la condición *c* (valor igual a 1 y el siguiente valor igual a 0).
En caso de que la alineación de la imagen sea izquierda se recorre cada renglón de *M* de izquierda a derecha desde D hasta que se cumpla la condición *c*. Posteriormente se guarda en I el índice de M que cumpla con la condición c. La posición donde se guardara el índice será (*I*(*m*)) .

En caso contrario donde la imagen sea alineación derecha, se recorre cada renglón de *M* de derecha a izquierda desde *D* hasta que cumpla la c y de igual forma se guarda el valor del índice en *I*(*m*).
El desplazamiento D y los índices *I* se agregan a una estructura de datos *r*. Con el desplazamiento y los índices, se obtienen los valores de la imagen original que representa el área de la mama y se guardan en la misma en otra estructura *r2*. Para almacenar los valores de la región de interés (la mama) por cada renglón en *F1* se recorre desde D hasta *I*(*m*) y cada valor es almacenado en *r2*.
Dichas estructuras son almacenadas en un archivo binario *ab*. La estructura *r* primero y después *r2*.
Dicho archivo *ab*, es posteriormente comprimido por un algoritmo de compresión sin pérdida.
Así, nos quedamos con un archivo comprimido *ab2* que contienen toda la información necesaria para reconstruir la mastografía digital. El diagrama general del algoritmo compresor se puede visualizar en la Figura 3.

3.2. Descompresor

Teniendo el archivo con los datos para la reconstrucción de la imagen, el algoritmo descompresor toma como entrada este archivo y la alineación de la imagen.

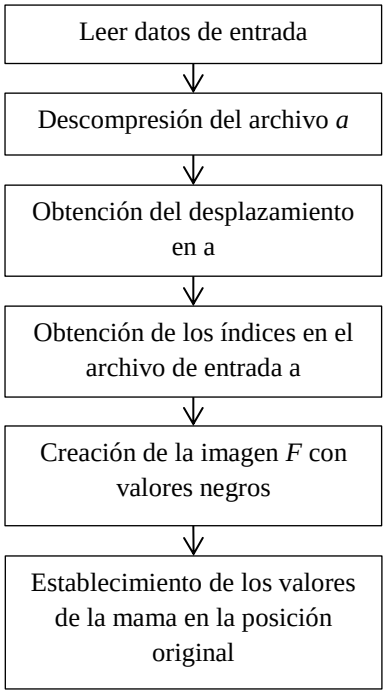


Figura 4. Diagrama de flujo del algoritmo descompresor

El algoritmo descompresor comienza por obtener los datos de entrada que será nuestro archivo binario comprimido que a su vez es salida del algoritmo compresor, el proceso inmediato siguiente es descomprimir al archivo con el algoritmo descompresor. Posteriormente se obtiene el valor del desplazamiento que es el primer elemento del archivo D y los índices I que son de tamaño m. Se lleva un contador *cont* para no perder la posición de los elementos leídos del archivo.
Se crea una imagen *F*(m,n) con todos los valores en 0 (negro).

4. RESULTADOS

Para demostrar la efectividad de nuestra propuesta y poder compararla, se utilizó la base de datos MIAS (Mammography Image Analysis Society). Esta es una base de datos pública de mastografías digitales. El contar con una base de datos al alcance de los investigadores se tiene la posibilidad de realizar comparaciones y evaluar algoritmos. Las imágenes de la base de datos MIAS tienen una resolución de 1024X1024 píxeles a 8 bits de profundidad.
Nuestra plataforma experimental cuenta con las siguientes características.
Procesador Intel Pentium con 2.40GHz de frecuencia, 4GB de memoria RAM y un disco duro de 500GB. El sistema operativo fue Ubuntu 12.04 y lenguaje de programación MATLAB.

Los experimentos se realizaron sobre imágenes de la base de datos MIAS utilizadas en el trabajo de Tayel et al. [3], esto nos permitió comparar los resultados de una forma concisa.
La escala de compresión (EC) fue medida por (Tamaño original /Tamaño del archivo comprimido). La primera columna de la tabla es el número de la imagen de la base de datos MIAS, la segunda columna son los resultados de JPEG2000 sin pérdida, la tercera columna es el resultado de la propuesta algorítmica híbrida de Tayel y Mohsen [3], la penúltima columna son los resultados de aplicar en el último proceso del algoritmo propuesto el codificador ZIP, la última columna son los resultados de nuestra propuesta con el algoritmo de compresión sin pérdida RAR.
Al comparar los resultados obtenidos con nuestra propuesta con, los resultados de aplicar JPEG2000 sin pérdida y la propuesta híbrida del trabajo [3], donde los resultados en promedio son 9.14:1, 7.23:1 y 7.80:1 respectivamente, el algoritmo propuesto muestra una EC superior al algoritmo híbrido de Tayel y Mohsen [3]. Nuestra propuesta también mostró resultados superiores que JPEG2000.

Comparativo de Algoritmos de Compresión en Mastografías Digitales*

Imagen	JPEG2000	HLC	P. ZIP	P. RAR
1	7.08	7.21	5.32	8.12
5	4.73	4.42	3.33	8.79
16	6.71	6.24	4.78	7.23
17	8.5	10.38	7.74	11.05
18	8.26	9.71	7.32	10
19	4.70	4.01	3.07	10.87
22	6.08	6.43	4.78	7.30
25	4.49	4.08	3.13	4.74
27	4.74	4.75	3.63	5.25
30	6.08	6.27	4.80	7.13
33	8.40	10.65	8.15	12.15
34	8.14	9.29	6.96	10.60
35	8.42	10.51	7.45	11.35
36	8.25	9.86	7.20	10.99
37	8.41	9.27	6.95	10.43
38	9.27	8.93	6.71	10.08
39	9.98	10.90	8.12	12.39
40	13.18	14.01	10.38	15.89
42	5.94	5.99	4.40	6.74
43	8.18	9.69	7.33	10.98
44	7.73	8.41	6.40	9.54
45	6.09	6.46	4.68	6.93
46	5.87	5.89	4.56	6.82
47	6.48	7.22	5.47	8.24
49	5.1	4.53	3.42	5.12
Promedio	**7.23**	**7.80**	**5.84**	**9.14**

*Resultados en EC obtenidos por los compresores JPEG2000, HLC, nuestra propuesta con la con el algoritmo de compresión ZIP y nuestra propuesta con el algoritmo de compresión RAR.

5. CONCLUSIONES

En el presente trabajo se propone un nuevo algoritmo híbrido para la compresión de imágenes de mastografías digitales. Debido a que la información que se pierde es la del fondo de la imagen la reconstrucción es sin pérdida relevante para el diagnóstico, puesto que la zona de la mama se recobra tal cual.

JPEG2000 es un algoritmo que ha mostrado buenos resultados en cuanto a escala de compresión. Por lo tanto se considera un buen punto de referencia. Como podemos ver en la tabla de resultados la propuesta de Tayel y Mohsen [3] supera en EC a JPEG2000 en promedio con 7.88% mientras que nuestra propuesta con RAR supera a JPEG2000 con el 26.42%. En los resultados se puede observar que nuestra propuesta con el compresor ZIP es inferior en escala de compresión incluso que JPEG2000 con el 23.80% mientras que el algoritmo RAR con el mismo archivo binario, obtuvo unan escala de compresión 56.50% superior al compresor ZIP, esto debido a que el algoritmo RAR fue favorecido con el tamaño y las características del archivo binario.

Como trabajos futuros proponemos usar algún tipo de compresión con pérdida en el área de la mama y posteriormente evaluar las diferencias con un algoritmo de detección de microcalcificaciones. Así como también probar con otro tipo de compresores en el área de la mama para evaluar y comparar los resultados.

6. AGRADECIMIENTOS

Se agradece a El Consejo Nacional de Ciencia y Tecnología (CONACyT) con número de beca 484968 y a La Universidad Autónoma de Sinaloa por el apoyo brindado.

7. REFERENCIAS

[1] Desantis, CarolSiegel, RebeccaBandi, Priti Jemal, and Ahmedin. Breast cancer statistics. CA: A Cancer Journal for Clinicians, 61:409-418, 2011.

[2] Holm LE, Callmer E, Hjalmar ML, Lidbrink E, Nilsson B, Skoog L. "Dietary Habits and Prognostic Factors in Breast Cancer". J Natl Cancer Inst 1989, 81-1218-23.

[3] M. Tayel and A. Mohsen, "Hybrid Lossless Compression of Breast Mammography," pp. 273–277, 2011.

[4] M. Tayel, A. Mohsen, "Breast Boarder Boundaries Extraction Using Statistical properties of Mammogram", Signal Processing (ICSP),IEEE 10th International Conference on Beijing China, 2010.

[5] H.-Y. Chan, H. Sari-Sarraf, B. I. Grinstead, and S. S. Gleason, "Content-Based Compression of Mammograms with Fractal-based Segmentation and a Modified JPEG2000," Opt. Eng., vol. 43, no. 12, p. 2986, 2004.

[6] M. Penedo, M. Souto, P. G. Tahoces, J. M. Carreira, J. Villalón, G. Porto, C. Seoane, J. J. Vidal, K. S. Berbaum, D. P. Chakraborty, and L. L. Fajardo, "Free-response Receiver Operating Characteristic Evaluation of Lossy JPEG2000 and Object-based Set Partitioning in Hierarchical Trees Compression of Digitized Mammograms.," Radiology, vol. 237, no. 2, pp. 450–457, 2005.

[7] Xu Weidong, Xia Shunren, "A Model Based Algorithm to Segment the Pectoral Muscle in Mammograms," Neural Networks and Signal Processing, 2003. Proceedings of the 2003 International Conference on , vol.2, no., pp.1163,1169 Vol.2, 14-17 Dec. 2003.

[8] P. G. M. Penedo, M. J. Lado, "Effects of JPEG2000 Data Compression on an Automated System for Detecting Clustered Microcalcifications in Digital Mammograms.pdf." IEEE, 2006.