



UNIVERSIDAD JUÁREZ AUTÓNOMA DE TABASCO
DIVISIÓN ACADÉMICA DE INFORMÁTICA Y SISTEMAS



CREACIÓN DE UN BANCO DE DATOS DE MAMOGRAMAS PREPROCESADOS

Trabajo recepcional bajo la modalidad de tesis
que para obtener el grado de

Licenciado en Sistemas Computacionales

Presenta

Omar Trinidad Gutiérrez Méndez

Directores

**Dra. Juana Canul Reich
M. en C. Guillermo de los Santos Torres**

Cunduacán, Tabasco

Agosto de 2013



UNIVERSIDAD JUÁREZ AUTÓNOMA DE TABASCO
DIVISIÓN ACADÉMICA DE INFORMÁTICA Y SISTEMAS



CREACIÓN DE UN BANCO DE DATOS DE MAMOGRAMAS PREPROCESADOS

Trabajo recepcional bajo la modalidad de tesis
que para obtener el grado de

Licenciado en Sistemas Computacionales

Presenta

Omar Trinidad Gutiérrez Méndez

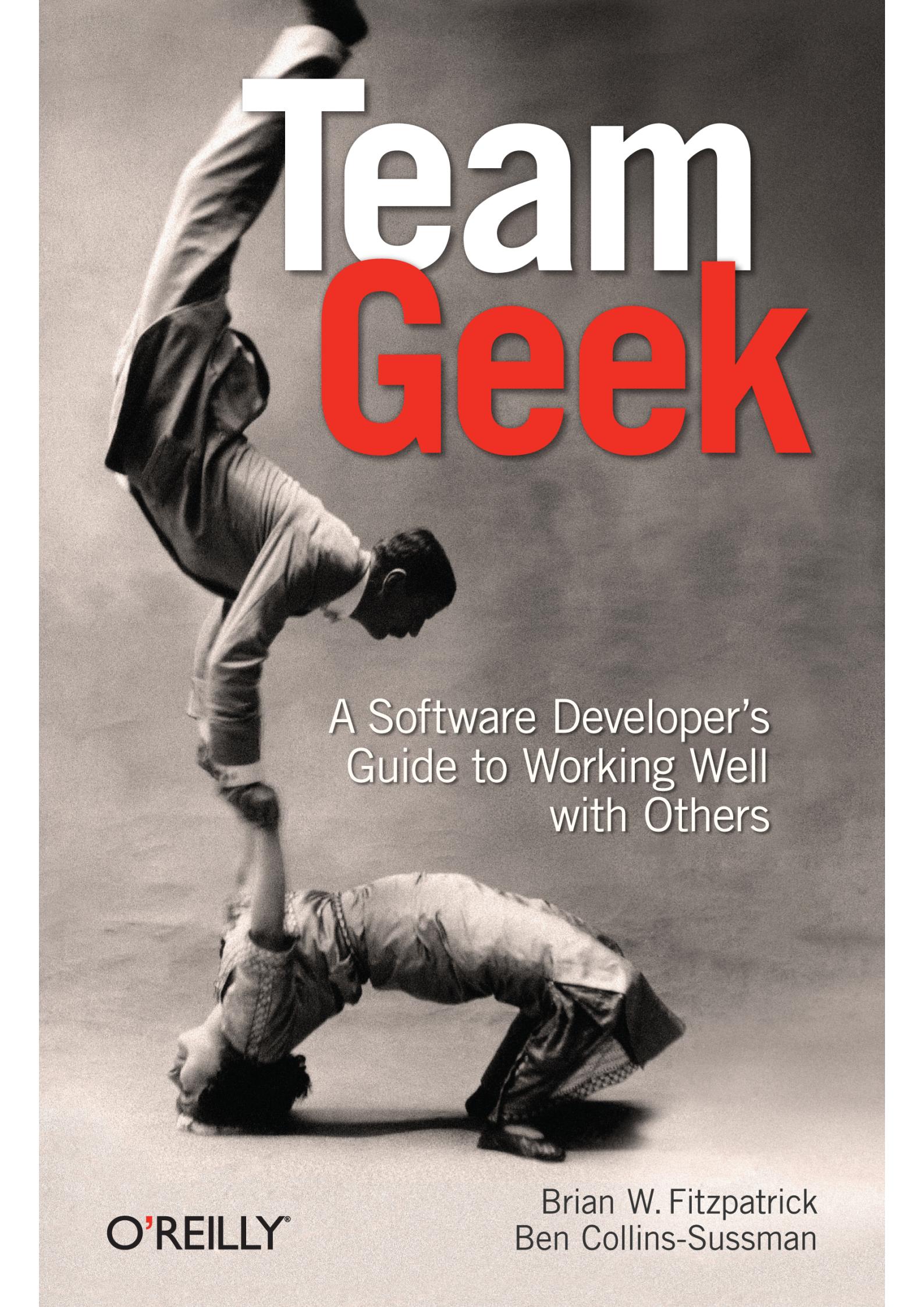
Directores

**Dra. Juana Canul Reich
M. en C. Guillermo de los Santos Torres**

Revisores

**M. en C. José Luis
Lic. Martha Patricia Silva Payró
M. en C. Cristina López Ramírez
Julian Assange
Nikola Tesla**

Team Geek



A Software Developer's
Guide to Working Well
with Others

O'REILLY®

Brian W. Fitzpatrick
Ben Collins-Sussman

Hacking Politics

How Geeks,
Progressives,
the Tea Party,
Gamers, Anarchists
and Suits Teamed Up
to Defeat SOPA and
Save the Internet

EDITED BY

DAVID MOON,

PATRICK RUFFINI,

AND DAVID SEGAL

Agradecimientos

A Guido van Rossum por inventar Python.

Dedicatoria

Índice general

Capítulo I. Generalidades

1.1	Antecedentes	1
1.2	Planteamiento del problema	2
1.3	Objetivos	2
1.3.1	Objetivo general	2
1.3.2	Objetivos específicos	3
1.4	Justificación	3

Capítulo II. Marco teórico

2.1	Breve introducción a las imágenes digitales	4
2.2	Introducción al procesamiento de imágenes	6
2.3	Imágenes médicas	7
2.3.1	Mamografías	8
2.3.2	Proceso de formación de una mamografía	8
2.3.3	Estándar DICOM	9
2.3.4	BIRADS	10
2.3.5	Sistemas CAD	11
2.4	Preprocesamiento de mamografías	12
2.5	Bases de datos similares	15

Capítulo III. Desarrollo

3.1	Etapa de recolección	18
3.2	Método	19
3.2.1	Reducción del área de trabajo	19
3.2.2	Conversión de bits	19
3.2.3	Eliminación de ruido	19
3.2.4	Mejora del contraste	20
3.2.5	Compresión	20

Capítulo IV. Resultados

Capítulo V. Conclusiones

5.1	Trabajos futuros	22
-----	----------------------------	----

Índice de figuras

2.1	Representación gráfica del <i>sistema de coordenadas de una imagen</i>	4
2.2	Resolución	6
2.3	Diferentes resoluciones. Imagen de Lenna obtenida de <i>The USC-SIPI Image Database</i>	6
2.4	Histogramas	7
2.5	Algunos tipos de imágenes médicas	8
2.6	Cuatro proyecciones en un estudio mamográfico	9
2.7	Etapas para la construcción de un sistema CAD	11
2.8	Remoción del músculo pectoral	14
3.1	Preprocessing method applied to each image in the collection	18
4.1	Compressed 8-bits mammogram and histogram	21

Índice de cuadros

2.1	Escala de grises	5
2.2	Etiquetas DICOM	10
2.3	Categorías de evaluación BIRADS	10
2.4	Algunas bases de datos de mamografías digitales	16
3.1	Etiquetas DICOM que permanecen en la imagen	19

Introducción

El documento está organizado como sigue: en el Capítulo I se da una revisión de las generalidades de nuestro proyecto, después se detallan algunos conceptos sobre imágenes digitales y también se hace una revisión de la literatura existente en el Capítulo II. En el Capítulo III se presenta el método utilizado. En el Capítulo IV se da cuenta de los resultados y finalmente en el Capítulo V se hace un recuento de lo que se hizo y se expone el trabajo futuro.

Resumen

En esta tesis ...

Proponemos un método híbrido de preprocesamiento de imágenes.

Finalmente también se expone una interfaz gráfica.

Capítulo I. Generalidades

1.1. Antecedentes

El cáncer de mama es un grave problema de salud pública, de acuerdo con la Organización Mundial de la Salud en 2008 el cáncer de mama fue el tipo de cáncer más común entre las mujeres y el que más muertes provocó [Boyle et al., 2008]. De acuerdo con el INEGI las estadísticas son similares en México [Brandan María, 2006, INEGI, 2013]. La detección temprana de este padecimiento es la mejor forma de reducir el número de decesos, a su vez, el método más comúnmente empleado para ello es el estudio de mamografías¹.

Las mamografías son exámenes radiográficos diseñados para detectar el cáncer de mama [Bushberg and Boone, 2011]. El estudio de mamogramas es una tarea a cargo de radiólogos, no es una actividad trivial, factores como la fatiga o el uso de equipos de calidad no estándar influyen en la precisión del diagnóstico. Dado que es una tarea subjetiva hay un margen de error en la opinión de los radiólogos, lo que provoca diagnósticos conocidos como *falsos positivos* que son una falsa alarma al paciente o aún peor *falsos negativos*, que evitan que el paciente conozca la gravedad de su caso y se someta a un tratamiento.

Los sistemas de *detección asistida por computadora* (CAD) son desarrollados con técnicas de reconocimiento de patrones para identificar regiones sospechosas en una imagen y alertar de ello al radiólogo. Por otro lado los sistemas de *diagnóstico asistido por computadora* que al igual que los primeros también son conocidos como CAD² van más allá y su fin es emitir un juicio sobre las lesiones, clasificando a estas como malignas o benignas [Castellino, 2005].

Los sistemas CAD fueron diseñados para ayudar a los radiólogos en la interpretación de imágenes médicas. Aunque la precisión de los sistemas CAD no es perfecta [Fenton et al., 2007] sirven como una segunda opinión para los radiólogos. El desarrollo de los sistemas CAD ha sido vertiginoso desde los años 80's [Giger et al., 2008].

¹Los términos mamografía, mamograma y mastografía son equivalentes y serán utilizados a lo largo de este documento indistintamente.

²Para distinguir entre los sistemas de detección asistida por computadora y los sistemas de diagnóstico asistido por computadora algunos autores suelen abbreviar los primeros como CADe y los últimos como CADi o CADx.

La elaboración de sistemas de este tipo es una tarea compleja, el primer reto a sortear es obtener imágenes de prueba. El acceso a imágenes médicas aun con propósitos de investigación es complicado debido a cuestiones de privacidad. Este trabajo aborda la construcción de una colección de mamogramas digitales preprocesados (y) de dominio público bautizada como *Colección de Mamogramas Digitales Preprocesados* (CPDM, por sus siglas en inglés: *Collection of Preprocessed Digital Mammograms*). Los estudios fueron recolectadas durante cuatro meses en Hospital de Alta Especialidad Dr. Juan Graham Casasús.

1.2. Planteamiento del problema

La concepción y creación de bases de datos de mamografías no es un tema nuevo [Nishikawa, 1998]. Nishiwaka plantea un enunciado interesante: *En el futuro, las bases de datos estarán constituidas por mamogramas digitales; actualmente (1998) mamogramas digitalizados están siendo utilizados*. Lo que esta tesis aborda es la creación de una colección de mamogramas preprocesados.

El desarrollo de sistemas CAD comprende varias etapas, la primera de ellas es el *preprocesamiento*. Algunas técnicas de preprocesamiento fueron abordadas en este trabajo. El preprocesamiento es la etapa previa al procesamiento de imágenes *per se*, de preparación [Ponraj et al., 2011].

Las mamografías son imágenes no convencionales.

1.3. Objetivos

1.3.1. Objetivo general

Crear un banco de datos de dominio público con mamografías digitales preprocesadas.

1.3.2. Objetivos específicos

- Aplicar los algoritmos adecuados para preprocesar los mamogramas de tal manera que mejore la apariencia visual de las mastografías, esto es, que se resalten las lesiones al ojo humano.
- Crear una ficha electrónica para cada caso que contenga datos relevantes como el diagnóstico médico.

1.4. Justificación

Aunque ya existen varias bases de datos, ninguna de ellas es una colección mexicana y pocas de ellas son latinoamericanas. La mayoría de las bases de datos públicas que existen ofrecen mamografías tradicionales *digitalizadas* con escáneres, una colección con imágenes totalmente digitales supondría mejores resultados en la ejecución de algoritmos.

También se pueden usar las imágenes para enseñanza y entrenamiento de radiólogos noveles. El proyecto es útil para la comunidad científica dedicada al procesamiento y análisis de imágenes médicas, más específicamente de imágenes mamográficas, como ya se comentó queda abierta la posibilidad de desarrollar un sistema CAD, lo que representa una ventaja potencial para los radiólogos.

Por otra parte, hay que señalar que la mayor parte de los sistemas CADs son comerciales y representan un gasto para los hospitales. México es un país en vías de desarrollo que adopta el papel de *consumidor* de tecnología, la comunidad científica trabaja en el desarrollo de algoritmos para la elaboración de sistemas CADs y otras tecnologías que permitan la reducción de la brecha tecnológica.

Creemos que una colección de mamogramas digitales es realmente útil a la comunidad

Capítulo II. Marco teórico

Desde tiempos inmemorables las imágenes han acompañado al ser humano, desde las pinturas rupestres en la Cueva de Altamira, pasando por los códices Mayas, los daguerrotipos, las cámaras *Pinhole* o los rayos X de Röntgen hasta las imágenes del Curiosity Rover de la NASA. En este capítulo se introducen conceptos básicos sobre las imágenes digitales, su procesamiento y su utilidad en la lucha contra el cáncer de mama. También se abordan ideas expuestas en trabajos similares.

2.1. Breve introducción a las imágenes digitales

Hay muchos tipos de imágenes digitales, en esta tesis nos enfocaremos en las mamografías. Las imágenes digitales están formadas por una unidad básica llamada *pixel*. Una computadora representa una imagen como un arreglo bidimensional de píxeles, cada posición de ese arreglo almacena valores numéricos. Con el fin de conocer la posición de un pixel las imágenes tienen un *sistema de coordenadas*. A diferencia del sistema cartesiano, el origen está en la parte superior izquierda (Figura 2.1).

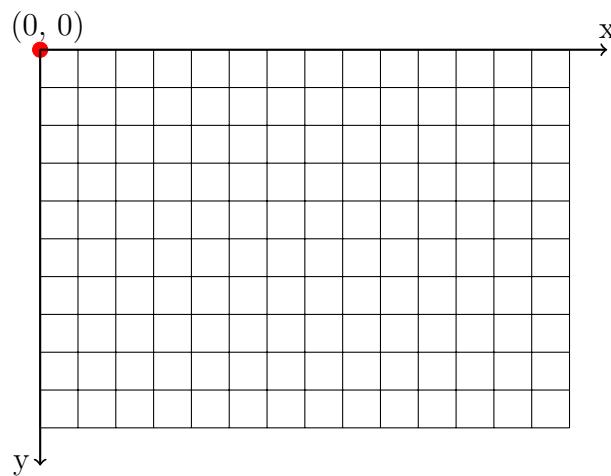


Figura 2.1: Representación gráfica del sistema de coordenadas de una imagen

La *adquisición de imágenes* es el proceso por el cual una escena se convierte en un arreglo de píxeles. El primer paso de este proceso se conoce como **muestreo** y consiste en la conversión de una distribución continua de luz a su representación discreta. El segundo paso es la *cuantificación*, que consiste en la conversión a una escala de enteros.

Cuadro 2.1: Escalas de grises [Burger and Burge, 2008]

Bits por pixel	Rango	Uso
1	0 ... 1	Imágenes binarias: documentos, fax
8	0 ... 255	Universal: Fotos, escaneos, impresiones
12	0 ... 4095	Alta calidad: Fotos, escaneos, impresiones
14	0 ... 16383	Profesional: Fotos, escaneos, impresiones
16	0 ... 65535	Altísima calidad: Medicina, astronomía

De acuerdo a lo anterior, una imagen también puede definirse como una función discreta como se muestra en la Ecuación 2.1.

$$f(x, y) = \begin{pmatrix} f(1, 1) & f(1, 2) & \cdots & f(1, y) \\ f(2, 1) & f(2, 2) & \cdots & f(2, y) \\ \vdots & \vdots & \ddots & \vdots \\ f(x, 1) & f(x, 2) & \cdots & f(x, y) \end{pmatrix} \quad (2.1)$$

En las imágenes *a color* cada pixel generalmente tiene tres canales, es decir, cada pixel almacena una tripleta de valores, las imágenes *en escala de grises* en cambio tienen un sólo canal por pixel. El rango de valores que puede tomar cada pixel es conocido como *bit de profundidad*. Cada pixel puede almacenar 2^k valores diferentes que son números enteros en el rango: $[0 \dots 2^k - 1]$ donde 0 es representa el color negro y 2^k el color blanco el blanco. La Tabla 2.1 sumariza información sobre las imágenes en escala de grises y en la Figura 2.2 podemos ver la misma imagen con diferentes niveles de grises.

Otros término básico para entender una imagen es la *resolución*, la resolución de una imagen especifica las dimensiones espaciales de la imagen en el mundo real y es dada por el número de elementos de la imagen por medida; por ejemplo, *puntos por pixel*, (dpi, por sus siglas en inglés). En la Figura 2.3 podemos ver la misma imagen con diferente resolución.

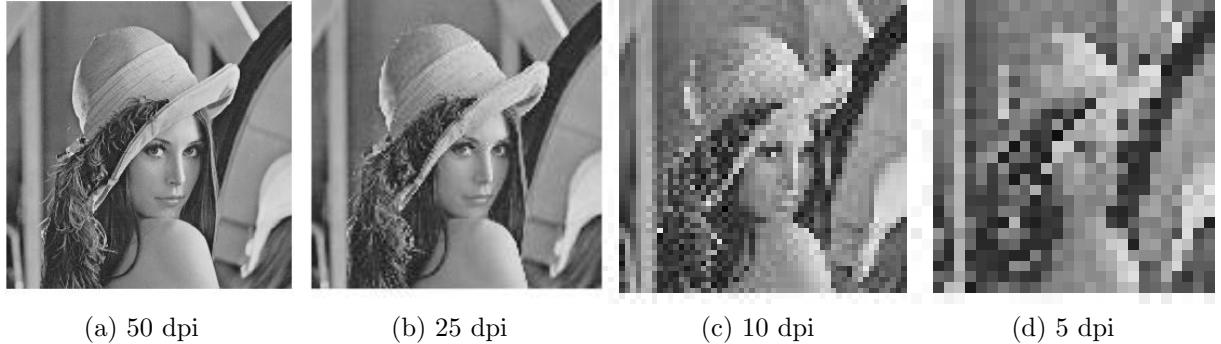


Figura 2.2: La imagen de Lenna obtenida de The USC-SIPI Image Database vista con distintos distintos valores dpi



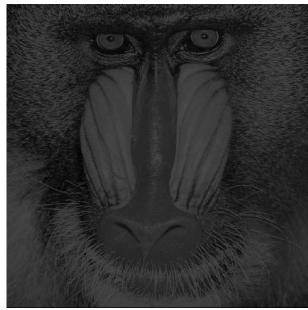
Figura 2.3: Diferentes resoluciones. Imagen de Lenna obtenida de The USC-SIPI Image Database

2.2. Introducción al procesamiento de imágenes

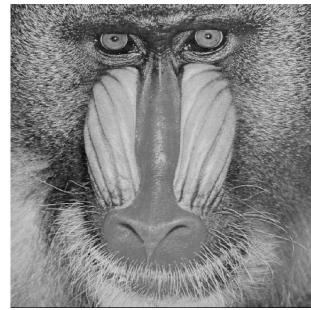
El *procesamiento de imágenes* es el conjunto de técnicas y algoritmos utilizados con un propósito doble: mejorar la apariencia visual de las imágenes al observador humano y preparar las imágenes para la medición de sus estructuras y características.

La modificación de *histogramas* es un recurso muy utilizado para evaluar

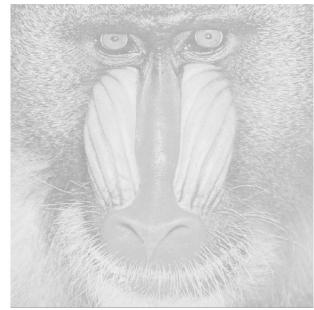
Un histograma es una distribución de frecuencia, el histograma de una imagen describe la frecuencia y la intensidad de valores que ocurren en una imagen [Burger and Burge, 2008]. Está demostrado que la ecualización de histogramas es un método efectivo para mejorar la calidad de las imágenes médicas.



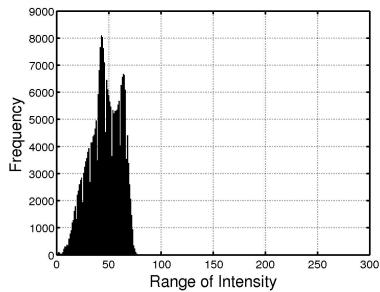
(a) Bajo contraste



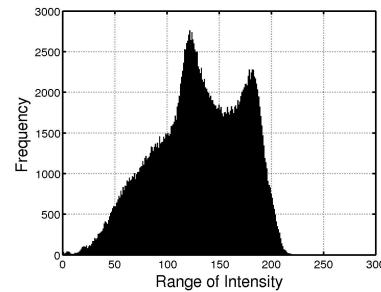
(b) Contraste normal



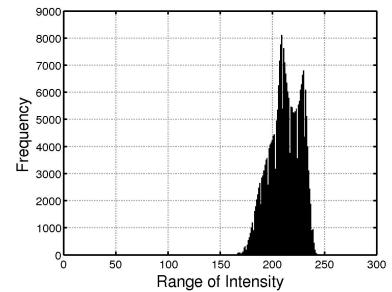
(c) Alto contraste



(d) Histograma de (a)



(e) Histograma de (b)



(f) Histograma de (c)

Figura 2.4: La misma imagen con diferentes contrastes y su respectivo histograma. En (a) se visualiza la imagen con un bajo contraste. Imagen del Mandril obtenida de The USC-SIPI Image Database

2.3. Imágenes médicas

Uno de los usos más importantes de las imágenes digitales recae en la medicina, existe una gran variedad de imágenes médicas. Podemos mencionar las **tomografías computarizadas** que son imágenes del corte transversal de alguna parte del cuerpo, son generadas con rayos X. Las **resonancias magnéticas** son generadas a partir de la propiedades magnéticas de un tejido. La teoría sobre la cual descansa la generación de imágenes de este tipos está ligada a la teoría especial de la relatividad y la mecánica cuántica.

Los **ultrasonidos** son generados a partir de ondas acústicas, su uso no es exclusivo de la medicina, también son usados en la sismología, el estudio de grietas microscópicas o en el estudio del fondo del mar. Las imágenes generadas por la **medicina nuclear** se valen del uso de isótopos radioactivos. Por último tenemos las **radiografías** (las mamografías son radiografías) [Suetens, 2009].

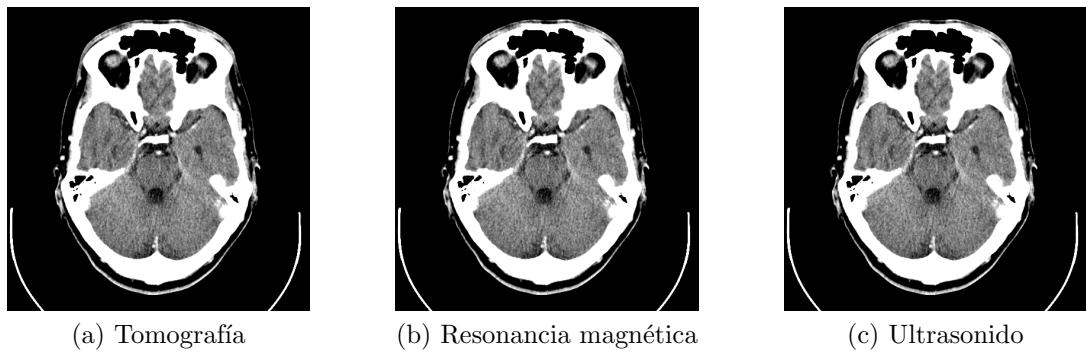


Figura 2.5: Diferentes imágenes médicas obtenidas de The Cancer Imaging Archive

2.3.1. Mamografías

El *cáncer de mama* es un padecimiento en el que se desarrollan células malignas en los tejidos de la mama, estas se duplican cada 100-300 días, un tumor de 1 cm ha realizado alrededor de treinta duplicaciones antes de alcanzar este tamaño, lo que implica que el cáncer de mama tiene, como mínimo, unos siete años de evolución. Tomando en cuenta lo anterior se entiende la importancia de la detección prematura de esta enfermedad [Brandan María, 2006].

Cada estudio mamográfico tiene dos proyecciones o vistas por cada mama. La proyección craniocaudal derecha (R-CC) e izquierda (L-CC) y la proyección mediolateral oblicua derecha (R-MLO) e izquierda (L-MLO). Estas cuatro proyecciones se pueden ver en la Figura 2.6. Las proyecciones CC se obtienen de arriba para abajo y las proyecciones MLO de afuera para adentro.

Actualmente se utilizan dos tipos de mamografías: las analógicas, mejor identificadas como *placas* y las digitales¹. Las tradicionales mamografías analógicas están reemplazando por las mamografías digitales. El uso de estas últimas representa una mejora en el proceso de adquisición de la imagen, su almacenamiento y visualización [Pisano, 2000].

2.3.2. Proceso de formación de una mamografía

Obtener una imagen mamográfica es un desafío debido a que la mama está constituida por tejidos similares entre sí y porque las lesiones buscadas por el radiólogo que indicarían

¹Las mamografías analógicas son conocidas como *Screen-Film Mammography* (SFM) y las digitales como *Full-Field Digital Mammography* (FFDM), el nombre de los equipos utilizados para generar estas imágenes es homónimo.

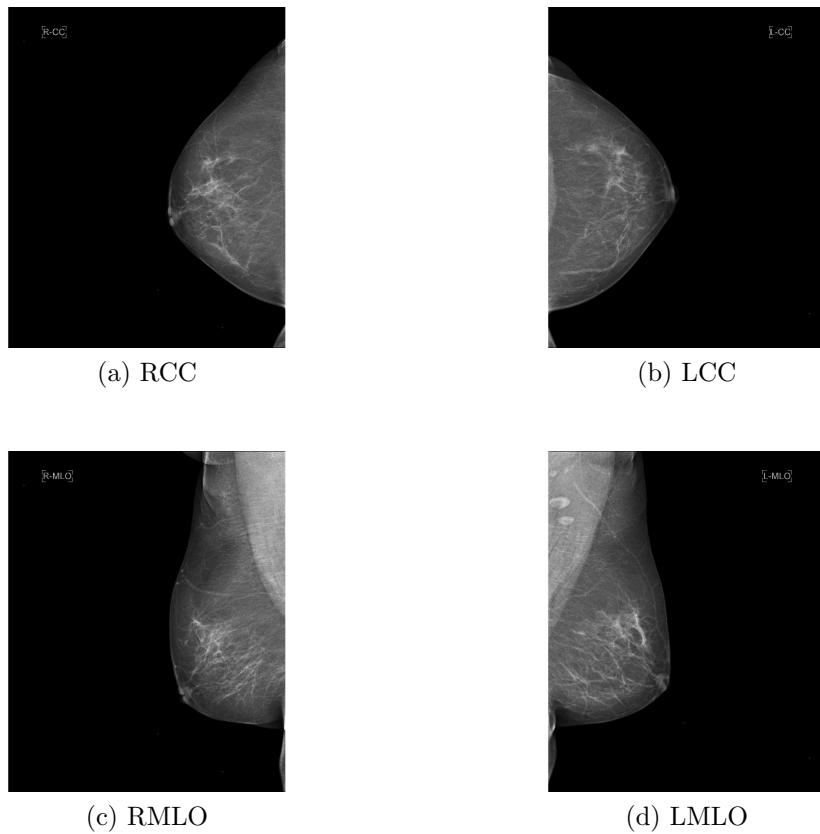


Figura 2.6: Cuatro proyecciones en un estudio mamográfico

la presencia de un tumor son pequeñas o similares al tejido normal [Brandan María, 2006]. Las mamografías son imágenes obtenidas al exponer la mama a una dosis leve de rayos X. Los mastógrafos disponen de un receptor que captura.

2.3.3. Estándar DICOM

DICOM² es el estándar en las imágenes médicas digitales, fue desarrollado por el Colegio Americano de Radiología (ACR³, por sus siglas en inglés) en conjunto con la Asociación Nacional de Fabricantes Eléctricos (NEMA⁴, por sus siglas en inglés) [NEMA et al., 2011]. DICOM opera la adquisición de la imagen, su transferencia, almacenamiento y visualización.

Los dispositivos extraíbles como los discos compactos organizan su contenido a través de un archivo llamado DICOMDIR que juega el rol de una pequeña base de datos de archivos DICOM o un índice de archivos DICOM, colocado en la raíz de un dispositivo

²Digital Imaging and COmmunication in Medicine

³American College of Radiology

⁴National Electrical Manufacturers Association

Cuadro 2.2: Algunas etiquetas DICOM

Etiqueta	Descripción
(0000, 0000)	Imágenes binarias: documentos, fax
(0000, 0000)	Imágenes binarias: documentos, fax

Cuadro 2.3: Categorías de evaluación BIRADS

Categoría	Descripción
0	Necesita evaluación adicional y/o más mamografías para comparar
1	Negativo
2	Hallazgo(s) benigno(s)
3	Hallazgo(s) benigno(s) probable(s). Se sugiere un seguimiento corto
4	Anormalidades sospechosas. Biopsia debe ser considerada
5	Altamente sugestivo de malignidad. Acciones apropiadas se deben tomar
6	Malignidad probada

extraíble. DICOMDIR organiza todos los datos del directorio en cuatro niveles: Paciente, Estudio, Serie e Imagen [Pianykh, 2011].

Los archivos DICOM contienen metainformación organizada en forma de etiquetas, en la Tabla 2.2 podemos ver algunas de las etiquetas principales de un archivo DICOM.

2.3.4. BIRADS

Como ya se mencionó antes, el diagnóstico de las mastografías lo llevan a cabo radiólogos, sin embargo, el tratamiento está a cargo de médicos especializados. Para evitar confusiones que puedan influir en el tratamiento el ACR desarrolló un léxico estandarizado llamado BIRADS⁵ [Reston, 2003], que provee la terminología básica y un sistema de clasificación para los estudios mamográficos. En la Tabla 2.3 podemos ver un resumen de las categorías de evaluación.

BIRADS también clasifica los tipos de lesiones, entre las lesiones descritas en la cuarta

⁵Breast Imaging Reporting and Data System Atlas

edición de BIRADS tenemos las *masas*, las *calcificaciones*, las *microcalcificaciones*, las *asimetrías focales*, las *asimetrías globales* y la *distorsión de la arquitectura*.

Una masa es una estructura tridimensional usualmente visible en dos proyecciones ortogonales. Las calcificaciones y microcalcificaciones son depósitos de calcio. Las asimetrías y la distorsión de la arquitectura no son masas.

2.3.5. Sistemas CAD

Un sistema CAD puede mejorar el desempeño de los radiólogos en la predicción del cáncer de mama. Varios sistemas de este tipo han sido desarrollados por la comunidad científica, para la detección de masas [Bellotti et al., 2006] o clusters de microcalcificaciones [Yu and Guan, 2000]. Sin embargo, sigue siendo un problema abierto, sobre todo con la aparición de bases de datos de mamografías digitales.

[Bozek et al., 2009]

Las etapas para la construcción de un sistema CAD generalmente son: preprocesamiento, selección de regiones de interés (ROI⁶), extracción de características y finalmente un algoritmo de aprendizaje supervisado, ver la Figura 2.7.

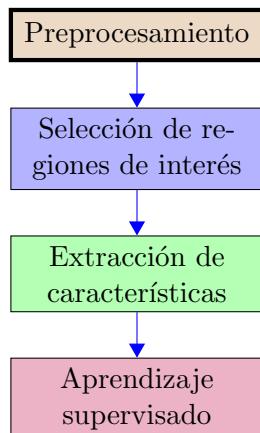


Figura 2.7: Etapas para la construcción de un sistema CAD

En este trabajo se aborda la primera etapa en la construcción de un sistema CAD. El **preprocesamiento** es la etapa previa al procesamiento de imágenes *per se*, el principal objetivo de esta etapa es mejorar la calidad de la imagen para que quede lista para su

⁶Regions of Interest

posterior procesamiento [Ponraj et al., 2011], también se utiliza para mejorar la visibilidad del observador [Rahmati et al., 2010].

2.4. Preprocesamiento de mamografías

Varios trabajos de investigación se han desarrollado sobre el preprocesamiento de mamografías. En esta sección hacemos una revisión de varios enfoques, algunos de los cuales se rescataron para crear un método híbrido que combina las fortalezas de estos.

El preprocesamiento de mamogramas engloba algunas tareas también válidas en otro tipo de imágenes como reducción del área de trabajo, remoción de ruido o mejora de contraste, no obstante, también existe el *preprocesamiento orientado a mamografías* que comprende tareas como orientación de la dirección de la mama, la supresión del músculo pectoral o la detección del pezón.

La *reducción del área de trabajo* es una tarea sencilla cuyo propósito es reducir la imagen al objeto de interés con el fin de aminorar el tiempo de ejecución de otros métodos subsiguientes y también mejorar su eficiencia. Este procedimiento es expuesto por [Holguín et al., 2006] y [Dehghani and Dezfooli, 2011].

En el proceso de adquisición de las imágenes es posible que obtengan algún tipo de contaminación, conocida como ruido. Los orígenes de ese ruido pueden ser errores de transmisión, sensores con mal funcionamiento, secciones de memoria defectuosas, entre otras cosas. En consecuencia, la *remoción de ruido* es indispensable antes de analizar los datos de las imágenes [Motwani et al., 2004].

En las imágenes mamográficas existen cuatro tipos de ruido: cuántico, electrónico fijo, señales secundarias y *quanta* secundario [Hashimoto, 2008]. El ruido cuántico es predominante en imágenes mamográficas, se genera debido a los rayos X con fotones de baja intensidad. Este tipo de ruido puede ser modelado por la distribución de Poisson. El ruido electrónico puede ser modelado como ruido gaussiano o aditivo, como se puede ver en la Ecuación 2.2:

$$g = u + n, \quad (2.2)$$

donde u es la imagen mamográfica contaminada con ruido cuántico, n es el ruido gaussiano incorporado a la imagen mamográfica g .

Dos Santos propone un método de dos pasos para reducir el ruido cuántico [Romualdo et al., 2009, dos Santos Romualdo et al., 2009], utilizando la transformada de Anscombe para convertir el ruido cuántico en ruido aditivo que es eliminado utilizando el filtrado Wiener. Su método fue evaluado utilizando un sistema CAD.

Naveed propone un método similar al anterior que en primera instancia detecta el ruido y luego lo filtra [Naveed et al., 2012]. La detección se lleva a cabo utilizando redes neuronales y para la remoción del ruido se utilizan tres filtros, que son *non local mean* (NLM), el filtro adaptativo de Wiener y un método basado en el filtro Frost. La remoción busca preservar los detalles. Este estudio propone la eliminación del ruido impulsivo y cuántico. Se evalúa el método comparando la precisión de algoritmos de *clasificación* en presencia y ausencia de ruido.

Otra tarea común en el preprocesamiento de mamografías es el *realce o mejora de contraste*. Una técnica muy socorrida para mejorar el contraste en las imágenes digitales es la Ecualización de Histogramas⁷ (EH), que se utiliza para redistribuir los niveles de grises en una imagen y como resultado hacer un ajuste de contraste, en otras palabras, con esta técnica las regiones con menos contraste obtienen más contraste y aquellas regiones con alto contraste pierden contraste.

En las imágenes médicas la distribución de los niveles de grises no es uniforme. Esta es la razón por la que el enfoque de Ecualización Local de Histogramas (LHE⁸) es más adecuado en este tipo de imágenes. En las imágenes médicas mejorar el contraste en pequeñas regiones es más importante que mejorar el contraste de forma global [Mohan and Ravishankar, 2013].

La Ecualización Adaptativa de Histogramas (AHE⁹) es una mejora de HE. AHE tiene algunas desventajas, lentitud y mejora de ruido, por lo que muchas variantes de AHE han sido propuestas [Pizer et al., 1987]. La variante más destacada es la técnica llamada Ecualización Adaptativa de Histogramas con Contraste Limitado (CLAHE¹⁰ [Pisano et al., 1998]. CLAHE ha sido ampliamente utilizado para modificar el contraste de mamografías [Pisano et al., 1998, Maitra et al., 2012].

⁷Este término es equivalente a Ecualización Global de Histogramas (EGH)

⁸Local Histogram Equalization

⁹Adaptive Histogram Equalization

¹⁰Contrast-Limited Adaptive Histogram Equalization

Rahmati propone una versión mejorada de CLAHE: FCLAHE [Rahmati et al., 2010]. Su método trabaja las desventajas de CLAHE, que mejora el -frente- y el -fondo- linealmente, de tal manera que el resultado es una imagen con alto contraste al -frente- y en el -fondo-. La mejora del fondo puede conducir a que los radiólogos identifiquen -falsos positivos-. Para probar la eficiencia del algoritmo propuesto ejecutan un algoritmo de segmentación sobre imágenes que fueron pasadas sobre CLAHE y sobre FCLAHE, obteniendo una mejora en el accuracy de FCLAHE.

La *remoción del músculo pectoral* en imágenes con proyección MLO es otra tarea de preprocesamiento orientado a mamogramas. Maitra propone un método de tres fases, la fase inicial es la mejora de contraste con el algoritmo CLAHE, después se aisla el músculo pectoral de la región de interés y finalmente se suprime utilizando el algoritmo *seeding region growing*. Su método pudo aislar el músculo pectoral en la mayoría de las imágenes de miniMIAS, ver Figura 2.8.

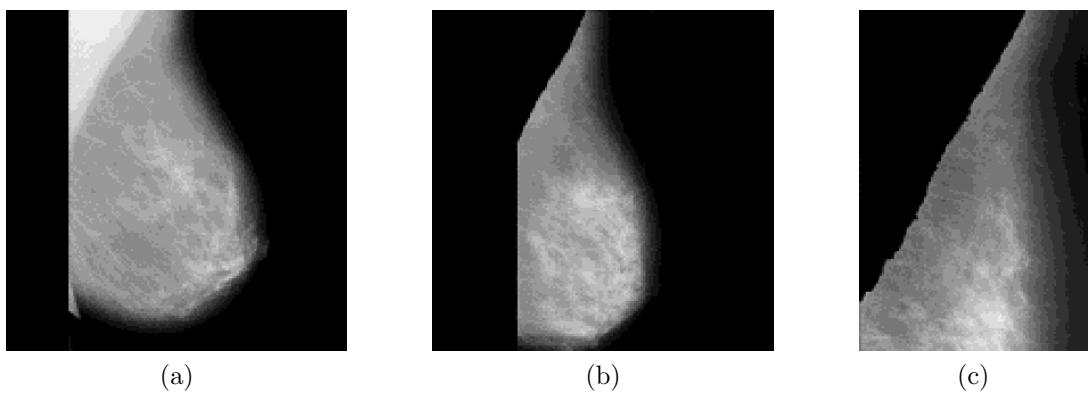


Figura 2.8: Remoción del músculo pectoral. En (a) se aprecia la imagen original, (b) muestra la remoción del músculo y (c) es un acercamiento

Akram también propone la remoción del músculo pectoral además de etiquetas y artefactos [Akram et al., 2013]. Su propuesta utiliza el método de contorno activo. Por su parte Mirzaalian plantea un algoritmo para la extracción del contorno de la mama y otro para la extracción del músculo pectoral [Mirzaalian et al., 2007]. Para extraer el contorno de la mama utiliza en conjunto la ecualización de histogramas, una técnica de convolución con una máscara, que sirve como un filtro de paso bajo, binarización y etiquetado.

Ponraj y Pisano hacen estudios extensivos del estado del arte del algoritmos utilizados para modificar imágenes mamográficas [Ponraj et al., 2011], [Pisano et al., 2000].

2.5. Bases de datos similares

Existen múltiples bases de datos, públicas o privadas, utilizadas en estudios relacionados a la detección del cáncer de mama. Las características de ellas son diferentes. En la Tabla 2.4 se muestran algunas de ellas.

MIAS [Suckling, 1994] aunque es el recurso más viejo, sigue siendo ampliamente utilizado. MIAS no tiene soporte actualmente. DDSM [Heath et al., 2000] es el recurso más utilizado. Es importante notar la presencia de dos recursos latinoamericanos, más específicamente, brasileños.

IRMA (Image Retrieval in Medical Applications) is a extensive project [Oliveira et al., 2008]

...

MIRaCLe (Mammography Image reading for Radiologists and Computers Learning Database) es un repositorio dinámico para el entrenamiento y evaluación de computadoras y radiólogos [Antoniou et al., 2009]. MIRaCLe cuenta con un sistema web que provee dos servicios: *Software de Clasificación y Evaluación*, donde el usuario puede realizar consultas, visualizar y escoger diferentes imágenes para probar algoritmos CAD y el *Evaluación de Radiólogos*, que provee funcionalidades para evaluar el desempeño de un radiólogo.

MIDAS (Mammographic Image Database for Automated Analysis) es una base de datos con muestras tomadas entre la población femenina de Brasil [Fernandes et al., 2012]. De esta base de datos destaca la inclusión de secuencias de datos genómicas de los pacientes y su apertura a la colaboración externa, -brand new innovation-. MIDAS cuenta también con un sistema web.

BancoWeb es una base de datos creada por LAPIMO (Laboratório de Análise e Processamento de Imagens Médicas e Odontológicas) [Matheus and Schiabel, 2011].

INbreast: [Moreira et al., 2012], es una base de datos , los autores sugieren características esenciales de una base de datos que conllevaría a la creación de sistemas CAD.

Magic-5 [Bellotti et al., 2004] es la evolución de GPCalma (Grid Platform for Computer Assisted Library for Mammography) [Lauria et al., 2006], una base de datos distribuida.

La Universidad de Calgary ha creado un banco de imágenes conocido como AMDI

Cuadro 2.4: Algunas bases de datos de mamografías digitales

Nombre	Número de casos	Número de imágenes	Proyecciones	Tipo de archivo	Lugar de origen	Año	ACR	BIRADS
mini-MIAS	161	322	MLO	PGM	Reino Unido	1994	yes	yes
DDSM	2620	10480	MLO & CC	LJPEG	Estados Unidos	1999	yes	yes
BancoWeb	320	1'400	MLO & CC	TIFF	Brasil	1999	yes	yes
INbreast	115	410	MLO & CC	DICOM	Portugal	2012	yes	yes
MIDAS	100	600	MLO	DICOM	Brasil	2011	yes	yes
IRMA	-	10'509	MLO & CC	LJPEG	Alemania	2008	yes	yes
MIRaCLE	196	10480	MLO & CC	LJPEG	Grecia	2009	yes	yes
Magic-5 (GPCalma)	967	3369	MLO & CC	DICOM	Italia	2002	no	yes

[Suri and Rangayyan, 2006] cuya función principal es ser un atlas que ayude a los radiólogos a encontrar casos similares cuando emitir un diagnóstico se torna complicado. Pero también sirve como un sistema educativo [Guliato et al., 2009].

Capítulo III. Desarrollo

En este capítulo se presenta el método empleado para preprocesar los mamogramas digitales.

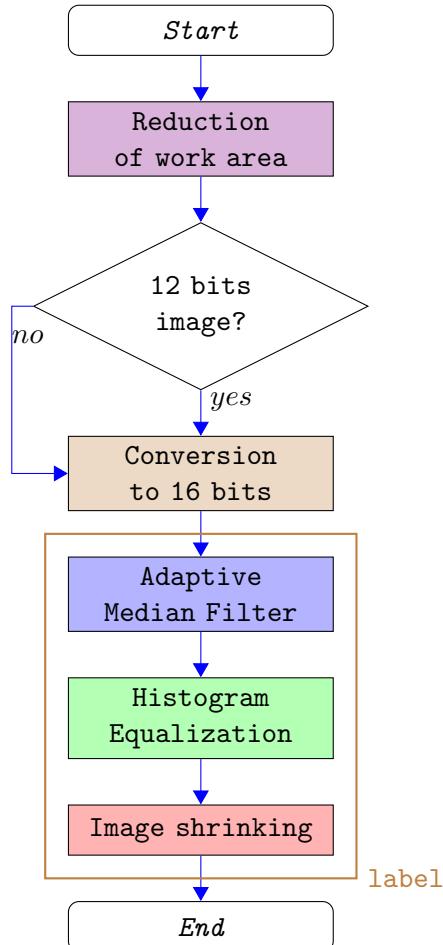


Figura 3.1: Preprocessing method applied to each image in the collection

3.1. Etapa de recolección

Las mamografías originalmente almacenadas en discos compactos, fueron extraídas y ... , removimos los datos privados de las imágenes, los únicos campos son los que se muestran en la 3.1

Cuadro 3.1: Etiquetas DICOM que permanecen en la imagen

Etiquetas	Descripción
(0000, 0001)	Bits alojados

3.2. Método

Se eligió un método híbrido

3.2.1. Reducción del área de trabajo

La primera etapa del método consiste en eliminar la región oscura que cubre gran parte de la imagen para acelerar el tiempo de ejecución de los procedimientos posteriores. El enfoque es similar al expuesto por Dehghani y Holguín [Dehghani and Dezfooli, 2011, Holguín et al., 2006].

3.2.2. Conversión de bits

Las mamografías que se obtuvieron son de 12 bits. Sin embargo al visualizarlas se obtiene una imagen opaca, esto es porque ...

3.2.3. Eliminación de ruido

Se elimina el ruido.

3.2.4. Mejora del contraste

Se mejora el contraste.

3.2.5. Compresión

```
1 % Get the size of the image
2 % Based on Abubaker code
3 [height width] = size(image);
4 imageCopy = repmat(uint8(0), height, width);
5 divider = 0.0;
6 maxLevel = double(usedGrayLevels);
7
8 while 1
9     divider = divider + 0.01;
10    if maxLevel/divider <= 255
11        break;
12    end
13 end
14 fprintf('text');
15 for h=1:1:height
16     for w=1:1:width
17         imageCopy(h, w) = image(h, w)/divider;
18     end
19 end
```

Capítulo IV. Resultados

La evaluación de resultados se realizó utilizando la opinión de los médicos y también utilizando los diagramas como herramienta evaluadora.

Cada caso preprocesado se almacenó en un servidor y está disponible en la siguiente dirección: www.casi.dais.mx.

El código fuente que se generó está disponible en línea bajo un controlador de versiones en el repositorio: github.com/omartrinidad/mammograms.

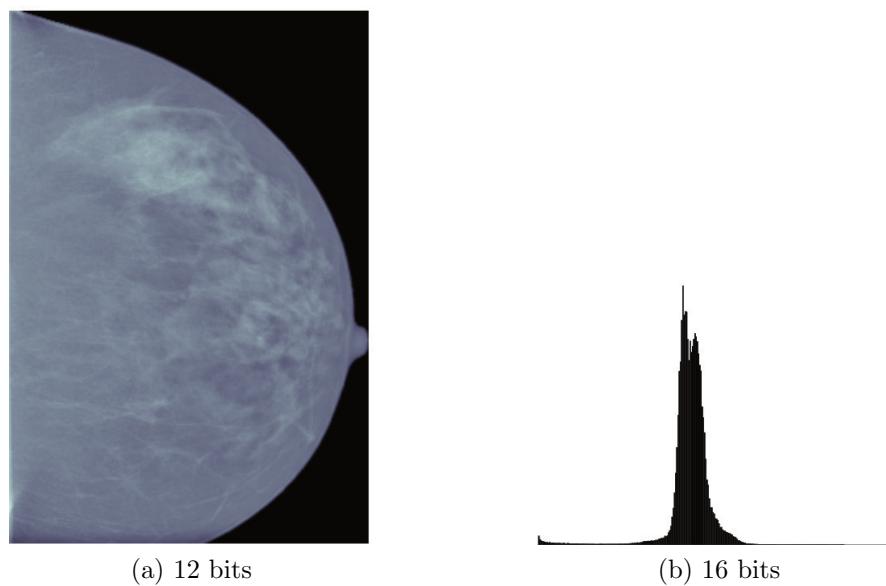


Figura 4.1: This description appears in above of the image.

Capítulo V. Conclusiones

Encontramos que

5.1. Trabajos futuros

Es posible hacer uso de algoritmos más efectivos para la eliminación del ruido y la mejora del contraste.

Trabajos similares al nuestro [Heath et al., 2000] han construido cosas como un servicio web.

El rango de aplicaciones médicas es amplio. El dataset está listo para usarse en etapas posteriores o ampliar la etapa de procesamiento.

Es posible crear un servicio web [referencia] que recupere las imágenes de forma efectiva (selección de casos), con una interfaz web responsiva. Previsualización. de estos casos.

Rescaling. Pasar de una resolución a otra sin problemas.

Preprocesamiento: eliminar el pectoral en las imágenes ¿cráneo caudal? Hacer que las imágenes queden en una sola dirección.

Eliminar el ruido cuántico, algunos trabajos que hablan sobre eso son ... el de Santos y el de Elsevier

Remover el músculo pectoral en vistas MLO como parte del preprocesamiento.

Bibliografía

- [Akram et al., 2013] Akram, F., Kim, J. H., and Choi, K. N. (2013). A preprocessing algorithm for the cad system of mammograms using the active contour method. *Applied Medical Informatics*, 32(2):1–13. (página 14).
- [Antoniou et al., 2009] Antoniou, Z. C., Giannakopoulou, G. P., Andreadis, I. I., Nikita, K. S., Ligomenides, P. A., and Spyrou, G. M. (2009). A web-accessible mammographic image database dedicated to combined training and evaluation of radiologists and machines. In *Information Technology and Applications in Biomedicine, 2009. ITAB 2009. 9th International Conference on*, pages 1–4. IEEE. (página 15).
- [Bellotti et al., 2004] Bellotti, R., Bagnasco, S., Bottigli, U., Castellano, M., Cataldo, R., Catanzariti, E., Cerello, P., Cheran, S., De Carlo, F., Delogu, P., et al. (2004). The magic-5 project: medical applications on a grid infrastructure connection. In *Nuclear Science Symposium Conference Record, 2004 IEEE*, volume 3, pages 1902–1906. IEEE. (página 15).
- [Bellotti et al., 2006] Bellotti, R., De Carlo, F., Tangaro, S., Gargano, G., Maggipinto, G., Castellano, M., Massafra, R., Cascio, D., Fauci, F., Magro, R., et al. (2006). A completely automated cad system for mass detection in a large mammographic database. *Medical Physics*, 33:3066. (página 11).
- [Boyle et al., 2008] Boyle, P., Levin, B., et al. (2008). *World Cancer Report 2008*. IARC Press, International Agency for Research on Cancer. (página 1).
- [Bozek et al., 2009] Bozek, J., Mustra, M., Delac, K., and Grgic, M. (2009). A Survey of Image Processing Algorithms in Digital Mammography. In *Recent Advances in Multimedia Signal Processing and Communications*, pages 631–657. Springer. (página 11).
- [Brandan María, 2006] Brandan María, V. Y. (2006). Detección del Cáncer de mama: Estado de la Mamografía en México. pages 147–162. (páginas 1, 8, 9).
- [Burger and Burge, 2008] Burger, W. and Burge, M. J. (2008). *Digital image processing: an algorithmic introduction using Java*. Springer. (páginas 5, 6).
- [Bushberg and Boone, 2011] Bushberg, J. T. and Boone, J. M. (2011). *The essential physics of medical imaging*. Lippincott Williams & Wilkins. (página 1).

- [Castellino, 2005] Castellino, R. A. (2005). Computer aided detection (cad): an overview. *Cancer Imaging*, 5(1):17. (página 1).
- [Dehghani and Dezfooli, 2011] Dehghani, S. and Dezfooli, M. A. (2011). A method for improve preprocessing images mammography. *International Journal of Information and Education Technology*, 1(1). (páginas 12, 19).
- [dos Santos Romualdo et al., 2009] dos Santos Romualdo, L. C., da Costa Vieira, M. A., and Schiabel, H. (2009). Mammography images restoration by quantum noise reduction and inverse mtf filtering. In *Computer Graphics and Image Processing (SIBGRAPI), 2009 XXII Brazilian Symposium on*, pages 180–185. IEEE. (página 13).
- [Fenton et al., 2007] Fenton, J. J., Taplin, S. H., Carney, P. A., Abraham, L., Sickles, E. A., D'Orsi, C., Berns, E. A., Cutter, G., Hendrick, R. E., Barlow, W. E., et al. (2007). Influence of computer-aided detection on performance of screening mammography. *New England Journal of Medicine*, 356(14):1399–1409. (página 1).
- [Fernandes et al., 2012] Fernandes, F., Bonifácio, R., Brasil, L., Guadagnin, R., and Lamas, J. (2012). Midas—mammographic image database for automated analysis. *Mammography-Recent Advances]. InTechOpen*, pages 243–60. (página 15).
- [Giger et al., 2008] Giger, M. L., Chan, H.-P., and Boone, J. (2008). Anniversary paper: History and status of cad and quantitative image analysis: the role of medical physics and aapm. *Medical physics*, 35:5799. (página 1).
- [Guliato et al., 2009] Guliato, D., Bôaventura, R. S., Maia, M. A., Rangayyan, R. M., Simedo, M. S., and Macedo, T. A. (2009). Indiam—an e-learning system for the interpretation of mammograms. *Journal of digital imaging*, 22(4):405–420. (página 17).
- [Hashimoto, 2008] Hashimoto, B. (2008). *Practical digital mammography*. Thieme. (página 12).
- [Heath et al., 2000] Heath, M., Bowyer, K., Kopans, D., Moore, R., and Kegelmeyer, P. (2000). The Digital Database for Screening Mammography. In *Proceedings of the 5th International Workshop on Digital Mammography*, pages 212–218. (páginas 15, 22).
- [Holguín et al., 2006] Holguín, L. G. A., Álvarez, D., and Guevara, M. L. (2006). Preprocesamiento de imágenes aplicadas a mamografías digitales. *Scientia Et Technica*. (páginas 12, 19).
- [INEGI, 2013] INEGI (visitado en agosto de 2013). www.inegi.org.mx. (página 1).
- [Lauria et al., 2006] Lauria, A., Massafra, R., Tangaro, S. S., Bellotti, R., Fantacci, M., Delogu, P., Torres, E. L., Cerello, P., Fauci, F., Magro, R., et al. (2006). Gpcalma: an

- italian mammographic database of digitized images for research. In *Digital Mammography*, pages 384–391. Springer. (página 15).
- [Maitra et al., 2012] Maitra, I. K., Nag, S., and Bandyopadhyay, S. K. (2012). Technique for Preprocessing of Digital Mammogram. *Computer Methods and Programs in Biomedicine*, 107(2):175–188. (página 13).
- [Matheus and Schiabel, 2011] Matheus, B. R. N. and Schiabel, H. (2011). Online mammographic images database for development and comparison of cad schemes. *Journal of Digital Imaging*, 24(3):500–506. (página 15).
- [Mirzaalian et al., 2007] Mirzaalian, H., Ahmadzadeh, M. R., Sadri, S., and Jafari, M. (2007). Pre-processing algorithms on digital mammograms. In *MVA*, pages 118–121. (página 14).
- [Mohan and Ravishankar, 2013] Mohan, S. and Ravishankar, M. (2013). Modified contrast limited adaptive histogram equalization based on local contrast enhancement for mammogram images. In *Mobile Communication and Power Engineering*, pages 397–403. Springer. (página 13).
- [Moreira et al., 2012] Moreira, I. C., Amaral, I., Domingues, I., Cardoso, A., Cardoso, M. J., and Cardoso, J. S. (2012). Inbreast: toward a full-field digital mammographic database. *Academic Radiology*, 19(2):236–248. (página 15).
- [Motwani et al., 2004] Motwani, M. C., Gadiya, M. C., Motwani, R. C., and Harris Jr, F. C. (2004). Survey of image denoising techniques. In *Proceedings of GSPX*, pages 27–30. Citeseer. (página 12).
- [Naveed et al., 2012] Naveed, N., Hussain, A., Arfan Jaffar, M., and Choi, T.-S. (2012). Quantum and impulse noise filtering from breast mammogram images. *Computer methods and programs in biomedicine*. (página 13).
- [NEMA et al., 2011] NEMA, N. E. M. A. et al. (2011). Digital imaging and communications in medicine (DICOM). *NEMA Standard Publications PS*, 3. (página 9).
- [Nishikawa, 1998] Nishikawa, R. M. (1998). Mammographic databases. *Breast disease*, 10(3):137–150. (página 2).
- [Oliveira et al., 2008] Oliveira, J. E. E., Gueld, M. O., de A. Araújo, A., Ott, B., and Deserno, T. M. (2008). Toward a standard reference database for computer-aided mammography. volume 6915, pages 69151Y–69151Y–9. (página 15).
- [Pianykh, 2011] Pianykh, O. S. (2011). *Digital Imaging and Communications in Medicine (DICOM): A practical introduction and survival guide*. Springer. (página 10).

- [Pisano, 2000] Pisano, E. D. (2000). Current status of full-field digital mammography1. *Radiology*, 214(1):26–28. (página 8).
- [Pisano et al., 2000] Pisano, E. D., Cole, E. B., Hemminger, B. M., Yaffe, M. J., Aylward, S. R., Maidment, A. D., Johnston, R. E., Williams, M. B., Niklason, L. T., Conant, E. F., et al. (2000). Image processing algorithms for digital mammography: A pictorial essay. *Radiographics*, 20(5):1479–1491. (página 14).
- [Pisano et al., 1998] Pisano, E. D., Zong, S., Hemminger, B. M., DeLuca, M., Johnston, R. E., Muller, K., Braeuning, M. P., Pizer, S. M., et al. (1998). Contrast limited adaptive histogram equalization image processing to improve the detection of simulated spiculations in dense mammograms. *Journal of Digital Imaging*, 11(4):193–200. (páginas 13, 13).
- [Pizer et al., 1987] Pizer, S. M., Amburn, E. P., Austin, J. D., Cromartie, R., Geselowitz, A., Greer, T., ter Haar Romeny, B., Zimmerman, J. B., and Zuiderveld, K. (1987). Adaptive histogram equalization and its variations. *Computer vision, graphics, and image processing*, 39(3):355–368. (página 13).
- [Ponraj et al., 2011] Ponraj, D. N., Jenifer, M. E., Poongodi, D. P., and Manoharan, J. S. (2011). A Survey on the Preprocessing Techniques of Mammogram for the Detection of Breast Cancer. *Journal of Emerging Trends in Computing and Information Sciences*, 2(12). (páginas 2, 12, 14).
- [Rahmati et al., 2010] Rahmati, P., Hamarneh, G., Nussbaum, D., and Adler, A. (2010). A new preprocessing filter for digital mammograms. In *Image and Signal Processing*, pages 585–592. Springer. (páginas 12, 14).
- [Reston, 2003] Reston, V. (2003). Birads Atlas. (página 10).
- [Romualdo et al., 2009] Romualdo, L. C., Marcel, A., and Homero, S. (2009). A pre-processing technique to improve the performance of a computer-aided detection scheme in mammography. (página 13).
- [Suckling, 1994] Suckling (1994). The mini-mias Database of Mammograms. *Excerpta Medica. International Congress Series*, (1069):375–378. (página 15).
- [Suetens, 2009] Suetens, P. (2009). *Fundamentals of Medical Imaging*. Cambridge University Press. (página 7).
- [Suri and Rangayyan, 2006] Suri, J. S. and Rangayyan, R. M. (2006). *Recent advances in breast imaging, mammography, and computer-aided diagnosis of breast cancer*, volume 155. SPIE press. (página 17).

[Yu and Guan, 2000] Yu, S. and Guan, L. (2000). A cad system for the automatic detection of clustered microcalcifications in digitized mammogram films. *Medical Imaging, IEEE Transactions on*, 19(2):115–126. (página 11).