UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE MEDICINA

Departamento de Radiología y Medicina Física

(Radiología)

TESIS DOCTORAL

Análisis y propuesta de métricas de calidad de imagen médica que mimetizan al observador humano

MEMORIA PARA OPTAR AL GRADO DE DOCTOR

PRESENTADA POR

Gabriel Prieto Renieblas

Director

Eduardo Guibelalde del Castillo

Madrid, 2017

Universidad Complutense de Madrid

Facultad de Medicina

Departamento de Radiología y Medicina Física

Programa en Ciencias Radiológicas

Análisis y propuesta de métricas de calidad de imagen médica que mimetizan al observador humano

Memoria para optar al grado de doctor presentada por

Gabriel Prieto Renieblas

Bajo la dirección del doctor

Eduardo Guibelalde del Castillo

Madrid, 2015

A Teresa

Agradecimientos

Esta memoria ha sido posible gracias a la participación de los coautores de los trabajos que la componen y de todos aquellos que han permitido, de manera directa o indirecta, su realización.

Deseo destacar en primer lugar a los profesores Eduardo Guibelalde y Agustín Turrero por su dirección y apoyo constante. No puedo dejar de señalar las largas conversaciones en persona y telefónicas con Agustín discutiendo desviaciones estándar, coeficientes de correlación y estadísticos de tenor parecido.

Quiero agradecer a todos mis compañeros del Departamento de Radiología y de Medicina Física sus sugerencias, indicaciones y experiencia que me han servido de guía en muchas ocasiones a lo largo de esta investigación, y especialmente a Alberto Muñoz, Margarita Chevalier, Alfonso Calzado, Víctor Delgado, José Miguel Fernández, Carlos Prieto, Luciano González y Eliseo Vañó.

Lejos geográficamente, pero presente en nuestro día a día, también quiero recordar a Irene Hernández por sus comentarios, sugerencias y amistad durante estos años.

De forma especial quiero dar las gracias, por las aclaraciones y consejos que he recibido, a los profesores Elizabeth Krupinski (University of Arizona), Zhou Wang (University of Waterloo), Sheila S. Hemami y David M. Rouse (ambos en University of Cornell) y Alan C. Bovik (University of Texas, Austin) acerca de los procesos de percepción automatizados y de las métricas desarrolladas por ellos. Sin sus orientaciones prácticas y teóricas esta tesis no hubiera sido posible.

Por último, no quiero olvidar a Mercedes Gálvez por su continua guía y apoyo en las procelosas aguas de la burocracia universitaria: sin ella, no habría sobrevivido entre tanto papel.

Informe del Director de la Tesis Doctoral

DATOS DE LA TESIS DOCTORAL	
Nombre del Doctorando	***Gabriel Prieto Renieblas***
Título de la Tesis	***Análisis y propuesta de métricas de calidad de imagen médica que mimetizan al observador humano***
Facultad o Centro	***Depto. Radiología. Facultad de Medicina***

DATOS DEL DIRECTOR DE LA TESIS DOCTORAL	
Nombre Completo	***Eduardo Guibelalde del Castillo***
Centro al que pertenece y dirección	***Depto. Radiología. Facultad de Medicina. Avda. Complutense s/n***
D.N.I/Pasaporte	***03805386J***
e-mail	***egc@ucm.es***

	VALORACIÓN DE LA TESIS			
	Muy Buena	***Buena***	***Suficiente***	***Deficiente***
Originalidad	***X***			
Definición Objetivos	***X***			
Metodología	***X***			
Relevancia Resultados	***X***			
Discusión / Conclusiones	***X***			

INFORME *(en caso necesario se podrán añadir más hojas):*

La presente tesis doctoral se presenta en formato publicaciones. En todas ellas el doctorando ha demostrado gran iniciativa y capacidad innovadora en los planteamientos llevando el peso en la publicación. La problemática desarrollada, percepción de calidad de imagen en radiodiagnóstico realizada por observadores humanos vs algoritmos matemáticos, es de amplia actualidad y plantea muchas cuestiones aún no resueltas. Destaca en este trabajo la excelente cooperación entre diferentes especialistas (físicos, médicos, radiólogos, estadísticos) lo que ha permitido un gran rigor en los planteamientos y éxito en la consecución de los objetivos previstos.

Madrid, a 22/10/2015

Fdo.:

Este impreso deberá entregarse al Departamento/Órgano responsable del Posgrado/ Comisión responsable del Programa de Doctorado, para su estudio y aprobación en la admisión a trámite de la tesis doctoral. Asimismo, deberá incluirse entre la documentación enviada a la Comisión de Doctorado para la designación del Tribunal y aprobación de la defensa de la Tesis Doctoral.

Tesis Doctoral en Formato de Publicaciones

Esta tesis doctoral se presenta en formato publicaciones, de acuerdo con el apartado 4.4 del Consejo de Gobierno de fecha 14 de octubre de 2008, en el que se modifica la normativa aprobada por la Junta de Gobierno con fecha 26 de mayo de 1999, en la que se aprueba la normativa de Desarrollo del Régimen relativo a la elaboración, tribunal, defensa y evaluación de la Tesis Doctoral del Real Decreto 778/1998, de 30 de abril (BOE del 1 de mayo) por el que se establece la ordenación de las enseñanzas universitarias oficiales de la Universidad Complutense de Madrid. Dichas publicaciones recogen todos los resultados que han sido obtenidos en los diferentes trabajos de investigación desarrollados con el fin de alcanzar el objetivo fijado para la realización de la tesis.

A continuación, se detallan los artículos que integran la tesis.

Publicaciones incluidas en esta tesis

I.- Prieto, G., Chevalier, M., & Guibelalde, E., 2008. A CDMAM Image Phantom Software Improvement for Human Observer Assessment. *Lecture Notes in Computer Science, Digital Mammography,* Volumen 5116, pp. 181-187. **Web of Science: 000258502700026**

II.- Prieto, G., Guibelalde, E., Chevalier, M. & Turrero, A., 2011. Use of the cross-correlation component of the multiscale structural similarity metric (R* metric) for the evaluation of medical images. *Med. Phys,* 38(8), p. 4512.7. **Web of Science: 000293417500009**

III.- Prieto, G., Chevalier, M. & Guibelalde, E., 2011. A software tool to compare contrast-detail detection in uniform and in real mammographic backgrounds. *Proc. SPIE,* Volumen 7966, pp. 122-128. **Web of Science: 000296320800045**

IV.- Prieto, G., Turrero, A., Muñoz, A., Gómez-León, N., Guibelade, E., 2015. Structural Similarity Index Family for Image Quality Assessment in Radiological Images. *Plos ONE.* **Revista indexada en Web of Science.** *En proceso de revisión editorial.*

Índice

1. Abreviaturas

4-	Basado en 4 componentes
BKE	Fondo conocido de manera exacta
BKS	Fondo conocido de manera estadística
BPF	Películas planas de hueso
CDMAM	Maniquí para mamografía de contraste-detalle
CPF	Películas planas de tórax
DICOM	Imagen digital y comunicación en Medicina
DQE	Eficiencia de detección cuántica
ESR	Sociedad Europea de Radiología
G	Gradiente
GB	Desenfoque gaussiano
GN	Ruido gaussiano
HVS	Sistema visual humano
IQM	Métrica de Calidad de Imagen
J2000	JPEG 2000
JPEG	Grupo conjunto de expertos en fotografía
M, MS	Multiescala
MOS	Media de evaluaciones subjetivas
MSE	Error cuadrático medio
MS-SSIM	Índice de similaridad estructural multiescala
MTF	Función de modulación de Transferencia
NEQ	Ruido equivalente cuántico
NPS	Espectro de ruido
NPWMF	Non Prewhitening matched filter
PACS	Sistema de Comunicación y Almacenamiento de Imágenes
PSNR	Relación de pico entre la señal y el ruido
r*	Índice de correlación cruzada de SSIM
R*	Índice de correlación cruzada multiescala de MS-SSIM
RM	Resonancia magnética
ROC	Característica operativa del receptor
SKE	Señal conocida de manera exacta
SKS	Señal conocida de manera estadística
SSIM	Índice de similaridad estructural

2. Resumen

La investigación que se presenta en este documento se centra en el paradigma de la percepción automática de la calidad de imagen médica, y en la correlación de dicha percepción con la percepción humana.

El análisis de la calidad de imagen médica tiene un lugar central en el diseño de sistemas de imagen para diagnóstico. El objetico de este análisis es, usualmente, el de diseñar una métrica capaz de evaluar la calidad de imagen percibida por un observador, una IQM por sus siglas en inglés (Image Quality Metric). Más aún, el objetivo de un gran número de investigadores es el de desarrollar métricas automatizadas capaces de reproducir los resultados que produciría un observador humano ante dichas imágenes. De forma prácticamente universal, estas métricas se desarrollan como programas informáticos, desarrollados en uno u otro lenguaje de programación. Hasta el momento solo se han obtenido éxitos parciales.

El número existente de aproximaciones a este problema y, por tanto, el número de algoritmos desarrollado es elevado; sin embargo, sigue siendo una cuestión abierta. En la literatura médica se encuentran dos aproximaciones claramente diferenciadas; una de ellas está basada en modelos de la función visual humana o en modelos ideales de observador (bien juntos o por separado). Estos modelos tratan de reproducir el procesado de la imagen en el observador desde su captación en el ojo hasta su procesado de alto nivel en el cerebro. Son modelos muy complejos, con una validez limitada y no han mostrado respuestas satisfactorias y, sobre todo, generalizables. Son estudios y modelos típicos en el campo de la imagen médica.

Por otro lado, los especialistas del mundo de las Telecomunicaciones han analizado la calidad de imagen desde un punto de vista más amplio, más enfocado en estudios de imágenes naturales (aquellas presentes en el entorno natural humano), y tanto en estudios de imagen fija como en vídeo. Muchos de estos análisis están basados en aproximaciones “top-down” al sistema visual humano. Estos modelos proponen hipótesis de carácter general acerca del funcionamiento del sistema visual humano y construyen modelos del mismo basándose en dichas hipótesis. Algunos de estos estudios han propuesto y desarrollado métricas muy bien correlacionadas con la percepción humana. Es quizá sorprendente que, hasta hace unos años, ha habido muy pocos estudios sobre la aplicación de estas métricas al campo de la imagen médica.

Dentro de este acercamiento, la métrica que ha tenido más éxito ha sido, sin ningún género de dudas, SSIM, presentada por Wang, Bovik y Simoncelli en el año 2004. Esta métrica se basa en la teoría propuesta por Wang y Bovik sobre el funcionamiento del sistema visual humano. Esta teoría afirma que nuestro sistema visual está especialmente adaptado para extraer información estructural de una imagen. Es una aproximación en la que se parte de una teoría del funcionamiento general del sistema visual humano, en lugar de deducir un esquema de funcionamiento a partir de sus elementos funcionales. A partir de esta métrica se ha desarrollado una amplia familia de índices que comparte la estructura básica con SSIM y que ha obtenido correlaciones crecientes entre los resultados de dichas métricas y los resultados del observador humano. Actualmente es la métrica más usada para medir la calidad de imagen percibida en la industria del vídeo por cable y por satélite.

Uno de los más prometedores miembros de esta familia ha sido el índice de correlación cruzada multiescala de SSIM, denotado como R* por sus autores, Rose y Hemami, en 2008. Su diseño se enfocó al problema de la percepción cerca del límite de visibilidad. Esta tarea tiene una gran

importancia en el análisis de las imágenes de maniquíes médicos y, en general, en el campo de la imagen médica.

No ha sido la única investigación que ha intentado mejorar los resultados de SSIM. Entre los muchos acercamientos que se han realizado, tenemos que destacar los tres siguientes:

a) Los basados en el análisis de gradientes de la imagen
b) Los basados en el análisis de las texturas de las imágenes
c) Los que simulan diferentes distancias de visualización

Estas tres aproximaciones, junto con R*, han mostrado muy buenos comportamientos en entornos de imagen ruidosos y desenfocados. Todas ellas se han probado con imágenes naturales (imágenes de nuestro entorno), pero hasta donde llega nuestro conocimiento, nunca se han analizado las posibles combinaciones de los cuatro acercamientos, ni con imágenes naturales ni con imágenes médicas.

Esta tesis ha analizado el comportamiento de las distintas métricas señaladas y ha propuesto nuevas métricas que combinan los cuatro acercamientos citados. El entorno de estudio ha sido el de imágenes médicas, en particular el de imágenes radiológicas. Los pasos que hemos seguido se describen a continuación.

Herramientas utilizadas

Al comienzo de esta investigación, nos encontramos con más de 250 herramientas capaces de visualizar, editar o extraer información de imágenes médicas. Realizamos una investigación exhaustiva, buscando un editor de imágenes DICOM que permitiera la modificación de dichas imágenes. Las modificaciones que preveíamos incluían, pero no se limitaban a, inserción de artefactos, de fondos anatómicos, patologías, diferentes tipos de ruido, etc. En esta fase inicial del proyecto buscábamos una herramienta software capaz de generar una base de datos de imágenes originales y modificadas para llevar a cabo experimentos de percepción. La herramienta seleccionada fue ImageJ, de Wayne Rasband. Esta herramienta permitía el control de píxeles individuales mediante programación Java, lo que si bien era complejo, nos permitía la mayor potencia posible de manipulación. Esta herramienta ha sido el elemento central de todos nuestros algoritmos, todos ellos desarrollados en Java y como plugins de ImageJ. Estos algoritmos han sido los que hemos aplicado para manipular y modificar nuestra base de datos de imágenes a lo largo de la investigación.

Análisis de imágenes de maniquíes: el problema de la memoria del observador humano

Se han diseñado varios maniquíes para el estudio de la calidad de imagen mamográfica: ACR TOR(MAM), CDMAM, etc. La tarea asociada a estos maniquíes es percibir, en las imágenes obtenidas de ellos con sistemas de mamografía, el contraste mínimo (contraste umbral) para cada diámetro de una serie de discos con distintos espesores insertados en el maniquí. Normalmente los discos están insertados en posiciones bien conocidas y la evaluación se basa en el paradigma SKE. La principal ventaja que presenta específicamente el maniquí CDMAM es el elevado número de celdas con discos (205) y que en cada una de ellas el disco test puede estar en una de cuatro posiciones distintas. El disco de referencia siempre se encuentra en el centro

de cada celda. Sin embargo, ya que un grupo de discos siempre se ve en la imagen y otro grupo de discos nunca se ve, el procedimiento de evaluación se restringe a examinar un pequeño número de celdas que incluyen los discos más críticos de percibir. Además, las tolerancias establecidas en algunos protocolos hacen que se reduzca todavía más el número de celdas que se evalúan. En consecuencia, el número final de celdas evaluadas puede ser muy pequeño y el operador puede memorizar perfectamente la posición esperada del disco en esas celdas. Este efecto memoria puede introducir una distorsión significativa en el proceso de evaluación.

En este punto del proyecto, el equipo desarrolló una herramienta software para mejorar las características del maniquí CDMAM. Dada una imagen digital del CDMAM, el programa cambia de forma automática la posición del disco de una a otra esquina de cada una de las celdas del CDMAM. Se puede seleccionar un ángulo fijo de giro o uno aleatorio, lo que hace imposible que un observador pueda usar su memoria para prever la posición final del disco en cada una de las celdas.

Se probaron dos algoritmos, ambos con éxito. Se hicieron análisis ROC partir de las respuestas de 36 observadores y los resultados concluyeron que las imágenes originales eran estadísticamente indistinguibles de las modificadas por nuestro programa. El área ROC fue de 0,507 ± 0,024 para el primer algoritmo y de 0,522 ± 0,026 para el segundo.

Este primer desarrollo nos permitió un acercamiento a entornos experimentales de percepción de calidad de imagen médica y fue un primer paso para desarrollar algoritmos complejos en Java.

El problema de la percepción automatizada

Desarrollamos un segundo trabajo con la intención de analizar el potencial de la métrica R*, de la familia SSIM, desarrollado por Rouse y Hemami, para predecir las prestaciones de un observador humano en tareas de detección de detalles, tareas muy cercanas a la diagnosis en rayos-x para determinadas técnicas de adquisición. Específicamente, se comprobó la efectividad de R* aplicando esta métrica a una tarea de detección contraste-detalle.

El umbral de contraste visible para R* se determinó con dos conjuntos de imágenes (conjunto 1 y conjunto 2) de un maniquí CDMAM. Los resultados de R* y de los observadores humanos se compararon con el método de Eficiencia Constante. Además, se comparó la respuesta de R* y de otros dos algoritmos usados de forma habitual para evaluar automáticamente imágenes radiográficas del maniquí CDMAM.

El estudio demostró que los resultados de un observador humano y de R* eran muy similares. Los coeficientes de correlación lineal entre R* y los observadores humanos en el conjunto 1 fueron de 9,84, con desviaciones medias del 16% para todas las imágenes. En el conjunto 2, el coeficiente de correlación encontrado era de 0,984 y desviaciones medias menores del 11%.

Estos resultados mostraron que R* podía usarse para mimetizar a los observadores humanos en ciertas tareas, como la determinación de curvas contraste-detalle en presencia de fondos uniformes. Además, el algoritmo que se diseñó basado en R*, mejoraba de forma significativa otras métricas y algoritmos usados en ese momento para evaluar imágenes del maniquí CDMAM.

También estos resultados afianzaron la posibilidad de aplicar la métrica R* al área de investigación de imagen médica, aplicando las adecuadas metodologías y condiciones experimentales.

Percepción en fondos uniformes frente a percepción en fondos anatómicos

La percepción humana de pequeños detalles de interés en imagen médica cambia en presencia de fondos anatómicos con estructuras. Hay pequeños, pero relevantes, detalles de imagen médica que quedan enmascarados por la presencia de estructuras anatómicas del paciente. Normalmente estos enmascaramientos son debidos a variaciones de la intensidad del detalle de interés por el citado fondo o bien, otras veces, debido al tamaño similar que presentan las estructuras anatómicas y los elementos relevantes.

Para analizar este problema, el equipo de esta investigación encontró de interés comparar la respuesta de un observador automatizado y de un observador humano ante el mismo conjunto de señales, bien insertadas en fondos mamográficos o bien insertadas en fondos uniformes. Esta comparación buscaba dos objetivos: el primero era analizar el umbral de reconocimiento de señal en función del fondo; el segundo era comparar la respuesta de un observador humano frente a un observador automatizado analizando el mismo conjunto de señales, bien embebido en un fondo mamográfico o bien embebido en un fondo uniforme.

Es bien conocido por la praxis médica y por la literatura científica que las características estructurales locales en fondos mamográficos aumentan el umbral de detección en tareas contraste-detalle. En consecuencia, resulta difícil, cuando no directamente imposible, inferir el límite de detección en los citados fondos mamográficos a partir de los datos obtenidos del análisis de imágenes de fondo plano, como las que producen las adquisiciones radiográficas del maniquí CDMAM.

Se desarrolló una herramienta software capaz de mezclar imágenes obtenidas de un CDMAM con imágenes de fondos mamográficos reales. Esto permitió desarrollar tareas SKE para ambos tipos de fondo y para observadores humanos y automatizados. La herramienta usaba imágenes CDMAM y las mezclaba con regiones de interés tomadas de mamografías reales. Tanto la región como el método de mezcla (sumando o multiplicando píxeles) se podía seleccionar por parte del operador. En la presente investigación se analizó un conjunto de 8 imágenes y se comparó la respuesta entre el observador humano y la métrica R*, ya analizada en un trabajo anterior.

Se observaron cuatro hechos relevantes:

a) La baja respuesta del observador humano en fondos mamográficos frente a la respuesta en fondos uniformes, debido al ruido estructurado que presentan los primeros. Este efecto está ampliamente señalado en la literatura médica.
b) El segundo hecho fue constatar la baja respuesta de R* en fondos mamográficos en relación con la respuesta en fondos uniformes. Este efecto también era esperable debido al citado ruido estructural del fondo mamográfico.
c) El tercer hecho fue que el umbral de contraste detectable aumentaba a medida que el tamaño de los discos crecía. Este efecto solo se manifestaba en diámetros superiores a 1 mm. Para diámetros inferiores, como era de esperar, al disminuir el diámetro crecía el umbral de detección. Este relativamente sorprendente tercer efecto ya ha sido

ampliamente descrito en la literatura y se debe a que las estructuras del fondo mamográfico tienen un tamaño similar al de los discos más grandes, por lo que se produce un efecto de enmascaramiento que puede parecer, en un primer acercamiento, paradójico.

d) El cuarto hecho fue constatar que la respuesta de la métrica R* y del observador humano eran similares.

La aplicación de la herramienta software a un problema de creación de imágenes fusionadas y para la comparación de la respuesta del observador humano y de la métrica R* mostró y reafirmó el potencial de esta familia de desarrollos para analizar el comportamiento de distintos tipos de observadores.

Ampliando el problema: diferentes tipos de imágenes y diferentes tipos de ruido

Al principio de este resumen describimos la familia de métricas SSIM. Como se ha comentado, estudios muy extensos de análisis de calidad percibida de imagen han demostrado que tanto SSIM como MS-SSIM mimetizan bastante bien la calidad de imagen percibida por un observador humano. Sin embargo, estas métricas muestran algunas limitaciones.

Algunos investigadores han encontrado que SSIM y MS-SSIM tienen un comportamiento errático con respecto al del observador humano en tareas de reconocimiento cercanas al límite de visibilidad de la señal. Este comportamiento invalidaría estas métricas para análisis de imágenes con regiones de interés cerca del límite de visibilidad, extremo muy usual en imágenes médicas radiológicas.

Algunos estudios muestran límites en el comportamiento de estas métricas analizando imágenes médicas. Otros estudios muestran que la correlación observador humano / métrica disminuye cuando se mide la calidad con imágenes ruidosas o desenfocadas.

Estos inconvenientes son factores que limitan la aplicación de estas métricas en el campo de la Imagen Médica y, en especial, en el de la Radiología. Hay imágenes radiológicas de interés en Medicina que muestran diferencias muy tenues entre un tejido sano y un tejido con determinadas patologías. Por otro lado, el ruido en la imagen y su desenfoque son factores de distorsión muy usuales en la experiencia diaria de un radiólogo.

Como ya se ha señalado, algunos autores han propuesto diversas modificaciones a las métricas SSIM y MS-SSIM para evitar las comentadas limitaciones. Rose y Hemami propusieron en 2009 un nuevo índice, R*, basado en el componente estructural de MS-SSIM. Postularon y probaron en una serie de experimentos que este componente podía evitar la falta de efectividad de MS-SSIM cerca del límite de visibilidad. También los autores de este documento han realizado pruebas extensas (citadas en la bibliografía) de este índice.

Chen y otros propusieron en 2006 una variación de SSIM, GSSIM, basada en análisis de gradientes de luz. Esta nueva métrica mejoró los resultados de SSIM con imágenes ruidosas y desenfocadas. Li y Bovik aplicaron en el año 2010 un modelo de cuatro componentes (4) basado en el análisis de texturas de las superficies y de los bordes entre regiones. Aplicaron este nuevo modelo a SSIM y MS-SSIM, junto con el citado análisis de gradiente desarrollado por Chen y otros. Este análisis multidimensional dio lugar a 8 nuevas métricas que han mostrado

características muy prometedoras y que eliminan las principales limitaciones de SSIM y MS-SSIM.

La intención de la última fase de la presente investigación fue analizar las citadas aproximaciones de forma simultánea y en todas las combinaciones posibles y comprobar su respuesta en un entorno médico. Esta aproximación dio lugar a un conjunto de 16 métricas, 8 ya conocidas y otras 8 desarrolladas por los autores de esta investigación.

Para comprobar la efectividad de estas nuevas IQM, hemos aplicado estas métricas a una tarea de doble estímulo sobre una base de datos de imágenes médicas radiológicas. Esta base de datos comprende distintas técnicas de adquisición (MR, placas simples de tórax, de extremidades, etc.). A su vez, estas imágenes fueron modificadas con cuatro tipos de distorsión: desenfoque gaussiano, ruido gaussiano, compresión JPG y compresión JG2000. A cada tipo de distorsión se le aplicaron cinco grados distintos de distorsión. Las imágenes fueron analizadas por un conjunto de radiólogos expertos con una metodología de doble estímulo y sus resultados se compararon con los obtenidos como resultado de aplicar las 16 métricas comentadas al mismo conjunto de imágenes.

Nuestros resultados experimentales mostraron que las lecturas humanas eran sensibles a la información de bordes relativa entre la imagen de referencia y las imágenes test. También esas mismas lecturas humanas mostraban sensibilidad particular ante los cambios de textura de una a otra imagen, así como a los cambios de zonas de borde a zonas lisas y viceversa.

Tenemos que llamar la atención sobre que ciertos estudios anteriores habían mostrado que el hecho de simular diferentes distancias de visualización entre el observador y la imagen (aproximaciones multiescala) tenían un efecto muy pequeño en la percepción de las señales. Por el contrario, nosotros hemos encontrado que esta aproximación multiescala es muy relevante, sobre todo en imágenes grandes (de alta resolución).

Nuestros resultados mostraron que varias métricas (4-G-SSIM, 4-MS-G-SSIM, 4-G-r* y 4-MS-G-r*) pueden usarse como subrogados de un radiólogo en tareas de análisis de calidad de la imagen médica. Especialmente, 4-MS-G-SSIM presenta unas excelentes correlaciones con los observadores humanos para todo tipo de imágenes y para todo tipo y grado de distorsión.

Conclusiones

1. Se pueden utilizar programas de manipulación de imagen para evitar los efectos de memoria de los observadores experimentados en la evaluación de imágenes de maniquíes de contraste-detalle, en particular el maniquí CDMAM, de amplio uso en mamografía. En particular, el programa desarrollado en este trabajo, puede ser utilizado por la comunidad científica para evaluar el citado efecto memoria.

2. Existen métricas de calidad de imagen provenientes de campos distintos del de la imagen médica, como las de la familia SSIM, que han sido muy poco estudiadas y aplicadas a problemas relacionados con este área. Nuestros estudios muestran que una de ellas, R*, puede ser utilizada para la evaluación automática de maniquíes.

3. Se pueden crear herramientas para generar imágenes híbridas de maniquíes y fondos anatómicos reales en mamografía. Estas imágenes se pueden utilizar para analizar la respuesta de un ser humano o de un sistema automatizado que aplica una IQM. Los resultados de la métrica elegida, resultan ser similares a aquellos obtenidos por el observador humano.

4. Existen métricas cuyos resultados analizando la calidad de una imagen médica tienen un comportamiento muy similar a los de un observador humano experto. Estos experimentos se han llevado a cabo sobre un amplio abanico de imágenes médicas y sobre un amplio conjunto de tipos de ruido relevantes en imagen médica.

Por último, queremos compartir nuestros resultados con nuestros colegas científicos y académicos. Todos los programas y algoritmos desarrollados a lo largo de esta investigación serán publicados en nuestra web (https://www.ucm.es/gabriel_prieto) al finalizar la lectura de esta tesis. Una parte de estos algoritmos ya está libremente disponible en la citada dirección web para la comunidad y han sido utilizados por varios grupos de investigación para el desarrollo de experimentos de percepción de calidad de imagen.

3. Introducción

La investigación que se presenta en esta tesis doctoral se centra en el problema de la percepción automatizada de la calidad de imagen y en su correlación con la percepción humana de esa misma calidad de imagen en el ámbito de la imagen médica.

El análisis de la calidad de imagen tiene un papel central en el desarrollo de sistemas de imagen médica. Usualmente el objetivo final de estos estudios es diseñar una métrica capaz de evaluar la calidad de imagen percibida por el observador, una métrica de calidad de imagen (IQM). Más aún, el objetivo de un gran número de investigaciones es desarrollar una métrica automatizada que, analizando la imagen por medio de programas informáticos, sea capaz de predecir los resultados producidos por un observador humano.

El número de algoritmos y aproximaciones que se ha utilizado es elevado y el problema es todavía una cuestión completamente abierta. Llamativamente, en la literatura científica existen dos aproximaciones claramente diferenciadas; la primera de ellas está basada en modelos de la función visual humana o bien en modelos de observadores ideales mezclados o no con la anterior aproximación. Son modelos que intentan reproducir el camino seguido por la imagen desde el ojo hasta los centros de percepción, recreando en algunos casos la respuesta neuronal o del propio sistema visual completo. Estos sistemas suelen ser altamente complejos, válidos normalmente en condiciones restrictivas y no han proporcionado modelos generalizables y satisfactorios.

La segunda aproximación se basa en modelos menos analíticos, con acercamientos de tipo “top-down” al funcionamiento del sistema visual humano. Algunos de estos acercamientos, como los algoritmos de la familia SSIM, han tenido éxitos enormes en el campo de las Telecomunicaciones, particularmente en la transmisión de imagen estática y de vídeo. Sorprendentemente, se ha hecho un uso muy reducido de estos algoritmos en el campo de la imagen médica. En este estudio se analizan las aplicaciones de esta familia de IQM a este campo de análisis. Destacamos que este equipo de investigadores fue prácticamente pionero a nivel mundial en la aplicación de las métricas SSIM a la imagen médica.

El problema de la percepción automatizada de la calidad de imagen

La revolución digital en Radiodiagnóstico ha generado enormes cantidades de imágenes en soporte digital. Este soporte digital ha permitido la visualización y mejora de la imagen radiológica mediante programas de visualización y de tratamiento de imagen. Además, la especificidad de las imágenes médicas, con su cabecera DICOM, necesita de programas especializados capaces de explotar y gestionar esta información.

También es de interés para diversas investigaciones, como la que anima esta tesis, disponer de herramientas capaces de manipular las imágenes y modificarlas. Estas modificaciones en una investigación avanzada deben permitir controlar la imagen a nivel de píxel y no solo mediante controles genéricos de contrastes, exposición, luz, etc., que, si bien de extremada potencia para el radiólogo, y con capacidad de manipulación global o zonal, no permiten manipular una imagen hasta el último detalle.

El problema con el que se enfrenta cualquier investigador en el campo del tratamiento digital de la imagen radiológica es el enorme número de programas distintos capaces de visualizar, manipular y extraer datos de este tipo de imágenes. En el momento de iniciarse esta investigación, en el año 2007, el número rondaba los 250.

Al comenzar esta tesis, se realizó una investigación exhaustiva del mejor programa para manipulación de imágenes médicas en ese momento (Prieto, et al., 2007). En ese momento se busca un visualizador y manipulador de imágenes DICOM que permita modificar una imagen e insertar en ella estructuras de interés radiológico, como lesiones, fondos anatómicos y sintéticos, artefactos, etc. El proyecto busca el mejor software de manipulación de imágenes para poder disponer de una base de datos de imágenes originales y modificadas, que permita realizar estudios de percepción de imagen, en un futuro tanto automatizada como humana. El programa elegido fue ImageJ, de Wayne Rasband (Rasband, 1997-2015), que sirvió como base para todos los programas, desarrollados en Java, utilizados para la manipulación de imágenes durante la presente investigación.

3.1. El problema de la memoria en la percepción de imagen de maniquíes por observadores humanos

El uso de maniquíes está ampliamente extendido en radiología, principalmente para el tratamiento de la calidad de imagen: TG-18, XCAT, ACR, TOR(MAM), CDMAM (estos tres últimos en el campo de la mamografía), etc. El procedimiento de uso es similar: ajustados determinados parámetros de control del sistema que se desea analizar, se toman imágenes radiológicas de los maniquíes. Posteriormente un observador (usualmente humano) analiza estas imágenes buscando el detalle de mínimo contraste o de mínima resolución que es capaz de ver, o bien analiza la distorsión de determinados elementos geométricos. Estos niveles mínimos de resolución, contraste o de otros parámetros se utilizan como indicadores de la calidad de imagen.

Un problema que aparece es que los usuarios que analizan el sistema de adquisición de imágenes utilizan básicamente los mismos maniquíes y, usualmente, el mismo modelo. Esto produce un efecto memoria en todos los observadores. El observador busca en determinadas posiciones los elementos mínimos que ya conoce que se van a encontrar en ese punto y analiza en detalle si el elemento es visible o no y si cumple las características de percepción deseadas. Esta memoria de localización induce un considerable efecto de distorsión sobre la lectura de la imagen del maniquí, ya que el usuario focaliza su atención en determinadas zonas de la imagen radiológica y dedica menos atención a otras zonas.

El maniquí CDMAN (Bijkerk, et al., 1993) (Bijkerk, et al., 2000a) (Bijkerk, et al., 2000b), maniquí contraste-detalle ampliamente usado en mamografía, contiene una matriz de 205 celdas cuadradas. Dentro de cada celda, en posiciones aleatorias, se colocan discos de oro de distinto diámetro y espesor (**Figura 2.a del trabajo II**). La posición del disco encontrada por el observador humano se compara con la posición real del disco en el maniquí. El porcentaje de aciertos y la visualización o no de los discos de mínimo espesor y diámetro y su posterior tratamiento estadístico, proporcionan ciertos parámetros de calidad de imagen.

El **trabajo I**, para evitar el comentado efecto memoria, crea una herramienta que cambia la posición de la imagen de los discos dentro de cada una de las celdas, colocándolos en posiciones

aleatorias o bien gira esas posiciones un número de grados determinado por el operador, en múltiplos de $\pi/2$ radianes. También presenta dos sistemas de giro: en el primero de ellos utiliza toda la celda como una unidad de imagen y la rota mediante el sistema elegido (aleatorio o múltiplo de $\pi/2$ radianes). En el segundo sistema extrae cuatro pequeños cuadrados de cada una de las esquinas de cada celda y las intercambia entre sí con el sistema seleccionado. Para averiguar si el efecto de manipulación era perceptible, se realizaron experimentos con 36 observadores, analizando las curvas ROC de respuesta de los mismos.

3.2. El problema de la percepción automatizada

El análisis de la calidad de imagen tiene un papel central en el diseño de sistemas de visualización de imagen para diagnosis clínica. Se ha realizado un enorme esfuerzo por desarrollar métricas correlacionadas con los estudios de imágenes obtenidas desde maniquíes o, directamente, con las prestaciones de diagnosis clínica de los sistemas de imagen médica. El objetivo último de estas métricas de calidad de imagen (IQM) es desarrollar un algoritmo capaz de evaluar y mimetizar la calidad percibida por un observador de una imagen diagnóstica.

En estudios de maniquíes, el uso de herramientas automatizadas que simulan el punto de vista del radiólogo podría evitar la variabilidad inter e intraobservador. También podría minimizar el gran número de imágenes y de observaciones, y el consiguiente gasto de tiempo, necesario para optimizar los parámetros de adquisición de imagen, para evaluar nuevos equipos o, incluso, para evaluar nuevas tecnologías. Hasta ahora solo se han conseguido éxitos parciales. La búsqueda de una IQM que esté completamente correlacionada con el sistema visual humano (HVS) y, particularmente, con el punto de vista del observador es todavía una cuestión abierta.

Ciertas métricas muy usadas, como la relación de pico entre la señal y el ruido (PSNR) o el error cuadrático medio (MSE) entre píxeles de dos imágenes son muy sencillas de calcular, pero no muestran buenas correlaciones con la calidad percibida por observadores humanos (Girod, 1993) ni son útiles para deducir la capacidad diagnóstica de un equipo de imagen médica (Burgess, 1999).

Existen otras métricas más relacionadas con las prestaciones físicas del sistema, como la función de modulación de transferencia (MTF), el espectro de ruido (NPS), el ruido equivalente cuántico (NEQ) y la eficiencia de detección cuántica (DQE). Todas ellas describen mucho mejor, desde una aproximación física, la formación de imagen y pueden ser usadas para predecir la respuesta del observador bajo el Modelo de Observador Ideal (Myers, 2000). Sin embargo, este modelo solo puede aplicarse a entornos donde la señal y el fondo deben ser conocidos de manera exacta o, al menos, de manera estadística (SKE/BKE - SKS/BKS) (Barrett, et al., 1986).

Existen otros modelos que tienen buena correlación con el observador humano y pueden ser aplicados a tareas más complejas que las SKE/BKE – SKS/BKS. Principalmente engloban el modelo canalizado de Fisher-Hotelling, el NonPreWhitening Matched Filter (NPWMF) y el Non-PreWhitening con un filtro de ojo (Eckstein, et al., 2000). Estos modelos son bastante útiles en la evaluación de la calidad de imagen para ciertos tipos de adquisición de imagen y para determinados tipos de ruido (Eckstein, et al., 2000).

Sin embargo, el uso de una IQM que pueda ser empleada de modo general y para distintos tipos de ruido sigue sin estar resuelta en el ámbito de la imagen médica.

Paralelamente a los estudios realizados en imagen médica, durante años se han realizado estudios por parte de especialistas del mundo de la imagen en Telecomunicaciones que han analizado el problema de la calidad de imagen desde un punto de vista más general, más enfocado a la calidad de imagen natural (aquellas de nuestro entorno), tanto en imagen fija como en vídeo. Estos estudios han llevado a la propuesta de métricas que correlacionan de forma bastante sólida con la percepción humana. Sin duda, la métrica más destacable de los últimos años ha sido el índice de similaridad estructural (SSIM) (Wang, et al., 2004). Esta métrica se basa en la teoría propuesta por Wang y Bovik (Wang & Bovik, 2002) que considera que el sistema visual humano (HVS) está muy adaptado para extraer información estructural de una imagen. Una amplia familia de métricas basada en SSIM se ha desarrollado a lo largo de los últimos años (Wang, et al., 2004) (Wang, et al., 2003) (Rouse & Hemami, 2009), etc. con prestaciones y correlaciones crecientes con el HVS. Uno de los elementos de esta familia es el coeficiente de correlación cruzada multiescala SSIM (R*) (Rouse & Hemami, 2009). Su diseño fue específico para extender las funcionalidades de SSIM a tareas cercanas al límite de percepción. Este tipo de tareas son de enorme interés en el análisis de imágenes de maniquíes y, en general, en el campo de la diagnosis médica.

El **trabajo II** aplica por primera vez en la literatura académica (hasta la fecha de su publicación y hasta donde nosotros hemos podido averiguar) una IQM de la familia SSIM al problema de la percepción de la imagen médica.

En este trabajo se analiza el comportamiento de la métrica R* frente al observador humano en una tarea de percepción contraste-detalle: la detección de discos en el maniquí CDMAM (Bijkerk, et al., 2000a). Esta tarea se halla, en su punto crítico, cerca del límite de percepción, ya que es la visión de los discos más pequeños o de menor espesor la que determina en última instancia la calidad de la imagen médica en aspectos de contraste-detalle. Así mismo, los resultados de R* se compararon con dos métricas ampliamente usadas para la evaluación de las imágenes CDMAM.

3.3. El problema de la percepción en fondos uniformes frente a fondos anatómicos reales

Es bien conocido que las características locales de las estructuras anatómicas en imágenes médicas reducen la percepción del observador humano de estructuras pequeñas y apenas perceptibles. Este hecho es especialmente cierto en mamografía (Grossjean & Muller, 2006) (Burgess, et al., 2001), debido a la complejidad del fondo anatómico y el pequeño tamaño de ciertas características relevantes como las microcalcificaciones, o el escaso contraste con el tejido circundante de ciertas estructuras tumorales. Por ello, las prestaciones del equipo mamográfico no pueden ser completamente extrapoladas a partir de los datos adquiridos de un maniquí como el CDMAM con un fondo uniforme, caracterizado por ruido blanco y con distribución uniforme.

Basándonos en estas premisas, resulta de interés comparar la respuesta de un sistema de imagen médica frente al mismo conjunto de señales, bien insertadas en un fondo plano como el de un maniquí o bien insertadas en un fondo anatómico real. Esta comparación tiene dos objetivos. El primero es analizar la variación de la respuesta del observador humano en uno u otro fondo. El segundo es comparar las prestaciones de un observador humano frente a una IQM, para averiguar la validez de esta última.

En el **trabajo III** se creó una herramienta capaz de mezclar fondos mamográficos reales con las imágenes de fondo uniforme obtenidas de un CDMAM. En estas imágenes estaban presentes las imágenes de los discos utilizados para la tarea contraste-detalle y, por tanto, esta tarea de percepción se podía lleva a cabo bien con el fondo mamográfico real o bien con el fondo uniforme del propio maniquí.

También se comparó la respuesta del observador humano con la de la métrica R*, aplicada en anteriores trabajos. El estudio produjo tres resultados:

a) Puso de manifiesto la baja respuesta del observador humano cuando se introducían fondos anatómicos reales, debido al ruido estructurado presente en la imagen. Este efecto ya se había descrito ampliamente en la literatura científica (Grossjean & Muller, 2006) (Burgess, et al., 2001) (Burgess, 2001).
b) Destacó la baja respuesta de la métrica R* en fondos mamográficos. Así mismo es destacable que la respuesta de la métrica R* era similar a la del observador humano para diámetros de disco menores o iguales a 1,25 mm.
c) Mostró cómo el observador humano mejoraba su respuesta, tanto de forma absoluta como con referencia a la métrica R*, con discos de diámetro superior a 1,25 mm. Este efecto ha sido encontrado por otros autores (Burgess, 2001) y compartimos con ellos la explicación de que para determinados tamaños de discos, la percepción humana se basa más en la percepción del borde del disco que en el contraste entre el disco y el fondo de la imagen, ya que para estos tamaños la percepción del borde era más manifiesta para todos los observadores.

3.4. El problema de los distintos tipos de ruido y los distintos tipos de imagen radiológica

Los tipos de ruido presentes en cada técnica de adquisición de imagen radiológica son de fuentes y características variadas: aparece ruido blanco, ruidos de reconstrucción en TC, desenfoque, etc. (Williams, et al., 2007) También existen ruidos de compresión para determinadas aplicaciones médicas, como la telemedicina, en la que es necesario aplicar sistemas de compresión JPEG o JP2000 para transmitir las imágenes (Johnson, et al., 2010) (Krupinski, et al., 2007) (Loose, et al., 2009). También la Sociedad Europea de Radiología (European Society of Radiology -ESR-, 2011) ha analizado y dado pautas para aplicar compresión irreversible a determinadas imágenes médicas, analizando su impacto en la diagnosis. A todas ellas, se une el ruido estructurado introducido por la propia imagen anatómica que limita la percepción de estructuras de interés en el diagnóstico.

El uso de maniquíes limita los tipos de ruido presentes en una imagen para el análisis de la calidad de la imagen, fundamentalmente el ruido estructurado. A este factor de variabilidad del ruido, se une la variabilidad de las técnicas de adquisición de imagen y la enorme diferencia en la tipología de las imágenes de salida. Así, en RM disponemos de series de imágenes de salida con resoluciones usuales de 512x512 píxeles. En radiografías de tórax, las imágenes de salida típicas son de 2500x2000 píxeles.

Como se ha comentado, el uso de una IQM que pueda correlacionar con el observador humano y, en particular, con el especialista médico, es de interés para automatizar procedimientos de optimización de imagen. Se han realizado muy pocos estudios del funcionamiento de una IQM sobre un alto grado de variabilidad de tipología de imágenes médicas y de ruidos. En general,

dichos estudios evalúan el funcionamiento de determinadas IQM con determinadas imágenes y con determinados tipos de ruidos, “sintonizando” la IQM para las especiales características de cada combinación.

EL **trabajo IV** presenta un análisis de un conjunto extenso de IQM perteneciente a la familia SSIM desarrollado durante los últimos años (Wang, et al., 2004) (Rouse & Hemami, 2009) (Prieto, et al., 2011) (Chen, et al., 2006) (Li & Bovik, 2010). Estas IQM han mostrado mejoras frente a otros miembros de la familia SSIM, especialmente en estudios de percepción en los que estaban presentes diversos tipos de ruido similares a los descritos en imágenes radiológicas. También estas nuevas IQM han introducido el efecto de percepción mejorada que presenta el observador humano en la presencia de bordes de fuerte contraste, en línea con los resultados obtenidos por nuestro grupo en el **trabajo III**. El trabajo extendió las pruebas de este conjunto de IQM ya desarrolladas a otras nuevas y originales, propuestas por el propio grupo de trabajo.

El conjunto de estas 16 IQM se aplicó a un amplio grupo de imágenes obtenidas con distintas modalidades de adquisición o de distinto tipo estructural: RM, CPF y BPF. También se aplicaron distintos tipos de ruido usuales en imagen médica o de interés para determinadas investigaciones: ruido blanco, desenfoque, compresión JPEG y compresión JP2000 (European Society of Radiology -ESR-, 2011) (Johnson, et al., 2010) (Burgess, 2001) (Williams, et al., 2007) (Krupinski, et al., 2007) (Loose, et al., 2009), todos ellos con distintos grados de intensidad.

Este conjunto de imágenes se evaluó por un grupo de expertos radiólogos, preguntando específicamente la utilidad médica de las distintas imágenes, modificadas con un tipo u otro de distorsión y con distintos grados de intensidad. Las respuestas de los expertos fueron comparadas con aquellas obtenidas por las distintas IQM.

Los resultados del experimento mostraron un subconjunto de cuatro IQM y, entre ellas, especialmente dos, que tuvieron excelentes correlaciones con los observadores humanos para todo tipo de distorsión y de adquisición. Notoriamente, destacaron aquellas métricas que insertaban y ponderaban en su algoritmo tres componentes:

- El análisis de bordes.
- El análisis multiescala, entendido como la simulación de la visualización de la imagen a distintas distancias por parte de un observador.
- El análisis de texturas.

4. Objetivos

El objetivo de este estudio es evaluar métricas ya conocidas de la familia SSIM y proponer otras nuevas de la misma familia desarrolladas por el equipo de trabajo, todo ello aplicado al campo de la imagen médica, tanto en estudios de maniquíes en mamografía, como en estudios generales de calidad de imagen radiológica real.

El objetivo global se ha abordado mediante aspectos específicos y de creciente acercamiento al objetivo global en cada uno de los artículos que componen este trabajo. Este acercamiento se resume en:

- Una manipulación de las imágenes de un maniquí contraste-detalle para evitar los efectos de memoria presentes en la evaluación de dicho maniquí por usuarios experimentados.

- Una comparación entre los rendimientos perceptuales humanos en la evaluación de un maniquí contraste-detalle frente a los rendimientos de una IQM automatizada no aplicada anteriormente a imágenes médicas.

- Una comparación entre los rendimientos humanos y de una IQM automatizada en imágenes de maniquíes usuales y en otras sintetizadas en las que se modificaban los fondos de dichos maniquíes por fondos anatómicos reales.

- Un análisis comparativo entre observadores humanos y 16 IQM (8 existentes en la literatura científica y 8 nuevas propuestas por este equipo de investigación) sobre un conjunto de imágenes médicas radiológicas de distintas modalidades (RM, BPF, CPF) y distorsionadas con distintos tipos de ruido (GB, GN, JPG, JP2000) y distintos niveles de los mismos.

5. Material y métodos

5.1. El problema de la memoria en la percepción de imagen de maniquíes por observadores humanos

Para la evaluación y manipulación de imágenes necesaria para realizar el **trabajo I**, se desarrollaron programas de tratamiento de imagen en Java como plugins dentro del entorno de desarrollo y visualización de ImageJ (Rasband, 1997-2015). Estos programas permitieron las manipulaciones descritas más adelante.

El maniquí CDMAM es un maniquí contraste-detalle ampliamente usado en mamografía. Está formado por una base de aluminio con discos de diámetro y espesor variables. El conjunto está ensamblado dentro de una carcasa de Plexiglás. Los discos están ordenados en 16 filas y 16 columnas, que forman una rejilla con 205 celdas (**Figura 2.a del trabajo II**). Dentro de una fila, el diámetro se mantiene constante, mientras que se incrementan en forma logarítmica los espesores. Dentro de cada columna, el espesor se mantiene constante, mientras que se incrementan de forma logarítmica los diámetros. Cada celda cuadrada formada en la intersección de las filas y las columnas, contiene dos discos de oro con el mismo diámetro y espesor. Uno de los discos, el de referencia, se sitúa en el centro de la celda; el otro, el disco de prueba que hay que buscar, se encuentra en una de las cuatro esquinas, a distancias conocidas de las esquinas de intersección de la rejilla y en puntos conocidos y determinados en el propio proceso de fabricación

Detección de la rejilla en la imagen. Para poder manejar las imágenes de los discos dentro de la imagen del maniquí, es necesario conocer de forma exacta la posición de los discos dentro de la imagen. El método usado en este trabajo se basa en el conocimiento previo completo de la geometría del CDMAM.

Seleccionamos una zona central del CDMAM y, dentro de ella, la primera columna de píxeles. Sobre esta columna seleccionamos en cada píxel un abanico de líneas de píxeles que barría una zona entre 35º y 45º, en pasos de un cuarto de grado. Sobre cada una de las líneas, sumamos la luminosidad de los píxeles. Valores máximos de esta suma indicaban que la línea escaneada por el algoritmo se superponía con una de las líneas de la rejilla del CDMAM y podíamos obtener el valor del ángulo de inclinación de las líneas de rejilla con pendiente negativa y el punto de comienzo de la línea de rejilla.

El algoritmo se repetía en la última columna del área seleccionada del CDMAM, en este caso barriendo ángulos entre 125º y 135º. De forma análoga al paso anterior, calculábamos la inclinación de las líneas de rejilla de pendiente positiva y su punto de comienzo.

Los dos valores de ángulo y punto de comienzo calculados en los dos lados de la región seleccionada del CDMAM nos proporcionan los valores del valor de la diagonal (D) de cada una de las celdas. Estos valores eran distintos en cada lado del maniquí, probablemente debido a la geometría del haz de rayos-x. Extrapolando estos datos hasta los extremos del maniquí obteníamos una imagen muy exacta de la posición de la rejilla.

Para asegurar los resultados, el programa escaneaba la imagen alrededor del punto de corte calculado de las líneas de rejilla con el borde de la imagen. Se examinaban +/- 10 píxeles alrededor de los puntos calculados y se repetía el proceso en los dos lados del CDMAM. La

distancia entre los puntos calculados y reales era de 0-1 píxeles. Solo en un número de casos menor del 1% la distancia fue igual o mayor que 2 píxeles.

Este algoritmo se mostró robusto. Se comprobó en 40 imágenes distintas captadas por diferentes sistemas de distintos fabricantes: (LORAD-HOLOGIC, GE MEDICAL SYSTEMS, AGFA, FUJI) y con distintos niveles de ruido. El índice de ruido, calculado como desviación estándar / valor medio del píxel medido en esquinas sin líneas de rejilla, tenía valores entre 0,010 y 0,025. Los ángulos de rejilla variaban entre 43º y 47º.

Manipulación de la imagen del maniquí. Para evitar el efecto memoria del observador, en el que recuerda las posiciones de los discos de test, desplazamos las posiciones de los discos de test dentro de la imagen. Usamos dos algoritmos diferentes, ilustrados en las **Figuras 2 y 3 del trabajo II**.

Cada celda tenía una forma real trapezoidal, no cuadrada, de forma que no podíamos usar el centro de cada celda como punto de rotación, ya que una manipulación de este tipo hubiera llevado a mostrar escalones en los puntos de encuentro de la rejilla. Para evitar este efecto, se trabajó en una zona segura de píxeles dentro de la celda que no tocaba la rejilla en sí.

El primer algoritmo, denominado "Un rombo", definía un cuadrado dentro de cada celda a la citada distancia segura (8 píxeles) y lo giraba en múltiplos de $\pi/2$ al azar en cada una de las celdas. El segundo método, denominado "cuatro de diamantes", definía cuatro pequeños cuadrados dentro de cada celda, centrado cada uno de ellos en la posición de los posibles centros de los discos del CDMAM y con un tamaño suficiente para englobar de forma segura el disco sin tocar la rejilla. Después, se intercambiaba la posición de estos discos dentro de cada celda y para todas las celdas del maniquí.

Para evaluar los resultados se llevó a cabo un experimento de tipo ROC. Para ello, se crearon dos conjuntos de imágenes. El conjunto 1 contenía 8 imágenes CDMAM obtenidas bajo distintos parámetros radiológicos. Sobre este conjunto 1 se aplicó el algoritmo "Un rombo" y se obtuvieron 8 imágenes del CDMAM modificadas con las posiciones de los discos colocadas al azar. El conjunto 2 se obtuvo de manera similar, pero aplicando el algoritmo "cuatro de diamantes".

Los dos conjuntos de imágenes se ordenaron internamente al azar y fueron examinados por dos grupos distintos de evaluadores. El primero estaba formado por 27 físicos médicos con una experiencia entre 2 y 10 años en control de calidad de equipos de radiología. El segundo grupo estaba compuesto por 9 estudiantes del Master de Física Médica de la Universidad Complutense de Madrid. Cada observador respondía a la pregunta "¿Cree usted que esta imagen ha sido modificada por un programa de ordenador en algún aspecto?". Téngase en cuenta que las imágenes, modificadas o no, estaban intercaladas al azar. Las imágenes fueron visualizadas en distintos monitores mediante el programa ImageJ. El test no tenía límite de tiempo de respuesta y los observadores podían utilizar cualquier herramienta de procesamiento de imagen incluida en ImageJ. La respuesta incluía un nivel de confianza de 0 a 4.

5.2. El problema de la percepción automatizada

La métrica R* (Rouse & Hemami, 2009) pertenece a un conjunto de IQM (la familia SSIM) que buscan la evaluación objetiva de la calidad de una imagen consistente con la apreciación subjetiva del observador humano. Estos algoritmos evalúan una imagen de prueba, X, con respecto a una imagen de referencia, Y, para cuantificar su similitud, evaluando esta similitud en una escala numérica. En este sentido, todas estas métricas se corresponden con tareas SKE, ya que la imagen de referencia siempre está disponible.

Procedemos a describir el funcionamiento interno de la métrica R*.

Sean X e Y imágenes que van a ser comparadas como matrices de píxeles y sea **x** = {x_i | i = 1, 2,…, N}, sea **y** = {y_i | i = 1, 2,…, N} pares de ventanas cuadradas de píxeles (procesadas como submatrices de píxeles) de X e Y respectivamente; **x** e **y** están localizadas en la misma zona en ambas imágenes. Se define el índice *r(x, y)* en términos de la desviación estándar entre píxeles σ_x and σ_y en las zonas **x** and **y,** y en función de la covarianza σ_{xy} de **x** e **y** como:

$$r(x,y) = (\sigma_{xy})/(\sigma_x \sigma_y)$$

Como se puede observar, si las submatrices **x** e **y** cubren el mismo objeto en la misma localización, r muestra un máximo.

r(x,y) toma valores entre -1 y 1. Cuanto más cercano sea el valor a 1, mayor es la similitud entre las submatrices **x** e **y**.

Cuando la señal, o bien la señal más el fondo son uniformes, σ_x o σ_y tienden a cero y el valor de r(x,y) se hace inestable. Este es el caso de submatrices evaluadas sobre señales de referencia uniformes, donde todos los píxeles toman el mismo valor y la varianza se hace cero. Para estos límites, se introducen excepciones en el cálculo del índice r(x,y). Son las siguientes:

a) Supongamos que $\sigma_x > 0$ y la submatriz **y** es uniforme. En estas condiciones, la submatriz **x** no es similar a la submatriz **y**, de forma que el valor de r(x,y) debe ser cero.
b) Cuando ambas submatrices tienen una varianza igual, el valor de r(x,y) deber ser 1, ya que las matrices son idénticas. Teniendo en cuenta estas excepciones, se redefine el índice *r(x,y)* como $r^*(x,y)$:

$$r^*(x,y) = \begin{cases} 0 \ \textit{para} \ \sigma_x > \sigma_y = 0, \ o \ \sigma_y > \sigma_x = 0 \\ 1 \ \textit{para} \quad \sigma_x = \sigma_y = 0 \\ r(x,y) \quad \textit{otros} \end{cases}$$

Puesto que esta métrica compara dos imágenes completas X e Y, las submatrices (**x, y**) se mueven a lo largo de X e Y calculando el valor de $r^*(x,y)$ para cada posición. El resultado final es el producto del valor de $r^*(x,y)$ en todas las posiciones.

La percepción de los detalles de una imagen depende, entre otros factores, de la resolución de la imagen y de la distancia de visualización a la misma (Wang, et al., 2003). El algoritmo propuesto incorpora M distintas distancias de visualización, simulando diferentes resoluciones

espaciales por reducción iterativa del tamaño de la imagen. Esta reducción se realiza en dos pasos:

a) Se aplica un filtro paso-bajo para reducir el ancho de banda de la señal y evitar efectos de *aliasing* al reducir el tamaño de la imagen. Este filtro se aplica antes de la citada reducción.
b) El tamaño de las dos imágenes (test y referencia) se reduce en un factor de 2 sin aplicar cálculo de valores medios, que ya no son necesarios por el paso del filtro paso-bajo.

Estos dos pasos se aplican iterativamente M-1 veces, ya que el tamaño original de la imagen se toma como la primera distancia de visualización. El valor final global del *índice de correlación cruzada multiescala de similaridad estructural*, la métrica R*, se obtiene multiplicando los valores de *r*(x, y)* obtenidos a cada una de las diferentes escalas, de acuerdo con la siguiente expresión:

$$R^* = \prod_{j=1}^{M} r_j^*(x, y)$$

Conjunto experimental de imágenes. Se descargó de la web de European Reference Organization for Quality Assured Breast Screeening and Diagnostic Services (EUREF) (Kassermeijer & Thijssen, 2010) un conjunto de 8 imágenes CDMAM en formato raw (set 1). Estas imágenes se obtuvieron con un equipo GE Senographe 2000D a 27kVp, 125 mAs y con una resolución de 1 píxel por cada 100 µm. Estas imágenes han sido evaluadas por cuatro observadores humanos experimentados como elemento de referencia para experimentos de percepción con las imágenes CDMAM. Cada observador evaluó dos imágenes diferentes una vez. Las lecturas obtenidas por estos observadores están disponibles en la misma web que las imágenes.

Se adquirió un segundo conjunto de 20 imágenes (set 2) CDMAM con un equipo Sectra MicroDove LD30 a 32kVp y una resolución de 1 píxel por cada 50 µm. La evaluación de estas imágenes se realizó por un panel de siete expertos. La experiencia de los observadores interpretando mamografías era, al menos, de 3 años.

Los dos conjuntos fueron evaluados de acuerdo con la metodología y reglas publicadas y descritas en el manual del maniquí (Kassermeijer & Thijssen, 2010). De acuerdo con esta metodología, el propósito de cada observación es determinar, para cada diámetro de disco, el nivel límite de percepción de espesor del disco. Para ello, en cada columna (diámetro constante) el último disco percibido con exactitud indica el límite de percepción.

Aplicación de la métrica R a la evaluación del CDMAM.* En el **trabajo I** se han descrito los algoritmos usados para la detección de la rejilla en la imagen de un CDMAM (Prieto, et al., 2009). Este algoritmo ha mostrado ser robusto incluso ante deformaciones geométricas de las rejillas de los maniquíes (Prieto, et al., 2010). Una vez obtenida esta referencia, se desarrolló un algoritmo en ImageJ que calculaba R* en cada una de las cuatro esquinas donde podía aparecer el disco. La imagen de referencia utilizada era un disco ideal creado por síntesis de ordenador (**Figura 1.a en el trabajo II**). El algoritmo barría la zona aproximada donde podía encontrarse cada disco, ya que las diferencias de fabricación de los distintos CDMAM y las ligeras distorsiones geométricas del proceso de adquisición hacían que las imágenes de los discos no estuvieran a distancias exactas y prefijadas de las esquinas de la rejilla. El barrido se hacía computando 25 submatrices de píxeles alrededor de la posición prevista, esto es, fijando un borde de 5 píxeles alrededor de ella. Una descripción gráfica del algoritmo se puede encontrar en la **Figura 1.b del trabajo II**.

Una vez computado el valor de R* en las cuatro esquinas, su valor máximo se tomaba como el de la posición del disco más probable predicha por la métrica R*. En este sentido, R* actuaba como un observador que intentaba dar una predicción en cada una de las celdas.

Para eliminar efectos aleatorios de lecturas más allá del límite de percepción, el propio manual del CDMAM (Bijkerk, et al., 2000a) indica que se apliquen reglas de corrección de los vecinos más cercanos, las reglas NNC (nearest neighbours correction). Básicamente estas reglas descartan aciertos que no se hallan rodeados de un número mínimo de aciertos, entendiendo que la percepción de un disco en medio de varios discos no percibidos, incluso de mejor calidad perceptual que el primero, corresponde al mero azar. Estas reglas NNC fueron aplicadas a las lecturas de observadores humanos y a las lecturas obtenidas por la métrica R*. En este sentido, su tratamiento fue similar al de un observador humano.

Los resultados producidos por los observadores humanos y por R* se compararon estadísticamente mediante coeficientes de correlación de Pearson. Además, también se compararon los resultados con dos algoritmos ampliamente usados en la evaluación de imágenes CDMAM: CDCOM (Karssemeijer & Thijssen, 1996) y PRCDCOM (Young, et al., 2006), realizándose una comparativa final entre todos los subconjuntos de imágenes y los cuatro métodos de observación: observador humano, CDCOM, PRCDCOM y R*. Esta comparativa se realizó mediante métodos de análisis de regresión. Los modelos obtenidos se linealizaron y la comparación de las correspondientes líneas de regresión se realizaron aplicando análisis ANOVA a las tablas obtenidas. Los análisis estadísticos se llevaron a cabo con los paquetes estadísticos SPSS y STATGRAPHICS.

5.3. El problema de la percepción en fondos uniformes frente a fondos anatómicos reales

Para la manipulación de imágenes necesaria en el **trabajo III**, se desarrollaron programas de tratamiento de imagen en Java como plugins dentro del entorno de desarrollo y visualización de ImageJ (Rasband, 1997-2015). Estos programas permitieron las manipulaciones descritas más adelante.

Selección del fondo. El programa desarrollado requiere una imagen CDMAM (**Figura 1 del trabajo III**) y una figura de un fondo mamográfico real, como la que se muestra en la **Figura 2 del mismo trabajo**. El operador selecciona mediante un doble clic la zona de interés que quiere mezclar con la imagen del maniquí. El lugar donde hace doble clic el operador corresponde con la esquina superior de una celda de dimensiones iguales a las del CDMAM mostrado. Nótese que el cálculo del tamaño de celda se realiza individualizadamente por programa para cada imagen CDMAM, ya que difieren notablemente dependiendo del equipo con el que se adquieren, aún para el mismo maniquí.

Sistema de mezcla. El programa permite dos sistemas de mezcla de la imagen del maniquí con el fondo anatómico, La primera está basada en suma lineal. El programa calcula el valor de luminosidad media de los píxeles del cuadrado seleccionado en el fondo mamográfico, y toma este valor como valor de referencia 0. Este valor medio se sustrae al valor de cada píxel en el cuadrado seleccionado. Este procedimiento proporciona un nuevo cuadrado con valores positivos y negativos de los píxeles. Estos valores son sumados, píxel a píxel, a cada una de las celdas que componen la imagen del CDMAM que se desea manipular. La razón para esta suma lineal es que, en los sistemas de mamografía, la luminosidad de los píxeles se obtiene de acuerdo

con el logaritmo de los valores de exposición. La mezcla de ambos fondos debería ser multiplicativa en el dominio de la exposición. Sin embargo, en el dominio de la luminosidad, esta mezcla multiplicativa corresponde a una suma. Se pueden encontrar ejemplos de este tipo de mezclas y de métodos más complejos en la literatura (Madsen, et al., 2006) (Saunders, et al., 2006) (Castella, et al., 2009).

El segundo sistema de mezcla de las imágenes del maniquí con los fondos mamográficos está basado en la multiplicación. El valor de luminosidad de los píxeles del fondo mamográfico se divide por el valor medio de los píxeles de la zona seleccionada. De esta forma, el valor del píxel en el nuevo cuadrado nos proporciona el porcentaje de transmisión de señal. Estos valores se multiplican píxel a píxel, a cada una de las celdas que componen la imagen del CDMAM que se desea manipular. Se pueden encontrar ejemplos de este y otros métodos más complejos en otros autores (Madsen, et al., 2006) (Saunders, et al., 2006) (Castella, et al., 2009).

Por último, se puede introducir un factor de atenuación en ambos métodos, para aumentar o disminuir la señal de fondo mamográfico añadida a la imagen del maniquí CDMAM.

Conjunto de imágenes. El primer tipo de manipulación se aplicó al conjunto de 8 imágenes descargadas de EUREF (Kassermeijer & Thijssen, 2010) y descritas en el **trabajo II**. Las imágenes con y sin manipulación fueron evaluadas por un observador con tres años de experiencia en evaluación de imágenes CDMAM y con unas prestaciones perceptuales similares a las de los observadores de referencia publicados en EUREF.

Análisis de imágenes por métodos automatizados. El algoritmo R*, descrito en el **trabajo II**, fue utilizado para analizar sus prestaciones en fondos mamográficos frente a los observadores humanos. Para comparar los resultados, hemos usado el método de Eficiencia Constante (Tanner & Birdsall, 1958) (Myers, et al., 1985). De acuerdo con este método la eficiencia de R* frente al observador humano (HO) se define como:

$Pc' = (Pc\ Human\ Observer / Pc\ R^*)^2$

Donde Pc representa la figura de mérito Proporción Correcta (de aciertos).

Si la métrica R* es un buen predictor del comportamiento humano, Pc' se debe mantener aproximadamente constante a lo largo de las distintas condiciones del experimento (Eckstein, et al., 2000), en nuestro caso, a través de los diferentes diámetros de los discos.

5.4. El problema de los distintos tipos de ruido y los distintos tipos de imagen radiológica

Para evaluar el grado de correlación de la respuesta de una métrica automatizada y el observador humano se analizaron las modificaciones más prometedoras que se habían realizado sobre la familia SSIM en los últimos años por la comunidad científica y se combinaron los distintos tipos de acercamiento. De esta forma se crearon IQM específicas que presentaban las características de cada uno de estos acercamientos. Básicamente se combinaron cuatro tipos de características:

a) SSIM vs r*. Análisis "S" frente a "r*" en el texto
b) Análisis multiescala vs simple escala. Análisis "M" en el texto
c) Análisis de gradientes (o su ausencia). Análisis "G" en el texto
d) Análisis de tipo de regiones (o su ausencia). Análisis "4" en el texto

Estos cuatro tipos de características integradas en todas las combinaciones posibles dieron lugar a un total de 16 IQM (8 de ellas ya conocidas y otras 8 propuestas en el **trabajo IV**) que fueron comparadas con la percepción de un grupo de radiólogos sobre un conjunto de imágenes médicas. Estas imágenes médicas, a su vez, fueron distorsionadas con distintos tipos de ruido y con diferentes grados de distorsión.

Las métricas

SSIM. (Wang, et al., 2004) Esta métrica evalúa una imagen test Y con respecto a una imagen de referencia X y cuantifica su similaridad visual. Respecto a su cálculo, es completamente similar al descrito para R* en el **trabajo II**; sin embargo, para cada submatriz de píxeles se calculan tres subíndices:

$l(x,y) = (2\mu_x\mu_y + C1)/(\mu_x^2 + \mu_y^2 + C1)$
$c(x,y) = (2\sigma_x\sigma_y + C2)/(\sigma_x^2 + \sigma_y^2 + C2)$
$r(x,y) = (\sigma_{xy} + C3)/(\sigma_x\sigma_y + C3)$

Donde μ_x y μ_y son los valores medios de las submatrices evaluadas, σ_x y σ_y la desviación estándar del valor de las submatrices de píxeles evaluadas. C1, C2 y C3 son constantes introducidas para evitar la inestabilidad del índice cuando los factores $(\mu_x^2 + \mu_y^2)$, $(\sigma_x^2 + \sigma_y^2)$ or $\sigma_x\sigma_y$ están próximos a cero.

El índice *l(x, y)* está relacionado con las diferencias de luminosidad entre ambas imágenes; el índice *c(x, y)* se relaciona con las diferencias en contraste. Por último, el índice *r(x, y)* ya explicado en el **trabajo II,** se relaciona con variaciones estructurales entre las submatrices evaluadas. La forma general del índice SSIM se define como:

$SSIM(x, y) = [l(x, y)]^{\alpha} . [c(x, y)]^{\beta} . [r(x, y)]^{\gamma}$

Donde α, β, y γ son parámetros que definen la importancia relativa de cada componente. Al igual que en el caso de R*, se calcula un índice global SSIM de la imagen multiplicando por el valor de SSIM para todas las submatrices evaluadas.

r*. (Rouse & Hemami, 2009) (Prieto, et al., 2011) Ya explicada en el **trabajo II**.

Multiescala. (Wang, et al., 2003) Ya explicada en el **trabajo II**.

Gradiente, G. (Chen, et al.) en el año 2006 desarrollaron una métrica denominada Gradient Structural SIMilarity (G-SSIM), basada en SSIM. Proponían que el sistema visual humano (HVS) es muy sensible a los bordes y a la información de contorno de las figuras y que estas partes son las que tienen mayor información estructural de la imagen. Para calcular este nuevo índice, sustituyeron las imágenes de test y referencia por sus mapas de gradiente, obtenidas mediante la aplicación de operadores Sobel a las imágenes originales. Sobre los mapas de gradientes, calcularon el índice SSIM, pero la componente de luminosidad se calculaba sobre las imágenes originales, mientras que las componentes de contraste y estructural se calculaban sobre los mapas de gradientes. El resto de las reglas del índice SSIM se aplicaban iterativamente hasta conseguir el valor de G-SSIM.

Cuatro componentes, 4. (Li & Bovik) en el año 2010 analizan el problema de falta de correlación entre la percepción humana y SSIM para imágenes muy desenfocadas o con grandes cantidades

de ruido gaussiano. Proponen un modelo que divide la imagen en regiones, de acuerdo con sus propiedades de uniformidad (textura) o de presencia de bordes. En su propuesta, SSIM se evalúa en estas regiones de la imagen de referencia y la de test, pero se le da un peso específico al valor de SSIM si la región de referencia pasa de tener bordes a no tenerlos en la imagen test, otro valor si el cambio es de región sin bordes a región con bordes y otros pesos para comparaciones en que el tipo de región no cambiaba. De acuerdo con este acercamiento, desarrollaron versiones modificadas de varios índices, en particular, propusieron las versiones 4-SSIM, 4-M-SSIM, 4-G-SSIM y 4-M-G-SSIM.

Basándose en estas métricas, el trabajo propone la combinación de todas las aproximaciones en una familia de 16 índices (**ver Tabla I del trabajo IV**) para su evaluación frente a observadores humanos.

Los observadores. Se seleccionó a cuatro doctores en Medicina, especializados en Radiología, con edades de 57, 35, 32 y 53 y una experiencia en diagnóstico radiológico de 31, 9, 6 y 27 años respectivamente. Denotamos a los sujetos por A, B, C y D.

La base de datos de imágenes. Las imágenes fueron escogidas de la base de datos del hospital Los Madroños (Brunete, Madrid), buscando ejemplos muy representativos de la práctica diaria de un radiólogo. Las imágenes fueron escogidas por un radiólogo con una experiencia de 27 años en este campo, el observador D.
Se seleccionaron tres subconjuntos de ocho imágenes cada uno:

a) BPF. Radiografías usuales de hueso: espalda, rodilla, pie, mano, muñeca, etc.
b) Imágenes de RM. Cráneo, espalda, cuello, etc. Se escogió una imagen representativa de cada estudio.
c) Radiografías de Tórax (CPF).

Las imágenes tenían 256 niveles de gris (imágenes de 8 bits). El tamaño en píxeles de cada imagen era distinto, dependiendo de la técnica de adquisición. El tamaño usual en píxeles para BPF era de 1400x1700, de 512x512 para RM y de 2500x2000 para CPF. Toda la información relativa a los pacientes fue anonimizada.

Tipos de distorsión. Las imágenes fueron distorsionadas con tipos de ruido usuales en un entorno radiológico o de interés para algunas aplicaciones médicas (Johnson, et al., 2010) (European Society of Radiology -ESR-, 2011) (Williams, et al., 2007) (Krupinski, et al., 2007) (Loose, et al., 2009). Los tipos de distorsión son:

a) Desenfoque gaussiano (GB). Se aplicó un kernel circular simétrico de tipo gaussiano con desviaciones estándar de entre 1 y 5 píxeles, aplicando la función de ImageJ “gaussian blur” (v. 1.44).
b) Ruido gaussiano (GN). La desviación estándar aplicada fue de entre 20 y 100, utilizando la función de ImageJ “Add gaussian noise” (v. 1.44).
c) Compresión JPEG (JPG). Se comprimieron las imágenes a ratios que oscilaban entre 0,12 y 0,15 bits por píxel, aplicando la función de Matlab “imwrite” (v. 8.0).
d) Compresión JPEG2000 (J2000). Comprimiendo a ratios de entre 0,01 y 0,04 bits por píxel, aplicando la función de Matlab “imwrite” (v. 8.0).

El número de escalones para cada distorsión se fijó en 5. Por tanto, el número total de imágenes distorsionadas fue de 24 imágenes originales x 4 tipos de distorsión x 5 niveles de distorsión = 480 imágenes.

La cantidad de distorsión empleada intenta reflejar un rango amplio de aspectos, desde imágenes ligeramente distorsionadas hasta niveles elevados de distorsión. Este rango se buscó específicamente para cubrir un amplio grado de percepción de los observadores, desde modificaciones apenas perceptibles por el usuario a distorsiones manifiestas incluso para un lego en la materia.

Metodología del experimento de percepción humana. Se dividieron las imágenes en subconjuntos de 24 imágenes. Cada uno de ellos incluía todas las posibles distorsiones de una imagen y la imagen original. El nombre del fichero informático asociado a las propias imágenes fue aleatorizado para evitar ningún tipo de correspondencia entre distorsión y nombre de fichero.

Cada conjunto de imágenes fue evaluado independientemente por los observadores A, B y C. El observador D fue excluido de esta etapa del experimento para evitar cualquier tipo de sesgo en la evaluación de imágenes, ya que había sido él el encargado de la selección previa.

Se usó un método de doble estímulo. Cada radiólogo disponía de una doble ventana en su dispositivo de visualización. La ventana de la izquierda muestra la imagen de referencia; la de la derecha, las imágenes distorsionadas en orden aleatorio. Los observadores debían responder a la siguiente cuestión:

> ¿Esta imagen es Mala (1), Pobre (2), Regular (3), Buena (4) u Óptima (5) para mi praxis médica? Siempre teniendo en cuenta que la Óptima sería la de referencia (**o indistinguible médicamente de ella**).

Es importante señalar que la intención del experimento no era que los observadores encontraran ligeras diferencias o parecidos entre las imágenes. La intención principal es averiguar la utilidad de la imagen distorsionada para la práctica médica. En este sentido medimos las pérdidas diagnósticas (Krupinski, et al., 2004) (Royal College of Radiologists -RCR, UK-, 2008), y esta intención se hizo claramente patente para los observadores.

No se estableció límite de tiempo para evaluar cada imagen y las sesiones de evaluación fueron de un tiempo máximo de 30 minutos, intentando evitar todo tipo de fatigas visuales. Hubo dos lecturas globales de todas las imágenes por todos los observadores, con un intervalo de seis meses entre ellas, para analizar las posibles variaciones intraobservador.

Medidas de IQM. Las IQM descritas se desarrollaron por nuestro equipo como plugin desarrollados en Java dentro del entorno de ImageJ (Rasband, 1997-2015). Estos programas proporcionaron los distintos valores de las IQM para los mismos conjuntos de imágenes que fueron sometidos a evaluación por los observadores.

Análisis estadístico. Se efectuó un primer análisis de la consistencia intraobservador comparando las dos diferentes medidas obtenidas con los citados seis meses de lapso. Este análisis de consistencia intraobservador se realizó mediante el uso de coeficientes kappa ponderados (Fleiss, 1981), aplicando pesos Cicchetti-Allison (Cicchetti & Allison, 1971). Para aplicar este análisis a cada observador, las puntuaciones “1”, “2”, “3”, “4” y “5” de todo el conjunto de imágenes se ponderó y se clasificó en una tabla de 5x5. Los valores de dicha tabla eran el número total de pares de lecturas concordantes en la primera y segunda ronda de lectura. Por consiguiente, el número total de pares para cada observador es de 480. Este análisis nos permite evaluar la consistencia interna de los observadores. La interpretación de los coeficientes obtenidos tuvo en cuenta la significancia estadística, el número de valoraciones (5)

y su prevalencia (Bakeman, et al., 1997). El análisis incluye el estudio de la consistencia intraobservador para cada uno de los tres tipos de imagen considerados por separado (CPF, RM, BPF) y la homogeneidad del estadístico kappa (Fleiss, 1981) en diferentes tipos de imágenes para un mismo observador.

Para el análisis de la relación entre los resultados de la evaluación automática de las métricas y la evaluación media (MOS) por parte de los observadores humanos, se utilizaron los coeficientes de Pearson (r), y Spearman (rs). El análisis basado en el coeficiente de Spearman es complementario, ya que el sistema de evaluación para ambos sistemas (IQM y MOS) y el adecuado tamaño de la muestra garantizan la pertinencia del uso del coeficiente de Pearson como estadístico principal. Se añadió un tercer estadístico, el error cuadrático medio (RMSE) entre los resultados de las métricas y la MOS. Para profundizar en el análisis, la relación entre los dos conjuntos de evaluaciones se analizó mediante métodos de regresión lineal, considerando las evaluaciones IQM como la variable independiente y el MOS como la variable dependiente. La pendiente b de la recta obtenida en esta regresión, y el punto de intercepción de la misma con el eje de ordenadas, a, proporcionaron información adicional de la asociación entre IQM y MOS. Con estas consideraciones, una pendiente cercana a 1, un valor de a cercano a cero, junto con elevados valores del coeficiente de Pearson y de Spearman, unidos a valores bajos de RMSE muestran un buen ajuste entre las métricas y los observadores.

El análisis estadístico se llevó a cabo utilizando los paquetes SPSS 22 y Epidat 4.1.

6. Discusión integradora

La creación de imágenes sintéticas o alteradas por ordenador suele presentar el problema del aspecto irreal o manipulado de dichas imágenes. Este efecto se hace más patente a observadores entrenados en el reconocimiento de determinado tipo de imágenes. Este mismo entrenamiento hace que los observadores experimentados recuerden perfectamente las posiciones y características de los elementos que forman parte de los maniquíes contraste-detalle, introduciendo distorsiones en el modo de evaluación de dichos maniquíes de acuerdo con la experiencia.

Para evaluar el efecto memoria en los observadores que analizan los maniquís contraste-detalle en radiología y para evaluar la viabilidad de determinados métodos de alteración de imagen, el **trabajo I** presenta una herramienta desarrollada por el equipo de investigación aplicada al maniquí CDMAM. Esta herramienta cambia la posición de la imagen de los discos dentro de cada una de las celdas, colocándolos en posiciones aleatorias con dos sistemas de giro distintos. Este procedimiento asegura que el método de elección forzada entre 4 alternativas que utiliza el CDMAM es válido, incluso cuando el maniquí es evaluado por expertos con gran experiencia en evaluación de este tipo de imágenes. El programa puede seleccionar un ángulo prefijado de rotación para todos los discos o un valor aleatorio, lo que hace imposible para un observador conocer la posición del disco test dentro de cada celda y, especialmente, en las celdas críticas cerca del límite de percepción. Se realizaron análisis ROC sobre la respuesta de 36 observadores. Dichos análisis mostraron que las imágenes originales y las modificadas por programa eran indistinguibles. El área bajo la curva ROC fue de 0,507+/- 0,024 para el primer sistema de giro y de 0,522 +/- 0,026 para el segundo sistema (**Tabla 1 del trabajo I**), indicando que no existe diferencia estadísticamente significativa para la percepción de los observadores entre las imágenes reales y las modificadas por programa.

Para analizar la posibilidad de una evaluación automatizada de la imagen de un maniquí por parte de una IQM, R* (no aplicada antes a imágenes médicas), evitando la participación del observador humano, se desarrolló el **trabajo II**. También se comparó el comportamiento de R* frente a dos programas ampliamente usados para evaluar el maniquí CDMAM: CDCOM y PRCDCOM. Los análisis se hicieron con dos conjuntos de imágenes (set 1 y set 2) analizados por dos conjuntos de observadores distintos.

Los resultados obtenidos demostraron que los resultados aplicando R* eran indistinguibles estadísticamente de los observadores humanos para los dos experimentos (**Figura 3.b. del trabajo II**), con un F-Test para el test de paralelismo entre observadores humanos y R* con $p > 0,71$ para el set 1 y $p > 0,08$ para el set 2). También se observó que los resultados de R* eran estadísticamente mejores (en el sentido de más cercanos a la percepción humana) que los otros dos algoritmos usados como referencia, CDCOM y PRCDCOM (**Figura 4 del trabajo II**).

Aplicando el test de paralelismo con respecto a los observadores humanos, ambos CDCOM y PRCDCOM mostraron su validez, con las imágenes del set 1, sin diferencias estadísticamente significativas en el test de paralelismo. Sin embargo, los valores del F-test para la hipótesis de paralelismo con el set 2 fueron de $p = 0,002$ para CDCOM y $p = 0,011$ para PRCDCOM, lo que invalida la tesis de paralelismos entre unos resultados y otros y, por tanto, su uso como subrogados del observador humano.

El aseguramiento de la calidad de imagen es uno de los pilares de la gestión de los sistemas de imagen médica. El uso de maniquíes contraste-detalle es de enorme utilidad en este aseguramiento, pero una buena parte de ellos presentan fondos uniformes sobre los que se detectan los elementos de contraste-detalle insertados. Las características de calidad de la imagen son luego extrapoladas al funcionamiento general del sistema. Sin embargo, las imágenes anatómicas reales presentan estructuras que enmascaran otras estructuras de bajo contraste o detalle de interés diagnóstico, como es el caso de determinadas masas tumorales en mamografía. Es el conocido por ruido estructural. En el **trabajo III** se creó una herramienta informática que sustituyó el fondo uniforme de un maniquí usado en mamografía, el CDMAM, por fondos tomados de mamografías reales. Las imágenes fueron analizadas por observadores humanos y por la IQM R*, utilizada en anteriores trabajos. La **Figura 7 del trabajo III** muestra las prestaciones relativas de ambos observadores, humano e IQM, en ambos fondos. Se observan tres hechos relevantes: el primero es la baja respuesta del observador humano en fondos mamográficos. Este efecto ha sido ampliamente descrito en la literatura. El segundo hecho es que R* muestra un comportamiento similar: su rendimiento se reduce en fondos mamográficos, tal y como era esperable debido al elevado nivel de ruido estructural presente. El tercer hecho destacable es que el umbral de detección de espesor para discos mayores de 1,25 mm crecía, al contrario de lo que pueda parecer intuitivo en una primera aproximación, ya que los discos son mayores en diámetro y, por tanto, quizá más perceptibles. La explicación, ya avanzada en otros trabajos (Grossjean & Muller, 2006) (Burgess, et al., 2001), es que el tamaño de los discos comienza a ser del tamaño de determinadas estructuras anatómicas presentes en el fondo real, que enmascaran la detección del disco.

Se calculó el valor de Eficiencia Relativa del observador humano frente a R*. Los resultados pueden verse en la **Figura 8 del trabajo III**. Esta figura muestra un efecto ya encontrado por otros autores (Grossjean & Muller, 2006) utilizando otras métricas: las prestaciones de la IQM se reducen para discos mayores de 1,25 mm con respecto a las del observador humano. Este efecto también fue encontrado por (Burgess, et al., 2001) aplicando también métricas diferentes. Opinamos, junto a estos autores que, para determinados tamaños de discos, la percepción humana usa más la información de cambio de bordes que la información de contraste entre el centro del disco y el fondo. Esta información fue muy relevante para las siguientes investigaciones del grupo de trabajo.

Las investigaciones relacionadas en los anteriores trabajos llevaron de forma natural a preguntarse una serie de cuestiones:

a) ¿Cuál sería el comportamiento de la métrica estudiada ante una modificación del algoritmo que incluyera el efecto de los gradientes, es decir, de los bordes de las estructuras, en la percepción?
b) ¿Qué otras modificaciones podían plantearse a la métrica R* para mejorar su rendimiento en fondos estructurados?
c) ¿Cómo se comportaría la métrica modificada ante varios tipos de fondos estructurados, y no solo en mamografía?
d) ¿Cómo se comportaría la métrica mencionada ante la inclusión de diversos tipos y niveles de ruido de interés en la investigación médica?

Estas preguntas llevaron a la elaboración del **trabajo IV**, cuya metodología y alcance ya se ha descrito. En este trabajo se puso de manifiesto la mejora de las métricas analizadas, entendida

"mejora" como "comportamiento similar al del observador humano" ante los siguientes componentes:

1) Análisis de texturas (componente "4"). La aplicación de este componente (**Figura 4 y Tabla VIII del trabajo IV**) siempre mejoraba el comportamiento de cualquier métrica analizada, con porcentajes de mejora de hasta el 60% en el coeficiente Pearson de correlación entre las lecturas humanas y automatizadas y, en general, en todos los estadísticos escogidos para medir el grado de correlación entre los observadores humanos y las IQM.

2) Componente multiescala, "MS". Con pequeñas excepciones, el acercamiento de una solución multiescala que simulaba distintas distancias de observación, mejoró el comportamiento de todos los estadísticos usados (**Figura 5 y Tabla VIII del trabajo IV**), con aumentos de hasta el 40% en alguno de ellos. Tan solo la métrica 4-r* presentaba un r ligeramente superior que 4-MS-r* (0,64 frente a 0,60). También la pendiente de la recta de regresión b empeoraba ligeramente con la métrica G-r* frente a la métrica MS-G-r*, pero solo en esta figura de calidad y con un valor de 0,82 frente a 0,79.

3) Componente "G". También este componente, que sobrepondera los bordes de las estructuras presentes en las imágenes mejora todas las métricas, con pequeñas excepciones (**Figura 6 y Tabla VIII del trabajo IV**). Sin embargo, el efecto no es tan acusado como con el componente MS o 4, excepto con el estadístico r, que mejora hasta en un 30% para algunas métricas. Otros estadísticos, como RMSE, b, etc., mejoraban hasta un máximo de un 15%. Las excepciones a la mejora de las métricas se produjeron en las métricas 4-MS-r*, MS-r* y 4-MS-SSIM, con pérdidas, exclusivamente del valor b, de un 3%, 3% y 8% respectivamente.

4) Componente estructural de SSIM, r*. Este componente mejoraba el valor de todos los estadísticos, con incrementos de hasta el 60% en el valor de r o del 40% en el valor de rs (**Figura 8 y Tabla VIII del trabajo IV**). Sin embargo, su comportamiento era menos uniforme y más errático que los otros tres componentes analizados. La métrica 4-MS-G-SSIM perdía prestaciones de forma general. MS-G-SSIM no cambiaba con respecto a MS-G-r*. MS-SSIM y 4-MS-SSIM mostraban una mejora general de los estadísticos analizados, pero el valor de rs decrecía un 10% y un 7% respectivamente. Las otras IQM mejoraban de forma clara sus prestaciones.

Si analizamos las prestaciones generales sobre todos los tipos de ruido aplicados y sobre todas las técnicas de adquisición estudiadas, la métrica más efectiva es 4-MS-G-SSIM, que supera claramente a las otras métricas en los valores de r, rs y RMSE (**Tabla V del trabajo IV**). Además, muestra unos excelentes valores de b (1,10) y de a (-0,04). La segunda métrica más efectiva es 4-MS-G-r*. Otras dos métricas, 4-G-SSIM y 4-G-r*, muestran resultados similares, aunque sus prestaciones son ligeramente inferiores a las dos primeras métricas referidas.

También se realizó un análisis por tipo de imagen del anterior conjunto de las cuatro mejores métricas. Las imágenes de RM produjeron los mejores resultados de correlación en términos de la combinación de los estadísticos r, rs y RMSE (**Tabla VI del trabajo IV**). Aunque las correlaciones globales eran menores para los estudios de CPF y de BPF, este efecto era debido al bajo resultado de las métricas 4-G-SSIM y 4-G-r*. Sin embargo, las dos primeras métricas del conjunto, 4-MS-G-SSIM y 4-MS-G-r* presentaron también buenos valores para estos dos tipos de imágenes.

Un resultado interesante del estudio fue la gran mejora del valor de r cuando se introducía el componente MS para analizar imágenes CPF y BPF. Como se puede ver en la **Figura 3 del trabajo IV**, cuanto más grande es la imagen, mayor es la mejora introducida por el componente multiescala. Como se ha descrito, el componente MS divide iterativamente hasta 5 veces el tamaño de la imagen y evalúa la métrica correspondiente para cada tamaño. Las imágenes de CPF tenían tamaños en la dimensión mayor de hasta 2.400 píxeles. Reducir iterativamente su tamaño 5 veces nos da una imagen final de 75 píxeles, todavía con información visual para un observador. Sin embargo, reducir 5 veces una imagen de RM produce imágenes de 16x16 píxeles, con escasa o nula información para un observador.

También se analizó el comportamiento de las métricas por tipo de imagen y por tipo de ruido añadido. En el análisis de las imágenes BPF (**Tabla VII del trabajo IV**) se observa que las métricas no multiescala empeoraban frente a sus simétricas multiescala, notablemente en presencia de desenfoque gaussiano. 4-MS-G-SSIM mostraba coeficientes r mayores a 0,87 para todo tipo de distorsiones.

Al analizar las imágenes de RM, las prestaciones de las cuatro métricas eran muy buenas (**Tabla VII del trabajo IV**) y comparables, con un valor de r entre 0,81 y 0,91 para todo tipo de distorsiones.

El análisis de las imágenes CPF mostró, al igual que con las imágenes BPF, que las métricas de escala simple tenían prestaciones claramente inferiores que sus simétricas multiescala (**Tabla VII del trabajo IV**). Nuevamente, este efecto era debido al bajo rendimiento de esas métricas en presencia de desenfoque gaussiano. 4-MS-G-SSIM mostró de nuevo excelentes resultados para todos los tipos de distorsión (0,85<= r<= 0,95).

7. Conclusiones

1. Se pueden utilizar programas de manipulación de imagen para evitar los efectos de memoria de los observadores experimentados en la evaluación de imágenes de maniquíes de contraste-detalle, en particular el maniquí CDMAM, de amplio uso en mamografía. Las alteraciones en las imágenes se pueden realizar con distintos métodos y ninguna de ella es perceptible para el observador, incluso aunque esté muy familiarizado con este maniquí. En particular, el programa desarrollado en este trabajo, puede ser utilizado por la comunidad científica para evaluar el citado efecto memoria.

2. Existen métricas de calidad de imagen provenientes de campos distintos del de la imagen médica, como las de la familia SSIM, que han sido muy poco estudiadas y aplicadas a problemas relacionados con este área. Nuestros estudios muestran que una de ellas, R*, específicamente diseñada para analizar señales en el límite de percepción, puede ser utilizadas para la evaluación automática de maniquíes. Los resultados son estadísticamente indistinguibles de los proporcionados por un observador humano, luego la tarea de percepción puede ser automatizada, con los consiguientes ahorros de tiempo.

3. Se pueden crear herramientas para generar imágenes híbridas de maniquíes y fondos anatómicos reales en mamografía. Estas imágenes se pueden utilizar para analizar la respuesta de un ser humano o de un sistema automatizado que aplica una IQM. Los resultados de la métrica elegida, resultan ser similares a aquellos obtenidos por el observador humano. Además, se pone de manifiesto el enmascaramiento de determinados elementos de contraste-detalle por el ruido estructurado del propio fondo mamográfico. Por último, se observa una sobre respuesta del observador humano en detalles de contraste de dimensiones elevadas con respecto a la métrica analizada, probablemente debida a la especial percepción del sistema visual humano ante estructuras con bordes contrastados.

4. Existen métricas cuyos resultados analizando la calidad de una imagen médica tienen un comportamiento muy similar a los de un observador humano experto. Estos experimentos se han llevado a cabo sobre un amplio abanico de imágenes médicas y sobre un amplio conjunto de tipos de ruido relevantes en imagen médica. La gran correlación mostrada por los componentes que analizan los cambios de textura, los bordes de las estructuras y que integran distintas distancias de observación hacen suponer que estos componentes tienen especial relevancia en el funcionamiento del sistema visual humano.

Por último, queremos compartir nuestros resultados con nuestros colegas científicos y académicos. Todos los programas y algoritmos desarrollados a lo largo de esta investigación serán publicados en nuestra web (https://www.ucm.es/gabriel_prieto) al finalizar la lectura de esta tesis. Una parte de estos algoritmos ya está libremente disponible en la citada dirección web para la comunidad y han sido utilizados por varios grupos de investigación para el desarrollo de experimentos de percepción de calidad de imagen.

8. Bibliografía

Bakeman, R., Quera, V., McArthur, D. & Robinson, B. F., 1997. Detecting sequential patterns and determining their reliability with fallible observers. *Psychological Methods,* Volume 2, pp. 357-370.

Barrett, H. H., Myers, K. J. & Wagner, R. F., 1986. Beyond signal detection theory. *Proc. SPIE 0626,* Proc. SPIE 0626, Application of Optical Instrumentation in Medicine XIV and Picture Archiving and Communication Systems(626), pp. 231-239.

Bijkerk, K., JM, L. & Thijssen, M., 1993. The CDMAM phantom: a contrast-detail phantom specifically for mammography. *Radiology,* Volumen 185, p. 395.

Bijkerk, K., JM, L. & Thijssen, M., 2000. *Manual CDMAM-phantom type 3.4,* Nijmegen, The Netherlands: Department of Diagnostic Radiology, University Medical Centre.

Bijkerk, K., JM, L. & Thijssen, M., 2000. Modifications of the CDMAM Contrast-Detail Phantom for Image Quality Evaluation of Full-Field Digital Mammography Systems. *Proceeedings of IWDM 2000,* Volumen IWDM 2000, pp. 663-640.

Burgess, A. A., Jacobson, F. L. & Judy, P. F., 2001. Human observer detection experiments with mammograms and power-law noise. *Med. Phys.,* 28(4), p. 419.437.

Burgess, A. E., 1999. The Rose model, revisited. *J. Opt. Soc. Am.,* A(16), pp. 633-646.

Burgess, A. E., 2001. Evaluation of detection model performance in power-law noise. *Proceedings of SPIE,* Volumen 4324, pp. 123-132.

Castella, C. y otros, 2009. Mass detection on mammograms: influence of signal shape uncertainty on human and model observers. *J. Opt. Soc. Am,* A(26), pp. 425-436.

Chen, G. H., Yang, C. L. & Xie, S. L., 2006. Gradient-based structural similarity for image quality assessment. *IEEE International Conference on Image Processing,* pp. 2929-2932.

Cicchetti, D. V. & Allison, T., 1971. A New Procedure for Assessing Reliability of Scoring EEG Sleep Recordings. *American Journal of EEG Technology,* Volume 11, pp. 101-109.

Eckstein, M. P., Abbey, C. K. & Bochud, F. O., 2000. A practical guide to model observers for visual detection in synthetic and natural noisy images. In: J. Beutel, H. Kundel & R. Van Metter, eds. *Handbook of Medical Imaging.* Bellingham WA: SPIE, pp. 593-626.

European Society of Radiology (ESR), 2011. Usability of Irreversible Image Compression in Radiological Imaging. A Position Paper by the European Society of Radiology (ESR). *Insights into Imaging,* 2(2), pp. 103-115.

Fleiss, J. L., 1981. *Statistical methods for rates and proportions.* New York: John Wiley & Sons.

Girod, B., 1993. What's wrong with mean-squared error. In: *Digital Images and Human Vision.* s.l.:MIT press, pp. 207-220.

Grossjean, B. & Muller, S., 2006. Impact of textured background on scoring of simultaed CDMAM phantom. *IWDM 2006,* Volumen 4046, pp. 460-467.

Johnson, J. P. et al., 2010. Using a visual discrimination model for the detection of compression artifacts in virtual pathology images. *IEEE Transactions on Medical Imaging,* 30(2), pp. 306-314.

Karssemeijer, N. & Thijssen, M. A. O., 1996. Determination of contrast-detail curves of mammography systems by automated image analisys. *Digital Mammography,* pp. 155-160.

Kassermeijer, N. & Thijssen, M. A. O., 2010. *CDCOM software, manual, and sample images.* [En línea]
Available at: www.euref.org
[Último acceso: 17 July 2010].

Krupinski, E. A. et al., 2004. Use of a human visual system model to predict observer performance with CRT vs LCD display of images. *J Digit Imaging,* 17(4), pp. 258-263.

Krupinski, E. A. y otros, 2007. Digital Radiography Image Quality: Image Processing and Display. *J Am Coll Radiol,* Volumen 4, pp. 371-388.

Li, C. & Bovik, A. C., 2010. Content-partitioned structural similarity index for image quality assessment. *Journal Image Communication,* 25(7), pp. 517-526.

Loose, R. et al., 2009. Compression of digital images in radiology results of a consensus conference. *Rofo,* 181(1), pp. 32-37.

Madsen, M. y otros, 2006. A new software tool for removing, storing and adding abnormalities to medical images for perception research studies. *Acad Radiol,* Volumen 1, pp. 305-312.

Myers, K. J., 2000. Ideal observer models of visual signal detection. In: J. Beutel, H. Kundel & R. Van Metter, eds. *Handbook of Medical Imaging, Physics and Psycophysics.* Bellingham WA: SPIE, pp. 558-592.

Myers, K. J. y otros, 1985. Effect of noise correlation on detectability of disc signals in m3edical imaging. *J. Opt. Soc. Am,* A(2), pp. 1752-1759.

Prieto, G., Chevalier, M. & Guibelalde, E., 2009. Automatic scoring of CDMAM using a model of the recognition threshold of the human visual system: R*. *Image Processing (ICIP), 2009 16th IEEE International Conference on,* pp. 2489-2492.

Prieto, G., Chevalier, M. & Guibelalde, E., 2010. A software tool to measure the geometric distortion in x-ray image systems. *Proc. of SPIE,* Volumen 7622, pp. 173-180.

Prieto, G., Chevalier, M. & Guibelalde, E., 2011. A software tool to compare contrast-detail detection in uniform and in real mammographic backgrounds. *Proc. SPIE,* Volumen 7966, pp. 122-128.

Prieto, G., Guibelalde, E. & Chevalier, M., 2007. *Manipulation of DICOM format images: Search of a visualization and manipulation environment for the generation of hybrid images.* Pisa, Italy, Xth EFOMP Congress.

Prieto, G., Guibelalde, E., Chevalier, M. & Turrero, A., 2011. Use of the cross-correlation component of the multiscale structural similarity metric (R* metric) for the evaluation of medical images. *Med. Phys,* 38(8), p. 4512.7.

Rasband, W. S., 1997-2015. *ImageJ, U. S. National Institutes of Health.* [Online]
Available at: http://rsb.info.nih.gov/ij/plugins/index.html [Accessed 12 5 2015].

Rouse, D. M. & Hemami, S. S., 2009. Analyzing the Role of Visual Structure in the Recognition of Natural Image Content with Multi-Scale SSIM. *Proceedings of SPIE,* Volume 6806.

Royal College of Radiologists (RCR, UK), 2008. *The adoption of lossy data compression for the purpose of clinical interpretation.* [Online] Available at: https://www.rcr.ac.uk/sites/default/files/publication/IT_guidance_LossyApr08_0.pdf [Accessed 12 5 2015].

Saunders, R., Samei, E., Baker, J. & Delong, D., 2006. Simulation of mammographic lesions. *Acad Radiol,* Volumen 13, pp. 860-870.

Tanner, W. P. & Birdsall, T. G., 1958. Definitions od d' and η as psychophysical measures. *J, Acoust. Soc. Am.,* Volumen 30, pp. 922-928.

Wang, Z. & Bovik, A. C., 2002. Why is image quality assessment so difficult?. *Proceedings of IEEE Int. Conf. Acoustics, Speech and Signal Processing,* Issue 4, pp. 3313-3316.

Wang, Z., Bovik, A. C., Sheikh, H. R. & Simoncelli, E. P., 2004. Image Quality Assesment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing,* Volumen 13, pp. 600-612.

Wang, Z., Simoncelli, E. & Bovik, A., 2003. Multi-scale structural similarity for image quality assessment. *Proceedings of the 37th IEEE Asilomar Conference on Signals, Systems and Computers,* pp. 529-554.

Williams, M. B. et al., 2007. Digital Radiography Image Quality: Image Acquisition. *J Am Coll Radiol,* Volume 4, pp. 371-388.

Young, K. C., Cook, J. J. H., Oduko, J. M. & Bosmans, H., 2006. Comparison of sofware and human onserversin reading images of the CDMAM test object to assess digital mammography systems. *Proc. SPIE,* Volumen 6142, p. 614206.

9. Otras publicaciones del autor relacionadas con el tema de la tesis

- **Prieto, G.**, Guibelalde, E. & Chevalier, M., 2007. *Manipulation of DICOM format images: Search of a visualization and manipulation environment for the generation of hybrid images.* Pisa, Italy, Xth EFOMP Congress.
- **Prieto, G.**, Chevalier, M. & Guibelalde, E., 2009. Automatic scoring of CDMAM using a model of the recognition threshold of the human visual system: R*. *Image Processing (ICIP), 2009 16th IEEE International Conference on,* pp. 2489-2492. **Web of Science: 000280464301072**
- **Prieto, G.**, Chevalier, M. & Guibelalde, E., 2009. An alternative approach to CDMAM analysis using a new perceptual algorithm: MS-SSIM*. *IFMBE Proceedings. World Congress on Medical Physics and Biomedical Engineering,* 25(2), pp. 518-521. **Web of Science: 000306060900145**
- **Prieto, G.**, Chevalier, M. & Guibelalde, E., 2010. A software tool to measure the geometric distortion in x-ray image systems. *Proc. of SPIE,* Volumen 7622, pp. 173-180. **Web of Science: 000285047200162**

10. Trabajo I

A CDMAM Image Phantom Software Improvement for Human Observer Assessment

Gabriel Prieto, Margarita Chevalier, and Eduardo Guibelalde

Dept. Radiología. Fac. Medicina. Universidad Complutense de Madrid.
28040 Madrid, Spain
gprietor@med.ucm.es, chevalier@med.ucm.es, egc@med.ucm.es

Abstract. A software tool is presented to improve the features of CDMAM image phantom by University Hospital Nijmegen. This software tool ensures that the 4-alternative forced choice method of CDMAM is actually kept, even when is being scored by highly expertise observers familiar on the test object pattern. For digital images, the developed software tool automatically changes the image position of the four corners. It can be selected a fixed rotation angle or a random one, so making impossible that any observer is able to remember the exact corner position of the target disc inside any cell. Two alternative successful algorithms have been tested. ROC curve analysis obtained by 36 observers shows that both original and computer-modified images are indistinguishable. The ROC area was 0.507±0.024 for first algorithm and 0.522±0.026 for the second one, indicating that there was no statistical difference between real and computer-modified images for both of them.

1 Introduction

Many phantoms have been designed to study mammographic image quality such as ACR, TOR(MAM) or CDMAM phantom [1 2 3]. The task with these phantoms is to obtain the minimum contrast (threshold) for each diameter of a series of discs with different contrasts. Usually, the discs are located in well known positions and the evaluation is based in the SKE paradigm. The main advantage of the CDMAM phantom (Nijmejen) is that discs are located in one of the four corners of the 205 cells in which the phantom is divided. However, due to a group of discs is always seen while other group is never seen, the evaluation procedure is focused to a less number of cells. In addition, the tolerances established in some protocols for some discs could reduce the evaluation to a smaller number of discs. In consequence, the memory effect can not be rejected.

2 Methods

We have developed our algorithms as a plugin inside ImageJ, the image manipulation program developed by Wayne Rasband [4]. Our software developments will be periodically updated in the ImageJ website http://rsb.info.nih.gov/ij/plugins/index.html, including object and source codes, instructions of use and several test images.

E.A. Krupinski (Ed.): IWDM 2008, LNCS 5116, pp. 181–187, 2008.

The CDMAM phantom consists of an aluminium base with gold discs of varying thicknesses and diameters, which is attached to a Plexiglas cover. The discs are arranged in 16 rows and 16 columns. Within a row, the disc diameter is constant, with logarithmically increasing thickness. Each cell contains two gold disks each with the same diameter and thickness. One disk, the reference signal, is in the centre of the cell; the other, the test signal, is in one of the four corners.

Detection of grid position. In order to manage the disks information from the phantom images, it is necessary to accurately detect the position of the phantom grid where gold discs are inserted. Several methods have been applied to find this position [5 6]. The method used in this work was designed to be simple and time effective on the basis that we have complete information about the geometry of the CDMAM.

We select a fixed ROI around the centre of the image with side dimensions of one third of the dimensions of the CDMAM. So, we can be sure that there are no unexposed bright areas or alphanumerical information that could affect our algorithms. Inside this area, we scan all pixels in the first and last columns. For each pixel we consider a fan of straight lines with origin at this pixel and ending on the other side of the ROI, within an angle range between 35 ° and 55º, stepping one quarter of degree (Fig. 1). For each of these straight lines we calculate the addition of the pixel values. Maximum values of these additions indicate where the grid lines lie and which one is their angle. Maximum values from the ROI's first column allow us to detect negative slope grid lines and with the maximum values from the last column we detect positive slope grid lines. Both values give us the diagonal length (D) of the phantom cells, different for each side of the phantom. This fact probably can be due to the x-ray beam geometry. Using both data (angle and D), we extrapolate until the edge of the grid in the full CDMAM image.

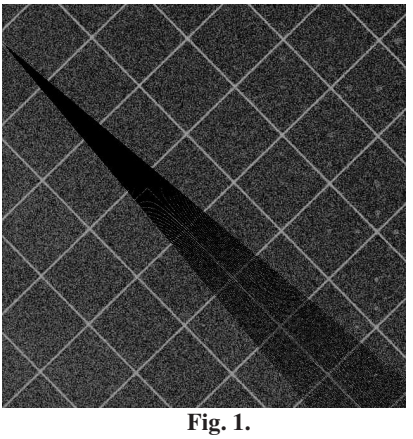


Fig. 1.

To assure the results, we scan around the intersection of each theoretical grid line with the edge. We repeat the process along the CDMAN grid edge in steps of D +/- 10 pixels around the expected points, looking for the better starting point of each straight line that better fits the grid. We run this process for both sides of the phantom. According the data of all straight lines, we calculate the intersection points of the grid. The distance between our calculated crossing points and the actual crossing points is between zero or one pixel. Only in a few cases for each image (<1%) this distance was equal or bigger than two pixels.

The main properties of this algorithm are:

a) Low computational consumption. The computing complexity for the central ROI calculations is of the order of 30 x n (where n is the number of pixels of the hole

image). This complexity is of the order of n/10 for computing the algorithm for the rest of the image.[1]

b) No need of any kind of pre-processing, even for very noisy images

c) Robustness of the algorithm under very different conditions. The percentage of success detecting the grid was 100%. We have tested different images (40) from instruments from different manufacturers and models (LORAD-HOLOGIC, GE MEDICAL SYSTEMS, AGFA, FUJI) and with different levels of noise. The noise index (Std. deviation / mean value of the pixel) measured at a corner without grid lines, alphanumeric symbols or graphics, had values between 0,010 and 0,025. The angle of the detected grid lines had values between 43º and 47º. The only error we found a few times was a maximum shift of +/- 2 pixels between the calculated crossing points of the grid and the actual ones.

Phantom image manipulation. To avoid the memory effect of the expert observer, we have moved the corner disks in some cells. We have used two different algorithms, very well illustrated in Fig. 2 and 3.

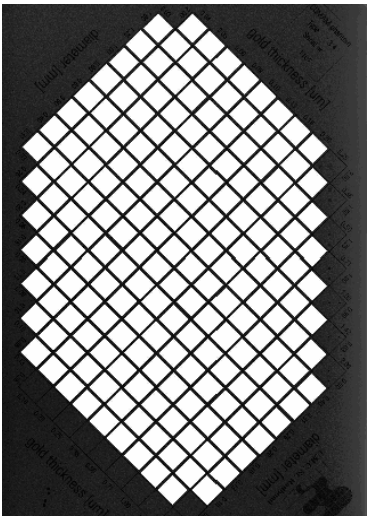


Fig. 2.

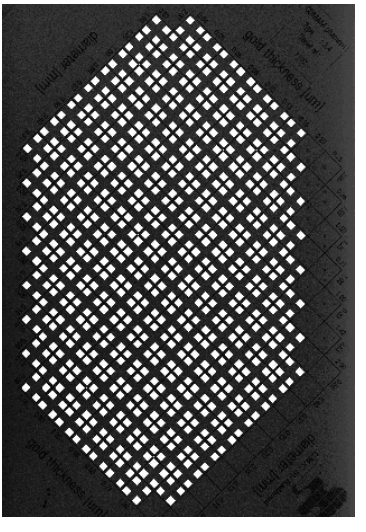


Fig. 3.

For both algorithms we have marked a "safe region" inside each cell where we can process the image without disturbing the grid itself. The cell shape is trapezoidal because of the geometry of the whole image acquisition system. In consequence we can not use the grid itself as rotation edge because each cell is not symmetrical with respect to their diagonals. To avoid this problem, we consider a square region inside each cell with a margin of 8 pixels from the upper crossing point toward the centre of the cell. The resulting square has a diagonal length of (D-16) pixels (where D= min. [left diagonal, right diagonal] pixels).

[1] For instance, the time consumption for an image of 1628 x 2280 pixels, with 16 bits per pixel, using the first algorithm was 0.53 seconds and 0.56 seconds using the second algorithm, in a laptop computer, Dell Inspiron 4400, processor Intel Centrino Core2 Duo T7200, 2 Ghz, 2 Gb RAM.

The first algorithm ("One rhombus") rotates each cell in steps of 90 degrees around its centre. In the second algorithm ("Four of diamonds"), we define four little rhombuses inside each cell, centred each one at the four possible centres of the CDMAM's disks and then we interchange these rhombuses between them inside the same cell.

The interface for some combinations is showed in Fig. 4 and 5

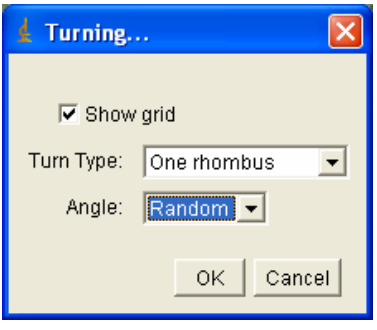


Fig. 4.

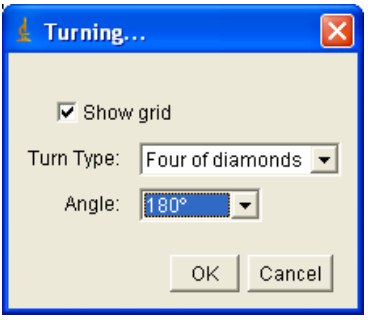


Fig. 5.

Both algorithms have some limitations:

a) The angle of the grid must lie between 35º and 55º, but these values simply are two parameters that can be changed easily.
b) The total image area should cover the total image area of the CDMAM with a tolerance less than a 10%. We will eliminate this limitation in the next program update.
c) There should not be great differences of uniformity inside each cell to be rotated. In this case, the program still works properly, but it is very easy to find out that the images are manipulated. We have defined two ROIs inside each cell (see Fig. 6 and 7). If the difference between the mean value of the pixels inside each ROI is greater than 2%, the manipulation is visible and this image could not use to run out our test. This limitation is a strong one related to the observer perception characteristics.

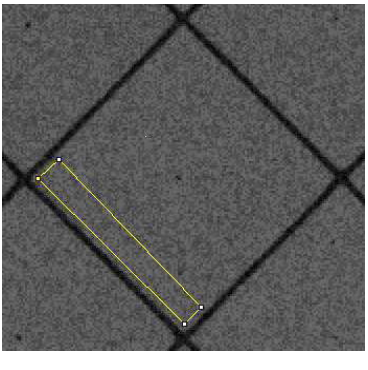


Fig. 6.

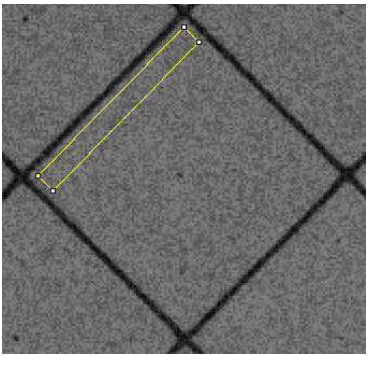


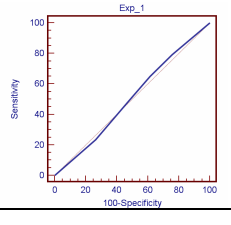
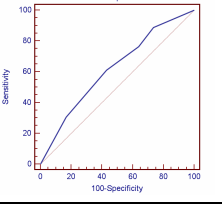
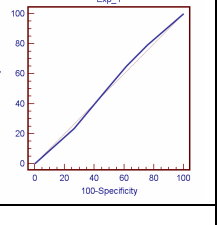
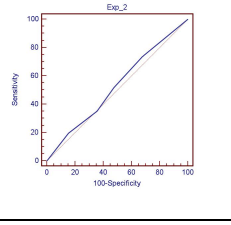
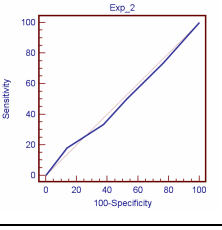
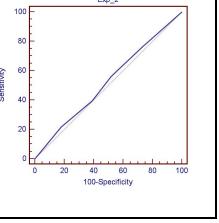
Fig. 7.

Algorithms evaluation. A ROC experiment has been developed to evaluate both algorithms. We arranged two set of images. Set #1 contained 8 different CDMAM images acquired under different radiological conditions. We obtained 8 modified images by applying algorithm "Rhombus" to each image. The complete set was formed with modified and non modified images randomly ordered. Set #2 was formed with the same 8 original CDMAM images and 8 modified images obtained by applying the algorithm "Four of diamonds". The complete set was randomly ordered. The two sets of images were presented to two different groups of observers. The first one was composed by 27 medical physicists with an experience between 2 – 10 years in quality control. The second group was composed by 9 students of a Medical Physics Master, with no experience in diagnostic radiology neither quality control. Each observer had to answer to the question "Do you think this image has been computer modified in any form?" The test was run in different displays and the observers could use any tool of the ImageJ viewer with no time limits. The answer included a confidence level, from 0 to 4.

3 Results

As it can be seen from ROC curves in Table 1, the computer-modification of images was indistinguishable for both groups and for both algorithms. Result may show a light significance in test #1 analyzed for students, but the significance is at the limit of random choice. Combining expert and non-expert observers, the ROC area was 0.507±0.024 for first algorithm (significance level P=0.7724 for area 0.5) and 0.522±0.026 (significance level P=0.4003 for area 0.5) for the second one, indicating that there was no statistical difference between real and computer-modified images for both of algorithms presented.

Table 1.

	Experts' group	Students' group	Adding results
Experiment #1			
Experiment #2			

4 Discussion

The main finding of this study is that there are algorithms that can be used to modify CDMAM images, moving the disk positions around the centre of the cells. The modified image is indistinguishable, even for expert observers.

The strongest limitation of the algorithms is associated to images with a great lack of background uniformity inside the cells. Theoretically, this limitation could be removed by moving the image of the gold disk inside each cell. This approach presents some practical problems. The first one is related to the accuracy in cutting the disc image. This accuracy should be equal or better than two pixels around to avoid the annular section around the circle with a different gray level than the one at the destination point. The second one concerns to the change on the contrast ratio of the disk respect to the background in the destination point due to the lack of uniformity. This change would be produced although the cut operation had enough accuracy. Both problems might be solved changing the mean pixel value of the disk we move according the ratio of the back signal between the source and the destination point.

Our next steps are to optimize this software tool to manage images with a high gradient of uniformity and after that we will run the memory test with expert radiologists. Our developments will be regularly updated in the plugins section of ImageJ website.

5 Conclusion

1) Both algorithms can be used to manipulate CDMAM images.
2) They can be useful in those cases in which a researcher suspects that the observer is memorizing the position of some disks, mainly the critical ones around the middle section of the CDMAM.
3) They can be used to investigate and quantify the memory effect in the radiologist community. We have made a preliminary test with students, trying to train them to memorize the position of the disks, not for explicit methods, but for indirect ones.

References

1. Bijkerk, K.R., Lindeijer, J.M., Thijssen, M.A.O.: The CDMAM phantom: a contrast-detail phantom specifically for mammography. Radiology 185(P), 395 (1993)
2. Bijkerk, K.R., Thijssen, M.A.O., Arnoldussen, T.J.M.: Manual CDMAN-phantom type 3.4., Department of Diagnostic Radiology, University Medical Centre, Nijmegen, The Netherlands (2000)
3. Bijkerk, K.R., Thijssen, M.A.O., Arnoldussen, T.J.M.: Modification of the CDMAN Contrast-Detail Phantom for Image Quality Evaluation of Full-Field Digital Mammography Systems. In: Yaffe, M. (ed.) Proceedings of IWDM 2000, pp. 663–640. Medical Physics Publishing, Madison, Toronto (2000)

4. Rasband, W.S.: ImageJ, U. S. National Institutes of Health, Bethesda, Maryland, USA (1997-2008), http://rsb.info.nih.gov/ij/
5. Karssemeijer, N., Thijssen, M.A.O.: Determination of contrast-detail curves of mammography systems by automated image analysis. In: Doi, K., Nishikawa, R.G., Scmidt, R.A. (eds.) Digital Mammography, pp. 155–160. Elsevier, Amsterdam (1996)
6. Rico, R., Muller, S.L., Peter, G.: Automatic scoring of CDMAN a dose study. Proc. SPIE 5034, pp. 164-173 (2003)

11. Trabajo II

Use of the cross-correlation component of the multiscale structural similarity metric (R* metric) for the evaluation of medical images

Gabriel Prieto,[a] Eduardo Guibelalde, and Margarita Chevalier
Department of Radiology, Faculty of Medicine, Complutense University, 28040 Madrid, Spain

Agustín Turrero
Department of Statistics and Operations Research, Faculty of Medicine, Complutense University, 28040 Madrid, Spain

(Received 26 July 2010; revised 9 June 2011; accepted for publication 10 June 2011; published 21 July 2011)

Purpose: The aim of the present work is to analyze the potential of the cross-correlation component of the multiscale structural similarity metric (R*) to predict human performance in detail detection tasks closely related with diagnostic x-ray images. To check the effectiveness of R*, the authors have initially applied this metric to a contrast detail detection task.
Methods: Threshold contrast visibility using the R* metric was determined for two sets of images of a contrast-detail phantom (CDMAM). Results from R* and human observers were compared as far as the contrast threshold was concerned. A comparison between the R* metric and two algorithms currently used to evaluate CDMAM images was also performed.
Results: Similar trends for the CDMAM detection task of human observers and R* were found in this study. Threshold contrast visibility values using R* are statistically indistinguishable from those obtained by human observers (F-test statistics: $p > 0.05$).
Conclusions: These results using R* show that it could be used to mimic human observers for certain tasks, such as the determination of contrast detail curves in the presence of uniform random noise backgrounds. The R* metric could also outperform other metrics and algorithms currently used to evaluate CDMAM images and can automate this evaluation task. [DOI: 10.1118/1.3605634]

Key words: MS-SSIM, model observer, mammography, image quality, CDMAM

I. INTRODUCTION

Image quality analysis plays a central role in the design of imaging systems for medical diagnosis. A great effort to develop meaningful metrics (lab and clinical), well correlated with imaging phantom studies and with clinical performance of the medical imaging systems has been made in the last few years. The final objective of these image quality metrics (IQM) is usually to design an algorithm able to score the perceived quality of a medical image. For phantom studies, the use of automatic tools that mimic the radiologist's point of view analyzing an x-ray image could avoid interobserver and intraobserver variability and minimize the great number of images, of observers and the great deal of time usually required to optimize the image acquisition parameters or to evaluate equipment or new technologies, for instance, by means of the receiver operating characteristic (ROC). So far only partial success has been achieved. The search for IQM that fully correlates with the quality perceived by the human visual system (HVS) and particularly with the radiologist's point of view is still an open question.

Certain widely used metrics such as the peak signal-noise ratio or the mean-squared error are very simple to calculate, but do not show a good correlation with the image quality perceived by human observers[1] and indeed they are not useful to deduce the capability of diagnostic equipment.[2] Other metrics closer to the actual performance of systems, such as the modulation transfer function, the noise power spectrum, the noise equivalent quanta, and the detection quantum efficiency describe much better the image formation process of the system and can be used not only to improve image quality but also to predict the observer response under the ideal observer model approach.[3] This model, based on the statistical theory of decision can only apply to simple tasks such as a "signal-known-exactly/background-known-exactly" ("SKE/BKE") detection task.[4] Moreover, the sensitivity of the ideal observer model is much higher than that of the human observer.

There are other models that have a better correlation with the human observer, which can also be applied to more complex tasks than SKE/BKE. These include mainly the channelized Hotelling observers, the nonprewhitening matched filter (NPW) and the NPW with an eye-filter.[5] However, for mammographic images, these models are not good predictors of human performance.[5]

There are other metrics such as the structural similarity (SSIM)[6] that have shown very good results mimicking the human performance in analyzing natural images in videos and still-images. These metrics are based on the perceptual visual theory proposed by Wang and Bovik[7] that considers the HVS highly adapted for extracting structural information from the scenes. A family of objective image quality assessment algorithms has been developed based on this premise.[6,8,9] They evaluate visual image quality by measuring the structural similarities between two images, one of them

being the reference one. This family includes the cross-correlation component of the multiscale structural similarity metric (R*),[9] that has been explicitly designed for recognition threshold tasks. Note that the radiologist's tasks usually use *reduced reference* or *no reference* metrics that require only a partial reference signal or none at all. However, in some specific situations, as the case presented in this paper, it is possible to model the "perfect image" and to use reference metrics to perform automatic tasks highly correlated with observer predictions.

Despite some criticisms of the SSIM family,[10] the R* metric shows some promising features that suggest the possibility of being successfully applied to medical image analysis tasks. As mentioned above, this family is designed and fully tested to analyze natural scenes, whose complexity is of the order or even greater than that of medical imaging. It has been successfully used for ensuring the quality and fidelity of the image in a large number of commercial and research applications. In particular, it surpasses most of the metrics currently used in the analysis of video and still image.[9] Moreover, some experiments prove that R* sensitivity for detecting image structures close to the perception threshold is analogous to that of human observers.[9]

To check the effectiveness of R*, we have initially applied this metric to a contrast detail detection task. For this, we developed an automatic evaluation tool based on the R* metric that was applied to score images of the CDMAM phantom.[11] Similarly to other authors,[12] we have made a comparison of our method with human-observer contrast-detail detection tasks as well as with other automatic evaluation algorithms based on the CDCOM software.[13,14]

II. THEORY

The R* metric belongs to the set of quality assessment (QA) algorithms that seek an objective evaluation of image quality consistent with subjective visual quality. These algorithms evaluate a test image X with respect to a reference image Y to quantify their similarity. In this sense, all of them (including R*) are signal known exactly (SKE) tasks. R* evaluates perceptual quality of the X image, referred to the test image Y, by computing a local spatial index, $r(x, y)$, that is defined[9] as follows:

X and Y being images to be compared (computed as matrixes of pixels) and $\mathbf{x} = \{x_i \mid i = 1, 2,\ldots, N\}$ and $\mathbf{y} = \{y_i \mid i = 1, 2,\ldots, N\}$ pairs of local square windows (computed as sub-matrixes of pixels) of X and Y, respectively, $\mathbf{x}$ and $\mathbf{y}$ are located at the same spatial position in both images. The index $r(x, y)$ is defined in terms of the pixel value standard deviations σ_x and σ_y at sub-matrixes $\mathbf{x}$ and $\mathbf{y}$ and the covariance σ_{xy} of $\mathbf{x}$ and $\mathbf{y}$:

$$r(x,y) = (\sigma_{xy})/(\sigma_x\sigma_y) \tag{1}$$

As can be seen, if sub-matrixes $\mathbf{x}$ and $\mathbf{y}$ cover the same object in the same location, r shows a maximum.

$r(x, y)$ takes values between -1 and 1. The closer the value of $r(x, y)$ to 1, the closer the similarity between sub-matrixes $\mathbf{x}$ and $\mathbf{y}$.

When the signal or the signal + background are uniform, σ_x or σ_y tend to be zero and the value of $r(x, y)$ is unstable. This is the case of sub-matrixes measured inside uniform reference signals, where all pixels take the same value and the variance is null. For these limits, the index calculation is made by supposing that $\sigma_x > 0$, and the sub-matrix $\mathbf{y}$ is uniform. Then, the variance of $\mathbf{y}$ is zero. Under these conditions, $\mathbf{x}$ does not correlate with $\mathbf{y}$, so the $r(x, y)$ value must be set to zero. When both sub-matrixes have equal variance, the $r(x, y)$ value must be set to 1. Thus, the alternative definition of the index is given as

$$r^*(x,y) = \begin{cases} 0 \text{ for } \sigma_x > 0 \text{ and } \sigma_y = 0, \text{ or } \sigma_y > 0 \text{ and } \sigma_x = 0 \\ 1 \text{ for } \sigma_x = \sigma_y = 0 \\ r(x,y) \text{ other} \end{cases} \tag{2}$$

As the model compares two images, the test (X) and the reference (Y), the sub-matrixes ($\mathbf{x}$, $\mathbf{y}$) are moved over X and Y and $r^*(x, y)$ values are calculated for each position. If X and Y contain the same object in the same location, $r^*(x, y)$ shows a maximum.

Detail perception depends, among other factors, on the resolution of the image and on the observer-to-image distance.[8] To incorporate M observer viewing distances, the algorithm simulates different spatial resolutions by iterative down-sampling in two steps: first, a low-pass filter is applied to reduce the bandwidth of the signal to avoid aliasing effects before the signal is down sampled, and second, the size of both images (reference and test) is reduced by a factor of 2, sub-sampling without any average (averaging is not needed after the low-pass filter is applied).

These two steps are iteratively applied $M-1$ times. (The original size of the image is taken as the first viewing distance. There is no need for downsampling for $M = 1$) The overall cross-correlation multiscale structural similarity metric R^* value is obtained by combining measurement at different scales according to the following expression:

$$R^* = \prod_{M}^{j=1} r_j^*(x, y) \tag{3}$$

III. MATERIALS AND METHODS

The CDMAM phantom (version 3.4, Artinis, St. Walburg 4, 6671 AS Zetten, The Netherlands) consists of an aluminum base with a matrix of gold disks of varying thicknesses and diameters, which is attached to a PMMA cover. The discs are arranged in a matrix of 16 rows by 16 columns. Within a row, the disk diameter is constant, with logarithmically increasing thickness. Within a column, the disk thickness is constant, with logarithmically increasing diameter. Each cell in the matrix contains two gold disks each with the same diameter and thickness. The reference signal is the disk at the center of the cell and the test signal is the disk randomly located in one of the four quadrants. The imaging task can be identified as a four-alternative-forced choice (4AFC) task, since the observer has to detect the quadrant of each cell in which a disk appears to be present. This phantom

is widely used and fully tested for image quality assessment in mammography.

A set of eight raw CDMAM images (set #1) were downloaded from the European Reference Organization for Quality Assured Breast Screening and Diagnostic Services (EUREF) web site.[14] The images were obtained with a GE Senographe 2000D at 27 kVp, 125 mAs and with a resolution of 1 pixel per 100 μm. The CDMAM images were scored by four experienced human observers. Each observer scored two different images once. The observer readouts are available at the same website.

A second set of 20 images (set #2) was obtained with another CDMAM unit. In this case the images were acquired with a Sectra MicroDose LD30 at 32 kVp and 50 μm pixel size. Scoring was performed by a panel of seven experts. Six observers scored three different CDMAM images once. The seventh observer scored two different CDMAM images once. The experience of the observers interpreting mammograms was at least 3 yrs.

Both data sets were evaluated according to the methodology, and rules for CDMAM scoring published and described in the phantom manual.[11] According to this methodology, the purpose of each observation is to determine, for each disk diameter, the threshold gold thickness (the "just visible" gold thickness). So in every column (same diameter) the last correctly indicated eccentric disk has been determined. Finally, the nearest neighbors correction (NNC) rules[11] are applied to the image readouts for smoothing the edges among cells that were correctly and noncorrectly evaluated. According to these rules, for every score there are three possibilities:

- True: the eccentric disk was indicated at the true position (TP).
- False: the eccentric disk was indicated at a false position (FP).
- Not: the eccentric disk was not indicated at all.

and two main rules:

- A"True" needs two or more correctly indicated nearest neighbors to remain a "True".
- A "False" or "Not" disk will be considered as "True" when it has three or four correctly indicated nearest neighbors.

These two main rules have minor and specific exceptions for those disks that have only two nearest neighbors (at the edges of the phantom).

The software tools here presented are written as a JAVA computer algorithm and integrated program (plug-in) for the display and image processing IMAGEJ software.[15] All images are captured or defined in a gray scale of 16-bits, with pixel values from 0 up to 65535.

III.A. R* metric application to CDMAM scoring

The first task to manage the disk information from the phantom images is the accurate detection of the grid line images, which form the matrix in which gold disks are distributed. Although several methods have been applied to find the grid position,[13,16] we used here an algorithm[17] developed by ourselves, which has been successfully proven even when slight distortions of the images are present.[18]

Once the grid lines are detected, the second step to be followed is the accurate detection of the disks in each matrix cell. The algorithm looks for the gold disks around the four quadrants near the grid crossing points by analyzing the structural similarity among the cells in the phantom image (image X in the "Theory" section) and in a reference mask image of the disks (image Y in the "Theory" section). The reference or mask image is a perfect white disk, with a pixel value of 65 535, inserted into a black background (margin), with a pixel value of 0, whose size matched the disk diameter to be evaluated [Fig. 1(a)].

The technical specifications of the phantom give the nominal disk distances from the grid crossing points. However,

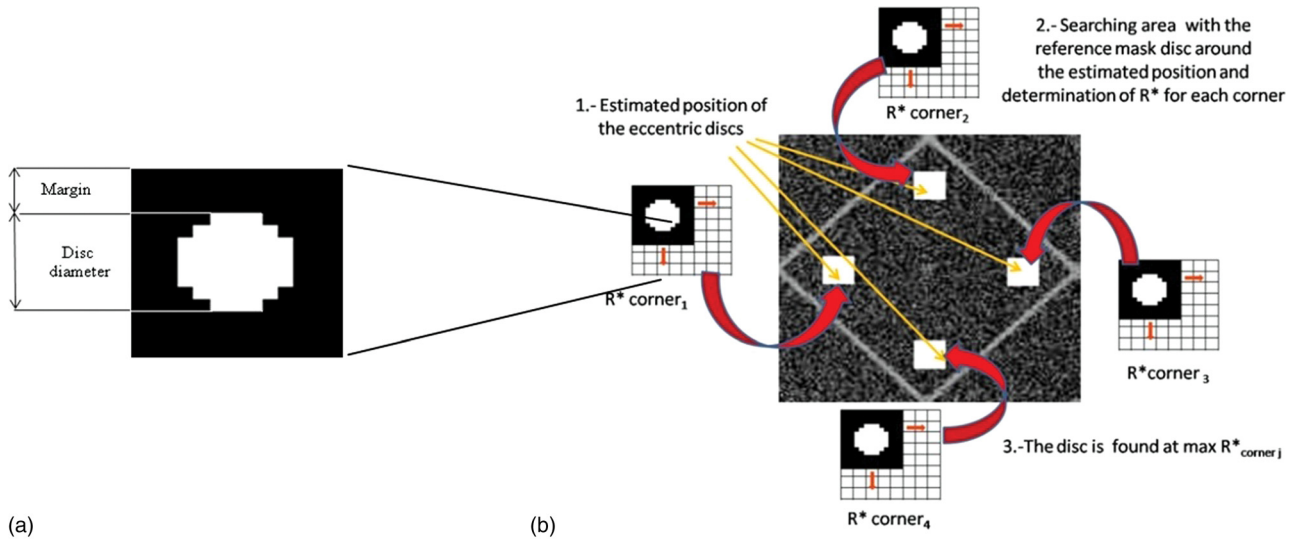
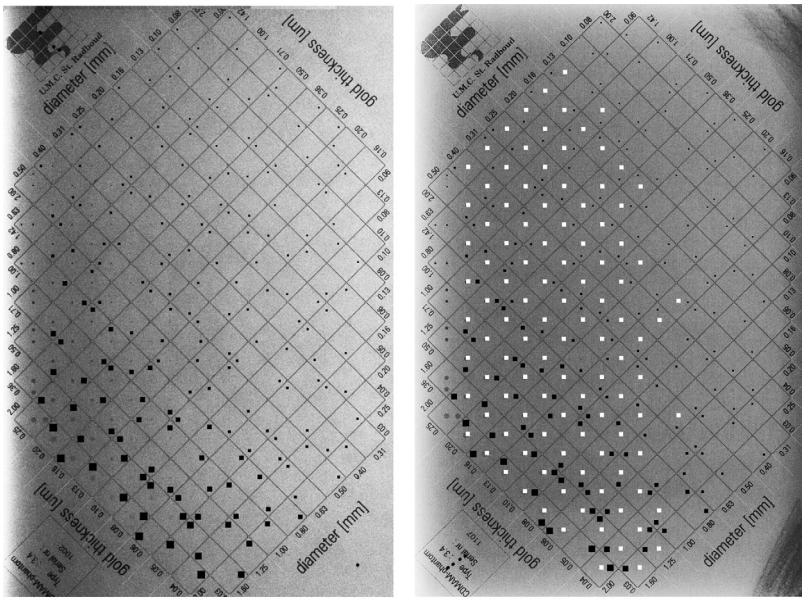


FIG. 1. Searching methodology. (a) Reference or mask image (b) Steps followed to search for the quadrant with the maximum R* i.e., most probably position of the eccentric disk.



(a) (b)

FIG. 2. (a) Graphical layout showing the predicted eccentric disk positions (black squares). (b) Graphical layout showing the correctly found eccentric disks. If the algorithm has found the eccentric disk, a white central square appears.

due to the manufacturing process, these distances can vary from unit to unit. Therefore, the R* value is calculated at 25 different positions (5×5) around the expected location of the disks at each cell quadrant [see Fig. 1(b)]. The maximum value of R* was adopted as the R* value for this cell quadrant. Then, the maximum value of the R* derived from the four quadrants determines the most probable position of the eccentric disk at each cell.

In the present work, the value of R* is calculated according to Eq. (3) where the value of M has been set to be a maximum of 5 for set #1 and of 6 for set #2, since after 5 and 6 (respectively) downsizing steps by a factor of 2, even the details of the largest disks disappear. (The larger disks of the CDMAM have a diameter of 2 mm. That means 20 pixels for set #1 and 40 for set #2, with resolutions of 100 and 50 μm per pixel, respectively. After 5 and 6, (respectively) downsizing by two, the diameter of these disks is less than 1 pixel and disappears in the image.)

Figure 2(a) shows the predicted disk location at each cell in a test image. Black squares are located at the quadrant with the maximum value of the R^*_j metric ($j = 1,..,4$). In Fig. 2(b) the white squares show the quadrants containing disks correctly identified (TP) by the algorithm (hits). Finally, the NNC rules were applied to the image readouts.

To compare the perception threshold of the R* algorithm and the human observer, Pearson correlation coefficients were calculated by comparing the thickness threshold for every image and for every diameter from the human observer and from R*, that is, this analysis was performed over the scatter plot of both variables (thickness and disk diameter) for the whole set of images.

The relationship between thickness and disk diameter was investigated by means of regression analyses in the two experiments. Comparisons of the models were carried out through the R^2 statistic. To overcome heterogeneity of variance, thickness data were log transformed.

III.B. Comparison with other methods

For comparison purposes, the sets of CDMAM images were also automatically evaluated by using two algorithms. The first one is CDCOM program,[13] which is a freely available[14] algorithm currently used for automatic evaluation of the CDMAM phantom. The second evaluation program, here named PRCDCOM, performs a smoothing and fitting of the readout matrixes produced by the CDCOM program following the procedures described by Young *et al.*[19] The threshold values derived with the two automatic methods were compared with those resulting from our algorithm.

The comparison of threshold values derived from the CDCOM and PRCDCOM methods with the values resulting from our algorithm and from the human observer was carried out through regression analyses. The obtained models were linearized and the comparison of the regression lines was studied by analyzing appropriate ANOVA tables. Statistical analyses were performed using SPSS® and STATGRAPHICS® statistical packages.

IV. RESULTS AND DISCUSSION

IV.A. R* metric application to CDMAM scoring. Threshold thickness calculations

Figure 3 shows, in a log–log graphic, the average threshold thickness for disk diameters ranging from 0.10 to 2.00 mm obtained by the experienced human observers (HO) and by applying the R* algorithm to the same sets of data.

IV.A.1. Correlation analysis

A strong linear relationship was observed between the thickness thresholds obtained from R* and human observers {Pearson coefficients $r = 0.9249$ in set #1 [Fig. 3(a)] and $r = 0.8922$ in set #2 [Fig. 3(b)]}.

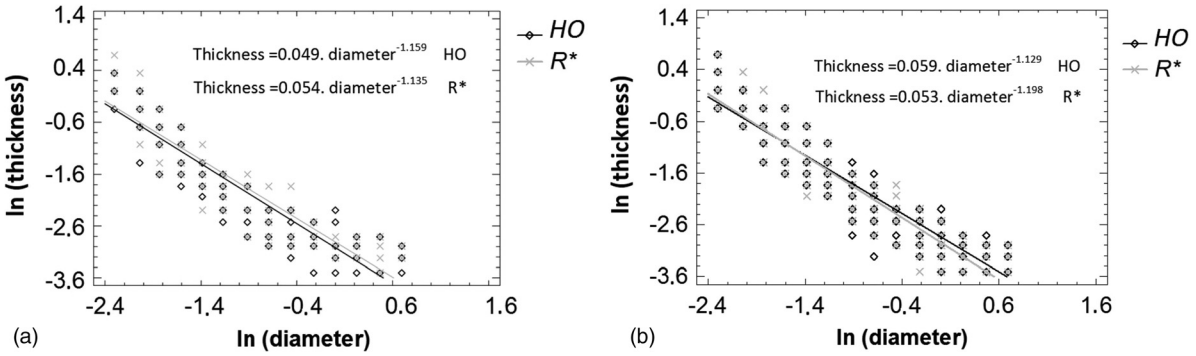


FIG. 3. Average threshold thickness as a function of the diameter from human observer (HO) and R* for set #1 (a) and set #2 (b).

IV.A.2. Regression analyses

The logarithm of thickness decreased with the disk diameter. The scatter plot suggests fitting a log–log model, that is, the approach is to consider a linear relationship among log-transformed variables; Figs. 3(a) and 3(b) show the results of these fits. The values of the R^2 statistic were 0.8624 and 0.8360 for HO and R*, respectively, in set #1 and 0.9020 and 0.8978 for HO and R*, respectively, in set #2. The statistical comparison of both regression lines shows no significant differences between them in set #1, according to the F-test statistics for the hypotheses of equality of intercepts and parallelism ($p = 0.1439$ and $p = 0.7117$, respectively) . The same comparison in set #2 shows results slightly nearer to statistical significance ($p = 0.085$ and $p = 0.065$), but always greater than statistical significance values ($p > 0.05$). These results suggest that R* could be used as a surrogate of the human observer with no evidence of statistical difference.

IV.B. Comparison with other methods

We have to point out at this juncture the range of validity for the CDCOM and the PRCDCOM programs. According to their developers,[13,19] these algorithms can only be applied to disk diameters equal to or smaller than 1.00 mm, so the graphics in Figs. 4(a) and 4(b) have been reduced from a maximum of 2.00 mm to a maximum of 1.00 mm to compare the four methods in the same range of experimental data.

The regression model which provides the best fit to the data derived from the four analyzed methods is the multiplicative or log–log model. Results are different for sets #1 and #2.

For set #1 [Fig. 4(a)], the four models fit quite well to the data (all R^2 statistics are greater than 0.93). The four regression lines are parallel with significant differences only between PRCDCOM and CDCOM threshold values (F-test statistic: $p = 0.0256$). Regarding the comparisons of intercepts, there are significant differences between the CDCOM method and the remaining ones (F-test statistics: all $p < 0.003$ for the equality of intercepts hypothesis). According to these results, PRCDCOM and R* could be adequate surrogates of the HO, but not CDCOM.

Similar results were found for set #2 [Fig. 4(b)] for the log–log model, (all R^2 statistics are greater than 0.96). In this case, the test for parallelism shows statistically significant differences between the CDCOM method and HO and R* methods ($p = 0.0002$ and $p = 0.0058$, respectively) and also between the PRCDCOM method and HO method (F-test statistic: $p = 0.011$). Regarding the comparisons of intercepts, there are significant differences between the CDCOM method and the remaining ones (all $p < 0.0002$).

Regarding Figs. 3(a) and 3(b), R* is valid for a larger range of diameters (up to 2.00 mm) than CDCOM and PRCDCOM with no statistically significant difference from the HO readouts.

According to these results, R* could be an adequate surrogate of the HO, but not PRCDCOM or CDCOM.

V. CONCLUSIONS

These results show that the R* metric can be used to mimic human observers for certain tasks, such as the

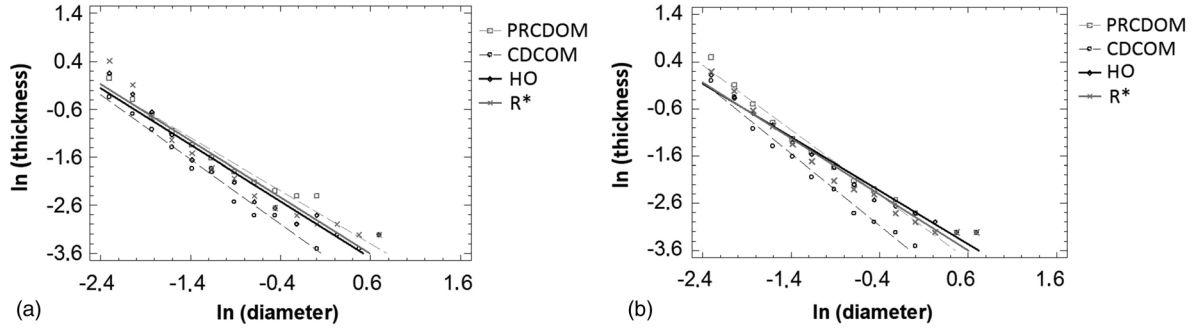


FIG. 4. Average threshold thickness as a function of the diameter from PRCDCOM, CDCOM, R*, and human observer (HO) for set #1 (a) and set #2 (b).

determination of contrast detail curves in the presence of uniform random noise backgrounds . The reliability of the results has been ensured by the similar threshold thickness obtained for each diameter by both observers, R* metric and HO, showing that both present a similar response independently of the signal, with no statistically significant difference.

Despite the fact that more samples and experiments should be carried out, the algorithm here designed based on R* metric could outperform other currently used metrics and algorithms used to evaluate CDMAM images, such as CDCOM and PRCDCOM and could be applied to the same range of disk diameters as the HO.

These results demonstrate the possibility of applying the R* metric to the medical imaging area of research applying adequate experimental conditions and methodology.

ACKNOWLEDGMENTS

The authors want to acknowledge the collaboration of Secträs Image Processing Department, especially Björn Svensson for his collaboration delivering images and human observer readouts for this experiment. They would like to thank the following people for their initial comments and opinions which encouraged us to deepen our knowledge of the possibilities and limitations of the metric considered here: David A. Clunie, Sheila S. Hemami, Elizabeth A. Krupinski, Wayne S. Rasband, David M. Rouse, and Zhou Wang. They also would like to thank Michael P. Kennedy for his help reviewing our English grammar and syntax.

[a]Author to whom correspondence should be addressed. Electronic mail: gprietor@med.ucm.es

[1]B. Girod, "What's wrong with mean-squared error," in *Digital Images and Human Vision,* edited by A. B. Watson (MIT, Cambridge, MA, 1993), pp. 207–220.

[2]A. E. Burgess, "The Rose model, revisited," J. Opt. Soc. Am. A **16**, 633–646 (1999).

[3]K. J. Myers, "Ideal observer models of visual signal detection," in *Handbook of Medical Imaging, Physics and Psycophysics*, edited by J. Beutel, H. Kundel, and R. Van Metter (SPIE, Bellingham, WA, 2000), Vol. 1, pp. 558–592.

[4]H. H. Barrett, K. J. Myers, and R. F. Wagner, "Beyond signal detection theory," application of optical instrumentation in medicine XIV and Picture Archiving and Communications (PACS IV) for medical applications, Newport Beach, CA. *Proceedings of the Society of Photo-optical Instrumentation Engineers,* (Bellingham, WA, 1986), Vol. 626, pp. 231–239.

[5]M. P. Eckstein, C. K. Abbey, and F. O. Bochud, "A practical guide to model observers for visual detection in synthetic and natural noisy images," in *Handbook of medical imaging, physics, and psychophysics*, edited by J. Beutel, H. Kundel, and R. Van Metter (SPIE, Bellingham, WA, 2000), Vol. 1, pp. 593–626.

[6]Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," IEEE Trans. Image Process. **13**, pp. 600–612 (2004).

[7]Z. Wang and A. C. Bovik, "Why is image quality assessment so difficult?," IEEE Trans. Acoust., Speech, Signal Process. **4**, 3313–3316 (2002).

[8]Z. Wang, E. Simoncelli, and A. Bovik, "Multi-scale structural similarity for image quality assessment," in *Proceedings of the 37th IEEE Asilomar Conference on Signals, Systems, and Computers*, Pacific Groove, Ca, IEEE (2003), pp. 529–554.

[9]D. M. Rouse and S. S. Hemami, "Analyzing the Role of Visual Structure in the Recognition of Natural Image Content with Multi-Scale SSIM," Proc. SPIE **6806**, 680615 (2009).

[10]R. Dosselmann and X. D. Yang, "A comprehensive assessment of the structural similarity index," *Signal, Image, and Video Processing* (Springer, London, 2009), Vol. 51, pp. 81–91.

[11]K. R. Bijkerk, M. A. Thijssen, and T. H. Arnoldussen, "Manual CDMAM-Phantom Type 3.4" (translation from the Dutch by S. van Woudenberg), University Medical Centre Nijmegen, July 2000.

[12]R. M. Gagne, B. D. Gallas, and K. J. Myers, "Toward objective and quantitative evaluation of imaging systems using images of phantoms," Med. Phys. **33**, 83–95 (2006).

[13]N. Karssemeijer and M. A. O. Thijssen, "Determination of contrast-detail curves of mammography systems by automated image analysis," in *Digital Mammography*, edited by K. Doi, R. Giger, R. M. Nishikawa, and R. A. Scmidt (Elsevier, Amsterdam, 1996), pp. 155–160.

[14]R. Visser and N. Karssemeijer, "CDCOM Manual: software for automated readout of CDMAM 3.4 images," CDCOM software, manual, and sample images are posted at www.euref.org, Last accessed June 2010.

[15]W. S. Rasband, ImageJ, U. S. National Institutes of Health, Bethesda, MD, http://rsb.info.nih.gov/ij/plugins/index.html 1997–2007, Last accessed June 2011.

[16]R. Rico, S. L. Muller, and G. Peter, "Automatic scoring of CDMAN a dose study," Proc. SPIE, **5034**, 164–173 (2003).

[17]G Prieto, M. Chevalier, and E. Guibelalde, "CDMAM image phantom software improvement for human observer assessment," in *Lecture Notes in Computer Science 5116 Digital mammography*, edited by E. A. Krupinski (Springer-Verlag, Berlin, Heidelberg 2008), Vol. 5116, pp. 181–187.

[18]G. Prieto, M. Chevalier, and E. Guibelalde, "A software tool to measure the geometric distortion in x-ray image systems," Proc. SPIE, **7622**, p. 173 (2010).

[19]K. C. Young, J. J. H. Cook, J. M. Oduko, and H. Bosmans, "Comparison of software and human observers in reading images of the CDMAM test object to assess digital mammography systems," Proc. SPIE, **6142**, 614206 (2006).

12. Trabajo III

A SOFTWARE TOOL TO COMPARE CONTRAST-DETAIL DETECTION IN UNIFORM AND IN REAL MAMMOGRAPHIC BACKGROUNDS

Gabriel Prieto, Margarita Chevalier, Eduardo Guibelalde
Dept. Radiología. Fac. Medicina. Universidad Complutense de Madrid.
28040 Madrid (Spain)
gprietor@med.ucm.es, chevalier@med.ucm.es, egc@med.ucm.es

ABSTRACT

A software tool is presented to merge CDMAM phantom images with real mammographic backgrounds. It allows SKE tasks in uniform and in real backgrounds. This kind of tasks can be used to compare human, human visual metric or model observer performance in detail detection using uniform or mammographic backgrounds.

As it is very well known, local characteristics of the structures in real mammographic backgrounds reduce the human performance in contrast-detail detection tasks. In consequence that performance cannot be inferred from the data acquired in white noise (flat) backgrounds such as a CDMAM phantom produces.

It is of interest to compare the response of a mammography system to the same set of signals, either embedded in flat or in real backgrounds. This comparison achieves two goals. The first one is to analyze the variation of the recognition threshold of the system for both backgrounds. The second one is to analyze the performance of a human observer or a model observer over the same set of signals, varying the nature of the backgrounds.

The software tool presented here uses CDMAM images to merge with a region of interest selected from a real mammography. This region as well as the mixing image method (basically adding or multiplying pixels) can be freely selected by the user. In this work a set of measurements of 8 images has been analyzed. We can preview the variation of the contrast-detail detection for a human observer and a human visual system metric (R*).

Keywords: Medical image perception, observer performance, mammography, CDMAM.

1. DESCRIPTION OF PURPOSE

The software tool presented here has been designed and developed to test the human observer performance in uniform and structured backgrounds, but also to compare the human observer behaviour vs. different Human Visual System metrics or model observers. By way of illustration of this ability, we have applied it to compare the performance of a Human Visual System metric, R* [1] and the human observer (HO) in real backgrounds. The main reason to choose this metric is that its behaviour in uniform backgrounds is very similar to that of the human observer.[2] Therefore, we have a reference to compare results when realistic backgrounds are considered.

Medical Imaging 2011: Image Perception, Observer Performance, and Technology Assessment,
edited by David J. Manning, Craig K. Abbey, Proc. of SPIE Vol. 7966, 79661D · © 2011 SPIE
CCC code: 1605-7422/11/$18 · doi: 10.1117/12.877460

2. MATERIAL AND METHODS

2.1. Background selection

The algorithm requires one CDMAM[3] image (Figure 1) and one background image, such as the one shown in Figure 2. The operator can choose by clicking the mouse, the upper corner of the ROI he wants to merge. First, the algorithm automatically locates the grid crossing points of a CDMAM image by applying a method previously designed.[2, 4, 5] It is demonstrated this method is simple, time effective and shows a very small error in the calculation of the crossing points (< 3 pixels) and the cell dimensions. According to these dimensions, the algorithm builds the rhombus that will be merged with each cell in the phantom and shows it to the operator (Figure 3). The operator can accept this ROI or can create a new one, clicking on the mammographic image on another position.

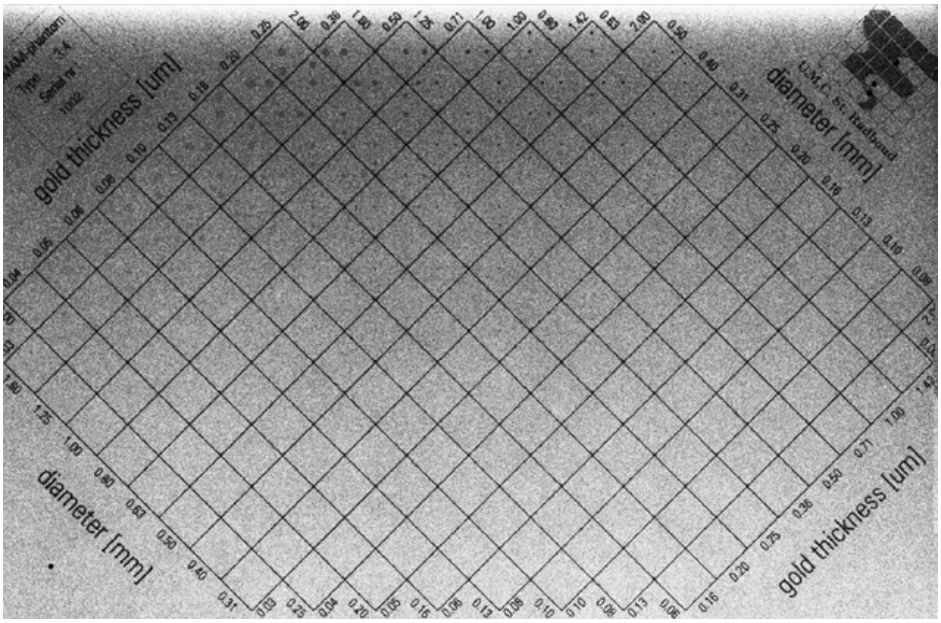


Figure 1. CDMAM image.

2.2 Merging methods

The system allows two methods to merge the phantom image with the mammographic background (Figure 4): linear additive or multiplicative. In case of linear additive, the algorithm computes the mean luminance value of the pixels of the selected rhombus in the mammographic background and takes this value as zero signal. This mean value is subtracted from the luminance value of each pixel of the selected rhombus. This gives a new rhombus with negative and positive values that is added, pixel by pixel, to each cell in the CDMAM image.

In the multiplicative method, the luminance value of the pixels in the mammographic background is divided by the average pixel value of the background (inside of the selected rhombus). These values are multiplied, pixel by pixel, to

each cell in the CDMAM image. In addition, it can be introduced an attenuation factor for both methods increasing or decreasing the percentage of merging (Figure 4).

The selected ROI will be merged with each one of the CDMAM cells. Figure 5 and 6 show respectively the output of the algorithm and a detail of this new phantom image.

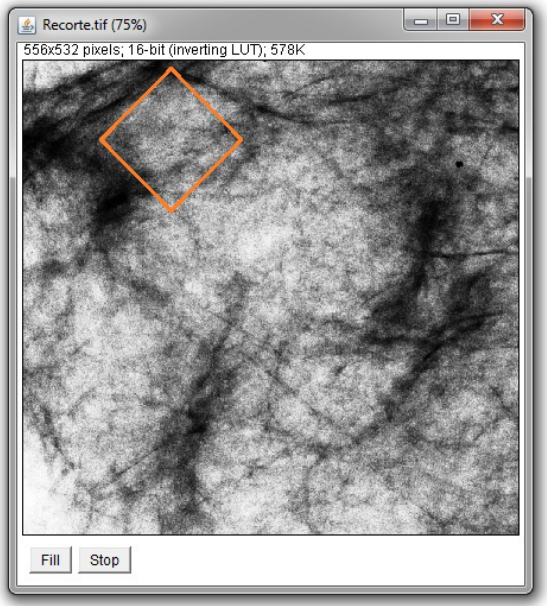


Figure 2. Mammographic image

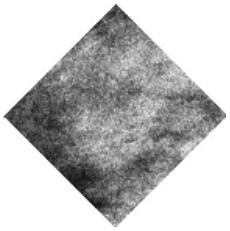


Figure 3. Detail. Rhombus selection inside the background image

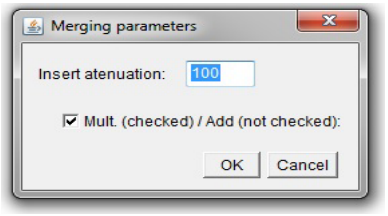


Figure 4. Merging method and percentage of attenuation

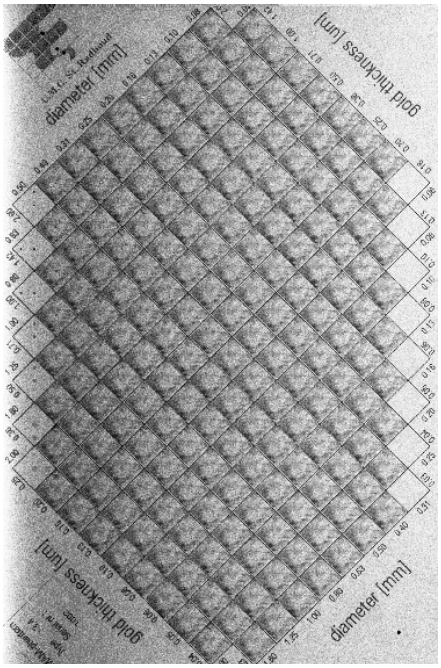


Figure 5. Result of merging: CDMAM image superimposed with mammographic background

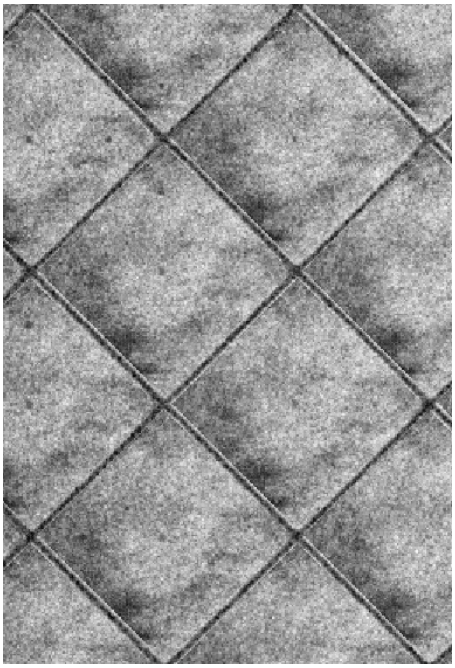


Figure 6. Result of merging, detail. Note the CDMAM discs inside the cells.

Every step in this method takes less than 0.5 seconds in a laptop with a processor Intel Core 2 Duo P8600, 4 GB RAM, Windows 7, Home Edition, 64-bits. We have developed our algorithms as a Java plugin to be used with ImageJ, the image manipulation program developed by Wayne Rasband. [6]

2.3. Set of images.

We have applied this software tool to one set of 8 CDMAM images, with the aim to obtain a preliminary result to test the efficiency of the method. These images were downloaded from the European Reference Organization for Quality Assured Breast Screening and Diagnostic Services (EUREF) web site.[7] The images were obtained with a GE Senographe 2000D at 27 kVp, 125 mAs and with a resolution of 1 pixel per 100 μm. These images were scored, according to the CDMAM rules of scoring,[3] by one observer with an experience of three years reading this kind of phantoms.

2.4. Performance comparison: relative efficiency.

To compare the performance of the HO and R*, we have applied the Constant Efficiency method. [8, 9] According to this method, the relative efficiency of R* versus the HO is defined as

$$Pc' = (Pc \text{ Human Observer} / Pc\ R^*)^2 \qquad (1)$$

where Pc represents the figure of merit Proportion Correct.[10]

If the model is a good predictor, the relative efficiency Pc' should be approximately constant across the different conditions of the experiment;[10] in this work these conditions are the different diameters of the CDMAM discs.

RESULTS AND DISCUSSION

Figure 7 shows the contrast-detail curves obtained in four different conditions:

1) **R* Uniform** is obtained applying a metric (R*) to the mentioned set.
2) **HO uniform** is the corresponding contrast-detail curve for a human observer.
3) **R* Mammo** is the contrast-detail curve of R* obtained merging the previous CDMAM images with a mammographic background (this background is shown in figure 4)
4) **HO Mammo** is the corresponding contrast-detail curve for a HO. Note that the performance of both observers is almost the same for all disc sizes in uniform backgrounds, according to previous results.[2]

We have also computed the Relative efficiency (Pc') of the HO versus the metric R*. These results are shown in Figure 8. **Pc' Uniform** is the Relative efficiency of HO versus the metric R* in uniform backgrounds. **Pc' Mammo** is the corresponding ratio in mammographic backgrounds. Continuous lines are guides to the eye with no theoretical significance.

Attending to figure 7, three relevant facts can be observed in mammographic backgrounds. The first one, as expected, is the lower response of the human observer in mammographic backgrounds related to uniform ones, due to the structured noise in the image.

The second fact is the lower response of the R* metric in mammographic backgrounds related to the response of R* metric in the uniform backgrounds, as can be expected, due also to the higher (structured) noise presented in the image.

The third fact is that contrast threshold increases as disc size increases for disc diameters greater than 1.25 mm, due to the masking effects of the structures of the mammographic background. These structures have a size similar to the greater discs. This fact has also been reported by several researchers.[11, 12]

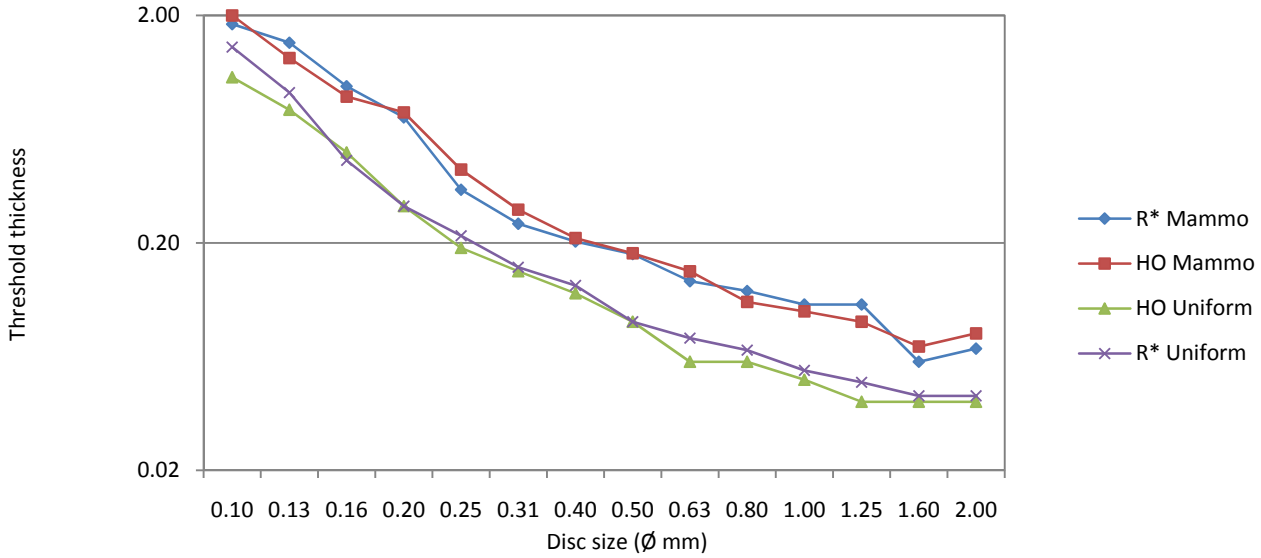


Figure 7. Performance variation in contrast-detail task for a human observer (HO) and a human visual system metric (R*) in uniform and mammographic (Mammo) backgrounds.

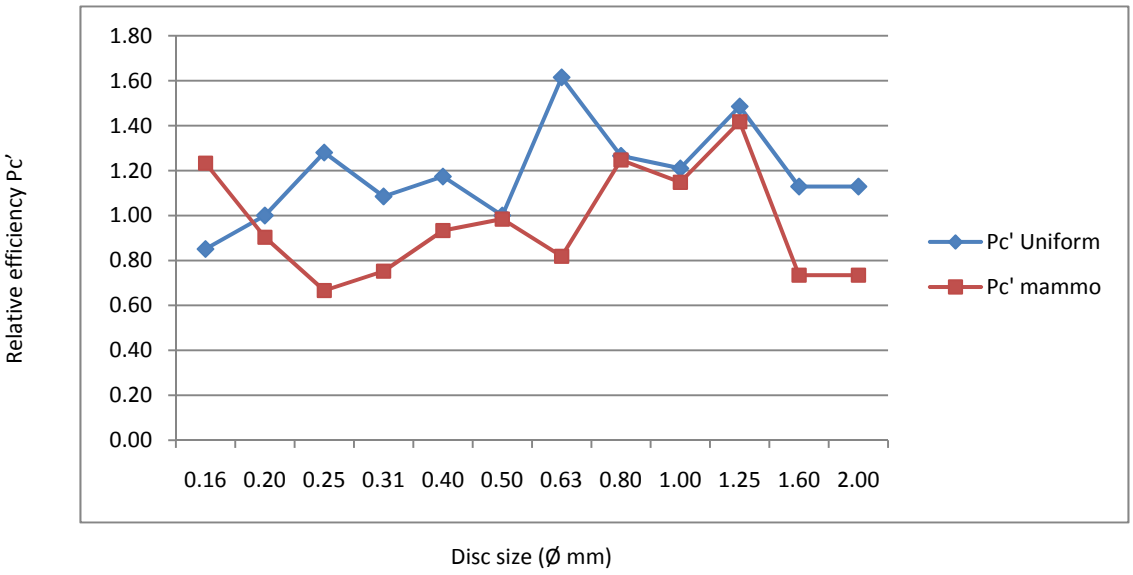


Figure 8. Relative efficiency Pc' in uniform and mammographic (Mammo) backgrounds.

Figure 8 shows that both observers have on average a similar sensitivity. Pc' has a mean value of 0.96 for mammographic backgrounds and 1.19 for uniform ones.

Figure 8 also shows that for diameters greater than 1.25 mm, the HO performance becomes much better than the R* metric performance. Other authors[11] have found a similar effect that is explained in terms of human readers performance. For details larger than 1mm, it seems that human readers could rely more on the disc edge than on the disc contrast. Moreover, a similar effect and explanation was also found by A. E. Burgess,[13] showing better human detection using as inserts flat discs, with a high gradient of contrast at the edge, than using spheres, with a lower gradient of contrast at the edge than a flat disc. This fact implies the need to run this kind of test with real lesions (microcalcifications and benign or malign masses) better than with unrealistic sharp edge discs.

CONCLUSIONS

The new software tool presented in this work can be used to generate hybrid images merging CDMAM images and real mammographic backgrounds and to compare the performance of different observers (human or automated) for contrast-detail detection. Its application to an actual problem confirms the results obtained, similar to others well known by the scientific community, and shows its potential as tool of analysis of the performance of different observers.

REFERENCES

[1] D. M. Rouse and S. S. Hemami, "Analyzing the Role of Visual Structure in the Recognition of Natural Image Content with Multi-Scale SSIM", *Proceeedings of SPIE* **6806** (2009).

[2] G. Prieto, M. Chevalier, and E. Guibelalde, "Automatic scoring of CDMAM using a model of the recognition threshold of the human visual system: R*", *Image Processing*, 2009. ICIP 2009. 16th IEEE International Conference on. 7-11 Nov. pp. 2489-2492, (2009).

[3] N. Karssemeijer, M.A.O. Thijssen "Determination of contrast-detail curves of mammography systems by automated image analysis", *Digital Mammography*, ed. Doi K, Giger R, Nishikawa, Scmidt R A., Elsevier, Amsterdam, pp. 155-160, (1996).

[4] G. Prieto, M. Chevalier, and E. Guibelalde, "A CDMAM Image Phantom Software Improvement for Human Observer Assessment". E.A. Krupinski (Ed.): *IWDM 2008*, LNCS **5116**, pp. 181–187, Springer-Verlag Berlin Heidelberg (2008).

[5] G Prieto, M Chevalier, and E Guibelalde, "A software tool to measure the geometric distortion in x-ray image systems", *Proceedings of SPIE*, **7622**, pp. 7622-173 (2010).

[6] W. S. Rasband, ImageJ, U. S. National Institutes of Health, Bethesda, Maryland, USA, http://rsb.info.nih.gov/ij/. (1997-2008).

[7] Sample images are posted at www.euref.org 2th December 2010.

[8] W. P. Tanner and T. G. Birdsall, ''Definitions of *d*' and η as psychophysical measures,'' *J. Acoust. Soc. Am*. **30**, 922–928 (1958).

[9] K. J. Myers, H. H. Barrett, M. C. Borgstrom, D. D. Patton, and G. W. Seeley "Effect of noise correlation on detectability of disc signals in medical imaging", *J. Opt. Soc. Am*. A, **2**, 1752 - 1759 (1985).

[10] M. P. Eckstein, C. K. Abbey, F. O. Bochud, "A practical guide to model observers for visual detection in synthetic and natural noisy images", in [*Handbook of Medical Imaging, Physics and Psychophysics*], edited by J. Beutel, H. Kundel, R. Van Metter (SPIE, Bellingham, WA, 2000), Vol 1, Physics and Psychophysics, 593-626 (2000).

[11] B. Grossjean and S. Muller, "Impact of textured background on scoring of simulated CDMAM phantom", Susan M. Astley et al (Eds.), *IWDM 2006*, LNCS **4046**, pp.460-467, Springer-Verlag Berlin Heidelberg (2006).

[12] A.E. Burgess, F.L. Jacobson, and P.F. Judy, "Human observer detection experiments with mammograms and power-law noise", *Med. Phys.*, 28(4), pp.419-437, (2001).

[13] A.E. Burgess, "Evaluation of detection model performance in power-law noise", Proceedings of SPIE, Vol.4324, pp123-132, (2001).

13. Trabajo IV

PLOS ONE

Structural Similarity Index Family for Image Quality Assessment in Radiological Images

--Manuscript Draft--

Manuscript Number:	
Article Type:	Research Article
Full Title:	Structural Similarity Index Family for Image Quality Assessment in Radiological Images
Short Title:	SSIM index family for Image Quality Assesment in Radiological Images
Corresponding Author:	GABRIEL PRIETO UNIVERSIDAD COMPLUTENSE DE MADRID Madrid, Madrid SPAIN
Keywords:	Perception and quality model; quality metrics and assessment tools; x-ray imaging; MRI; SSIM
Abstract:	Objectives: The Structural Similarity Index (SSIM) family is a set of metrics that has shown good agreement with human observers in tasks using reference images. These metrics analyze the viewing distance, the edge information between the reference and the test images, the changed and preserved edges, the textures, and the structural similarity of the images. In this paper seven novel metrics, based on that family, are proposed. This new set of metrics, together with another nine well-known SSIM family metrics, was tested to predict human performance in some specific tasks closely related with the evaluation of radiological medical images. Materials and Methods: We used a database of radiological images, comprising different acquisition techniques (MR and Plain Films). This database was distorted with four different types of distortions (Gaussian blur, Gaussian noise, JPEG, and JP2000), and five different levels of degradation. These images were analyzed by a board of radiologists with a double-stimulus methodology and their results were compared to those obtained from the 16 metrics analyzed and proposed in this research. Results: Our experimental results showed that the human observer readings were sensitive to the edge information between the reference and the test images, the changed and preserved edges, and the textures. Previous studies, that mixed these techniques, have shown the low relevance of using multi-scale approaches, simulating different viewing distances from the image to the observer. On the contrary, we have found the superiority of this approach over single-scale approaches, which take into account only one viewing distance. Conclusion: These results showed that several metrics (4-G-SSIM, 4-MS-G-SSIM, 4-G-r*, and 4-MS-G-r*) can be used as good surrogates of a radiologist to analyze the medical quality of an image in an environment with a reference image. Specially 4-MS-G-SSIM keeps an excellent performance for all types of image and distortion.
Order of Authors:	GABRIEL PRIETO
	EDUARDO GUIBELALDE
	ALBERTO MUÑOZ
	NIEVES GOMEZ-LEON
	AGUSTIN TURRERO
Opposed Reviewers:	
Additional Information:	
Question	**Response**
Financial Disclosure Please describe all sources of funding that have supported your work. A complete funding statement should do the	The autors received no specific funding for this work

GABRIEL PRIETO RENIEBLAS
DEPARTAMENTO DE RADIOLOGÍA Y MEDICINA FÍSICA.
FACULTAD DE MEDICINA. UNIVERSIDAD COMPLUTENSE DE MADRID.
AVDA. COMPLUTENSE s/n. 28040. MADRID. SPAIN.
gprietor@med.ucm.es Tlf.: +34913941555 Mobile: +34609362029

October, 18th 2015

Dear Sirs,

I am attaching the paper entitled "Structural Similarity Index Family for Image Quality Assessment in Radiological Images" by Gabriel Prieto, Agustín Turrero, Alberto Muñoz, Nieves Gómez-León, and Eduardo Guibelalde for submission to Plos ONE with the aim of being published in your highly appreciated journal.

We suggest Mr. Christian von Falck (*Hannover Medical School, Department of Diagnostic and Interventional Radiology)* as Editor. We also suggest as Editor those related in the on-line form we have filled: Mr. Derek Abbott, Mr. Kevin J. Croce, Mr. Nathan Jeffery, and Mr. Daniel L. Rubin.

Sincerely,

- Gabriel Prieto Renieblas -

Structural Similarity Index Family for Image Quality Assessment in Radiological Images

Gabriel Prieto[1*], Agustín Turrero[2], Alberto Muñoz[1], Nieves Gómez-León[3], Eduardo Guibelalde[1]

1 Department of Radiology, Faculty of Medicine, Complutense University, 28040 Madrid, Spain

2 Department of Statistics and Operations Research, Faculty of Medicine, Complutense University, 28040 Madrid, Spain

3 Department of Radiology, Princesa Hospital. Autónoma University, 28049 Madrid, Spain

Abstract

Objectives: The Structural Similarity Index (SSIM) family is a set of metrics that has shown good agreement with human observers in tasks using reference images. These metrics analyze the viewing distance, the edge information between the reference and the test images, the changed and preserved edges, the textures, and the structural similarity of the images. In this paper seven novel metrics, based on that family, are proposed. This new set of metrics, together with another nine well-known SSIM family metrics, was tested to predict human performance in some specific tasks closely related with the evaluation of radiological medical images.

Materials and Methods: We used a database of radiological images, comprising different acquisition techniques (MR and Plain Films). This database was distorted with four different types of distortions (Gaussian blur, Gaussian noise, JPEG, and JP2000), and five different levels of degradation. These images were analyzed by a board of radiologists with a double-stimulus methodology and their results were compared to those obtained from the 16 metrics analyzed and proposed in this research.

***Results*:** Our experimental results showed that the human observer readings were sensitive to the edge information between the reference and the test images, the changed and preserved edges, and the textures. Previous studies, that mixed these techniques, have shown the low relevance of using multi-scale approaches, simulating different viewing distances from the image to the observer. On the contrary, we have found the superiority of this approach over single-scale approaches, which take into account only one viewing distance.

Conclusion: These results showed that several metrics (4-G-SSIM, 4-MS-G-SSIM, 4-G-r*, and 4-MS-G-r*) can be used as good surrogates of a radiologist to analyze the medical quality of an image in an environment with a reference image. Specially 4-MS-G-SSIM keeps an excellent performance for all types of image and distortion.

Citation:

Editor:

Received: **Accepted:** **Published:**

Funding: This research was developed without any specific fund.

Competing interests: The authors have declared that no competing interests exist.

* Corresponding author. E-mail: gprietor@med.ucm.es (GP)

Introduction

Image quality analysis plays a central role in the design of imaging systems for medical diagnosis. The final objective of this image quality analysis is usually to design a metric able to score the perceived quality of a medical image: an image quality metric (IQM). So far only partial success has been achieved.

Certain widely used metrics such as the peak signal-noise ratio (PSNR) or the mean square error (MSE) are very simple to calculate, but do not show a good correlation with the image quality perceived by human observers [1] and they are not useful to deduce the diagnosis capability of diagnostic equipment [2].

Other metrics closer to the actual performance of systems, such as the Modulation Transfer Function (MTF), the Noise Power Spectrum (NPS), the Noise Equivalent Quanta (NEQ), and the Detection Quantum Efficiency (DQE) describe much better the image formation process of the system and can be used to predict the observer response under the Ideal Observer Model approach [3]. However, this model can only be applied to tasks such as a "signal-known-exactly/statistically / background-known-exactly/statistically" ("SKE/BKE" or "SKS/BKS") detection task [4].

There are other models that have a good correlation with the human observer which can also be applied to SKE/BKE or SKS/BKS tasks or even more complex tasks. These include mainly the Fisher-Hotelling channelized models, [5] the Non-PreWhitening Matched Filter (NPWMF), [6] and the NPW with an eye-filter [7]. These models attempt to

reproduce human performance in different tasks, taking into account functions to mimic the contrast sensitivity function of the human eye (eye filter) or neuronal visual perception paths (channels). These models are quite useful in image quality assessment for certain acquisition techniques and types of noise [8]. However, in this study we are looking for a general index of image quality, independent of the acquisition technique or the type of noise present in the image, despite the fact this index could be less accurate for a certain type of noise or for a certain acquisition technique than these models.

In the perceptual visual theory proposed by Wang et al. [9] the human visual system (HVS) is considered to be highly adapted for extracting structural information from a scene, and therefore a measure of Structural SIMilarity (SSIM) should be a good approximation of perceived image quality. A family of objective IQM has been developed based on this premise [10, 11, 12, 13, 14, 15]. They evaluate visual image quality by measuring the structural similarities between two images, one of them being the reference one. A multi-scale version of SSIM (MS-SSIM) has also been proposed [10].

Results in large studies have shown that SSIM and MS-SSIM mimic quite well the perceived quality of an image by a human observer. However, they show some limitations:

1) Some researchers have found [11] that SSIM and MS-SSIM do not perform so well for recognition threshold tasks (tasks near the perception limit), which invalidate their application to the analysis of images with regions of interest at the limit of visibility.
2) Some studies show limits in the performance of these indexes analyzing medical images [16, 17].
3) Other studies show that the correlation between SSIM and MS-SSIM and human observers decreases when they are used to measure the quality of blurred and noisy images [13, 15].

These drawbacks are limiting factors in the medical imaging area, specifically in Radiology. Radiological images of medical interest show subtle differences between the image with no pathological findings and the image that shows these findings. Blur and noise are some of the most usual distortion factors in a day-to-day radiological practice.

Some authors have proposed some modifications of SSIM and MS-SSIM to avoid these limitations. Rouse and Hemami [11] proposed a new IQM, r*, based on the structural component of MS-SSIM that could avoid the lack of effectiveness near the recognition threshold. Chen et al. [13] proposed a gradient-based SSIM (G-SSIM) that improves the SSIM results in blurry and noisy images. Li and Bovik [15] applied a four-component model based on the texture

and edge regions of the image. They applied this model to SSIM and MS-SSIM, getting eight new IQM. These three approaches have shown promising features to overcome the limitations of SSIM and MS-SSIM.

The aim of the present work is to analyze the potential of these modifications in the SSIM family, testing in a medical environment a complete set of proven and new IQM proposed here, the latter created by combination of all the related approaches.

To check the effectiveness of these IQM, we have applied these metrics to a double-stimulus task with a database of radiological images. We have compared these results with those obtained from a board of expert radiologists.

Materials and Methods

Metrics

SSIM

The Structural SIMilarity index [9] evaluates a test image X with respect to a reference image Y to quantify their visual similarity. In this sense, it is a Signal Known Exactly (SKE) task. SSIM evaluates the quality of the X image, referred to the test image Y, by computing a local spatial index that is defined as follows:

X and Y being images to be compared (computed as matrices of pixels) and $x = \{x_i \mid i = 1, 2,\ldots, N\}$ and $y = \{y_i \mid i = 1, 2,\ldots, N\}$ pairs of local square windows (computed as sub-matrices of pixels) of X and Y respectively, x and y are located at the same spatial position in both images. SSIM is defined in terms of the average pixel values, μ_x and μ_y, the pixel value standard deviations σ_x and σ_y at patches x and y and the covariance (cross-correlation) σ_{xy} of x and y through the following indexes:

$$l(x,y) = (2\mu_x\mu_y + C1)/(\mu_x^2 + \mu_y^2 + C1) \quad (1)$$

$$c(x,y) = (2\sigma_x\sigma_y + C2)/(\sigma_x^2 + \sigma_y^2 + C2) \quad (2)$$

$$r(x,y) = (\sigma_{xy} + C3)/(\sigma_x\sigma_y + C3) \quad (3)$$

where C1, C2, and C3 are constants introduced to avoid instabilities when $(\mu_x^2 + \mu_y^2)$, $(\sigma_x^2 + \sigma_y^2)$ or $\sigma_x\sigma_y$ are close to zero.

The l(x,y) index is related with luminance differences, c(x,y) with contrast differences, and r(x,y) with structure variations between x and y.

The general form of the SSIM index is defined as:

$$SSIM(x, y) = [l(x, y)]^{\alpha} \cdot [c(x, y)]^{\beta} \cdot [r(x, y)]^{\gamma} \qquad (4)$$

where α, β, and γ are parameters that define the relative importance of each component. SSIM(x,y) ranges from 0 (completely different) to 1 (identical patches). Finally a mean SSIM index is computed to evaluate the global image similarity.

Despite its simple mathematical form, SSIM objectively predicts subjective ratings as well as more sophisticated IQM [9] even for medical images [18, 19, 20]. However, SSIM does not match very well the observer's prediction in noisy and blurred images, images near the recognition thresholds, or for some medical images. Some modifications have been proposed to avoid these limitations.

Multi-scale index: MS-SSIM

Detail perception depends, among other factors, on the resolution of the image and on the observer-to-image distance. To incorporate M observer viewing distances, Wang et al. developed a multi-scale SSIM (MS-SSIM) index [10]. MS-SSIM simulates different spatial resolutions by iterative down-sampling and weighting the different values of each component of SSIM (luminance, contrast, and structure) at different scales. This index has been proven to be more accurate than SSIM for certain conditions [10, 11].

The recognition threshold: r*

Rousse and Hemami [11] proposed a cross-correlation multi-scale structural similarity metric (r*) based on the structural component of MS-SSIM (r(x,y) in (3)). They found out that the structural component was more related to human perception (for images near the recognition threshold) than the complete MS-SSIM metric. They proposed the use of the structural component r(x,y) with light modifications, avoiding the use of C3 in (3), and giving alternate definitions of r(x,y) to avoid division by zero.

Several studies in the medical imaging field have shown good results of this metric in certain tasks near the limit of visibility [21, 22, 23].

Improving Badly Blurred Images: G-SSIM

Chen et al. [13] developed a metric named Gradient Structural SIMilarity (G-SSIM) based on SSIM. They proposed that the HVS should be very sensitive to the edge and contour information, and these parts should be the most important structural information of an image. They substituted the images to be compared by their gradient maps, obtained applying Sobel operators across the original images. The luminance component is calculated with the original images, but the contrast and structural components are calculated with the gradient maps of these images. After that, they apply the usual SSIM rules to calculate the G-SSIM value. Their results showed an improvement of SSIM and MS-SSIM.

Four Components: 4-SSIM, 4-MS-SSIM, 4-G-SSIM, 4-MS-G-SSIM.

Li and Bovik [15] faced the lack of effectiveness of SSIM and MS-SSIM considering a four component model that classified local image regions according to edge and smoothness properties. In their studies, SSIM values are weighted by region type. According to this approach, they developed modified versions of SSIM, MS-SSIM, GSSIM, and MS-G-SSIM: 4-SSIM, 4-MS-SSIM, 4-G-SSIM, and 4-MS-G-SSIM and compared the performance of the whole set. Their experiments, applying these metrics in the LIVE Image Quality Assessment Database [24], showed that 4-SSIM, 4-MS-SSIM, 4-G-SSIM, and 4-MS-G-SSIM were more consistent with human observers than any other metrics.

Based on these proposals, we have applied a complete set of IQM (Table 1) to test the combination of these four approaches: 4-component approach (4), gradient approach (G), multi-scale approach (MS), and basic SSIM index (S) or structural approach (r*). Note that the first eight IQM in Table 1 were tested by Li and Bovik [15] with the LIVE database. This database includes pictures of faces, people, animals, nature scenes, man-made objects, etc., but no medical image. We developed the last new seven IQM to test the performance of the structural component r*.

Table 1. Set of IQM to be tested

	Metrics based on the three components of SSIM: luminance, contrast and structure.
SSIM	The original Structural SIMilarity index.
G-SSIM	Calculates SSIM over the gradient version of the image.
MS-SSIM	Multi-scale version of SSIM.
MS-G-SSIM	Multi-scale version of G-SSIM.
	4-component versions (weighting region type) of the four previous metrics
4-SSIM	Weights the values of the SSIM map according to the change (or preservation) of the original image's texture.
4-G-SSIM	Equal to 4-SSIM, but the original images are replaced by their gradient version.
4-MS-SSIM	Multi-scale version of 4-SSIM. 4-SSIM is calculated for every scale and then pooled according to the MS-SSIM rules.
4-MS-G-SSIM	Multi-scale version of 4-G-SSIM.
	Metrics based on the structural component of SSIM: r*
r*	The structural component of SSIM index, as was proposed by Rouse and Hemami.
G-r*	Calculates r* over the gradient version of the image.
MS-r*	Multi-scale version of r*. Is equivalent to the R* index proposed by Rouse and Hemami.
MS-G-r*	Multi-scale version of G-r*.
	4-component versions (weighting region type) of the four previous metrics
4-r*	Weights the values of the r* map according to the change (or preservation) of the original image's texture.
4-G-r*	Equal to 4-r*, but the original images are replaced by their gradient version.
4-MS-r*	Multi-scale version of 4-r*. The 4-structural component is calculated for every scale and then pooled according to the MS-r* rules.
4-MS-G- r*	Multi-scale version of 4-G- r*.

Experiments

The observers

Four medical doctors were selected, all of them specialized in Radiology. They were 57, 35, 32, and 53 years old and they had an experience in diagnostic radiology in hospitals of 31, 9, 6, and 27 years, respectively. We denote them as observers A, B, C, and D, respectively.

The database

The images have been collected looking for usual and representative examples in the day-to-day medical practice of a radiologist. The specimens of the database were collected by observer D and checked for their suitability to the referred day-to-day medical practice. Three subsets of eight images (each one) were selected:

1) Bone plain films (BPF). Usual bone radiographies: back, knee, foot, hand, wrist, etc.
2) Magnetic resonance (MR). Head, back, neck, etc. A representative slice for each selected case.
3) Chest plain films (CPF).

The colour depth was 8-bit (256 gray-levels) for each image. The size of each kind of image was different, depending on the acquisition technique. The usual size for each type of image was (in pixels) 1400x1700 for BPF, 512x512 for MR, and 2500x2000 for CPF. All patient identifiers were removed from the images. Fig 1 shows one specimen of each subset.

Fig 1. Examples of the different subsets.

Image distortion types

The images were distorted with some kind of distortions that are usual in a radiological environment [25] or are of interest for some medical applications: Gaussian blur, white noise, JPEG compression, and JPEG2000 compression [16, 18, 26, 27].

a) Gaussian blur (GB). The images were distorted with a circular symmetric Gaussian kernel with standard deviation ranging from 1 to 5 pixels, using the ImageJ (v. 1.44) function "Gaussian blur".
b) White noise. Gaussian noise (GN) of standard deviation between 20 and 100, using the ImageJ (v. 1.44) function "Add Gaussian noise".
c) JPEG compression (JPG). Compressed at bit rates ranging from 0.12 bpp to 0.15 bpp using the Matlab (v. 8.0) function imwrite.
d) JPEG2000 compression (J2000). Compressed at bit rates ranging from 0.01 bpp to 0.04 bpp, using the Matlab (v. 8.0) function imwrite.

The amount of distortion is intended to reflect a broad range of visual appearances, from light differences to strong distortions. This broad range of distortions was designed to manage the observer and IQM response from the near to

the supra-threshold problem. The number of steps for each distortion type was fixed at five. The total number of distorted images was 24 original images x 4 types of distortions x 5 levels of distortion = 480 images.

Fig 2 shows an example of the different distortion levels (referred to Gaussian noise) applied to a MR specimen. The first image (top-left) shows the image without any distortion.

Fig 2. Gaussian noise applied to an image belonging to the MR subset.

Test methodology

24 sets of images were arranged, one set for each original image. Each set comprised all of the distorted images obtained from each original image and the original itself. The name of the images was randomized.

Each set of images was independently evaluated by observers A, B, and C. Observer D was excluded of this evaluation in order to avoid bias, because he was the observer that selected the images. The images were displayed in the usual medical environment of these radiologists, trying to mimic their day-to-day medical experience.

We used a double-stimulus methodology. Each radiologist had a double-window space on their displays. The left one showed the reference image, without any kind of distortion. The right window showed the distorted images in a random sequence. The experts reported their answer based on the following instruction:

"Rate the quality, **for your medical practice,** of the distorted image on the following scale: bad (1), poor (2), intermediate (3), good (4) or excellent (5), always taking into account that the optimum level (5) is the level of the reference image **or that of any image medically indistinguishable from it**".

It is of importance to point out that the intention of the experiment was not to find out subtle differences between images by the observers, or visual similarities or dissimilarities between images (visual losses). The main intention was to determine the helpfulness of the image for a medical practice, for a diagnosis: we were measuring the diagnostic losses [28, 29], and this was the aim reflected in the instructions given to the experts. This intention was made clear to the observers involved in our experiment.

No time limit was fixed. The usual reviewing time for each image was in a range between 3 to 8 seconds. Usually, the poorer the image quality, the less time consumed to answer. Each evaluation session was not longer than 30 minutes and the lapse between sessions was, at least, 15 minutes, trying to avoid any kind of visual fatigue.

There were two reviews of the images, the second one six months later than the first one, in order to test the intra-observer variability.

Measures

The complete set of images was analyzed with algorithms developed by ourselves as a plugin in ImageJ, v. 1.44 [30], in order to get the value of the 16 proposed metrics for each image. These values were distributed in an interval between 0 and 1. These values were compared with those obtained by the human observers in average, Mean Opinion Scores (MOS), also scaled between 0 and 1.

Statistical analysis

Selection of observers

Once the second review of the images was completed, an analysis of the intra-observer consistency was performed using the weighted kappa coefficient [31] by using Cicchetti-Allison weights [32]. To apply it for each observer, the scores "1", "2", "3", "4", and "5" from the whole sample of images were pooled and classified to produce a 5x5 table, with entries of the table being the number of concordant or discordant pairs according to the first and the second readings. Consequently, the total number of pairs in each observer´s table was 480. This analysis enabled us to select the consistent or trustworthy observers. The interpretation of the obtained coefficients bore in mind the statistical significance, the number of scores, 5, and their prevalence [33]. The analysis included the study of the intra-observer consistency for each one of the three types of images separately, and the homogeneity of kappa statistics [31] through different types of images, for each observer separately.

The readings of the second review were used to evaluate the inter-observer agreement or variation. The generalized kappa statistic [31] and the Friedman two-way analysis of variance were applied. A Friedman test was used since we were employing a randomized complete block design where each image behaves as a block [34]. A total of 480 images were used for these analyses which included the kappa coefficient of every score separately, together with its corresponding jackknife confidence interval [35].

Performance Measures

For the analysis of the relation between the image scores, provided by the metrics, and the corresponding MOS provided by the observers, Pearson, r, and Spearman, rs, correlation coefficients were used. Spearman's analysis remained complementary since the definition of both scores and the high sample size guarantee the adequacy of the Pearson statistics. A third statistic was the root-mean-squared-error (RMSE) between metric scores and MOS. To deepen the assessment of the performance of the IQM, the relationship between both scores was analyzed by means of linear regression analyses, considering IQM scores as independent or predictor variable and the MOS as dependent variable. The slope b and the intercept a of the line gave additional measures to the association degree between IQM and MOS. A slope close to 1 and an intercept close to 0, together with large values of r and rs, and a small value of RMSE, will show a fairly good metric-observers agreement. In this context, RMSE measures the variability of the data with respect to the bisector of the first quadrant. The mean and the standard deviation (SD) of the MOS and the IQM scores, for each group of images, are included as descriptive measures to show over or underrating of the metrics versus observers.

This statistical analysis was achieved by means of SPSS 22 and Epidat 4.1 statistical packages.

Results

Selection of observers

The results from the analysis of the intra-observer agreement for all images are shown in Table 2. As expected, the kappa coefficients are strongly significant, $p<0.0001$. The confidence intervals and especially their lower limits, greater than 0.54, lead to the conclusion that all observers agree and, therefore, they have been kept for further analysis.

By type of image, the highest values of the kappa coefficient were reached for MR and BPF, but there were no statistically significant differences between the kappa coefficients corresponding to the four types of images, for each observer separately (all $p>0.10$).

Table 2. Weighted kappa coefficient (Cicchetti weights), standard error (SE), 95% confidence interval, z-statistic and p-value from the double reading by every single observer.

	W. Kappa	SE	95% CI	z-statistic	p-value
Obs. A	0.658	0.027	(0.605; 0.711)	18.40	<0.0001
Obs. B	0.662	0.027	(0.609; 0.715)	18.11	<0.0001
Obs. C	0.603	0.028	(0.548; 0.658)	16.43	<0.0001

Application of the Friedman test did not find significant differences between observers (test statistic = 1.90, p=0.39). Therefore, one could conclude the agreement between the three observers. The scores from the three radiologists selected, A, B, and C, were used to evaluate the quality of the images; and the average of these scores made up the MOS.

The results from the application of the generalized kappa statistic to these observers are shown in Table 3. The most frequent categories for all observers were scores 2 and 3. These differences in marginal totals produce a few substantial changes in the prevalence of the extreme categories against the central ones by which the global kappa is reduced [36]. Based on this, we could conclude a moderate-good concordance between the observers, kappa=0.595, 95% CI: (0.555; 0.635).

Table 3. Generalized kappa statistic, 95% jackknife confidence interval, z-statistic and p-value.

Category	Kappa	95% CI	z-statistic	p-value
Score 1	0.727	(0.662; 0.792)	27.59	<0.0001
Score 2	0.566	(0.505; 0.627)	21.49	<0.0001
Score 3	0.532	(0.469; 0.595)	20.18	<0.0001
Score 4	0.575	(0.505; 0.645)	21.82	<0.0001
Score 5	0.625	(0.514; 0.736)	23.71	<0.0001
Global kappa	0.595	(0.555; 0.635)	41.87	<0.0001

IQM performance

In the first step, taking into account the whole set of images, we have selected the most accurate metrics based on the values of the r, rs, and RMSE statistics, and also we have compared the mean values of IQM versus MOS. This last condition is intended to not discard those metrics, with relatively high correlation coefficients, but showing overrates (or underrates) leading to an increase of the RMSE. Table 4 shows the measures of performance for the 16 metrics.

A comparison between the IQM mean values and MOS (0.41) shows that almost all metrics overrate versus the observers. Only r* and 4-G-r* provide similar mean values somewhat lower (0.37) and G-r* underrates clearly (0.21).

In order to select the more accurate IQM, the statistics combination (r, rs >0.65, RMSE<0.25) was chosen. The justification of these thresholds, apparently not very demanding, is due to the heterogeneity of images: type of image, type of distortion, and size in pixels. Taking into account these thresholds, only 4-G-SSIM and 4-G-r* met these requirements.

Table 4. r, r_s, mean, SD, b, a, and RMSE. 480 images. IQM vs MOS.

	r	r_s	mean	SD	b	A	RMSE
SSIM	0.35	0.44	0.73	0.31	0.31	0.18	0.46
G-SSIM	0.42	0.43	0.51	0.27	0.44	0.19	0.31
MS-SSIM	0.46	0.60	0.88	0.17	0.74	-0.23	0.53
MS-G-SSIM	0.59	0.67	0.74	0.20	0.82	-0.19	0.39
4-SSIM	0.54	0.60	0.68	0.26	0.58	0.02	0.37
4-G-SSIM	**0.67**	**0.66**	0.45	0.24	0.78	0.06	**0.22**
4-MS-SSIM	0.55	0.74	0.88	0.15	1.02	-0.48	0.52
4-MS-G-SSIM	**0.75**	**0.81**	0.72	0.19	1.10	-0.38	**0.36**
r*	0.58	0.56	0.37	0.22	0.75	0.13	0.24
G-r*	0.60	0.57	0.21	0.20	0.82	0.24	0.30
MS-r*	0.59	0.58	0.65	0.20	0.82	-0.12	0.33
MS-G-r*	0.64	0.63	0.51	0.23	0.79	0.01	0.24
4-r*	0.64	0.66	0.59	0.21	0.83	-0.07	0.28
4-G-r*	**0.69**	**0.66**	0.37	0.23	0.82	0.11	**0.21**
4-MS-r*	0.60	0.70	0.80	0.16	1.03	-0.41	0.44
4-MS-G-r*	**0.71**	**0.76**	0.66	0.21	0.94	-0.20	**0.31**

Observers: mean 0.41, SD = 0.28

However, the metrics 4-MS-G-SSIM and 4-MS-G-r*, which apply the multi-scale component to the previously selected metrics, were also selected for a deeper analysis. Despite the gap in the mean value (0.72 and 0.66 versus 0.41), which explains the high RMSE values (0.36 and 0.31) and the low values of a (-0.38 and -0.20), it should be noted that this gap is uniform throughout all the scores (b = 1.10 and 0.94).

One way to improve these two metrics would be to correct their values through a change of origin. In particular, we used the mean difference IQM-MOS as the value for change, by subtracting that value from all the scores of each metric. This operation will produce an increase of the intercept, equaling the mean values, without changing the rest of the performance statistics (which are invariant to changes of origin, see Table 5). Specifically, this correction produces a decrease of the RMSE value for 4-MS-G-SSIM and 4-MS-G-r*, and now both IQM can meet the requirements we have fixed for a further analysis. 4-G-SSIM and 4-G-r* metrics do not require a similar correction due to the proximity of the means of both to the observer (0.45 and 0.37 versus 0.41), these differences being in terms of Cohen's d effect

sizes [37] of 0.16 and 0.19 respectively, indicating a "small" effect size (d <0.2). We used this index to be independent of sample size.

Table 5. r, r_S, mean, SD, b, a, and RMSE for four IQM. 480 images.

	r	r_s	mean	SD	b	A	RMSE
4-G-SSIM	0.67	0.66	0.45	0.24	0.78	0.06	0.22
4-MS-G-SSIM-0.31	0.75	0.81	0.41	0.19	1.10	-0.04	0.18
4-G-r*	0.69	0.66	0.37	0.23	0.82	0.11	0.21
4-MS-G-r*-0.25	0.71	0.76	0.41	0.21	0.94	0.03	0.20

Observers: mean = 0.41, SD = 0.28

The most effective IQM is 4-MS-G-S (after correction). It outperforms the other metrics in r, rs and RMSE. It shows an excellent value of slope b and intercept a. The second most effective is 4-MS-G-r*. 4-G-S and 4-G-r* show a similar result and their performance is slightly lower than the performance of the other two metrics. We analyzed in depth this subset of metrics.

Analysis by type of image

Table 6 shows the results, by type of image, of the four selected metrics. The scores of 4-MS-G-SSIM and 4-MS-G-r* have been modified by subtracting from their scores the mean difference IQM-MOS for each type of image. The values of these subtrahends are listed within the table. Note that this correction can always be applied to the metrics in a day-to-day radiological practice, because the type of image is well known before the acquisition of the image. In that sense, it is a numeric constant included in the algorithm itself. As has been shown earlier in this document, 4-G-SSIM and 4-G-r* metrics do not require a similar correction due to the proximity of the means of both to the MOS.

Table 6. r, r_S, mean, SD, b, a, and RMSE, for four metrics. Results by type of image.

Metric	mean - MOS	Type	r	r_s	mean	SD	b	a	RMSE
4-G-S									
		BPF	0.63	0.60	0.44	0.25	0.71	0.09	0.23
		MR	0.83	0.84	0.51	0.25	1.01	-0.08	0.18
		CPF	0.51	0.49	0.40	0.21	0.63	0.15	0.23
4-MS-G-S									
	0.31	BPF	0.75	0.82	0.40	0.20	1.03	-0.01	0.19
	0.36	MR	0.86	0.90	0.43	0.15	1.66	-0.29	0.18
	0.26	CPF	0.73	0.78	0.40	0.18	1.02	-0.01	0.17
4-G-r*									

		BPF	0.64	0.59	0.36	0.22	0.79	0.12	0.22
		MR	0.86	0.88	0.45	0.24	1.06	-0.04	0.15
		CPF	0.56	0.52	0.30	0.21	0.69	0.19	0.24
4-MS-G-r*									
	0.24	BPF	0.69	0.74	0.40	0.24	0.80	0.09	0.21
	0.32	MR	0.85	0.87	0.43	0.15	1.70	-0.31	0.19
	0.18	CPF	0.76	0.82	0.40	0.19	1.01	0.00	0.17

MOS: BPF (0.40), MR (0.43), CPF (0.40).

As expected, 4-MS-G-SSIM and 4-MS-G-r* provide, for all types of image, much better results (attending to RMSE) than those of the non-modified version of them.

MR images provide the best results for all metrics in terms of combination r, rs, RMSE. Although CPF and BPF images provide the worst agreement (due to the worst performance of 4-G-SSIM and 4-G-r*), 4-MS-G-SSIM and 4-MS-G-r* show good agreement for every kind of image.

The MS component dramatically improves the Pearson coefficient for CPF and BPF and keeps the results for MR. Fig 3 highlights the evolution of the Pearson coefficient by type of image and compares the single and multi-scale version of the selected metrics.

Li and Bovik found [15] that 4-G-SSIM performed better than 4-MS-G-SSIM, suggesting that multi-scale was not of great importance for the performance of an IQM. That result, remarkably, can be consistent with the one obtained by us. Li and Bovik tested their metrics against the LIVE Image Quality Assessment Database [24]. This database consists of a set of images, with sizes in pixels from 480 to 768 in width and from 480 to 512 in height. The dimension in pixels of our set of images is 1400x1700 for BPF, 512x512 for MR, and 2500x2000 for CPF.

As can be seen in Fig 3, the larger the images (CPF and BPF), the better the improvement achieved with multi-scale (MS option). According to the theory shown by Wang et al. [10] the multi-scale factor improves the results for larger images (BPF, CPF), due to the fact that the MS component adds the different viewing distances as a factor of the human reading. This approach divides iteratively the image by a factor of two up to five times. For small images, such as those belonging to the LIVE Database or those included in our experiment acquired by MR, the size of the image after five downsizings by a factor of two is of the order of 16 pixels. Downsizing images of about 2.400 pixels (those belonging to the CPF set) gives a final size of about 75 pixels. This size carries much more information for the HVS than images of 16 pixels.

Fig 3. Influence in the Pearson coefficient of the multi-scale component for the selected metrics.

Type of distortion and type of image

In our experiment four types of distortion have been applied to each type of image. In the analysis of real medical practice images, this feature cannot be fully controlled before the acquisition of the image. It seems reasonable, however, to analyze the performance of the selected metrics for each type of distortion within each image type. In this way we can check the robustness or not of the metrics, based on the distortion of the image for each group of images (BPF, MR, and CPF). To simplify, Table 7 presents these results for r values.

Table 7. r value for the four selected metrics. Results by type of image and distortion (Dist.). Number of images by type of image and distortion=40.

Type	Dist.	4-G-S	4-MS-G-S	4-G-r*	4-MS-G-r*
BPF					
	GB	**0.38**	0.75	**0.24**	0.67
	GN	0.81	0.89	0.75	0.81
	J2000	0.81	0.88	0.65	0.80
	JPG	0.86	0.87	0.75	0.83
MR					
	GB	0.91	0.91	0.89	0.88
	GN	0.90	0.86	0.88	0.83
	J2000	0.89	0.91	0.90	0.86
	JPG	0.89	0.89	0.91	0.85
CPF					
	GB	**0.25**	0.85	**0.19**	**0.55**
	GN	0.87	0.95	0.88	0.95
	J2000	0.89	0.90	0.88	0.90
	JPG	0.85	0.88	0.81	0.89

Very good agreement, $\mathbf{r} \geq 0.85$. Good agreement, $0.75 \leq \mathbf{r} < 0.85$. Fairly good agreement, $0.65 \leq \mathbf{r} < 0.75$. Poor agreement (bold), $\mathbf{r} < 0.65$

BPF. Fig 3 showed the worst performance of 4-G-SSIM and 4-G-r* compared against their multi-scale versions, 4-MS-G-SSIM and 4-MS-G-r*. Table 7 shows that this behavior is mainly due to the low performance of the single-scale IQM when GB distortion is present. Excluding GB distortion, BPF images show similar results for the other three types of distortion, but non-homogeneous for the four considered metrics: 4-G-SSIM and 4-MS-G-r* provide similar performance ($0.80 \leq r \leq 0.86$), better than 4-G-r* ($0.65 \leq r \leq 0.75$). 4-MS-G-SSIM shows the best results ($r \geq 0.87$).

MR. The good performance of the four IQM metrics with MR images, shown in the raw analysis of these images (Table 6), remains for the four types of distortion ($0.83 \leq r \leq 0.91$). Thus, we can conclude that the performance of the four metrics does not depend on the type of distortion for these images, and is optimal and uniform among them.

CPF. Fig 3 showed the worst performance of 4-G-SSIM and 4-G-r* compared against their multi-scale versions, 4-MS-G-SSIM and 4-MS-G-r*. Table 7 shows that this behavior is mainly due to the low performance of the single-scale IQM when GB distortion is present. Excluding this distortion, the four metrics show optimal performance ($0.81 \leq r \leq 0.95$) with CPF images, similar to that obtained with MR images. 4-MS-G-SSIM again provides optimal results in all kinds of distortion ($0.85 \leq r \leq 0.95$)

The influence of GB

Li and Bovik showed [15] that the multi-scale approach has no advantage when a GB distortion is applied to a set of images. On the contrary, we have found in our experiment that the multi-scale approach largely improves the performance of 4-G-SSIM and 4-G-r* when a GB distortion is applied. This apparent disparity can be due to the different resolution of some images of our set and the different levels of distortions. First, the multi-scale approach improves the quality of IQM for the largest images, CPF and BPF sets, showing a good agreement with the multi-scale theory [10]. Second, the distortion degree of our images is much slighter than the corresponding one in Li and Bovik's study.

The influence of the different components (G, 4, MS, and r*) over the complete set of metrics

Some IQM components overestimate their mean value and others underestimate it. In order to make a comparison between them, this behavior penalizes the RMSE value. To show a uniform set of data, it is of interest to rebuild Table 4 correcting linearly the mean value of every IQM by the difference between this mean value and the MOS for the complete dataset of images. This correction minimizes the RMSE values for all the metrics. These results for the quality statistics r, rs, b, and RMSE are shown in Table 8.

Table 8. R, r_s, and RMSE for the 16 IQM. Mean value correction applied. IQM vs MOS

IQM	r	r_s	b	RMSE
SSIM	0.35	0.44	0.31	0.29

G-SSIM	0.42	0.43	0.44	0.25
MS-SSIM	0.46	0.60	0.74	0.22
MS-G-SSIM	0.59	0.67	0.82	0.20
4-SSIM	0.54	0.60	0.58	0.22
4-G-SSIM	0.67	0.66	0.78	0.18
4-MS-SSIM	0.55	0.74	1.02	0.20
4-MS-G-SSIM	0.75	0.81	1.10	0.18
r*	0.58	0.56	0.75	0.20
G-r*	0.60	0.57	0.82	0.20
MS-r*	0.59	0.58	0.82	0.20
MS-G-r*	0.64	0.63	0.79	0.19
4-r*	0.64	0.66	0.83	0.19
4-G-r*	0.69	0.66	0.82	0.18
4-MS-r*	0.60	0.70	1.03	0.19
4-MS-G-r*	0.71	0.76	0.94	0.20

To compare the different components of the IQM, we have grouped together pairs of metrics that change from one to the other only in one component. So, Fig 4 compares the effect of the 4-component in r, rs, b, and RMSE, grouping SSIM and 4-SSIM, MS-SSIM and 4-MS-SSIM, and so on. Fig 5 shows the effect of the MS component, Fig 6 the G component effect, and, finally, Fig 7 shows the variation between SSIM and r*. The influence of every component is shown in percentage of variation for r, rs, and RMSE. The percentage of variation of RMSE has been multiplied by -1, in order to show positive values when RMSE decreases with the related component, and negative values when it increases. The influence on slope, b, is shown as a percentage of variation with respect to the value "1".

Fig 4. Effect of component 4. Relative percentage increase in the quality statistics of the IQM metrics

Fig 5. Effect of component MS. Relative percentage increase in the quality statistics of the IQM metrics.

Fig 6. Effect of component G. Relative percentage increase in the quality statistics of the IQM metrics.

Fig 7. Effect of component r* vs SSIM. Relative percentage increase in the quality statistics of the IQM metrics.

4. Regarding Fig 4 and Table 8, one fact stands out: this component always improves the values of r, rs, b, and RMSE, with no exception.

MS. This component always improves the values of r, rs, b, and RMSE with some minor exceptions: 4-r* has a Pearson correlation coefficient slightly higher than 4-MS-r* (0.64 vs. 0.60). The value of b worsens with the MS component for the metric G-r* (0.82 vs 0.79).

G. This component always improves the values of r, rs, b, and RMSE with some minor exceptions: the value of b worsens for the metrics 4-MS-r*, MS-r*, and 4-MS-SSIM by 3%, 3%, and 8% respectively. The overall improvement of this component is notoriously lower than that from 4 or MS or r* components.

r*. This component improves the values of r, rs, b, and RMSE, but is more erratic than the other three components. 4-MS-G-SSIM decreases their overall performance. MS-G-SSIM does not change, on average, its performance. MS-SSIM and 4-MS-SSIM show an overall improvement, but rs decreases by 10% and 7% respectively. The other IQM improve clearly their performance.

Discussion

4-MS-G-SSIM provides optimal results in all kinds of distortion and images. The second most effective IQM is 4-MS-G-r*. 4-G-SSIM and 4-G-r* show an identical result and their performance is a little bit lower than the performance of the other two metrics.

For MR images, the four metrics show similar behavior. For CPF and BPF images (the largest images in the set), 4-MS-G-SSIM shows a better performance than the other three IQM, especially than those metrics that use a single-scale approach (4-G-SSIM and 4-G-r*). Specifically, the worst results are those that include GB distortion on images of BPF and CPF in 4-G-SSIM and 4-G-r* metrics ($r<0.39$ for all of them).

Those metrics that apply the 4 and the G component show the best performance among the complete IQM set. Those results are consistent with previous papers [13, 15] and show a strong correlation of the HVS with gradients (G component), and edge and smoothness properties (4 component) in the images.

Previous studies [15] have shown the irrelevance of using the multi-scale (MS) approach in large databases. On the contrary, we have found the superiority of this approach over the single-scale approach. This fact, previously explained, can be due to the large size of some images (CPF, BPF) included in our database.

The use of the structural component of SSIM (r*) instead of the use of the complete Structural Similarity Index (SSIM) shows a slight advantage. This result shows a good agreement with previous studies near the recognition threshold [11, 21, 22]. Despite this fact, the effect of the component r* is less than the other three components (4, G, MS). The best metric (4-MS-G-SSIM) applies the SSIM component instead of the r* component, showing lighter, but better, results than its counter partner 4-MS-G-r*. It should be taken into account that the present set of images is far from the supra threshold problem that can be found in other databases like the LIVE Database. However, neither does the present database meet the criteria of the near threshold problem proposed by Rousse and Hemami [11] and applied in the quoted works [21, 22] which revealed a superior performance of r* vs. SSIM. Our database shows few differences between images with different distortion levels, but these distortion levels can be easily recognized, unlike the recognition threshold levels. Further analyses could show the behavior of the structural component with the distortion levels, but this is not the aim of the present work, focused on stronger distortions.

Conclusions

We can conclude that components 4, G, and MS show a strong agreement with the HVS, and 4-MS-G-SSIM can be used as a good surrogate of a human observer to analyze the medical quality of a general radiological image in an environment with a reference image and with simple types of noise. 4-MS-G-r*, 4-G-SSIM and 4-G-r* also show results that are consistent with human subjectivity in a wide set of medical images.

We are aware that some model observers could be more accurate in reproducing human perception for certain tasks, for certain types of noise or for certain acquisition techniques, all of them more specific for some set of radiological images. Our aim in this study has been to find a general index that can be a good surrogate of the human observer in a wide range of medical imaging situations.

Last, but not least, we want to share our efforts with our scientific colleagues. The whole set of programs and algorithms we have applied in this study, will be be freely available in our website (https://www.ucm.es/gabriel_prieto) for the scientific community.

Acknowledgments

The authors thank Prof. A. C. Bovik and C. Li for their support clarifying to us some aspects of the four-component (4) model applied to SSIM.

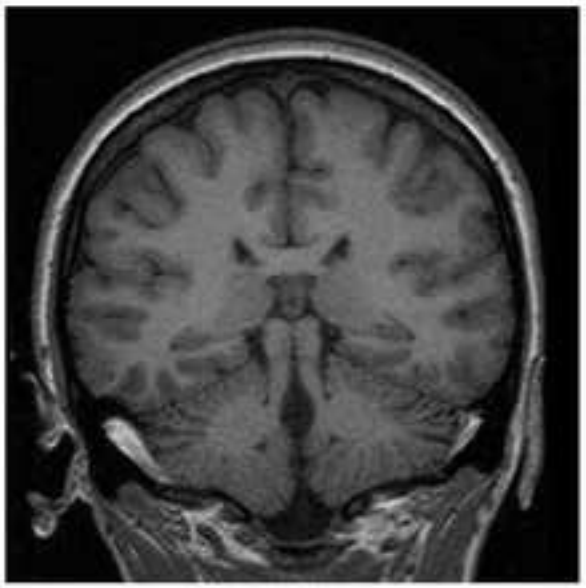
References

1. Girod B, What's wrong with mean-squared error. In: Watson AB, editor. Digital Images and Human Vision. MIT press; 1993. p. 207-220.
2. Burgess AE. The Rose model, revisited. J. Opt. Soc. Am. 1999 A (16), p. 633– 646.
3. Myers KJ. Ideal observer models of visual signal detection. In : Beutel J, Kundel H, Van Metter R, editors. Handbook of Medical Imaging, Physics and Psycophysics. SPIE, Bellingham, WA ; 2000. P. 558-592.
4. Barrett HH, Myers KJ, and Wagner RF. Beyond signal detection theory. Application of Optical Instrumentation in Medicine XIV and Picture Archiving and Communications (PACS IV) for Medical Applications, Newport Beach, CA, 1986. Proceedings of the Society of Photo-optical Instrumentation Engineers, Bellingham, WA, 626, 231–239;1986.
5. Fiete RD, Barrett HH, Smith WE, and Myers KJ. The Hotelling trace criterion and its correlation with human observer performance. J. Opt. Soc. Am. 1987. A (4) p. 945–953.
6. Wagner RF, Brown DG, and Pastel MS. Application of information theory to the assessment of computed tomography. Med. Phys. 1979. (6). p. 83–94.
7. Eckstein MP, Abbey CK, Bochud FO. In: Beutel J, Kundel H, Van Metter R, editors. A practical guide to model observers for visual detection in synthetic and natural noisy images. Handbook of Medical Imaging, Physics and Psychophysics. SPIE, Bellingham, WA, 2000. (1) Physics and Psychophysics. p. 593-626.
8. ICPFU Report 54. Medical Imaging – The Assessment of Image Quality. Bethesda. MD: International Commission on Radiation Units and Measurements (1996).
9. Wang Z, Bovik AC; Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. Image Processing, IEEE Transactions on. 2004. (13). no.4. p. 600-612.
10. Wang Z, Simoncelli E, and Bovik AC. Multi-scale structural similarity for image quality assessment. Proceedings of the 37th IEEE Asilomar Conference on Signals, Systems and Computers; 2003. p. 529–554.
11. Rouse DM and Hemami SS. Analyzing the Role of Visual Structure in the Recognition of Natural Image Content with Multi-Scale SSIM. Proceedings of SPIE. 2009. p. 6806.
12. Brooks AC, Zhao XN, Pappas TN. Structural similarity quality metrics in a coding context: exploring the space of realistic distortions. IEEE Transactions on Image Processing. 2008. (17). no. 8. p. 1–12.
13. Chen GH, Yang CL, Xie SL, Gradient-based structural similarity for image quality assessment. IEEE International Conference on Image Processing; 2006. p. 2929–2932.
14. Sampat MP, Wang Z, Gupta S, Bovik AC, Markey MK. Complex wavelet structural similarity: a new image similarity index. IEEE Transactions on Image Processing; 2009. (18), no. 11, p. 2385-2401.
15. Li C and Bovik AC. Content-partitioned structural similarity index for image quality assessment. Journal Image Communication; 2010. (25), no. 7, p. 517-526.
16. Johnson JP, Krupinski EA, Yan M, Graham AR, Weinstein RS. Using a visual discrimination model for the detection of compression artifacts in virtual pathology images. IEEE Transactions on Medical Imaging, 2010. (30), no. 2, p. 306-314.
17. Kim B, Lee H, Joong K, Seo J, Park S, Shin YG, and Lee KH. Comparison of Three Image Comparison Methods for Visual Assessment of Image Fidelity of Compressed Body CT Images. Med. Phys. 2011. (38), no. 2, p. 836-844.
18. European Society of Radiology (ESR). Usability of Irreversible Image Compression in Radiological Imaging. A Position Paper by the European Society of Radiology (ESR). Insights into Imaging; 2011. 2.2: p.103–115. PMC.
19. Kowalik-Urbaniak IA, Brunet D, Wang J, Vrscay ER, Wang Z, Koff D, Smolarski-Koff N, et al. The quest for 'diagnostically lossless' medical image compression: a comparative study of objective quality metrics for compressed medical images. SPIE Medical Imaging: Image Perception, Observer Performance, and Technology Assessment. 2014 Feb. San Diego, CA. 2014.

20. Kowalik-Urbaniak IA, Castelli J, Hemmati N, Koff D, Smolarski-Koff N, Vrscay ER, et al. Modelling of subjective radiological assessments with objective image quality measures of brain and body CT images. International Conference on Image Analysis and Recognition; 2015 June 22-24; Niagara Falls, Ontario; 2015.
21. Prieto G, Guibelalde E, Chevalier M, Turrero A. Use of the cross-correlation component of the multiscale structural similarity metric (R* metric) for the evaluation of medical images. Med Phys. 2011. (38), no. 8, p. 4512-7.
22. Von Falck C, Rodt T, Hartung D, Meyer B, Wacker F, Shin HO, A systematic approach towards the objective evaluation of low-contrast performance in MDCT: combination of a full-reference image fidelity metric and a software phantom. European Journal of Radiology; 2012. (81), no. 11, p. 3166-71.
23. Von Falck C, Bratanova V, Rodt T, Meyer B, Waldeck S et al. Influence of Sinogram AffiMRed Iterative Reconstruction of CT Data on Image Noise Characteristics and Low-Contrast Detectability: An Objective Approach. PLoS ONE 8. 2013. (2): e56875. Doi:10.1371/journal.pone.0056875.
24. Sheikh HR, Wang Z, Cormack L, Bovik AC, LIVE Image Quality Assessment Database [cited July 2014]. Available from: http://live.ece.utexas.edu/research/quality
25. Williams MB, Krupinski EA, Strauss KJ, Breeden WK, Rzeszotarski MS, Applegate K, et al. Digital Radiography Image Quality: Image Acquisition. J Am Coll Radiol; 2007. (4), p. 371-388.
26. Krupinski EA, Williams MB, Andriole K, Strauss KJ, Applegate K, Wyatt M, et al. Digital Radiography Image Quality: Image Processing and Display. J Am Coll Radiol; 2007. (4), p. 389-400.
27. Loose R, Braunschweig R, Kotter E, Mildenberger P, Simmler R, Wucherer M. Compression of digital images in radiology results of a consensus conference. Rofo; 2009. 181(1), p. 32-37.
28. Krupinski EA, Johnson JP, Roehrig H, Nafziger J, Fan J, Lubin J. Use of a human visual system model to predict observer performance with CRT vs LCD display of images. J Digit Imaging; 2004. 17(4). p. 258–263.
29. Royal College of Radiologists (RCR, UK). The adoption of lossy data compression for the purpose of clinical interpretation, 2008 [cited May 2015]. Available from: https://www.rcr.ac.uk/sites/default/files/publication/IT_guidance_LossyApr08_0.pdf.
30. Rasband WS, ImageJ, U. S. National Institutes of Health, Bethesda, Maryland, USA. 1997 – 2015. [cited Jul 2014]. Available from: http://rsb.info.nih.gov/ij/plugins/index.html
31. Fleiss JL. Statistical methods for rates and proportions. John Wiley &Sons, New York; 1981.
32. Cicchetti DV and Allison T. A New Procedure for Assessing Reliability of Scoring EEG Sleep Recordings. American Journal of EEG Technology. 1971. (11), p. 101–109.
33. Bakeman R, Quera V, McArthur D, Robinson BF. Detecting sequential patterns and determining their reliability with fallible observers. Psychological Methods; 1997. p. 357–370. doi:10.1037/1082-989X.2.4.357.
34. Woolson RF and Clarke WR. Statistical methods for the analysis of biomedical data, 2nd ed. John Wiley & Sons, New York; 2002.
35. Efron B and Tibshirani RJ. An introduction to the bootstrap. Chapman & Hall, New York; 1993.
36. Sim J and Wright C. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Physical Therapy; 2005. 85 (3), p. 257-268.
37. Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. Hillsdale, Lawrence Earlbaum Associates, New Jersey, 1988.

Click here to download Figure: Fig 1.tif

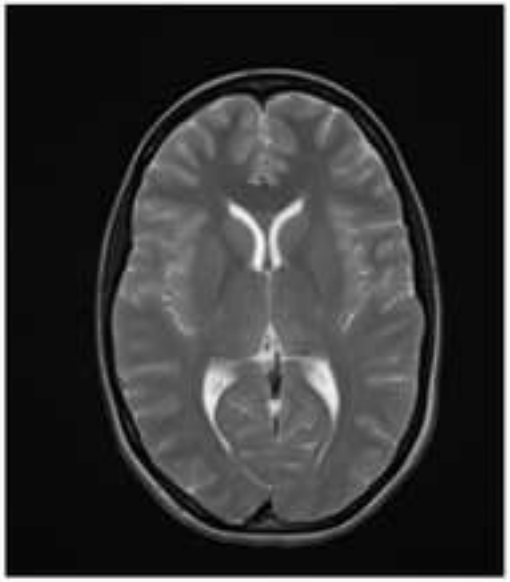
BPF



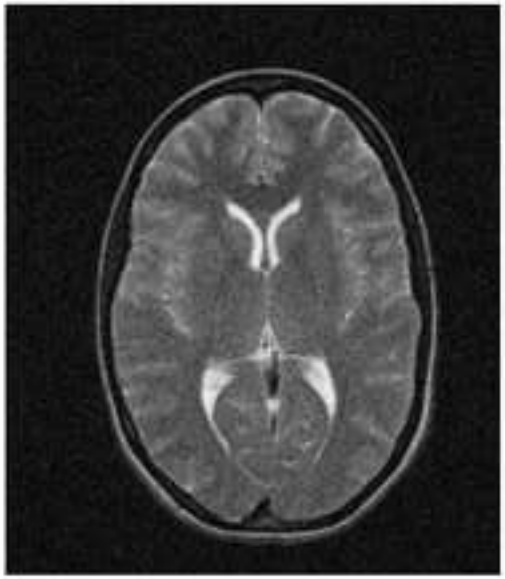
MR

CPF

Click here to download Figure: Fig 2.tif



No distortion.



sd=20

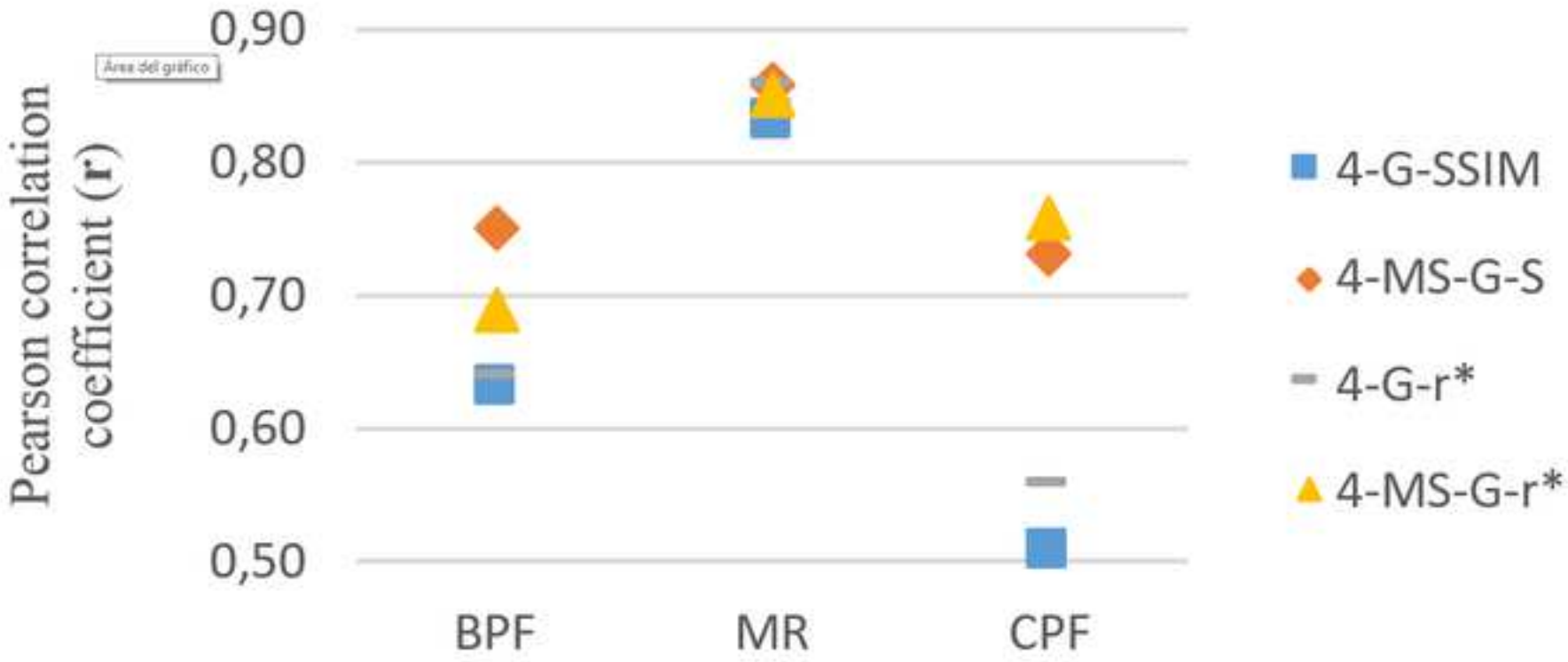
sd=40

sd=60

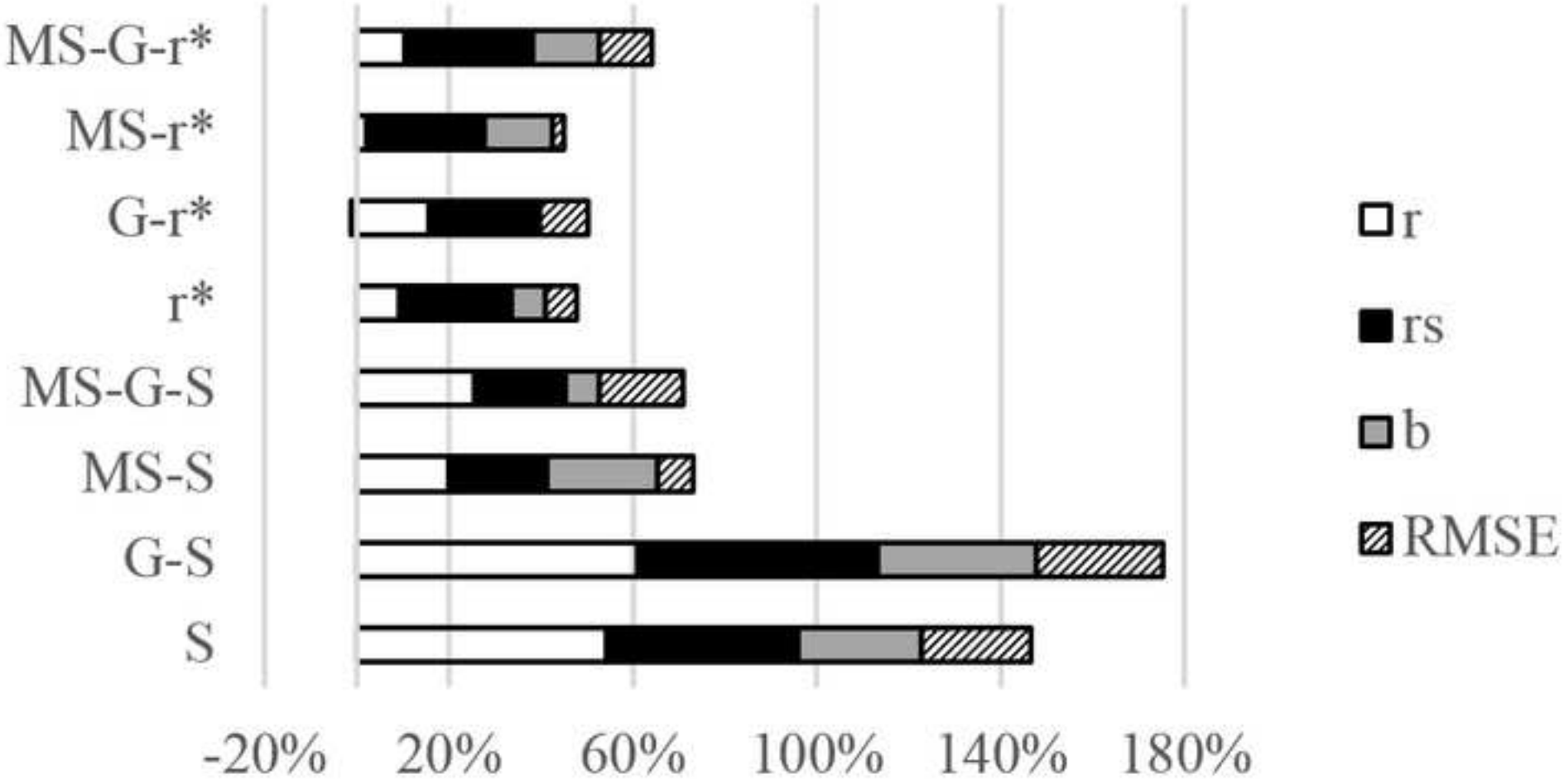
sd=80

sd=100

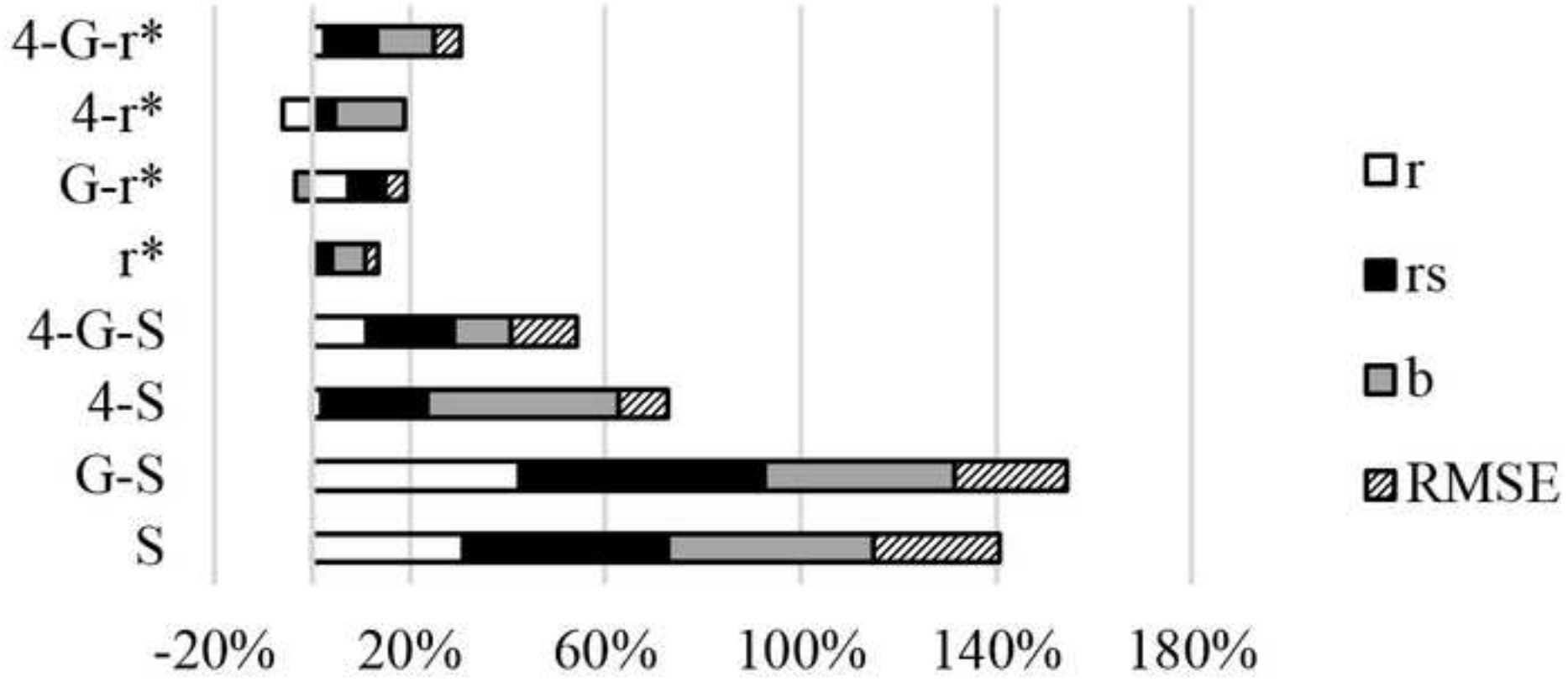
Click here to download Figure: Fig 3.tif



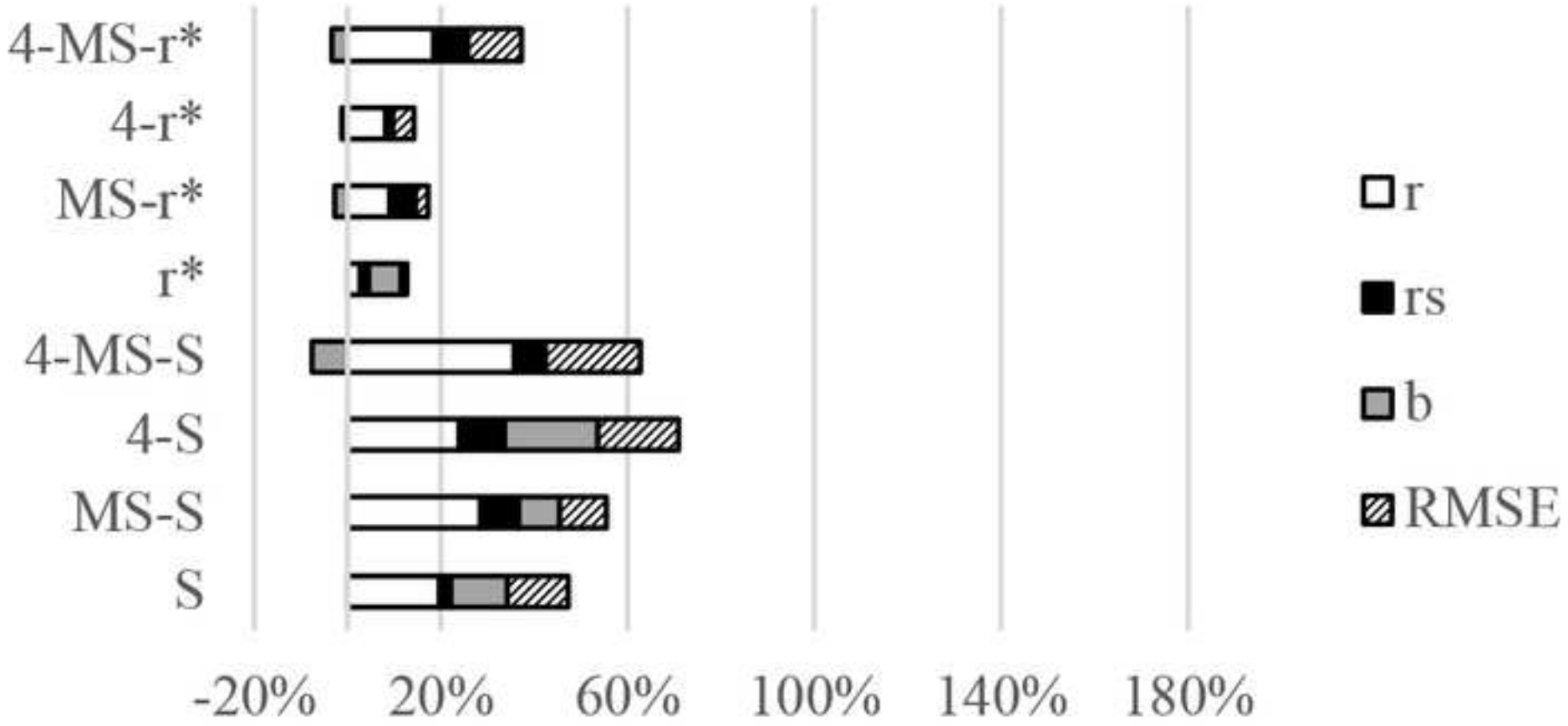
Click here to download Figure: Fig 4.tif



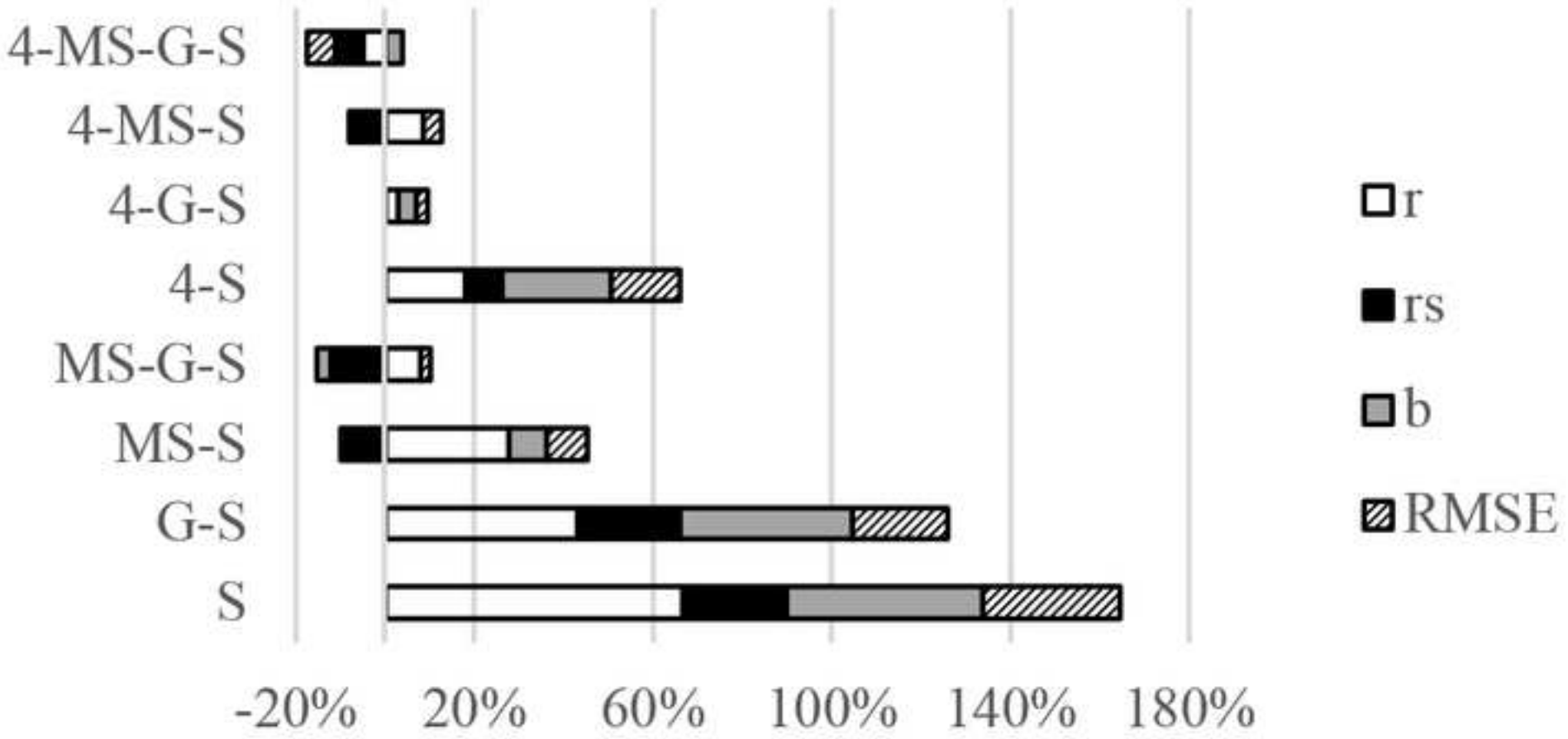
Click here to download Figure: Fig 5.tif



Click here to download Figure: Fig 6.tif



Click here to download Figure: Fig 7.tif



14. Summary

Analysis and proposals of image quality metrics that mimic the human observer

The research shown in this document is focused on the automated perception of the image quality in medical imaging and its correlation with the human perception of this image quality.

Image quality analysis plays a central role in the design of imaging systems for medical diagnosis. The final objective of this image quality analysis is usually to design a metric able to score the perceived quality of a medical image: an image quality metric (IQM). Moreover, the goal of a great number of researches is to develop an automated metric capable of mimic the results produced by a human observer. Usually these metrics are developed as computerized algorithms. So far, only partial success has been achieved.

The number of algorithms and approaches in the literacy is high and the problem is still an open question. Remarkably, in the scientific literacy there are two approaches clearly different; one is based on human visual function models or on ideal observer models, (considered together or not). They are models that try to reproduce the image process from the eye to the perception centers in the brain, even modelling (in some approaches) the neuronal response or the whole visual system. These models are usually very complex, with a limited validity and they have not shown generalized and satisfactory responses. These models are typical in the medical image field.

For years, many specialists in Telecommunications have analyzed the quality image problem from a broader point of view, more focused on the studies of natural images (those from the natural environment), in still image as well as in video image. Many of these studies have been based on "top-down" models of the human visual system. These models propose some hypotheses about the general function of the human visual system and build their human visual models according to these hypotheses. Some of these studies have proposed metrics that correlate very well with the human perception. Surprisingly, so far, there are a few studies and applications of these metrics in the medical imaging field.

Attending to this approach, the most successful metric developed lately has been the Structural SIMilarity Index (SSIM) proposed by Wang et al. in 2004. This metric is based on the Wang and Bovik's theory about the human visual system. This theory states that our visual system is highly adapted to extract structural information from an image (note the "top-down" approach). A broad family of metrics has been developed based on this metric, with increased correlations between the metrics and the human perception.

One of the most promising members of this family is the cross correlation multiscale coefficient of SSIM, the so called R*, developed by Rouse & Hemami in 2009. Its design was focused around the problem of perception near the limit of visibility. This task is of great importance in medical phantom analysis and, generally speaking, in the field of medical imaging.

There have been other researches trying to improve the results of SSIM. We have to highlight three approaches: those based on image gradient analyses, those based on image textures analysis and those based on analysis that simulate different viewing distances. These three approaches, together with R*, have shown promising features in noisy and blurred

environments. They have been tested with natural images but never together, that is, analyzing the behavior of the metrics mixing different combinations of the four approaches.

This thesis has analyzed the behavior of different metrics (all of them belonging to the SSIM family) and proposes new metrics that combine the related approaches. We have studied different medical images in different environments. The steps can be summarized as follows.

Tools

At the beginning of this research, there were more than 250 different tools capable of visualizing, editing, or extracting information from medical images. An exhaustive research was performed, looking for a DICOM image editor that allowed the modification of a medical image. This modification included (but not only) insertions of artifacts, anatomic backgrounds, pathologies, different types of noise, etc. This project was looking for the best software tool to produce an image database with original and modified images to carry out perception experiments. The selected tool was ImageJ, by Wayne Rasband. This tool was the backbone for all our algorithms (developed in Java as plugins of ImageJ), that we have applied to manipulate all the images tested throughout this research.

Phantom image analysis: the human memory problem

Many phantoms have been designed to study mammographic image quality such as ACR, TOR(MAM) or CDMAM phantom. The task with these phantoms is to obtain the minimum contrast (threshold) for each diameter of a series of discs with different contrasts. Usually, the discs are located in well-known positions and the evaluation is based in the SKE paradigm. The main advantage of the CDMAM phantom is that discs are located in one of the four corners of the 205 cells in which the phantom is divided. However, due to a group of discs is always seen while other group is never seen, the evaluation procedure is focused on a little number of cells. In addition, the tolerances established in some protocols for some discs could reduce the evaluation to a smaller number of discs. In consequence, the memory effect in the observer cannot be rejected.

A software tool was developed to improve the features of CDMAM image phantom. This software tool ensured that the 4-alternative forced choice method of CDMAM is kept, even when is being scored by highly expertise observers familiar on the test object pattern. For digital images, the developed software tool automatically changed the image position of the four corners. It could be selected a fixed rotation angle or a random one, so making impossible that any observer was able to remember the exact corner position of the target disc inside any cell. Two alternative successful algorithms had been tested. ROC curve analysis obtained by 36 observers showed that both original and computer-modified images are indistinguishable. The ROC area was 0.507±0.024 for first algorithm and 0.522±0.026 for the second one, denoting that there was no statistical difference between real and computer-modified images for both algorithms.

This first development allowed us an initial approach to experimental environments on image quality perception and was a first step to develop complex algorithms in Java.

The automated perception problem

We developed a second work with the aim to analyze the potential of the cross-correlation component of the multi-scale structural similarity metric (R*metric, developed by Rouse and Hemami), to predict human performance in detail detection tasks closely related with diagnostic x-ray images. To check the effectiveness of R*, we have initially applied this metric to a contrast detail detection task.

Threshold contrast visibility using the R* metric was determined for two sets of images (set 1 and set 2) of a contrast-detail phantom (CDMAM). Results from R* and human observers were compared attending to the contrast threshold and using the Constant Efficiency method. A comparison between the R* metric and two algorithms, currently used to evaluate CDMAM images was also performed.

Similar trends for the CDMAM detection task of human observers and R* metric were found in this study. Threshold contrast visibility values using R* were very close to those obtained by human observers with average deviations less than 16% for the images in set 1 (linear coefficient of correlation of 0.984) and 11% for those in set 2 (linear coefficient of correlation of 0.993).

These results using R* showed that it could be used to mimic human observers for certain tasks, such as the determination of contrast detail curves in the presence of uniform backgrounds. The algorithm here designed based on the R* metric could outperform other metrics and algorithms currently used to evaluate CDMAM images and could automate this evaluation task.

These results proved the possibility of applying the R* metric, belonging to the SSIM family, to the medical imaging area of research applying adequate experimental conditions and methodology.

Perception in uniform backgrounds vs. perception in anatomical backgrounds

The human perception of small details of interest changes in the presence of structured backgrounds. Tiny, but medically significant signals, are masked in the presence of anatomical structures, due to changes in the signal intensity or due to the presence of background structures with the same size of the relevant signal.

To analyze this problem, it was of interest to compare the response of a mammography system to the same set of signals, either embedded in flat or in real backgrounds. This comparison achieved two goals. The first one was to analyze the variation of the recognition threshold of the system for both backgrounds. The second one was to analyze the performance of a human observer or a model observer over the same set of signals, varying the nature of the backgrounds.

A software tool was developed to merge CDMAM phantom images with real mammographic backgrounds. It allowed SKE tasks in uniform and in real backgrounds. This kind of tasks can be used to compare human, human visual metric or model observer performance in detail detection using uniform or mammographic backgrounds.

As it is very well known, local characteristics of the structures in real mammographic backgrounds reduce the human performance in contrast-detail detection tasks. In consequence that performance cannot be inferred from the data acquired in white noise (flat) backgrounds such as a CDMAM phantom produces.

The software tool used CDMAM images to merge with a region of interest selected from a real mammography. This region as well as the mixing image method (adding or multiplying pixels) could be freely selected by the user. In this work a set of measurements of 8 images had been analysed. We could preview the variation of the contrast-detail detection for a human observer and a human visual system metric (R*).

Four relevant facts could be observed in mammographic backgrounds. The first one is the lower response of the human observer in mammographic backgrounds related to uniform ones, due to the structured noise in the image. This effect has been described widely in the literature.

The second fact is the lower response of the R* metric in mammographic backgrounds related to the response of R* metric in the uniform backgrounds, as can be expected, due also to the higher (structured) noise presented in the image.

The third fact is that contrast threshold increases as disc size increases for the largest disc diameters, due to the masking effects of the structures of the mammographic background. These structures have a size similar to the greater discs. This fact has also been shown by several researchers.

The fourth fact was the similar behaviour shown by the metric R* and the human observer.

This software tool could be applied to generate hybrid images merging CDMAM images and real mammographic backgrounds and to compare the performance of different observers (human or automated) for contrast-detail detection. Its application to an actual problem validated the results obtained, similar to others well known by the scientific community, and showed its potential as tool of analysis of the performance of different observers.

Broadening the problem: different types of noise and different types of images

Earlier in this document we described the SSIM family. Results in large studies have shown that SSIM and MS-SSIM mimic quite well the perceived quality of an image by a human observer. However, they show some limitations:

Some researchers have found that SSIM and MS-SSIM do not perform so well for recognition threshold tasks (tasks near the perception limit), which invalidate their application to the analysis of images with regions of interest at the limit of visibility.

Some studies show limits in the performance of these indexes analysing medical images. Other studies show that the correlation between SSIM and MS-SSIM and human observers decreases when they are used to measure the quality of blurred and noisy images.

These drawbacks are limiting factors in the medical imaging area, specifically in Radiology. Radiological images of medical interest show subtle differences between the image with no pathological findings and the image that shows these findings. Blur and noise are some of the most usual distortion factors in a day-to-day radiological practice.

Some authors, as it has been said before, have proposed some modifications of SSIM and MS-SSIM to avoid these limitations. Rouse and Hemami proposed a new IQM in 2009, r*, based on the structural component of MS-SSIM that could avoid the lack of effectiveness near the recognition threshold. This index has been broadly tested by the authors of the present job.

Chen et al. proposed in 2006 a gradient-based SSIM (G-SSIM) that improves the SSIM results in blurry and noisy images. Li and Bovik applied in 2010 a four-component model based on the texture and edge regions of the image. They applied this model to SSIM and MS-SSIM, getting eight new IQM. These three approaches have shown promising features to overcome the limitations of SSIM and MS-SSIM.

The aim of the last step of this research was to analyze the potential of these modifications in the SSIM family, testing in a medical environment a complete set of 8 proven and 8 new IQM proposed here, the latter created by combination of all the related approaches.

To check the effectiveness of these IQM, we have applied these metrics to a double-stimulus task with a database of radiological images. This database comprised different acquisition techniques (MR and Plain Films). The images in the database were distorted with four different types of distortions: Gaussian blur, Gaussian noise, JPEG, and JP2000, and five different levels of degradation. These images were analysed by a board of radiologists with a double-stimulus methodology and their results were compared to those obtained from the 16 metrics analysed and proposed in this research.

Our experimental results showed that the human observer readings were sensitive to the edge information between the reference and the test images, the changed and preserved edges, and the textures. Previous studies, that mixed these techniques, have shown the low relevance of using multi-scale approaches, simulating different viewing distances from the image to the observer. On the contrary, we have found the superiority of this approach over single-scale approaches, which take into account only one viewing distance.

These results showed that several metrics (4-G-SSIM, 4-MS-G-SSIM, 4-G-r*, and 4-MS-G-r*) can be used as good surrogates of a radiologist to analyze the medical quality of an image in an environment with a reference image. Specially 4-MS-G-SSIM keeps an excellent performance for all types of image and distortion.

Conclusions

1. Some algorithms may be used to manipulate phantom images, in order to avoid the memory effects in users with a high degree of expertise analysing these phantom images. These image alterations can be developed through different methods and none of them is noticeable by the user.

2. There are image quality metrics coming from different fields of those of medical imaging that can be applied to certain medical tasks. This thesis is focused on the analysis of one of the most successful family: the SSIM family. Specifically, the R* metric, designed to manage signals near the perception limit, can be applied for the automated evaluation of medical phantoms. Its results are statistically indistinguishable from those obtained from a human observer, so this metric could be applied to automated perception tasks.

3. Software tools can be developed to generate hybrid images, merging phantom images (specifically CDMAM images) and real mammographic backgrounds to compare the performance of different observers (human or automated) for contrast-detail detection. The R* metric and its application to an actual problem shows similar results to others well known metrics developed by the scientific community, and shows its potential as tool of analysis of the performance of different observers. Besides, our study stands out the masking effect in the signal recognition due to structures with similar size to that of the signals.

4. There are metrics whose behaviour is very similar to that shown for a human expert analysing the quality of a medical image. These experiments have been running throughout a wide database of medical images and a broad set of different types of noise and distortions of interest in medical imaging. The high correlation shown by those components, focused on the analysis of texture changes, edges of the structures, different viewing distances and cross-correlation between images, support the theories that state that these components are of great relevance for the human visual system.

Last, but not least, we want to share our efforts with the scientific community. The whole set of programs and algorithms we have developed and applied in this study, will be freely available in our website (https://www.ucm.es/gabriel_prieto) for the scientific community. At this time, some part of them are already freely available in our website, and some of these algorithms have been applied by several groups of research through different studies of image quality perception.