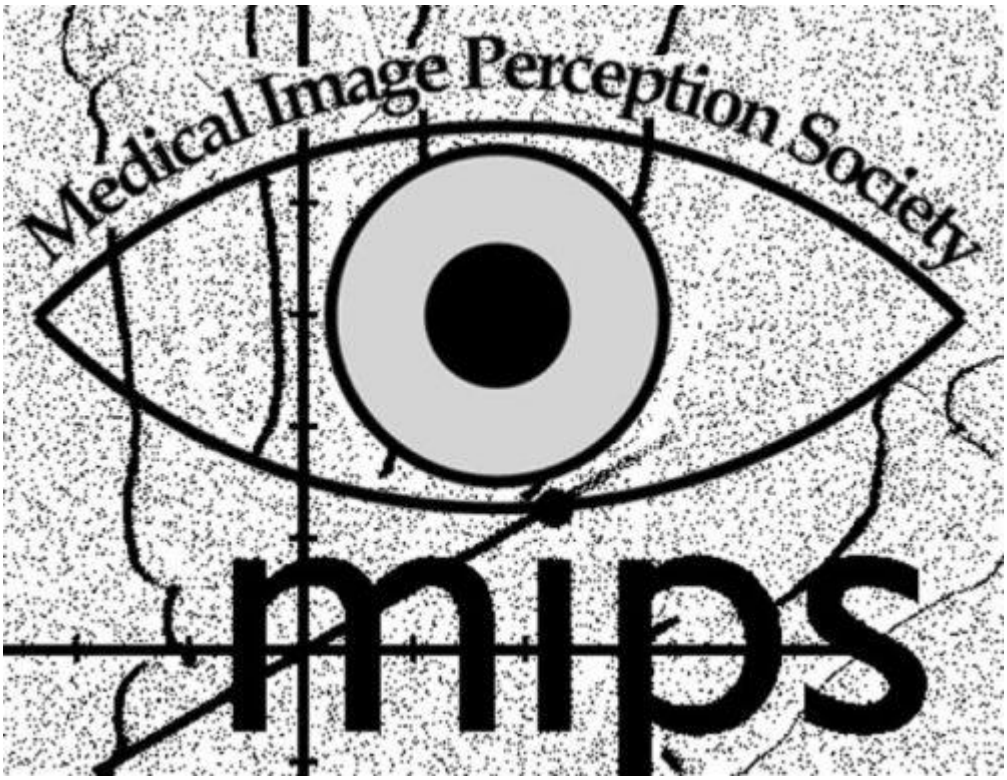
MIPS XVII

July 12-15, 2017

Whitehall Houston Hotel

Houston, TX



ORGANIZING COMMITTEE

Tamara Haygood, PhD, MD

The University of Texas MD Anderson Cancer Center

Mia Markey, PhD

The University of Texas Austin

Mini Das, PhD

University of Houston

Howard Gifford, PhD

University of Houston

Gezheng Wen, MS

The University of Texas Austin

Elizabeth A. Krupinski, PhD

Emory University

KEYNOTE PRESENTATION

Saturday July 15th 9:20 am

"Image Perception in Ophthalmology: Practices, Challenges, and Computational Approaches"

Helen Li, MD

Houston Methodist Weill Cornell Medical College

Baoxin Li, PhD

Arizona State University

Dr. Helen Li specializes in ocular oncology and vitreoretinal diseases. She founded the Community Retina Group in Houston after almost 20 years of clinical practice at the University of Texas Medical Branch, where she received numerous teaching awards from residents and fellows. Dr. Li served as director of the UTMB retina service and founded the retina fellowship program. She regularly publishes her research in leading ophthalmology journals and presents on retinal diseases and ocular tumors at national and international scientific conferences. She is an active member of the Macular Society, Retina Society, the International Society of Ocular Oncology, the American Uveitis Society, the Association of Research in Vision, and the American Telemedicine Association. Dr. Li has served as principle investigator of National Eye Institute, industry and foundation sponsored research and clinical trials. Her research projects include the ocular complications of AIDS trial and the diabetic retinopathy clinical research network study. She is currently interested in validating biomarkers for prognosis of melanoma in the eye and evaluating novel therapeutics for various types of ocular cancer.

Dr. Baoxin Li is currently a professor and the chair of the Computer Science & Engineering Program and a Graduate Faculty Endorsed to Chair in the Electrical Engineering and Computer Engineering programs. From 2000 to 2004, he was a Senior Researcher with SHARP Laboratories of America, where he was the technical lead in developing SHARP's HiIMPACT Sports™ technologies. He was also an Adjunct Professor with the Portland State University from 2003 to 2004. His general research interests are on visual computing and machine learning, especially their application in the context of human-centered computing. He won twice SHARP Labs' President's Awards, in 2001 and 2004 respectively. He also won SHARP Labs' Inventor of the Year Award in 2002. He is a recipient of the National Science Foundation's CAREER Award. He holds 16 issued US Patents. His work has been featured on NY Times, EE Times, MSNBC, Discovery News, ABC News, Gizmodo India, and many other media sources. His research interests are in computer vision and pattern recognition, image/video processing, statistical methods in visual computing.

DINNER SPEAKER

Friday July 14th 7:15 pm

"**Medical Care of Allied Prisoners of War in WWII Germany"**

Tamara Haygood, PhD, MD

The University of Texas MD Anderson Cancer Center

Tamara Haygood, PhD, MD is currently an Associate Professor in the Department of Diagnostic Radiology, Division of Diagnostic Imaging, The University of Texas MD Anderson Cancer Center, Houston, TX. She is a native Texan and came to Anderson from a solo private practice in LaGrange, Texas. She received her undergraduate and PhD degrees from Rice University in History and her MD from The University of Texas Houston and did her residency at the University of New Mexico. Her research interests lie in decision-making and medical image perception, radiologic reporting, and medical care of Allied prisoners of war in WWII Germany. When not at work, she enjoys gardening, travel, walking, and her family, friends, and pets.

MIPS SCHOLARS

Koos van Geel *Maastricht University*

Amareswararao Kavuri *University of Houston*

Kristina Landino *George Washington University*

Lucie Leveque *University of Hull*

William Nesbitt *University of Houston*

Krista Nicklaus *University of Texas Austin*

Sean Rose *University of Chicago*

Lauren Williams *University of Utah*

Hanshu Zhang *Wright State University*

Scholars are supported in part by NIBIB/NCI R13EB024389

Wednesday July 12, 2017			
		Start Time	**End Time**
Welcome Reception (light hors d'oeuvres)		5:00 pm	7:00 pm
Thursday July 13, 2017			
Tamara Miner Haygood & Mia K. Markey	Welcome & Announcements	8:00 am	8:20 am
***Scientific Session 1: Methodology* Chair: Brandon Gallas**			
Frank W. Samuelson	ROC curves from MAFC experiments using a sorting algorithm	8:20 am	8:40 am
Tamara Miner Haygood	Memory bias in observer-performance literature	8:40 am	9:00 am
Stephen L. Hillis	What aspect of reader performance are we interested in?	9:00 am	9:20 am
Newcomer Introductions (Elizabeth A. Krupiniski)		9:20 am	9:50 am
***Scientific Session 2: Errors I* Chair: Mark F. McEntee**			
Mark F. McEntee	Examining the "gambler's fallacy" in radiology	9:50 am	10:10 am
Trafton Drew	The cost of distraction: Quantifying the effects of interruption during diagnostic radiology using mobile eye tracking	10:10 am	10:30 am
Coffee Break & Poster Viewing		10:30 am	11:00 am
***Scientific Session 3: Mammography I* Chair: Robert Nishikawa**			
Kristina Landino	Comparing salience detection algorithms applied to mammograms	11:00 am	11:20 am
Ali Avanaki	Towards an anthropomorphic model observer for spiculated masses	11:20 am	11:40 am
Delgermaa Demchig	Automatic segmentation of the dense tissue in digital mammograms for BIRADS density categorization	11:40 am	12:00 pm
Lunch Break		12:00 pm	1:30 pm
***Scientific Session 4: Mammography II* Chair: Lonie Salkowski**			
Hayden Schill	Detecting the "gist" of breast cancer in mammograms	1:30 pm	1:50 pm
Jay Hegdé	Quantitative characterization of eye movements during 'deep learning' of diagnostic features in mammograms	1:50 pm	2:10 pm
Elizabeth A. Krupinski	Human, animal & computer-based medical image interpretation: What can we learn?	2:10 pm	2:30 pm
***Scientific Session 5: Ultrasound* Chair: Howard C. Gifford**			
Lucie Lévêque	Quality assessment of ultrasound video for medical tele-assistance	2:35 pm	2:55 pm
Yasser M. K. Omar	Automatic Selection Of The Best Despeckle Filter Of Ultrasound Images	2:55 pm	3:15 pm
Speed Mentoring (Tamara Miner Haygood)		3:15 pm	3:40 pm
Coffee Break & Poster Viewing		3:40 pm	4:00 pm
***Scientific Session 6: Surgery* Chair: William Auffermann**			

Koos van Geel	Neural correlates of expertise in radiology and surgery: Toward ecological validity in fMRI research	4:00 pm	4:20 pm
Krista M. Nicklaus	Towards augmented reality visualization of mastectomy specimens for breast reconstruction surgery	4:20 pm	4:40 pm
Dinner on your own			
Friday July 14, 2017			
Elizabeth Krupinski	MIPS Business Meeting	8:00 am	8:20 am
***Scientific Session 7: Breast Tomsynthesis* Chair: Francine Jacobson**			
Krista M. Nicklaus	A human observer study of multi-lesion detection in digital breast tomosynthesis	8:20 am	8:40 am
Amareswararao Kavuri	Understanding noise power spectrum in light of human observer detection in tomosynthesis	8:40 am	9:00 am
Stephen H. Adamo	Mammography to tomosynthesis: Comparing two-dimensional to three-dimensional visual search in radiologists and undergraduates	9:00 am	9:20 am
William H. Nisbett	Understanding the impact of local texture features on search and localization in digital breast imaging	9:20 am	9:40 am
Mia K. Markey	Scavenger Hunt Results	9:40 am	9:50 am
Coffee Break & Poster Viewing		9:50 am	10:15 am
***Scientific Session 8: Training & Education* Chair: Trafton Drew**			
Lonie R Salkowski	Cognitive processing differences of experts and novices when correlating anatomy and cross sectional imaging	10:15 am	10:35 am
Koos van Geel	Training the evaluation of radiographs: Normal-abnormal proportion differentially influences sensitivity and specificity	10:35 am	10:55 am
Mark McEntee	Do Nigerian radiographers have potential to interpret chest radiographs?	10:55 am	11:15 am
William F. Auffermann	Search pattern training for central line positioning on chest radiography	11:15 am	11:35 am
Group Photo		11:35 am	11:50 am
Lunch Break		11:50 am	1:20 pm
***Scientific Session 9: Errors II* Chair: Murray Loew**			
Jeremy M. Wolfe	Mixed hybrid search: A model system to study incidental finding errors in radiology	1:20 pm	1:40 pm
Lauren Williams	Interruptions in diagnostic radiology: Is there a tradeoff between speed and accuracy?	1:40 pm	2:00 pm
Volunteers!	2019 Meeting Bids	2:00 pm	2:15 pm
***Scientific Session 10: Pathology & Dermatology* Chair: Tamara Miner Haygood**			
P. Suhail Parvese	Towards automated analysis of nucleolar and centromere shapes in indirect immunofluorescence	2:15 pm	2:35 pm

Brandon D. Gallas	MRMC analysis of mitotic counts	2:35 pm	2:55 pm
N. Punitha	An approach to classify dermoscopy images using dynamic restricted Boltzmann machine	2:55 pm	3:15 pm
Coffee Break & Poster Viewing		3:15 pm	3:40 pm
***Scientific Session 11: Localization* Chair: Mini Das**			
Craig K. Abbey	Estimated templates for forced-localization tasks in ramp-spectrum noise	3:40 pm	4:00 pm
Craig K. Abbey	Observer effects in 3D search from classification images	4:00 pm	4:20 pm
Alexandre Ba	Peripheral vision implication in the search and recognition of low contrast hepatic metastasis in abdominal CT scans: Preliminary study with an eye-tracker	4:20 pm	4:40 pm
Houston Historical Walking Tour (separate registration required)		5:00	6:00
Conference Dinner with Speaker (separate registration required) At the hotel after the walking tour		6:30	8:30
Saturday July 15, 2017			
Mini Das & Howard C. Gifford	Announcements	8:00 am	8:20 am
***Scientific Session 12: Model Observers* Chair: Craig Abbey**			
Sean D. Rose	Parameter selection for linear iterative image reconstruction in breast tomosynthesis with the non-prewhitening and Hotelling observers	8:20 am	8:40 am
Miguel A. Lago	Foveated model observer predicts dissociation of signal detectability across 2D and 3D images	8:40 am	9:00 am
Howard C. Gifford	The role of pre-whitening in visual-search models of human observers	9:00 am	9:20 am
***Keynote Speakers* Chair: Elizabeth Krupinski**			
Helen Li & Baoxin Li	Image Perception in Ophthalmology: Practices, Challenges, and Computational Approaches	9:20 am	10:05 am
Coffee Break		10:05 am	10:30 am
***Scientific Session 13: Interpretation* Chair: Mia K. Markey**			
Tamara Miner Haygood	Consultation and citation rates for older imaging studies and documents in radiology	10:30 am	10:50 am
Elizabeth A. Krupinski	Musculoskeletal discomfort in radiologists	10:50 am	11:10 am
Hanshu Zhang	Single display versus dual displays: A cognitive modeling perspective	11:10 am	11:30 am
Mark McEntee	Australian breast reader assessment strategy on mammographic improves radiologists' test reading performance	11:30 am	11:50 am
Mia K. Markey & Tamara Miner Haygood	Wrap-Up & Adjourn	11:50 am	12:00 pm

ROC Curves from MAFC Experiments Using a Sorting Algorithm

Frank W. Samuelson, Ph.D.
Division of Imaging, Diagnostics, and Software Reliability, OSEL/CDRH/FDA

Rationale

Sometimes the area under the ROC curve, a summary of reader performance, is calculated from experiments that collect rating data from observers and sometimes from multiple alternative forced choice (MAFC) experiments. Estimates of AUC between these types of studies may be different because of reader behavior. We would like to compare the form of ROC curves between such studies also. Generally ROC curves can not be generated from MAFC data, so we need to create a method to create ROC curves.

Methods

We developed an algorithm for producing ROC curves from MAFC experiments. The algorithm is efficient in terms of the number of MAFC trials or observations required. This algorithm is similar to common computer sorting algorithms, because constructing an ROC curve is equivalent to sorting the images from least suspicious to most suspicious. The algorithm selects which M images should be presented to the observer for each observer input. As the observer compares and selects images, the algorithm sorts the images by completing a success matrix, and presents the observer with images that are most likely to be informative in sorting. For complete sorting all images need to appear in multiple MAFC trials. To avoid the repeated sequential display of the same image, the algorithm does not use pivots, which is different from typical sorting algorithms. The algorithm can utilize a comparison of an arbitrary M images rather than the typical comparison of two elements. To reduce the number of required MAFC trials required we preferentially select images with few trials and images that have similar estimated ranks. The number of MAFC trials may be additionally reduced with the prior use of a human-like model observer.

Results

Simulations demonstrate that the algorithm is relatively efficient. For an equal number of N/2 signal present (SP) and N/2 signal absent (SA) images, a most efficient sorting algorithm can sort all images in approximately $0.85\, N \cdot \log_2 N$ 2AFC trials. Our algorithm can sort the N/2 SP images with respect to the N/2 SA images to generate a ROC curve in approximately the same number of trials. At the time of the conference we will present results from studies with real readers.

Conclusions

We developed an efficient sorting algorithm for the construction of ROC curves from MAFC experiments. This will allow us to compare forms of ROC curves from data collected in different kinds of reader study experiments.

Memory Bias in Observer-Performance Literature

Tamara Miner Haygood, PhD, MD; Samantha Smith

Department of Diagnostic Radiology, UT M.D. Anderson Cancer Center, Houston, Texas, USA;
Tamara.Haygood@mdanderson.org

Rationale

Observer-performance studies using human observers often test the ability of the observer to accomplish a certain task such as detection of pulmonary nodules or breast masses under a variety of different conditions.. When the images being interpreted between one condition and the next are similar, the results can be affected by an observer's recognition of an image. There is relatively little advice available in the radiology literature on how to prevent recognition memory. We studied what published authors have done in this regard.

Methods

We searched AJR online using the terms "observer performance" and "observer study" to identify articles reporting observer-performance experiments. We identified 50 articles dating between 1973 and 2016. We extracted from each the number of reported experiments and five factors related to the authors' approach to recognition memory. These included 1) separate or other viewing of the tested conditions, 2) random or other ordering for tested condition, 3) random or other ordering for cases within sessions, 4) availability of clinical information.

Results

In these 50 articles, 57 separate experiments were reported.

Based on how the tested conditions were presented to viewers, these experiments fit into three categories: 1) Sequential-viewing experiments in which the second tested condition would normally be used as an adjunct to the first, for example, computer-assisted detection in mammography. There were 11 of this type of experiment. 2) Experiments in which memory for images was irrelevant such as alternative forced-choice or multi-point ranking experiments. Sixteen experiments were in this category. 3) Experiments in which image memory remained a potential source of bias, of which there were 30. These experiments are considered in this report.

Among these 30 experiments, tested conditions were presented in a counter-balanced order in 8 experiments, in the same order for each reader in 9 experiments, in random order in 3 experiments, in unique orders in 3 experiments. 7 reports did not indicate in what order the tested conditions were presented.

In the 40 experiments using separate viewing of tested condition, individual images were ordered randomly in 17 and in a pseudo-random fashion in 3. Twenty reports did not indicate how they were ordered.

Within a reading session, 13 experiments presented cases in random order, 2 presented them in a pseudo-random order, and 15 did not indicate how they were ordered.

18 experiments used a time lapse between viewings. Time lapses ranged from 1 to 730 days, 73.7 days average, 21 days median. 11 reports did not indicate if a time lapse was used, and one used sequential viewing.

Readers were blinded to patient information in 10 experiments. Three reports said there was no clinical information given but did not specifically say that patient identifying information was unavailable. In 15 reports there was no indication of whether patient information was available. 2 experiments used phantoms.

Conclusions

In the spirit of providing full details of how an experiment was conducted, many of these published papers would have benefitted from a more precise discussion of how they avoided memory bias.

What aspect of reader performance are we interested in?

Stephen L. Hillis, PhD, Departments of Radiology and Biostatistics

Rationale

There seems to be no end to the debate over which reader performance estimator should be used for comparing modalities in a diagnostic radiologic imaging study. Should one use a receiver-operating-characteristic (ROC), free-response ROC (FROC), region-of-interest (ROI), or location ROC (LROC) method? If one uses ROC, should one use ROC AUC, pAUC,..., etc.?

Methods

I propose that the reason that this debate never ends is because there is not agreement on what information is being sought. More specifically, if we cannot agree on what parameter (known in statistical jargon as an estimand) we want to estimate, then there can never be agreement on which estimator to use, since the estimators mentioned above estimate different parameters, i.e., different aspects of reader performance. Astonishingly, there is little discussion at this meeting or other radiologic meetings that I've attended as to what information is being sought. A primary purpose of this talk is to encourage such discussion.

Results

To illustrate the importance of knowing what performance information is being sought, I consider three different situations – screening mammography, diagnostic mammography, and CAD – and discuss, from my understanding, what aspect of reader performance is most relevant for each situation. I also discuss which of the estimators mentioned above appears to best provide the desired reader performance information. I suggest a new method that is an extension of LROC that allows the research flexibility in specifying the reader performance aspect of interest and demonstrate this method with an example.

Conclusions

The purpose of this talk is primarily to point out a large deficiency in the field of diagnostic radiology – the absence of discussion as to what makes one reader's performance better than another reader's performance for different situations. Because currently available reader-performance estimators (such as the ROC AUC and the JAFROC statistic) estimate different aspects of reader performance, there cannot be agreement as to which is most appropriate unless

there is agreement on what we want to know. Hopefully, as a result of this talk there will be more discussion regarding what reader performance information is sought for various situations. Once it is clear what information is being sought, then it will be possible determine which of the currently available estimators is most appropriate, or whether there is a need for developing a new estimator that provides more relevant information.

Examining the ‘gambler’s fallacy’ in radiology

Mark F. McEntee, Ph.D.[1], Trafton Drew, Ph.D.[2], Avigael Aizenmann, Ph.D.[3], Ann Carrigan, Ph.D.[4], Ernest Ekpo, Ph.D.[1], Jeremy Wolfe, Ph.D.[5]
[1]Department of Health Sciences, University of Sydney; [2]Department of Psychology, University of Utah; [3]Department of Vision Science, University of California Berkeley; [4]Department of Cognitive Science, Macquarie University; [5]Visual Attention Lab, Harvard University

Rationale

Humans have a tendency to behave as if past events influence events, even when the events are independent, a finding known as the ‘gambler’s fallacy'. A classic example of this fallacy is the tendency of the betting public to bet heavily on ‘red’ on a roulette wheel after a string of ‘black’ outcomes (Clotfelter & Cook, 1993; Tversky & Kahneman, 1971). While it is likely that most radiologists are aware of this fallacy, these tendencies are sufficiently ingrained that they may exert implicit pressure on behaviour even if one has explicit knowledge.

Methods

To test this idea, gave radiologists explicit instructions and asked them to examine single images from 30 mammograms. Mammograms will be displayed on two separate, single 5-megapixel monitors. The software used was (Håkansson et al. 2010). There will be 3 cases with pathology-proven masses distributed amongst the cases. Placement of the cases will be pseudorandom, constrained so that one positive case follows within two cases of another positive case. The gambler’s fallacy prediction would be that, immediately after finding a positive case, (1) time spent on a case should decline, (2) criterion should become more conservative, and (3) the chance of a false negative or mislocalization error should increase.

Results

The proportion of hits and false alarms was 0.5 and 0.07 respectively. Our original hypothesis was that the Gambler’s fallacy would result in an increase of false negative errors for the p2 case, however, the case ratings for p2 (on a modified BIRADS scale) do not significantly differ from p1 ratings. This is likely due in part to radiologists assuming such a high prevalence of positive cases, eliminating the Gambler’s Fallacy. The Gambler’s Fallacy would also predict that a positive case succeeding another positive case would have a decrease in reaction time, as radiologists may have a predetermined notion that the succeeding case is most likely negative. There is no difference in the average reaction time for cases p1 and p2, suggesting radiologists viewed both positive cases as independent events. The failing of the Gamblers’ Fallacy in this paradigm is likely reflected in the high false alarm rate.

Conclusion

Laboratory experiments struggle to demonstrate the gamblers' fallacy as in the experimental setting radiologists’ prevalence assumption results in a high false alarm rate.

The cost of distraction: Quantifying the effects of interruption during diagnostic radiology using mobile eye tracking

Trafton Drew[1], Lauren Williams[1], Booth Aldred[2], Marta Heilbrun[2], Satoshi Minoshima[2]

[1]University of Utah
[2]University of Utah School of Medicine

Rationale

Radiologists are frequently interrupted while performing tasks where a simple mistake may have dire consequences for the patient. What are the costs of these sorts of interruptions? The cognitive psychology literature suggests that there will be clear costs on both the speed and accuracy when tasks are interrupted. Moreover, some observational research has found that the number of discrepant diagnoses increases with the number of phone calls during a given shift (Balint et al., 2014). By some estimates, ~30% of all errors during diagnostic radiology are due to perceptual errors (Berlin 2007). We wondered whether increasingly frequent interruptions that occur in the reading room could be contributing to this problem.

Methods

Thirty-two radiologists (Rs) participated in two experiments: 16 at the University of Utah (Experiment 1) and 16 while attending RSNA (Experiment 2). The experiment took place at a modified workstation with 2 (RSNA) or 3 (Utah) monitors. Rs were given a worklist and told to read through the cases as quickly and accurately as possible. The worklist was populated with a mixture of volumetric (e.g. Chest CT) and 2d (e.g. chest radiograph) images. Rs were asked to dictate their impressions of each case. Diagnostic accuracy was coded based on dictation. In Experiment 1, Rs were interrupted on two cases by a phone call. Upon answering the phone, a pre-recorded message asked them to find a patient in a different worklist and provide a quick diagnosis for a patient. In Experiment 2, Rs were interrupted by a Research Assistant who asked them to stop reading the case they were working on and fill out a form with demographic information. In both experiments, the cases that were interrupted were manipulated across Rs so that an equal number of Rs saw each critical case with and without an interruption. We monitored eye-position throughout both experiments using mobile eye-tracking glasses.

Results

In Experiment 1, we observed a significant increase in the amount of time spent on cases that were interrupted (Mean: 529s Interrupted, 367s Uninterrupted: $t(14)=4.5$, $p<.001$), but no cost on diagnostic accuracy ($t(14)=1.17$, p=n.s.). The observed time cost may have been driven by an increased proportion of time spent looking at the dictation screen and a decreased amount of time examining medical images on interruption trials. In Experiment 2, there was no time cost associated with the interruption (Mean: 365 Interrupted, 416s Uninterrupted: $t(14)=-1.75$, p=n.s.). Similarly, the proportion of time spent looking at medical images was unaffected by the interruption.

Conclusions

Our results clearly indicate that the nature of the interruption has important implications for how it will influence performance. We observed a large time-cost when the interruption involved looking at additional medical images and none when it did not. This suggests that the degree of overlap between the primary task and the interrupting task is an important predictor for the resultant cost of interruption. In ongoing work with naïve observers performing an analogous task, we are currently examining whether this prediction holds true.

Comparing salience detection algorithms applied to mammograms

Kristina Landino, B.S., and Murray Loew, Ph.D.
Department of Biomedical Engineering, George Washington University, Washington, DC 20052

Rationale

Salience in imaging is defined as the extent to which an object in an image catches the eye of the viewer. Currently, several software packages exist which calculate salience using a wide range of models and implementations. For example, in several of the models examined here, the software creates a series of maps for individual salience features like orientation and intensity, and then combines those individual feature maps into an overall map of salience for the entire picture. Differences exist between these models in the way feature maps are calculated, in the way they are integrated into a single map, and in the ways that various types of noise are minimized. Differences also exist in the use of mid-level and high-level features like horizon line detection and facial and person recognition software. In some algorithms, the extent to which any one feature affected the final salience map was dependent, through the use of covariance, on the other features. While in other algorithms, features contributed independently while in others they behaved additively or competitively. In yet other models, neural networks are used to create a series of layers, each of which transforms the data and finds the most salient points in an image. Those models focus on finding semantic objects, or objects defined by a set of attributes, as well as low-level features. In total, this paper compares 15 models, including our own algorithm, and compares the models' accuracies when applied to a common database of images. Additionally, previous work has shown a correlation between the salient points in a mammogram and presence of a mass, and the inverse relationship between salience and time-to-first detection of a mass by a human observer. Here we apply several state-of-the-art software packages to a database of mammograms and compare their accuracies in detecting masses in mammograms.

Methods

Each method was tested on a set of 322 mammograms to assess its capability to detect masses in breast tissue. Additionally, all models were tested against MIT's training set of images and eye-track data to verify previous results. To understand the significance of the results, we used the Kullback-Leibler divergence, Pearson correlation coefficient, and several other measures including area under the receiver operating characteristic curve (AUC) to compare the salience maps created by the software packages to a baseline map created by Gaussian smoothing of eye-track data.

Results

We will present the results of this work, rank the various algorithms, and comment on the ability of different salience approaches to be successful when identifying masses in mammograms. We found that software packages, which were most capable of detecting salience as defined by eye-tracking data, were capable of detecting masses.

Conclusions

The results of our study show that salience software can potentially be a useful tool when attempting to identify masses in breast tissue. In addition, the software studied here may also be expanded to video, allowing for salience detection of motion features in medical video.

Towards an anthropomorphic model observer for spiculated masses

Ali Avanaki, PhD, Kathryn Espig, MSc, Albert Xthona, MSc, Tom Kimpe, PhD, MBA

Barco Healthcare

Rationale

As compared to other abnormalities in mammograms, spiculated masses (SpMs) are more likely to be malignant. This motivated prior research on computer-aided detection of SpMs[1,2], using Hough or Radon transforms to detect the spicules, and/or using alternative image analysis[3,4,5]. However, we found no prior study in the literature addressing the perception or detection of SpMs by human observers. An anthropomorphic model observer for SpMs may be used in virtual clinical trials for optimization of medical imaging and visualization systems.

Methods

We adapt Barten's model[6] for visibility of sinusoidal patterns to the anthropomorphic detection of SpMs, consisting of a central mass, assumed to be generally round and at a known location, from which several spicules emanate, as follows. We theorize that the detection of a SpM is equivalent to the detection of its central mass (i.e., central mass contrast with respect to its surround should exceed a certain threshold) *and* the detection of several spicules. The latter is modeled by angularly unwinding the concentric rings surrounding the central mass and inspecting whether there is enough contrast for visibility of each spicule in several adjacent rings. By allowing small shifts between the visible activities detected in different rings, the non-radial spicules can be also detected. Note that the viewing distance may be considered optimal for viewing the central mass or the spicules but not both. Validation will be conducted by gauging the visibility of SpMs by humans[7] and comparing those against the detection probabilities produced by the model observer.

Results and conclusion

Our preliminary results are promising. This is a work in progress and it is too early to draw a conclusion. For example, the validation against human observers (to be performed) will most likely reveal that the method of combining the results of spicule detection and central mass detection into a single SpM detection probability depends on the specific human observer to be modeled and the viewing condition (e.g., viewing distance or display contrast) which determine the amount of detection-related information in the visual elements of the SpM (i.e., each of the spicules and the central mass).

---

[1] Karssemeijer, N. (2002). Detection of Masses in Mammograms In R. N. Strickland editor, *Image-Processing Techniques for Tumor Detection* (pp. 187-212). New York, NY: Marcel Dekker.

[2] Sampat, M. P., Markey, M. K., Bovik, A. C. (2005). Computer-Aided Detection and Diagnosis in Mammography In A.C. Bovik editor, *Handbook of Image and Video Processing* (pp. 1195-1217). Burlington, MA: Elsevier.

[3] Muralidhar, G. S., *et al* (2010). Snakules: A model-based active contour algorithm for the annotation of spicules on mammography. *IEEE Transactions on Medical Imaging*, *29*(10), 1768-1780.

[4] Sampat, M. P., Bovik, A. C., Whitman, G. J., & Markey, M. K. (2008). A model-based framework for the detection of spiculated masses on mammography. *Medical physics*, *35*(5), 2110-2123.

[5] Krylov, V. A., & Nelson, J. D. (2014). Stochastic extraction of elongated curvilinear structures with applications. IEEE Transactions on Image Processing, 23(12), 5360-5373.

[6] Barten, P. G. (1999). *Contrast sensitivity of the human eye and its effects on image quality* (Vol. 72). SPIE press.

[7] Avanaki, A., Espig, T., Xthona, A., & Chesterman, F. (2017). How does display brightness affect lung CT reading? SPIE MI live demo workshop.

Automatic Segmentation of the Dense Tissue in Digital Mammograms for BIRADS Density Categorization

Delgermaa Demchig MD, Ziba Gandomkar MS, Patrick C. Brennan PhD

Medical Image Perception and Optimization Group (MIOPeG), Faculty of Health Sciences, University of Sydney, Sydney, NSW, Australia

Rationale

Currently, the Breast Imaging Reporting and Data System (BIRADS) density categorization is the most popular tool for density assessment among radiologists. However, it is subject to inter-observer variabilities. Therefore, different automated methods have been proposed for dense tissue segmentation. In [1], a technique based on modeling of breast tissue using a Gaussian mixture model was proposed to segment the fibroglandular tissue in digitized mammograms. We modified and extended this method to segment the dense tissue in digital mammograms and then classified them to different BIRADS density categories.

Methods

Three readers were asked to evaluate 150 craniocaudal (CC) digital mammograms and assign a BIRADS density score to each mammogram. The majority voting was used to determine the label of each image. Half of the cases were cancer-containing while rest of them were normal. The images were collected from nine different machines from seven manufacturers. The steps of the dense tissue segmentation are shown in Figure 1. Briefly, mammograms were filtered using a median filter and then the breast mask was found by thresholding. The mixture of Gaussian distributions was fitted to the grey-level histogram of breast tissue. The appropriate value for the number of components in the model was found iteratively. Finally, based on the fitted model, a threshold was selected to segment the dense area.

In order to find whether the percentage of dense tissue differed significantly among different BIRADS categories, the Kruskal-Wallis H-test was utilized. Pairwise comparisons between different categories were done using the rank-based Tukey-Kramer test.

We compared two different methods for classification of mammograms into four BIRADS categories. First, we thresholded the percentage density into four levels. The cut-off values for

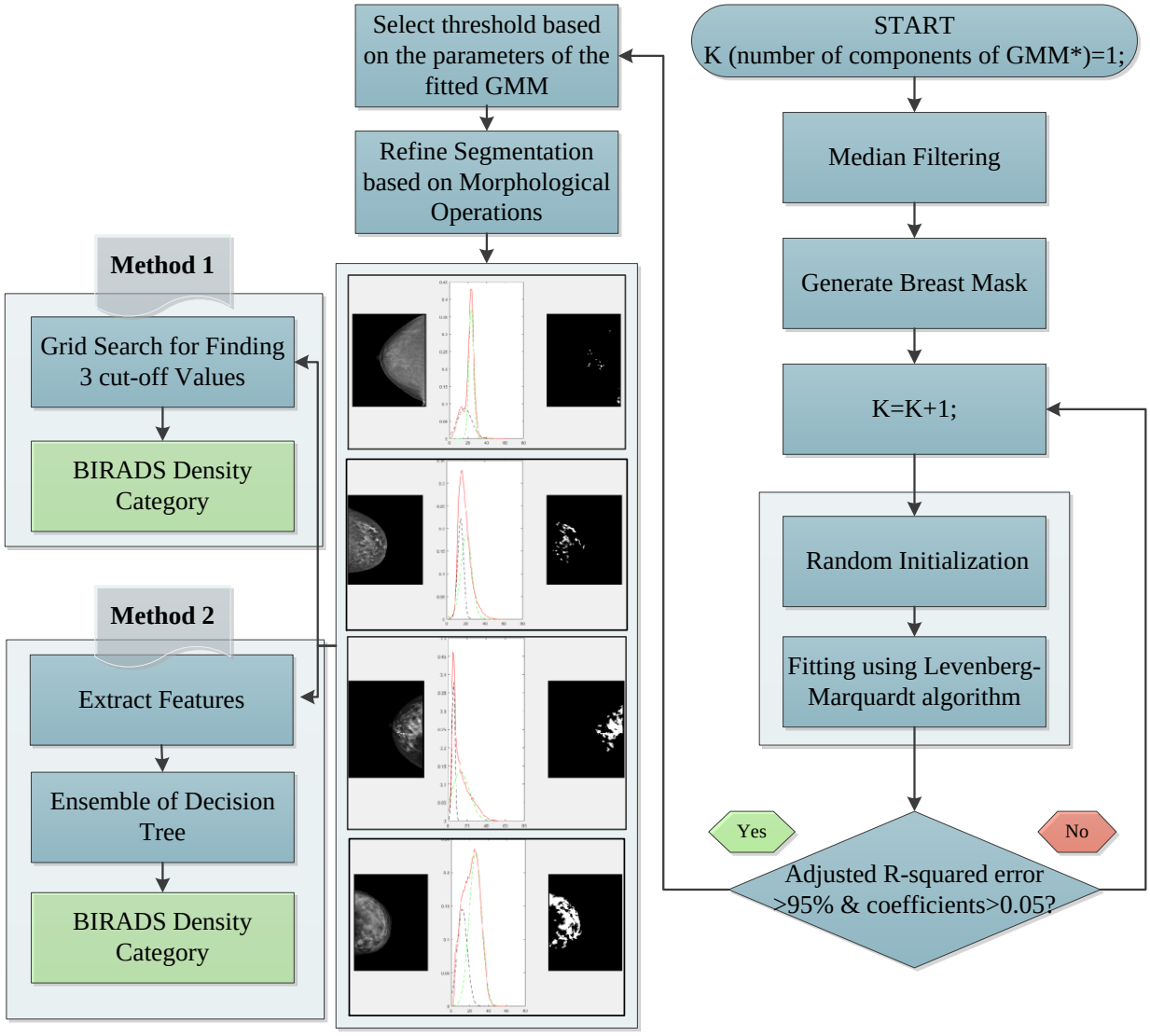
thresholding was found by grid search method. Second, we extracted 21 textural feature [2-4] and three first order statistical features (mean, standard deviation, skewness) from both fatty and dense tissues and fed these features, along with area of dense tissue, total breast area, and percentage density into an ensemble of decision trees for classification. The leave-one-out cross-validation was used to evaluate the method. The statistical analysis and implementation of the algorithm was performed in MATLAB environment.

Results

The percentage density differed significantly among different BIRADS categories ($\chi^2(3) = 89.9$, $p<0.0001$) and differences between all pairs were significant. The first method resulted in a correct classification rate (CCR) of 66.7% for predicting consensus of three radiologists' BI-RADS categories (BIRADS-I: 79.2%, BIRADS-II: 83.1%, BRADS-III: 31.3%, BRADS-IV: 52.2%) while the second method's CCR was 82.7% (BIRADS-I: 79.2%, BIRADS-II: 90.1%, BRADS-III: 75.0%, BRADS-IV: 73.9%). For two-category classification, where BIRADS-I was combined with BIRADS-II (low density) and BIRADS-III with BIRADS-IV (high density), CCRs were 90.0% (high: 95.8%, low: 80.0%) and 90.7% (high: 93.7%, low: 85.5%) respectively for method 1 and 2.

Conclusions

The proposed automatic method was able to predict radiologist-based BIRADS density categories by using both the percentage density with textural and intensity-based features. It can be hypothesized that radiologists consider both amount of dense tissue and tissue texture in density assessment.



*Gaussian mixture model

Figure 1- Steps of the proposed algorithm

References

[1] Ferrari, R. J., Rangayyan, R. M., Borges, R. A., & Frere, A. F. (2004). Segmentation of the fibro-glandular disc in mammograms using Gaussian mixture modeling. Medical and Biological Engineering and Computing, 42(3), 378-387.

[2] Haralick, R. M., & Shanmugam, K. (1973). Textural features for image classification. IEEE Transactions on systems, man, and cybernetics, 3(6), 610-621.

[3] Soh, L. K., & Tsatsoulis, C. (1999). Texture analysis of SAR sea ice imagery using gray level co-occurrence matrices. IEEE Transactions on geoscience and remote sensing, 37(2), 780-795.

[4] Clausi, D. A. (2002). An analysis of co-occurrence texture statistics as a function of gray level quantization. Canadian Journal of remote sensing, 28(1), 45-62.

Detecting the "gist" of breast cancer in mammograms

Hayden Schill BS[1], Anne-Marie Culpan PhD[2], Jeremy M. Wolfe PhD[1,3], Karla K. Evans[4] PhD

[1]Department of Surgery, Brigham & Women's Hospital; hschill@bwh.harvard.org
[2]Division of Biomedical Imaging, University of Leeds; A.M.Culpan@leeds.ac.uk
[3]Department of Ophthalmology and Radiology, Harvard Medical School; jwolfe@partners.org
[4]Department of Psychology, University of York; karla.evans@york.ac.uk

Rationale

The visual system quickly extracts the global structure and statistical regularities of everyday scenes, allowing us to 'get the gist' of our environment before attention captures the details. Gist processing gives us an entry into the meaning of a scene: E.g. On first glance, this is a store where I might find something I want to buy. How might gist processing apply in medical image perception? E.g. This is an image in which I might find something of clinical significance.

Methods I

To investigate this, Evans et al. (2013) presented radiologists and cytologists with normal and abnormal medical images for 250 to 2,000 msec and asked them to rate the level of abnormality and to mark the most likely location of the lesion on a blank image, after this brief exposure.

Results

Results showed that abnormalities could be detected at above chance levels under these conditions even with subtle signs of cancer (at 500 msec, radiologists viewing bilateral mammograms d' = 1, cytologist viewing Pap smears d'=1.2). Radiologists were at chance when they attempted to localize the lesion, suggesting the gist signal is a global/texture signal, rather than an occasional lucky fixation on the lesion. Radiologist continued to perform above chance with images of a single breast. Thus, the signal is not an asymmetry between the breasts nor is it correlated with measures of breast density. Radiologists can distinguish normal from abnormal patients at above chance level even when the 'abnormal' breast is the breast contralateral to the cancer with no presence of the lesion in the image (Evans et al, 2016). These and other findings indicate that some aspect of the texture of the breast, other than the appearance of a lesion, can indicate abnormality. Is that signal present before the cancer, itself, appears?

Methods II

Radiologists were presented with bilateral mammograms that had been acquired 3 years prior to the mammograms that showed visibly actionable cancer. Thus, the abnormal cases were "normal" mammograms from patients who would become "abnormal" in 3 years, developing breast cancer. These cases were intermixed with completely normal mammograms. Radiologists were asked to rate the abnormality of the images on a 0-100 scale after exposure for 500 msec and later rate density of the same images.

Results

The results revealed an ability to distinguish images of the breasts of normal patients from those who would later develop cancer. The signal is small (d'=0.2) but statistically significant ($p < 0.001$). Even though radiologists were viewing images taken 3 years prior to any visible signs of cancer being detected, they were able to classify images as normal or abnormal at above chance levels. These decisions were not based on their breast density judgements.

Conclusion

This further supports the hypothesis that radiologists have access to a global, non-selective signal of abnormality. If that signal could be reliably detected by humans or by computational systems, it could be a valuable part of the effort to assess an individual woman's risk factors and detect cancer early.

Quantitative Characterization of Eye Movements During 'Deep Learning' of Diagnostic Features in Mammograms

Jay Hegdé (PhD)
Department of Ophthalmology, Augusta University, Augusta, GA, USA

Rationale

We have previously shown that implicit statistical learning of abstract patterns, or 'deep learning', can be used to train naïve, non-professional subjects to reliably detect anomalies in mammograms [1]. In the present study, we tested the hypothesis that eye movement patterns change in a learning-dependent fashion during the learning.

Method

We used our previously described deep learning methodology [1-3] to train naïve adult subjects (N = 9) with no previous radiological training to detect anomalies in actual screening mammograms. Another 5 subjects were similarly trained using digitally synthesized, perceptually metameric counterparts of the actual mammograms [1,3]. Subjects were trained to a criterion of d' >= 1.5 (p < 0.05). Eye movements were monitored throughout the training using a high-resolution (2000 Hz) video eye tracker.

Results

Eye movement patterns elicited by actual *vs*. synthetic mammograms were statistically indistinguishable (principal components analysis (PCA), test for linear separability, p > 0.05). Lengths of eye movement trajectory as well as the number of microsaccades roughly followed an inverted 'V' pattern over the course of training, whereby they rapidly rose at the outset of the training, peaked during the steepest part of the learning curve, and fell steadily to asymptotically low levels as the subjects reached asymptotic performance. Trials in which subjects reported finding no anomaly elicited trajectory lengths twice as long and microsaccades twice as frequent as the trials in which subjects reported finding an anomaly. PCA of eye movement trajectories showed that the eigenvalues of the microsaccadic components were inversely correlated with performance during the given block (r = -0.53, df = 1347, p << 0.05).

Conclusions

Taken together, our results indicate that eye movements during the acquisition of diagnostic expertise using mammograms follow a common statistical pattern across subjects. Detailed scrutiny of the images, mediated by microsaccades, is prevalent during the learning phase, but not during the asymptotic phase. Thus, after training, the "gist" of the image may be evident to the expert viewer without the necessity for detailed scrutiny of the image.

References

[1] Hegdé, J. “Role of Statistical Learning in Radiological Diagnosis of Cancer” *MIPS XVI* (2015).
[2] Chen, X., and Hegdé, J. “Learning to break camouflage by learning the background,” *Journal of Vision* (2012).
[3] Hegdé, J., and Arienzo, D. “Neural Substrates of Camouflage-Breaking” *Journal of Vision* (2016).

Acknowledgements

This study was supported by Army Research Office (ARO) grants W911NF-11-1-0105 and W911NF-15-1-0311 and Defense University Research Instrumentation (DURIP) grants W911NF-12-1-0319 and W911NF-14-1-0447 to J.H.

Human, Animal & Computer-Based Medical Image Interpretation: What Can We Learn?

Elizabeth A. Krupinski [1*], Michel de Lange,[2] Victor M. Navarro [3], Edward A. Wasserman [3], Richard M. Levenson [4]

[1] *Department of Radiology & Imaging Sciences, Emory University; ekrupin@emory.edu*
[2] *Department of Statistics University of Amsterdam michel_de_lange@yahoo.co.uk*
[3] *Department of Psychological & Brain Sciences, University of Iowa; victor-navarro@uiowa.edu; ed-wasserman@uiowa.edu*
[4] *Department of Pathology & Laboratory Science, UC Davis; levenson@ucdavis.edu*

Rationale

It is well known that human observers, even the most expert mammographers, can miss lesions in breast images as well as make false positives. There has been much research into the nature and causes of these errors, but there is still much to learn. We have a set of mammographic images that in separate studies were evaluated by mammographers, pigeons (whose visual systems are similar in many ways to those of humans), and a machine learning algorithm. Perhaps by studying these different "observers" we can learn more about the causes of interpretation errors and thereby develop new tools and techniques to obviate them.

Methods

Two sets of mammograms, one with masses and one with microcalcification clusters in half the images, were viewed by the three sets of observers in 3 independent studies: 1) mammographers (faculty and senior residents), 2) pigeons, and 3) a machine learning algorithm. Detection performance was measured for the microcalcification images and discrimination performance (benign vs malignant) for the mass images.

Results

For all three sets of "observers", performance was better with the calcification than for the mass images and in all cases, even with the pigeons, performance was significantly better than chance. In the pigeon and machine learning studies, performance effectively transferred from training to test sets. In the pigeon study there were inter-observer differences similar to those seen in human observer studies. Some of the pigeons readily learned the task and transferred their acquired skills to new test images; while others had difficulty learning the task and generalizing to new images.

Conclusions

Analysis of image interpretation data by "observers" other than experienced trained observers may help us better understand the mechanisms of medical image perception and may prove useful in quality assessment by serving as surrogates in several types of studies. The inter-observer differences observed in the pigeon study parallel what is often observed in radiology residents – many tend to excel in some image interpretation tasks or modalities, but never excel in others.

Quality assessment of ultrasound video for medical tele-assistance

Lucie Lévêque[1] (MSc), Yongqiang Cheng[1] (PhD),
Christine Cavaro-Ménard[2] (PhD), Hantao Liu[3] (PhD)

[1]*School of Engineering and Computer Science, University of Hull, United Kingdom*
[2]*LARIS Laboratory, University of Angers, France*
[3]*School of Computer Science and Informatics, Cardiff University, United Kingdom*

Rationale

In a typical tele-assistance practice, medical video signals are communicated remotely in real time and are therefore vulnerable to distortion due to data compression and transmission. It is highly desirable to understand how tele-assistance practitioners perceive the quality of videos, and consequently to improve the clinical practice.

Methods

A perception experiment was conducted with radiologists assessing quality of various ultrasound videos. Four source videos were extracted from four distinctive ultrasound scans from Angers University Hospital. They were compressed by two different compression schemes (i.e., H.264 and HEVC) at various compression ratios, yielding 32 stimuli including originals. Eight radiologists participated in the experiments, scoring the overall quality of each video.

Results

An ANOVA (Analysis of Variance) was performed by selecting the perceived quality as the dependent variable, the video content and compression as fixed independent variables, and the participants as random independent variable. The results show that there is no significant difference between participants in scoring quality, and that content and compression are statistically significant. For each compression scheme, the perceived quality monotonously increases with the increase of bit rate. Using the same bit rate, HEVC gives better perceived quality than H.264. We also applied two widely recognised objective quality metrics developed for natural images/videos to our new database. The Pearson correlation between the predictions of PSNR and the MOSs is 0.58, and 0.71 in the case of SSIM.

Conclusions

Our results provide insights into how distortion in ultrasound videos affect the quality of experience of radiologists in the practice of tele-assistance. We demonstrate that compression and video content have a significant impact on the perceived quality, and that the objective quality assessment contains plenty of headroom for further improvement.

Automatic Selection Of The Best Despeckle Filter Of Ultrasound Images

Yasser M. K. Omar, Ph.D.
College of Computing and Information Technology Arab Academy for Science Technology and Maritime Transport Cairo, Egypt
dr_yasser_omar@yahoo.com

Rationale

Ultrasound image is considered as widely available medical image. The images are produced by the interference echoes of a transmitted waveform which is used in various medical imaging devices such as x-ray, CT, and, MRI scanners. It is very safe for human. Devices of ultrasonic are frequently used by healthcare professionals. The main applications of ultrasound during this time were to develop SONAR for underwater navigation, communication and to detect other vessels.

Ultrasound image is contained noise called “speckle noise”. Speckle noise in ultrasound images reduces the contrast and quality of resolution. Speckle noise is a multiplicative noise which is difficult to remove compared to additive noise. The speckle noise is converted to additive noise by applying log transformation. Thus, speckle noise can be removed from ultrasound image. The numbers of techniques have been proposed for despeckling noise filter in ultrasound image. The most commonly used despeckle noise filter techniques: linear filter technique, non-linear techniques, diffusion filter technique and wavelet filter technique. Experts face a lot of difficulties for selecting the appropriate technique manually. Although there are different despeckle noise techniques to remove noise but they are not suitable to work with all images.

Methods

This paper contributes in Ultrasound images to help expertise selecting best despeckle noise technique. The paper starts by implementing the main four despeckle techniques. The results are evaluated based on the expertise opinion. Moreover, to tackle the goal of automatic selecting the appropriate technique we extract features such as entropy, homogenous, contrast, mean, variance, energy and, correlation. The extracted features represented as input and the dominant expertise opinion represented as an output were submitted to various machine learning algorithm for training and testing.

Results

The machine learning was used with SVM and Decision tree. As the result, it was concluded that to select the best technique. The best accurate and efficient models were obtained from linear SVM and also were recorded in the testing and validating accuracy rate of 98%.

Conclusions

So the results shows that we can building a classifier system that enables to calculate the best technique for removing specular noise based on the features extracted from the input image and SVM.

Neural correlates of expertise in radiology and surgery: Towards ecological validity in fMRI research

Ellen M. Kok[1] (PhD), Anique B. H. de Bruin[1] (PhD), Ide Heyligers[1,2] (PhD), Andreas Gegenfurtner[3] (PhD), Simon G.F. Robben[4] (MD, PhD), Koos van Geel[1,4] (MD), Diana Dolmans[1] (PhD), Jeroen J.G. van Merriënboer[1] (PhD), Bettina Sorger[5] (PhD)

1. *School of Health Professions Education, Maastricht University, Maastricht*
2. *Zuyderland Medisch Centrum, Heerlen*
3. *Technische Hochschule Deggendorf, Deggendorf*
4. *Department of Radiology, Maastricht University Medical Center, Maastricht*
5. *Department of Cognitive Neuroscience, Maastricht University, Maastricht*

Rationale

Functional Magnetic Resonance Imaging (fMRI) is a non-invasive neuroimaging method to measure brain function. 'Active' brain regions need more oxygen, so changes in blood oxygenation as measured with fMRI indicate alterations in regional brain activation. Thus, fMRI could provide a useful method to reveal differences in brain activation between experts, intermediates and novices in a specific medical domain (e.g., radiology or surgery), which could allow us to refine and extend our theories of visual and motor expertise. FMRI research, however, poses several challenges to expertise research: Traditionally, experiments performed in an MR scanner are highly artificial, and tasks are restricted and repetitive. Further, the possibilities for collecting behavioral responses other than button presses are extremely limited. On the other hand, expertise research stresses the importance of ecologically valid tasks. In two running fMRI studies, aiming at high ecological validity, we investigate brain-activation differences related to different expertise levels in radiology and surgery.

Methods

We investigate expertise differences using fMRI in two real-life tasks: diagnosing chest radiographs, and passive viewing of conducting surgical procedures. For both tasks, we included participants of three expertise levels. In radiology, we look at differences between laypeople (n = 11), beginning residents (n = 10) and senior residents (n = 7). In surgery, we contrast novices, medical students (n = 10), residents (n = 10) and surgeons (n = 10). We use localizer tasks to localize brain regions of interest that were selected based on literature. These brain regions are expected to demonstrate differences in brain-activation related to the different expertise levels during the performance of the ecologically valid tasks. In the radiological experiment, 66 chest radiographs were presented for two seconds each. After presentation, participants were presented with a possible diagnosis and were asked to indicate whether this diagnosis was correct (or not) via button presses. All 66 chest radiographs were subsequently presented again but now for ten seconds each, and participants were again required to indicate the correctness of the diagnosis. For the surgery task, we video-taped three different surgical procedures in orthopedic surgery from the perspective of the surgeon (using a go-pro camera). We selected 60 five-second fragments that showed a single

movement (e.g., stitching a wound). As a control condition, we also video-taped and presented in the scanner 60 fragments of everyday activities (e.g., opening a jar).

Conclusions

We collected data of 58 participants in total. Data analysis is ongoing. In this presentation, we will discuss our methodological approach for acquiring high-quality fMRI data while aiming for high ecological validity, and provide some preliminary results.

Towards Augmented Reality Visualization of Mastectomy Specimens for Breast Reconstruction Surgery

Krista M. Nicklaus[1], Ali Naqvi[1], Mary Catherine Bordes[2], Greg P. Reece[3], Mia K. Markey[1,4]
[1] *Department of Biomedical Engineering, The University of Texas at Austin*
[2] *Department of Behavioral Science, The University of Texas MD Anderson Cancer Center*
[3] *Department of Plastic Surgery, The University of Texas MD Anderson Cancer Center*
[4] *Department of Imaging Physics, The University of Texas MD Anderson Cancer Center*

Rationale

Autologous breast reconstruction often requires multiple procedures in order to achieve an acceptable aesthetic result, increasing risks and costs to the patient. The number of subsequent aesthetic revisions depends on the effectiveness of the initial reconstruction procedure, which relies on the surgeon's ability to plan how they will reform the breast shape. Currently, reconstructive surgeons rely on pre-operative patient photographs and measurements, which do not account for the anatomical changes to the breast during and after mastectomy. The extracted mastectomy specimen can provide information about the post-mastectomy breast, but is unfortunately unavailable to reconstructive surgeons due to the need for prompt histological analysis. Our hypothesis is that providing reconstructive surgeons with an intra-operative stereoscopic image of the mastectomy specimen using augmented reality glasses will enhance their ability to make surgical decisions to improve the aesthetic outcome and reduce the number of subsequent procedures.

Methods

We have previously collected 3D scans of mastectomy specimens with a handheld 3D scanner (Go!Scan 3D Scanner, Creaform, Canada) from 12 patients undergoing mastectomy at The University of Texas MD Anderson Cancer Center. Image processing of the 3D scans was performed with Meshlab. With these preliminary images, we used the UCSF Chimera package to create stereoscopic rotating images of the mastectomy specimens and presented them to reconstructive surgeons at MD Anderson with Epson Moverio BT-200 smart glasses (Epson, Japan) under mock surgery conditions. The Moverio BT-200 smart glasses have a see-through display, stereoscopic viewing, and adequate processing capabilities for displaying images. The surgeons' perception of image quality and system effectiveness was evaluated using the established System Usability Scale and through open-ended interview questions.

Results

We displayed four rotating image movies that display different image rotation sequences, measurements of the specimen, color schemes, and varied specimen sizes to determine what features are optimal for 1) learning from the image and 2) utility in the operating room. The movies are generated automatically by running a Python script through UCSF Chimera. The reconstructive surgeon is ready to shape the reconstructed breast approximately 2-5 hours after the mastectomy is performed. Thus, the minimum time available to obtain the 3D specimen scan, process the image, and generate the movie is 2 hours, so we aimed to implement as much automation as possible. Figure 1 shows a frame of a movie with a high contrast color scheme and specimen measurements. The 3D viewing function of the Moverio smart glasses combines the left and right images for a stereoscopic visualization. Surgeons' concerns with the system include the impact of lighting in the OR on their ability to see the image clearly and hands-free operation of the AR glasses during surgery.

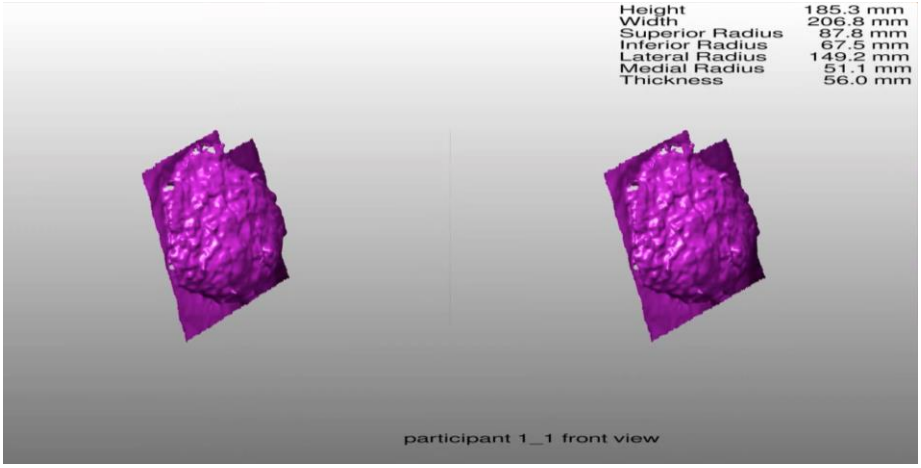


Figure 1: Stereoscopic movie of a rotating mastectomy specimen image with related measurements.

Conclusions

We successfully created a stereoscopic visualization system that allows 3D scans of mastectomy specimens to be viewed on Epson Moverio BT-200 smart glasses. Reconstruction surgeons evaluated the system for effectiveness and future usability in an intra-operative setting. Future work includes optimization of movie schemes and smart glasses function based on surgeon input, improving hands-free operation, and developing the possibility of user-interaction.

A human observer study of multi-lesion detection in digital breast tomosynthesis

Gezheng Wen[1,2], Krista M. Nicklaus[3], Tamara Miner Haygood[2], Mia K. Markey[3,4]

[1]Electrical and Computer Engineering, The University of Texas at Austin;
[2]Diagnostic Radiology, The University of Texas MD Anderson Cancer Center;
[3]Biomedical Engineering, The University of Texas at Austin;
[4]Imaging Physics, The University of Texas MD Anderson Cancer Center

Rationale

Multifocal and multicentric breast cancer (MFMC) is defined as two or more tumor foci within a single breast. A diagnosis of MFMC significantly impacts treatment planning. The long-term objective of our research is to optimize digital breast tomosynthesis (DBT) for detection of MFMC. Our vision is that DBT has the potential to improve the detection of multiple breast lesions and may offer advantages such as fewer false-positive findings, lower cost, and better accessibility. Prior efforts to identify DBT system geometries that optimize image quality only considered unifocal breast cancer scenarios, and DBT system geometries that yield images that are informative for the task of detecting unifocal breast cancer may not necessarily be informative for the task of detecting MFMC. This study focuses specifically on the impact of different DBT system geometries on radiologists' detection performance for MFMC.

Methods

Five radiologists specializing in breast imaging and/or having experience with DBT interpretation were recruited. 3D anthropomorphic computational phantoms with a random number of embedded synthetic breast lesions were scanned by a simulated DBT system to simulate MFMC cases. Four regions of interest (ROIs) are extracted at each possible lesion location to represent the reconstructed slice. The task of the radiologists was to read the DBT images for detecting multiple lesions, and to report the presence or absence of a lesion using an ordinal scale. We evaluated four DBT system geometries to investigate two key factors of image acquisition: narrow-arc geometry vs. wide-arc geometry, and large vs. small projection angular increment. We estimated the area under ROC curve (AUC_{ROC}) and under alternative response ROC curve (AUC_{AFROC}) as the figures of merit for the observer performance in making image-level and location-specific detection decisions, respectively.

Results

For the narrow-arc geometries, the observers achieved higher AUC_{ROC} and AUC_{AFROC} for the MF cases than for the MC cases. However, for the wide-arc geometries, the observers achieved higher AUC_{ROC} and AUC_{AFROC} for the MC cases than for the MF cases. This conflict suggests that the narrow-arc geometry may be more effective for detecting MF lesions while the wide-arc geometry may be more effective for MC lesions. Moreover, for both MF and MC cases, the rank ordering of the DBT geometries by AUC_{ROC} was not the same as that by AUC_{AFROC}. This suggests that the optimal designs of DBT would change if the clinical task of interest changes.

Conclusions

We present a human observer study that investigates DBT system geometries for detecting MFMC lesions. We have shown that the DBT geometries may not be equally efficient for MF cases and MC cases. We have also shown that the optimal geometry of DBT may vary when the task of clinical interest changes.

Understanding noise power spectrum in light of human observer detection in Tomosynthesis

Amareswararao Kavuri[1], Nathaniel R. Fredette[1] and Mini Das[1,2,*]
[1]*Department of Biomedical Engineering,*
[2]*Department of Physics*
University of Houston, Houston, TX
** Email: mdas@uh.edu*

Rationale:

The anatomical noise power spectrum of a mammographic image has been shown to obey a power law relationship by several investigators. This frequency distribution of the noise in the image, which captures the anatomical structure and variability of the breast, has been characterized by the parameter β or the slope of the power law spectrum. Researchers have predicted that lower anatomical structural noise as indicated by a lower value of β would indicate better abnormality detection performance. In our initial studies, we see that β varies with different acquisition parameters in digital breast tomosynthesis (DBT). We propose to investigate how β varies with DBT acquisition parameters and reconstruction methods. Our studies will also shed light to the extent to which β can characterize detectability via comparison with human observer studies.

Methods:

Anatomical variability is characterized via the noise power spectrum in literature as $\boldsymbol{NPS_a = \alpha f^{-\beta}}$, where NPS_a is the anatomical noise power spectrum in the region of the breast dominated by structural noise, f is the spatial frequencies represented and β is the slope of log-log plot of NPS_a. Regions of size 64x64 (17.28x17.28mm) are selected with 50% overlap in the breast region of simulated slices. A Hanning data taper is applied to reduce spectral leakage and then the power spectrum is calculated using a Fourier transform. The radially averaged noise power spectrum is then plotted on a log-log plot. The slope of linear portion of this plot represented by powers in the ranges of 0.15 – 0.7 cycles/mm is computed as the β of the region. An average β is then calculated from all the selected regions of the image.

Results:

We simulated projections using anthropomorphic breast phantoms generated by Bakic et al. at University of Pennsylvania, reconstructed these images and then calculated average β values for sets of images which were acquired and reconstructed using exact same strategies. Multiple strategies were examined in the complete study, which included varying project arcs (between 30° and 90°) and number of projections (between 3 and 51). Our preliminary results show that lower β values do not necessarily indicate signal detectability in DBT images. Studies included considerations of varying breast densities and noise levels as well.

Conclusions:

From our initial findings, it appears that lower β does not always imply better lesion detectability. An alternative use for the parameter β hinges on its attempt to capture anatomical structure of the entire breast. This could be still a valid application in virtual clinical trial like simulation platforms. Due to its dependence on acquisition parameters, the noise power spectra anatomical noise cannot be used as an indicator to predict performance of different modalities such as mammography, DBT and breast CT. Further investigations are underway to examine other aspects related to the image noise power spectrum that may control ultimate detectability.

Mammography to Tomosynthesis: Comparing Two-Dimensional to Three-Dimensional Visual Search in Radiologists and Undergraduates

Stephen H. Adamo[1] Ph.D., Justin M. Ericson[1] Ph.D., Joseph C. Nah[1] MA, Rachel Brem[2] MD, & Stephen R. Mitroff[1] Ph.D.
[1]Department of Psychology, The George Washington University
[2]Department of Radiology, The George Washington University

Rationale:

Radiological techniques for breast cancer detection are currently transitioning from relying primarily on mammography, two-dimensional (2D) image of breast tissue, to being complemented by tomosynthesis, a technique that creates a three-dimensional (3D) image of the breast. With tomosynthesis, a radiologist can search through multiple layers of depth to evaluate the image(s) with greater fidelity. While there is a clear benefit of tomosynthesis with a reduction in false positives (e.g., Durand et al., 2015), it can take significantly longer (e.g., Bernardi et al., 2012) and it is unclear what other factors of a 3D search environment might affect performance.

This project sought to better understand 3D search by evaluating commonly studied factors in 2D search (e.g., response time and search accuracy) in both professional radiologists and non-professional searchers with the ultimate goal of creating a search environment that can inform radiology by using non-radiographs and non-professionals. To accomplish this, we created a 3D search program that emulates tomosynthesis while allowing for flexibility to manipulate factors such as set size (i.e., number of items in the display) and clutter.

Methods:

Data were collected from 29 radiologists from the 2016 Radiological Society of North America conference and 31 undergraduate students from The George Washington University. Observers were asked to search in both 3D and 2D environments. Importantly, the program emulated tomosynthesis, but did so with simplified stimuli that was accessible to both professional and non-professional observers. Observers had 60 seconds per trial. In the 3D images, they could traverse throughout the sphere, moving from one search display "slice" to the next and search for a target "T" amongst distractor "Ls" (see Figure 1B). On half of the trials, the 3D sphere was "compressed" into a 2D image akin to how a breast image can be viewed using tomosynthesis or as a mammogram (Figure 1A). There were 24 trials divided equally between 3D and 2D with targets present on half of the trials in each condition.

Results:

A series of 2x2 ANOVAs revealed a main effect of condition (3D vs 2D)—professionals and non-professionals had significantly more correct rejections in target-absent trials, more hits in target-present trials, and had fewer false alarms (i.e., false positives) in 3D compared to 2D. There were no significant between-subject (professionals vs. non-professionals) or interaction effects. Importantly, observers also took significantly longer to search on target-absent trials in 3D compared to 2D with no significant between-subject or interaction effects.

Conclusions:

The results demonstrated a clear reduction in false alarms (false positives), improvement in hit rates, and improvement in correct rejections when searching in 3D. However, observers took longer. Theoretically, these results suggest a potential speed/accuracy trade-off when searching in 3D compared to 2D. Practically, the results typically found when comparing tomosynthesis to mammography were obtained within computerized 3D and 2D searches with both professionals and non-professionals. This program might provide an easier alternative to running radiologists with real medical images and can help to further identify key advantages and potential pitfalls of tomosynthesis in relation to mammography.

Figure 1

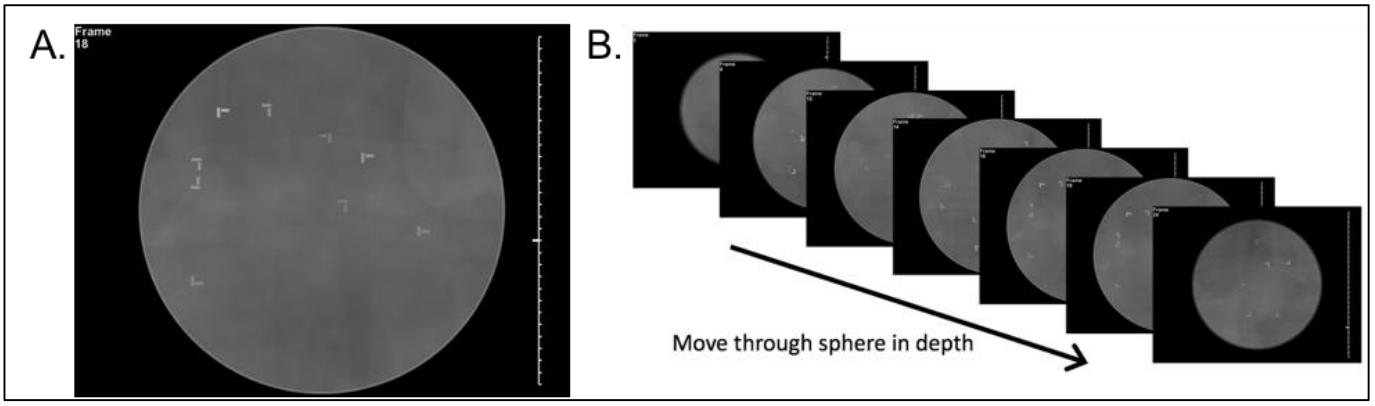


Figure 1. (A) Sample image of a single layer of the search space. A target "T" is on the right. (B) depictions of the 3D nature of the search space and how observers can search from one slice to another in the sphere.

Understanding the impact of local texture features on search and localization in digital breast imaging

William H. Nisbett[1], Amareswararao Kavuri[2], Mini Das[1,2,*]

[1] *Department of Physics*
[2] *Department of Biomedical Engineering*
University of Houston, Houston, TX

[*] *Email: mdas@uh.edu*

RATIONALE:

The notions of image quality and appearance are essential to the assessment of medical imaging techniques and systems. In the field of perception especially, these concepts attract much discussion because of their influence on observer performance. Though texture features extracted from medical images have previously been correlated with risk in various studies, the impacts of texture features on search and localization in these images have yet to be rigorously probed. In this study, we will investigate the effects of both changes in breast properties and changes in acquisition and reconstruction parameters on image texture. Furthermore, by comparing these texture descriptors with changes in human observer LROC with respect to the same parameters, we hope to shed some light on the impact of certain acquisition geometries and reconstruction parameters on detection.

METHODS:

A serial cascade model for image generation was utilized to create some 6,000 simulated DBT images of the phantoms provided by Bakic et al. from the University of Pennsylvania. These images were reconstructed for a variety of DBT and phantom parameters such as angular span, number of projections, and breast density. Thirteen texture features were calculated for randomly selected 100x100 pixel ROIs within the fibroglandular tissue of all images. Additionally, a human observer LROC study was conducted for images of 60^{o} span, a combination of all densities, and varying number of projections.

RESULTS:

Four of the thirteen texture features show strong correlation with the area under the LROC curve (correlation coefficient > 0.75) when both are plotted against the number of projections used in image reconstruction. Moreover, these four texture parameters exhibit significant changes when acquisition, reconstruction, and phantom parameters were altered.

CONCLUSIONS:

Based on the trends of the texture descriptors, there does indeed appear to be some relationship between human observer LROC data and the texture of images used in the studies. Furthermore, it is clear that DBT parameters as well as breast density impact the texture of images significantly.

Cognitive processing differences of experts and novices when correlating anatomy and cross sectional imaging

L R Salkowski, MD MS PhD
Department of Radiology, University of Wisconsin-Madison

RATIONALE

The ability to correlate anatomical knowledge and medical imaging is foundational to radiology. Experts do this well, but we have little understanding about how this occurs. Even more problematic, we don't know how novices assimilate this understanding. It is difficult to teach concepts of anatomy and imaging correlation when we don't know the degree of understanding of our novices.

METHOD AND MATERIALS

Ten radiologist experts (average age 47.4 years with 13.5 average years of experience; 9 males, 1 female) and 11 senior medical student novices (average age 27.4 years; 65 males, 6 females) performed a simulation localizing axial and sagittal computed tomography images within a human simulation torso. This study was IRB approved. Data was collected on image orientation, time, and correctness of localization. Participants were encouraged to do think alouds during the simulation sessions. The transcripts were coded and assessed for emerging themes. The simulation data was assessed with one-way and two-way ANOVAs. Chi-square analysis was preformed on the qualitative action codes. Significance was assessed at $p<0.05$.

RESULTS

In support of the literature, experts are significantly faster at making decisions on medical imaging than novices ($p<0.001$). Quickness is only one factor. When localizing an image in the body, experts rely on organ substructures ($p<0.0001$) whereas novices weigh heavily on size or amount of an organ in the image ($p<0.001$). Experts are more likely to use the correct terminology ($p<0.001$), whereas novices are more likely to misinterpret the anatomy ($p=0.002$) and use non-anatomic descriptive cues (color, blobs, patterns) to describe what they are viewing ($p=0.004$). Experts notice patterns on medical imaging not common to novices. When performing fine-tuning adjustments during a localization, experts isolate a structure with a narrow zone of change ($p<0.001$), compared to novices who use the shape or size of an organ ($p<0.001$) or trial and error methods ($p<0.001$) when performing the same tasks.

CONCLUSION

There are differences in the cognitive processing of experts and novices with respect to meaningful patterns, organized content knowledge and the flexibility of retrieval. Presented are some novice-expert differences in image processing. This study investigated extremes, opening an opportunity to investigate the sequential knowledge of residents, and where educators can help intervene in this learning process.

Training the evaluation of radiographs: Normal-abnormal proportion differentially influences sensitivity and specificity

Koos van Geel[1,2] (MD), Ellen M. Kok[2] (PhD), Abdullah Aldekhayel[1] (BSc), Simon G.F.[1] Robben (MD, PhD), Jeroen J.G. van Merriënboer [2] (PhD)

1. *Department of Radiology, Maastricht University Medical Center, Maastricht*
2. *School of Health Professions Education, Maastricht University, Maastricht*

Rationale

Medical images, such as radiographs, are increasingly requested in everyday clinical practice and therefore medical students need to be trained in evaluation of images. Training in image evaluation usually consists of first a lecture and second a practice phase. It is uncertain if this lecture-first, practice second order is most advantageous for training students to evaluate images. Medical students, despite being novices in x-ray evaluation, already have some knowledge of anatomy and pathophysiology. An order with a practice phase prior to a lecture may enhance learning as students actively use their knowledge prior to passively acquiring expert information. Additionally, training generally concentrates on abnormalities, while x-rays in everyday clinical practice are predominantly normal. The influence of practice first versus lecture first and the normal-abnormal proportion on evaluation of images by medical students is examined.

Methods

103 3rd-year medical students trained chest radiograph (CXR) interpretation by watching a video lecture on basic CXR interpretation followed by 20 practice cases (lecture-first order), or practicing these cases before watching the lecture (practice-first order). After each practice case students were presented the right answer. The proportion of normal-abnormal x-rays (30% vs. 70% normal x-rays during practice phase) was manipulated to make a 2x2 between-subjects design. After their respective training students made a post-test of 20 cases (60% normal) and sensitivity (correctly identified abnormal x-rays / total abnormal x-rays) and specificity (correctly identified normal x-rays / total normal x-rays) were measured.

Results

Mean sensitivity was .97 (*SD* = *.06*) for the lecture-first/30% normal-group; .98 (*.05*) for the lecture-first/70% normal-group; .90 (*.08*) for the practice-first/30% normal-group, and .89 (*.09*) for the practice-first/70% normal-group. On sensitivity, there was no interaction effect ($F_{1,99} = .17$, $p = .68$) nor a main effect of order ($F_{1,99} = .02$, $p = .90$), but a main effect of proportion of normal-abnormal x-rays ($F_{1,99} = 25$, $p < .01$, $\eta_p^2 = .20$) in favor of 30% normal practice was found.

Mean specificity was .57 (.18*)* for the lecture-first/30% normal-group; .74 (*.13*) for the lecture-first/70% normal-group; .51 (*.17*) for the practice-first/30% normal-group, and .67 (*.12*) for the practice-first/70% normal-group. On specificity, there was no interaction effect ($F_{1,99} = .04$, $p = .85$) but main effects of both order ($F_{1,99} = 4.24$, $p = .04$, $\eta_p^2 = .04$), in favor of lecture first, and proportion

of normal-abnormal images ($F_{1,99} = 30.1$, $p < .01$, $\eta_p^2 = .23$), in favor of the 70% normal images practice, were found.

Conclusion

Contrary to our hypothesis, the practice-first order was not superior for sensitivity or specificity. Students may have needed the guidance of a lecture first for an effective practice phase. It should be explored what the effect of adding a third practice phase would be. Furthermore, the results show that proportion normal and abnormal x-rays in practice can differentially influence students` scores on sensitivity and specificity. Increasing the proportion of normal images in training might be useful to align image evaluation training better to the needs of everyday clinical practice.

Do Nigerian Radiographers have potential to interpret chest radiographs?

Ernest U. Ekpo, PhD[1,3]; Nneoyi O. Egbe, PhD[1]; Bassey E. Akpan, BSc (Hons)[2]; Mark F. McEntee, PhD[3]

[1]*Department of Radiography and Radiology, University of Calabar, PMB 1115, Calabar, Nigeria*
[2]*Clinical Applications Unit, GE Healthcare, Victoria Island, Lagos, Nigeria*
[3]*Discipline of Medical Radiation Science, Faculty of Health Science and Brain and Mind Centre, University of Sydney, 75 East Street, Lidcombe, Sydney, NSW 2141, Australia*

Rationale

The high population to radiologist ratio (1:700,000) in Nigeria has negatively impacted radiology service delivery. The dearth of radiologists, increasing demand for radiological services, and pressure from patients for their X-ray reports has led to an all-comers situation where Nigerian radiographers perform X-ray interpretation in private settings. However, their ability to report X-ray has not been explored and requires consideration. This work aims to assess the performance of Nigerian radiographers in chest X-ray interpretation and parameters associated with performance.

Methods

A test-set containing 50 posteroanterior (PA) was used for the study. Fifty-eight (58) self-selected radiographers read the test-set; 23 of these had no pathology (normal) and 27 had features of chest pathology (abnormal). Each abnormal radiograph had one abnormality, and the case mix was a combination of pulmonary and cardiac pathologies. All participants self-reported their age, gender, academic qualification, number of years since qualification, sector of practice (public or private) and employment status and previous training (formal or informal) in X-ray interpretation. The 50 radiographs were presented in a random order, and readers independently reviewed and reported them. No information about the number of normal and abnormal cases and the types of abnormalities in the test-set was disclosed. Receiver operating characteristic (ROC) analysis was used to assess reader performance. Participants were grouped according to the sector of practice (public vs private), age (>32 vs. <32 years), years qualified (>5 vs<.5

years), and previous training in X-ray reporting, and Mann–Whitney U-test was used to compare reader groupings.

Results

A total 51 radiographers completed the reading, and 2,550 readings were made. Readers were aged 26 to 60 years (mean- 534.7 years), with 46% being 32 years or younger. Years of experience ranged from 3 to 20 years (mean-59.4 years), with 29 (56.9%) working in the private setting and 22 in public hospitals. Radiographers' sensitivity ranged from 63.6 [95% CI: 0.522–0.828] to 100 [95% CI: 0.929–1.000] (Mean- 76.9 [95% CI: 0.658–0.864]). Their specificity ranged from 64.3 [95% CI: 0.483–0.796] to 95.7 [95% CI: 0.929–1.000] (Mean- 79.8 (95% CI: 0.658–0.864). The mean false positive rate was 20.2%. Only years of experience as radiographer ($p = 0.005$) and private practice ($p = 0.004$) were positively associated with performance.

Conclusion

Findings are encouraging and demonstrate that even without formal training, this self-selected cohort of Nigerian radiographers can appreciably report chest radiographs. Formal training of radiographers in image interpretation should help in improving radiological service delivery in this region.

Search Pattern Training for Central Line Positioning on Chest Radiography

William F. Auffermann, MD, PhD; Elizabeth A. Krupinski, PhD; Srini Tridandapani, MD, PhD

Department(s) & Institution(s) (For all authors): Department of Radiology and Imaging Sciences, Emory University School of Medicine, 1365 Clifton Road NE, Atlanta GA 30322, USA

Rationale:

Knowledge about the mechanisms of medical image perception have been extensively studied, but only recently used to develop focused perceptual educational tools. Many medical personnel are expected to be able to evaluate a chest radiograph for critical abnormalities before a final interpretation is rendered by a radiologist. One such interpretation task is the evaluation of central lines for appropriate positioning. The main goal of this study is to examine if focused search pattern training improves the ability of a novice to evaluate the position of central lines on chest radiographs.

Methods:

Eighteen healthcare trainees and practitioners were enrolled, 5 were radiology technologists, 13 were nurse practitioner students. Participants were asked to localize the tip of central catheters on chest radiographs, record their confidence in localization, and determine whether or not the line was correctly positioned. The timing of search pattern training varied between control and experimental groups. An attentional control was provided for the group not receiving training. Performance at line positioning relative to training was examined. Specific metrics considered include: fraction of cases with correct tip localization, confidence in tip localization, and fraction of cases correctly identified as normally positioned catheters. Statistical significance was tested using the Wilcoxon rank-sum test. P-values of greater than 0.05 were considered statistically significant.

Results:

Difference in median fraction of correctly localized catheter tips for control and experimental groups were 0.0 and 0.5, $p = 0.5000$ and 0.3805 respectively. Difference in the median confidence during localization for the control and experimental groups were 0.0 and 0.15, $p = 0.8792$ and 0.4355 respectively. Differences in true negative fraction for correct categorization of line positioning (correctly positioned versus malpositioned) for control and experimental groups were -0.0056 and 0.0722, $p = 0.7798$ and 0.2261 respectively.

Conclusions:

The improvement in performance was greater in the experimental group when compared with the control group. However, these differences were not statistically significant at the Type I error level of 0.05. One possible explanation for these results is that the small sample size was not adequate to demonstrate a statistically significant training effect. Further study using a larger sample size may be useful to further examine this question. These results suggest that our knowledge of medical image perception may be useful for developing further educational tools for training in medical image perception and interpretation.

Mixed hybrid search: A model system to study incidental finding errors in radiology.

Jeremy M Wolfe PhD
Department of Surgery, Brigham & Women's Hospital
Departments of Ophthalmology & Radiology, Harvard Medical School

Rationale:

When a radiologist examines an image with one goal in mind (e.g. Does this patient have lung cancer?), s/he is also asked to report any "incidental findings" – findings that might be clinically significant even if they are not the reason that the study was ordered (e.g. a broken rib). When such incidental findings are missed, there are can be negative medical consequences for the patient and negative legal consequences for the physician. If we want to reduce incidental finding errors, it would be useful to have a way to study the process that produces them without needing to use scarce radiologist time. Thus, our goal in this project is to develop a 'model system' that can be used to study incidental findings in non-experts.

Methods:

To this end, we have developed the "mixed hybrid search" task. In standard visual search, observers look for one type of target. In hybrid search, observers look for an instance of any of several specific targets held in memory (Find this rabbit, this truck, and this key). Hybrid search is so-named because it combines visual and memory search. Reaction times (RTs) in hybrid search increase linearly with the visual set size and linearly with the log of the number of targets held in memory. The same pattern is seen with search for categorical targets (e.g. find any cat, bottle, or dessert), though categorical targets produce longer RTs than specific targets. To simulate the incidental finding situation, in the mixed hybrid search paradigm, observers search for any of three specific and three categorical targets. Specific targets are the analog of the radiologist's specific task. Categorical targets are the analog of the incidental findings. They are known to the observer but are less well-defined than the specific targets. Observers memorize the targets for a given block of trials and then search through 300 displays, half of which contain one target. In a second experiment, the categorical targets appear on only 20% of target-present trials, mimicking the fact that incidental findings will be relatively rare.

Results:

When categorical and specific targets are mixed within a block, observers miss more than twice as many categorical targets as they do specific targets. Observers miss fewer

categorical targets if all targets in a block are categorical. Observers miss the fewest targets when all are specific. In Experiment 2, a mixed block with 4X as many specific targets as categorical targets produces a high miss rate for categorical targets (38%, > 7X the rate for the more common specific targets), mimicking the pattern of incidental finding errors in radiology.

Conclusions:

Mixed hybrid search has properties that make it a plausible model system for incidental findings. If further studies confirm that this paradigm captures important aspects of the problem, we can use mixed hybrid search to test interventions that could reduce the incidental error rate in the lab and in the clinic.

Interruptions in Diagnostic Radiology: Is there a Tradeoff between Speed and Accuracy?

Lauren Williams, B.S. and Trafton Drew, Ph.D.

Department of Psychology, University of Utah

Rationale

Radiologists work in highly disruptive, high stakes environments. A recent observational study found that an interruption occurs every 12.1 minutes during regular business hours (Ratwani, Wang, Fong, & Cooper, 2016). Furthermore, the shifts with greater phone-call volume are associated with an increase in the number of discrepancies between readings (Balint, et al., 2014). In our recent work, we used an experimental design to quantify the effects of interruptions on search through chest CT scans with simulated nodules (Williams & Drew, 2017). We found that interruptions led to an increase in search time, but there was no effect on diagnostic accuracy. Eye-tracking measures revealed an increase in refixation rate in the moments following the interruption, which predicted an individual's overall time cost. This suggests that the increase in search time is caused by an inability to remember which regions of the image were searched prior to the interruption. However, despite this impairment, an equal number of abnormalities were detected across both conditions. Given an unlimited search time, observers might have been able to avoid errors by spending more time on the interrupted cases. However, radiologists are often under substantial time constraints and might not have sufficient time to recover from every interruption. The goal of the current study was to investigate the effects of interruptions when there is a finite amount of time available to search each image. We hypothesized that these time constraints would lead to more errors for interrupted cases.

Methods

Novice participants (n=24) searched through 20 chest CT scans for artificial lung nodules. During half of the CT scans, search was interrupted by a series of 10 true or false math equations between 20 and 40 seconds following search onset. Participants were allotted 60 seconds to complete each CT scan and received a 15 second warning before the time was up. The time spent on the math problems did not count toward the allotted search time, and participants were instructed to be as accurate as possible on both tasks.

Results

There were no differences in nodule detection rate between interruption (M = 54.8%, SD = .17) and control (M = 53.5%, SD = .15) trials, t(23) = .51, p = 61. The average number of false alarms

per case did not differ between interruption (M = .31, SD = .71) and control (M = .25, SD = .65) trials, $t(23) = 1.13$, $p = .27$.

Conclusions

Despite the connection between errors and interruptions found in observational studies, we have consistently failed to find this relationship using an experimental design (Williams & Drew, 2017; Aldred, et al., 2016). In the current study, we found that interruptions did not increase error rates even when observers searched under time constraints. Future studies should further investigate the factors that lead to errors in diagnostic radiology. Not all interruptions are equally disruptive; one promising line for future research will be to examine whether interruptions that involve stronger connections to medical image perception may produce the predicted costs in diagnostic accuracy.

Towards automated analysis of nucleolar and centromere shapes in Indirect immunofluorescence images

P. Suhail Parvaze[1], S.S. Suganthi[2] and S. Ramakrishnan[1]

[1] Biomedical Engineering Group, Indian Institute of Technology Madras, Chennai, India.
[2] Tata Elxsi limited, IITM Research Park, Chennai, India
E-mail: suhailsp@gmail.com[1], suganthi.ss@tataelxsi.co.in[2], sramki@iitm.ac.in[1]

Introduction:

Automated analysis of HEp-2 images is considered topical in contemporary research owing to its ability to aid physicians to a greater extent. Delineation of nucleus and the classification of patterns in these images are necessary steps towards computer aided diagnosis. In this work, an attempt has been made to analyze two staining patterns namely, nucleolar and centromere using shape based features and machine learning techniques.

Methods:

598 HEp-2 cell images acquired from the public database are considered for the study. The images are preprocessed using contrast enhancement technique. The nucleolar and centromere patterns are extracted by multiplying with the ground truth binary mask. Conventional geometric, Laplace Beltrami Eigen value and distance based features are derived from the segmented nucleus and two level thresholded images. Finally, the patterns are classified using Multi-Layer Perceptron, Support Vector Machine and Random Forest algorithm.

Results and discussions:

HEp-2 cells are observed to be low in contrast especially in intermediate intensity images. Therefore HEp-2 cell images are contrast enhanced and the nucleus regions are extracted using the ground truth binary mask. Six shape based features are extracted at three levels namely ground truth, first and second level thresholded masks. Total of 18 features are fed to the classifiers in order to differentiate the patterns. The classifier performance is validated using accuracy, precision and recall measures. The classification accuracy of 91.5% is observed using multilayer perceptron whereas random forest and SVM resulted in an accuracy of 90.5% and 87% respectively.

Conclusions:

In this study, centromere and nucleolar patterns of HEp-2 cells are classified using shape based features. After preprocessing, nucleus is segmented and shape based features are extracted. Results show that the extracted features are capable of differentiating centromere and nucleolar patterns with a maximum accuracy of 91.5% using MLP classifier. Thus, it appears that the proposed framework can be used for screening autoimmune diseases automatically and further aid the diagnosis procedure.

Key Words: HEp-2 cells, IIF, Laplace Beltrami, random forest, support vector machine and multilayer perceptron

MRMC Analysis of Mitotic Counts

Brandon D. Gallas, PhD and Weijie Chen, PhD
Division of Imaging, Diagnostics, and Software Reliability, OSEL/CDRH/FDA

Rationale

We wanted to evaluate the survival outcome prognostic ability of mitotic counts reported by pathologists evaluating glass slides stained with pHH3 (phospho-histone H3) compared to those of the standard stain, H&E (Hemotoxylin & Eosin). The pHH3 stain reacts to cells undergoing mitosis; it stains them red. Counting mitoses is expected to be easier for pHH3 as it is a color detection task rather than a challenging morphologic discrimination task. Furthermore, we wanted to compare the counts that come from evaluating the glass slides on the microscope to those from evaluating whole slide images (WSI's) on a digital display.

Methods

We conducted a study with 12 pathologists and 113 patients. The patients were canines diagnosed with oral melanoma. Survival data included date of death by melanoma for 30 patients, death by unknown cause for 27 patients, and last live contact for 10 patients. A pHH3 slide and H&E slide were prepared for each patient and these slides were scanned with an Aperio AT2 scanner to produce corresponding WSI's. The pathologists were veterinary pathologists distributed across four sites. Data collection followed clinical practice in which pathologists count mitoses in 10 consecutive, non-overlapping, high-power fields of view, starting in an area of high mitotic activity. High-power fields of view are typically those resulting from a 40X objective and a 10X eyepiece, corresponding to approximately 0.24-0.26 mm^2 of tissue. Each pathologist determines the fields of view to evaluate. Counting mitoses with WSI's was done on a digital display by creating circular annotations outlining fields of view with areas equivalent to those on the microscope. These annotations were saved to allow us to investigate the overlap of fields of view chosen by each pathologist. Prognostic performance was evaluated in terms of Harrell's C statistic, the area under the receiver operating characteristic curve (AUC), sensitivity, and specificity. To compare counts from glass slides and the microscope to counts from WSI, we used the (pairwise) probability of concordance, percent difference, and the coefficient of variation. Established multi-reader multi-case (MRMC) analysis methods were used to analyze prognostic performance, while new MRMC analysis methods were developed for the other comparisons.

Results

Preliminary findings show that there is a significant amount of within- and between-reader variability, and that different pathologists choose different fields of view. Analyses are ongoing and more results will be presented at the conference.

Conclusions

Selection of fields of view may be a significant source of the variability observed in the data collected. We intend to complement this study with one where the readers read the same fields of view in both modalities.

An approach to classify dermoscopy images using Dynamic Restricted Boltzmann Machine

Punitha N[1] and Ramakrishnan S[2]

[1,2] Biomedical Engineering Group, Department of Applied Mechanics, Indian Institute of Technology Madras. E-mail: npunitha92@gmail.com[1], sramki@iitm.ac.in[2]

Introduction:

Dermoscopy is a noninvasive diagnostic imaging technique which provides 20-70% magnification of skin surface and enables diagnosis of melanoma. However, this process is highly subjective and scarcely reproducible. Hence an automated analysis is required to improve the accuracy of detection. A unique handcrafted feature is not sufficient for precise diagnosis. Therefore, without extracting handcrafted features, machine learning technique can be used to increase the accuracy of classification.

Methods:

In this work, an attempt has been made to analyze dermoscopy images and classify melanoma using Dynamic Restricted Boltzmann Machine (DRBM). The dermoscopy images for analysis are obtained from publically available online PH^2 database which includes 160 benign and 40 malignant 8-bit RGB images. A typical melanoma image is characterized by multiple hypo-pigmented regions with irregular shape and boundary. The proposed framework uses DRBM to identify unknown critical features from the images and classify them. In DRBM the learning rate is adaptive and ten-fold cross validation is used for training and testing the network.

Results and discussions:

The dermoscopy images for analysis are fed as input to DRBM. In the first layer of DRBM, dimensionality reduction takes place where the raw image information is abstracted into 200 features. Visualization of weights from DRBM shows that certain rows have different distribution than other rows which indicates that some critical features are extracted from the original inputs. The DRBM network is trained for 300 epochs and the average reconstruction error is calculated at the end of each epoch. It is observed that the error gradually reduces and saturates after 200 epochs. It is also seen that the weights in the output nodes are distinct for benign and malignant classes which form the basis of the classification.

Conclusions:

In this study, automatic classification of melanoma is performed using DRBM. The learning rate is adapted to identify critical features from the dermoscopy images. It is found that DRBM is able to distinguish between benign and malignant classes effectively with accuracy of 85.5%. Hence it appears that the proposed framework can be used in the automatic classification of melanoma.

Key Words: Dermoscopy, Melanoma, Dynamic Restricted Boltzmann Machine

Estimated Templates for Forced-Localization Tasks in Ramp-Spectrum Noise

Craig K. Abbey[a], PhD, Frank W. Samuelson[b], PhD, Rongping Zeng[b], PhD, John M. Boone[c], PhD, Miguel P. Eckstein[a], PhD, and Kyle Myers[b], PhD.

[a]Department of Psychological and Brain Sciences, University of California Santa Barbara
[b]Division of Imaging Diagnostics and Software Reliability, United States Food and Drug Administration
[c]Departments of Radiology and Biomedical Engineering, University of California Davis

Rationale: In this study we examine localization performance in Gaussian random textures that simulate limiting effects in computed tomography (CT). These include blur from the system transfer function, background variability from normal anatomical structures, ramp-spectrum acquisition noise, and apodization used to control noise in the images. Understanding the effects of these components on how observers perform visual tasks in the images is important for optimizing tomographic imaging systems for best task performance and for validating such improvements through the use of model observers.

The classification image technique we use directly estimates the weighting function used by observers for forced localization tasks under the assumption of a scanning linear template. We are particularly interested in the effect of apodization on observer templates since this represents a point of control in the imaging process.

Methods: In this study we use the classification image approach to estimate scanning linear templates used by human observers in tasks that have different signal sizes, different amounts of background variability, and different levels of apodization. We also compute observer efficiency with respect to the ideal observer.

The classification image methodology uses noise fields from the incorrect localizations to build an estimate of the weights used by the observer to perform the task. The basic idea is that incorrect localizations occur in regions of the image where the noise field matches the weighting profile, thereby eliciting a strong internal response.

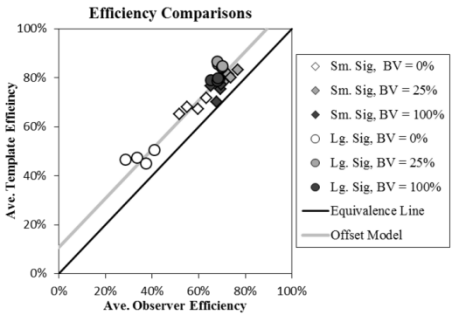


Fig. 1. Scatterplot of observer efficiency and template efficiency. The average efficiency of the observers in each task are plotted with the average efficiency to the estimated templates. The data are reasonably well fit (r^2 = 90.5%) by an offset of 10.8% in efficiency.

Results: Average observer efficiency varies substantially across the different conditions from 28% to 82%. The estimated templates offer some direct mechanisms for explaining this variability. As shown in Figure 1, when the templates are used as a scanning template model, they explain about 90% of the variability in the average observer efficiency data.

Conclusions: In these studies, the classification images are a useful way to characterize and investigate human observer performance. The information we find from this study may be used to construct model observers that utilize image information in ways that are similar to humans.

Observer Effects in 3D Search from Classification Images

Craig K. Abbey, PhD, Miguel A. Lago, PhD, and Miguel P. Eckstein, PhD.
Department of Psychological and Brain Sciences, University of California Santa Barbara

Rationale: In this study we examine search performance for 3D localization tasks in Gaussian random textures in which subjects are able to freely scroll through the image as part of their search for the target. We investigate two target sizes, corresponding to 1mm and 4mm diameter spheres that have been blurred with a system MTF. Targets are embedded in two different noise textures (white noise and $1/f^3$ power-law) for a total of four conditions.

The classification image technique directly estimates the weighting function used by observers for this task, and allows us to probe the depth direction, which is not directly views in our display, relative to the lateral directions witch are viewed. We are particularly interested in whether subjects can efficiently integrate across multiple slices in depth as part of performing the localization task.

Methods: We have adapted the classification image technique to 3D search tasks. Our particular task is a 3D forced localization task, in which a subject searches a 3D volume, and indicates the location of a target signal. Subjects know that only one target is always present in an image at an unknown location. They make a single localization response indicating the position of the target. The image display we use allows subjects to freely scroll through the volumetric image, and a localization response is made through a mouse-click on the image. Localization responses are considered correct if they are close to the target center (within 6 voxels).

The classification image methodology uses noise fields from the incorrect localizations to build an estimate of the weights used by the observer to perform the task. The basic idea is that incorrect localizations occur in regions of the image where the noise field matches the weighting profile, thereby eliciting a strong internal response.

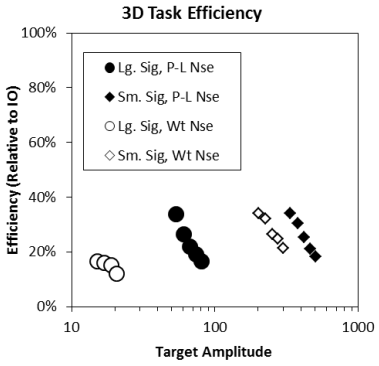


Figure 1. Localization Efficiency Plots. The plots show relatively low statistical efficiency compared to 2D localization tasks in similar backgrounds, suggesting inefficient search in 3D images.

Results: Our results consist of reporting on observer performance through statistical efficiency with respect to the ideal observer, and showing the weighting schemes estimated by the classification images. An example of subject efficiency in the 3D task is plotted in Figure 1

Conclusions: At this stage the main conclusion is that the classification image technique can be used to investigate the spatial weighing – including weighting across slices – used by observers in 3D free-search tasks. At the meeting we will describe more fully the 3D classification images derived from our experiments.

Peripheral vision implication in the search and recognition of low contrast hepatic metastasis in abdominal CT scans: preliminary study with an eye-tracker

Alexandre BA[1], MSc., Sabine SCHMIDT[2], M.D., Francis R. VERDUN[1], PhD, François BOCHUD[1], PhD
[1]*Institute of Radiation Physics, Lausanne University Hospital, Lausanne, Switzerland*
[2]*Department of Radiology, Lausanne University Hospital, Lausanne, Switzerland*

Rationale

While many imaging modalities are three-dimensional, few studies have examined scrolling in volumetric images through eye-tracking experiments.
CT examinations convey large amount of data and prohibits radiologists to closely scrutinize all anatomical regions with their high-resolution fovea. Therefore, radiologists process CT volumetric images at least partly with their peripheral vision.
It is highly probable that peripheral vision plays an important role when radiologists explore CT examinations composed of hundreds of slices and under time constraints. However, few is known about its implication, for example, in the search of hepatic metastases.
Peripheral vision will be characterized in terms of eccentricity, the distance between the center of the fovea and a given point on the field of vision in angular units. This preliminary study aims at answering the following question: What is the eccentricities' range of radiologists' saccades when they search for low contrast hepatic metastases in volumetric CT scans?

Methods

We designed an experiment which tracks the radiologists' visual fixations and saccades in multiple CT slices. We instructed the readers to perform a free search of multiple metastases and estimated their diagnostic performance.
Regarding the peripheral visual characterization, we estimated the range of eccentricities used during the search and the recognition processes.
Stimulus material is composed of abdominal CT scans from our local database with synthetic signals mimicking hepatic metastases.

Results

The experiment is still in progress and the results will be presented during MIPS conference.
The measured distribution of eccentricities will give a quantitative assessment of radiologists peripheral processing in CT volumetric images. We also expect that the distribution of

eccentricities varies according to radiologists' search strategy in volumetric dataset (scanning or drilling).

Conclusion

These results will help to develop model observer based image quality metrics, to better predict human observer classification performance in volumetric images. Next experiment will investigate similar task and will involve more radiologists and additional parameters estimation like accuracy and scrolling patterns.

Parameter Selection for Linear Iterative Image Reconstruction in Breast Tomosynthesis with the Non-prewhitening and Hotelling Observers

Sean D. Rose (BS), Ingrid Reiser (PhD), Emil Y. Sidky (PhD), and Xiaochuan Pan (PhD)
The University of Chicago Dept. of Radiology MC-2026, 5841 S. Maryland Avenue, Chicago IL, 60637.

Rationale

Iterative image reconstruction for digital breast tomosynthesis (DBT) involves a variety of parameter and implementation choices including voxel size, voxel aspect ratio, and regularization strength. Exploration of the corresponding parameter spaces is warranted for every algorithm, task, and system design under consideration. Efficiently computable simulation-based image quality metrics are needed to facilitate this task.

The purpose of this work is the development and comparison of two task-based image quality metrics for assessing the effect of regularization strength on microcalcification detectability in DBT reconstruction.

Methods

Two task-based image quality metrics are investigated: a region-of-interest (ROI) Hotelling observer for a signal-known-exactly/background-known-exactly (SKE/BKE) detection task and an ROI non-prewhitening (NPW) observer. The noise model is additive Gaussian in the sinogram domain with mean equal to variance, thus approximating a Poisson distribution. A simulation study is performed with the two metrics in which regularization strength is varied for Tikhonov penalized least-squares reconstruction (PLS), for which the reconstruction optimization problem is

$$\underset{x}{\operatorname{argmin}} \ \|Ax - b\|^2 + (\lambda\|A\|)^2\|x\|^2$$

where A is the linear forward model, b is the sinogram data, and x is an image estimate. The metrics are applied to the task of microcalcification detection, which is modeled using a 0.32mm diameter high-contrast Gaussian signal. The metrics are calculated in closed form, as opposed to with estimates using noise realizations, to facilitate efficient investigation of parameter spaces involved in reconstruction. Trends in the ROI-HO and ROI-NPW metrics are compared with 3D reconstructions from ACR mammography accreditation phantom data acquired with a Hologic Selenia Dimensions DBT system.

Results

The efficiencies — squared ratios of signal-to-noise ratio (SNR) in the image domain to SNR of the HO in the data domain — of the Hotelling and non-prewhitening observers are shown as a function of regularization strength in the top panel of Fig. 1. Real data ACR phantom reconstructions are shown in the bottom panel. A back projection reconstruction is included for reference. We note that the PLS solution limits to the back-projection image, up to scale, as $\lambda \to \infty$.

The efficiency of both observers tends to increase with increasing regularization strength until a point of saturation. The value of λ at which equal efficiency is achieved by the ROI-NPW can be up to 3 times larger than the corresponding value for the ROI-HO. The ROI-NPW and ROI-HO observer efficiencies both saturate at a value of 0.99 achieving close to the maximum attainable value of 1.0. In the real data reconstructions, the reconstructed specks appear more conspicuous as the noise level is reduced by increasing regularization strength.

Conclusions

The efficiency curves for both the ROI-HO and ROI-NPW observer suggest that information relevant to task performance is better preserved with increasing regularization but also suggest a point at which increasing regularization yields diminishing returns. The ROI-HO outperforms the ROI-NPW at all regularization strengths, suggesting prewhitening does impact performance of the investigated task in DBT. The trend of increasing conspicuity with increasing regularization in the ACR data reconstructions appears to coincide with the ROI-HO and NPW-HO efficiency trends.

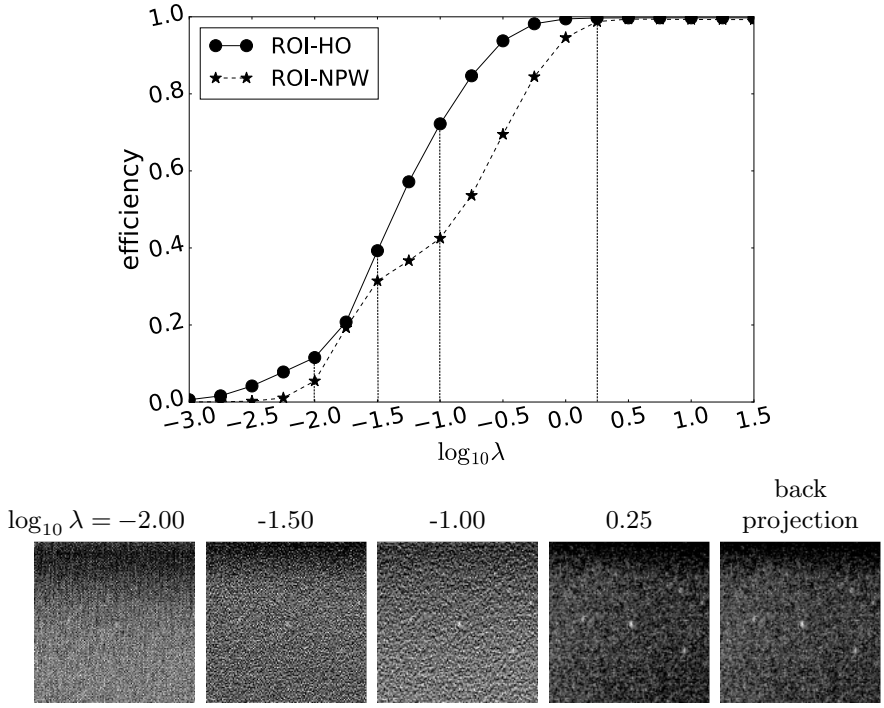


Fig. 1: Top: ROI-HO and ROI-NPW for detection of a 0.32mm calcification as a function of regularization strength in PLS reconstruction. Vertical lines mark regularization strengths used in phantom reconstructions. Bottom: PLS reconstructions of 0.32mm specks of ACR phantom at different regularization strengths. Display windows were chosen subjectively to maximize visibility of specks.

Foveated model observer predicts dissociation of signal detectability across 2D and 3D images

Miguel A. Lago, Craig K. Abbey, Miguel P. Eckstein
Department of Psychological & Brain Sciences, University of California Santa Barbara, Santa Barbara, CA. 93106, USA

RATIONALE

Model observers have been successful at predicting human observer performance for detection and search for signals in 2D noisy images. We argue that search in 3D images represents a paradigm shift in medical imaging because radiologists do not typically exhaustively scrutinize all regions of interest with the high-resolution fovea. Instead, for the detection of signals, observers must rely on peripheral retinal regions, and process the visual information with reduced spatial detail. We hypothesize that the peripheral processing can have important influences in the search of small signals in 3D images. These influences are not captured by current model observers or by the ideal observer for search. We propose foveated model observers that can correctly predict the search of small signals in 3D search.

METHODS

A 3D synthetic noisy background was generated using statistics similar to the x-ray mammograms (noise power spectrum = $1/f^{2.8}$). A mass and microcalcification-like signal were embedded in the 3D background. Human observers were instructed to find the signals in the 3D volume. In a separate condition, observers searched for the signals in a 2D slice. The signals were present with a 50% of probability. For the 2D case, we presented the central slice in which the signal appears. We quantified accuracy (yes/no task) in detecting each of the two signals in both 2D and 3D images. We implemented standard 2D/3D detection model observers including the Hotelling Observer (HOT), the Channleized Hotelling (CHO), the Non-Prewhitening model (NPW) and the NPW with eye filter (NPWE). In addition, we implemented an ideal observer for search. Finally, we implemented new foveated model observers (Foveated-Channelized Hotelling and Foveated NPWE) that take into account the varying spatial resolution across the visual field. Performance predictions were obtained for all models for the four experimental conditions.

RESULTS

Results showed that the small microcalcification-like signal is more highly detectable than a larger mass-like signal in 2D search, but its detectability largely decreases (relative to the larger signal) in the 3D search task. All standard detection model observers (HOT, CHO, NPW, NPWE) as well as the ideal observer for search did not predict the drastic decrease in microcalcification detectability in 3D search. The foveated model observers that take into account the varying resolution processing could predict the human experimental results.

CONCLUSION

The interaction of the properties of the visual periphery, the spatial frequency content of the signal and the observer search patterns have important influences on search in 3D images. Our findings show that these influences cannot be captured by current detection model observers nor by an ideal observer for search. In contrast, a new family of foveated model observers that account for the inhomogeneous visual processing across the retina might be important for assessment of medical image quality in 3D images.

The Role of Prewhitening in Visual-Search Models of Human Observers

Howard C. Gifford, Ph.D.[1]*, Kheya Banerjee[1], Zohreh Karbaschi[1] and Mini Das, Ph.D.[1,2]
[1]*Department of Biomedical Engineering, University of Houston;*
[2]*Department of Physics, University of Houston;*
**Corresponding author: hgifford@uh.edu*

Rationale

Mathematical observers use prewhitening to decorrelate statistical image noise during the target-detection process. That human observers can prewhiten has been indicated by past comparison studies with channelized Hotelling (CH) observers. Largely based on known-target detection at a fixed location, these studies have led to human-observer models in which channel prewhitening is degraded by added internal noise. A visual-search (VS) observer with intrinsic uncertainties may provide a more precise assessment of humans' prewhitening abilities.

Methods

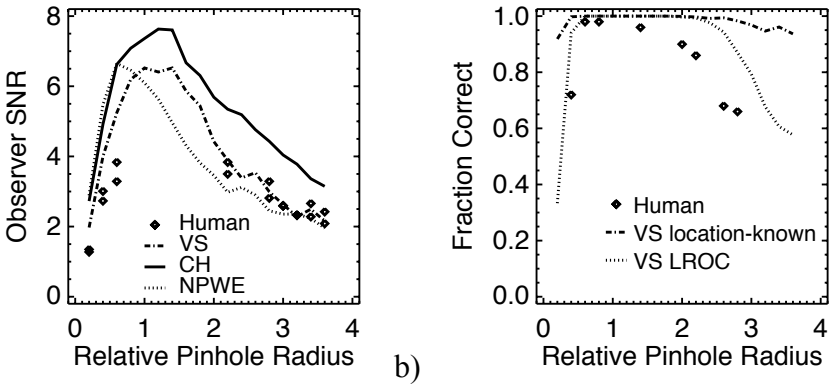
Our VS observer applied a feature-based linear discriminant solely at candidate locations obtained from an initial image search. The observer was tested against humans in localization ROC (LROC) and location-known two-alternative forced-choice (2AFC) studies. These studies used images created with a statistical lumpy background model [from Rolland and Barrett] which simulated single-pinhole planar imaging of 2D phantoms with or without a Gaussian target. Pinhole size was the study variable for this abstract. With 2AFC images, the VS observer sought the maximally suspicious candidate within a search radius about the fixed location. The 2AFC study also included CH and NPWE (nonprewhitening + eye filter) observers that only analyzed the target center location. The CH and VS observers both used Abbey's sparse set of three difference-of-Gaussian channels/features. The VS observer applied different levels of prewhitening and performance-based adaptive feature selection (AFS).

Results

As shown in Fig. 1, the VS observer with full prewhitening gave close agreement with the 2AFC human results but overestimated human LROC performance. With partial prewhitening (or standardization), the VS observer gave close 2AFC agreement but underestimated human LROC performance. Standardization and AFS together produced good agreement in both studies.

Conclusion

Prewhitening and AFS may offer complementary means of specifying important features for a task. This initial work is being augmented with studies based on other variables such as background lumpiness, target geometry and diagnostic task.



a) b)

Fig.1: Observer performance as a function of pinhole radius. a) SNRs obtained by transforming fractions correct from the 2AFC study. Note that two data points for the human observers [at pinhole radii 0.6 and 1.2] mapped to infinity. b) LROC fractions correct for human and VS observers [VS 2AFC performance also shown]. In both studies, the VS observer applied full channel prewhitening.

Consultation and Citation Rates for Older Imaging Studies and Documents in Radiology

Tamara Miner Haygood, PhD, MD [1]; Barry Mullins, MD[1]; Jia Sun, PhD[2]; Behrang Amini, MD PhD[1]; Priya Bhosale, MD[1]; Hyunseon C. Kang, MD, PhD[1], Tara Sagebiel, MD[1]; Bilal Mujtaba, MD[1]

[1]Department of Diagnostic Radiology, UT M.D. Anderson Cancer Center, Houston, Texas, USA;
[2]Department of Biostatistics, UT M.D. Anderson Cancer Center, Houston, Texas, USA;
Tamara.Haygood@mdanderson.org

Rationale

We noticed that in quality assurance conferences the consensus conclusion was often that the mistake could have been avoided if more prior images or documents had been consulted. It was assumed that anything that was not specifically cited in the report had not been consulted. At the same time, the first author noted while reading imaging studies that she did not always cite in the report older studies or documents that she consulted. Therefore, was it safe to assume that an image or document that is not cited was also not consulted? It is this question that this investigation addresses.

Methods

In this IRB approved study, one observer watched the board-certified radiologists while they interpreted imaging studies and issued reports. He recorded what type of study was being interpreted (either computed tomography [CT], magnetic resonance imaging [MRI], or conventional radiography [x-ray]). He also recorded the number and type of prior imaging studies and documents that were consulted during the interpretation. These observations were then compared with the signed report to determine how many of the consulted imaging studies and documents were cited.

Results

Five board-certified radiologists issued 62 reports. Of the studies reported upon, 44 were CTs, 7 were MRIs, and 11 were radiographs. While issuing these 62 reports, the radiologists consulted 198 previous imaging studies and 285 documents ($p=0.0017$).

Of the 198 previous imaging studies that the radiologists consulted, 116 (58.6%) were cited in a report. Of the 285 documents consulted, 3 (1.1%) were cited in a report. This difference was statistically significant. ($p<0.0001$).

The interpreting radiologists consulted a variety of documents. Many were radiology reports, but the majority were not. For example, of the documents consulted in interpretation of the included CT scans, 60.5% were something other than a radiology report.

There was not a 1-to-1 correlation between numbers of imaging studies and numbers of radiology reports consulted. When CTs were being interpreted, the radiologists consulted 148 prior imaging studies but 92 reports. When MRIs were being interpreted, they consulted 27 prior imaging studies but 10 reports, and when conventional radiographs were being interpreted, they consulted 23 prior imaging studies and only 2

radiology reports. Similar results were seen when looking at the behavior of individual radiologists except that one radiologist consulted slightly more radiology reports (42) than imaging studies (39).

Radiologists tended to consult the same type of imaging study as that which they were interpreting. When CTs and MRIs were being interpreted, this relationship was statistically significant. As an exception to this pattern, when interpreting x-rays the radiologists consulted a different type of imaging study rather than another x-ray during 9 interpretations and consulted an x-ray in 8 interpretations.

Conclusions

It cannot be safely assumed that an older radiologic image or medical document was not consulted during radiologic interpretation merely because it is not cited in the report. Radiologists often consult more old studies than they cite, and they do not cite the vast majority of prior documents that they consult.

Musculoskeletal Discomfort in Radiologists

Elizabeth A. Krupinski [1], Rebecca L. Seidel, MD[1]

[1] *Department of Radiology & Imaging Sciences, Emory University; ekrupin@emory.edu, rseidel@emory.edu*

Rationale

There has been increasing concern in recent years that in the PACS environment radiologists are spending more time sitting at their workstations engaged in repetitive tasks (e.g., scrolling through images), and this can lead to musculoskeletal (MSK) injuries, fatigue and poor health outcomes. The goal of this study was to assess the prevalence of musculoskeletal discomfort, its intensity and the degree to which it interferes with their work using a validated survey tool.

Methods

In this IRB approved study, the Cornell Musculoskeletal Discomfort Questionnaire was distributed electronically to all members (faculty, fellows, residents) of our large radiology department. Additional demographic questions were included. Reponses were anonymous unless they opted to receive a $5 gift card.

Results

There was a 33% response rate (n = 99), 39% female and average age of 36.94 (range 26-61). 80% reported spending more than 7 hours/day at their workstation with 52% spending 100% of the time seated. Females were significantly more likely to report discomfort in the right shoulder, left shoulder, and left forearm; those spending > 7 hours in the right shoulder; board certified > 10 years in left upper and forearm; and those > 90% seated in left shoulder and upper back. In terms of degree of discomfort (slightly, moderately, very uncomfortable), females were significantly more likely than males to report neck, lower back and hip/buttocks pain as moderate/very uncomfortable. With respect to discomfort interfering with work, 53% of those with neck, 41% with low back, and 40% with upper back pain said it at least slightly interfered with work. 11% with right wrist pain said it interfered substantially. Those board certified more than 10 years were more likely to report neck pain as interfering as were older radiologists.

Conclusions

Results confirmed previously published findings of MSK pain among radiologists, and added new insight symptom frequency and impact on work. Statistically significant demographic factors included gender, age, years in practice, time spent seated and hours at the workstation. These differences may be due to furniture design that favors a male body habitus or differences in position and posture between genders. Based on these results recommendations regarding improved workplace design can be implemented.

Single display versus dual displays: A cognitive modeling perspective

Hanshu Zhang[1], Joseph W. Houpt, Ph.D.[1]

[1] *Department of Psychology, Wright State University; zhang.180@wright.edu, joseph.houpt@wright.edu*

Rationale

Radiologists often work with multiple display workstations. However, few studies have assessed radiologists' performance with different configurations for multiple displays. We briefly summarize studies we ran comparing performance with multi-sensor imagery in a single fused image display to side-by-side displays. Our assessment is based on a mathematical cognitive modeling framework, which we propose as a method assessing radiologists' performance in future studies.

Methods

In the side-by-side display conditions, observers could strategically choose to process the images either sequentially or simultaneously. We examined their strategies using Systems Factorial Technology (hereafter, SFT). SFT is a framework that analyzes information processing from several perspective: architecture (serial/parallel),workload capacity (limited/unlimited/super), stopping rule (self-determination/exhaustively), stochastic dependence(dependent/independent). Under appropriate expermental conditions, interaction contrasts of the mean (MIC) and survivor function (SIC) of response times can be applied to measure architecture and stopping rule. The capacity coefficient measures workload capacity, that is, it indicates the efficiency of observers processing multiple sources of information relative to processing them in isolation. The efficiency comparison is between the predicted performance assuming unlimited-capacity, independent and parallel (UCIP) processing and the actual performance. We applied SFT to image processing covering images with different complexities: Landolt C, direction (left/right) decision, and a more applied weapon/non-weapon discrimination. Each stimulus was imaged with a long-ware infrared (LWIR) sensor and a standard visible-spectrum sensitve camera. Any time multiple source images were presented side-by-side, the information was redundant, meaning the observer could respond as soon as the discrimination was made using either source. In the single display, images were fused using Laplacian Pyramid Transform as a combination of information from both sensors.

Results

All observers were limited capacity indicating they were less efficient with two image types than with single-sensor images. Across observers, there was evidence of different strategies, inicated by differences in architecture and stopping rule. Based on our results, for visible and LWIR images in these tasks, we suggest side-by-side display is a better choice.

Conclusions

Given our success with assessing performance in that domain, we believe SFT is a promising tool for future assessemnt of radiologists' workstation configurations.

Australian Breast Reader Assessment Strategy on mammographic improves radiologists' test reading performance

Wasfi I Suleiman PhD, Mohammad A Rawashdeh PhD, Sarah J Lewis PhD, Mark F McEntee PhD, Warwick Lee MD, Kriscia Tapia, and Patrick C Brennan PhD

Background

Error and variability in mammography interpretation are frequently reported, however, factors responsible for these are unclear. It is important to explore parameters that impact upon performance as well as ways of improving performance and reducing inter-reader variability. To provide feedback on performance, and explore parameter to reduce error in mammography interpretation, Breast Reader Assessment Strategy (BREAST) was established in Australia in 2009. BREAST was designed to complement existing BreastScreen Australia quality assurance and quality improvement activities. This work aims to assess whether radiologists who regularly undertake the Breast Reader Assessment Strategy (BREAST) demonstrate improvement in performance over time.

Materials and Methods

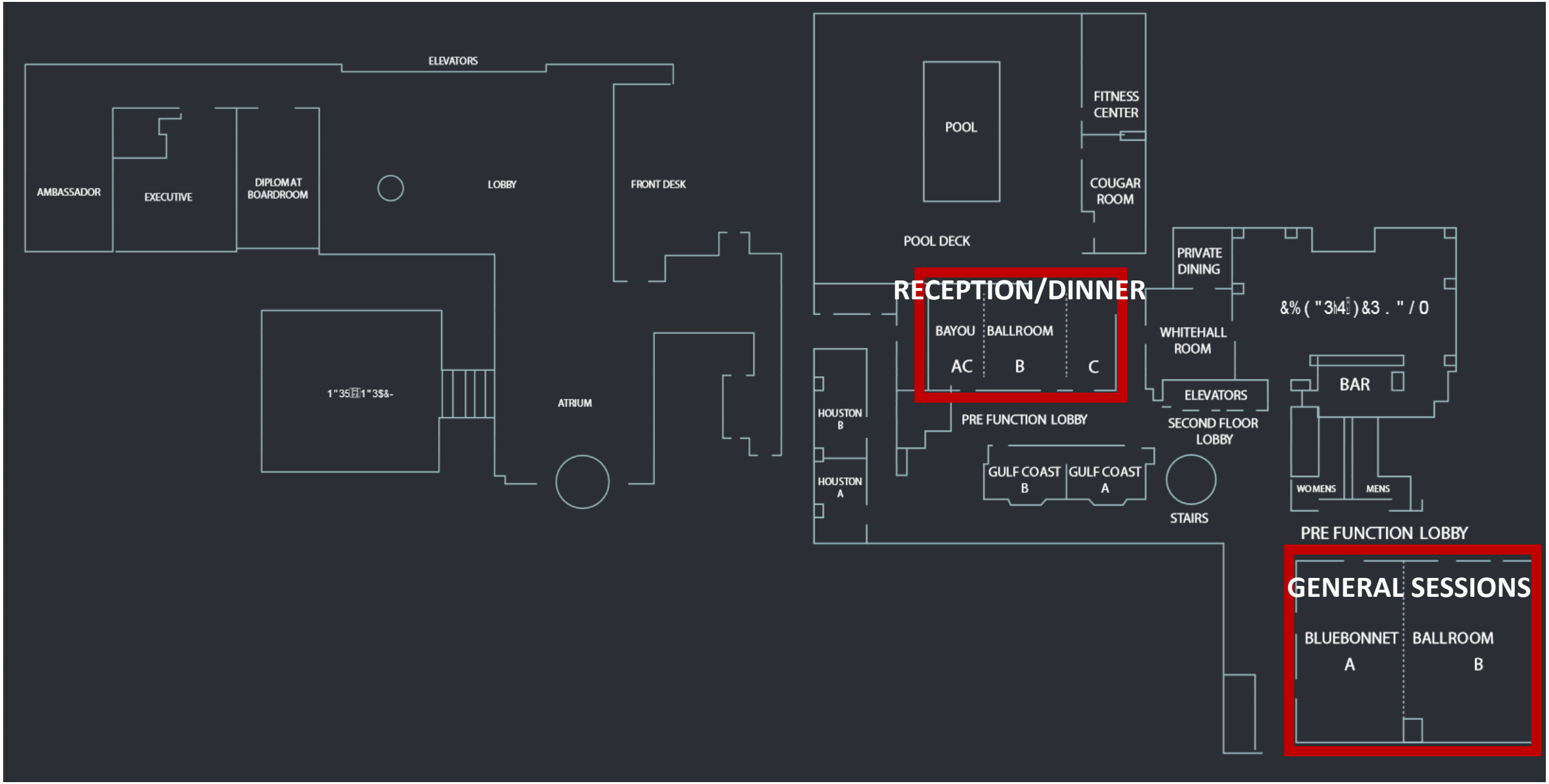
Fourteen Royal Australian and New Zealand College of Radiology (RANZCR) radiologists independently assessed a year-specific BREAST mammographic test-set in each of the years, 2011, 2012 and 2013. The mean sensitivity, specificity, location sensitivity, JAFROC FOM and inferred ROC AUC were calculated and compared.

Results

Significant increases were noted in mean sensitivity ($p = 0.01$), specificity ($p = 0.01$), location sensitivity ($p = 0.001$), JAFROC FOM ($p = 0.001$) and ROC AUC ($p = 0.001$) between 2011 and 2012. There were also increases in mean sensitivity ($p = 0.002$), specificity ($p = 0.001$), location sensitivity ($p = 0.001$), JAFROC FOM ($p = 0.001$) and ROC AUC ($p = 0.001$) between 2011 and 2013.

Conclusion

These findings demonstrate that regardless of experience, radiologists who undertake the BREAST programme demonstrate significant improvements in test-set performance during a 3-year period. BREAST show great promise and demonstrates that carefully constructed test-sets for education can improve lesion detection with Digital Mammography in test-set conditions.


ELEVATORS
AMBASSADOR
EXECUTIVE
DIPLOMAT BOARDROOM
LOBBY
FRONT DESK
1"35 1"3$&-
ATRIUM
POOL
FITNESS CENTER
COUGAR ROOM
POOL DECK
RECEPTION/DINNER
BAYOU
BALLROOM
AC
B
C
PRIVATE DINING
WHITEHALL ROOM
&% ("3 4) &3 . " / 0
BAR
ELEVATORS
SECOND FLOOR LOBBY
HOUSTON B
HOUSTON A
PRE FUNCTION LOBBY
GULF COAST B
GULF COAST A
STAIRS
WOMENS
MENS
PRE FUNCTION LOBBY
GENERAL SESSIONS
BLUEBONNET A
BALLROOM B

Journal of Medical Imaging

Call for Papers

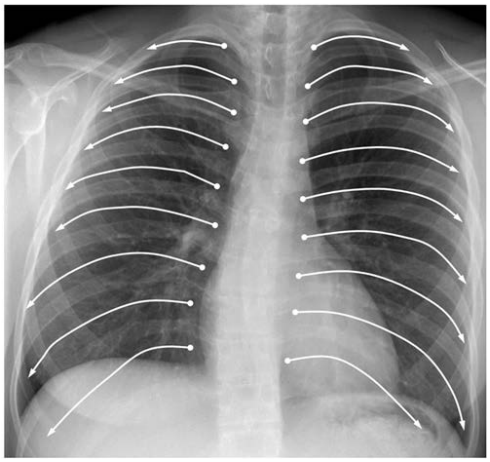
Published by SPIE

Medical Image Perception and Observer Performance

Guest Editor:

Elizabeth A. Krupinski, PhD, Emory University

Medical image perception research measures the human observer's ability to perform specific diagnostic tasks using real or simulated medical images and compares the observer performance with predictions from quantitative models.



From W. F. Auffermann et al, "Teaching search patterns to medical trainees in an educational laboratory to improve perception of pulmonary nodules," *J. Med. Imag.* **3**(1), 011006 (2015).

The theme of this JMI special section is image perception and observer performance research on detection and discrimination of abnormalities; cognition, psychophysics and behavior; perception errors; visual search patterns; human and ideal observer models; computer-based perception; impact of display and ergonomics; image processing; and assessment methods, metrics, and statistics. Although radiology imaging is a key focus, we are particularly interested in articles that deal with other medical imaging specialty applications such as pathology, ophthalmology, dermatology, and telemedicine.

This special section is open to everyone, and but especially encourages relevant submissions from the Medical Image Perception Conference (MIPS XVII). The Medical Image Perception Conference is a biennial conference dedicated to bringing together people interested in human and computer perception of medical image information and related subjects, such as detection and discrimination of abnormalities, cognitive and psychophysical processes, perception errors, search patterns, human and ideal observer models, computer-based perception (CAD and CADx), impact of display and ergonomic factors on image perception and performance, role of image processing on image perception and performance, and assessment methodologies.

For more information on submission, please see the journal web site at http://spie.org/JMIauthorinfo. Please indicate in your cover letter that the submission is for this special section. All submissions will be peer-reviewed. Peer review will commence immediately upon manuscript submission, with a goal of making a first decision within 6 weeks of manuscript submission. Special sections are opened online once a minimum of four papers have been accepted. Each paper is published as soon as the copyedited and typeset proofs are approved by the author.

Manuscripts due 15 December 2017.

For information on how to prepare a manuscript for JMI, please visit our website www.spie.org/JMIauthorinfo.

SPIE.

MIPS XVII SURVEY

1) Overall, how would you rate the scientific quality of the presentations?

Excellent Very Good Good Fair Poor

2) How likely are the presented talks going to impact your research?

Very Likely Likely Neutral Not Likely Very Unlikely N/A

3) How likely are the presented talks going to impact your clinical practice?

Very Likely Likely Neutral Not Likely Very Unlikely N/A

4) How would you rate the diversity of the presentation topics?

Not Diverse Enough Just Right Too Diverse

5) How would you rate the length of the talks (15 minutes + 5 minutes for questions)?

Not Long Enough Just Right Too Long

6) How would you rate the length of the meeting (2.5 days)?

Not Long Enough Just Right Too Long

7) How many previous MIPS meetings have you attended?

0 1-2 3-5 6-8 8-10 11-13 > 14

8) How likely are you to attend future MIPS meetings?

Very Likely Likely Neutral Not Likely Very Unlikely

9) Are you a member of MIPS?

Yes No

10) What is your main area of interest/expertise?

Observer Performance Visual Search Technology Assessment Observer Models

Assessment Methods Observer Expertise Training & Education Human Factors

CAD & Other Computer-Based Decision Aids Statistical Techniques

11) Comments & Suggestions:

Or go online & complete

https://radiologyemory.qualtrics.com/jfe/form/SV_3CATaBcD7zWk86x

Scavenger Hunt

Goal: Get to know each other by finding out who belongs to the clues provided! How you get your answers is up to you. Prizes will be awarded based on number correct! Please complete and hand in to Elizabeth, Mia, Tamara, Mini or Howard no later than Thursday 4:00.

Clue	Person
I have expertise in machine learning and data mining	
I worked at Disney World as a character all throughout college	
Only recently took a liking to Brussels sprouts	
Won 3 medals at a karate tournament	
I love to travel! Recently I`ve been to Chernobyl and Iran	
I was too involved in the construction of my addition	
I have a degree in Chinese	
I have batted on the field at Fenway Park	
I am the first generation in my family NOT to have been educated in a one-room schoolhouse	
Stories are my data, data with a soul	
After practicing medicine for more than 15 years, I have recently completed coursework to finish a PhD	
I went to Iran	
I am the East District Chair of the North American Board of the Union for Reform Judaism	
I have parachuted and hang glided	
I donated 2 feet long hair to pediatric cancer patients	

Your name _____