Medical Image Perception Society (MIPS) Conference XVI

Ghent Marriott Hotel

10 Korenlei, B-9000 Ghent, Belgium

June 3 - 5, 2015

MIPS XVI local hosts and organizers

Asli Kumcu, MSc (Ghent University)
Federica Zanca, PhD (University of Leuven)
Ljiljana Platiša, PhD (Ghent University)

MIPS XVI Organizing Committee

Elizabeth Krupinski, PhD (University of Arizona)
Ljiljana Platiša, PhD (Ghent University)
Federica Zanca, PhD (University of Leuven)
Asli Kumcu, MSc (Ghent University)
Hilde Bosmans, PhD (University of Leuven)
Ewout Vansteenkiste, PhD (Ghent University)
Tom Kimpe, PhD (Barco N.V.)
Bart Goossens, PhD (Ghent University)

MIPS XVI Sponsors

Gold sponsors

Barco N.V., Kortrijk, Belgium

Research Foundation - Flanders (FWO), Brussels, Belgium

Silver sponsors

Holoxica Ltd, Edinburgh, United Kingdom

Bronze sponsors

iMinds - Flanders’ digital research center, Ghent, Belgium

Scholarships

The MIPS Scholars for MIPS XVI are supported in part by the
National Institutes of Health (Grant # R13EB020477)

MIPS XVI Student Scholars:

Amber J. Gislason-Lee, University of Leeds, United Kingdom

Ann Carrigan, Macquarie University, Australia

Djordje Starčević, University of Novi Sad, Serbia

Ellen Kok, Maastricht University, the Netherlands

Felipe Parages, Illinois Institute of Technology, USA

Frank Schebesch, University of Erlangen-Nuremberg, Germany

Irene Hernández-Girón, Leiden University Medical Center, the Netherlands

Manzoor Razaak, Kingston University, London

Maram Alakhras, University of Sydney, Australia

Qu Xiaoxia, Beijing Institute of Technology, China

Stephen Littlefair, University of Sydney, Australia

Vladimir Ostojić, University of Novi Sad, Serbia

List of Abstracts

Design, development and evaluation of a novel digital pathology workstation: The Leeds Virtual Microscope

Rebecca Randell[1] (PhD), Rhys Thomas[2] (PhD), Roy Ruddle[2] (PhD),
Darren Treanor[1,3] (FRCPath, PhD)

[1] *Faculty of Medicine, University of Leeds, Leeds, UK*
[2] *Faculty of Engineering, University of Leeds, Leeds, UK*
[3] *Leeds Teaching Hospitals NHS Trust, Leeds, UK*

Rationale

Digital pathology (whole slide imaging) has the potential to significantly improve the practice of diagnostic pathology. However, its adoption has been slow. A key barrier to the adoption of whole slide imaging has been its relative ineffiency compared to the microscope (up to 60% slower [1]), leading to rejection of the technology by its intended users. We describe a 4 year project to design a novel digital pathology workstation employing user-centred design to make a digital pathology workstation as fast as, or faster than, a microscope.

Methods

We created a multi-disciplinary team including expertise in pathology, health technology assessment, graphics and human computer interaction. We employed an iterative process of information gathering, system design, prototyping and experimental evaluation with professional users.

Results

Our studies of pathologists work practices showed the complex nature of work at the microscope, building a diagnosis from multiple sources of information and with significant switching between activities. Analysis of videos showed that pathologists spent about 60% of their time using the microscope, often in a "hands free" way, allowing them to concentrate on the pathological features without significant effort controlling the instrument [2]. The final workstation was evaluated with 12 expert users. The time taken to report a complex cancer resection (with 12-25 slides) was similar on the workstation and the microscope. With the workstation, pathologists spent a significantly greater proportion of the total task time viewing slides and revisited slides more often [3].

Conclusions

A multi-disciplinary user-centred approach produced to a novel digital pathology workstation which in experimental evaluations had similar performance to the microscope and was acceptable to pathologists.

References

[1] Treanor D, Quirke P. The Virtual Slide and Conventional Microscope - a Direct Comparison of Their Diagnostic Efficiency. J Pathol 2007;213:7a.

[2] Working at the microscope: analysis of the activities involved in diagnostic pathology. Randell R, Ruddle RA, Quirke P, Thomas RG, Treanor D. Histopathology. 2012 Feb;60(3):504-10.

[3] Diagnosis of major cancer resection specimens with virtual slides: impact of a novel digital pathology workstation. Randell R, Ruddle RA, Thomas RG, Mello-Thoms C, Treanor D. Hum Pathol. 2014 Jul 2

Bias and the expert witness in radiological malpractice litigation

Stephen Littlefair MSc[1], Claudia Mello-Thoms PhD[1], Warren Reed PhD[1], Patrick Brennan PhD[1]

[1] *Medical Image Optimisation and Perception Group (MIOPeG), Discipline of Medical Radiation Sciences, Faculty of Health Sciences,* ***University of Sydney****, Australia*

Rationale

Misdiagnosis of disease in a medical image is likely due to search, perception or cognition errors. An additional contributing factor is the poor quality of clinical information given to the radiologist by the referrer. In certain malpractice cases, the radiologist (defendant) who misses a chest lesion may only have been provided with significantly limited information. However, at trial, the expert witness for the plaintiff is privy to a complete medical history and is cognisant of the location of the now known tumour. This unfair advantage allows the expert witness to focus on the particular area of interest.

Methods

15 board certified radiologists were asked to locate lung field malignancies on a dataset of 40 adult chest images (50% abnormal). There was only one nodule on each radiograph but the radiologists could choose as many suspect locations as they wished on each image. A general, poor quality clinical history was given: *cough for 3+ weeks"*. This was called the "defendants read"

4-6 weeks later the radiologists were asked to look at the same dataset (they were unaware that the dataset was unchanged). The radiologists were given the following information: "*all of these patients have a lung tumour diagnosed on a subsequent radiograph 6 months later".* This was termed the "expert witness read."

Results

We are currently in the process of collecting the data for this study, but we are confident that data analysis will be concluded by conference time. Our previous experiments indicated that when a radiologist had a prior expectation of abnormality there was a significant increase in the number of false positive decisions (P= 0.02).

In this particular study we will compare the performance of radiologists with differing clinical information to investigate whether the expectation of abnormality and prior knowledge affects the decision making of a potential expert witness.

Conclusions

Data collection is ongoing. We seek to test the hypothesis that expert-witness testimony is influenced by prior expectation. The results will be presented and if required potential solutions will be proposed.

Searching for the prevalence effect

Todd S. Horowitz[1] (PhD)

[1]*National Cancer Institute, National Institutes of Health, Bethesda, Maryland, U.S.A, Earth*

Rationale

Many medical image tasks are characterized by low prevalence; e.g., prevalence in mammography is approximately .0049. While some studies show that low prevalence leads to more missed cases [1], other studies report no prevalence effect [2]. Meanwhile, basic cognitive psychology research has found robust evidence for a prevalence effect in visual search [3]. I conducted a meta-analysis of prevalence effects in both cognitive psychology and medical image perception to ask whether there is a prevalence effect for measures of detectability and/or criterion, and whether visual search findings would generalize to medical image perception.

Methods

I compiled signal detection parameters *d'* (detectability) and *c* (criterion) from 13 cognitive and 10 medical image studies. I selected comparisons of low (0.02 to 0.30, m = 0.06, sd = 0.08) to medium (0.40 to 0.50, m = 0.49, sd = 0.04) prevalence; ranges reflect the most frequently sampled prevalence values. I computed Hedges g for both parameters, and weighted each effect size by the inverse of the variance.

Results

There was a significant prevalence effect for both cognitive (g = -2.91, 95% CI [-3.16,-2.65]) and medical image (g = -1.06, 95% CI [-1.43,-0.69]) studies. Observers were more conservative when targets were rare. In contrast, prevalence had no significant effect on detectability in either domain (cognitive: g = -0.14, 95% CI [-0.32,0.03]; medical: g = -0.22, 95% CI [-0.56,0.11]).

Conclusions

These findings help resolve some of the disagreement within the medical image perception literature on the prevalence effect. Many researchers have looked for an effect on detectability (e.g., [2]), which turned out to be small and not significant. However, there was a substantial effect on criterion, leading to more missed targets at low prevalence. Critically, these results suggest that we can use basic visual search experiments to help illuminate mechanisms underlying prevalence effects in medical image interpretation, despite large superficial differences in visual stimulus complexity, response procedures, and observer expertise.

References

[1] K. K. Evans, R. H. Tambouret, A. Evered, D. C. Wilbur, and J. M. Wolfe, "Prevalence of Abnormalities Influences Cytologists' Error Rates in Screening for Cervical Cancer," *Arch. Pathol. Lab. Med.* **135**, 1557–1560 (2011).

[2] D. Gur, H. E. Rockette, D. R. Armfield, A. Blachar, J. K. Bogan, G. Brancatelli, C. A. Britton, M. L. Brown, P. L. Davis, et al., "Prevalence Effect in a Laboratory Environment," *Radiology* **228**, 10–14 (2003).

[3] J. M. Wolfe, T. S. Horowitz, M. J. van Wert, N. M. Kenner, S. S. Place, and N. Kibbi, "Low target prevalence is a stubborn source of errors in visual search tasks.," *J. Exp. Psychol. Gen.* **136**, 623–638 (2007).

Regression models for analyzing visual grading studies

Örjan Smedby[1,2] (Dr.Med.Sci.), Seyed Ehsan Saffari[2] (Ph.D.), Áskell Löve[3] (M.D.),
Mats Fredrikson[2] (Ph.D.)

[1]Royal Institute of Technology (KTH), Stockholm, Sweden
[2]Linköping University, Linköping, Sweden
[3]Lund University, Lund, Sweden

Rationale

For optimizing and evaluating image quality in medical imaging, one can use visual grading experiments, where observers rate some aspect of image quality on an ordinal scale. To analyze the grading data, several regression methods are available, and this study aimed at empirically comparing such techniques, in particular when including random effects in the models, which is appropriate for observers and patients.

Methods

Data were taken from a previous study of CT of the brain in 40 patients where 6 observers graded (on an absolute scale) or ranked (sorted in quality order) the image quality of four imaging protocols, differing in radiation dose and image reconstruction method, using three different criteria [1]. The models tested included linear regression, the proportional odds model for ordinal logistic regression, the partial proportional odds model, the stereotype logistic regression model and rank-order logistic regression (for ranking data). In the first two models, random effects as well as fixed effects could be included; in the remaining three, only fixed effects. Goodness of fit was evaluated with Akaike's information criterion (AIC) and McFadden's Pseudo R^2. The potential for dose reduction was estimated with the technique previously proposed by our group [2].

Results

In general, the goodness of fit showed small differences between the models with fixed effects only. For the mixed-effects models, higher AIC but lower Pseudo R^2 was obtained, which may be related to the different number of parameters in these models. With the ranking data, the rank-ordered logistic regression model yielded higher AIC and Pseudo R^2 values than the other models. The estimated potential dose reduction by new image reconstruction methods varied only slightly between models.

Conclusions

We suggest using the ordinal logistic regression model with mixed effects, which can handle ordinal data and random effects appropriately [3]. For rank-order data, the rank-ordered logistic regression model appears to be most appropriate.

References

[1] Löve Á, Siemund R, Höglund P, Van Westen D, Stenberg L, Petersen C, et al. Hybrid iterative reconstruction algorithm in brain CT: a radiation dose reduction and image quality assessment study. Acta Radiol. 2014 Mar;55(2):208-17.

[2] Smedby Ö, Fredrikson M, De Geer J, Borgen L, Sandborg M. Quantifying the potential for dose reduction with visual grading regression. Brit J Radiol 2013;86:31197714.

[3] Smedby Ö, Fredrikson M, De Geer J, Sandborg M. Visual grading regression with random effects. Proc. SPIE 8318, 831805 (2012)

A framework to design experiments to link human and model observers for image quality analysis

R.W. Bouwman[1], R.E. van Engen[1], M.J.M. Broeders[1], G.J. den Heeten[1],K.C. Young[2], D.R. Dance[2], W.J.H. Veldkamp[1,3]

[1]Dutch reference centre for screening (LRCB)
[2]National coordinating centre for physics in mammography (NCCPM)
[3]Leiden University medical centre (LUMC)

Rationale

Statistical model observers are candidates for the evaluation of the image quality (IQ) of processed and/or reconstructed images. Our goal is to introduce statistical model observers for IQ analysis in quality control procedures. The IQ is then estimated using a figure of merit (FOM) resulting from model observers scoring the images. One of the requirements for this FOM is that there exists a sufficient correlation between the performance of human and model observers. Only then can an observed change in IQ be linked to a change in clinical performance. We describe a framework to determine the design parameters for experiments to examine the correlation between model and human observers. This comparison can be made with different model observers, background structures, noise correlations or signals.

Methods

In the framework a two alternative forced choice (2-AFC) detection experiment is simulated using signal detection theory and assuming a linear relation between the performances of models and human observers. To set-up the simulations a pre-study needs to be performed to estimate inter-observer variation and to estimate the expected performance difference for the conditions under investigation. The impact of different design parameters (number of observers, images and data points) can then be simulated. To validate the proposed framework the results of simulated experiments were compared with results of a limited human-model observer experiment. For this purpose we studied the detection of disks of different contrasts and sizes in clustered lumpy backgrounds (CLB), two different model observers and a fixed number of data points.

Results

The impacts of the number of observers and images on the correlation between human and model observers were similar for real and simulated experiments. It was found that increasing the number of observers and images resulted in correlation parameters with smaller variation and a better fit. In the example based on the pre-study further improvements of the goodness of fit were not found after approximately 200 images and four observers.

Conclusion

The proposed framework was found to be effective in estimation of the number of observers and images needed for comparing correlations between human and model observers. This framework facilitates choices regarding the use and limitations of model observers for IQ-analysis based on the correlation with human observers.

Assessment of CT image quality in all reconstruction planes using an updated NPWE model observer

Julien G. Ott[1] (MSc), Fabio Becce[2] (MD), Eric Dugert[2] (MD), François O. Bochud[1] (PhD) and Francis R. Verdun[1] (PhD)

[1] Institute of Radiation Physics, CHUV, Lausanne, Switzerland
[2] Department of Diagnostic and Interventional Radiology, CHUV, Lausanne, Switzerland

Rationale

CT images are traditionally reconstructed and analysed in the axial plane. However, in clinical practice, images may have to be visualised in the coronal and/or sagittal planes, particularly in cardiovascular, thoracic and musculoskeletal imaging.

The recent introduction of iterative reconstructions (IR) has helped to significantly reduce radiation dose, but with a possible change in image quality. While the impact of IR on CT image quality has been widely studied in the axial plane, this has yet to be done for the coronal and sagittal planes.

Methods

Image quality phantoms were scanned in all three planes at a $CTDI_{vol}$ of 7.3mGy using a HD 750 GE scanner. Each data set was reconstructed using a bone convolution kernel, with both classical (FBP) and IR algorithms (ASIR 40 and 80 as well as MBIR). NPS (Noise Power Spectra) and MTF (Modulation Transfer Functions) were then calculated. An additional image quality metric called Target Transfer Function (TTF) was also computed. TTF was obtained using a custom-made phantom in which several contrast differences were measured. Then, applying mathematical treatment to the data allowed to estimate spatial resolution when taking contrast transfer into account. We finally used those data to compute and compare the detectability of a 1-mm articular cartilage lesion in all three planes. Detectability was estimated in the Fourier space, with an updated NPWE model observer relying on the TTF metric.

Results

We found that CT images exhibit significantly different spatial resolution depending on whether it was estimated using the TTF or MTF. This can be explained by the non-linearity of the different kernels and reconstruction algorithms used.

Furthermore, a significant reduction of both spatial resolution and noise was observed in the coronal and sagittal planes compared with the axial plane. This eventually led to a loss in detectability as estimated by the NPWE model observer in the coronal and particularly the sagittal plane compared with the axial plane. In addition, a drastic enhancement of detectability was observed in all reconstruction planes when switching from FBP to MBIR.

Conclusions

CT images acquired under the same conditions and reconstructed in different planes will exhibit significant differences in terms of detectability and, therefore, diagnostic image quality. Thus, viewing clinical examinations in different planes has an impact on diagnostic accuracy and one must be careful when switching from the axial to the coronal or sagittal plane. Nevertheless, our results suggest that the use of IR (particularly MBIR) may compensate for this loss of detectability.

Image Shuffling Saves Time in Mammography

Trafton Drew, PhD[1]; Avi M. Aizenman, BS[2]; Matthew B. Thompson, PhD[2,3]; Mark D. Kovacs, MD[4]; Mike Trambert, MD[5,6,7]; Murray Reicher, MD[7]; Jeremy M. Wolfe, PhD[2,8]

[1]University of Utah; [2]Brigham and Women's Hospital; [3]Queensland University; [4]University of California, San Francisco; [5]Cottage Health System; [6]The Sansum Clinic; [7]DR Systems; [8]Harvard Medical School

Rationale

When astronomers search for newly appearing or moving objects, they co-register two or more images taken from the same view at different times and "shuffle" between the different images. One might imagine that this technique would be helpful in screening mammography as well, but no two mammograms, even of the same breast, are the same. It is known that substantial changes between two images can be missed if the images are misaligned ("change blindness"). In spite of this, might shuffling two images still be superior to standard Side-By-Side (SBS) viewing? Recent work from our lab suggests that shuffling photographs taken from slightly different locations leads to superior performance compared to SBS. Similarly, Riley and colleagues presented data during MIPS 2013 that suggested that shuffling images led to superior performance vs. SBS when searching artificial noise stimuli. Would these results with naïve observers generalize to radiologists viewing medical images? In mammography, we developed a technique called image shuffling where a metafile associated with each image was used to automatically sort new and prior exams so that images could be sequentially viewed. This method enables efficient flickering of breast imaging exams.

Methods

Twenty-Four radiologists participated in the experiment while attending large radiology meetings (ARRS and RSNA). Each radiologist viewed four-view screening mammograms each with a prior comparison mammogram. They evaluated two practice and ten experimental exams. Radiologists were asked to treat the experiment as if they were performing screening mammography. The experimental exams consisted of 5 proven normal exams and 5 biopsy-proven subtle breast cancers. Radiologists were randomly assigned to view each case in either SBS or Shuffle viewing mode. They were asked to go as quickly as possible without sacrificing accuracy. Each case was rated on the BIRADS scale, with the quadrant of detected lesion, and nature of detected lesion (mass vs. calcification) indicated as well. Case-level accuracy was determined based on BIRADS rating.

Results

Radiologists reached their decisions significantly faster when viewing mammograms in Shuffle mode ($F(1,23)=5.15$, $p<.05$). On average, Shuffle cases were completed 14 seconds faster than cases viewed SBS, a 15% reduction of total time spent. Performance, as measured by correctly calling back cases that contained an abnormality, was somewhat higher in Shuffle mode but the study lacks the power to determine if this is a reliable effect ($F(1,23)=1.02$, $p>.05$). Similarly, significant differences were not found for d' (discriminability index), nor c (criterion).

Conclusions

Shuffle mode led to significantly faster performance with at least equivalent diagnostic accuracy when viewing this enriched sample of mammograms containing subtle abnormalities. Time savings were found even though successive mammograms were not perfectly aligned. Given the ever-increasing case load for radiologists, this simple manipulation of how the images are viewed could save valuable time in clinical practice, allowing radiologists to read more cases, or spend more time on difficult cases. Future work will extend this finding to other modalities, such as chest radiographs, and explore whether automated co-registration may further enhance the benefits of the Shuffle viewing mode.

Using Mobile Technology (and Big Data) to Inform Radiological Research

Stephen R. Mitroff12 (Ph.D.), Adam T. Biggs1 (Ph.D.), Justin M. Ericson1 (Ph.D.), Jonathan Winkle1(B.S.), Stephen H. Adamo1 (B.S.), & Emma Wu Dowd1 (M.Sc.)

[1]*Center for Cognitive Neuroscience, Duke University, Durham NC, U.S.A.*
[2]*MedStar Health Research Institute, Washington D.C., U.S.A.*

Rationale

Academic radiology and cognitive psychology research have helped reveal the underlying causes of certain types of radiological search errors; however, several sources of these errors have remained elusive. In the current presentation, we will discuss how data collected from millions of searchers across billions of trials can inform specific radiological search problems that have been especially hard to address in laboratory or clinical settings. Specifically, we will present data that informs search errors related to *satisfaction of search* and rarely-appearing targets.

Methods

We have partnered with Kedlin Co., the makers of a smartphone app called *Airport Scanner*, to obtain "big data." *Airport Scanner* is a game where the player serves as an airport security officer and searches for contraband in simulated carry-on bags. The game contains numerous elements that are ideal for research endeavors - a variable number of targets per bag, a variable number of distractors per bag, multiple levels with varying difficulty, hundreds of different target types and distractor types, a secondary distraction task, etc. We have access to over 2 billion trials from over 7 million devices, and we have used this unique dataset to address questions that have been previously intractable [1-4].

Results

We will discuss several findings relevant to radiological search. For example, we have demonstrated that *satisfaction of search* (an increased risk of missing a target after having already found another target) is partially caused via a 'perceptual set' mechanism - after finding a target, you are more likely to find other targets that are perceptually and conceptually similar [1]. Likewise, we will show how target frequency (how often a specific target appears across all searches) can greatly affect search accuracy [3].

Conclusions

It is vital to minimize radiological search errors, but this can only be done by understanding the causes of each error type. We will present a novel technique for investigating the general search behaviors that can underlie radiological search errors. This approach complements and expands current research endeavors, and most importantly, can address previously intractable problems.

References

[1] Biggs, A. T., Adamo, S. H., Dowd, E. W., & Mitroff, S. R. (in press). Examining perceptual and conceptual set biases in multiple-target visual search. *Attention, Perception, & Psychophysics*.

[2] Biggs, A. T., Adamo, S. H., & Mitroff, S. R. (2014). Rare, but obviously there: Effects of target frequency and salience on visual search accuracy. *Acta Psychologica*, 152, 158-165.

[3] Mitroff, S. R., & Biggs, A. T. (2014). The Ultra-Rare-Item effect: Visual search for exceedingly rare items is highly susceptible to error. *Psychological Science*, 25(1), 284-289.

[4] Mitroff, S. R., Biggs, A. T., Adamo, S. H., Dowd, E. W., Winkle, J., & Clark, K. (2014). What can 1 billion trials tell us about visual search? *Journal of Experimental Psychology: Human Perception & Performance.*

Role of Statistical Learning in Radiological Diagnosis of Cancer

Jay Hegdé (PhD)

Department of Ophthalmology, Georgia Regents University, Augusta, GA, USA

Rationale

The tremendous variation among radiological images means that diagnosing an anomaly is fundamentally a statistical decision-making process. It also implies that acquiring the diagnostic expertise must involve statistical learning, whereby the radiological trainee must learn, however implicitly, what a given type of anomaly “looks like” and how it differs from normal variations in the underlying tissue. We have recently demonstrated the existence of such “deep learning” mechanisms in the analogous context of learning to recognize camouflaged objects in natural scenes [1]. In the present study, we tested the hypothesis that naïve subjects can learn to detect malignancies in mammograms solely by learning the statistical properties of malignant and non-malignant breast tissue.

Methods

Using a modification of our previously described image synthesis approach [1], we generated a large number of synthetic mammograms using, as input, actual screening mammograms with or without a malignancy. During each trial, naïve subjects with no previous radiological training viewed a randomly drawn image for 500 ms and reported, using a key press, whether or not the given image contained a malignancy. During the training blocks, subjects received a visual feedback at the end of every trial. The trials during the testing blocks were identical, except that that the subjects received no feedback. Subjects trained in blocks of 50 trials each until they reached asymptotic performance.

Results

Before the training, subjects performed at chance levels ($d' \approx 0.0$; $p >> 0.05$) as expected. During the training, subjects’ diagnostic performance improved rapidly, reaching a criterion level ($d' = 2.1$; $p < 0.05$) in as few as 7 training blocks (mean, 11.2 blocks; maximum, 21 blocks; N =5). Similar results were obtained even when the subjects were never shown the same image twice, so that the only way they could learn the diagnostic task was to learn the statistical properties of the training images.

Conclusions

Our results suggest that statistical learning can account, at least in part, for how radiological diagnostic expertise is acquired. They also suggest the possibility that statistical learning can be used as an effective pedagogical tool in radiological training.

References

[1] Chen, X., and Hegdé, J., “Learning to break camouflage by learning the background,” 23(11), 1395-1403 (2012).

This work was supported by the U.S. Army Research Laboratory and the U.S. Army Research Office Grant W911NF-11-1-0105 to J. H.

How much image noise can be added in cardiac X-ray imaging without loss in perceived image quality?

Amber J. Gislason-Lee (MSc), Asli Kumcu[a] (MSc), Stephen M. Kengyelics (MSc), Laura A. Treadgold (PhD), Andrew G. Davies (MSc)

Division of Biomedical Imaging, University of Leeds, UK
[a]iMinds-IPI, Ghent University, Belgium

Rationale

X-ray imaging systems which provide images in real-time are essential for diagnosis and interventional treatment of coronary heart disease. Cardiologists use live, moving images of the coronary arteries called angiograms for diagnosis. X-ray exposure may cause damaging short term effects such as skin burns and long term genetic effects such as cancer. As interventional procedures become longer, more complicated, and more frequent, it is increasingly important to minimize X-ray dose whilst maintaining adequate image quality. Technical image quality measurements including noise are generally used to determine optimal dose levels, and they can be accurately measured using static phantom images, however it is not well understood how changes in these measurements are perceived by a clinician. For example, when treating a patient's heart, a clinician may not notice image degradation caused by reducing the X-ray dose. This study aims to determine how much noise can be added to a patient image without altering the perceived quality of the image. Noise is directly related to radiation dose, therefore results may demonstrate potential for a reduction in radiation dose used for cardiac interventional procedures; this would benefit patients and personnel.

Methods

Image noise was added to five percutaneous coronary interventional (PCI) patient angiograms, selected to represent the range of adult cardiac patient sizes and to include angular cardiac views commonly used in clinical practice. Incremental amounts of computer-generated quantum noise were added to the angiograms – frame by frame, pixel by pixel - to simulate corresponding levels of dose reduction. Ten cardiologists, radiologists and radiographers working in a cardiac catheter lab viewed image pairs, selecting the preferred image in a two alternative forced choice staircase (1 up / 3 down) psychophysics experiment; each pair had the original and a degraded image. After a training period the level of image degradation was varied based on the previous response, to determine the point of subjective equality. The software used was written in MATLAB specifically for this task.

Results

The median point of subjective equality was 33% ± 16% dose reduction for the five PCI patients.

Conclusions

Results demonstrated scope to increase noise of cardiac X-ray images by up to 33% before it is noticeable by clinical professionals, indicating a potential for 33% dose reduction.

BREAST 2011-2014: Scientific achievements and future directions

Brennan P.C.[1] (PhD), Tapia K (BA)[1], Trieu P.D. (MDR)[1], Ryan J. (PhD)[2],
Lee W.B. (MBBS FRACR)[1]

[1]Faculty of Health Sciences, University of Sydney, Australia. [2]Ziltron, Dublin, Ireland

Rationale

Mammographic imaging remains a front line for detecting breast cancer with over 1.5m women X-rayed in Australia every two years. Nonetheless, approximately 30% breast cancers are being missed on a single read. The Breast Screen Reader Assessment Strategy (BREAST) is an innovative quality assurance and competence programme developed in Australia by the University of Sydney, BreastScreen NSW, and Ziltron, to a) monitor the performance of radiologists involved in BreastScreen, b) establish the source of mammographic errors and, c) explore solutions to reduce errors. BREAST presents a range of test sets, with at least one of these having been completed by 70% of Australian and New Zealand BreastScreen clinicians. This has resulted in 50,000 ratings or data inputs, thus constituting a valuable resource to clinicians and scientists world-wide.

Methods

Radiologists judge test sets each containing 60 cases and are asked to record their decisions using Ziltron's online software framework. Instantly upon completion of a test set, readers are provided with ROC, JAFROC, sensitivity, and specificity scores describing their performance. Reader-specific image files are immediately generated and readers can review all the images they have just read and examine truth and correct and incorrect decisions made for each image. The data generated by users are de-identified and securely stored in the BREAST database.

A BREAST Access and Management Committee (BAMC) was established and policy and procedures developed in 2014 to facilitate global access to BREAST-generated data.

Results

The data arising from research projects using the BREAST initiative has led to an improved understanding of: the usefulness of digital breast tomosynthesis in the Australian context; optimising radiologic reading conditions for experimental activities; cancer presentations that impact diagnostic performance; intelligent pairing of radiologists for double reading; the impact of breast density on radiologic diagnosis; and radiologists' characteristics that improve diagnostic efficacy.

Conclusion

The avid acceptance of BREAST by Australian radiologists has led to useful scientific outputs. In addition to being a reflective teaching and assessment tool, the program has improved our knowledge around issues important to breast cancer diagnosis. Future projects are welcomed. International implementations and future directions will be discussed.

An overview of BREAST - a multi-parametric scheme designed to increase radiologist performance in breast cancer detection

Brennan P.C.[1] (PhD), Trieu P.D. (MDR)[1], Tapia K. (BA)[1], Ryan J. (PhD)[2], Lee W.B. (MBBS FRACR)[1]

[1]Faculty of Health Sciences, University of Sydney, Australia. [2]Ziltron, Dublin, Ireland

Rationale

Breast cancer is one of the most common types of cancer diagnosed in women. The rate of females developing breast cancer in Australia is one in nine and the risk of mortality is one in thirty seven. The success of screening programs depends on the accurate image interpretation of radiologists implying that if reader efficacy is monitored and individual-specific errors highlighted, underperformance can be identified and addressed. For the purpose of improving the capability of radiologists in breast cancer detection, a novel web-based solution known as Breast Screen Reader Assessment Strategy (BREAST) has been introduced through a collaboration between the University of Sydney, BreastScreen NSW and an information technology partner, Ziltron. Over the last 3 years, BREAST has received highly positive feedback from radiologists and it is considered an essential strategy in improving the readers' skill in breast cancer detection. Along with clinical benefit, the scheme substantially supports a wide range of research activities in medical image perception through a large number of radiologist interactions with a high quality image database.

Methods

BREAST provides clinically-relevant, cancer enriched image test sets which consist of 60 cases with two view digital bilateral mammograms in each set. With access to Ziltron's online system, readers are able to investigate image case sets at the full resolution available through local PACS systems. In each test set, radiologists are asked to record whether the case is normal or abnormal and for each abnormal finding that would warrant further mammographic assessment, lesion localization and BIRADS score are required. Once each set is completed, readers will receive instant feedback about their performance for each case and an overall report which details the number of correct recall cases, percentage of correct negative cases, percentage of correctly identified lesions, ROC, JAFROC scores along with potential reference levels of good performance (25th, 50th, 75th percentiles). The data generated by readers is de-identified and used for clinical and research purposes.

Results

After 3 years, the BREAST programme is available and well accepted in all states across Australia and New Zealand. Over 500 readings have been performed by more than 200 screen readers, accounting for approximately 70% of BreastScreen readers. According to radiologists, BREAST test sets are significant to their on-going professional development. 86% of readers surveyed highlighted the value of this learning activity as an effective strategy for training. The data from BREAST has served our research community well with a series of publications in leading radiologic journals

Conclusion

The introduction of BREAST in Australia is recognized as being important to scientists, clinicians and BreastScreen managers. With the high level of reader engagement, it is anticipated that this novel approach will assist the optimization of mammographic readings, and increase the radiologists' ability to detect breast cancer. We welcome collaborations and ideas to expand the use of the BREAST program.

The effect of mammographic breast density in the digital imaging era

Dana S. AL Mousa[1] (MSc), Claudia Mello-Thoms[1] (PhD), Elaine A. Ryan[1] (PhD), Patrick C. Brennan[1] (PhD)

[1] *Department of Medical Imaging and Radiation Sciences, Faculty of Health Sciences, University of Sydney, NSW, Australia*

Rationale

To understand the effect of mammographic density on radiologic diagnostic efficacy and behaviour in the digital imaging era.

Methods

This work consists mainly of two studies: the first is a pilot study on a set of 55 digital mammograms examined by 22 radiologists. Mammograms were classified into low- and high mammographic density to investigate radiologists' performance defined by: sensitivity, location sensitivity, specificity, receiver operating characteristic (ROC Az) curves and jackknife free-response receiver operator characteristics (JAFROC) figures of merit (FOM). The second study included a set of 150 digital mammograms examined by 14 radiologists (six and eight radiologists who read more and less than 2000 cases per year), seven of which underwent eye-tracking recording. Images were classified into low- or high-mammographic density, in order to investigate the impact of density on radiologists' performance and visual search patterns. Lesions overlaying were compared to those outside the dense fibroglandular tissue.

Results

The pilot study showed significant increases in sensitivity (P= 0.02) and ROC Az (P=0.0001) in high- compared to low- mammographic density cases. These findings were supported by the second study results where radiologists who read more than 2000 per year showed significantly higher JAFROC FOM (P=0.04) for high- compared to low- density mammograms. All radiologists and those radiologists reading more than 2000 cases per year showed increased radiologists' performance with high- compared to low- mammographic density cases when lesions are overlaying the fibroglandular tissue. Eye-position data showed a significant increase in time to first hit when lesions are located outside- compared to overlaying- fibroglandular tissue (P=0.001) in both low- and high-mammographic density cases. In addition, dense areas of breast parenchyma and lesion areas when they are overlaying the fibroglandular region attracted radiologists' visual attention.

Conclusions

The findings showed that in modern digital imaging systems, increased mammographic density improved the performance of experienced radiologists which may be linked to the change in their visual search patterns when interacting with high mammographic density images. These findings challenge our current understanding of breast density impact on diagnostic efficacy and have important implications for symptomatic and screening mammography.

Impact of Fatigue on Satisfaction of Search (SOS)

Elizabeth A. Krupinski[1] (PhD), Kevin S. Berbaum[2] (PhD), Kevin Schartz[2] (PhD), Alexandra Schaeffer[1], Robert Caldwell[2] (MFA), Mark Madsen[2] (PhD)

[1]*Department of Medical Imaging, University of Arizona, Tucson, AZ USA*
[2]*Department of Radiology, University of Iowa, Iowa City, IA USA*

Rationale

Our long-term objective is to understand and overcome SOS errors. Studies have revealed that SOS occurs for both perceptual *and* cognitive reasons including faulty scanning, changes in visual search, faulty pattern recognition, and faulty decision making. The influence of fatigue on SOS has yet to be investigated.

Methods

20 observers at 2 institutions examined a series of 70 CR chest exams, 38 with diverse abnormalities and 32 without disease. In the SOS condition a simulated solitary pulmonary nodule was inserted and in the non-SOS condition no nodules were added. Readers viewed the cases in the SOS and non-SOS conditions after a long day interpreting clinical cases. Data were analyzed using ROC techniques and compared to a related SOS study in which observers viewed the cases during normal working hours (not fatigued).

Results

The non-fatigued ROC study revealed that the SOS condition results in a shift in reporting criteria, especially for residents – there is a reduction in their willingness to respond rather than a reduction in accuracy. The fatigue SOS study revealed a similar shift in criteria rather than a reduction in accuracy, again more prevalent for residents than attendings. Median inspection times for (1) normal cases without nodules, (2) normal cases with added nodules, (3) abnormal cases without added nodules and (4) abnormal cases with added nodules showed that the addition of nodules did not significantly affect inspection time, with non-SOS readings requiring 48.3 sec on average and SOS readings requiring 46.5 sec. Presence of a native abnormality required greater inspection time ($p < 0.0001$) with 40.2 sec required without native abnormalities and 54.5 sec with.

Conclusions

This study suggests that the SOS effect, as opposed to what was observed in previous SOS studies, may be due more to a shift in reporting criteria than a reduction in accuracy due to faulty search or detection mechanisms. Fatigue does not appear to change the nature of this observation. Given the complicated nature of running a fully-crossed non-SOS/SOS study with the same observers before and after a long day of work it is not possible to directly assess whether fatigue results in an increase in this criteria shift. Residents have a larger shift in criteria than attendings.

Teaching Search Patterns to Medical Trainees in an Educational Laboratory

William F. Auffermann (MD/PhD), Brent P. Little (MD), Srini Tridandapani (MD/PhD)

Department of Radiology and Imaging Sciences, Emory University School of Medicine
1365 Clifton Road NE, Atlanta, GA 30322, USA

Rationale

Much is known in the field of medical image perception about the way in which experts and novices interpret images. This raises the question of whether our knowledge can be used to develop educational tools to help healthcare trainees become better at image interpretation. The goal of this project is to demonstrate that showing a comprehensive lung scan pattern to healthcare trainees in a multi-subject computer laboratory improves a subject's ability to identify pulmonary nodules on chest radiographs (CXR), and allows acquisition of useful data for perception research.

Methods

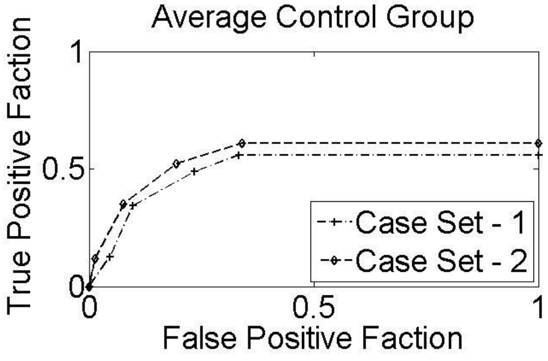
This study was deemed institutional review board exempt. Subjects were split into control and experimental groups. Each group had their own room with one subject per computer. CXRs were divided into 2 case sets of equal size and nodule prevalence. Subjects were presented randomized images, asked to mark a nodule if present, and give an estimate of their confidence using a 5 point scale. The experimental group was shown a comprehensive lung search/scan pattern between the first and second case sets, the control group was not. Performance was quantified using localization receiver operator characteristic (LROC) analysis and differences in the area under the LROC curve (ΔAUC).

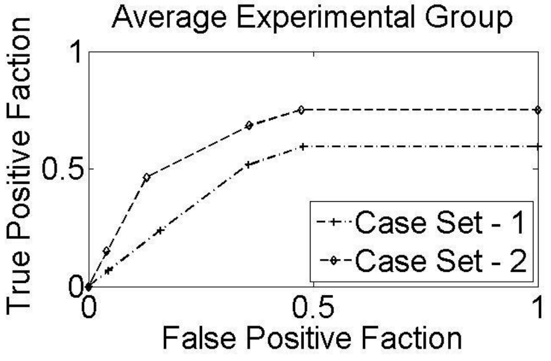
Results

Control Group: There was no difference in subject performance between case sets, ΔAUC = 0.0559, p = 0.1253. Experimental group: There was an improvement in subject performance after training, ΔAUC = 0.1539, p = 0.0012.

Conclusions

Search pattern training may be taught in a laboratory environment to improve healthcare trainees' ability to identify pulmonary nodules on CXRs. Our knowledge of medical image perception may potentially be used to design tools for training healthcare practitioners to be better at medical image interpretation.





Identifying error types in radiological image interpretation of learners

Cécile J. Ravesloot[1] (MD), Anouk van der Gijp[1] (MD), Marieke F. van der Schaaf[2] (PhD), Josephine C.B.M. Huige[1] (MD), Olle ten Cate[3] (PhD), Koen L. Vincken[4] (PhD), Christian P. Mol[4], Jan P.J. van Schaik[1] (MD, PhD)

[1]Radiology department, University Medical Center Utrecht, The Netherlands
[2]Department of Education, Utrecht University, The Netherlands
[3]Center for Research and Development of Education, University Medical Center Utrecht, The Netherlands
[4] Image Sciences Institute, University Medical Center Utrecht, The Netherlands.

Rationale

Errors occur in different phases of the image interpretation process of learners. Insight in the error types made by learners is crucial for giving effective feedback. Using a step-by-step-questions assessment procedure to identify error types in image interpretation and reveal partial knowledge of or hidden errors in the interpretation process, we investigated which error types can be identified in the image interpretation process of radiology clerks and the reliability of this procedure in terms of inter-rater agreement.

Methods

Hundred-nine radiology clerks took a radiology image interpretation test consisting of ten CT image cases and one to three X-ray cases. The image interpretation questions concerned step-by-step questions: labelling an abnormality (perception), describing the abnormality (analysis) and giving a diagnosis and/or advice (synthesis) [1]. Errors were coded as perception, analysis, synthesis or undefined errors by two independent observers. A hidden error was identified if a correct diagnosis was given, based on an incorrect perception or analysis. Partial knowledge was identified if an incorrect diagnosis was given based on a correct perception and/or analysis.
Consensus was reached after discussion, in case of discrepancies. Prevalence of error types and inter-rater reliability of the procedure were calculated.

Results

With our step-by-step questions procedure applied to 1351 cases, 831 errors were identified. 638 errors were found in the process of image interpretation (77%), of which 29.6% were perception errors, 15.7% analysis errors and 31.7% synthesis errors. The step-by-step questions revealed hidden errors in 125 cases (9%) and partial knowledge in 243 cases (18%). We found a mean inter-rater reliability of Cohen's $\kappa = 0.8$.

Conclusions

A step-by-step question approach can reliably distinguish perception, analysis and synthesis errors. Besides, the approach reveals hidden errors and partial knowledge of students.

References

[1] van der Gijp A, van der Schaaf MF, van der Schaaf IC, Huige JC, Ravesloot CJ, van Schaik JP, et al. Interpretation of radiological images: towards a framework of knowledge and skills. Adv Health Sci Educ Theory Pract. 2014.

Multi-faceted evaluation demonstrates a variety of benefits when digital breast tomosynthesis is used with digital mammography

Maram M Alakhras[1] (M.S.), Claudia Mello-Thoms[1,3], (PhD) Mary Rickard[1,2] (MD), Roger Bourne[1] (PhD), Patrick C Brennan[1] (PhD)

[1]Medical Image Optimisation and Perception Research Group, University of Sydney, Australia
[2]Sydney Breast Clinic, Sydney, NSW, Australia
[3] Department of Radiology, University of Pittsburgh, Pittsburgh, U.S.A.

Rationale

The major shortcoming of digital mammography (DM) is tissue overlap due to the two dimensional (2D) nature of the images produced. This may result in obscuring of lesions and, consequently missed cancers and, alternatively, false positive diagnoses. Digital breast tomosynthesis (DBT) is a novel three dimensional imaging technology which reduces tissue overlap, thus improving breast cancer detection and identification of normal tissue. This study provides an assessment of the performance of DBT in conjunction with DM in terms of localisation performance, confidence level of radiologists in scoring breast cases, and the radiologists' ability to identify lesion type.

Methods

The study included 50 cases (27 cancer, 23 normal/benign), each using both DM and DBT modalities. Twenty three experienced breast radiologists interpreted all cases and gave each case a confidence score (1-5) where 1 = "normal", 2 = "benign", 3 = "equivocal", 4 = "suspicious, and 5 = "malignant". If a lesion was marked the type was reported (stellate mass, round mass, non specific density, architectural disturbance or microcalcifications). Statistical analyses were performed to compare jackknife free response receiver operator characteristics figure of metric (JAFROC FOM), confidence level of radiologists, and lesion type identification for DM+DBT compared with DM alone.

Results

Use of DM+DBT resulted in significantly improved JAFROC FOM (0.745 vs 0.621, $p< 0.001$) and higher confidence levels in scoring cancer (3.977 vs 3.635, $p< 0.0001$) and normal cases (2.731 vs 2.881, $p= 0.0179$) compared with DM alone. Adding DBT to DM increased the number of correctly identified stellate breast lesions (20.3% of stellate masses were missed on DM while correctly marked on DM+DBT compared with only 1.2% missed on DM+DBT and correctly marked on DM alone).

Conclusions

This study suggests that adding DBT to DM significantly improves: diagnostic performance for breast cancer diagnosis; confidence level in scoring normal/benign and cancer cases; and identification of stellate masses. Whilst further research involving a higher numbers of specific lesion types is required, our multi-faceted approach confirms the value of DBT when used with DM.

Investigating texture characteristics and the association to detections rate in screening mammography

Mohammed A. Rawashdeh, PhD, Warwick B. Lee, PhD, Sarah Lewis, PhD, Warren Reed PhD and Patrick C. Brennan, PhD

Department of Medical Imaging and Radiation Sciences, Faculty of Health Sciences, University of Sydney, Australia

Rationale

Variations in the performance of expert breast readers are well known, but how lesion types and texture types influence such variations is under explored. This large study, involving 129 readers, investigates the relationship between the image texture characteristics of malignant breast lesions and their rate of detection by radiologists reporting screening mammograms. Institutional review board ethics was granted for the study.

Methods

The test set comprised of 60 cases with 20 positive (biopsy proven) cases, 16 with a single mass and 4 cases with multicentric masses (resulting in a total known 24 cancers overall). 129 readers, all breast imaging radiologists, were asked to identify and locate the lesions. Each lesion was then ranked according to a detectability rating (the number of observers who correctly located the lesion divided by the total number of observers). A lesion was considered correctly detected when it received a confidence scores between 3 – 5 and the location was correctly marked by reader within 2 cm from the center of the lesion. The lesion centres were determined by two experienced radiologists who did not participate as readers in the study. Additionally, the detection rate of each cancer was correlated with following quantities:

- Type of lesion present including mass lesions (n=8), calcifications (n=8), architectural distortion (AD) (n=4) and unspecific density (NSD) (n=4).
- A grey-level co-occurrence matrix (GLCM): Analysis of the texture characteristics of each lesion and of adjacent tissue performed using Matlab (version 7.13, by Math-Works).

Results and Conclusions

The mean age of the readers was 50.7 years, the mean years since qualification was 14.4, and the mean number of years of experience in reading mammograms was 10.9. The sample had a mean number of hours per week reading breast images of 12.2, with an average of 4005 mammographic readings per year. The “detectability rating” varied from 40% to 95% with a median 67% for all lesions. The mean detectability rating for masses was 0.73, for calcifications 0.48, for AD 0.43, and for unspecific density 0.67. Kruskall-Wallis test was used to determine statistically significant differences amongst the 4 lesions types. The Kruskall-Wallis test showed significant differences among the 4 groups for all lesion types (mass lesions, H=63.587, P<0.0001; calcifications, H=28.751, P<0.0001; architectural distortion, H=39.137, P<0.0001; NSD, H=25.234.137, P<0.0001). The work still in progress and the texture data will be presented in the conferences.

Systematic viewing strategies in Radiology: seeing more, missing less?

Ellen M. Kok[1] (MSc), Halszka Jarodzka[2] (PhD), Anique B.H. de Bruin[1] (PhD), Hussain BinAmir[3],
Simon G.F. Robben[4] (PhD), Jeroen J.G. van Merriënboer[1] (PhD)

[1]Department of Educational Research and Development, Maastricht University, Maastricht
[2]Welten Institute, Open University of the Netherlands, Heerlen
[3]International Track Medicine, Maastricht University
[4]Department of Radiology, Maastricht University Medical Center

Rationale

Radiology textbooks and websites recommend a systematic viewing strategy to interpret (chest) radiographs, in which a list of anatomical areas is consistently inspected according to a fixed order. This is supposed to ensure full inspection of the images (i.e. full coverage), and in turn, improve diagnostic performance. We investigated the assumption that a systematic viewing strategy can lead to increased coverage of the image, which in turn leads to improved performance. In order to do so, we investigated whether medical students with no experience in radiology could be trained in systematic and full-coverage viewing strategies.

Methods

Seventy-five 2nd year medical students underwent training in either systematic, full-coverage (no focus on systematicity) or non-systematic viewing. The content of the trainings was the same in three groups; only the viewing strategy taught was varied between groups. The training consisted of a 40 minute video that explained and showed the viewing strategy and 5 practice items on which feedback was given. After the training, participants interpreted 22 chest radiographs during which we measured their eye movements using a 250 Hz Remote eye tracker from SMI. The chest radiographs presented different types of abnormalities, ranging from very subtle to clearly visible. Nineteen of the images contained more than one abnormality. We measured the amount of systematic viewing (using Levenshtein Distance) and average percentage coverage of images. Participants were required to click on all abnormalities they saw. We used this data to calculate sensitivity and specificity.

Results

We found a significant correlation between the amount of systematic viewing and coverage, but no relationship between either of these variables and performance. Participants in the systematic viewing condition viewed images more systematically that participants in the other two conditions. The percentage coverage was highest in the systematic viewing and the full-coverage viewing conditions. However, systematic viewing and non-systematic viewing yielded similar sensitivity, the full-coverage viewing condition yielded lowest sensitivity. No significant differences were found in specificity.

Conclusions

Although eye tracking data shows that we succeeded in teaching participants systematic, full-coverage or non-systematic viewing, we did not find a relationship between systematic viewing, coverage and diagnostic performance. These data question systematic viewing as the “gold standard” in radiology teaching.

Polar-Map Model Observer for Optimization of 3D SPECT-MPI Reconstruction

Felipe M. Parages[1] (MS), J. Michael O'Connor[2] (PhD), P. Hendrik Pretorius[2] (PhD),
Jovan G. Brankov[1] (PhD)
[1]ECE Department, Illinois Institute of Technology, Chicago, IL.
[2]Department of Radiology, University of Massachusetts Medical School, Worcester, MA.

Rationale

Model observers (MO) are widely used in medical imaging to act as surrogates of human observers in task-based image quality evaluation. In SPECT myocardial perfusion imaging (MPI), a realistic task-based approach involves detection and localization of perfusion defects, as well as a subsequent assessment of lesion severity. In this paper we explore a machine-learning MO based on Naive-Bayes classification (NB-MO) of polar-map features for these diagnostic tasks, with the goal of finding the optimal range for some reconstruction smoothing parameter, namely: cutoff frequency of ramp-filter in FBP, width of 3D Gaussian post-reconstruction filter in OSEM.

Dataset and Methods

Our simulated dataset (280 cases) included lesions with different sizes, perfusion-reduction ratios, and locations. Averaged projections were reconstructed using FBP and OSEM methods, with several pre- and post-reconstruction smoothing levels, respectively, that we aim to optimize. For each case, five human specialists (physicians) were presented with the typical SPECT clinical view (SA, HLA and VLA), and were required to score each LV segment with a discrete value score ranging from 0 (normal perfusion) to 4 (absent uptake). A Naive-Bayes classifier is then trained using polar-map image features and the corresponding human scores given for each segment. Next, it is tested over polar-maps not seen during training, *aiming to predict each human observer separately*. Finally, a multi-reader multi-case analysis of alternative free-response ROC (AFROC) curve was performed for NB-MO and human observers. NB-MO was validated following two training strategies: 1) training with one FBP dataset 2) training with one FBP and one OSEM dataset. For comparison, we also report performances of a non-prewhitening (NPW) MO applied on polar-map images.

Results

Results on Fig. 1 show agreement between performance of NB-MO and humans, as well as optimal smoothing values whose ranges correlate with those typically recommended for clinical practice, regardless of the training strategy followed.

Conclusions

A machine-learning MO based on Naive-Bayes classification may be successfully used for optimization of SPECT-MPI reconstruction algorithms, acting as a human observer surrogate in tasks that involve detection, localization and assessment of cardiac perfusion defects.

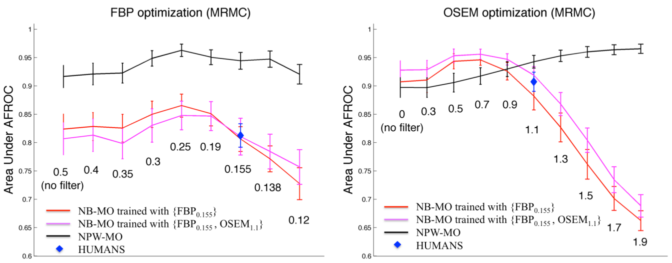


Figure 1. Multi-reader multi-case (MRMC) AFROC performance for different smoothing levels in FBP (left) and OSEM (right) reconstructions, predicted by NB-MO and NPW-MO. Figure also shows performance of human readers for reconstructions used in the human observer study, namely: FBP(cutoff fc = 0.155) and OSEM(Gaussian $\sigma = 1.1$). Error bars indicate ± 1 standard deviation.

What do models of visual perception and recognition see

Aude Oliva[1] (PhD)

[1] *Computer Science and Artificial Intelligence Laboratory, MIT*

Rationale

With the success of new computational architectures for visual processing (e.g. convolutional neural networks), and access to image databases with millions of labeled examples, the state of the art of computational vision and cognition is advancing rapidly. One important factor for continued progress is to understand the representations that are learned by these models.

Methods & Results

Here, I will show that meaningful parts and diagnostic objects naturally emerge from training these neural networks to perform visual scene classification, demonstrating that the same network can perform both scene (the whole image) recognition and object localization (the parts that compose the scene) in a single forward pass.

Conclusion

I will discuss how this approach can be extended to other domains of visual knowledge, like medical images classification and segmentation.

An internal-noise observer equivalent to CSF-based 3D anthropomorphic observer

Ali R. N. Avanaki (PhD)[1], Kathryn S. Espig (MS)[1],
Tom R. L. Kimpe (PhD)[2], and Andrew D. A. Maidment (PhD)[3]
[1]Barco Healthcare, Beaverton, OR [2]Barco Healthcare, Kortrijk, Belgium
[3]University of Pennsylvania, Department of Radiology, Philadelphia, PA

Rationale

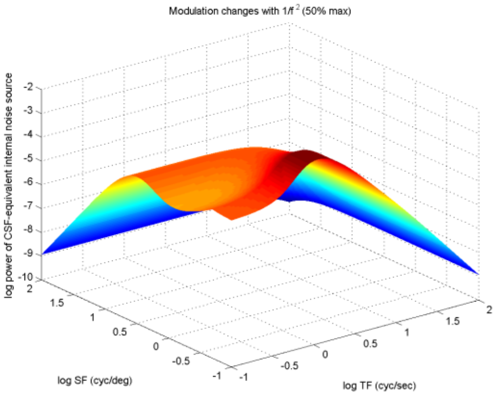
For one class of numerical observers, internal noise models are calibrated to match the performance of human observers for a given detection scenario [Lu & Dosher JOSA-A 1999]. Another class of anthropomorphic numerical observers are modeled after the properties of human visual system (HVS) such as contrast sensitivity function (CSF) [Avanaki *et al* SPIE MI 2014]. We investigate the relationship between the two classes by deriving an internal noise observer equivalent to our anthropomorphic HVS-based observer.

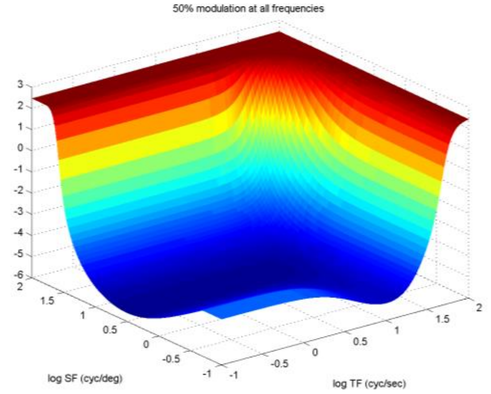
Methods

We calculated the performance of a typical human observer, for detection of a tonal spatiotemporal signal in Gaussian noise, using the following two methods and equated the results. The methods are based on either Barten's spatiotemporal CSF with psychometric function [Barten 1999] or a matched filter detector with a white additive internal noise source. Thus, the power of the noise source in the second observer as a function of signal properties was calculated.

Results

Under the assumption that detection of a signal may be decomposed to the detection of its spatiotemporal frequency components, the noise power of the CSF-equivalent internal-noise observer may be calculated for a given image signal. Internal noise "spectrums" calculated for two representative spatiotemporal frequencies of the signal are depicted below: image signal spectrum either changed with $1/f^2$ (left) and or was constant (right).


Modulation changes with $1/f^2$ (50% max)


50% modulation at all frequencies

Conclusions

Since the power of noise in the internal-noise equivalent observer is a function of the spatiotemporal frequency and modulation of the signal, the two classes of observers are not equivalent. In other words, additive or additive-multiplicative internal noise models and white internal noise are inadequate in predicting the performance of a HVS-based numerical observer. It is not possible to define one single internal-noise equivalent observer that will match the HVS-based numerical observer for any image signal.

Training Readers to Use Multi-Level ROC Scores

Brandon D. Gallas (PhD), Qi Gong (MS), and Kyle J. Myers (PhD)

[1]*Division of Imaging, Diagnostics, and Software Reliability, OSEL/CDRH/FDA, Silver Spring, MD, USA*

Rationale

A multi-level diagnostic score is not universally part of a standard clinical report, but multi-level scores are required for ROC studies. Additionally, from an analysis point of view, it is generally better for readers to use many levels; fewer levels lead to ROC curves that are poorly sampled. In practice, clinicians can be uncomfortable reporting multi-level (ROC) scores and, consequently, researchers can be uncomfortable asking for them. In this presentation we will discuss reader training, data collection, and their impact on the quality of ROC scores, curves, and results.

Methods

The context for this work is an FDA project called VIPER, the validation of imaging premarket evaluation and regulation. VIPER investigates whether a significant difference in performance between two imaging modalities measured in a large prospective clinical trial can also be achieved in a much smaller, controlled, cancer-enriched lab study. To support this end and other related questions, we have designed a reader study to evaluate the impact of different study populations: we vary the prevalence of cancers (enrichment) and the distribution of non-cancers (screening population versus a stress test). The VIPER reader study compares the diagnostic ability of readers using full-field digital mammography (FFDM) and screen-film mammography (SFM) for the task of cancer detection in the subpopulation of women with dense breasts.

The reader study is quite large. There are five study conditions, each with 20 readers and more than 400 cases: a "low" (10%) prevalence screening study, a moderate (31%) prevalence screening study, a moderate (31%) prevalence stress test, a high (50%) prevalence screening study, and a high (50%) prevalence stress test. For reference, the prevalence observed in the DMIST trial, a prospective study, was about 0.5%-0.8%, depending on how the reference information is applied.

The reader training for the VIPER reader study included two documents. The first was a one-page description of the study purpose and the images (source, size, approximate prevalence, and distribution). The second was a three-page document of instructions for reporting the multi-level diagnostic score, i.e., the ROC score. The ROC data collection method was different from many other studies. It first asked for a clinical recall decision. If the reader decided "do not recall", they were then asked for a score between 1 and 100: 1 equaled "Most Normal" and 100 equaled "Least Normal". If the reader decided "recall", they were then asked for a score between 101 and 200: 101 equaled "Least Suspicious" and 200 equaled "Most Suspicious".

Results

Data collection is still ongoing, but preliminary results show that readers in the VIPER study were able to spread their ROC scores out over many levels, yielding ROC curves that were sampled well. We will discuss the impact of good ROC data on the study results.

Conclusions

Reader training and data collection can be designed and implemented to yield ROC curves that are sampled well. The associated time and cost to adequately train the readers are nearly insignificant next to the time and cost of executing the study.

A New Automated Evaluation Methodology of CT Image Quality Based on Contrast-Detail Measurements

Haney Alsleem[1](PhD), Robert Davidson[2](PhD)

[1]*Radiological Sciences, College of Applied Sciences,* University of Dammam
[2]*Head, Discipline of Medical Radiation Science,* Charles Sturt University

Rationale

The essential principle of maintaining lower radiation dose and optimum image quality is to understand the effects of exposure factors on image quality. The evaluation method of low contrast detail (LCD) detectability performance—particularly the automated approach—is a good choice for deep understanding the influences of exposure parameters on image quality. However, this method requires a certain specification of an LCD phantom and dedicated software that are not commercially available. The study aimed to develop a new methodology of evaluation and optimisation of computed tomography (CT) image quality based on LCD detectability performance.

Methodology

A new phantom was designed to obtain CT images of LCD. The specifications of the phantom design were optimised to satisfy the requirement of the new evaluation methodology of LCD detectability performance and based on evaluation of the limitations of available phantoms and the standard recommendations of phantom manufacturing. The phantom was manufactured with the cooperation of Artinis Medical Systems (Zetten, The Netherlands). A dedicated software was developed with the cooperation of Artinis Medical Systems to objectively evaluate the obtained CT images of the new phantom. The LCD detectability performance of CT images were measured by calculating the CT inverse image quality figure (CT IQF_{inv}). The new methodology was validated by determining the influences of exposure factors of kVp and mAs, slice thicknesses and objects location within the phantom on the image quality in terms of CT IQF_{inv} measurements. The validation was based on software and radiographers' scoring results.

Results

A new method of calculating the IQF_{inv} values for CT images, CT IQFinv, was developed based on the method of calculating the IQF_{inv} in digital radiography (Equation 1). A further requirement was the linear interpolation of the Hounsfield Units of the phantom's objects to account for both positive and negative contrast values.

$$\boldsymbol{IQFinv = \frac{100}{\sum_{i=0}^{8} Li \bullet Di,th}} \qquad \text{(Equation 1)}$$

where Li,th is threshold of the linear interpolation contrast values, and Di,th is threshold of detail sizes. CT IQFinv values were obtained objectively by the software and subjectively from radiographers' assessment. The results from radiographers and software showed that the new methodology of CT image quality assessment was sensitive to changing kVp, mAs and slice thicknesses.

Conclusion

The developed automated assessment methodology of LCD detectability performance in CT has the potential to effectively evaluate the effects of protocol parameters on image quality of different CT scanners and systems. The new phantom needs further improvement and the software should be also improved to increase the sensitivity and accuracy of their performance. Wider range of different kVp, mAs, slice thicknesses and other protocol parameters and different CT scanners should be also examined in future studies to ensure that the results conform to theory in a wider range of variables.

Computing contrast ratio in medical images using local content information

B. Ortiz-Jaramillo[1] (M.Eng.), A. Kumcu[1] (M.S.), L. Platisa[1] (PhD), W. Philips[1] (PhD)

[1]TELIN-IPI-iMinds, Ghent University

Rationale

Image quality assessment in medical applications is often based on quantifying the visibility between a structure of interest such as a vessel, termed foreground (F) and its surrounding anatomical background (B), i.e., the contrast ratio. A high quality image is the one that is able to make diagnostically relevant details distinguishable from the background. Therefore, the computation of contrast ratio is an important task in automatic medical image quality assessment.

Methods

We estimate the contrast ratio by using Weber's law in local image patches. A small image patch can contain a flat area, a textured area or an edge. Regions with edges are characterized by bimodal histograms representing B and F, and the local contrast ratio can be estimated using the ratio between mean intensity values of each mode of the histogram. B and F are identified by computing the mid-value between the modes using the ISODATA algorithm. This process is performed over the entire image with a sliding window resulting in a contrast ratio per pixel.

Results

We have tested our measure on two general purpose databases (TID2013 [1] and CSIQ [2]) to demonstrate that the proposed measure agrees with human preferences of quality. Since our measure is specifically designed for measuring contrast, only images exhibiting contrast changes are used. The difference between the maximum of the contrast ratios corresponding to the reference and processed images is used as a quality predictor. Human quality scores and our proposed measure are compared with the Pearson correlation coefficient. Our experimental results show that our method is able to accurately predict changes of perceived quality due to contrast decrements (Pearson correlations higher than 90%). Additionally, this method can detect changes in contrast level in interventional x-ray images acquired with varying dose [3]. For instance, the resulting contrast maps demonstrate reduced contrast ratios for vessel edges on X-ray images acquired at lower dose settings, i.e., lower distinguishability from the background, compared to higher dose acquisitions.

Conclusions

We propose a measure to compute contrast ratio by using Weber's law in local image patches. While the proposed contrast ratio is computationally simple, this approximation of local content has shown to be useful in measuring quality differences due to contrast decrements in images. Especially, changes in structures of interest due to low contrast ratio can be detected by using the contrast map making our method potentially useful in X-ray imaging dose control.

References

[1] Ponomarenko N. et al., "A New Color Image Database TID2013: Innovations and Results," Proceedings of ACIVS, 402-413 (2013).

[2] Larson E. and Chandler D., "Most apparent distortion: full-reference image quality assessment and the role of strategy," Journal of Electronic Imaging, 19 (1), 2010.

[3] Kumcu, A. et al., "Interventional x-ray image quality measure based on a psychovisual detectability model," MIPS XVI, Ghent, Belgium, 2015.

**This work was performed within the PANORAMA project (co-funded by grants from Belgium, Italy, France, the Netherlands, the United Kingdom, and the ENIAC Joint Undertaking).*

Automatic detection of collimation field in digital radiographic images

Vladimir Ostojic[1] (MSc), Djordje Starcevic[1] (MSc), Vladimir Petrovic[1] (PhD)

[1]*Telecomms and Signal Processing Group, Faculty of Technical Sciences, Novi Sad, Serbia*

Rationale

Use of collimators for narrowing the x-ray beam allows focusing radiation on a specific part of the anatomy and prevents excessive radiation from reaching the patient. However, as digital radiography detectors create image by detecting radiation on its entire surface, parts of detector surface shielded by collimator correspond to low-intensity background in resulting image. This redundant background causes diagnostic distraction and hampers signal processing algorithms, such as automatic tone scale adjustment. Manual cropping of directly irradiated section is time-consuming and impractical, so we propose an automatic method.

Methods

Directly irradiated field in most modern radiography images can be approximated as a rectangular region. Analysis of image gradient magnitude and orientation defines the set of line candidates. Our research showed that edges of the region as well as irradiated anatomy are efficiently depicted using Frobenius norm of Hessian (FNH). Algorithm proposed uses the fact that region edges are lines whose neighborhood contains large portion of surrounding FNH and which represent substantial pixel intensity and FNH changes. Edge candidates which fail to satisfy collimation edge properties are disregarded. Final edge set is chosen through the maximization of objective function parameterized by region pixel intensity, percentage of image FNH contained in the region, average FNH per region pixel and average FNH per edge.

Results

Algorithm was tested on the expert annotated database of 680 images of various anatomies, obtained using three different digital radiography detectors. It contains images with and without collimation, as well as with and without rotated collimation region. Algorithm's performance was evaluated using Dice and Tanimoto metrics, maximum absolute difference (MAD) between manually annotated and automatically determined collimation region and percentage of correct automatic annotations. Automatic annotation was considered correct if MAD was lower than a threshold defined as 5% of minimum collimation region dimension limited to the interval [0.5 cm, 1 cm]. Average Dice metric on the test base was 0.99 ± 0.04, average Tanimoto metric was 0.97 ± 0.05, average MAD was 4.76 ± 15.64 mm. Percentage of correctly annotated regions is 96.03 %.

Conclusions

We proposed a Frobenius norm of Hessian based algorithm for automatic collimation field detection. Our algorithm used estimated collimation field edge and region analysis to determine the best collimation edge candidates set. Algorithm's evaluation was performed through comparison to an expert annotated database.

Homomorphic Anti-scatter Grid Artefact Removal

Djordje Starcevic[1] (MSc), Vladimir Ostojic[1] (MSc), Vladimir Petrovic[1] (PhD)

[1]Telecomms and Signal Processing Group, Faculty of Technical Sciences, Novi Sad, Serbia

Rationale

X-ray imaging is the most frequent medical imaging method which suffers from artefacts originating from the Compton Effect which produces scattered x-ray radiation resulting in lowered image contrast. To counteract this effect, anti-scatter grids are used to prevent scattered radiation from reaching the detector, but in turn create an artefact of their own in the form of parallel stripes in the image. These artefacts have a distracting effect on the observer and reduce the fidelity of the image, so should be removed using digital image processing techniques.

Methods

Due to their periodic nature, grid artefacts can be represented in the form of Fourier series and thus removed using notch filters which eliminate appropriate harmonics. Multiplicative grid model paired with homomorphic filtering is used as it is shown to perform better than additive model [1]. The removal of the first harmonic alone was proven insufficient in cases where the second harmonic aliases to high frequencies. Therefore we propose an algorithm to adaptively estimate parameters of Gaussian notch filters used to remove the first two artefact harmonics. The parameters are estimated in 1D spectral domain while the filtering is performed in 2D spectral domain. Tests were performed on several detector types paired with a grid of 3.6 lp/mm frequency, which is sparse in respect to the detector pixel size.

Results

Results show that artefact removal by the proposed algorithm outperforms state of the art methods [1],[2] in cases where the second artefact harmonic aliases to high frequencies. Lower complexity version of the algorithm, which uses 1D removal of first two harmonics, was inadequate as it caused ringing artefacts and was dropped in favor of 2D removal. Furthermore, it was experimentally confirmed that homomorphic filtering gives better results than original image domain filtering.

Conclusions

We proposed a homomorphic Gaussian notch filtering algorithm for anti-scatter grid artefact removal. State of the art algorithms which remove only the first artefact harmonic were tested, proven unsatisfactory in case of high frequency aliasing of the second harmonic and outperformed by our algorithm which removes the first two artefact harmonics.

[1] Dong Sik Kim et al. “Grid artifact reduction based on homomorphic filtering in digital radiography imaging” Proc. SPIE 8668, Medical Imaging 2013: Physics of Medical Imaging, 86682C (March 6, 2013); doi:10.1117/12.2006760.

[2] Lin, Chih-Yang et al. “A Study of Grid Artifacts Formation and Elimination in Computed Radiographic Images.” Journal of Digital Imaging 19.4 (2006): 351–361. PMC. Web. 13 Jan. 2015.

Does binocular disparity impact the contrast sensitivity function?

Johanna Rousson[1,2] (MS), Jérémy Haar[1] (MS), Ljiljana Platiša[2] (PhD), Bastian Piepers[1] (MS), Tom Kimpe[1] (PhD), Wilfried Philips[2] (PhD)

[1] Barco NV, Healthcare Division, President Kennedypark 35, 8500 Kortrijk, Belgium
[2] IPI-TELIN-iMinds, Ghent University, St-Pietersnieuwstraat 41, 9000 Ghent, Belgium

Rationale

In order to ensure successful and reliable diagnosis, accurate calibration of medical displays is required. Typically, knowledge of the contrast sensitivity function (CSF) describing human eye ability to detect a low contrast pattern stimulus is crucial to develop calibration algorithms. Over the last decades the 2D CSF and its dependence to parameters such as the mean luminance, the stimulus size, and eye disorders have been intensively studied. Although 2D imaging remains more widespread than 3D imaging in diagnostic applications, 3D imaging systems are already being used and studies reveal that they could improve diagnostic performance. Nevertheless, very few studies have examined the CSF in stereoscopic viewing (hereafter 3D CSF). To know whether binocular disparities may impact the CSF, we investigated the relationship between the well-known 2D CSF and the 3D CSF.

Methods

Seventeen human observers tested for their normal visual acuity and stereovision participated into subjective experiments following a 3-down 1-up staircase. In the staircase experiment, the contrast of the stimulus was either decreased or increased depending on the observer's response to the preceding stimulus: target visible or target invisible. The stimuli were computer-generated stereoscopic images comprising a vertically oriented 2D Gabor patch as the target. The experiment was performed for seven different frequencies (0.4; 1; 1.8; 3; 4; 6.4; 10) expressed in cycles per degree (cpd), and two depth planes (the plane of the display, DP:0, and the depth plane lying 171 mm behind the display plane, DP:171). At DP:171 the spatial frequency was adapted to account for the increase in perceived viewing distance, and therefore to have constant spatial frequency across DPs. The stimuli were 1920x1200 pixel large images displayed on a 24 inch full HD stereoscopic surgical monitor using a patterned retarder. The experiments were conducted in a controlled environment with an ambient light of 0.8 lux.

Results

Computed medians and first and third quartiles as well as results of Friedman significant testing suggest that at low frequencies ($f \leq 1.8$ cpd) the CSF is significantly lower for DP:171 (3D CSF) than for DP:0 (2D CSF). However, at a frequency of 10 cpd the analysis indicated a significant improvement of the 2D contrast sensitivity (CS) compared to the 3D CS. For all the other spatial frequencies, the CS is not affected by the introduction of binocular disparities.

Conclusions

Differences in location of elements between the retinal images are likely to induce a loss in CS at low frequencies. As a consequence the suggested difference between the 2D CSF and the 3D CSF may have important implication in medical display market in the sense that new calibration algorithms would have to be developed for medical displays based on binocular disparity.

Interventional X-ray quality measure based on a psychovisual detectability model

A. Kumcu (MS), B. Ortiz-Jaramillo (MEng), L. Platisa (PhD), B. Goossens (PhD), W. Philips (PhD)

iMinds-TELIN-IPI, Ghent University, Ghent, Belgium

Rationale

Classical estimates of diagnostic performance – model observers – typically test subtle signals at threshold contrast perception. This approach may not be suitable for real-time quality assessment of medical imaging systems in which observers operate at suprathreshold contrast levels, such as interventional X-ray. Automatic dose control mechanisms for these systems adjust patient dose based on pre-determined patient thickness/dose curves and measurement of average gray levels in the acquisition [1], and may overestimate the dose needed to conduct the clinical task on a given patient or region. We present a real-time task-based quality measure that aims to estimate the minimum dose needed to obtain suprathreshold contrasts of target objects (vessels). This measure may be incorporated in a feedback loop for dose reduction while ensuring sufficient image quality for the clinical task.

Methods

The quality measure was built from two components: (1) a detectability function which models the clinical task of target detection, consisting of a set of psychometric functions that predict the detectability of vessel-like targets given a set of image features such as contrast and noise [2], and (2) an algorithm which measures these features on X-ray sequences. The psychometric functions were two-parameter psychometric Weibull functions fit to 1-down/1-up staircase results. Image parameters were varied to represent realistic image content as measured on interventional X-ray phantom and patient images. Background parameters were varied to correspond to changes in dose level: two levels of additive uncorrelated Gaussian noise including a noise-free background, and four background luminance levels. Sloan letters were used as targets. Test images were presented with static (still) and dynamic (25 frames per second) noise. The second component was an algorithm that estimated image features such as target contrast [3], background luminance, and noise variance. The detectability (0 to 100%) of image pixels was determined from the corresponding psychometric functions, given the measured image feature values. The quality measure was defined as the ratio of pixels with 100% detectability to the total number of detectable pixels. The change in contrast or noise, corresponding to a change in dose, needed to reach 99.5% detectability was also determined. The quality model was compared to a subjective quality study conducted on a chest phantom with contrast-filled cardiac arteries acquired at 12 dose levels on a Philips Allura interventional X-ray system.

Results

Preliminary results indicate that the model has a monotonic relationship with subjective quality preferences of interventional cardiologists. Further experiments are currently in progress.

Conclusions

We present an image quality measure which may be used in a real-time interventional X-ray dose control loop. While initial results are promising, further research is needed to validate the approach.

References

[1] Gislason, A. J., Hoornaert, B.; Davies, A. G. & Cowen, A. R., “Allura Xper Cardiac System Implementation of Automatic Dose Rate Control,” Philips Healthcare Technical Report, 2011

[2] Kumcu, A., Platiša, L., and Philips, W., "Effects of static and dynamic image noise and background luminance on letter contrast threshold", QoMEX, 2015

[3] Ortiz-Jaramillo, B., Kumcu, A., Platisa, L., and Philips, W., “Computing contrast ratio in medical images using local content information,” MIPS XVI, Ghent, Belgium, 2015

This work was performed within the PANORAMA project (co-funded by grants from Belgium, Italy, France, the Netherlands, the United Kingdom, and the ENIAC Joint Undertaking).

Comparing Diagnostic Image Reading Performance in Laboratory and Clinical Studies

Frank Samuelson[1] (PhD), Craig Abbey[2] (PhD), Xin He[1] (PhD)

[1]US Food and Drug Administration, Silver Spring, MD, USA
[2]University of California Santa Barbara, Santa Barbara, CA, USA

Rationale

Before using a new diagnostic imaging device in a clinical study on a large number of patients, it is common to perform controlled laboratory studies with a limited number of readers and cases comparing the new device to the standard of care. These studies measure the performance of a set of readers using the devices to diagnose a set of patients enriched with a high prevalence of diseased cases. After the acceptance of a new technology into widespread clinical use, observational studies of performance in the population are conducted, which measure the actual clinical performance. Ideally the pre-clinical laboratory study will predict the relative effects measured in the observational study. We want to know which measures of reader performance in such a pre-clinical study are predictive of future observational clinical studies.

Methods

Comparing performance measures in controlled laboratory studies and observational clinical studies can be difficult because these studies have very different disease prevalences, and in general the prevalences and false negative rates in the clinical studies are not known. To avoid this problem we use ratios of performance statistics between modalities within any one study, similar to the method of Baker and Pinsky (2001). We examine ratios or percent changes in false positive rates and true positive rates. Using simple assumptions and models, we also examine changes in expected utility and area under the ROC curve (AUC). In this presentation we examine data from several published laboratory and clinical studies of digital mammography systems, breast ultrasound systems, and a digital breast tomosynthesis system.

Results

We find that utilizing ratios is a very useful way to compare performance measures from multiple studies. We find that the percent changes in some performance metrics, such as expected utility and AUC are consistent across both laboratory and clinical studies. Changes in other metrics, such as the false positive rate and true positive rate can differ greatly between laboratory and clinical studies.

Conclusions

Using estimates of ratios or percent changes we can compare statistics among many pre-clinical laboratory studies and subsequent clinical observation studies. We conclude that percent changes in expected utility and AUC are reproducible measures, even across studies with greatly different prevalence, and therefore these performance measures have predictive power. Predicting changes in true positive, false positive, recall, or detection rates remains challenging.

Measuring the relative impact of various artifacts on the perceived quality of MR images

Hantao Liu[1] (PhD), Christine Cavaro-Ménard[2] (PhD), Jean-Yves Tanguy[3] (MD), Ken Hawick[1] (PhD)

[1]Department of Computer Science, University of Hull, Hull, United Kingdom
[2]Laboratory LARIS, University of Angers, Angers, France
[3]Department of Radiology, Angers Hospital, Angers, France

Rationale

Magnetic resonance (MR) imaging is vulnerable to various artifacts, which potentially cause inefficient and/or inaccurate diagnosis. These artifacts, in general, may be categorized as "structured" or "unstructured"; or classified as "white" or "colored". In current MR imaging systems, there are circumstances (e.g., the change of the acquisition bandwidth or the sequence parameters) whereby one type of artifact can be traded off with another type of artifact. Hence, to support these choices the relative impact of "structured" versus "unstructured" or "white" versus "colored" artifacts on perceived image quality needs to be investigated.

Methods

We stimulated four types of artifacts: a white unstructured artifact (i.e., white noise), a colored unstructured artifact (i.e., colored noise), a white structured artifact (i.e., edge ghosting) and a colored structured artifact (i.e., ghosting). They were varied at two different levels of energy in the signal distortion, and then linearly added to an original image. A set of eight high-quality MR images of different content was selected, and each was degraded with four types of artifacts once at a low energy level and once at a high energy level. We used four difference versions of colored noise applied at the same energy level, since its appearance is strongly affected by a dedicated randomization procedure needed for the simulation. As a result, the test database consisted of 112 stimuli (i.e., 8originals×2energy levels×7distortion versions). A perception experiment was conducted with thirteen radiologists rating the quality of test images, using a simultaneous-double-stimulus protocol.

Results

Preliminary results indicate that there is a significant difference in quality between the four types of artifacts, despite the fact that they are applied at exactly the same energy level. At a different aggregation level (as shown in Fig. 1), "spectral coloring" or "structuredness" is used as the classification variable; and each of these variables has a statistically significant effect on the perceived quality.

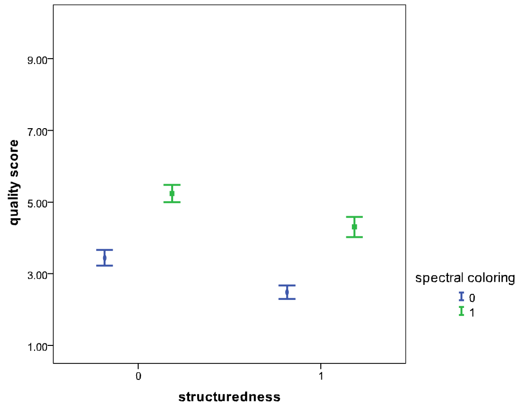


Fig. 1. Scatter plot of perceived quality for structuredness and spectral coloring of artifacts (averaged over all image content and the two energy levels). The "spectral coloring" refers to "1" for colored artifacts, and "0" for white artifacts. The "structuredness" makes "1" for structured artifacts, and "0" for unstructured artifacts.

Conclusions

This study suggests that different types of artifacts statistically significantly impact the perceived quality of MR images. Edge ghosting deteriorates the image quality most, followed by white noise, ghosting and colored noise. "Colored" artifacts deteriorate quality less than "white" artifacts, while "unstructured" artifacts deteriorate quality less than "structured" artifacts.

Expert detection and localisation in rapid presentation of mammograms

Ann J. Carrigan[1,2] (BAppSc (MRS); GradDipAppSc (Med US); BSc Psych (Hons)), Susan G. Wardle [1,2] (PhD), Claudia Mello-Thoms [3] (PhD), Anina N. Rich [1,2] (PhD).

[1] *Perception in Action Research Centre & Department of Cognitive Science, Macquarie University, Australia*
[2] *ARC Centre of Excellence in Cognition & Its Disorders, Macquarie University, Australia*
[3] *Medical Image Optimisation and Perception Group (MIOPeG), Discipline of Medical Imaging and Radiation Sciences, Faculty of Health Sciences, University of Sydney.*

Rationale

Experienced radiologists are able to detect lesions at above-chance levels when reading images such as mammograms in 200-250 ms[1,2]. Currently there are two models of radiologist visual search, both of which assume that the 'gist' of an image is used in diagnosis, but they make different predictions about what this wholistic information provides in terms of *localisation* of the abnormality. Here, we test the competing predictions from these models to understand the relationship between detection based on 'gist' and localisation of abnormalities in radiology.

Methods

Eighteen breast radiologists performed 96 trials viewing digital single left medio-lateral oblique breast mammograms (for 250 ms, 500 ms, or 750 ms). Half of the images contained a suspicious abnormality (BI-RADS 4 and above) and half were normal. They were asked whether the image was normal or contained a suspicious mass (detection) and to click with the mouse on the location on a subsequent outline of the image (localisation). We tested for (a) detection performance across durations; and (b) localisation performance when detection was correct.

Results

Detection performance: Radiologists were able to detect abnormalities at above chance levels even at our shortest duration (250 ms 58% correct; 500 ms 71% correct; 750 ms 66% correct). Measurements of sensitivity (*d* prime) (250 ms: 1.28; 500 ms: 1.67; 750 ms: 1.44) were above chance but did not vary across durations.
Localisation performance: On trials where radiologists correctly detected an abnormality, they could also correctly localise the mass on more than 60% of these trials across durations (250 ms: 60%; 500 ms: 65%; 750 ms: 66%); performance did not significantly improve across duration.

Conclusions

Experienced breast radiologists can extract lesion presence information about abnormalities in brief durations. We found that when a mass was correctly detected, observers also had accurate information about the location of that abnormality on at least 60% of the trials. The results inform our understanding of the information that experts extract from a mammogram in the earliest stages of visual processing, and may have implications for teaching in applied settings.

References

[1] Kundel, H.L. & Nodine, C.F. (1975). Interpreting chest radiographs without visual search. Radiology, 116, 527-532.
[2] Evans, K.K., Georgian-Smith, D., Tambouret, R., Birdwell, R.L., & Wolfe, J.M. (2013). The gist of the abnormal: Above-chance medical decision making in the blink of an eye. Psychonomic Bulletin Review, *20*, 1170 – 1175.

Computer aided FCD lesion detection based on T1 MRI data

Xiaoxia Qu[1,2](MS), Ljiljana Platiša[2] (PhD), Bart Goossens[2] (PhD),Tingzhu Bai[1](PhD), Karel Deblaere[3] (PhD) and Wilfried Philips[2](PhD)

[1]*School of Optoelectronics, Beijing Institute of Technology, Beijing, China*
[2]*IPI-TELIN-iMinds, Ghent University, Ghent, Belgium*
[3] *Ghent University Hospital, Department of Radiology, Ghent University, Ghent, Belgium*

Rationale

Focal cortical dysplasia (FCD) is a frequent cause of epilepsy and can be detected using brain magnetic resonance imaging (MRI). The FCD lesions in MRI images are characterized by blurring of the gray matter/white matter (GM/WM) junction, cortical thickening and hyper-intensity signal within lesional region compared with other cortical regions. However, detecting FCD lesions by means of visual inspection can be a very difficult task for radiologists because the lesions are very subtle. To assist physicians in detecting the FCD lesions more efficiently and reduce the false positive regions resulted from the existing methods [1], we propose an algorithm for automated FCD detection based on T1 MRI data.

Methods

The proposed computer aided diagnosis (CAD) technology mainly contains the measurement of difference in feature values of subject and healthy controls (DFSH) and classification using the Combination of Multiple classifiers (CMC) method, and is called DFSH-CMC. To increase the difference between FCD and non-FCD regions, the feature maps including gray matter thickness map, gradient map, relative intensity map and gray/white matter boundary width map are computed from the T1 MRI data of 41 subjects (10 patients and 31 healthy controls). The DFSH maps are measured after registering all data into a standard space. Different single (not combined) classifiers are applied for classifying voxels of MRI data into FCD (positive) or non-FCD (negative): naive Bayesian, linear discriminant analysis (DA), quadratic DA and Mahalanobis DA classifiers. To lower the number of false positive (FP) voxels resulted from the single classifiers, we utilize the CMC method to reclassify the voxels classified as positive by the single classifiers. Each subject is classified as a patient if the subject's image has voxels classified as FCD, otherwise, as a healthy control.

Results

The proposed method has correctly identified 8 out of 10 FCD patients and 30 out of 31 healthy controls. Compared to the feature maps, the DFSH maps are able to better differentiate between FCD and non-FCD regions. The single classifiers could correctly identify voxels within FCD regions as positive, but the number of FP voxels is large. Using CMC method, most of FP voxels resulting from the single classifiers are correctly reclassified as negative.

Conclusions

The proposed DFSH-CMC algorithm shows promise to become a valuable tool for automated detection of FCD lesions based on T1 MRI data. In future, the detection of FCD lesions using the multi-modal MRI data (e.g. fluid attenuated inversion recovery MRI and T1 weighted MRI) will be considered for further improving the FCD detection performance of the DFSH-CMC.

References

[1] Antel S. B., Collins D.L. and Bernasconi N. et al, “Automated detection of FCD lesions using computational models of their MRI characteristics and texture analysis”, NeuroImage, 19(4),1748-1759, 2003.

Eye-tracking search strategies for teaching breast ultrasound?

Ann Carrigan[1,2] (BAppSc (MRS); GradDip Medical Sonography; BSc Psych (Hons)), Patrick Brennan [3] (PhD), Mariusz Pietrzyk[3] (PhD), Jill Clarke [3](BAppSc (Hons), GradDip Medical Sonography, MHlthScEd), Eugene Chekaluk[4] (PhD)

[1]Perception in Action Research Centre & Department of Cognitive Science, Macquarie University, Australia
[2] ARC Centre of Excellence in Cognition & Its Disorders, Macquarie University, Australia
[3] Medical Image Optimisation and Perception Group (MIOPeG), Discipline of Medical Imaging and Radiation Sciences, Faculty of Health Sciences, University of Sydney.
[4]Department of Psychology, Macquarie University, Australia

Rationale

Today's university students are digital natives who experience and demand sophisticated electronic learning tools and facilities. University teachers are increasingly interested in those that provide effective pedagogical strategies. One such learning device could be developed from eye position tracking information that identifies preferred search strategies. Such an example was detected during a study of expert performance of medical sonographers [1] examining breast images.

Methods

The original study [1] recruited thirty participants who were tested before and after a 4 hour working session. Each studied fifty randomised breast ultrasound images that varied on whether a cancer was absent or present and on degree of detection difficulty. Some participants were eye-tracked, including the most experienced (>25 years) and the least experienced (3 months). Eye-tracked heat maps of these two subjects underwent thorough interpretive analysis using a qualitative analysis software package (NVivo10).

Results

Heat map analysis of the expert's eye positions showed almost exclusive attention on the breast cancer, as expected. The novice heat map however convincingly reflected the teaching strategies she had been exposed to. In breast ultrasound, teaching starts with the five anatomical breast 'layers': skin, subcutaneous tissue (with Cooper's ligaments), glandular tissue layer, retro-mammary layer and pectoral muscle. Students of diagnostic ultrasound are taught to examine each layer carefully for features of breast cancer, as whilst the primary site is usually within the glandular tissue, crucial diagnostic information can be gleaned from each layer.

Conclusions

This study suggests that a worthwhile pedagogical strategy for teaching breast ultrasound may result from students investigating eye-tracking heat maps demonstrating thorough search strategies, such as those found in the least experienced sonographer. This heat map will be compared to a typical breast ultrasound anatomy lesson and the implications of using eye position search data for pedagogical purposes will be explored.

References

[1] Carrigan, A., Chekaluk, E., Brennan, P., Clarke, J. & Pietrzyk, M. 2013, 'Expert performance of medical sonographers: an evaluation of physical and psychological variables ', Medical Imaging Perception Conference Abstract. Available: http://home.comcast.net/~eakmips/documents/MIPS-XV-2013Abstracts.pdf, The George Washington University, Washington DC.

A FROC study on the influence of breast thickness on simulated lesion detection

Salvagnini E.(M.S.)[1,3], Bosmans H.(PhD)[1,2], Van Ongeval C.(MD, PhD)[2], Van Steen A.(MD)[2], Michielsen K.(M.S.)[1], Cockmartin L.(M.S.)[2], Struelens L.(PhD)[3]. and Marshall N.W.(PhD)[2]

[1]KU Leuven, Department of Imaging and Pathology, Belgium;[2]UZ Leuven, Department of Radiology, Belgium; [3]SCK•CEN, Boeretang 200, Mol 2400, Belgium

Rationale

Theoretical indexes of detectability, such as signal-difference-to-noise ratio (SDNR) and threshold gold thickness, show a decrease of object detectability/visibility as test-object thickness increases due to the assessment of the automatic exposure control in mammographic systems. It is not proven whether clinical data present the same trend. The aim of this work was to investigate the influence of compressed breast thickness on lesion detectability for microcalcification clusters and masses.

Methods

520 cranio-caudal lesion-free mammograms were divided in 4 thickness groups (T-groups), T1 ≤ 29 mm, T2 = 30 - 49 mm, T3 = 50 - 69 mm, T4 ≥ 70 mm. Each group included 130 images in which 70 simulated lesions (35 microcalcification clusters and 35 masses) were inserted. Each lesion was inserted in four mammograms (one for each thickness group) which presented the same BI-RADS density score and an area with the same local glandularity percentage (g%) calculated with Volpara sofware. Keeping these characteristics the same is fundamental for minimizing the effect of background variability on the lesion detection, isolating in this way the effect of thickness. Four radiologists performed a free search study and rated their confidence using a five-point rating scale. Afterwards, a JAFROC analysis was applied. The alternative free-response receiver operating characteristic (AFROC) area for each T-group and the p-value together with the 95% confidence interval (CI) were calculated for all paired comparisons of thickness groups.

Results

Considering all lesions, the AFROC area decreases from 0.802 to 0.553 with increasing thickness from T1 to T3 while the area for T4 (0.565) was found almost equal to T3 (0.553). All p-values were found smaller than 0.05 and the 95% CI did not include zero except for the comparison of T3-T4 where the p-value was equal to 0.7757 and the 95% CI of the difference was [-0.0909, 0.0680]. Results of the groups T3 and T4 are considered not significantly different, while the decreasing trend from T1 to T3 is significant. The results for microcalcifications showed a similar trend: the AFROC area is reduced from 0.749 to 0.506 when breast thickness increases. The p-values and the 95% CI show no significant differences for the T1-T2 and T3-T4 comparisons (p-values > 0.05). Each of the other pairs of thickness groups was found significantly different (p-values < 0.0032). Results for masses showed a different trend, the T1 AFROC area (0.855) is larger than all the other T-groups. T2, T3 and T4 present similar AFROC areas, 0.609, 0.561 and 0.623 respectively. The p-values confirmed these results with values <0.05 when T1 is compared to all other groups while there is no significant difference for the other pairs of T-groups.

Conclusions

Lesion detectability decreases significantly when breast thickness increases. This effect is mostly visible on microcalcification clusters whose detection is known to be quantum noise limited. The increase of thickness means an increase of scatter and hence of noise. Detection of masses seems to be less affected by thickness increase. These results suggest that a different assessment of the mammographic exposure settings for thicker breasts could be set-up in order to improve lesion detection for these types of breast.

Prospects and challenges of eye tracking in volumetric images

Antje C. Venjakob

Technische Universität Berlin, Department of Psychology and Ergonomics, Chair of Human-Machine Systems

Rationale

While visual search, perception and cognition are relatively well studied phenomena regarding 2-dimensional images, a lot less is known about these in volumetric data. Eye tracking is a suitable method for the study of this, but it faces particular challenges that need to be addressed in the multi-slice context.

Methods

In the course of four experimental studies that used eye tracking to study visual search processes, perception and cognition in stack mode CT images, a number of specific challenges have been identified and reviewed.

Results

The analysis of factors that need to be addressed in novel ways when conducting eye tracking during the interpretation of multi-slice compared to single-slice images, yielded amongst others the following results:

- Incorporation of smooth pursuit eye movements into data analysis
- Definition of the size of AOIs, particularly for FP and TN decision sites
- Need for custom made software to map fixations to slices
- Integration of traditional eye tracking parameters with parameters that assess scrolling behavior
- Finding a suitable number of cases for an experiment
- Standardization of imaging material

Conclusions

The factors that have been identified will be presented, highlighting why they pose challenges to eye tracking research. Possible solutions to the imposed problems will be discussed. The talk will finish by highlighting the prospects that eye tracking in volumetric images offers and why the effort to overcome the challenges is warranted.

Where do neurologists look when analyzing neonatal EEG? An eye-tracking study

Marc Gibaud[1] (MD), Christine Cavaro-Ménard[2] (PhD), Sylvie Nguyen[1,2] (PhD-MD)

[1] *Department of Child Neurology, Angers Hospital, France*
[2] *Laboratory Laris, University of Angers, Angers, France*

Rationale

Electroencephalography in newborn infants has become valuable as a noninvasive screening tool for infants at high risk of perinatal injuries. Continuous electroencephalogram (EEG) provides useful information that reflects the function of the neonatal brain. The EEG may assist in determining brain maturation and identifying focal or generalized abnormalities. It is useful in assessing prognosis for neonates at risk for neurological sequelae.

But neonatal EEG differs significantly in many ways from EEG of older children and adults. There are numerous features that are age-specific and change almost week-to-week in the preterm infant. Some patterns may be normal at one age and abnormal if they persist for several weeks. This makes it difficult to interpret EEG results. In this study, we analyzed the visual behavior of experts during neonatal EEG interpretation using an eye-tracking system.

Methods

A set of 8 EEG traces of premature neonates was selected. We limited the analysis to the most relevant 10 minutes fragment of each 40 minutes EEG trace. On each EEG fragment, all specific patterns and wave-forms were detected, identified (pattern type), described (start and end of the pattern) and registered in a database, with the help of an expert in neonatal EEG analysis.

The 8 EEG fragments were interpreted by 7 experts in neonatal EEG analysis using an appropriate visualization software (as shown in Fig.1).

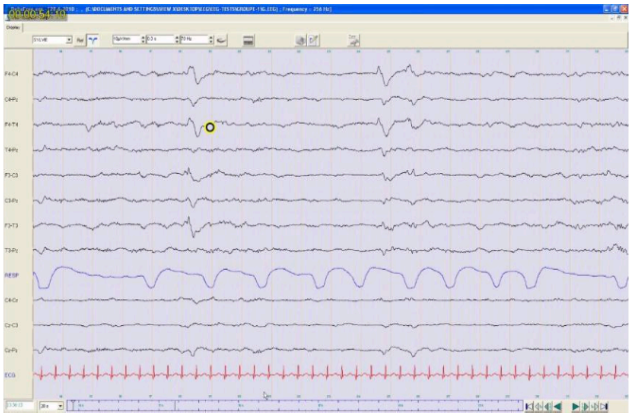


Fig 1: One 20s window of one neonatal EEG using the DeltaMed software

Eye tracking data were recorded and analyzed using a bottom-up approach (data driven using the fixation points) and a top-down one (knowledge driven using the patterns database).

Results

The results of the bottom-up analysis showed that the observers spent most of their time watching the center of the screen. Moreover, we noted the tendency towards a decrease in interpretation time when the trace was considered as normal.

The results of the top-down analysis showed that there is no direct correlation between the number of specific patterns seen and the relevance of the diagnosis. Moreover, experts spent more time on patterns located in the first half of the trace.

Conclusions

This study provides information on the visual behavior of experts during neonatal EEG analysis. This knowledge is critical to the design of EEG viewing devices. Moreover, perceptual feedback is beneficial for neurologist's training, a major challenge in neonatal EEG interpretation.

Quality of experience for remote surgery. A preliminary study for abdominal surgery

Lucie Lévêque[1] (MS), Christine Cavaro-Ménard[1] (PhD), Patrick Le Callet[2] (PhD), Emilie Lermite[3] (PhD-MD)

[1]Laboratory LARIS, University of Angers, Angers, France
[2]IRCCyN, UMR 6597, University of Nantes, France
[3]Department of Digestive Surgery, Angers Hospital, Angers, France

Rationale

Tele-assisted surgery (also known as “remote surgery”) involves an expert who gives real-time instructions to a less experienced surgeon, or the only available person on site, or even a robotic interface, in order to perform surgery on a patient even though they are not physically in the same location. Usually, it involves the transmission of the video of the operation scene and the video of the medical (anatomical, functional, physiological) data. Remote surgery should meet the same standards that traditional healthcare does in terms of efficiency, effectiveness and quality of clinical processes. All these requirements can be seen as elementary components participating in the quality of experience (QoE) in an ad hoc medical application scenario. This latter should not be assimilated to quality of service (QoS) as it clearly refers to the experience of the user, while QoS mostly focuses on the system. QoE studies should be conducted with an appropriate methodology incorporating user involvement and digital data and their relation with QoS.

Methods

To conduct our study, we recorded four abdominal open surgeries. We limited the analysis to the most relevant 20 second-fragment, creating 4 short videos referred to as “original videos”. To simulate transmission distortions, we used H264 video compression format, we changed bit rate and frame rate using a video converter (SUPER) and packets were randomly removed using a specific software. We selected 5 bit rates (1Mbps, 512kbps, 350kbps, 256kbps and 128kbps) and 2 frame rates (30fps and 15fps). The test database consisted of 32 distorted videos. Experiments were lead in two phases: a semi-structured interview - using MASK-I (Method for Analyzing and Structuring Knowledge) and CTA (Cognitive Task Analysis) methods - and a questionnaire on distorted videos, both with medical experts.

Results

According to the semi-structured interviews, remote surgery seems to be feasible if the expert knows enough information about the patient’s clinical case. Surgeons emphasized that latency in audio transmission between local surgeon and remote expert must be low enough to enable quick reactions. However, the latency threshold depends on the medical act, and the RTT (round-trip time) must be compatible with the performed act. Moreover, a video of the whole scene is not necessary throughout the course of the surgery. Quantitative results from questionnaires showed that below 350kbps, video quality is not sufficient to be used for tele-assistance and that a good color rendering appeared important to locate the organs.

Conclusions

QoE has a potential relevancy to optimize and understand the technical transmission chain from the final task viewpoint which is one of the most important factors to adopt telemedicine.

A Motion based Video Quality Metric for Cardiac Ultrasound Videos

Manzoor Razaak and Maria G. Martini

Wireless and Multimedia Networking research group, Kingston University, London

Rationale

Medical video quality evaluation is often done using state of the art video quality metrics which mainly assess the perceptual quality of the video without considering the aspect of diagnostic quality. A diagnostic-quality oriented medical video quality metric can be beneficial in medical video quality evaluation. One of the ways of achieving diagnostic quality oriented evaluation is developing content-aware video quality metrics. Towards this approach, we propose a full reference, content-aware video quality metric designed for quality evaluation of cardiac ultrasound videos. The proposed metric considers the motion information of each cardiac cycle of the ultrasound video to evaluate the diagnostic quality.

Materials and Methods

The Horn and Schunck optical flow method is used to estimate the magnitude of the motion vectors of each pixel of a frame of both the reference video and the impaired video. A weighted response for each pixel of a frame is obtained by applying a Gaussian weighting function to each pixel of a frame of the video using a window based approach. The mean squared error for each frame between the reference and the impaired video is computed for all the frames in a single cardiac cycle and represents the error index for that cardiac cycle. The maximum of the error indices from all the cardiac cycles is then used to represent the quality measure of the cardiac ultrasound video. The proposed metric was tested on 24 cardiac ultrasound videos, i.e. three videos compressed at eight different quality levels, using the High Efficiency Video Coding (HEVC) standard. Subjective quality evaluation of the videos was done by four medical experts. The proposed metric scores were correlated with the subjective scores of the medical experts. Further, the correlation scores of the proposed metric were compared with seven state-of-the-art video quality metrics.

Results

The results of our experiments found that the proposed metric shows consistently high correlation with the subjective scores. The proposed metric also outperforms most of the state-of-the-art metrics considered in our tests. The Pearson correlation and the Spearman correlation of the proposed metric with subjective scores of the medical experts were found to be 0.94 and 0.93 respectively. On the other hand, correlation scores of the popular metrics like SSIM, PSNR, and VQM were 0.93, 0.90, and 0.92 respectively (Pearson Correlation).

Conclusions

The proposed metric presents a diagnostic-quality oriented video quality metric for cardiac ultrasound videos. The results of our tests showed that the motion in cardiac ultrasound videos can be effectively utilized for a reliable objective evaluation of diagnostic quality. The approach of using specific characteristics of the medical video can enable design and development of more diagnostic-quality oriented video quality metrics.

Looking without seeing or not believing your eyes? An eye-tracking study on diagnosing X-rays

Laura Zwaan[1] (PhD), Indra Pieters[2] (MD, PhD), Daniel Schreij[3] (PhD), Abel Thijs[4] (MD, PhD), Jan Theeuwes[3] (PhD), Cornelis van Kuijk[2] (MD, PhD), Danielle Timmermans[1] (PhD), Artem V. Belopolsky[3] (PhD)

[1] VU University Medical Center, EMGO Institute for Health and Care Research, Amsterdam
[2] VU University Medical Center, Department of Radiology and Nuclear Medicine, Amsterdam
[3] Cognitive Psychology Department, VU University Amsterdam
[4] VU University Medical Center, Department of Internal Medicine, Amsterdam

Rationale

Diagnostic errors in radiology are not uncommon. In order to reduce diagnostic errors, it is important to determine the underlying causes. The goal of this eye-tracking study was to determine the influence of clinical information that matched or mismatched the disease on the X-ray on diagnostic performance in a realistic setting. Furthermore, the effect of the number of abnormalities on diagnostic performance was examined.

Methods

We tracked eye-movements of 25 radiologists while they examined 48 chest X-rays (12 without abnormalities, 24 with one abnormality and 12 with two abnormalities). The clinical information was manipulated such that it either matched or mismatched the abnormality on the X-ray.

All radiologists were presented with 12 match and 12 mismatch cases each having a single abnormality. In 12 cases in which there were two abnormalities, the radiologists were presented with clinical information that matched only one of them. Based on the fixation duration the errors were classified into search errors, recognition errors and decision errors.

Results

Matching clinical information more often led to a correctly reported abnormality than mismatching clinical information (66.5% versus 57.6% correct, $p<0.005$). Overall, the recognition errors were the most common error type (74.5%), followed by decision errors (20.4%) and search errors (5.1%).

There was a significant interaction between the number of abnormalities and the error type ($p<.05$). Specifically, more decision errors were made in the two abnormality cases (14.8%) than in the one abnormality cases (5.6%). The cases were not different in the number of recognition or search errors. There was also a significant interaction between the clinical information and the error type ($p<.05$). Specifically, more decision (7.3% vs. 13.1% for match and mismatch, respectively) and recognition errors (31.9% vs 42.5% for match and mismatch, respectively) were made when the clinical information did not match the abnormality.

Conclusions

The results show that diagnostic errors are primarily due to a failure in recognizing the abnormality that is being fixated. This is consistent with inattentional blindness. Mismatching clinical information increases such recognition failures. The number of abnormalities had the strongest effect on how a diagnostic decision was reached based on the available information. Specifically, the presence of a second abnormality lead to discarding clinically relevant information: the radiologist ”did not believe his own eyes”.

Interventional Images on the iPad

Rachel Toomey[1] (PhD), David Leong[2] (MSc), Michael Evanoff[3] (PhD), John Ryan (PhD)[1,4], Eoin Kavanagh[4] (MB BCh BAO FRCPI FFR RCSI), Fracis Zarb (PhD) [5], Jonathan McNulty (PhD)[1], Louise A Rainford[1] (PhD)

[1]School of Medicine and Medical Science, University College Dublin, Dublin 4, Ireland
[3]Analogic, 8 Centennial Drive, Peabody, MA 01960, U.S.A.
[3]American Board of Radiology, 5441 E. Williams Circle, Tucson, Arizona, U.S.A.
[4]Ziltron Ltd., Limerick, Ireland
[5]Department of Radiology, Mater Misericordiae University Hospital, Eccles St., Dublin, Ireland
[6]Department of Radiography, Faculty of Health Sciences, University of Malta, Malta

Rationale

A previous research study by our group into the suitability of the iPad 3rd Generation (Apple, Cupertino, CA) for American Board of Radiology examinations[1] prompted further study into the preferences of, and perception of image quality by, vascular/interventional (VIR) radiologists. This study aimed to examine whether these radiologists preferred the iPad or a secondary class monitor for different cases, and what factors influenced these choices.

Methods

Seventeen examining radiologists with the American Board of Radiology each reviewed twenty VIR examination cases side-by-side on an iPad and a calibrated secondary class LCD monitor (ViewSonic, Walnut, VA). The iPad had higher max and min luminances, although the luminance ratios of the devices were very similar. Participants rated the visibility of the relevant image features on a scale of 1 (unacceptably poor) to 5 (excellent). They also stated which display (if any) they thought performed better for each case and why, and simple thematic analysis was applied.

Results

Participants ranked visualization of pertinent image features highly for both devices overall, with no statistically significant differences noted ($p=0.58$). Thematic analysis of comments showed that the devices were most commonly considered equally useful. However, where a preference was expressed, the iPad received a large majority of favorable comments concerning spatial resolution and contrast resolution / brightness, and the monitor in terms of display size.

Conclusions

The iPad was well received, although the smaller display caused problems for some cases where multiple images were displayed side by side. A large majority of positive comments concerning luminance characteristics related to the iPad, despite its very similar luminance ratio to the monitor and lack of DICOM calibration. This may have implications for clinical image presentation.

References

[1] Toomey RJ, Rainford LA, Leong DL, Butler ML, Evanoff MG, Kavanagh EC, Ryan JT. "Is the iPad suitable for image display at American Board of Radiology examinations?" AJR:203(5):1028-33 (2014).

Variability in radiologists' diagnoses and representation of pathologies

Rachel Toomey[1] (PhD), Joanna Lowe[1] (BSc), Michael Evanoff[2] (PhD), Eoin Kavanagh[1,3] (MB BCh BAO FRCPI FFR RCSI), Louise A Rainford[1] (PhD)

[1]School of Medicine and Medical Science, University College Dublin, Belfield, Dublin 4, Ireland
[2]American Board of Radiology, 5441 E. Williams Circle, Tucson, Arizona, U.S.A.
[3]Department of Radiology, Mater Misericordiae University Hospital, Dublin, Ireland

Rationale

Variability between physicians has been reported in radiological diagnosis of various image types, including chest radiographs. This study aimed to investigate variations in diagnosis, and to try to understand the source of this variation by identifying the features radiologists considered most important in diagnostic decision making. Furthermore, the study employed a novel methodology, with the hope that examining how radiologists represented pathologies and relevant image features might yield some information about their interpretative process.

Methods

Twenty-two board-certified radiologists were each provided with an iPad with fifteen chest radiographs, a pencil and eraser, and a printed booklet. The booklet contained the same fifteen images, but with the anatomy within the ribs removed completely to allow space for drawing.Participants viewed each image and were asked to summarise their diagnosis, and draw/sketch the features that allowed them to make that diagnosis, on the corresponding page of the booklet. The diagnoses and characteristics of the drawings across participants were then compared.

Results

Data analysis is ongoing at the time of writing. Preliminary analysis showed substantial variation in the diagnoses made for some images. Drawings also varied significantly between radiologists for many images [Fig 1]; however, drawings associated with certain diagnoses (regardless of which image the diagnosis was asociated with) were sometimes very similar. For example, pulmonary nodules or masses were almost always represented simply by an empty circle in the relevant area of the chest, with little attempt to replicate the actual shape of the mass or other detail.

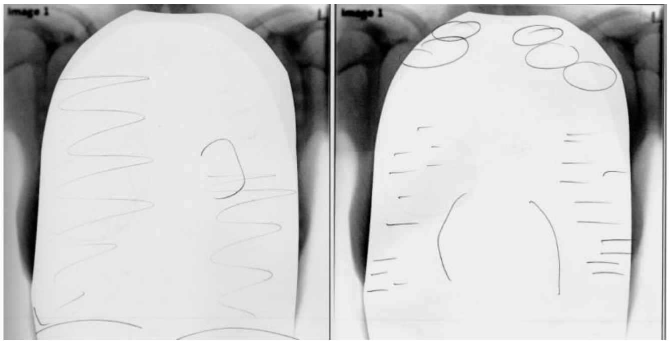


Fig 1. Two representations of the same image

Conclusions

This study adds to literature describing inter-radiologist variability. Further research into the perceptual processes that lead to varying diagnoses from the same image is encouraged. The drawings, although not very specific, provided an interesting insight into how radiologists represented pathologies.

3D printed lung phantom for clinical image quality assessment

J. Michiel den Harder[1] (PhD), Irene Hernandez-Giron[2,3] (MSc), Alfonso Calzado[3] (PhD),
Jacob Geleijns[1] (PhD), Wouter J.H. Veldkamp[1] (PhD)

[1]Radiology department, Leiden University Medical Center, Leiden, the Netherlands
[2]Física Mèdica, Universitat Rovira i Virgili, Tarragona, Spain
[3]Departamento de Radiología, Universidad Complutense de Madrid, Madrid, Spain

Rationale

Traditionally, quality assessment of diagnostic imaging systems in radiology aims at verification of technical parameters such as resolution and signal-to-noise ratio. Recent efforts focus on assessing whether the image quality is well optimized for the intended clinical task [1]. Anthropomorphic phantoms can be used to mimic clinical settings [2]. Here, we present an inexpensive, reproducible, and easy to handle 3D printed phantom mimicking a blood vessel structure of a human lung. 3D printed spheres can be inserted to mimic nodules. Alternatively, image data of nodules may be isolated and inserted elsewhere in the image retrospectively, either in Radon or in Image space.

Methods

A 3D blood vessel structure was designed using a custom-made algorithm in MATLAB® (MathWorks®, Natick, MA), converted to an STL file using 3D Slicer [3] and printed on a ProJet™ 3D printer (3D Systems, Rock Hill, SC). Additionally, spherical nodules of different diameters (5-8-10 mm) were printed. Scans of the phantom with and without inserted nodules were made on a Toshiba Aquilion 16 CT scanner (Toshiba Medical Systems, Tokyo, Japan) using an isotropic resolution of 5 mm. The phantom images were visually compared to images of a patient's lung. Subtraction images of the phantom with and without an inserted nodule were obtained, both in Radon and Image space.

Results

The phantom images had similar appearance as the images of the patient's lung. The noise level within the nodule (5.0 HU) was comparable to the noise level at the same location while the nodule was absent (5.5 HU). Subtraction of the phantom images with and without the inserted nodule resulted in a clear visualization of the nodule only, apart from minor registration errors. Performing the subtraction in Radon or in Image space resulted in a similar representation of the nodule.

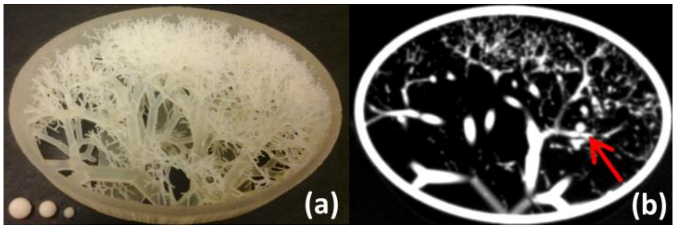


Figure 1: (a) Picture of the 3D printed anthropomorphic lung phantom and nodules and (b) CT scan of the phantom with inserted 5 mm nodule (arrow).

Conclusions

We have presented an inexpensive, reproducible, and easy to handle 3D printed anthropomorphic lung phantom for assessment of clinical image quality. The successful subtraction of images of the phantom with and without the inserted nodule and the fact that the noise level was independent on the presence of the nodule suggests that retrospective insertion of nodules may be realistic. Near future investigations will include using other materials for mimicking nodules.

References

[1] Hernandez e.a., BJR2014;87:20140014.
[2] Solomon e.a., SPIE 9033;doi:10.1117/12.2043555.
[3] Fedorov e.a., MRI2012 Nov;30(9):1323-41, www.slicer.org.

Variation in Paediatric CT dose Distribution in some Nigerian Tertiary Health Institutions and its Radiological Implications

C.A. Aborisade[1] (PhD), F.A. Balogun[2](PhD), C.O. Famurewa[3] (MBChB)

[1]Department of Physics, Obafemi Awolowo University, Ile-Ife. Nigeria
[2]Center for Energy Research and Development, OAU, Ile-Ife. Nigeria
[3]Deparment of Radiology, Obafemi Awolowo University Teaching Hospital Ile-Ife. Nigeria

Rationale

The paediatrics form a critical group in radiation exposure and any procedure that may lead to reduction exposure must be critically studied. In this study we look at the pattern of exposure of children undergoing CT procedures in some of our tertiary institutions. Three tertiary hospitals, Obafemi Awolowo University Teaching Hospital Complex, (OAUTHC), Ile-Ife, Lagos University Teaching Hospital, (LUTH), Lagos and University of Ilorin Teaching Hospital (UITH), Ilorin were studied.

Methods

An important aspect of this research work is to compute the effective dose from the equivalent dose obtained from the hospitals using the thermoluminescent dosimeters (TLD) placed on the paediatric patients during CT examination, subsequently the cancer risk associated with the procedure was computed using the lifetime attributable cancer mortality risks per unit dose as function of age at a single acute exposure as estimated by the National Academic of Science BEIR V (Biological Effects of Ionizing Radiations) committee, [1]. An average of 158 paediatric CT examination were estimated from each hospital.

Results

The values of the equivalent dose (mSv) for abdominal CT scan from OAUTHC ranged from 23.49 - 55. 25, with a mean of 37.52±14.36; head CT scan ranged from 10.07 – 69.94, with a mean of 39.99±13.77 and chest CT scan ranged from 8.60 – 31.94 with a mean of 14.51±12.63. The values of the equivalent dose (mSv) for abdominal CT scan from UITH ranged from 14.92 – 32.42, with a mean of 23.21±7.52; head CT scan ranged from 11.11 – 45.69, with a mean of 37.72±12.74 and chest CT scan ranged from 7.61 – 21.64 with a mean of 12.71±4.53. The collection of data from LUTH is in progress. The result of the cancer risk estimate showed that the paediatric patients who did abdominal CT scan had the highest cancer risk ranging from digestive cancer 357 per thousand patients to lung cancer risk of 4 per thousand patients, followed by the female patients who did the chest CT scan ranging from breast 97 per thousand patients to digestive cancer risk of 2 per thousand patients. The estimated cancer risk from patients who had skull CT scans ranging from leukemia 70 per thousand patients to lung 1 per thousand patients. The result also showed that the risk for the digestive, leukemia and lung is independent of the age while the risk for breast increases from 6 years old to 15 years.

Conclusions

The dose and procedures varies widely from one hospital to another. The study concluded that there was an urgent need for standardization of procedures in CT paediatric radiology.

References

[1] Brenner J.D, Eliston C.D., Hall E.J and Berdon W.E. Estimated Risks of Radiation-Induced Fatal Cancer from Pediatric CT, AJR 2001; 176; 289-296.

Training data selection for machine learning based model observer

Iris Lorente (M.S.), Jovan G. Brankov (Ph.D.)
ECE department, Illinois Institute of Technology, Chicago, IL

Rationale

The use of model observers (MO) has been of great importance in assessment of medical image quality by reducing the number of human studies, which are costly and time-consuming. Some types of MO require tuning the model by using a set of images previously evaluated by humans. Ideally, one would like to undergo the minimum number of human studies while obtaining consistent and generalized model parameters. In this paper, we explore two methodologies to select a small subset of images to perform human study and be used to train a learning model observer (LMO) based on the *Relevance Vector Machine* (RVM).

Methods

In this work, we use a human observer study from a Monte Carlo simulated SPECT myocardial perfusion acquisition. Projections were reconstructed with an OSEM algorithm (one, five and ten iterations steps) and post-reconstruction low-pass filtered by 3-D Gaussian filters using six different cutoff frequencies. So in total we had 18 reconstruction strategies. For each reconstruction strategy, 25 noise realization with defect and 25 without defect were simulated. Six human expert (physician) observers evaluated perfusion defect visibility in each image by giving a discrete confidence rating from 1 (defect definitely absent) to 6 (defect definitely present) in a SKE/LKE environment. As a MO, an RVM is trained for each human using 4 channel features. We explore two approaches to select the reconstruction strategies set of images to be used for training the LMO. The first is an active learning (AL) method based on adding to the training set, in each iteration, the reconstruction strategy set of images whose averaged posterior predictive variance is minimum. This selection approach aims to expand fast the predictive range since the samples that are outside of it have minimum variance due to specifics of RVM methodology. The second approach is based on finding the subset of strategies whose pooled images, viewed as a feature-space distribution, are the most 'similar' to the feature-space of the pooled remaining images. This 'similarity' is measured by a combination of Frechet distances between training-testing feature space and training-testing feature space with or without defect. Note that this methods does not need human observer data.

Results

To evaluate the performance of the proposed methods we established a baseline using exhaustive random selection and report average AUC prediction error, in Fig 1, as well as the error bars indicate median and 95% interval of confidence as a function of the number of selected reconstruction strategy in training. As an initial training set for AL we selected the reconstruction strategy whose parameters are in the middle of the reconstruction parameters space. Fig 1 shows that both explored approaches belong to the low tail of the random selection meaning that there is better agreement between human observer and MO measured by the AUC.

Conclusions

The results indicate that both methodologies can reduce the needed number of human studies while still providing an accurate model observer. The feature space based selection makes no use of human information in all the selection process and adds, slight, improvement in performance with respect to AL. Individual image selection strategies will be explored by the time of the conference.

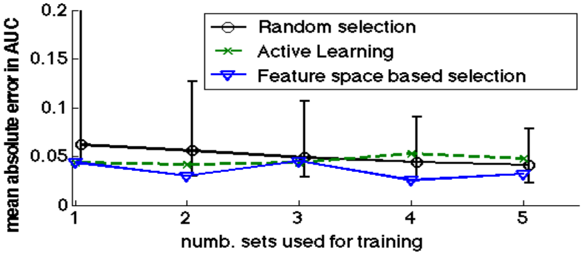


Fig. 1. Comparison between random selection and proposed methodologies. The performance of the MO is measured by the averaged difference in AUC between the human and its MO. Error bars indicate median and 95% interval of confidence.

Model observer performance in detection tasks in mammography: 2D vs reconstructed planes in tomosynthesis

Irene Hernández-Girón[1] (MSc), Margarita Chevalier[2] (PhD), Maria Castillo[2] (MSc), Julio Valverde Morán[3], Julia Garayoa[3] (PhD)

[1]Radiology department, Leiden University Medical Center, Leiden, the Netherlands
[2]Departamento de Radiología, Universidad Complutense de Madrid (UCM), Spain
[3]Servicio de Protección Radiológica, Hospital Universitario Fundación Jiménez Díaz, Madrid

Rationale

Digital Breast Tomosynthesis (DBT) is becoming more relevant due to the results of clinical studies showing its ability to improve breast lesion detection. Different manufacturers have commercialized DBT systems that present important differences regarding the acquisition geometry, reconstruction algorithms as well as detector technologies. The final DBT image quality is the result of a high number of variables which makes it difficult to define optimal conditions that would be valid for all the systems. Thus, it is important to develop objective methods to analyze the image quality.

Methods

Images of the TORMAM phantom (Leeds Test Objects Ltd, Boroughbridge, UK) were acquired for this study. The phantom is 15 mm thick and includes several test objects inmersed in an uniform background.The relevant objects for this study were 18 low contrast objects (3 mm diameter) with different nominal contrasts (0.5, 1, 1.5, 2, 3, 4). Images of the plate placed on top of 3 cm of PMMA were acquired with two systems: Hologic Selenia Dimensions and Fujifilm Amulet Innovality. Twenty 2D and 3D phantom images were acquired using the automatic exposure control. The settings for 2D were W/Rh, 29 kV, 67 and 80 mAs in both systems and for 3D they were W/Al, 32 kV, 32 mAs and 31 kV and 50 mAs respectively. Contrast to noise (CNR) was mesured on each object for the 2D images and for the reconstructed plane on focus and averaged for the 20 acquisitions. A non-prewhitening matched filter model observer (NPW) was used to analyze the detectability of the objects as a function of nominal contrast. This model cross-correlates a template with samples extracted from the images with object present or absent. Detectability indexes (d') were also calculated. CNR and d' values were analyzed for each manufacturer comparing the values in the 2D images and the reconstructed planes.

Results

CNR increased in the reconstructed planes images compared to 2D for the six contrast levels and both systems. The improvement in CNR was between 1.8 and 2.2 times (system 1) and between 2.4 and 2.8 times (system 2). The NPW d' values showed also the same trend for both systems (Fig 1) with an improvement in the objects detection as contrast increased and of the reconstructed planes images vs 2D mammograms.

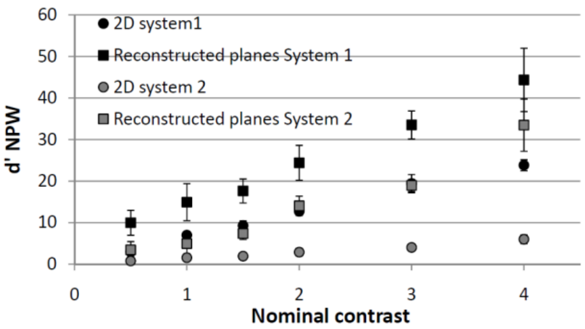


Figure 1. Detectability index (d') of the NPW model as a function of object contrast for 2D and reconstructed planes

Conclusions

The detectability of the low contrast objects in the TORMAM phantom, based on a model observer, was improved in the reconstructed planes compared to the 2D mammograms in two different DBT systems. Further research has to be made with anthropomorphic phantoms, as the objects studied in this work do not represent realistically breast tissue or patient lesions.

Thermography for breast cancer detection: basis, methods, and human/computer performance

Murray H. Loew (PhD), Mahsa Alborz (MS), Shijian Fan (MS), Sragvi Tirumala

Department of Biomedical Engineering, George Washington University, Washington, D.C. USA

Rationale

After several decades of effort to make thermography a reliable adjunctive modality in breast cancer screening and/or diagnosis, researchers still have been unable to demonstrate real clinical value of the method, even as software and camera technology improved. Recent work [1,2] has provided new physiological/mechanical models and a theoretical basis for making quantitative inferences about tumors at depth, using thermograms acquired under realistic conditions of pose and thermal environment. The present study aims to test whether those inferences can be made accurately and practicably in a clinical setting.

Methods

An infrared camera (8-12 μ, ~ 30 mK resolution, 640 x 480 pixels) will be used to image pairs of breasts of women volunteers who are visiting the breast clinic for routine screening or for recall. Image acquisition will be conducted under controlled ambient conditions and after stabilization of skin surface temperature. The images will be analyzed, with no manual intervention, using measures of ipsilateral localized temperature increase and by comparison to a corresponding region on the contralateral breast. Those measures, in combination with other image features (texture, shape) will yield a computed index of abnormality that can be compared to truth as provided by mammography, other imaging methods, or biopsy. The images also will be displayed in grayscale and in color for parallel evaluation by radiologists.

Results

The pilot study is ongoing and results are expected in spring 2015. In addition to estimating the performance (sensitivity, specificity, ROC statistics) of the automated method, the study will examine the effect of alternative color mappings on human-observer interpretations.

Conclusions

Our results will be compared to those of other recent investigations, with respect to both performance and degree of manual intervention.

References

[1] L. Jiang, W. Zhan, and M. H. Loew, "Modeling static and dynamic thermography of the human breast under elastic deformation," *Phys. Med. Biol*., Vol. 56, No.1, 2011, pp. 187-202.

[2] L. Jiang, W. Zhan, and M. H. Loew, "Toward understanding the complex mechanisms behind breast thermography: an overview for comprehensive numerical study," *Proc. SPIE* 7965, Medical Imaging 2011, 79650H (2011).

Interpretation of Multireader multicase analysis parameters, with applications to simulation models and sample size estimation

Stephen L. Hillis[1] (PhD)

[1] *Department of Radiology, University of Iowa, Iowa City, Iowa, U.S.A.*

Rationale

The Obuchowski-Rockette (OR) and the related Dorfman-Berbaum-Metz (DBM) methods have been the most frequently used analysis method for multireader multicase (MRMC) radiologic imaging data, with the DBM method being a special case of the OR method. The OR method variance component and correlation parameters have meaningful interpretations and yield important information about the structure of the data. Understanding the meaning of these parameters is useful for designing and sizing future studies. Furthermore, the ability to design simulation studies that emulate real-data studies is necessary for empirical validation of the OR, DBM, and other MRMC analysis methods. Specifically, it is important that simulated data result in OR parameter estimates similar to those encountered with real data.

Methods

I show how meaningful interpretation of the OR parameters follows directly from the OR model. I then present formulas that express the OR parameters in terms of the parameters of the Roe and Metz (RM) model for simulating MRMC data. These formulas allow us to determine the OR parameter values that correspond to each of the RM model simulation combinations. As a result, we can (1) determine how realistic the RM model is; (2) investigate the relationship between OR parameters and number of cases, number of readers, type of ROC curve, and other RM parameters; and (3) calibrate the RM model to have corresponding OR parameters that match real data sets.

Results

The OR correlations and reader and reader-by-modality variance components are essentially independent of case and reader sample size. In addition, the reader and reader-by-modality variance components are also essentially independent of the estimation method (e.g., the variance components will be similar for MLE and trapezoid AUC outcomes.) The error variance is dependent on case sample size but not reader sample size; it is also affected by the type of estimation. Based on these results, I discussed strategies for setting the values of the OR components for sizing future studies. Some of the RM model simulation combinations show unrealistically high reader variance, and an error in the covariance formula in the Roe and Metz paper is revealed that often results in an unnatural ordering of the covariances.

Conclusions

Formulas defining the relationship between OR parameters and RM simulation parameters were presented. These formulas reveal how OR parameters are effected various factors (e.g., reader and case sample sizes), which in turn suggest strategies for setting the values of the OR components for sizing future studies. In addition, the formulas showed there is a need for recalibration of the RM model to make it more in agreement with real data studies.

Observer performance estimation in digital breast tomosynthesis based on analysis of curvature features

Frank Schebesch[1] , Wei Wei[2], Anna Jerebko[2] , Michael Kelm[2], Lesley Cockmartin[3], Hilde Bosmans[3], Andreas Maier[1], Joachim Hornegger[1], Thomas Mertelmeier[2]

[1] Pattern Recognition Lab, Department of Computer Science, University of Erlangen-Nuremberg, Germany
[2]Healthcare, Siemens AG, Erlangen, Germany
[3]UZ Leuven, Department of Radiology, Belgium

Rationale

The detection of lesions in medical images by a human observer involves a combination of visual perception and highly trained interpretation capabilities of the human brain. Common model observers are using a template-matching approach that can be tuned to predict human task-based performance as measured by psychophysical tests. However, a correct implementation needs a lot of training data and specific local ground truth knowledge. In many technical evaluations a preliminary estimation of the expected observer performance appears to be beneficial, e.g. to compare different reconstruction algorithms and image acquisition parameters. Our goal is to create a method for an automated task-based evaluation of image quality for digital breast tomosynthesis applicable to a variety of reconstruction algorithms and acquisition schemes without the need to retrain or tune the model observers.

Methods

A tomosynthesis slice of a phantom image is filtered to enhance areas of high lesion localization probability and a NPW model observer is applied on the filtered and unfiltered image. Instead of using a template trained to a specific lesion structure, the observer is based on a circular calcification template. For the filter the second order derivatives of scale-space representations of the image are analyzed using the Determinant of Hessian (DoH) method and Eigen-analysis of the Hessian both based on Frangi's vesselness filter [1]. The phantom consists of a breast-shaped acrylic container of thickness 48 mm filled with water and acrylic spheres of different diameters [2]. Microcalcifications of diameter 90 µm to 250 µm are added as targets. Images were acquired on a Siemens Mammomat Inspiration Tomosynthesis system in DBT mode under automatic exposure control (AEC) and at half and double AEC dose and reconstructed with two different algorithms. The microcalcifications are assumed to be close-to-circular shaped in the lesion localization probability map of the phantom slice produced by the filter. An ROC analysis is performed on three sets of each acquisition and algorithm. The separability of the observer response is tested by splitting the image into two areas corresponding to positives and negatives. For evaluation four experienced human readers rated their confidence on the detection of the microcalcifications on three randomly chosen image series for each acquisition mode and algorithm combination.

Results

The comparison of the ratings between humans and the proposed model observer indicates that the estimated observer performance is sensitive to the qualitative differences in dose variations and reconstruction algorithms similar to human preferences. The response of the lesion template applied to the filtered image yields an average AUC scoring that correlates by 0.93 ($p < 0.01$) with the averaged estimation of the human observers. In contrast, the ROC analysis on the filter output alone yielded a correlation of 0.82 ($p < 0.05$) and the template response on the original image 0.85 ($p < 0.04$).

Conclusions

This study gives an indication that the combination of linear model observers with a structural template and a curvature filter provides better differentiation between true and false positives with respect to detectability of lesions than either of these methods alone. The results show that this approach correlates well with human observer estimations of image quality in images showing groups of microcalcifications. Further evaluation with different templates, lesion types and different physical phantoms will clarify how sensitive this general approach is to lesion shape and size.

References

[1] Alejandro F. Frangi, Wiro J. Niessen, Koen L. Vincken, and Max A. Viergever. Multiscale vessel enhancement filtering. In William M. Wells, Alan Colchester, and Scott Delp, editors, Medical Image Computing and Computer-Assisted Intervention – MICCAI'98, volume 1496 of Lecture Notes in Computer Science, pages 130–137. Springer Berlin Heidelberg, 1998.

[2] L. Cockmartin, H. Bosmans, and N. W. Marshall. Comparative power law analysis of structured breast phantom and patient images in digital mammography and breast tomosynthesis. Medical Physics, 40(8):081920–1–17, 2013.

Detectability of malignant mass models in 2D mammograms, differentiated towards size and background glandularity

Alaleh Rashidnasab[1] (PhD), Frédéric Bemelmans[2] (M.S.), Elena Salvagnini[1] (M.S.),
Nicholas W. Marshall[2] (M.S.) and Hilde Bosmans[1,2](PhD)

[1]Department of Imaging and Pathology, KU Leuven, Herestraat 49, 3000 Leuven, Belgium
[2]Department of Radiology, UZ Leuven, Herestraat 49, 3000 Leuven, Belgium

Rationale

To compare the detectability of breast cancers masses with ill-defined margins and spiculations as a function of their in plane size, thickness and glandularity of the background in 2D-mammography by means of a focused 4-alternative forced choice (4AFC) paradigm and human observers.

Methods

Regions of interest (ROI) were extracted from normal digital mammograms and their glandularity was measured using Volpara. The selection included 60 ROIs with low glandularity (<15%), 60 ROIs with medium glandularity (15%-30%) and 20 ROIs with high glandularity (>30%). Simulated 3D breast mass models, validated for realistic appearance in 2D mammography, were used for insertion into the ROIs. Five ill-defined margin masses and five spiculated masses with original outer size of ~8-11mm in x-y plane were selected. The masses were digitally scaled in all 3 dimensions to create masses of ~6-8 mm, ~4-6 mm, ~2-3 mm in-plane size. Thickness of the core of the mass (size in z direction) was calculated for all scaled masses. A validated physics-based insertion framework was used that accounts for polychromatic X-ray spectra, local glandularity, MTF and scatter of the imaging system being modelled (i.e. Mammomat Inspiration, Siemens, Germany). For each ROI, one ill-defined margin mass and one spiculated mass were randomly selected and inserted into the ROI. Insertion in the same ROI was repeated for the resized versions of each mass. 1120 (2×560) resultant ROIs with inserted masses were then processed and used as signal present images for 4AFC study. Six medical physicist participated in the reading test. Percentage correct (PC) in detection of masses were determined as a function of the background glandularity, type of mass, outer mass size in x-y plane and thickness of the core of the mass in z direction. Average PC of all observers and 95% confidence interval (CI) were calculated using bootstrapping.

Results

The results show, as expected, that detectability of the masses decreased from an overall detectability of 80% to 59% as the glandularity of the background increased. Detectability also decreased from 80% to 45% with decreasing outer size of the masses. Comparison of two types of masses shows higher detection for ill-defined masses than spiculated masses for similar outer sizes of the masses. However, for the same thickness of core (size in z direction) there was no significant difference in detectability between ill-defined and spiculated masses.

Conclusions

The presented 4AFC observer study and simulated data can be used to investigate threshold values for breast mass detection in mammographic background differentiated towards mass size, mass type and background type. In addition, we aim to use these results to test new model observers. Both issues are the purpose of ongoing research.

A survey of holographic 3D visualisation and perception of medical image datasets

Dr Javid Khan[1] (EngD), and Prof. Gordon Findlater[2] (PhD)

[1]Holoxica Ltd, Scottish Microelectronics Centre, Alexander Crum Brown Rd, Edinburgh, U.K
[2]School of Biomedical Sciences, Edinburgh University, Teviot Place, Edinburgh, U.K

Rationale

Advanced medical imaging technologies were pioneered over the past few decades; winning Nobel Prizes for CT (Cormack and Hounsfiled, 1979) and MRI (Mansfield, 2003). The pace of innovation has been extensive in terms of size, safety, speed, accuracy and resolution. However, it is surprising to find that corresponding advances in 3D displays have not matched the rapid pace of these developments. Half of the human brain is devoted to visual processing, however, conventional presentation of this information is in 2D. Current stereo 3D with eyewear only presents an illusion of 3D, which is largely unacceptable for clinical settings. This paper surveys true 3D image visualisation to see how this can benefit the perception of volumetric images from medical scanners.

Methods

The scientific literature suggests that the best means of 3D visualisation is via holographic and similar approaches [1]. This technology has matured to a level where static 3D images can be created from DICOM images using digital holograms in full colour 3D. These have been used for surgery, forensics, training and outreach. The next stage is to develop a dynamic volumetric display that can interface directly to a medical scanner. This work is mainly conducted by the military, high-end research laboratories and small innovative companies. The key question is what the real tangible benefits of such visualisation are.

Results

A recently declassified report by the USAF Research Laboratory discusses the relative benefits of 3D visualization vs. 2D [2]. It concludes that 3D is overall 75% better than 2D for specific applications including: spatial manipulation, finding, identifying or classifying objects. In medicine, work by Fraunhofer HHI shows using 3D displays increases the speed of surgical procedures by 15% and improves the accuracy of surgery by up to 20% in terms of incisions, stitching and navigation through the body [3].

Conclusions

The literature strongly suggests that visualisation of medical volumetric datasets in true 3D is beneficial. These are important for the interpretation of medical imaging data as well as diagnostics, treatment and surgery. Finally, we are seeking quantifiable methodologies of 3D image assessment as well as collaborators for trials and validation of upcoming 3D holographic video display technology.

References

[1] V. M. Bove, *Display Holography's Digital Second Act,* Proc. IEEE, **100**(4), 918–928 (2012).
[2] J. P. McIntire, P. R. Havig, and E. E. Geiselman, *Stereoscopic 3D displays and human performance: A comprehensive review,* Displays, **35**(1), 18–26 (2014).
[3] P. Fraunhofer, *New opportunities for 3D technology in medicine - Research News* (Mar 2013).

4D CT Cardiac image enhancement for Subjective Medical Quality Perception of the Left Ventricle

Hrvoje Leventić[1], Časlav Livada[1], Irena Galić[1], Vladimir Zlokolica[2], Lazar Velicki[3], Danilo Babin[4], Ratko Obradović[2], Bogoljub Mihajlović[1]

[1]Faculty of Electrical Engineering, University of Osijek, Croatia
[2]Faculty of Technical Sciences, University of Novi Sad, Serbia
[3]Institute of Cardiovascular Diseases, Sremska Kamenica, Serbia
[4]TELIN, Ghent University, Belgium

Rationale

Recently it has been described [1] that even though MRI heart imaging is the gold standard for cardiac structure and function evaluation, MRI is not always feasible in some patients. For such patients CT is a viable alternative, which however can introduce significant quality degradation, especially in preferable case of low level of radiation. In this paper we investigate and assess image enhancement approaches to be applied to image slices of 4D CT technology to facilitate decisions made by cardiologists. This applies to both medical doctor's subjective visual quality opinion as well as semi-automatic computer vision tasks (in this case segmentation of heart left ventricle) which efficiency is again to be assessed by the medical doctors. The improved quality assessment by cardiologists should aid more precise detection of heart failure, for which left ventricle dysfunction is often the main cause.

Methods

We have processed 4D CT slices with image enhancement methods related to contrast enhancement and de-noising algorithms. The CT data bases were selected by medical doctors (MD), specialized in cardiac surgery. We explore the usage of existing multi-scale methodologies (Laplacian Pyramid and Non-decimated wavelet transform) for local image enhancement and de-noising but also develop our own algorithm for contrast enhancement and de-noising algorithm for this special case of cardiac images. The explored algorithms for image enhancement were applied with different parameters tunable my MDs, which have provided subjective quality assessment for the purpose of LV segmentation and 3D volume determination.

Results

Preliminary results indicate that our proposed MD-quality-assessment driven image enhancement algorithm significantly improves the image quality and its usability for more precise LV registration.

Conclusions

Medical quality assessment is highly important for image enhancement algorithm optimization.

References

[1] D. Mangalat, A. Kalogeropoulos, V. Georgiopoulou, A. Stillman, and J. Butler, "Value of Cardiac CT in Patients With Heart Failure," *Curr. Cardiovasc. Imaging Rep.*, vol. 2, no. 6, pp. 410–417, Dec. 2009.

A comparative study for image quality of a cone beam computed tomography scanner and a multislice computed tomography scanner for paranasal sinus imaging

Jens De Cock[1] (MD), Federica Zanca[1, 2] (MD), Ruben Pauwels[2] (PhD),
Robert Hermans[1, 2] (MD, PhD)

[1]Department of Radiology, University Hospitals Leuven, Leuven, Belgium
[2]Imaging and Pathology Department, KU Leuven, Leuven, Belgium

Rationale

To evaluate image quality of a state of the art cone beam computed tomography (CBCT) system and a multislice computed tomography (MSCT) system in patients with sinonasal poliposis.

Methods

In this retrospective study two radiologists evaluated 57 patients with sinonasal poliposis who underwent a CBCT or MSCT sinus examination, along with a control group of 90 patients with normal radiological findings.

Visual grading with respect to the visibility of previously determined structures, as well as overall image quality was used, using absolute grading. Rating was based on a 5 point scale: 1 = the structure cannot be identified; 2 = insufficient image quality, the structure is identifiable, but without visibility of details; 3 = acceptable image quality, the structure is assessable in most of the details; 4 = good image quality, the structure is assessable in detail; and 5 = excellent image quality, perfect delineation of the structure.

The results of the two readers were reported separately. The Mann-Whitney test was used to assess statistical differences between MSCT and CBCT scores.

Kendall’s coefficient of concordance in interreader agreement for overall image quality was calculated.

Results

Overall image quality in CBCT was scored significantly higher than in MSCT in patients with normal radiologic findings (p-value: 0.00001). In patients with sinonasal poliposis, MSCT scored significantly higher than CBCT (p-value: 0.00001).

Conclusions

CBCT and MSCT are both suited for the evaluation of sinonasal poliposis. In patients with sinonasal poliposis, clinically important structures of the paranasal sinuses can be better delineated with MSCT, whereas in patients without sinonasal poliposis, CBCT turns out to define the important structures of the sinonasal region better.

Diminishment of Recognition Memory for Radiographs over Time

Karla K. Evans, Ph.D.* Christina Thomas, B.S.** Tara L. Sagebiel, M.D.** Diana M. Palacio, M.D.**
Myrna C. Godoy, M.D.** Tamara Miner Haygood, Ph.D., M. D.**

**Department of Psychology, University of York, Heslington, York, UK*
*** Department of Diagnostic Radiology, UT M.D. Anderson Cancer Center, Houston, TX, USA*

Rationale

Radiologists' ability to discriminate recently encountered images from new images exceeds chance only modestly. Recall is distinctly better with everyday scenes than with radiographs. We studied 1) whether better memory for everyday scenes would persist with a less variable set of everyday scenes and a more varied set of radiographs and 2) the degree of memory degradation that would occur with a longer time lapse between first and second viewings.

Methods

Images evaluated included: 216 musculoskeletal (MSK) radiographs and 216 forest scenes. These images were randomly divided into subgroups for each participant. The radiologists first viewed72 study images immediately followed by 72 test images, half of which were also in the first set and half of which were new. This was done for both MSK radiographs and for forest scenes.
After 30 to 112 days (mean 41.7 days), 72 study images were again followed by 72 test images, half of which were again new, but now with a time lapse of 27 to 68 days (mean 49.9 days) between study and test phases. For each radiologist there was no overlap between images viewed in the immediate and delayed recall phases.
Our 11 participating radiologists were all ABR certified attending radiologists with 4.5 to 38 years (15.8 mean) of experience after residency.

Results

Memory for MSK images was better than for forest scenes ($F(1, 10) = 6.61$, $p<0.28$) in the immediate recall phase. Readers averaged 67% correct answers when viewing forest scenes and 77% correct answers for radiographs. This percentage of correct answers was greater than in previously published studies which evaluated recall of chest radiographs. With delayed testing, there was a distinct decline in memory for both radiographs and forests ($F(1,10)=116.74$, $p<0.00001$) , and the difference in memory between the two types of images disappeared, with chance performance for both at approximately 50% percent correct for both types.

Conclusions

1. Recognition memory is better for more varied than for more homogenous scenes. 2. Even with a varied test set, a delay of only a few weeks between viewings is enough to extinguish recognition memory.

Varied Frequency Information in Mammography

Karla K. Evans, Ph.D. *, Jeremy M. Wolfe ***, Tayler M. Schwartz, B.S.** Dianne Georgian-Smith, M.D. *** Rosalind P. Candelaria, M.D.** Mark J. Dryden, M.D.**
Tamara Miner Haygood, Ph.D., M. D.**

**Department of Psychology, University of York, Heslington, York, UK*
*** Department of Diagnostic Radiology, UT M.D. Anderson Cancer Center, Houston, TX*
****Harvard Medical School & Department of Diagnostic Radiology, Brigham and Women's Hospital, Boston, MA*

Rationale

Radiologists are able, with a great deal of accuracy, to recognize masses and architectural distortion on mammograms in less than a second of viewing time. Little is known about the type of information on the image that allows them to do this. We compared rapid identification of masses and architectural distortion on mammograms in their usual state and after suppression of the high or low frequency information.

Methods

Nine attending breast radiologists viewed bilateral mammograms in either CC or MLO projection. Image pairs were shown in random order and displayed for ½ second. There were 120 image pairs, each of which was shown with the usual appearance, with only high-frequency information displayed, and with only low-frequency information displayed, for 360 total image pairs. Half contained masses or architectural distortion; half did not. Radiologists indicated with a mouse click the location of suspected abnormalities and also their level of confidence on a scale of 0 – 100.

Results

Radiologists did best when viewing the intact images containing both high and low spatial frequency information (D' = 1.06). When viewing the altered images, they were better at finding abnormalities with the high-frequency images than with the low-frequency images. The D' for high spatial frequencies was .97, and the D' for low-spatial frequencies was .26.

Conclusions

Masses and architectural distortion were better recognized with high-frequency than low-frequency information. We suspect this is because the spiculations associated with architectural distortion are often slender and fine and therefore would be more obvious on higher-frequency images.

Comparing detection of architectural distortion to other cancers

Wasfi I. Suleiman, MSc; Mark F. McEntee, PhD; Sarah J. Lewis, PhD; Mohammad A Rawashdeh, PhD; Patrick C. Brennan, PhD

Medical Image Optimisation and Perception Group (MIOPeG), and the Brain and Mind Research Institute, The Faculty of Health Sciences, The University of Sydney.
Postal Address: Medical Imaging and Radiation Science, Block M, Level 2, Cumberland Campus, 75 East Street, Lidcombe, NSW 2141, Australia.

Introduction

Screening mammography is not a perfect tool since the identification of subtle lesions is challenging and, as a result, readers fail to detect 10-30% of malignancies. In contrast to masses and calcification, Architectural Distortion (AD) does not present as an increased density in mammography and therefore is reportedly the most difficult type of tumour to detect and the most commonly missed abnormality. The current study seeks to examine the difficulty of detecting AD and to examine readers' performance in detecting AD compared with other cancers when using digital mammography.

Methods

Forty-one experienced breast screen readers (20 US and 21 Australian) were asked to read a single test set of 30 digitally acquired mammographic cases. Twenty cases had abnormal findings (10 with AD, 10 non-AD) and 10 cases were normal. Each reader was asked to locate and rate any abnormalities. Lesion and case-based performance was assessed. For each group of readers (US; Australian; combined) jack-knife free-response receiver operating characteristic, figure of merit (JAFROC, FOM) and inferred receiver operating characteristic area under curve (ROC, Az) were calculated. Sensitivity, location sensitivity, false positive (FP) and false negative (FN) responses were compared between images groups using Kruskal-Wallis statistics.

Results

For lesion-based analysis, significantly lower location sensitivity ($P \leq 0.0001$) and higher false negative values ($P \leq 0.0001$) were shown on AD images compared with non-AD images for all reader groups. The case based analysis demonstrated significantly lower sensitivity ($P=0.02$) and higher false negative values ($P \leq 0.0002$) for AD compared with non-AD lesions for the combined group, along with lower false negative values for non-AD lesions compared with all images ($P=0.02$) for all reader groups.

Conclusions

AD remains a challenging task for readers.

Key points:

- Architectural Distortion remains a challenging task for readers with digital acquisition.
- Significantly lower location sensitivity scores for the AD were shown.
- Significantly lower false negative values for non-AD values were shown.

3D rendering in Virtual Colonoscopy: utility and impact

S Gryspeerdt, Ph Lefere

AZ Delta, Roeselare, Belgium
VCTC, Hooglede, Belgium
stefaan.gryspeerdt@skynet.be
www.vctc.eu

Virtual colonoscopy (VC) is appealing as a non-invasive total colon examination technique, using dedicated software allowing visualization and navigation through the colon of the patient.

The main challenge is how to efficiently, but comprehensively, visualize all interesting features. 2D displays have significant limitations for this purpose, because 2D displays only show a 2D projection of a 3D scene. 2D projection doesn't allow us to see the parts of the dataset away from our direct view. Thus 2D imaging alone seems to be at risk of missing subtle lesions.

In recent years, research in VC has focused on real-time volume rendering of the colon, improving visualization of the complex colon wall. Different 3D display methods have been proposed, all focusing on improved 3D visualization using a virtual interior navigation or even dissection of the human colon.

The main purpose of such 3D analysis is to reduce the reading time and improve sensitivity. The ideal 3D display mode shows the complete colonic surface (hence, no polyps can be missed), in a time efficient way, and without image distortion (thus polyps can be recognized as such).

Compared to 2D analysis, hidden spots and distortion remains the main challenge in 3D volume rendering. Moreover, since dose reduction is a main concern in VC, increased image noise might further increase distortion, thus inducing false positive or negative findings.

We will compare 3D rendering techniques with 2D visualization, evaluate both techniques as for time efficiency , compare results in different groups of patients (high vs low polyp prevalence), and finally look at the techniques how to overcome problems related to image noise in current ultra-low dose VC.

An observer performance study in mammography: experience from the radiologist

Van Ongeval Chantal, PhD

Department of Radiology, UZ Leuven campus Gasthuisberg, Herestraat 49, 3000 Leuven

In digital mammography optimization of image processing is possible. Image processing aims to visualize radiographic (raw) images in such a way that small masses and microcalcifications can be detected by the radiologist. In addition, lesions and normal anatomical structures should be visualized without any artificial enhancement. Evaluating the quality of an image processing algorithm based on contrast-detail measurements is difficult, as the background is homogeneous. The clinical part of the evaluation of image quality can be realized in different ways. Firstly, a set of image quality criteria can be used. The concept behind these criteria is that images showing all normal structures with high detail and contrast will perform similarly for (subtle) lesions. Secondly, lesions can be simulated in raw data images and comparative studies with different processing algorithms can be done. A major advantage of the first approach is that it can be used by the radiologist each time whenever an old algorithm is optimized or when new digital mammography systems enter the market, and the evaluation is not time consuming. In case of simulation of lesions, a lot of interactions by physicists, scientist and radiologists are necessary and the number of images necessary for the evaluation is high. The methods for the assessment of image-processing algorithms are the receiver operating characteristic (ROC), the free-response receiver operating characteristic (FROC) and a visual grading analysis (VGA). As ROC an FROC analysis are used to measure the ability of the radiologist to detect and interpret breast lesions like microcalcifications and masses, VGA is based on the scoring of the visualization of normal anatomical structures. In case of studies on the effect of different image processing algorithm on the visualization and characterization of simulated lesions in a set of images ROC/FROC and VGA can be used and their relationship can be investigated.

In a study of Zanca et al. microcalcifications were retracted out of breast biopsy specimen and were simulated in digital images, different processing algorithms were applied on these images and radiologists were asked to find the simulated lesions, to localize them and to score the degree of malignancy. This approach was very similar to the daily clinical work: finding small lesions in casu microcalcifications in a mammogram. Finding lesions is more objective compared to the evaluation of the anatomical structures. The VGA evaluation is much more subjective, is difficult to quantify and is depending on the "taste" of the radiologist. An alternative is the use of VGA by comparing the same image processed with two different algorithm. In the paper of Zanca et al. on the comparison of visual grading and free-response ROC analyses for the assessment of image-processing algorithms in digital mammography, the results of the FROC did not follow the VGA results. It was therefore questioned if the fulfillment of the investigated quality criteria by VGA evaluation is sufficient for accurate detection of microcalcifications. The main reason for this difference is the different clinical task: in VGA the task is the evaluation of the visibility of the normal anatomy, in the FROC study the task is the detection of microcalcifications.

Meanwhile the image quality criteria with the VGA approach were part of the typetest for digital mammography in Belgium and in Europe (Euref) and proved its utility in the evaluation of different processing algorithm. Different optimizations of image processing algorithms were realized due to these evaluations.

Concerning the clinical task, later evaluation studies performed by Shaheen et al and Salvagnini et al. learned that the simulation of masses was more difficult and the evaluation was different compared to the search of microcalcifications. This is correlated to the fact that low density masses are very similar to glandular islands who can be dispersed in the breast parenchyma while for microcalcifications, the differences between physiological calcifications and clusters of calcifications are much clearer.

Concerning the evaluation of the impact of optimization on the detectability, the ROC/FROC approach remains necessary and more investigation on the optimal simulation of different lesions and automated search will be asked in the future.

References

Carton AK, Bosmans H, Van Ongeval C, Souverijns G, Rogge F, Van Steen A, Marchal G. Development and validation of a simulation procedure to study the visibility of micro calcifications in digital mammograms. Med Phys. 2003 Aug;30(8):2234-40.

Van Ongeval C, Van Steen A, Geniets C, Dekeyzer F, Bosmans H, Marchal G. Clinical image quality criteria for full field digital mammography: a first practical application. Radiat Prot Dosimetry. 2008;129(1-3):265-70.

Zanca F, Jacobs J, Van Ongeval C, Claus F, Celis V, Geniets C, Provost V, Pauwels H, Marchal G, Bosmans H. Evaluation of clinical image processing algorithms used in digital mammography. Med Phys. 2009 Mar;36(3):765-75.

Zanca F, Van Ongeval C, Claus F, Jacobs J, Oyen R, Bosmans H. Comparison of visual grading and free-response ROC analyses for assessment of image-processing algorithms in digital mammography. Br J Radiol. 2012 Dec;85(1020):e1233-41.

Zanca F, Hillis SL, Claus F, Van Ongeval C, Celis V, Provoost V, Yoon HJ, Bosmans H. Correlation of free-response and receiver-operating-characteristic area-under-the-curve estimates: results from independently conducted FROC/ROC studies in mammography. Med Phys. 2012 Oct;39(10):5917-29

Shaheen E, De Keyzer F, Bosmans H, Dance DR, Young KC, Van Ongeval C. The simulation of 3D mass models in 2D digital mammography and breast tomosynthesis. Med Phys. 2014 Aug;41(8):081913.

Salvagnini E., Bosmans H., Van Ongeval C., Van Steen A., Michielsen K., Cockmartin L., Struelens L. and Marshall N.W., "A FROC study on the influence of breast thickness on simulated lesion detection", MIPS conference, Ghent, June 2015, Oral Presentation

Salvagnini E., Bosmans H., Van Ongeval C., Van Steen A., Michielsen K., Cockmartin L., Struelens L. and Marshall N.W., "Impact of compressed breast thickness on detectability of simulated lesions: a clinical trial", ECR conference, Vienna, March 2015, poster presentation.