KEY ISSUES AND SPECIFICITIES FOR THE OBJECTIVE MEDICAL IMAGE QUALITY ASSESSMENT

Lu Zhang, Christine Cavaro-Ménard, Patrick Le Callet

KEY ISSUES AND SPECIFICITIES FOR THE OBJECTIVE MEDICAL IMAGE QUALITY ASSESSMENT

Lu Zhang, Christine Cavaro-Ménard and Patrick Le Callet

LISA, EA 4094, University of Angers; IRCCyN, UMR 6597, University of Nantes, France

ABSTRACT

Though several objective image quality assessment methods originally proposed for natural images and videos have been used in the context of medical images, some important specificities usually ignored have to be considered. This paper presents a review on some key issues (diagnostic task, pathology, figure of merit, expertise, validation subjective experimental protocol, etc.) that must be considered for the objective quality assessment of still radiographic images acquired from the acquisition systems of varied imaging modalities.

1. INTRODUCTION

While video quality assessment has been recently addressed in the context of medical applications, it has been often limited to medical videos such as telesurgery, endoscopy, etc.. In this paper, we focus on other field that is still radiographic image acquired from the acquisition systems of varied imaging modalities, such as CT (Computed tomography), MRI (Magnetic Resonance Imaging), ultrasound, projection (plain) radiography, fluoroscopy, PET (Positron Emission Tomography), SPECT (Single Photon Emission Computed Tomography), etc.

Several objective image quality assessment methods originally proposed for natural images and videos have been used in the context of medical images [1, 2], but some important specificities and key issues of objective medical image quality assessment have however been ignored. These actually make objective medical image quality assessment differs with the objective assessment of natural images and videos extending this latter to medical "natural" video content, such as coronary angiography, ultrasound video, etc..

While it has been poorly addressed by natural image quality assessment community, task-based approaches [3, 4] are fundamental and popular in the context of medical imaging. The underlying paradigm is to quantify the quality of a particular image by its effectiveness with respect to its intended task. Natural images quality assessment is usually focused on the perception of impairments, while the quality assessment of medical images normally focus on the radiologists diagnostic task performance. It is more understand-

able in terms of subjective assessment method, as illustrated in Fig. 1 (a), radiologists (human observers) are asked to perform the same diagnostic task given different medical imaging systems, which could be acquisition devices, image post-processing algorithms or image visualization systems. A better task performance means a better diagnostic accuracy, thus the system that enables radiologists to gain a better task performance or to spend less time for interpretation with the same diagnostic reliability is said to be better. Concerning objective assessment method, the key problem lies in modeling the diagnostic task process of radiologists by a numerical observer. Once the numerical observer performs approximately radiologists, it can evaluate different medical imaging systems using the same paradigm as the subjective assessment method, as illustrated in Fig. 1 (b).
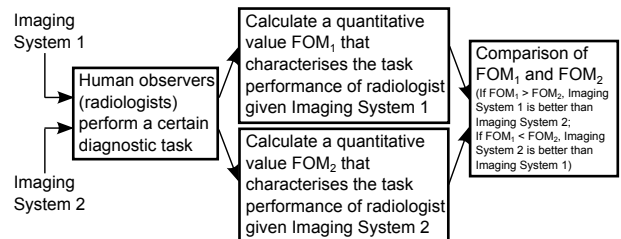
(a) Subjective assessment of medical imaging systems
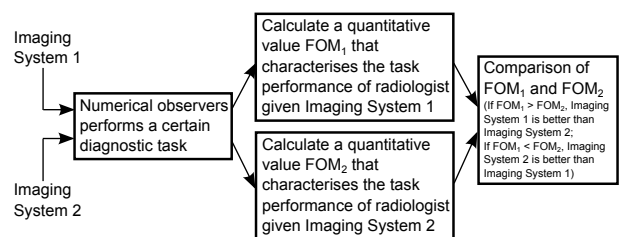
(b) Objective assessment of medical imaging systems

Fig. 1. Example of assessing two different medical imaging systems, which could be acquisition devices, image post-processing algorithms or image visualization systems.

Note that the objective to develop a predictive numerical observer of human task performance is not to substitute radiologists in the daily diagnosis, but to select a better medi-

cal imaging technique or system which can help radiologists ameliorate their diagnostic accuracy and efficiency.

This paper gives a review on some key issues and specificities to be considered in the implementation of a numerical observer (an objective medical image quality assessment method), whose flowchart is shown in Fig. 2. Each of the key issues will be presented through successive stages of the flowchart in Section 2, then subjective experiments with their results to illustrate some of the issues will be described in Section 3.
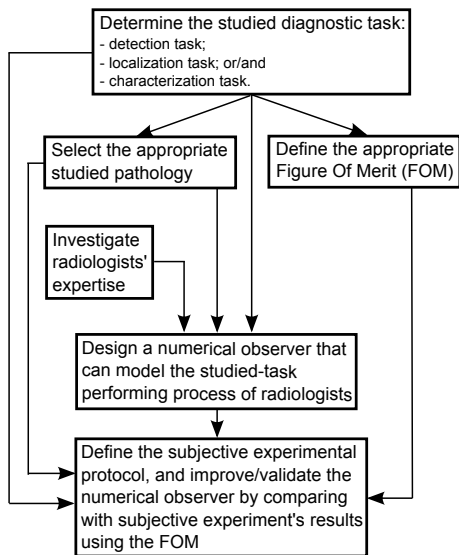

Fig. 2. The flowchart of implementing a numerical observer (an objective medical image quality assessment method).

2. KEY ISSUES FOR THE IMPLEMENTATION OF A NUMERICAL OBSERVER

2.1. Diagnostic tasks

The first thing to determine in the flowchart of implementing a numerical observer is the studied diagnostic task.

It is useful to well characterize the diagnostic process in order to model it. One widely accepted way [5, 6] is to divide the diagnostic process into three tasks: the detection task, the localization task and the characterization task. The detection task requires simply a confidence rating concerning the presence of an abnormality, e.g. a lesion or a nodule. The localization task consists in indicating the locations of abnormalities. The characterization task, related to assessing the different elements of the abnormality appearance, normally involves a linguistic response describing distinctive characteristics or essential features of abnormalities.

Modeling the entire diagnostic process is a very complicated problem, especially for the characterization task, this could explain why numerical observers are limited in

task range so far. Most of them deal with the detection task [7, 8], only several numerical observers are concerned with the localization task [9], and none investigates the characterization task to our knowledge.

Only after the studied task is determined, the studied pathology can be chosen appropriately (cf. Section 2.2) and the figure of merit (FOM, cf. Section 2.3) can be chosen correspondingly.

2.2. Appropriate pathology for different diagnostic task

Note that different imaging modalities will be favorable for the diagnosis of different pathologies, which present specific physical and physiological phenomena in each imaging modality [10]. Thus the studied pathology should be chosen by considering both the studied task and the studied modality.

As far as the detection and localization tasks are concerned, the abnormality of an appropriate pathology must not be too conspicuous. Otherwise, observers (radiologists or numerical observers) could always detect and localize the abnormalities easily with any imaging system. In that way, different imaging systems can not be differentiated through different task performances of observers. Taking Magnetic Resonance Imaging (MRI) modality for example, Fig. 3 shows MR images of two pathologies: Multiple Sclerosis (MS) and High-Grade Glioma (HGG). It is not hard to see from Fig. 3 (a) that the HGG lesion is so obvious on any one of the six sequences that even a person without any medical background can tell where it is. The evidence of the HGG lesion would not change even with moderate degradation of the imaging system for detection and localization tasks. In contrast, the MS lesions are much more subtle and difficult to be perceived, as seen from Fig. 3 (b). Furthermore, an accurate detection and localization of MS lesions is the first and the most important step for the diagnosis and treatment of MS patients. Thus if we choose Magnetic Resonance Imaging (MRI) as the studied modality, MS could be considered as one of the appropriate pathologies for the detection and localization tasks.

However, the appropriate pathology may be the exact opposite for the characterization task. An appropriate pathology for the characterization task should be a pathology for which the abnormality appearance examination has a direct and dominant effect on the diagnosis and treatment of patients, which is the case for HGG [10]. In the characterization task, on MR images a lot of lesion aspects should be evaluated on different sequences, such as : (1) Does the lesion appear as an iso-signal, a hyper-signal or a hypo-signal? In neuroradiological terminology, "iso-signal", "hyper-signal", "hypo-signal" indicate that the lesion has a equal, higher or lower intensity than the white matter, respectively. (2) Is the lesion texture homogeneous or heterogeneous? (3) Is the lesion shape a circle, an ellipse or other irregu-
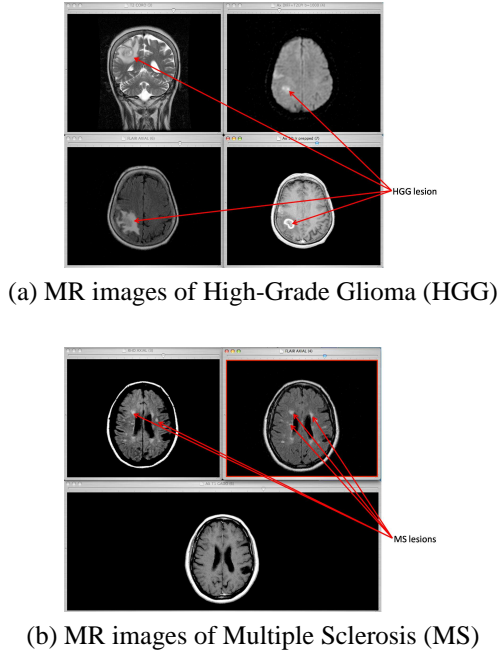
(a) MR images of High-Grade Glioma (HGG)

(b) MR images of Multiple Sclerosis (MS)

Fig. 3. MR images of two pathologies, Multiple Sclerosis (MS) and High-Grade Glioma (HGG), on different sequences.

lar shape? (4) Is the lesion outline thick, thin or irregular, clear or obscure? ... The MR images in Fig. 3 (a) can indeed provide the information for these questions in the characterization task.

Note that in MRI, a "sequence" is a subtle combination of radiofrequency pulses and gradients. Whatever the type of sequence, the aim is to favor the signal of a particular tissue (contrast), as quickly as possible (speed), while limiting the artifacts and without altering the signal to noise ratio [11]. The influence of using different sequences is out of the scope of this paper.

Once the pathology is selected, its lesion could be simulated by mathematical model in order to facilitate the design and the test of a numerical observer.

2.3. Figures of merit

A figure of merit (FOM) is a a quantitative value used to characterize a certain diagnostic task performance, derived from the responses of observers (radiologists or numerical observers) on a set of test images. It is essential for comparing or assessing different imaging systems, as well as for validating a numerical observer by comparing its FOM results with those gotten from subjective experiments.

Different FOMs have been proposed for different diagnostic tasks. The receiver operating characteristic (ROC) curve [12] is a FOM for the detection task, where a binary positive/negative decision is made by comparing observers

confidence rating with a criterion. A positive decision is called a true positive (TP) when the gold standard is also lesion-present, otherwise it is called a false positive (FP). A negative decision is called a true negative (TN) when the gold standard is also lesion-absent, otherwise it is called a false negative (FN). A ROC curve is a graphical plot of the fraction of TPs out of the actual positives (TPF) vs. the fraction of FPs out of the actual negatives (FPF), which is a comparison of two operating characteristics (TPF and FPF) as the criterion changes, as illustrated in Fig. 4. The area under the ROC curve (AUC) is a summary statistic of the ROC curve.
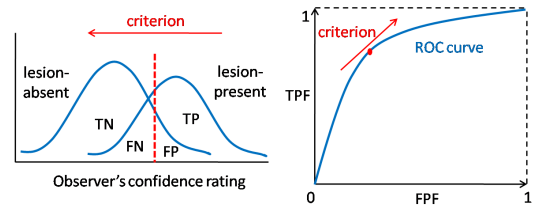

Fig. 4. An example of ROC curve, which is a graphical plot of the fraction of TPs out of the actual positives (TPF) vs. the fraction of FPs out of the actual negatives (FPF), which is a comparison of two operating characteristics (TPF and FPF) as the criterion changes.

The ROC curve is a classic approach in signal detection theory and is widely used in diverse domains. But in medicine and radiology, in order to evaluate more completely the diagnostic task performance, it is necessary to include the localization task besides the detection task. In addition, the localization is very important for diagnosis in neurology, the localization of the lesion is essential for optimal surgery. Therefore various extensions to ROC methods have been proposed to address these limitations of the classic ROC paradigm. There are two cases for the localization task: (1) Only one lesion is possibly present on an image, and observers have to locate only one lesion on each test image; (2) Multiple lesions are possibly present on an image, and observers have to locate each one of the lesions on each test image. The Localization ROC (LROC) curve [13] was proposed as the FOM for the first case. The Free-response ROC (FROC) [14] and the Alternative Free-response ROC (AFROC) [15] were proposed as the FOMs for the second case.

Recall that the observer's responses for the detection and localization tasks are a marked rate concerning the presence of each lesion, combined with the corresponding marked coordinates (indicating the lesion center) of each lesion. The above three FOMs classify the marked coordinates, regardless of the marked rates, as a TP or FP by comparing the distance between the marked point and the actual center of the lesion to an "acceptance radius". Then if the correspond-
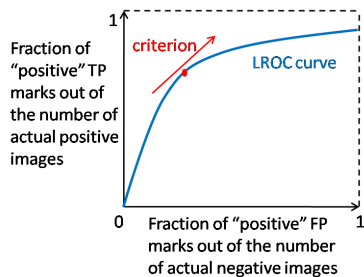
Fig. 5. An example of LROC curve, which is a graphical plot of the fraction of positive TP marks out of the number of actual positive images vs. the fraction of positive FP marks out of the number of actual negative images.
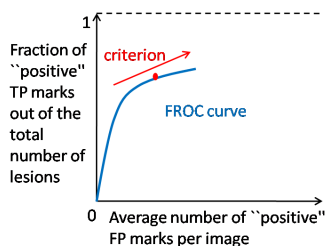

Fig. 6. An example of empirical FROC curve, which is a graphical plot of the fraction of "positive" TP marks out of the total number of lesions vs. the average number of "positive" FP marks per image.

ing rate is greater than the criterion, the TP or FP mark is classified as "positive", and "negative" in reverse.

A LROC curve uses the fraction of "positive" TP marks out of the number of actual positive images as its ordinate and the fraction of "positive" FP marks out of the number of actual negative images as its abscissa, as illustrated in Fig. 5. An empirical FROC curve (cf. Fig. 6) is a graphical plot of the fraction of "positive" TP marks out of the total number of lesions vs. the average number of "positive" FP marks per image. Note that the area under the empirical FROC curve can no longer be a FOM to summarize FROC curve, since a larger value of area under the empirical FROC curve can result either from an increase in TPs with correct localization or an increase in the number of FPs on each image[16]. However, in the literature we can still find some FOMs summarizing FROC data, such as the "augmented area under FROC curve" [14]. The AFROC curve (cf. Fig. 7) differs in the abscissas definition, compared to the FROC curve. For a certain criterion, the AFROC only considers one FP mark, the one with the highest rating, on each test image. Then it uses the number of highest rated "positive" FP marks divided by the total number of test images as its abscissa. The AFROC tends to be more stable and have a higher statistical power [15].
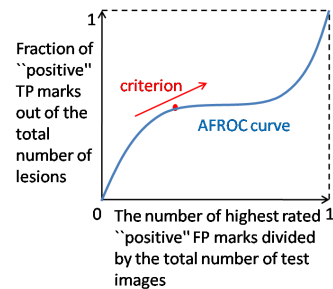

Fig. 7. An example of AFROC curve, which is a graphical plot of the fraction of "positive" TP marks out of the total number of lesions vs. the number of highest rated "positive" FP marks divided by the total number of test images.

2.4. Expertise

In order to design a numerical observer that can model the diagnostic process of radiologists, there is one more thing to be considered and investigated for medical image quality assessment than for natural image quality assessment, which is the influences of expertise.

While the end users of natural images or videos are naive observers, those of medical images are radiologists. Their expertise influences their task performance, especially in the cognitive process to interpret perceived information. For example, it has been showed in [17] that experience combined with training provides the basis for generating efficient visual search strategies and developing distinctive conceptual criteria for perceptual differentiation and interpretation of true breast masses from image artifacts and structured noise that mimics breast abnormalities; and it is found in [18] that an important aspect of the development of expertise is improved pattern recognition (taking in more information during the initial Gestalt or gist view) as well as improved allocation of attention and visual processing resources; an illustration about the influence of expertise on the perception of MR images will be presented in Section 3.

2.5. Subjective experimental protocol

The last thing to be considered to implement a numerical observer is the subjective experimental protocol.

The subjective experimental protocol for validating a objective quality assessment method is different in the context of medical images, compared to natural images. In natural image quality evaluation experiments, the investigators are normally asked to give a score on the quality, and the mean opinion score (MOS). However, in the context of medical images, the response depends on the studied task, as discussed in Section 2.1. Then different task performances can be quantified using the corresponding FOM.

As far as the detection and localization tasks are con-

cerned, a good subjective experimental protocol can be found in [19], which is close to the real clinical paradigm. One thing worthy of remark pointed in [19] is that if the image background is not clinical, but simulated, there is no justification for using radiologists to interpret such images. Since there is nothing in radiologists' training and education background that specially qualifies them to read such images, using radiologists in this mode is wasteful of their precious time and effort that they would focus better on providing timely diagnosis and treatment solutions to their patients. Thus recent investigations use more elements of real clinical images and abnormalities rather than just mathematically simulated backgrounds and targets [20].

3. AN ILLUSTRATION

Here we give an example to illustrate the investigation of the expertise (one of the key issues). Details about these experiments can be found in [21].

Our goal is to propose a numerical observer that can approach radiologists' task performance. We chose the detection task as the studied task, MRI as the studied modality and MS as the studied pathology. MS lesion is simulated by a two-dimensional elliptical Gaussian function. Before going further into the numerical observer modeling and the validation of it by using the ROC, we need inevitably to investigate the influences of the expertise.

Two subjective experiments have been conducted: Exp 1 consisted of six white matter blocks being selected within the white matter (without anatomical information), , as illustrated in Fig. 8 (a). Exp 2 comprised of one axial, cerebral slice, with six lesions located in separated areas of the slice (with anatomical information), as illustrated in Fig. 8 (b). One expert (more than 10 years experience), three radiologists (less than 5 years experience) and eight nave observers (without any prior medical knowledge) have performed these experiments. These human observers have been asked different questions dependent upon level of experience; the three radiologists and eight nave observers were asked if they were aware of any hyper-signal (this is considered as the sensation stage of visual processing), while the expert was asked if a clinical sign was present (in this case the expert needs to consider the clinical implications of a signal and interpret the findings cognitively, thus this is considered as the perception stage of visual processing).

Since many objective assessment studies were carried on a relatively homogeneous region (e.g. simulated background and the white matter blocks in our Exp 1), while radiologists normally view images in their entirety, we think that it is also important to know if the anatomical information changes the radiologists' behavior. Thus the analysis of the visibility thresholds, gotten from psychometric

curves for each participant and experiment, allows us to investigate not only the influence of expertise at two stages of visual processing, but also the influences with and without anatomical information.

Results indicate that at the sensation stage, radiologists have better detectability of simple hyper-signals than nave observers and anatomical information does not influence their sensation performance; at the perceptual stage, experts knowledge appears to influence the interpretation of the hyper-signals and an elevation of the detection threshold was observed, in addition, more anatomical information contributes to a reduction of the perception threshold.
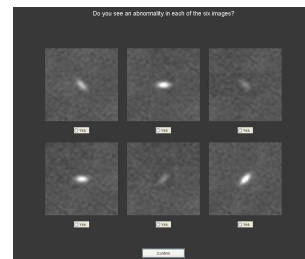
(a) GUI of Exp 1
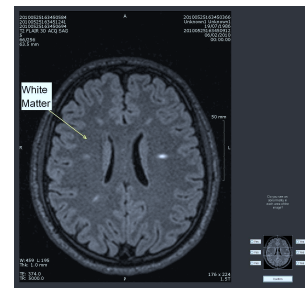
(b) GUI of Exp 2 (the arrow indicats the white matter)

Fig. 8. Graphical user interface (GUI) of two subjective experiments, figures from [21].

4. CONCLUSION

This paper addresses the key issues in the context of objective medical image quality assessment. On the whole, before going deeper into quality assessment we should firstly determine the diagnostic task and the modality, depending on that we can then reasonably choose the pathology and the figure of merit. We should also study the expertise in radiology for developing a numerical observer that can model the radiologists (human observers). Finally, the subjective experiment should be conducted to approximate the clinical practice as much as possible, in order to validate the numerical observer.

5. REFERENCES

[1] Nirmala S. R, Dandapat S., and Bora P. K., "Performance evaluation of distortion measures for retinal images," *International Journal of Computer Applications*, vol. 17, no. 6, pp. 17–23, March 2011, Published by Foundation of Computer Science.

[2] Olshen R. A Cosman P., Gray R. M, "Evaluating quality of compressed medical images: Snr, subjective rating, and diagnostic accuracy," in *Proceedings of the IEEE*, 1994, vol. 82, pp. 919–932.

[3] H.H. Barrett and K.J. Myerrs, *Foundations of Image Science*, John Wiley and Sons, Inc., Hoboken, New Jersey, USA, 2004.

[4] E.A. Krupinski and Y. Jiang, "Anniversary paper: Evaluation of medical imaging systems," *Medical physics*, vol. 35, pp. 645, 2008.

[5] Christine Cavaro-Menard, Jean-Yves Tanguy, and Patrick Le Callet, "Eye-position recording during brain mri examination to identify and characterize steps of glioma diagnosis," in *SPIE Medical Imaging*, 2010, vol. 7627, pp. 76270E–76270E–8.

[6] Louise Dickinson, Hashim U. Ahmed, Clare Allen, Jelle O. Barentsz, Brendan Carey, Jurgen J. Futterer, Stijn W. Heijmink, Peter J. Hoskin, Alex Kirkham, Anwar R. Padhani, Raj Persad, Philippe Puech, Shonit Punwani, Aslam S. Sohaib, Bertrand Tombal, Arnauld Villers, Jan van der Meulen, and Mark Emberton, "Magnetic resonance imaging for the detection, localisation, and characterisation of prostate cancer: Recommendations from a european consensus meeting," *European Urology*, vol. 59, no. 4, pp. 477 – 494, 2011.

[7] Jeffrey P Johnson, Elizabeth A Krupinski, Michelle Yan, Hans Roehrig, Anna R Graham, and Ronald S Weinstein, "Using a visual discrimination model for the detection of compression artifacts in virtual pathology images.," *IEEE Transactions on Medical Imaging*, vol. 30, no. 2, pp. 306–314, 2011.

[8] Fangfang Shen and Eric Clarkson, "Using fisher information to approximate ideal-observer performance on detection tasks for lumpy-background images," *Journal of the Optical Society of America A*, vol. 23, no. 10, pp. 2406–2414, 2006.

[9] S. Park, E. Clarkson, M. A. Kupinski, and H. H. Barrett, "Efficiency of the human observer detecting random signals in random backgrounds," *Journal of the Optical Society of America A*, vol. 22, no. 1, pp. 3–16, 2005.

[10] Karen L. Salzman Anne G. Osborn and A. James Barkovich, *Diagnostic Imaging: Brain, 2nd Edition*, Amirsys, Salt Lake City, UT, USA, 2009.

[11] "Advanced techniques in MRI, lecture 03," http://faculty.ksu.edu.sa/aalfaraj/RAD465/RAD465-Lecture03.pdf.

[12] John A. Swets, *Signal Detection Theory and ROC Analysis in Psychology and Diagnostics: Collected Papers*, Lawrence Erlbaum Associates, Mahwah, New Jersey, 1996.

[13] H C Gifford, M A King, R G Wells, W G Hawkins, M V Narayanan, and P H Pretorius, "LROC analysis of detector-response compensation in SPECT," *IEEE Transactions on Medical Imaging*, vol. 19, no. 5, pp. 463–473, 2000.

[14] Andriy I. Bandos, Howard E. Rockette, Tao Song, and David Gur, "Area under the free-response ROC curve (FROC) and a related summary index," *Biometrics*, vol. 65, pp. 247–256, 2009.

[15] Dev P. and Chakraborty, "Validation and statistical power comparison of methods for analyzing free-response observer performance studies," *Academic Radiology*, vol. 15, no. 12, pp. 1554 – 1566, 2008.

[16] Xin He and Eric Frey, "ROC, LROC, FROC, AFROC: an alphabet soup.," *Journal of the American College of Radiology JACR*, vol. 6, no. 9, pp. 652–655, 2009.

[17] C.F. Nodine, H.L. Kundel, S.C. Lauver, and L.C. Toto, "Nature of expertise in searching mammograms for breast masses," *Academic radiology*, vol. 3, no. 12, pp. 1000–1006, 1996.

[18] E.A. Krupinski and R.S. Weinstein, "Changes in visual search patterns of pathology residents as they gain experience," in *Proceedings of SPIE*, 2011, vol. 7966, p. 79660P.

[19] "How to conduct a free-response study," http://www.devchakraborty.com/HowToConduct.html.

[20] E.A. Krupinski and K.S. Berbaum, "The medical image perception society update on key issues for image perception research1," *Radiology*, vol. 253, no. 1, pp. 230–233, 2009.

[21] L. Zhang, C. Cavaro-Ménard, P. Le Callet, and L. H. K. Cooper, "The effects of anatomical information and observer expertise on abnormality detection task," in *in Proc. SPIE Medical Imaging*, February 2011, vol. 7966.