Technical University of Denmark

Ultrasound Image Quality Assessment

A framework for evaluation of clinical image quality

Hemmsen, Martin Christian; Pedersen, Mads Møller; Nikolov, Svetoslav; Nielsen, Michael Bachmann; Jensen, Jørgen Arendt

Publication date:
2010

Document Version
Publisher's PDF, also known as Version of record

Citation (APA):
Hemmsen, M. C., Pedersen, M. M., Nikolov, S., Nielsen, M. B., & Jensen, J. A. (2010). Ultrasound Image Quality Assessment: A framework for evaluation of clinical image quality. Abstract from SPIE, San Diego, .

Ultrasound Image Quality Assessment: A framework for evaluation of clinical image quality

Martin Christian Hemmsen[a,b], Mads Møller Pedersen[c], Svetoslav Ivanov Nikolov[b], Michael Backmann Nielsen[c], Jørgen Arendt Jensen[a]

[a]Center for Fast Ultrasound Imaging, Technical University of Denmark, 2800 Lyngby, Denmark
[b]B-K Medical A/S, Mileparken 34, 2730 Herlev, Denmark
[c]Department of Radiology, Rigshospitalet, 2100 Copenhagen, Denmark

ABSTRACT

Improvement of ultrasound images should be guided by their diagnostic value. Evaluation of clinical image quality is generally performed subjectively, because objective criteria have not yet been fully developed and accepted for the evaluation of clinical image quality. Based on recommendation 500 from the International Telecommunication Union - Radiocommunication (ITU-R) for such subjective quality assessment, this work presents equipment and a methodology for clinical image quality evaluation for guiding the development of new and improved imaging. The system is based on a BK-Medical 2202 ProFocus scanner equipped with a UA2227 research interface, connected to a PC through X64-CL Express camera link. Data acquisition features subject data recording, loading/saving of exact scanner settings (for later experiment reproducibility), free access to all system parameters for beamformation and is applicable for clinical use. The free access to all system parameters enables the ability to capture standardized images as found in the clinic and experimental data from new processing or beamformation methods. The length of the data sequences is only restricted by the memory of the external PC. Data may be captured interleaved, switching between multiple setups, to maintain identical transducer, scanner, region of interest and recording time on both the experimental- and standardized images. Data storage is approximately 15.1 seconds pr. 3 sec sequence including complete scanner settings and patient information, which is fast enough to get sufficient number of scans under realistic operating conditions, so that statistical evaluation is valid and reliable.

Keywords: Ultrasound imaging, Methodology for clinical quality assessment, Statistical analysis

1. INTRODUCTION

Researchers of new ultrasound imaging methods are interested in assessing the clinical quality of their method to increase the impact and attention it receives by manufacturers and other researchers. Such assessment of clinical quality is generally performed subjectively, because objective criteria have not yet been fully developed and accepted for the evaluation of clinical image quality. One major limitation with subjective assessment is, if the opinion is just based on an impression of quality, the usefulness of the assessment may be questionable (Vucich 1979, Barrett and Myers 2004, Månsson 2000). When judged by task-based critera - for example by the opinion of the radiologist relating to his/her ability to perceive certain anatomical details or features in the image and his/her confidence on the perception of these details, the assessment is more relevant.[1] Major difficulties accessing ultrasound data in the laboratory and clinic has not only limited the basic research, but also hindered the clinical testing of new ultrasound applications. In order to access raw ultrasound data, researchers have worked with ultrasound manufacturers to build custom ultrasound systems such as RASMUS,[2] but due to the size of the scanner it is inaccessible to the clinic. Recently a number of research interface platforms for

Further author information: (Send correspondence to Martin Christian Hemmsen)
Martin Christian Hemmsen: E-mail: mah@elektro.dtu.dk, Telephone: (+45) 45 25 57 38
Mads Møller Pedersen: E-mail: phd@medit.dk, Telephone: +45 35 45 18 23
Svetoslav Ivanov Nikolov: E-mail: sin@bkmed.dk, Telephone: +45 44 52 82 07
Michael Backmann Nielsen: E-mail: mbn@dadl.dk, Telephone: +45 35 45 35 45
Jørgen Arendt Jensen: E-mail: jaj@elektro.dtu.dk, Telephone: +45 45 25 39 24

clinical ultrasound scanners has been developed for systems such as Hitachi HiVision 5500,[?] Siemens Antares[3] and the Ultrasonix 500.[4] With the introduction of research interface platforms on clinicaly available scanners it is now possible to acquire and store data. However, for a system to be suitable for acquisition of data for clinical evaluations, the system has to keep factors, such as identical transducer, region of interest and recording time constant on both images. Another system requirement is the ability to get sufficient number of scans under realistic operating conditions, so that the statistical evaluation is reliable. Thus the data acquisition should, be capable of acquiring and storing sufficiently enough data, fast enough to conduct an ultrasound examination with multiple image sequences. The objective of this work is to develop a methodology and equipment for image quality evaluation for guiding the development of new and improved imaging methods.

2. EVALUATION METHODOLOGY

The main issue in performing a structured and fair comparison between images is to keep factors, such as transducer, scanner, region of interest and recording time constant. Other issues to consider is to get sufficient number of scans under realistic operating conditions and separating the developer and assessor in the evaluation process. To fullfill these demands we propose that evaluations of new methods is conducted in a three stage research, as illustrated on figure 1:

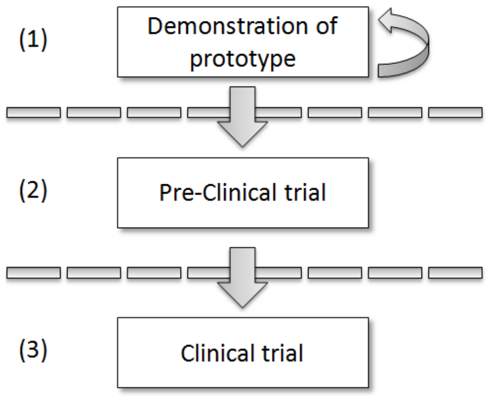


Figure 1: Diagram of the methodology

1. **Demonstration of prototype**, is the stage where developers demonstrate new imaging methods with measurements on phantoms and a few in-vivo images to demonstrate a workable prototype. In a collaboration between the developer and the ultrasound specialists, the new method's parameters are iteratively optimized to achieve the best possible setup. This stage ends and a pre-clinical study is started once all parameters are fixed.

2. **Pre-clinical trial**, is the stage where the relevance of a clinical investigation is tested. The necessary number of patients for the real clinical study is determined. This stage ends and the clinical trial begins when an exact clinical protocol is developed. It describes the method and its parameters in such a degree that the developer is and should be left out in the active part of the following research and should not have any influence on the outcome of the research in either data acquisition, any form of processing of it or evaluation.

3. **Clinical trial**, is the stage of research that determines the statistical significance of the new method. Assessment of the method is performed by a number of ultrasound specialists independent to the method. Furthermore, the assessors must be separeted from the specialists performing the ultrasound scanning, blinding them from the acquisition and any form of processing of it.

The evaluation methodology should ensure the validity of the assessment, as it separates the developer, investigator, and assessor once a research protocol has been established. This separation eliminates any confounding influence on the result from the developer and new processing schemes is not driven by the developers, but by the clinical value.

3. SYSTEM DESCRIPTION

3.1 Data Acquisition

The Ultrasound Research Interface (URI) consists of a commercially available ultrasound scanner (2202 ProFocus with a UA2227 research interface, BK-Medical, Herlev, Denmark) and a standard pc. The pc is connected to the scanner though a X64-CL Express camera link (Dalsa, Waterloo, Ontario, Canada) that allows the acquisition of digital beamformed RF echo data.

Figure 2 illustrates a simplified signal flow through the scanner to the research interface. A set of broadband pressure pulses centered at 2-10 MHz are transmitted into the tissue. As these pressure waves propagate, they are partially reflected at interfaces formed by two materials having different acoustic impedances. The transducer, in receive mode, detects the reflected echos as they impinge on the individual elements. Each of these signals arriving from the transducer elements are then processed by a beamformer to form one coherent signal. To e.g. form a 2-D image, this process is repeated for multiple angles or spatial positions. We refer to the echo data at the individual elements as "element RF data", element because it is the output of a single element, RF because the data spectrum is in the radio frequency band. The processed signal output from the beamformer is called "beamformed RF data" - this is the data that is accessible using the URI, and it will hereafter be referred to as RF data.

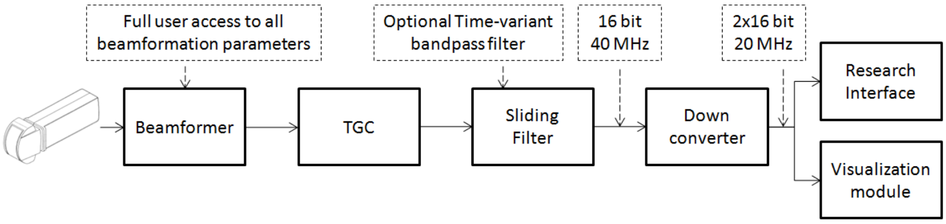


Figure 2: Simplified signal flow through the scanner

The RF data accessible through the URI is complex baseband signals stored as signed 16 bit integers, digitized at a rate of up to 40 samples/microsecond for all beam lines in a frame over a range of up to 22 cm. Users can also acquire pre-beamformed data by adjusting the receive aperture to a single element of interest. Acquired data is minimally processed because, aside from an optional time-variant bandpassfilter and beamforming, the only other processing is application of the time gain compensation (TGC) and transformation to I/Q data.

The acquisition of data is controlled via an in-house data grabber software module that features loading and saving of exact scanner settings for later experiment reproducibility. The data grabber module further enables the user to operate in two different modes:

1. **Standard mode**, in this mode the scanner is operating in factory default mode and standard scanner operation is available.

2. **Extended mode**, in this mode the user interface on the scanner is extented to enable control of various scanner settings, such as shoot sequence, receive- and transmit profiles, excitation waveforms and apodization functions.

Scanning in Standard mode the scanner is FDA approved and the grabber software captures standardized images as found in the clinic. Operating in Extended mode gives free access to all system parameters for beamformation, pulse shaping, and is applicable for clinical use. This enables researchers to capture experimental data that can be processed offline to evaluate new processing or beamformation methods. See Table 1 for a description of a subset of the parameter controls available for B-mode data acquisition in the extended scanner mode.

Data may be captured interleaved, switching between multiple basic mode setups, to maintain identical transducer, scanner, region of interest, and recording time on both the experimental- and standardized images. A basic mode setup is defined by the acquisition type, such as B-mode, M-mode, CFM-mode, power doppler

Table 1: Description of a subset of the parameter controls available in the extended scanner mode.

Parameter	Description
Dynamic focusing and dynamic apodization	Receive aperture dynamic focusing and aperture growth can be disabled individualy. When disabled, receive aperture size and focal position are fixed.
F#	Receive and transmit aperture size can be adjusted individualy. A maximum of 64 active elements is posible in standard mode and 128 elements in a synthetic aperture setting where rf data are acquired over two excitations.
Receive apodization	Receive apodization can be choosen from a fixed list of standard curves such as uniform or hamming weighting, or defined as a vector of element weights. If defined as a vector the curve can vary between individual image lines.
Receive time delay profile	Receive time delay profile can be specified individualy for each image line when dynamic focusing is disabled.
Line density	The image line density can be choosen from a range of one-half element pitch to two element pitch, in increments of one-half element pitch.
Speed of sound	Speed of sound can be defined in the interval from 1080 m/sec to 2500 m/sec.
Excitation waveform	Excitation waveform can be specified with a time resolution of 8.3 nsec and amplitude $\pm$ 1 or 0

mode or transverse oscillation. The ability to capture data interleaved enables processing on identical data in different ways, for assessment of different processing schemes.

The URI gives the researcher a high flexibility and enables multiple examinations to be performed in short time. The short time between examinations allows for a large database of processed images to be build; suitable for assessments where the specialists are off-site and where people who assess quality of the images must be independent of the aquisition. The length of the data sequences is only restricted by the memory on the external PC (one possible setup could be 20 seconds of interleave B-mode acquisition each with 192 image lines and depth of 11 cm). The data storage time is approximately 15.1 seconds for a 3 sec interleaved B-mode sequence including complete scanner settings and patient information. It is fast enough to obtain a sufficient number of scans under realistic operating conditions for valid and reliable statistical evaluations.

3.2 Data Management

Important aspects of data recording for clinical evaluations, is the ability to study under which conditions data were recorded (to be able to draw any conclusions from the data) and experiment reproducibility (to be able to reproduce the conclusions).

Data management is split into three new file formats

1. **RF data**, a file format with zero compression is developed to store the RF data from the scanner. The file format enables the user to load specified frames from a long data sequence without loading the entire data set first. RF data are stored as complex baseband signals as signed 16 bit integers.

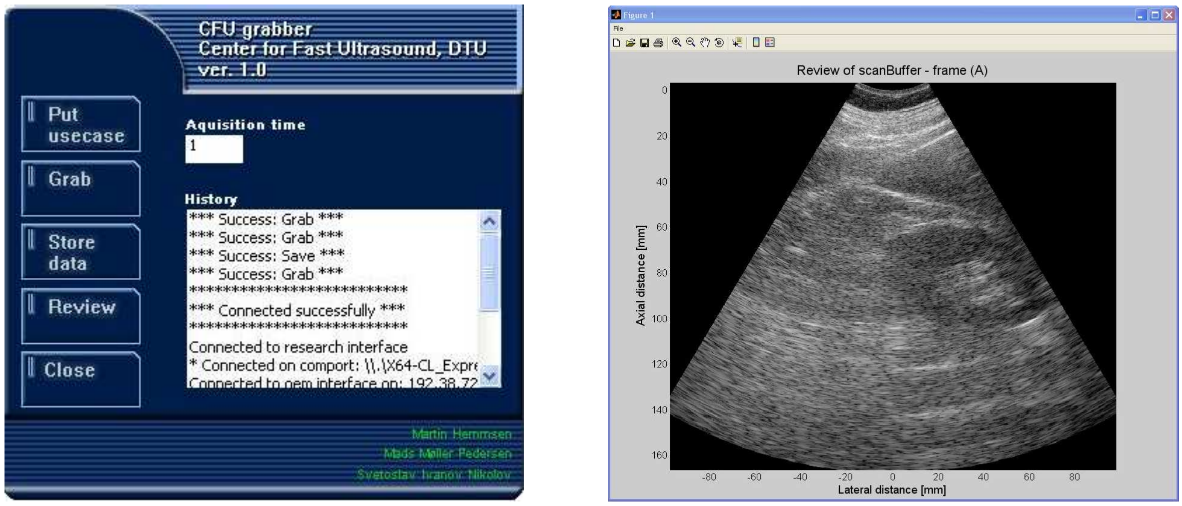
2. **Scanner parameters**, are stored at recording time. The parameter set is a complete description of the scanner setup and includes information such as beam geometry, probe name, transmit frequency, and TGC settings. The scanner parameters aids the user to redo experiments, generate images from the RF data, as well as creating simulation comparisons using tool such as Field II.[5]

3. **User Interface setup**, are stored at recording time. The parameter set is a full description of the user specified scanner setup and includes information such as zoom, overall gain, persistence, and various other visualization settings.

As a separate part of the URI, an open source, Matlab toolbox for basic file handling of the files collected with the URI is developed and available at http://server.elektro.dtu.dk/www/mah/. The file handling uses an open format develped in C++, available as a library and source code.

3.3 Graphical User Interface

The URI provides a simple graphical user interface, which offers the capability to load a given predefined scanner parameter set on the scanner, grab data to memory, review acquired B-mode data, and save data to disk. Figure 3a illustrates the GUI interface. Figure 3b illustrates a review of a B-mode scan of the right kidney and part of the right liver lobe. The URI implements a process running in the background featuring a service for communication with Matlab. See section 3.4 for more details.



(a) (b)

Figure 3: (a) Illustrates the Graphical User Interface which offers the capability to load scanner settings, grab data, save data and review data. (b) illustrates a review image of a B-mode scan in a standard Matlab figure.

The GUI is implemented in C++ and source code is available at http://server.elektro.dtu.dk/www/mah/, enabling users who are familiar with the C++ programming language to immediately develop customized client applications.

3.4 Matlab Control

As a separate part of the URI, an open source Matlab-based toolbox for remote control of the scanner was developed. The tools are available at http://server.elektro.dtu.dk/www/mah/, and provide access to a communication library developed in C++, see Table 2 for a subset of communication commands.

Table 2: Description of a subset of the commands available in the communication library.

Command	Description
Grab	This command initiate the URI to grab data to memory. The duration in seconds is specified as argument two, e.g. `TCPClient('Grab',10)`.
Review	This command initiate the URI to scan convert the data stored in memory and display the first B-mode frame, e.g. `TCPClient('Review')`.
Save	This command initiate the URI to acquire the scanner settings and save them along with the data stored in memory to disk, e.g. `TCPClient('Save','C:test.cfu')`. The resulting files saved to disk is test.cfu, test.dat and test.oem.
Put Usecase	This command loads a complete scanner parameter set on the scanner, e.g. `TCPClient('Put_usecase','test.dat')`.
OEM message	This command queries a message to the scanner and waits for reply, e.g. `TCPClient('OEM Message','Query:Gain')`.

Because the files are open source, users can download the toolbox and make customized functions that e.g. sets the scanner in a certain mode or build scripts for automization of recording procedures with e.g. varying parameters between each data acquisition.

3.5 Data Analysis

Based on earlier publications of studies of clinical evaluation between pairs of sequences[6] and recommended testing procedurs according to recommendation 500 from ITU-R[7] for subjective quality assessment, we propose a methodology for the assessment of subjective image quality and penetration depth of medical ultrasound imaging.

3.5.1 Movie generation

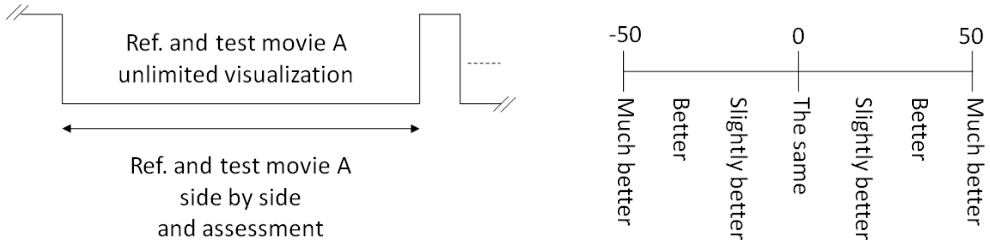
Scan line conversion and movie generation are performed in Matlab. The movies are generated using Matlabs build-in functions `avifile` and `addframe`, using zero-compression, to generate Windows AVI files. Data from an acquisition with multiple parameter setup is split into two movies, one for each parameter setup. In this way it is possible to generate both single image movies and paired movies where images are shown side-by-side.

3.5.2 Image quality assessment

The presentation method for assessment of image quality combines elements of the simultaneous double stimulus for continuous evaluation (SDSCE) method (ITU BT.500-11, Section 6.4) and the double stimulus continuous quality scale (DSCQS) method (ITU BT.500-11, Section 5). For reference, it may be called the simultaneous stimulus relative quality scale (SSRQS) method.

As with the SDSCE method, each trial will involve a split-screen presentation of material from two movies. One of the movie sources will be the reference (i.e., source movie), while the other is the test movie. The reference could be a conventional setup or the setup to compare against, and the test movie is the method under investigation. For both methods the parameters are optimized according to the diagnostic performance of the recording medium. Unlike the SDSCE method, observers will be unaware of the scanner conditions represented by the two members of the movie pair and the left-right placement of the movies are randomized.

As with the DSCQS method, a test session comprises a number of presentations, each with a single observer. Unlike the DSCQS method where the assessor only observes the stimulus two times and rates each stimuli, the assessor is free to observe the stimuli until a mental measure of relative quality associated with the stimulus is obtained. Figure 4a shows a basic test cell illustrating the presentation structure of reference and test material. Reference and test movies are displayed as matching pairs side-by-side with random left-right placement. Stimuli are visualized in a palindromic (playback may be reversed in time) display fashion in order to minimize discontinuity at the joints.



(a) (b)

Figure 4: Image quality assessment. (a) Basic test cell illustrating the presentation structure of reference and test material. Reference and test movies are displayed as matching pairs side-by-side with random left-right placement. Assessors are free to observe the stimuli until a mental measure of relative quality associated with the stimuli is obtained. (b) Visual analog scale (VAS) for image quality comparison between left and right stimuli.

The most often used criteria for manufacturers to implement new processing methods in their equipment is better diagnostic value compared to the existing method. Accordingly, a stimulus comparison scale, as described in ITU BT.500-11, Section 6.2, is recommended to be used. The specific judgement scale used is a non-categorical (continuous) scale, as described in ITU BT.500-11, Section 6.2.4.2, for reference it may be called Visual Analog Scale (VAS). During introduction of the assessors to the system and the rating methods, VAS is described with the same number of labels as on the ITU-R categorical comparison scale but with slighty modified labels (much better, better, slightly better, the same, slightly better, better, much better) to report the existence of perceptible quality differences and allow the random left-right placement of the stimuli. After introduction and during assessment the labels are hidden to avoid categorized data and to get a smoother distribution. Figure 4b shows the associated VAS for image quality comparison between left and right stimuli.

Judgement sessions consists of a series of assessment trials. These should be presented randomized, blinded, and independently of each other and, preferably, in a different random sequence for each observer. As with the judgement method described in ITU-R TG6/9[8] Section 7.1.1.3, each session shall involve two types of trials: test trials and check trials. However, each trial involves the display of the full width of the stimuli. The purpose of the check trial is to assess a measure of judgement bias. For each method under investitation, the following test trials are required for each test sequence:

Table 3: Description of the required test trials for each test sequence under investigation.

Left stimuli	Right stimuli
Reference sequence	Test sequence
Test sequence	Reference sequence

Preferably, there would be at least 2 repetitions of each of the cases above. For each method under investigation, the following check trials are required for each test sequence:

Table 4: Description of the check trials for each test sequence under investigation.

Left stimuli	Right stimuli
Reference sequence	Reference sequence
Test sequence	Test sequence

Again, preferably there would be at least 2 repetitions of each of the cases above.

The judgement sessions should be divided into sittings not more than one hour in duration separated by 15-minute rest periods. Assessors are instructed to evaluate which of the two presented stimuli is better on a visual analog scale. Figure 5 illustrates the GUI assosiated with the rating process of image quality. The assessment of penetration depth follows the assessment of image quality.

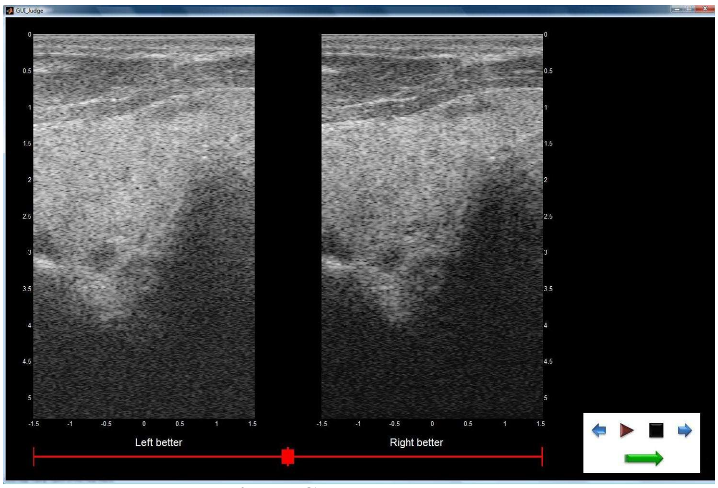
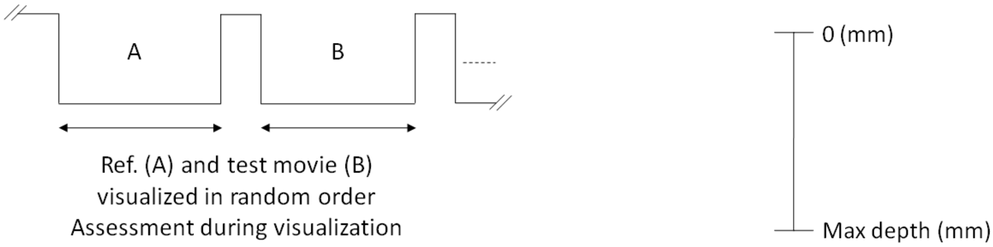


Figure 5: Illustration of the GUI assosiated with the rating process

3.5.3 Penetration assessment

The presentation method for assessment of penetration depth combines elements of the double stimulus continuous quality scale (DSCQS) method (ITU BT.500-11, Section 5) and the non-categorical judgement methods (ITU BT.500-11, Section 6.1.4.3). For reference, it may be called the sequential stimulus absolute scale (SSAS) method.

As with the DSCQS method, a test session comprises a number of presentations, each with a single observer. Unlike the DSCQS method where the assessor only observes the stimulus two times and rates each stimuli, the assessor is free to observe the stimuli until a mental measure of penetration depth associated with the stimuli is obtained. Figure 6a shows a basic test cell illustrating the presentation structure of reference and test material. Reference and test movies are displayed in a randomized sequential order. As with the SSRQS method stimuli are visualized in a palindromic display fashon. Observers will be unaware of the scanner conditions represented by the shown stimuli. Figure 6b shows the associated absolute penetration scale.



(a) (b)

Figure 6: Penetration depth assessment. (a) Basic test cell illustrating the presentation structure of reference and test material. Reference and test movies are displayed individualy in randomized order. Assessors are free to observe the stimuli until a mental measure of penetration depth associated with the stimuli is obtained. (b) Measure of penetration depth.

Besides the already described evaluation method in section 3.5.2 for comparison of image quality of new processing methods with conventional methods, it's interesting to investigate the penetration depth. Accordingly, a non-categorical judgement method as described in ITU BT.500-11, Section 6.1.4.2 is recommended to be used. The specific judgement scale used is a numerical scale, where assessors assign a value to each stimuli that reflect its penetration depth. The range of values are restricted to the same dimension as the dimension of the stimuli (e.g. 0 mm to 100 mm). During introduction of the assessors to the system and the rating methods, the assessors were asked: "After what depth is the image quality not usable for reliable diagnostic use?". After assessment the differences between depths in matching image pairs (reference and test stimuli) are used for the statistical analysis in order to avoid the bias from different assessors, who undoubtedly would have different opionions on how to answer the posed question.

Judgement sessions consists of a series of assessment trials. These should be presented randomized, blinded, and independently of each other and, preferably, in a different random sequence for each observer. For each method under investitation, the following test trials are required for each test sequence:

Table 5: Description of the required test trials for each test sequence under investigation.

Stimuli
Reference sequence
Test sequence

Preferably, there would be at least 2 repetitions of each of the cases above.

The judgement sessions followes the assessment of image quality and follows the quidelines described in section 3.5.2. Assessors are instructed to evaluate at what depth the image quality is no longer usable for reliable diagnostic use on a numerical scale, where they assess the sequence by placing a horizontal bar at the respective depth. Figure 7 illustrates the GUI assosiated with the rating process of penetration depth.

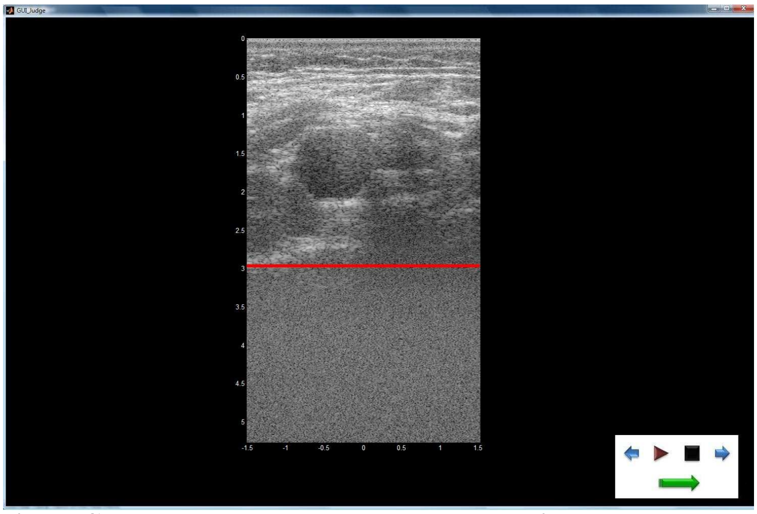


Figure 7: Illustration of the GUI assosiated with the assessment of penetration depth. The horizontal line (illustrated at 3 cm) is placed at the respective depth where the assessor evaluates the image quality no longer usable for reliable diagnostic use.

3.5.4 Statistical analysis

The following analysis is applicable to the results of SSRQS method and SSAS method for the assessment of image quality and penetration depth. In the first case, image quality is rated on a continuous scale indicating differences in image quality for the reference movie and test movie. The scale is defined as integer values between -50 and 50. In the second case penetration depth is rated on continuous scale indicating at what depth image quality is no longer usable for reliable diagnostic use. The readings from the scale is in millimeters between 0 and an arbitrary maximum equal to the size of the movies. Common for both methods is variations in the resulting distributions due to the differences in judgement between assessors and the effect of a variety of conditions associated with the experiment, for example, the use of several movies.

A test will consist of a number of judgement sessions, L, each with independent assessors. At each session, N independent sequence pairs will be presented, in some cases each pair will be presented a number of times, R.

Image quality

The statistical analysis of image quality is introduced, for each assessor, to test for any significant intraobserver variability with a Student's (one sample two-sided) t-test on the two cases from the test trials (Table 3). Secondly, judgement bias confined as a left-right bias for each assessor with a Student's (one sample two-sided) t-test on the two cases from the check trials (Table 4) is tested. Any assessors with a significant bias or significant variability shall be excluded in further investigations.

Since each assessor most likely has his own interpretation of the visual analog scale and shows different degrees of attraction to the center point in side-by-side image quality comparisons, no assumptions of normal distributed data can be made. Consequently, Wilcoxon signed-rank test with continuity correction could be used. The p-values of the pooled data should be corrected for multiple comparison using the Bonferroni method (Pedersen et al, 2006).

For a further detailed analysis of the distribution of ratings we propose to examine the median, 5% and 95% fractiles, with their associated confidence intervals. For each judgement session the standard error derived from N independent values with spread SD can traditionally be calculated as:

$$\sigma_i = \frac{Z * SD}{\sqrt{N}} \tag{1}$$

Where Z is 1.253 for the median and 2.108 for the 5% and 95% fractiles assuming a symmetrical (not skewed) distribution.

The standard deviation is the best measure of spread of an approximately normal distribution. This is not the case when there are extreme values in a distribution or when the distribution is skewed, in these situations interquartile range or semi-interquartile are preferred measures of spread. Interquartile range is the difference between the 25th and 75th centiles. Semi-interquartile range is half of the difference between the 25th and 75th centiles (StatsDirect).

For all assessors, the average of the median, 5% and 95% fractiles are then calculated. The standard error for each average is given as:

$$\sigma = \frac{\sqrt{\sum \sigma_i^2}}{L} \tag{2}$$

The confidence interval for the average of the median, $\bar{\mu}$, and each fractile can then be expressed as:

$$[\bar{\mu} - \delta, \bar{\mu} + \delta] \tag{3}$$

where:

$$\delta = t_{0.95}\sigma \tag{4}$$

The values of $t_{0.95}$ to be used in a confidence interval can be looked up in a table of the t distribution.

Penetration depth

The statistical analysis of penetration depth is performed with a student's (one sample two-sided) t-test on the resulting differences between sequence pairs, assuming normal distribution. In case of a significant difference it is relevant to examine the distribution of ratings and calculate the median, the 5% and 95% fractile together with their respective standard errors to be able to associate a confidence interval.

It is proposed to use the 95% confidence interval which is given by:

$$[\bar{\mu} - \delta, \bar{\mu} + \delta] \tag{5}$$

where:

$$\delta = t_{0.95}\sigma_m \tag{6}$$

The values of $t_{0.95}$ to be used in a confidence interval can be looked up in a table of the t distribution. The standard error σ_m can be derived from $M = N * R * J$ independent values with standard deviation σ and can traditionally assuming normal distribution be calculated as:

$$\sigma_m = \frac{Z * \sigma}{\sqrt{M}} \tag{7}$$

Where Z is 1.253 for the median and 2.108 for the 5% and 95% fractiles assuming a symmetrical (not skewed) distribution.

4. RESULTS

A system for acquisition and statistical evaluation of image sequences has been developed, based on a commercial available ultrasound scanner connected to a standard pc. Data acquisition features subject data recording, loading / saving of exact scanner settings for later experiment reproducibility, free access to all system parameters for beamformation and is certified for clinical use. The free access to all system parameters enables the ability to switch between standard mode and extended mode to capture standardized images as found in the clinic and experimental data from new processing or beamformation methods. Data may be captured interleaved, switching between multiple setups, to maintain identical transducer, scanner, region of interest and recording time on both the experimental- and standardized images. Data storage time is approximately 15.1 seconds pr. 3 sec sequence including complete scanner settings and patient information, which is fast enough to get sufficient number of scans under realistic operating conditions, so statistical evaluation is valid and reliable.

5. CONCLUSION

This work presents a methodology for clinical evaluation of image quality, which addresses the main problems in assessing clinical ultrasound image quality. The evaluation methodology should ensure the validity of the assessment, as it separates the developer, investigator, and assessor once a research protocol has been established. This separation eliminates any confounding influence on the result from the developer and new processing schemes is not driven by the developers, but by the clinical value.

We further present a research platform with free access to all system parameters for beamforming and with certification for clinical use. The clinical usability of the scanner, including the frame rate, is unaffected by activating the research interface.

The capabilities of the research interface module are fourfold; it allows one to:

- Acquire beamformed RF data to a file or memory on a remote pc running Matlab for offline processing. RF data are stored as complex baseband signals as signed 16 bit integers with a sampling rate of up to 40 MHz.
- Free access to all system parameters for beamforming and with certification for clinical use.
- Save and Load complete scanner parameters for experiment reproducibility.
- Remote control of scanner setup and acquisition from Matlab, enabeling automation of parameter studies.

As the core capabilities (saving and loading of complete scanner settings and interleaved RF data acquisition between multiple scanner setups) are available through a simple user interface on a standard pc, the research interface is well suited to obtaining data for clinical trials.

We believe that the research interface platform and the methodology for performing clinical evaluation of image quality can contribute to accelerated advancements in the diagnostic value of ultrasound imaging by allowing more ultrasound researchers to test and clinically evaluate promissing new methods in a standardized way.

ACKNOWLEDGMENTS

This work is sponsored by grant from the Danish Science foundation and BK Medical. Special thanks to Theis Lange for discussion and valuable input.

REFERENCES

[1] Tapiovaara, M., "Review of relationships between physical measurements and user evaluation of image quality," *Radiat Prot Dosimetry* **129 (1-3)**, 244–248 (2008).

[2] Jensen, J. A., Holm, O., Jensen, L. J., Bendsen, H., Pedersen, H. M., Salomonsen, K., Hansen, J., and Nikolov, S., "Experimental ultrasound system for real-time synthetic imaging," *IEEE Ultrason. Symp.* , 1595–1599 (October 1999).

[3] Brunke, S., Insana, M., Dahl, J., Hansen, C., Ashfaq, M., and Ermert, H., "Errata - an ultrasound research interface for a clinical system," *Ultrasonics, Ferroelectrics and Frequency Control, IEEE Transactions on* **54**, 198–210 (January 2007).

[4] Wilson, T., Zagzebski, J., Varghese, T., Chen, Q., and Rao, M., "The ultrasonix 500rp: A commercial ultrasound research interface," *Ultrasonics, Ferroelectrics and Frequency Control, IEEE Transactions on* **53**, 1772–1782 (October 2006).

[5] Jensen, J. A., "Field: A program for simulating ultrasound systems," **10th Nordic-Baltic Conference on Biomedical Imaging, Vol. 4, Supplement 1, Part 1**, 351–353 (1996b).

[6] Pedersen, M. H., Gammelmark, K. L., and Jensen, J. A., "In-vivo evaluation of convex array synthetic aperture imaging," *UMB* **33**, 37–47 (2007).

[7] "Recommendation 500-11: Methodology for the subjective assessment of the quality of television pictures," *ITU-R* (1974-2002).

[8] "Expert viewing to assess the quality of systems for the digital display of motion pictures in theatres," *ITU-R TG6/9 (Digital Cinema)* (2002).