Development of a Database Driven Statistical Quality Control Framework for Medical Imaging Systems

Xinhong Ding, *IEEE*, A. Hans Vija, *IEEE*, Johannes Zeintl, *Student Member, IEEE*, Aarti Kriplani

Abstract– **Phantom studies are typically the first tests when the performance of the imaging system needs to be verified or when a problem needs troubleshooting. It is not always clear whether the resulting image indicates a systematic problem with the system. Often perceived artifacts are only due to residual calibration errors or noise (quantum and system noise), yet the system being at nominal performance.**

The goal is to develop a tool that facilitates the image quality assessment by comparing the performance of a test system to a set of reference systems for a comprehensive diagnosis of the image formation chain beyond NEMA [3].

In this paper, we report the development of a database oriented statistical quality control framework for medical imaging systems. Central to this framework are three components: (a) MiQ, a phantom analysis tool [2], which processes data from various common phantoms and extracts quantitative image quality features, (b) a database which stores images and associated image quality data, and (c) a statistical inference engine which performs various statistical tests using information from the database.

To demonstrate feasibility, we randomly selected five new normal Symbia® SPECT•CT systems from the assembly line and performed three Tomo acquisitions on each using a Data Spectrum cylindrical phantom with hot and cold inserts. We also simulated one "abnormal" system by introducing axial shift for some projection data acquired on that system. Tukey-Kramer multiple comparison test is used to rank the five reference systems and to test the difference between the "abnormal" system and the reference systems.

Manuscript received November 14, 2008.

Dr. Xinhong Ding is with Siemens Medical Solution USA, Inc., Molecular Imaging, Hoffman Estates, IL 60192, USA (e-mail: xinhong.ding@siemens.com).

Dr. A. Hans Vija is with Siemens Medical Solution USA, Inc., Molecular Imaging, Hoffman Estates, IL 60192, USA (e-mail: hans.vija@siemens.com).

Johannes Zeintl is a PhD student with the University of Erlangen-Nuremberg, Institute of Pattern Recognition, Erlangen, Germany (e-mail: Johannes.Zeintl@nuklear.imed.uni-erlangen.de).

Dr. Aarti Kriplani is with Siemens Medical Solution USA, Inc., Molecular Imaging, Hoffman Estates, IL 60192, USA (e-mail: aarti.kriplani@siemens.com).

I. INTRODUCTION

Image quality assessment is performed to either evaluate the system's performance and/or quality state, or with a given system at a given quality state, to evaluate the information content of the image for a specific task.

Typically, phantom studies are the first experiments to be conducted for quality control (QC) to verify the performance of the imaging system or to troubleshoot a problem. However, it is not always clear whether the resulting image indicates a systematic problem with the system. Often perceived artifacts are only due to residual calibration errors or quantum and system noise.

Similar difficulty exists in the assessment of image quality when noise is inherent in the image. It is not always clear if the difference in observed images is caused by the deviation of the system (including algorithms), or simply due to the fluctuation of the noise in the data. Therefore a comparison to a database is preferable, to assess if the sample of an image is consistent with the ensemble. The ensemble is either a digital database or the experience of an expert reader.

Medical imaging assessment is traditionally performed by human observers based on individual experience. While this is still necessary, the individual human observer has some limitation in objective image quality assessment. One of the major limitations is the inability to correlate the observed image quality with historical data on the same or on different scanners of the same type. In addition, in many cases, different human observers may draw different conclusions for the same studies.

It is thus highly desirable to develop a tool that facilitates image quality assessment for a comprehensive diagnostic of the image formation chain beyond NEMA [3], and which takes into consideration the statistical nature of image quality assessment as exemplified by a collection of empirical reference data.

Toward this goal, we have developed a statistical image quality assessment framework that explores the knowledge

drawn from an image quality (IQ) database. Figure 1 below illustrates the basic concept of this framework.

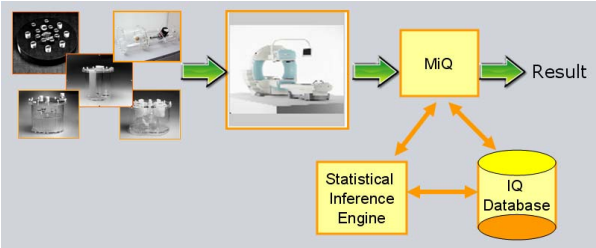


Figure 1. Concept of image quality assessment framework.

As shown in Figure 1, this framework is designed for image quality assessment based on various phantoms. It has three main components:

(a) MiQ, a phantom analysis tool [2], which processes data from various common phantoms and extracts image quality features,
(b) an image quality database which stores images and associated image quality data, and
(c) a statistical inference engine which performs various statistical testing using information from the database.

II. DATABASE DRIVEN STATICAL FRAMEWORK

A. *Phantom Analysis Tool*

This tool analyzes several commonly used phantoms and has the following main functionalities:

- image registration (manual or automatic)
- automatic segmentation
- quantitative characterization of image features
- comparison to image quality database

We now briefly list the phantoms and the corresponding quantitative features the tool extracts.

The NEMA 3 cylinder phantom (Figure 2) (Data Spectrum, Hillsborough, NC) was originally designed for PET acceptance testing with NEMA standard [3]. But it can also be used to evaluate the count rate, uniformity, attenuation compensation, and scatter compensation for SPECT.

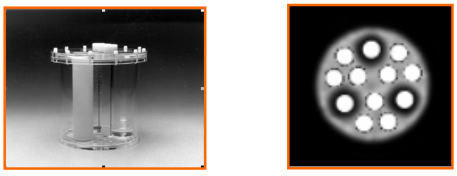


Figure 2. Data Spectrum NEMA 3 cylinder phantom and the region of interests (ROIs) generated by MiQ phantom analysis tool.

For NEMA 3 cylinder phantom, the MiQ phantom analysis tool first automatically locates the three cylinders corresponding to air, bone, and water, respectively, based on the segmentation done on CT images. The tool then defines nine background regions as shown in Figure 2. Based on this region of interests (ROIs), the air to background count ratio, the bone to background count ratio, and the water to background count ratio are calculated. In addition, the following positive uniformity and negative uniformity are also calculated, where the positive uniformity is defined as:

$$U_p = +100 \cdot \frac{max\{C_1, \cdots, C_9\} - \hat{C}}{\hat{C}}(\%), \quad (1)$$

and the negative uniformity as:

$$U_n = -100 \cdot \frac{\hat{C} - min\{C_1, \cdots, C_9\}}{\hat{C}}(\%), \quad (2)$$

where C_i is the average count of i^{th} background ROI, and $\hat{C}$ is the average of all C_i's.

The RMI 467 Electron Density CT (EDC) phantom (Figure 3) manufactured by Gammex (Middleton, Winsconsin, USA) contains various rods of different material. It is mainly used in conjunction with a CT scanner to establish the relationship between the electron density of various tissues and their corresponding CT numbers. This phantom can be used to evaluate the image quality of the CT as well as the accuracy of the corresponding mu-map values at different energy levels.

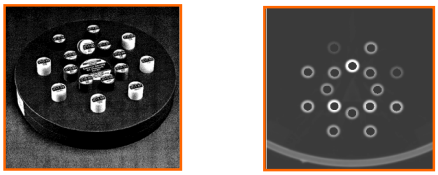


Figure 3. Gammex EDC phantom and the segmentation generated by MiQ phantom analysis toolkit.

For the EDC phantom, the MiQ phantom analysis tool first automatically segments the rods based on the CT images. It then identifies the material of each rod, and records the corresponding HU values of these rods. The measured HU values are then transformed to proper mu-map values for given energy level.

Figure 4 shows the cardiac torso phantom manufactured by Data Spectrum together with the 17 segment polar map generated by the MiQ phantom analysis tool. To generate this polar map, the tool must first segment the myocardium. This was achieved through a proper 3D registration of the phantom data with a reference synthetic heart model which has known orientation and other myocardial parameters. Based on segmented myocardium, the MiQ phantom analysis tool can generate several quantitative image features such as the uniformity of the polar map, the wall thickness of the myocardium, and the volume of the myocardium [2].

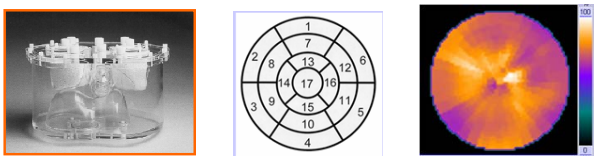


Figure 4. Data Spectrum cardiac torso phantom and the 17 segment polar map generated by MiQ phantom analysis tool.

In addition to the static cardiac torso, the MiQ phantom analysis tool is also capable of processing the dynamic cardiac torso phantom (Figure 5) manufactured by Data Spectrum.

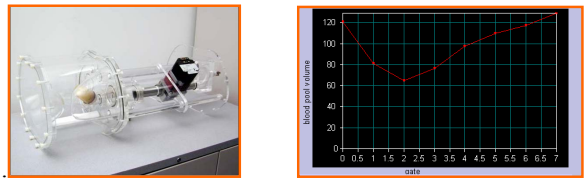


Figure 5. Data Spectrum dynamic cardiac torso phantom and the ejection fraction curve generated by MiQ phantom analysis tool.

This dynamic cardiac torso phantom is mainly used to assess image quality for gated cardiac studies. The MiQ phantom analysis tool uses the same technique described previously for the static cardiac torso phantom to locate the myocardium in each gate. The tool then estimates the volume of the myocardium in each gate using a curve fitting technique to overcome the problem that the image under consideration may not be contiguous. The MiQ phantom analysis tool then derives the ejection fraction based on the estimated volumes (Figure 5).

Figure 6 below illustrates the idea of the curve fitting method to estimate the missing part in a circular short axis heart slice.

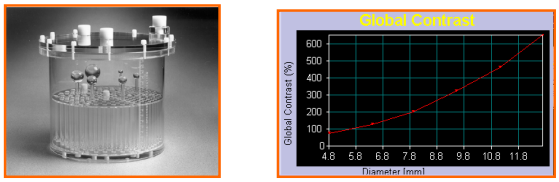
Figure 6. Left: the short axis view of a slice of cardiac phantom data. Right: curve fitting to estimate the radius of the circle.

Let (x_i, y_i), i=0,…,n be the set of available data (the left image in Figure 6). Then the objective of the curve fitting is to estimate the radius of a circle that best fits the data. The least square estimate of this radius is calculated as follows:

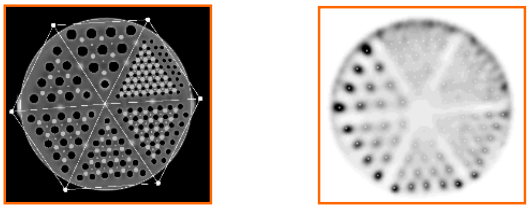
$$\hat{r} = \frac{\sum_i (x_i - x_o)cos(\theta_i) + \sum_i (y_i - y_o)sin(\theta_i)}{n} \qquad (3)$$
$$\theta_i = tan^{-1}(\frac{y_i}{x_i})$$

where (x_o, y_o) is the center of the circle.

Another commonly used phantom in quantitative image quality assessment is the cylindrical phantom with rod or sphere inserts made by Data Spectrum (Figure 7). The MiQ phantom analysis tool is also capable of processing this phantom. This tool first uses the CT image to find the center locations of each rod and to define the associated background (Figure 7(c)). These rod and background positions are then mapped from CT to SPECT images by applying the proper coordinate transformation and registration between the two modalities. For each sector of the rods, the MiQ phantom analysis tool then calculates, among other things, the contrast C(%) and the uniformity U(%). Figure 7(b) gives an example of the contrast curve as it is plotted across the diameters of different rods.



(a) Phantom (b) Contrast curve



(c) CT segmentation (d) SPECT segmentation

Figure 7. Data Spectrum cylindrical phantom and the quantitative features generated by MiQ phantom analysis tool.

The contrast C and the uniformity U in a sector with n rods and m background ROIs are calculated as follows:

$$C = 100 \cdot \frac{ave\{A_1, \cdots, A_n\} - ave\{B_1, \cdots, B_m\}}{ave\{B_1, \cdots, B_m\}}(\%), \qquad (4)$$

where A_i and B_j denotes the average counts inside the i^{th} rod and j^{th} background ROI, respectively.

$$U = 100 \cdot \frac{max\{B_1, \cdots, B_n\} - ave\{B_1, \cdots, B_n\}}{ave\{B_1, \cdots, B_n\}}(\%), \qquad (5)$$

where B_i denotes the average counts inside the i^{th} ROI.

B. Database

The image quality database is based upon the existing *syngo*® Dicom database that comes with the Siemens Symbia® SPECT•CT system. Unlike the traditional image database, the quantitative image quality features generated by MiQ phantom analysis tool will also be part of the database.

Figure 8 below shows the three main components of this database.

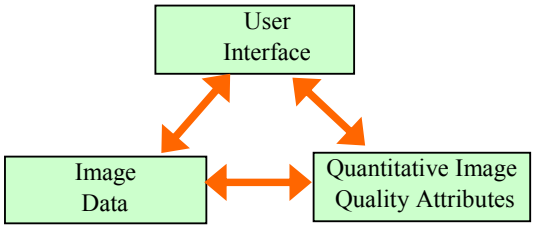


Figure 8. Three main components of the data base

C. *Statistical Inference Engine*

Many problems in quantitative image assessment can be formulated as a multiple comparison problem. For example, the quality of an unknown system can be assessed by comparing certain image quality features measured on this particular system to that of several known systems. As another example, one might be interested in comparing different images acquired from different acquisition protocols, or different images reconstructed using different algorithms or parameters. It is well known that for multiple objects, the statistical comparison in general can not be simply reduced to multiple pair-wise statistical comparison [4].

Instead of using the traditional approach of hypothesis testing which is built on distribution moments, we propose to utilize the multiple comparison method approach which is built on order statistic [4], because we believe it is more suitable to this problem.

III. Materials and Methods

A. *Data Acquisition and Processing*

To demonstrate feasibility, we randomly selected five new Symbia SPECT•CT systems from the assembly line, each with proper clinical QC, and acquired SPECT data on each using the Data Spectrum cylinder phantom with the Deluxe hot and cold rod inserts. The images were then reconstructed using Siemens Flash3D iterative reconstruction without attenuation correction or scatter correction. A CT scan of the cylinder phantom was performed. The acquisition was repeated three times, with all acquisition and reconstruction parameters held constant, and not optimized for best performance. We also randomly selected a test system and y-shifted (axial) some of the projection frames to simulate a poorly calibrated "abnormal" test system.

The images from the five reference systems (from left) and one test system are shown in Figure 9.

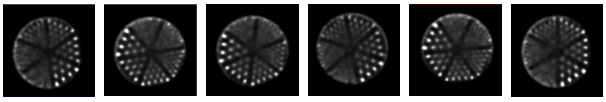


Figure 9. Images from reference systems and one test system

Clearly, a human observer in general can not tell, based on these images alone, if the differences among them are statistically significant or not.

B. *Quantitative Image Quality Feature Extraction*

The reconstructed images (NM) were then processed by the MiQ phantom analysis tool which registered the NM data with the CT, segmented the rods based on the CT images, and generated for instance the contrast C and uniformity U measures as described in the previous section.

After normalization to the max value, the raw measured data is shown in Tables 1 and 2 below.

TABLE 1. NORMALIZED HOT RODS CONTRAST DATA

Acquisition	Systems					Abnormal
	1	2	3	4	5	Test system
1	0.84	0.5	0.8	0.78	1	0.83
2	0.94	0.87	0.76	0.76	0.99	0.61
3	0.91	0.86	0.78	0.72	0.94	0.85

TABLE 2. NORMALIZED NON-UNIFORMITY DATA

Acquisition	Systems					Abnormal
	1	2	3	4	5	Test system
1	0.89	1	0.9	0.8	0.84	0.92
2	0.85	0.84	0.96	0.84	0.87	1.09
3	0.81	0.89	0.98	0.82	0.92	1.23

C. *Multiple Comparison*

Using the data from table 1 and table 2, we first conduct the following hypothesis test:

$$H_o: \mu_1 = \mu_2 = \cdots = \mu_5$$
$$H_a: \text{at least two } \mu_s\text{' are different} \qquad (6)$$

to answer the question:
a) Are the five systems statistically identical in terms of the measurement quantities?

We then compare the abnormal test system with the five reference systems by testing the hypothesis:

$$H_o: \mu_1 = \mu_2 = \cdots = \mu_5 = \mu_{test}$$
$$H_a: \text{at least two } \mu_s\text{' are different} \qquad (7)$$

to answer the question:
b) Is the unknown system different from the reference systems?

The following Tukey-Kramer multiple comparison test is used to do the test and to rank the systems.

$$\bar{y}_i - \bar{y}_j \pm q_{\alpha,\nu,k}\sqrt{\frac{1}{2}s^2(\frac{1}{n_i}+\frac{1}{n_j})}, \qquad (8)$$

where k is the number of treatments, $\bar{y}_i$ and $\bar{y}_j$ are the sample means of treatments i and j with sample size n_i and n_j, respectively, s^2 is an estimate of the experimental error, and $q_{\alpha,\nu,k}$ is the tabulated value of the studentised-range statistics at the α confidence level with ν degree of freedom [4].

IV. Result

Ranking of the reference systems: at α=0.05, for both the contrast data and the non-uniformity data, the test shows no difference for the five reference systems. The systems can be ranked as S2 ≤S4 ≤S3 ≤S1 ≤S5 based on Tukey's 95% confidence intervals, where Si denotes the i^{th} system.

Testing the "abnormal" system: at α=0.05, the contrast data comparison test does not demonstrate significant difference (sig=0.10); but the non-uniformity data comparison test does demonstrate a significant difference (sig=0.02) between the "abnormal" system and the reference systems. In particular, Tukey's 95% confidence intervals reveal that the "abnormal test system" is different from system 1 and 4. It is interesting to note that the simple paired t-test, which treats the five reference systems as a whole, will not detect the significant difference at this level.

V. Future Work

The next step is to test the sensitivity of the method by systematically introducing imperfections of increasing magnitude. We subsequently wish to catalogue artifact features and correlate them to system defects or suboptimal calibrations.

VI. Acknowledgment

We thank Dr Ron Malmin, Dr Manjit Ray, and Phillip Ritt for various help in collecting the data. We thank Dr Eric Hawman and Dr Amos Yahil for stimulating discussions. We also thank Dena Holzer for administrative support and editing.

References

[1] A. H. Vija, E. G. Hawman, and J. C. Engdahl, "Analysis of a SPECT OSEM reconstruction method with 3D beam modeling and optional attenuation correction: phantom studies." Nuclear Science Symposium Conference Record, 2003 IEEE Volume 4, Issue, 19-25 Oct. 2003 Page(s): 2662 - 2666 Vol.4.

[2] J. Zeintl, A. H. Vija, J. T. Chapman et al., "Quantifying the Effects of Acquisition Parameters in Cardiac SPECT Imaging and Comparison with Visual Observers." Nuclear Science Symposium Conference Record, 2006. IEEE Volume 6, Issue, Oct. 29 2006-Nov. 1 2006 Page(s): 3251 - 3257.

[3] National Electrical Manufacturers Association (NEMA). Performance Measurements of Scintillation Cameras. Standards Publication NU 1-2007.Washington DC: NEMA; 2007.

[4] Design of Experiments: Ranking and Selection. edited by Thomas J. Santner and Ajit C. Tamhane, Marcel Dekker Inc. 1984.