Image Quality Testing: Selection of Images for Assessing Test Subject Input

John Jendzurski and Nicholas G. Paulter
National Institute of Standards and Technology
Gaithersburg, USA

Eddie Jacobs
University of Memphis
Memphis, USA

Francine Amon
SP Technical Research Institute of Sweden
Borås, Sweden

Al Bovik and Todd Goodall
University of Texas at Austin
Austin, USA

Abstract— **Determining image quality is dependent to some degree on human interpretation. Although entirely subjective methods of evaluating image quality may be adequate for consumer applications, they are not acceptable for security and safety applications where operator interpretation may lead to missing a threat or finding threats where they do not exist. Therefore, methods must be developed to ensure that the imagery used in security and safety applications are of sufficient quality to allow the operator to perform his job accurately and efficiently. NIST has developed a method to quantify the capability of imagers to provide images of sufficient quality to allow humans to perform specific perception-based tasks. A one-time human-perception based step is required that results in perception coefficients that are combined with lab-measured objective image quality indicators (IQIs) to calculate image quality. This work uses a d′ evaluation method to examine the performance of test subjects in the human-perception based step, which was identification of a fire hazard in a set of grey-scale infrared images.**

Keywords-image quality; human performance; infrared imaging

I. BACKGROUND

The ability to objectively qualify imagers for a given security or safety function is typically, if not always, based on image quality results deduced from human perception studies. Human perception varies significantly, from person to person, and time to time for a given person. Consequently, without an enormous number of human test subjects, it is not likely that the results of human perception testing will yield an accurate, reproducible, and consequently reliable measure of the quality of an image and of the ability of an imager to provide a quality image. We examined this problem and conceived an alternative process, one that relies on objective, reproducible, and accurate laboratory-based testing of specific image quality indicators (IQIs) and on a one-time human perception study. The human perception study results in coefficients that multiply the values of lab-measured IQIs to achieve a single parameter that can describe the performance of the imaging system:

$$P_{perf} = f\left(\sum_{i=1}^{N} c_i x_i + \sum_{i,j=1}^{N} c_{ij} x_i x_j + \ldots + \sum_{i,j,\ldots,n=1}^{N} c_{ij\ldots N} x_i x_j \ldots x_n\right) \quad (1)$$

where c_i are the human perception coefficients, x_i are the lab-measured IQI values, N is the number of IQIs that describe the performance of the imaging system, and P_{perf} is the target imager performance value. This formula includes all possible IQI product terms.

This method was applied to the thermal imaging cameras (TICs) used by firefighters, which will be the focus of this paper. For the TICs, we defined four IQIs that describe its performance: spatial resolution, thermal contrast, noise, and brightness [1]. All four of these IQIs can be objectively measured in the lab using fixed , well characterized targets. However, these values are not directly useful to determine whether an image is of sufficient quality for a human to perform a perception task because of the interdependence of the IQIs and because the TIC operators (firefighters) are trained to seek clues in the images.

In the computation of (1), it was noted that only the first and second order terms had any significant contribution (less than 0.1 % of the total value) to P_{perf} [1]. Consequently, we redefine P_{perf} as:

$$P_{perf} = f\left(\sum_{i=1}^{N} c_i x_i + \sum_{i,j=1}^{N} c_{ij} x_i x_j\right) \quad (2)$$

II. ASSESSING TEST SUBJECT PERFORMANCE

A. Image Set and Test Subject Selection

The extraction of accurate and representative values of c_i is necessary to predict the ability of an imager to provide images of sufficient quality for an operator to perform a perception-based task. Determining the c_i is dependent on test subject input, therefore, it is necessary to assess whether the test subject inputs are valid (not simply guesses). The focus of this

paper deals with selecting an appropriate image set, S_{tsub}, for assessing test subject performance. If the test subject is determined to be guessing, then that test subject's inputs are not used to compute the c_i. The images in S_{tsub} should not be of such low quality that the ability of test subject to find a threat or hazard is no better than chance. This does not mean that images of exceedingly poor quality should not be shown to the test subject for computing P_{perf}, but the evaluation of the test subject's ability to provide useful information should not be based on such poor images. If a test subject is determined to be guessing, use of their results in the perception testing will adversely affect P_{perf}.

Concerns with the selection of an appropriate number of test subjects are described in [1], and in industry the recommendation ITU-R BT 500.11 states that at least 20 subjects are required to provide a reliable quality score. However, because we are focused on security and safety imagery, in which the operators typically are well trained, the training is consistent across the community, and the operators' performance is very similar, it is possible to use a small set (< 20) of test subjects to represent the larger population. This is contrary to the usual consumer application and interpretation of image quality because consumer performance is extremely variable and training is nonexistent. In that case, thousands of test subjects may be required.

A parameter often cited to evaluate the performance of a test subject in identifying a threat is d′, which is given by:

$$d' = Z(r_d) - Z(r_{fa}) \tag{3}$$

where r_d is the hit rate, r_{fa} is the false alarm rate, and $Z(\,)$ is the inverse of the cumulative Gaussian distribution. We will use this formula because of its widespread use and acceptance. The hit rate and false alarm rates are typically scored using the signal (item of interest) and response matrix shown in Figure 1.

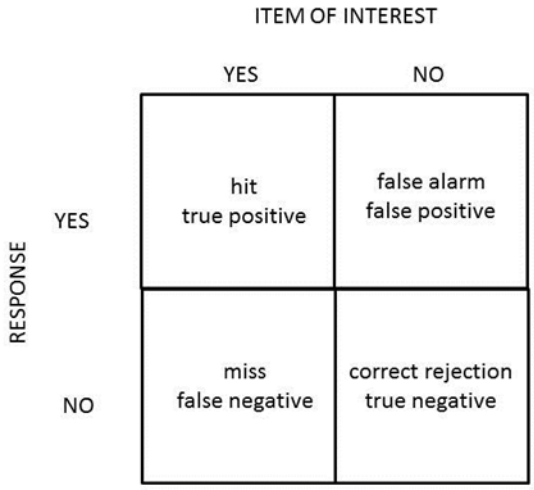

Figure 1: Signal and response matrix

The parameter d′ is used to determine which test subjects produced consistently high quality results. In order to improve the perception model that was based on (2) using the results from all of the test subjects, the results from these chosen test subjects can be used to build a more accurate perception model that does not include the negative effects of guessing.

B. Experiment

The images used were long-wave infrared grey-scale images showing residential and office indoor scenes. They were produced by an infrared (IR) imager having the following specifications: 640x480 array, 17 μm detector pitch, 50 mK noise equivalent temperature difference (NETD), dynamic range > 14 bit, and temperature range of -40 □C to 2000 □C. This imager's performance is superior to that of the TICs and so images produced by the IR imager were used as reference images. The grey-scale intensity of the reference image is based on the temperature of the objects in the scene. There were 180 reference images used in the experiments. Of this set, 150 contained an object that was a thermal hazard and 30 contained no thermal hazard but did contain at least two other innocuous thermal objects. Consequently, the d′ values we compute relate to a function of identification and not detection, the latter of which is the typical application of d′. The value of d′ for these two applications can be quite different, where the d′ for detection can be much larger than that for identification [2]. The reference images were not used to compute d′. Instead, the reference images were operated on to yield a much larger set of reduced quality images. Trained firefighters with experience in the use of TICs were the test subjects.

Two different sets of experiments were performed using the same 180 reference images. Phase 1 took place in 2007 and Phase 2 in 2011. Each reduced-quality image that was presented to a firefighter was obtained by varying computational parameters to yield the desired IQI values for that image. Since there are four IQIs, the total possible number of variations are dependent on the IQI range and increment. To limit the number of images presented to the firefighter, a reduced set of IQI values were selected. These values were determined to span the range of IQI space expected from a TIC. In the first set (Phase 1) of human subject observation experiments, the 180 reference images were modified to yield 25 different IQI sets per image, for a total of 4500 reduced-quality images. The values of IQI values in a given set are the same for all images. For example, the IQI values in set i $(i = 1, \ldots, 25)$ are nominally the same for any image $(j = 1, \ldots, 180)$. In the Phase 2 experiments, the IQI range was expanded to allow for extremely poor-quality images because there was a concern the Phase 1 IQI space was not adequate. In Phase 2, there were 55 different IQI sets per image, giving a total of 9900 reduced-quality images.

In both experiments, the firefighter was instructed to select a location in each image where the thermal hazard is thought to be located or to select the "no hazard" button if no thermal hazard is thought to be present.

C. Measuring d′

The images available for testing contained at least three thermal objects, where one may have been a thermal hazard. Consequently, the ability of the test subject to perform an identification function and not a more simple detection function was computed. As noted earlier, these different functions will yield different values of d′.

Using the following procedure, we defined for each image a detection area around the thermal hazard which corresponded to a correct threat identification. For each image, the centroids of threats and non-threats were computed by first converting the pristine grayscale images to binary images using a global threshold calculated using Otsu's method [3]. Then all connected regions in the binary image composed of less than eight pixels were disregarded. Regions were designated a threat or non-threat based on our knowledge of the thermal hazard location in the scene, and the centroids of the threat and non-threats were then computed. Finally the detection area was defined by a circle centered at the centroid of the threat object with a radius that was equal to half the distance between the centroid of the threat and the centroid of the nearest non-threat.

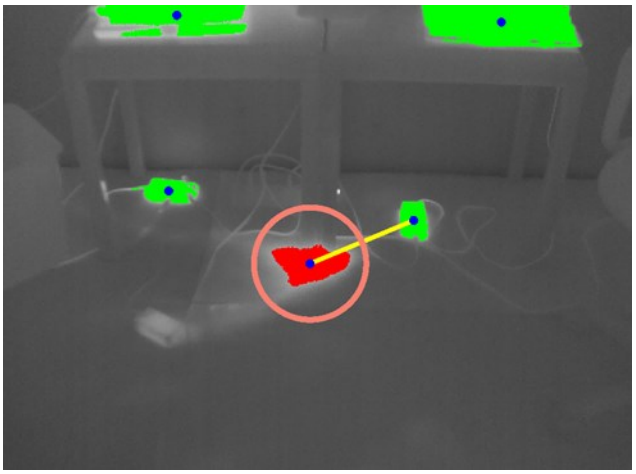

Figure 2: Pristine image showing the process of defining the detection area (indicated by the salmon colored circle).

With the detection area now defined, it was possible to view the test results and degraded images interactively and compare detection probabilities with the statistical properties of the degraded and pristine images to gain more insight. While viewing the images and test results, it was clear that for some degraded images the probability of detection was less than reciprocal of the number of objects in the image and therefore no better than simply guessing at random. The test results for these images were therefore not used in the computation of d′.

	Phase 1	**Phase 2**
Total number of distorted images used in experiment	4500	9900
Number of images for which the probability of detection was lower than random guessing	1185	7762

Table 1: Distorted images that produced probabilities of detection below the probability of guessing at random were not used in the computation of d′.

Values of d′ for the Phase 1 and Phase 2 experiments were calculated using (3). The results are shown below in Figures 3 and 4. It is important to note that these values of d′ are based on correct or incorrect identification of the threat objects' locations as opposed to a simpler determination of whether a threat was or was not present in the image.

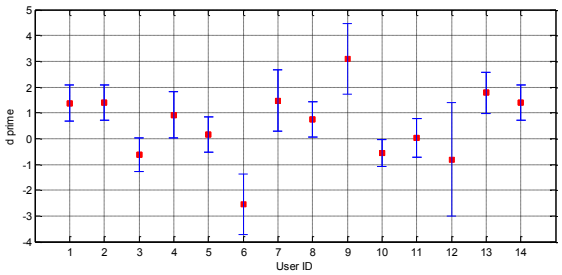

Figure 3: d′ calculation for Phase 1. The red squares indicate the average value for d′, and the blue lines indicate the standard deviation of the values.

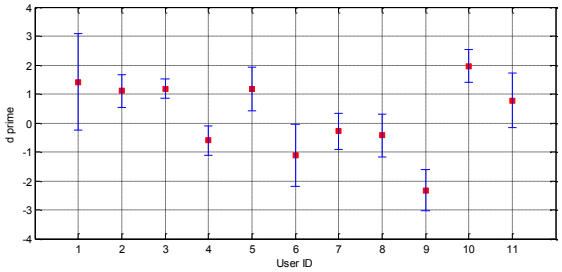

Figure 4: d′ calculation for Phase 2. The red squares indicate the average value for d′, and the blue lines indicate the standard deviation of the values.

III. Observations and Lessons Learned

The d′ calculations may be used to determine which firefighters' results are more relevant to the computation of the human performance coefficients. Previously we computed the human perception coefficients using the results of all firefighters. In light of the d′ calculations, we can compute the human perception coefficients again, this time disregarding results from firefighters who performed poorly. Poor performance here is interpreted to be at or below zero d′ values.

In the future experiments additional images could be presented that enable d′ for detection (in addition to d′ for identification) to be computed. These would have only one thermal object (no clutter), and rather than locate the threat, the firefighter would be instructed to simply indicate if the object is or is not a threat.

A real-time computation of d′ may also be implemented to gauge the attentiveness of the firefighters while the experiment is being conducted so that their attention may be redirected as necessary.

REFERENCES

[1] F.K. Amon, D. Leber, and N. Paulter, “Objective Evaluation of Imager Performance, “2011 Fifth International Conference on Sensing Technology, 28 Nov 2011, Palmerston North, New Zealand, pp. 52 – 57.

[2] S. Dehaene, L. Naccache, G.L. Clec’H, E. Koechlin, M. Mueller, G. Dahaene-Lambertz, P.-F. van de Moortele, and D.L. Bihan, “Imaging unconscious semantic priming,” Nature, Vol. 395, Oct 1988, pp. 597 – 600.

[3] Otsu, N., "A Threshold Selection Method from Gray-Level Histograms," IEEE Transactions on Systems, Man, and Cybernetics, Vol. 9, No. 1, 1979, pp. 62 – 66.