Size Discrimination in Computed Tomographic Images
Effects of Feature Contrast and Display Window

STEVEN E. SELTZER, MD, RICHARD G. SWENSSON, PHD, PHILIP F. JUDY, PHD, AND RICHARD D. NAWFEL, MS

Seltzer SE, Swensson RG, Judy PF, Nawfel RD. Size discrimination in computed tomographic images: effects of feature contrast and display window. Invest Radiol 1988;23:455–462.

Studies show that features on computed tomographic (CT) images in clinical formats become less detectable when the images are produced with wider CT display windows. We studied the effects of feature contrast and the display window on observer performance in higher-order tasks that involved discriminating small size differences between features on CT images. The features to be discriminated were pairs of disks (9.0 or 9.5 mm in diameter) superimposed on CT images of water phantoms. Sets of image stimuli for two different types of size-discrimination tasks were generated with various CT contrasts specified for the superimposed features and were produced on film transparencies with display windows ranging from 90 to 2880 Hounsfield units (HU) in width. Observers' performance improved with increasing CT contrast in both size discrimination tasks. Unlike performance in feature-detection tasks, however, size discrimination was unaffected by changing the CT display window over a factor of 16 (from 90 to 1440 HU). Performance fell only at the widest display window (2880 HU), for which CT noise was essentially invisible. These results suggest that the effect of changing the CT display window may depend on the spatial frequency content of image information required for a given task.

Key words: observer performance; psychophysics; perception.

From the Department of Radiology, Harvard Medical School and Brigham and Women's Hospital, Boston, Massachusetts.

Supported in part by NIH Grants R01 CA40444 and R01 CA 43114 from the National Cancer Institute.

Reprint requests: Steven E. Seltzer, MD, Department of Radiology, Harvard Medical School, 25 Shattuck Street, Boston, MA 02115.

Received October 9, 1987, and accepted for publication, after revision, January 14, 1988.

THE INTERPRETATION of medical images relies on a variety of perceptual and cognitive processes that radiologists use to detect and identify abnormal findings. A long-term goal of our research is to understand these fundamental processes and how they are affected by the image information. Such an understanding may lead to the development of imaging and display techniques that can minimize interpretation errors.

The radiologist's ability to detect any type of focal abnormality will always be limited by physical factors that determine a lesion's visibility on the image. Small or low contrast lesions, for example, must be differentiated from noise variations that are present in the surrounding background of normal tissue. In many clinical situations, however, the radiologist's problem is not object visibility but the proper interpretation of some perceived object, ie, an image feature that is easily visualized and not attributable to random image noise. A detection decision means that the radiologist has chosen to interpret the given feature as pathologic, rather than as normal.

The decision that a lesion is present often may be made by comparing some questionable object to other features on the image, features that are known (or assumed) to represent normal anatomic structures. On a computed tomographic (CT) image of the liver or on a radiograph of the chest, for example, the radiologist may decide that a suspect feature represents a tumor because it is too large, of inappropriate brightness, or too peripheral to be a blood vessel.[1,2] On a mammogram, the radiologist may decide that a

suspicious feature is a tumor mass, rather than an agglomeration of ducts and normal parenchyma, because of its contour, density, form and size.[3] The perceptual processes involved in making such complex differentiations among the visible image features may differ considerably from those used in deciding whether or not an object is present at a given location on the image.

Laboratory experiments to evaluate systems for producing medical images have used simple detection tasks. To evaluate the noise characteristics of medical images, investigators have studied the visibility of features superimposed on uniform-noise backgrounds.[4-7] Typically, a feature's detectability is measured by how accurately observers can differentiate samples of noise (ie, particular locations or image areas) that contain the specified feature from samples that do not. Aside from any additional limitations that the visual display may impose, an observer's ability to distinguish between real and random variations on images appears to be governed by physical and statistical considerations. Experiments with features superimposed on CT images show that the measured changes in a feature's detectability are closely related to physical (cross-correlation) calculations of its signal-to-noise ratio (SNR) on the CT image. Values of SNR from a realized cross-correlator could predict the detectability effects produced by manipulating a feature's size and CT contrast (feature-background difference in CT number), the CT noise level and the noise correlations.[8-10]

Only recently have perceptual tasks that involve discriminating physical differences between separate features been employed in the evaluation of imaging systems,[11] although such discrimination tasks are common in psychophysical experiments with noiseless visual stimuli.[12-15] Just as the image noise limits a feature's visibility (and hence its detectability), it also will limit the ability to differentiate between visible image features that differ in their physical characteristics (eg, size, shape, contrast or edge sharpness). By defining a physical calculation of the Task SNR, a stochastic limitation on performance accuracy can be quantified for the particular task of discriminating a specified difference between the possible features in a noise background. In a recent experiment, such calculations were used to predict observers' discrimination of size differences between pairs of features superimposed on CT images.[16] The measured indices of feature discrimination for observers were proportional to calculations of the Task SNR for an image cross-correlator; both types of indices increased with the actual difference in feature size and with the features' CT contrast.

For CT images produced in clinical formats (film transparencies), several investigators report that observer performance can be limited by the choice of visual-display parameters. In particular, the use of wide CT display windows appears to reduce feature visibility,[17] and substantially degrades the detectability of large and small lesions.[18,19] The large reduction in feature detectability cannot be explained by a decreased SNR for images displayed at wider CT windows, which simply scale the CT variations (both real and random), or by the effects of unscaled physical noise added in the production of film transparencies.[20] Rather, these display effects seem to reflect some loss in the human observer's ability to visualize a lesion's small change in CT contrast when the displayed contrast variations (in film optical density) are compressed by using wider CT display windows (eg, greater than 120 HU). Some investigators[21-23] attribute these (and similar) results to sources of variability within the human observer (observer noise), assumed to represent further additive noise from the detection process. Similar display-scale effects on detection performance have been reported for features in other types of noise backgrounds,[23,24] but these effects have not been investigated using other types of perceptual tasks.

Perceptual distinctions among the visible objects on an image also might be sensitive to display-scale effects, eg, to the contrast scaling produced by changing the display window of a digital image. Because any physical differentiation between image features (eg, a difference in size) is masked by image noise, it (like feature detection) also will be limited by the features' physical contrast. For this reason, the compression of CT variations by wide display windows—which degrades an observer's ability to detect features—might also degrade an observer's ability to discriminate the differences between features on CT images.

This experiment investigated how changes in features' CT contrast and the CT display window altered observers' ability to discriminate a small size difference between pairs of features superimposed on CT images. In one discrimination task, observers considered only pairs of features that differed in size and simply had to determine which feature was smaller. In a more complex discrimination task, all of the possible pairs of larger and smaller features were considered; observers had to determine whether the two features were the same (both small or both large) or different in size. Within each task, different experimental conditions independently varied both the features' CT contrast and the width of the CT display window by factors of two in HU. These manipulations produced subsets of conditions that scaled the displayed CT noise by a factor of 32, while maintaining the displayed contrast of the features at a constant optical density in the image transparencies.

Materials and Methods

Visual Stimuli

The stimuli used in these experiments were generated from CT images of water phantoms, taken with a Delta 2020 scanner (Technicare, Solon, OH). The noise-power spectrum of these images has been measured and previously reported[25]; the pixel standard deviation was 15 HU. The stimulus images presented to the observers consisted of hexagonal arrays of six circular pedes-

tal areas (64 mm in diameter), radially equidistant from a seventh pedestal at the center of each image. The six outer pedestals contained the noisy stimuli to be discriminated and were created by adding a constant value of 55 HU to the pixel CT numbers (whose expected value was 0) in those areas of the image. The center pedestal, used to present noiseless examples of the possible stimuli, was created by replacing the pixel CT numbers within that circular area by the constant value of 55.

The stimulus pairs of dark features were produced by subtracting two image profiles from each pedestal area. These were the profiles of disks, blurred by the CT scanner's edge-response function and scaled to produce a specified level of CT-contrast. The two possible disk diameters were 9.0 mm and 9.5 mm, as scaled to the pixel size of the CT scanner (0.5 mm). For each stimulus pair, the two superimposed profiles were centered 19 mm apart on the pedestal's horizontal diameter. In the Two-Alternative (2-A) Task, the pairs of features always differed in size, with the smaller (9.0 mm) feature equally likely to appear on the left as on the right. In the Same-Different (S-D) Task, the two feature sizes were independent and equiprobable in both positions (same-size and different-size pairs were equally likely).

The sets of images, each containing 120 stimulus pedestals, were produced by means of different selected combinations of the features' CT contrast and the CT window width for the two discrimination tasks. Each set of images was photographed onto 4-on-1 film transparencies by a clinical multiformat camera (Deltamat 50 Technicare, Solon, OH) with 256 grey levels. The center of the display window (grey level 128) was set at the pedestal's CT value, which limited the maximum possible change in contrast to 50% of the range in grey levels. To avoid thresholding of the pixel CT values within dark features, the displayed contrast of these features never exceeded 30% of the maximum range. Densitometric measurements (from step wedges superimposed on the CT images) were used to convert the changes in display level to differences in optical density (OD) on the film transparencies. These measurements also verified that the grey level steps produced linear changes in OD over the used range (grey levels 1 to 196), which extended more than four standard deviations (of the pixel noise) above the mean (pedestal) level for individual pixels in any set of images.

Each set of image stimuli had features superimposed at one of eight CT contrast values (differing by factors of 2), and was displayed using one of six window widths (also differing by factors of 2). The particular combinations of these parameters were chosen to produce five different levels for the features' displayed contrast, each achieved with six different levels for the displayed CT noise. Table 1 gives the values of CT contrast that were required, for the six different display windows, to produce each level of displayed feature contrast (percent contrast or difference in film OD). Figure 1 shows examples of these stimuli, the same image taken from four different conditions in the S-D Task.

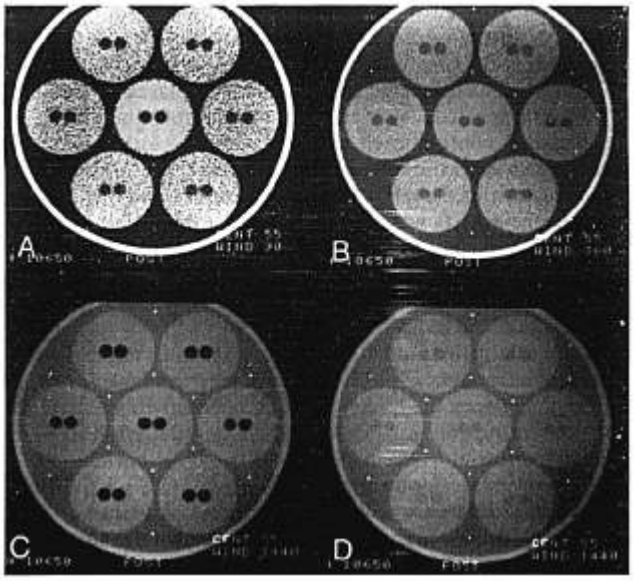

Fig. 1. The visual stimuli. (A), (B), (C), (D) show different values of window width, CT contrast (in HU) and display contrast (in film OD) used for the S-D Task in this experiment. (A) has CT contrast = 26.6 HU, display contrast = 0.70 OD, and window width = 90 HU. (B) has CT contrast = 26.6 HU, display contrast = 0.184 OD, and window width = 360 HU. (C) has CT contrast = 425 HU, display contrast = 0.184 OD, and window width = 1440 HU. (D) has CT contrast = 26.6 HU, display contrast = 0.055 OD, and window width = 1440 HU. The comparison among (A), (B), and (D) illustrates a window manipulation for a constant CT contrast (26.6 HU). (A) and (C) have the same display contrast (OD = 0.700), achieved with different combinations of CT contrast and window width. The comparsion between (C) and (D) illustrates the effect of a change in CT contrast at a constant window width (1440 HU).

TABLE 1. Values of Feature CT Contrast Used with the Six Display Windows to Produce Each Specified Level of Displayed Feature Contrast

Display Window Width (HU)	Displayed CT Noise Level*,† OD	%	CT Contrasts (in HU) For Each Specified Feature Contrast in Displayed CT Images* 0.700 OD 29.5%	0375 OD 14.8%	0.184 OD 7.4%	0.098 OD 3.7%	0.055 OD 1.8%
90	.367	16.7	26.6‡	13.3	6.7	3.3	1.7
180	.183	8.3	53.2	26.6	13.3	6.7	3.3
360	.092	4.2	106.4	53.2	26.6‡	13.3	6.7
720	.046	2.1	212.8	106.4	53.2	26.6	13.3
1440	.023	1.0	425.5‡	212.8	106.4	53.2	26.6‡
2880	.011	0.5	851.1	425.5	212.8	106.4	53.2

*Levels of contrast or noise in the displayed CT images are specified both as percentages of the displayed grey-level range (the window width in HU) and in the equivalent units of film optical density (OD).

†The pixel noise standard deviation (15 HU), as scaled by the display window.

‡Conditions used for the stimulus examples shown in Fig. 1.

Each successive row in Table 1 doubles the width of the CT display window, a scaling manipulation that decreases both the noise level and the feature contrast by factors of 2 in the displayed CT images. This inherent reduction in the displayed feature contrast can be avoided, however, by a compensatory doubling of the features' CT contrast each time the window width is doubled. Table 1 also gives the values of CT contrast that maintained the five predetermined levels of displayed feature contrast on the film transparencies. The sets of images used for the easier, 2-A Task were the 24 conditions in Table 1 that had values of the feature CT contrast between 6.7 and 212.8 HU; those used for the more difficult, S-D Task were the 20 conditions with CT contrasts between 26.6 and 851.1 HU.

Image calculations quantified how the stimulus manipulations changed the amount of physical information available for differentiating between the binary decision alternatives specified for each discrimination task. These calculations, described in a previous experiment,[16] were based on the two cross-correlations of image pixel values (for the two feature-locations in each pair) with the expected difference profile of the two possible features. As for the model outlined in the Appendix, the relevant decision variable depended on the discrimination task; it was either the difference in the two realized cross-correlations for each pair of features (2-A Task) or the absolute value of that difference (S-D Task). This physical decision variable provided an estimate of the Task SNR in each stimulus condition,[16] which characterized how accurately the image calculations could separate the binary alternatives defined for that discrimination task. The values of Task SNR were directly proportional to the features' CT contrast in both discrimination tasks. They increased from 1.2 to 38.2 in the 2-A Task and from 1.5 to 75.5 in the S-D Task and were not altered by the display window transformations of the pixel CT variations.

Previous investigations of physical quantization effects in digital images[26] indicate that display scaling does not affect SNR calculations or observer detection performance until the decreasing magnitude of the (scaled) noise standard deviation begins to approach the size of the discrete, grey level steps. The relatively high CT noise (15 HU/pixel), combined with the 256 grey levels in our stimulus images, kept even the widest display window (2880 HU) well within this quantization limit.

Observer Procedures

The images were read on a standard clinical illuminator that was masked to eliminate glare. Although there was no limit on viewing time, an entire set of images (120 stimulus pedestals) was always completed in a single reading session. As in a natural reading situation, the observers were free to move and to change positions; in addition, they were provided with a minification lens that could rapidly extend the effective viewing distance by a factor of about three.

All 44 sets of images from both tasks were read by three experienced observers (including one radiologist) who were familiar with CT images. Two of the observers completed all 24 image sets from the 2-A Task before reading the 20 sets from the S-D Task; the third observer reversed this procedure. Observers read these image sets in different order, constrained such that successive conditions changed in all three of the manipulated variables: the CT contrast and displayed contrast of the features and the CT display window.

The simpler 2-A Task presented only pairs of stimulus features that differed in size; observers used a six-category rating scale to indicate the likelihood that the smaller feature was located on the right, rather than on the left, within a given pair. The S-D Task presented both same size and different size pairs of stimulus features, however, and these two alternatives were equiprobable. For each stimulus pair, observers rated the likelihood that the two features differed in size, using a six-category scale, and then indicated which feature appeared to be the smaller in that particular noise background. The instructions asked observers to adopt rating criteria that produced roughly equal numbers of judgments in all six ordinal rating categories. Since the actual stimulus probabilities were known, the observers could use an initial inspection of the images to adjust their rating criteria appropriately for each condition.

Analysis of Performance

The analysis of observers' discrimination performance was based entirely on the likelihood ratings they assigned to the particular binary decision alternatives that were defined for each task. A rating receiver operating characteristic (ROC) curve that described the separation between these binary alternatives was generated from an observer's ratings of the 120 stimulus pairs within each set of images. Higher ratings in the 2-A Task indicated increasing confidence that a given pair of features (known to differ in size) had the smaller feature located on the right. The ROC curve was the proportion of such pairs (right smaller) that were correctly detected at each rating criterion, plotted against the corresponding proportion of falsely detected pairs (left smaller). In the S-D Task, higher ratings indicated increasing confidence that the two features of a given pair differed in their actual sizes. The ROC curve was the proportion of such different size pairs rated above each ordinal cutoff, plotted against the corresponding proportion of actual same-size pairs (both features large or both small). The analysis did not consider observers' judgments about the smaller feature's location in the S-D Task.

The realized ROC curves (for each observer, task and condition) were fitted using a previously developed, maximum likelihood procedure.[27-29] This procedure assumes that the true ROC curve is a linear function in normal-normal probability coordinates; it uses the sets of ordinal ratings assigned to the binary decision alternatives to estimate the linear-ROC parameters that characterize the two (assumed normal) distributions for the observers' decision variable. The separability of these binary decision alternatives was measured by the d_a index of discrimination accuracy, which is a monotonic transformation of the area below the observer's fitted ROC curve (the inverse normal transformation of this area, multiplied by $\sqrt{2}$).[30] This d_a index can be interpreted as an estimate of perceptual SNR for the observer's decision variable in the specified, binary decision task.

For sets of rating data that represent high levels of discrimination performance, the standard maximum likelihood procedure (described above) may be unable to obtain a two-parameter fit to the ROC curve. Such data sets often can be successfully fitted with a single-parameter version of this procedure, which uses a reduced set of likelihood equations to fit a symmetric ROC curve (assuming that the linearized ROC curve has slope = 1.0 and location parameter d_a). When a set of ratings is consistent with perfect discrimination performance, however, the ROC curve cannot be fitted by any parametric procedure and the d_a index is indeterminate. This was the case for nine sets of the observers' rating data in the present experiment, obtained from five conditions in the 2-A Task.

Results

Figures 2 and 3 present values of mean $d_a \pm$ SE, for the three observers' separate estimates, obtained from 19 of the

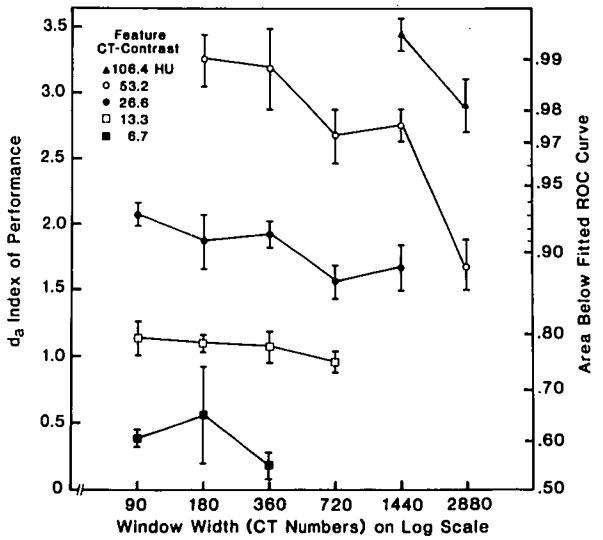

Fig. 2. Size discrimination performance in the Two-Alternative (2-A) Task. Mean values of the fitted indices of d_a (from the three observers' ROC curves) are plotted as a function of the window width (HU on a log scale) for features superimposed at five different levels of CT contrast. Vertical bars around each point indicate plus and minus one standard error of the mean-d_a estimate (from the interobserver variance in d_a).

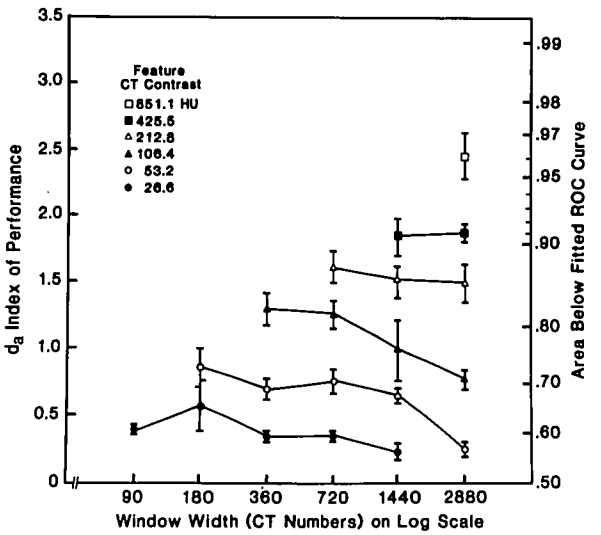

Fig. 3. Size discrimination performance in the Same-Different (S-D) Task. Mean values of the three observers' fitted indices of d_a are plotted as a function of the window width for features superimposed at six different levels of CT contrast. Vertical bars indicate ± SEM (from the interobserver variance).

24 conditions in the 2-A Task (Fig. 2) and from 20 conditions in the S-D Task (Fig. 3). These figures plot mean d_a as a function of the display-window width (on a log scale) and connect data points from conditions at each specified value of CT contrast for the superimposed features. Observers' size discrimination performance appeared to depend much more on the CT contrast of the features than on their displayed contrast in the film transparencies. In both tasks, the observers' performance improved with increasing feature CT contrast, which increased the Task SNR, but remained relatively constant across most manipulations of the display window, which did not affect task SNR.

Although the features were superimposed at lower CT contrasts in the simpler 2-A Task, the manipulations of CT contrast drove the observers' size-discrimination performance beyond its measurable range, given the sample sizes used for these image stimuli (60 of each type). As the feature CT contrast increased from 6.7 to 53.2 HU, the mean estimate of d_a increased from 0.37 ± 0.27 to 3.27 ± 0.28 (for windows of 180 and 360 HU), and the area below the observers' ROC curves increased from 0.60 to 0.99 (Fig. 2). Figure 2 does not include the data from the five conditions in which one or more of the three observers achieved perfect discrimination performance (indeterminately high d_a). This occurred when the feature CT contrast increased to 106.4 HU (windows of 360 and 720 HU) and 212.8 HU (windows of 720, 1440 and 2880 HU).

Manipulating the width of the display window from 90 to 1440 HU had little measurable effect on observers' size discrimination performance in the 2-A Task. However, that performance declined sharply for the two sets of images displayed with the widest window (2880 HU), in which the CT noise variations were essentially invisible (Fig. 2). These decreases in mean d_a, produced by doubling the window width (from 1440 to 2880), were about equal to the effects of reducing the values of feature CT contrast by one-half. That is, performance with the 106.4 HU features decreased to the level obtained for 53.2 HU features at narrower windows, whereas performance with the 53.2 HU features fell to that obtained for the 26.6 HU features. Once the width of the display window increased beyond 1440 HU, therefore, the observers' size discrimination performance in the 2-A Task appeared to depend on the features' displayed contrast (in OD units) on the film transparencies.

Size discrimination performance also was directly related to feature CT contrast (and hence to Task SNR) in the S-D Task, although much larger values of CT-contrast were required to drive observers' performance across the measurable range in d_a (Fig. 3). At the 1440 HU display window, an increase in the feature CT contrast from 26.6 to 425.5 HU increased the mean d_a estimates from 0.24 ± 0.06 to 1.85 ± 0.15, which increased the area below the observers' ROC curves from 0.57 to about 0.90. As seen in Fig. 3, the display window manipulations had little effect on observers' performance in this S-D task, even when the window's width increased to 2880 HU. Features that were superimposed at the highest CT contrast (851.1 HU) could not be displayed at windows smaller than 2880 HU without thresholding the CT values for many image pixels (by the window's lower boundary). At window 2880, however, these 851.1 HU

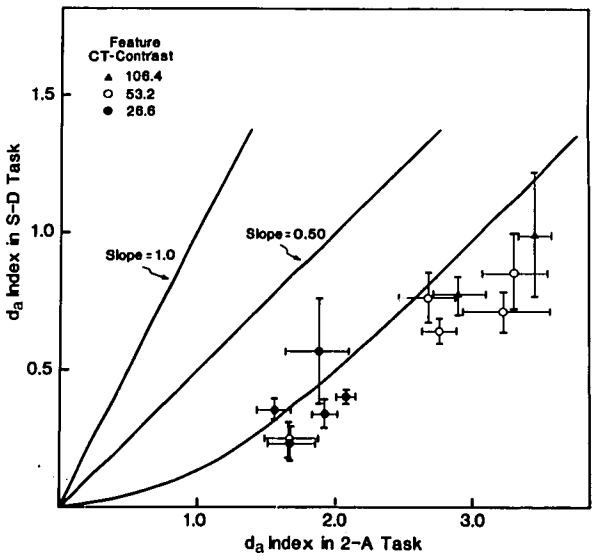

Fig. 4. Comparison of size discrimination performance in the 2A and S-D Tasks. Each data point plots the three observers' mean values of d_a in the two different tasks for one of the 12 physically identical stimulus conditions. The value of d_a in the S-D Task is plotted as a function of its value in the 2-A Task. The smooth curve shows the expected theoretical relation between these performance measures for the observer model described in the Appendix. After an initial slow increase, the predicted function becomes essentially linear with a slope of about 0.5 (see Appendix).

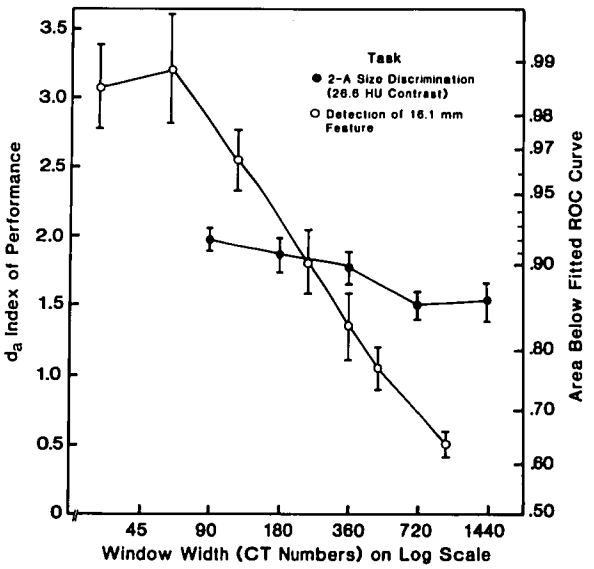

Fig. 5. Comparison of display window effects for detection and size discrimination performance. Mean values of the fitted d_a indices are plotted as a function of the window width for ten observers who read stimuli from a previous feature detection experiment[19] (16.1 mm features; CT contrast = 31.7 HU, Task SNR = 5.7), and for five observers in the present size discrimination experiment (2-A Task; CT contrast = 26.6 HU, Task SNR = 4.8).

features produced a mean d_a of 2.46 ± 0.18. This was substantially better than performance with the 425.5 HU features at display window 1440, which had the same displayed contrast (0.70 OD units) but more displayed CT noise (0.023 vs 0.011 OD/pixel) on the film transparencies.

For physically identical conditions, ie, those using the same feature CT contrasts and display windows to produce the images, observers' performance always was considerably better in the simpler 2-A Task than in the S-D Task. This change in performance could be predicted from a model of the observer's decision process that quantifies the increase in difficulty for the more-complex S-D Task (see Appendix). Figure 4 plots the relation between these mean values of d_a (S-D Task vs. 2-A Task) for the 12 comparable sets of images that yielded estimates of d_a from all three observers in both tasks.

The smooth curve in Fig. 4 is a theoretical relation between the expected values of d_a in these two size discrimination tasks, from the observer model described in the Appendix. The model predicts a slow initial increase in d_a for the S-D Task, as d_a increases in the 2-A Task, which becomes an essentially linear function for which the slope is about 0.5 (see Appendix). Although this theoretical relation has no fitted parameters, it provides a fairly good description of the large measured change in observers' performance between the two size discrimination tasks. Since the measured data points in Fig. 4 tend to fall below this theoretical curve, however, the S-D Task was more difficult than the model would predict, given observers' ability to perform the simpler 2-A Task with physically identical stimuli.

Discussion

Display Window Effects

This experiment shows that observers' ability to discriminate small size differences between features in CT images was almost unaffected by changing the CT display window from 90 HU to 1440 HU in width, which decreased both the feature contrast and CT noise by a factor of 16 in the displayed CT images. These results are quite unlike the large display-window effects found for observers' detection of features in similar CT images.[18,19]

Figure 5 compares detection and discrimination, plotting values of mean $d_a \pm$ SE from both a size discrimination task and a feature detection task as a function of the display window width. The stimulus formats and the physical characteristics of the CT images were identical in both tasks. The size discrimination results show data for five observers in the present 2-A Task (CT contrast = 26.6 HU), which augment the results shown in Fig. 3 by data from two additional observers. The feature detection results show data for ten observers who read sets of image stimuli taken from a previous detection experiment.[19] In this detection

task, observers rated the likelihood that a low-contrast, disk feature (16.1 mm in diameter) was superimposed at the center of each stimulus pedestal. As seen in Fig. 5, wider CT display windows produced a dramatic decrease in feature detectability (visibility on the CT image), but essentially no change in feature discriminability.

These different results for feature detection and size discrimination tasks demonstrate that the perceptual effects of particular image manipulations (eg, changes in display parameters) may depend on the specific perceptual tasks that observers are asked to perform. No physical analysis of the image information can explain these results, since display window manipulations did not affect the realized calculations of SNR in either the detection or discrimination tasks. Factors that affect the visibility of features on displayed images, such as the display window width, might not affect the appreciation of those physical characteristics serving to distinguish between features that are easily visible. Similar results were reported by Tsui et al,[11] who found that image spatial resolution is critical in making subtle distinctions between objects of different shapes but relatively less important to their overall visibility.

Higher Order Perceptual Tasks

A characterization of tasks by Hanson[31] suggests how the present results might generalize to other types of perceptual tasks. He classified tasks as lower order or higher order, depending on how the task performance is limited by the image noise variations at different spatial frequencies, ie, the noise-power spectrum (NPS). Lower order tasks, such as feature detection and brightness discrimination, use information at lower spatial frequencies to determine the amplitude of the feature. Higher order tasks, such as size discrimination, use information in a broad band of spatial frequencies to determine the precise locations of the edges of features. The convolution-back-projection algorithm used in reconstructing CT images leads to a ramp-shaped NPS that increases with spatial frequency (up to the Nyquist-frequency). Hanson's analysis implies that any white noise (constant NPS) added to this CT image noise would produce a relatively greater reduction in the task SNR for a lower order task than for a higher order task. If the added white noise were unaffected (unscaled) by changes in the CT-image display, then changing the CT display window, which scales the CT noise, would have a greater effect on performance in detection tasks (lower order) than in size discrimination tasks (higher order).

Unscaled additive noise, attributed to intraobserver variability (observer noise), is a mechanism that has been invoked to explain why display scaling degrades observers' detection performance, both for CT images[22] and for images containing spatially uncorrelated (white) noise.[21,23] Hanson's analysis, applied to the NPS for CT images, suggests how an observer noise mechanism might be able to explain the results shown in Fig. 5, ie, display window effects that are large in feature detection tasks but small in size discrimination tasks. Such a pattern of results would be predicted from the ramp-like NPS of CT images if the observer noise simply added constant noise power at all spatial frequencies. This assumption of a constant observer NPS implies that display window effects should be predictable from the NPS of the physical images, given the spatial frequency requirements of the particular task. For images with spatially uncorrelated noise (white), however, the assumption of a constant observer NPS predicts no task-dependent effects of scaling the image noise, eg, by changing the display window. In particular, display window manipulations of uncorrelated noise images should have precisely the same effects on observers' performance in both detection and size discrimination tasks.

Conclusions

Most laboratory detection tasks model a highly simplified interpretation problem that is rarely encountered in clinical practice, ie, disease indicated by the presence of some low-visibility object (feature) in an otherwise uniform noise background. More commonly, radiologists must perform higher order perceptual tasks to determine that an abnormality is present. That is, the features that provide information about the presence of disease often are readily visible on the image, but these pathologic features must be distinguished from those that represent the normal anatomic structures. The relevant physical information always is degraded by noise, which will limit the radiologist's ultimate efficiency in performing any perceptual task.

Images produced by digital techniques (eg, CT, MRI or DSA) can be displayed with expanded contrast, which would improve the radiologist's efficiency for simply detecting the presence of a low contrast lesion in a uniform background. However, these window manipulations may not improve actual diagnostic efficiency if the essential elements of the clinical diagnostic task are higher order perceptual tasks. Our results for CT images suggest that variations in the CT display window will not limit the radiologist's diagnostic efficiency as much as would be indicated by simple detection experiments.

Appendix

This section presents a heuristic model for observer performance in the present size discrimination tasks. The model is used to predict the expected d_a index of discrimination accuracy in the S-D task from its value in the 2-A task with physically identical stimuli. The same model also can be applied to cross-correlation calculations on realized CT images, which are known to satisfy the model's assumptions.[16] This is the procedure that was used to estimate the Task SNR for conditions in the present experiment. A previous experiment,[16] which varied the sizes and CT contrasts of features in S-D size discrimination tasks, found that the estimates of Task SNR were proportional to the d_a indices of performance accuracy obtained from observers' fitted rating-ROC curves.

An observer's perception of the four possible pairs of the two feature sizes can be represented as four different distributions in a two-dimensional

space, for which the x and y axes relate to the observer's separate perceptual impressions about the left and right features of each pair. We assume these distributions are bivariate-normal in form with a correlation of 0 (ie, x and y are independent). Changing the feature size in a given location, from smaller to larger, is assumed to increase the mean of the normal variable by the same amount for both x and y without changing the standard deviation. It is convenient to set $x = 0$ and $y = 0$ midway between the two mean values on each axis. The 2-A Task, which presented only the two pairs of different size features, produces two distributions in this space, for which the expectations lie along the 45° negative diagonal ($y = x$). The S-D Task, which presented all four feature-pairs, produces four distributions: the two distributions for different size pairs (included in the 2-A Task) and the two distributions for same size pairs, located along the 45° positive diagonal ($y = x$).

The decision variable required for these size discrimination tasks is the difference of the two perceived features sizes, $v = y-x$, which represents a projection of the two-dimensional distributions into a single dimension (on the negative diagnonal). For the 2-A Task, this process yields two normal distributions of a unidimensional decision variable (v), associated with the standard ROC curve. The value of d_a inferred from the observer's fitted ROC curve in the 2-A Task is an estimate of the separation of the two distribution means (at $\pm d_a/2$), with v scaled to a standard deviation unit.

For the S-D Task, this projection of the four two-dimensional distributions yields only three distinct (equal-variance, normal) distributions of the unidimensional variable $v = y-x$. The two types of different size pairs produced separate distributions with positive and negative means ($\pm d_a/2$, identical to those in the 2-A Task), but both types of same size pairs yield an identical difference distribution on v with a mean of 0. The relevant decision variable for distinguishing these two types of different size pairs from the same size pairs, ie, the binary alternatives specified for the observer's rating task, is $|v|$, the absolute value of this difference variable. This is because larger perceived differences in either direction (both positive and negative values of v) are more likely to arise from different size pairs than from the same size pairs.

The absolute value decision variable ($|v|$) is positively skewed and far from normally distributed, particularly for the distribution of same size pairs. Consequently, d_a(S-D), the accuracy index for an observer's rating decisions in the S-D Task, was obtained from direct calculations of the ROC curve generated from the distributions of the two binary alternatives on the variable $|v|$. (Because the distributions of the two types of different size pairs are symmetric about $v = 0$, both map into the identical distribution on $|v|$.)

The three underlying normal distributions on v, with mean values $-d_a/2$, 0, $+d_a/2$ and unit standard deviations, were specified by an assumed value of d_a in the 2-A Task. These specified distributions were used to derive the two distributions of $|v|$, for same size and different size pairs, which then generated the predicted ROC curve for same-different decisions based on this absolute value decision variable. The area below this expected ROC curve was obtained by direct calculation (using the trapezoidal rule), and converted into an estimate of d_a(S-D) for the S-D Task by taking the inverse normal transformation of this area, multiplied by $\sqrt{2}$.[30] The theoretical curve in Fig. 4 was produced by using various assumed values for d_a(2-A) to calculate the corresponding values of d_a(S-D) predicted by this model.

Figure 4 shows that d_a(S-D) is predicted to increase slowly for small values of d_a(2-A) but to approach a linear function with a slope of about 0.5 as d_a(2-A) becomes larger. Some insight about the form of this function can be obtained by considering the function predicted for a simpler, known direction S-D Task. The simpler task would specify the actual size of one feature in each pair a priori; then v, the difference variable for a given feature pair, could arise from only two possible distributions, ie, with means 0 or $d_a/2$ in one case and means 0 or $-d_a/2$ in the other. The d_a index of performance in this known direction S-D Task would be exactly 1/2 its value in the 2-A Task, making the predicted function d_a(2-A)/2. As seen in Fig. 4, the uncertainty about the direction of the possible size difference (in the present S-D Task) decreased this predicted function by about a constant amount (0.52) for larger values of d_a: ie,

$$d_a(\text{S-D}) \approx [d_a(\text{2-A})/2] - 0.52, \text{ for } d_a(\text{2-A}) > 2.0.$$

References

1. Moss AA. Computed tomography of the hepatobiliary system. In: Moss AA, Gamsu G, Genant HK, eds. Computed tomography of the body. Philadelphia: W.B. Saunders, 1985:665.
2. Simon G. Principles of chest x-ray diagnosis. London: Buterworths, 1978:19-20.
3. Tabar L, Dean PB. Teaching atlas of mammography. New York: Thieme-Stratton, 1985:18-19.
4. Strum R, Morgan R. Screen intensification systems and their limitations. AJR 1949;62:617-634.
5. Gould R, Judy P, Baarngard B. Noise characteristics of a microchannel plate image intensifier. Med Phys 1978;5:115-119.
6. Cohen G, BiBianca F. The use of contrast-detail-dose evaluation of image quality in computed tomographic scanner. J Comput Assist Tomogr 1979;3:189-195.
7. Hay GA. Traditional x-ray imaging. In: Wells PNT, ed. Scientific basis of medical imaging. New York: Churchill-Livingstone, 1982:37-38.
8. Judy PF, Swensson RG, Szulc M. Lesion detection and signal-to-noise ratio in CT images. Med Phys 1981;8:12-23.
9. Judy PF, Swensson RG. Detection of small focal lesions in CT images: effects of reconstruction filters and visual display windows. Br J Radiol 1985;58:137-145.
10. Judy PF, Swensson RG. Detectability of lesions of various sizes on CT images. Society of Photo-Optical Instrumentation Engineers 1985;535:38-42.
11. Tsui BMW, Metz CE, Beck RN. Optimum detector spatial resolution for discriminating between tumor uptake distributions in scintigraphy. Phys Med Biol 1983;28:775-788.
12. Engle FL. Visual conspicuity, visual search and fixation tendencies of the eye. Vision Res 1977;17:95-108.
13. Jenkins SE, Cole BL. The effect of the density of background elements on the conspicuity of objects. Vision Res 1982;22:1241-1252.
14. Cole BL, Jenkins SE. The effect of variability of background elements on the conspicuity of objects. Vision Res 1982;22:261-270.
15. Cohn TE, Makous W. Detection and identification. J Opt Soc Am [A] 1985;2:1455.
16. Judy PF, Swensson RG. Size discrimination of features on CT images. Society of Photo-Optical Instrumentation Engineers 1986; 626:225-230.
17. Hemmingsson A, Jung B, Naslund L, Ytterbergh C. Perceptibility of experimental and clinical lesions in the CT image with and without image processing. Acta Radiol [Diagn] 1981;22:67-76.
18. Warren RC, Pandya YV. Effect of window width and viewing distance in CT display. Br J Radiol 1982;55:72-74.
19. Twible DA, Judy PF, Swensson RG. Effects of the CT display window on the detectability of large and small lesions. Society of Photo-Optical Instrumentation Engineers 1984;486:204-208.
20. Judy PF, Swensson RG, Kijewski MF, Seltzer S. Perceptual evaluation of a laser multiformat camera. Radiology 1986;161(P):150.
21. Goodenough DJ, Rossman K, Lusted LB. Factors affecting the detectability of a simulated radiographic signal. Invest Radiol 1973;8:339-344.
22. Warren RC. Detectability of low-contrast features in computed tomography. Phys Med Biol 1984;29:1-13.
23. Ishida M, Doi K, Loo N, Metz CE, Lehr JL. Digital image processing: effect on detectability of simulated low-contrast radiographic patterns. Radiology 1984;150:569-575.
24. Burgess AE. On observer internal noise. Society of Photo-Optical Instrumentation Engineers 1986;626:208-213.
25. Kijewski MF, Judy PF. The noise-power spectrum of CT images. Phys Med Biol 1987;32:565-575.
26. Burgess A. Effect of quantization noise on visual signal detection in noisy images. J Opt Soc Am [A] 1985;2:1424-1428.
27. Dorfman DD, Alf E. Maximum-likelihood estimation of parameters of signal-detection theory and determination of confidence intervals-rating method data. J Math Psych 1969;6:487-496.
28. Grey DR, Morgan BJT. Some aspects of ROC curve-fitting: normal and logistic models. J Math Psych 1972;9:128-139.
29. Swets JA, Pickett RM. Evaluation of diagnostic systems: methods from signal detection theory. New York: Academic Press, 1982: 208-232.
30. Judy PF, Swensson RG. Display thresholding of images and observer detection performance. J Opt Soc Am [A] 1987;4:954-965.
31. Hanson KM. Variations in task and the ideal observer. Society of Photo-Optical Instrumentation Engineers 1983;419-60-67.