Object detection in cluttered infrared images

Kjell Brunnström, MEMBER SPIE
Bo N. Schenkman
Bengt Jacobson
Acreo AB
Electrum 236
164 40 Stockholm
Sweden
E-mail: Kjell.Brunnstrom@acreo.se
bosch@nada.kth.se

Abstract. Implementation of the Johnson criteria for infrared images is the probabilities of a discrimination technique. The inputs to the model are the size of the target, the range to it, and the temperature difference against the background. The temperature difference is calculated without taking the background structure into consideration, but it may have a strong influence on the visibility of the target. We investigated whether a perceptually based temperature difference should be used as input. Four different models are discussed: 1. a probability of discrimination model largely based on the Johnson criteria for infrared images, 2. a peak signal-to-noise ratio model, 3. a signal-to-clutter ratio model, and 4. two versions of an image discrimination model based on how human vision analyzes spatial information. The models differ as to how much they try to simulate human perception. To test the models, a psychophysical experiment was carried out with ten test persons, measuring contrast threshold detection in five different infrared backgrounds using a method based on a two-alternative forced-choice methodology. Predictions of thresholds in contrast energy were calculated for the different models and compared to the empirical values. Thresholds depend on the background, and these can be predicted well by the image discrimination models, and better than the other models. Future extensions are discussed.
[DOI: 10.1117/1.1531637]

Subject terms: infrared; Johnson; detection; psychophysics; spatial; vision model; masking.

Paper 020015 received Jan. 17, 2002; revised manuscript received Jul. 11, 2002; accepted for publication Jul. 15, 2002.

1 Introduction

With a camera that detects infrared (IR) radiation and shows it on a display, it is possible to see objects that the naked eye cannot see. Such cameras sense energy in the thermal IR wavelength region and show the temperature of different objects. The temperature scale is converted into a color scale or a gray scale on a display and in this way an image is obtained. This kind of camera can, for example, image a human through smoke in a burning house or heat leaking from a house. It is also possible to image objects when there is no reflected light, for example, at night.

To predict whether a particular camera is suited for a specific application, for example surveillance of military targets by human observers, it is common to analyze whether the targets in question can be detected, recognized, or identified in the specified situation. At least since 1958, the so-called Johnson criteria (Johnson[1]) have been used to specify the performance when using IR cameras. Johnson criteria determine 50% discrimination of a target. Johnson divided visual discrimination into four subtasks, namely detection, orientation, recognition, and identification. This work was intended for image intensifier systems, but was later extended to include IR camera systems. Johnson used scale models of eight military vehicles and one soldier against a homogeneous background (Holst[2]). Observers were asked to specify the lowest contrast where they could detect, orient, recognize, or identify the objects. For this lowest contrast, determined for each category and object, a bar pattern was shown on a monitor. The bar pattern was changed in spatial frequency until it was just resolvable, and the number of bar cycles that fitted into the smallest dimension of the object was noted (see Fig. 1). The average number of cycles across the minimum dimension of all the objects in the different categories is shown in Table 1, taken from Johnson[1] and Holst.[2] The detection criterion for 50% detection at 1.0 cycles is reportedly used for low or medium clutter images. Other frequency values have been recommended for higher clutter levels.

Since Johnson, much work has been done improving the Johnson criteria. The number of cycles per minimum dimension for each task has been changed to today's industrial standard,[2] where detection is 1.0, recognition 4.0, and identification 8.0 cycles, respectively. Orientation is not included. Another way to determine the minimum dimension has also been introduced. The Johnson criteria with this kind of minimum dimension are called two-dimensional Johnson criteria and are calculated by taking the square root of the length and height of the object.

For relating the equivalent number of cycles of the Johnson criteria to a camera, specific transfer function is measured, i.e., the minimum resolvable temperature difference (MRTD or MRT). This is a measure of an observer's ability to resolve a four-bar pattern through an IR camera under test.[3] MRTD is a sensor parameter that is a function and not just a single value.[4] The function gives the relation between the lowest temperature difference in a target that can be resolved on a monitor by an observer at different

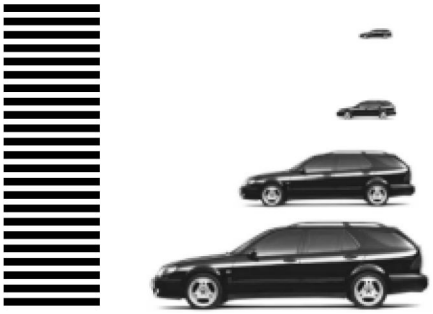

Fig. 1 The principle behind the Johnson criteria, being based on the number of cycles, shown by the bar pattern to the left at a distance, where it will be just resolvable (about 3 m). The images to the right representing detection, orientation, recognition, and identification, where the height of each image corresponds to a number of cycles in the bar pattern.

spatial frequencies of the target (e.g., as cycles/milliradian).

One apparently proper implementation of the MRTD/Johnson criteria, which has been proposed, is based on the probabilities of discrimination (PD) technique.[4] The recognition of objects using a target acquisition system is modeled by the sensor, its MRTD, the Johnson criteria, the atmosphere, and the object characteristics. These characteristics provide a model for estimating the probability of object detection, recognition, or identification and is by Driggers et al.[4] called probabilities of discrimination. A typical PD curve will have the probability (e.g., of detection) plotted as a function of range.

The input to a PD model is range, size, and the temperature difference between the object and background (ΔT). The output can be the probability of detection for different ranges, (Fig. 2). The Johnson detection criterion is defined for objects with homogeneous backgrounds. This is not often the case in practice. The background is often filled with clutter. As clutter increases, the ability to discern an object decreases. The objects have to be larger, and the number of cycles needed for detection must be increased.

A good model of the visual system should be able to predict how a human observer detects an object in a cluttered environment. This should be more difficult than detection of the same object against a homogeneous background.

One way to improve this model is to change the input ΔT so that it takes the background scenery of the image into account. This new ΔT can be called perceptual ΔT, (Fig. 3). The perceptual ΔT should predict a constant value if the range, size, and temperature are kept constant and only the structure of the background is changed. We investigate four different models for computing ΔT for different backgrounds with varying clutter, and compare them to threshold data from human observers. The comparison is done by letting the models predict the contrast detection thresholds of the human observers against different backgrounds. If this can be shown, it is a matter of calibration to produce the proper output value for predicting the perceptual ΔT. The models in the remainder of the article are mainly discussed in relation to contrast rather than temperature differences.

For comparison with an original model, the probabilities of discrimination technique is represented with a constant model. That is, whatever the background, it will always give the same temperature difference.

One of the models that may be used for this task is a spatial model of the human visual system, which takes account of the contrast sensitivity function and of the masking function of human vision (Ahumada and Beard[5]).

A third measure of the detection ability of objects of a human observer in a cluttered background is to use a measure related to a signal to noise ratio. One such measure is the signal to clutter ratio (SCR) (see Schmieder and Weathersby[6]). This model was not originally formulated for contrast detection, but since it is an attempt to measure clutter, we wanted to investigate if it could predict contrast detection as well.

A fourth measure is the peak signal to noise ratio (PSNR), which is a measure of difference that is often used for benchmarking in image quality studies.

2 Models

The traditional PD techniques take neither background nor perceptual considerations into account. The other models tested here take these parameters into account in varying degrees.

To find a suitable model for estimating the influence of the background on detection, different approaches are possible. The resulting models range from being computationally noncomplex and nonperceptually based to being more computationally complex and perceptually based. They are, in order of complexity: PD; PSNR; SCR; and two computational variants of one image discrimination model, image

Table 1 The four Johnson criteria discrimination levels and the corresponding average number of cycles across the minimum dimension. The table is taken from Holst.[2]

Discrimination Level	Meaning	Cycles Across Minimum Dimension
Detection	An object is present.	1.0
Orientation	The object is approximately symmetrical or unsymmetrical and its orientation may be discerned.	2.5
Recognition	The class to which the object belongs, e.g., tank, truck, man.	4.0
Identification	The object is discerned with sufficient clarity to specify the type, e.g., T-52 tank, friendly jeep.	8.0

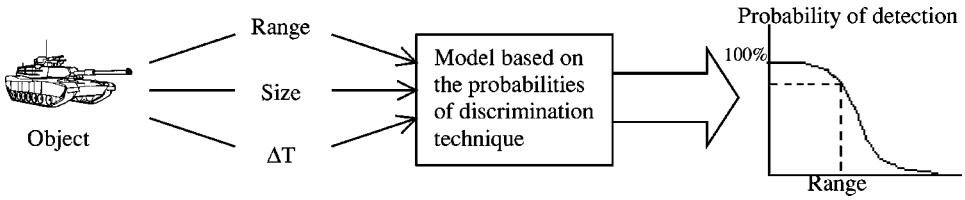

Fig. 2 A probabilities-of-discrimination model, input and output.

discrimination model, version 1 (IDM1) and image discrimination model, version 2 (IDM2).

2.1 *Probabilities of Discrimination*

The temperature difference, ΔT, in the PD technique is estimated without taking the structure of the background into account. When images are presented to an observer, these temperature differences are represented by contrasts against the background on the screen. We assume a black and white presentation, where warmer regions are shown as brighter than cooler areas. In comparison between other models, we represent the temperature difference calculation in the PD technique with a model consisting of the contrast of the target against different backgrounds, sometimes called the constant model. This model could be written

$$M(I_t, I_b) = C, \tag{1}$$

where I_t is the image containing the target and I_b is the background image, making up the model response M, which is equal to a constant C.

The constant in the model could, for instance, be set to the average temperature difference or contrast between the signal and the different backgrounds.

2.2 *Peak Signal to Noise Ratio*

The PSNR, which can be defined as

$$M(I_t, I_b) = 10 \cdot \log_{10}\left(\frac{255}{\mathrm{MSE}}\right)$$
$$\mathrm{MSE} = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}[I_t(i,j) - I_b(i,j)]^2 \tag{2}$$

is a common nonperceptual, physical measure for evaluating the influence of distortion on image quality. I_t and I_b are the values for the pixels of the target image and for the background image, respectively. It is an open issue whether the difference should be computed directly in the pixel values themselves or the more meaningful luminance values that is the physical measure of what is reaching the human eye, which we have used here. PSNR gives the estimate of the difference between the target image and the background image. Since the two images are exactly the same apart from the area of the target, this will be the only part that affects the result in the calculations. Therefore, the influence of the structure of the background outside the target is not considered by this model, although they are relevant for most detection tasks, especially those where no images are compared. The computational step for this model is as follows.

1. Convert the input images to luminance L_i using the measured gamma function γ of the screen,[7]
   $$L_i = \gamma(I_i), \tag{3}$$
   where I_i is one of the input images and $i = o, t$, i.e., the original or the target image.
2. Calculate the PSNR using Eq. (2).

2.3 *Signal to Clutter Ratio*

A way to estimate the influence of the background or the clutter in it is to calculate statistics of the distribution of the pixels in the image. Schmieder and Weathersby[6] suggested a measure that they called signal to clutter ratio and which was defined as follows,

$$\mathrm{SCR} = \frac{|\max I_t - \mathrm{mean}\, I_b|}{\mathrm{clutter}}$$
$$\mathrm{clutter} = \left(\frac{1}{N}\sum_{i=1}^{N}\sigma_i^2\right)^{1/2}, \tag{4}$$

where I_i is the maximum target value, I_b is the background mean, and σ_i is the standard deviation of the pixels. σ_i is calculated over an area, which is twice the size of the area of the target, and when that double area is divided into N search areas. This should give higher weights to clutter approximately the same size as the target. Both for PSNR and

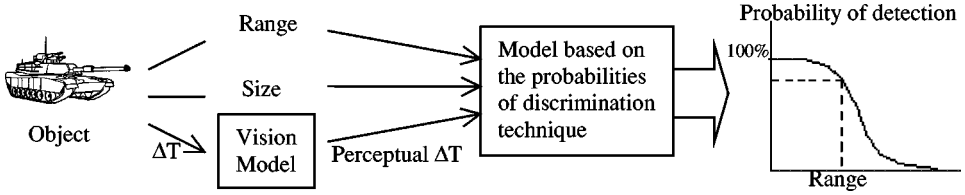

Fig. 3 An improvement of the probabilities-of-discrimination technique by addition of a vision model that takes care of the effects of the background scenery of the image.

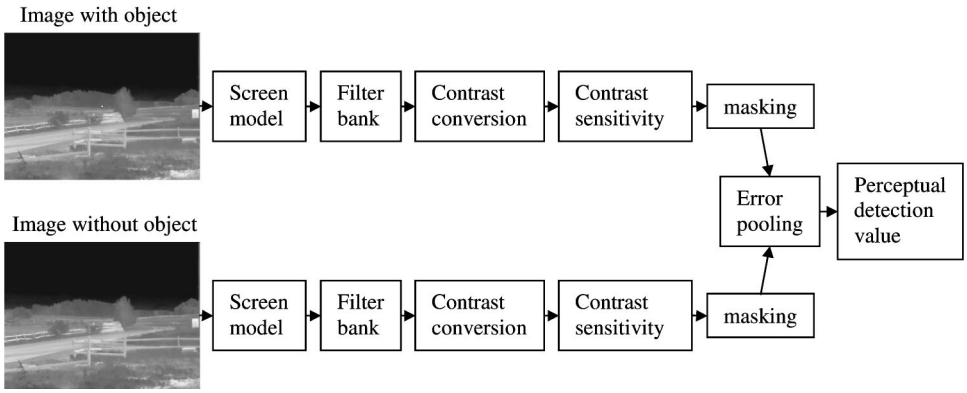

Fig. 4 Block diagram of an image discrimination model.

SCR, one may discuss if the measure should be calculated for the pixel values themselves or for the luminance values. For both of these model values we have used the luminance values in our calculations, since they represent what an observer may actually see. The computational steps for SCR are analogous to those for PSNR before.

2.4 *Image Discrimination Models, IDM1 and IDM2*

Image discrimination models are perceptually based models for estimating perceived difference between two images, see e.g., Eriksson and Andrén,[8] Lubin,[9] Daly,[10] and Watson and Solomon.[11] In many cases the intended application for such a model has been to measure image distortion. The differences between two images are spread throughout the image. In target detection the only thing that is different between the images is the target. Image discrimination models are also suited for predicting this kind of difference, i.e., the presence of a target in one of the images.[5,12–15] A generic model usually contains the computational modules, as shown in Fig. 4.

The screen model takes care of the transformation of the pixel values into corresponding luminances. Many models divide the processing into channels of different frequencies and orientations. The contrast values are then computed for each point in the image. Each spatial frequency is then adjusted for the human sensitivity for this particular frequency, i.e., the contrast sensitivity function (CSF) is used. The contents of the image, especially if it contains a large amount of high contrast, have been shown to affect the contrast detection thresholds,[16,17] a phenomenon usually called masking. The difference is then summed together, most commonly by Minkowski summation, also called vector summation, which is used to describe distances between objects in a multidimensional space. The dimension of the space is selected, usually as the one that provides an adequate fit to the objects described.

In an evaluation among several masking strategies on medical images, Eckstein, Ahumada, and Watson[13] found that models using wide band masking, such that the masking at a particular spatial location was contributed to by the activity in all frequency and orientation channels, performed the best in contrast detection tasks. Ahumada and Beard[5] proposed the simplifying assumption that the weighting between the channels in the pooling is the same and is normalized to one. This makes the whole processing independent of frequency and orientation, which makes the computational complexity substantially lower.

The following computations are performed for each input image. The image data is converted to luminance, according to Eq. (3).

1. The luminance images are converted into contrast images, using
$$C_i=\frac{L_i-\mathrm{LP}(L_o)}{\mathrm{LP}(L_o)}, \tag{5}$$
where LP is a lowpass filtering operation implemented as a gaussian filter and this is done on the original image only.
2. The contrasts are then adjusted using a model of the contrast sensitivity function (CSF). In our implementation we used a model by Barten.[18] This gives c, where
$$c_i=\mathrm{CSF}(C_i). \tag{6}$$
3. After the contrasts have been weighted by the CSF, the detection is predicted by
$$d'=\beta\left(\sum_{i=1}^{M}\sum_{j=1}^{N}\left\{\frac{|c_t(i,j)-c_b(i,j)|}{\left[1+\left(\frac{c_{\mathrm{rms}}}{b}\right)^2\right]^{1/2}}\right\}^{\beta}\right)^{1/2}, \tag{7}$$
which is a Minkowski sum with parameter β (here equal to 2) over all pixels in the region of interest. c_t is the contrast of the target image and c_b is the contrast of the background image. c_{rms} is the rms value of the contrasts of the background image and is the method used in this model to estimate the masking effect of the background.[5] The parameter b is used to set the level when the masking sets in; 0.007 has been used in this work. Two different methods of calculating c_{rms} have been proposed and will be investigated here.[5] In the first method the rms is calculated for the whole image and the same value is then used for all points in Eq. (7). We call this method IDM1. The other method is to square the contrasts and then to low-pass filter the squared contrasts. This will give a local estimate of the rms for each point. In Eq. (7), c_{rms} is then exchanged for $c_{\mathrm{rms}}(i,j)$. This method is called IDM2 and is similar to that of Ahumada and Beard.[19]

2.5 *Model Predictions*

Targets that are barely visible or just noticeable have a perceived visibility with a just-noticeable-difference set equal to 1 jnd between the target and the background. If a model can predict a constant output for various inputs with the target contrasts, which are barely visible, then such a model could be used for predicting target detection. This could be stated with notation taken from signal detection theory

$$d'_{\mathrm{mod}}(I_t, I_b) = 1, \tag{8}$$

where I_t and I_b are the intensities for target and background, respectively.

It is difficult to interpret the performance of a model for the deviations of d'_{mod} from 1. However, if the target contrast that gives $d'_{\mathrm{mod}} = 1$ is computed, then these results could be compared directly with the target contrast that is obtained experimentally. Furthermore, for most models, this may only be numerically computed.

A unit conversion parameter a has been added to all the models, so that $d'_{\mathrm{mod}} = a \cdot M$. This parameter converts the output of the model into units of jnd. This parameter has been estimated, except for the constant model, from the data by

$$a = 10^{-\mathrm{median}[\log_{10}(\mathbf{m})]}, \tag{9}$$

where **m** is a vector of model responses to input with the target signal contrast at threshold (see Brunnström et al.[20] for a more detailed discussion of this parameter.

To calculate the inverses of the models, their responses were computed for several inputs around the threshold, and then a low-order polynomial function was fitted to the data. For all models, except PSNR, a linear model was sufficient. For PSNR, a second-degree polynomial was used. The values for the PD model were set to the average contrast difference between the object and the background.

The performances of the observers were estimated in contrast energy, which is an integral value of the amount of contrast stimulating the eye in time and space. It is defined as

$$E = A \cdot t \sum_{i=1}^{M} \sum_{j=1}^{N} C(i,j)^2, \tag{10}$$

where A is the area in the visual angle of one pixel in deg^2, and t is the duration of the presentation is seconds. These values may be presented in the unit of decibel Barlow (dBB), which is defined as

$$\mathrm{dBB} = 10 \cdot \log_{10}\left(\frac{E}{E_0}\right)$$
$$E_0 = 10^{-6}\,\mathrm{deg}^2 \cdot s, \tag{11}$$

where E_0 is the strength of stimuli reported by Watson, Barlow and Robson[21] to have the lowest detection threshold. We chose the unit dBB for the evaluation of the data and for the presentation of the model predictions.

3 Experiment

The aim of the experiment is to determine the contrast energy of an object, placed in different background scenes, that an observer can detect with a higher probability than just guessing. This is usually set to a probability of detection that is at least equal to 50% correct choices of the test person. The test persons in the present study were asked to decide in which of the two presented images that the object was. The luminance value of this object was changed until a 50% detection probability was obtained. This luminance value is considered to be at the detection threshold, since it is detected in half of the presentations. These values were then used for computing contrast energies for comparison with the models' predictions.

We note that the term ''detection'' is used with different meanings in different contexts. Here we have chosen the meaning used in psychophysics, which is related to the sensory threshold concept. In a military context using IR, this study could be said to be concerned with hot spot detection.

3.1 *Images*

3.1.1 *Camera*

The camera used was a quantum well infrared photo (QWIP) detector camera from FLIR Systems, Danderyd, Sweden, with a detector chip made at Acreo, Kista, Sweden. The detector has 320×240 pixels, is Stirling cooled, and has a spectral range of 8 to 9 μm. The image performance of the camera has a thermal sensitivity of 0.03 °C at 30 °C, and an object temperature range from -20 to

Fig. 5 Image 1 ''timber,'' used in the experiment. The target object is added to the right.

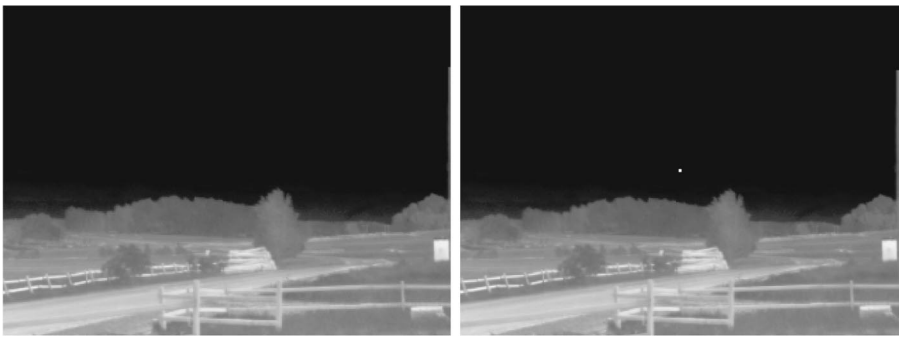

Fig. 6 Image 2, "sky," used in the experiment. The target object is added to the right.

+80 °C. In the camera there is a 170-MB PC-card disk, where 1000 images can be stored. The image file contains a 14-bit image and parameter block with all relevant information at the time of storage. The lens system of the camera consists of two different lenses, one with a 20-deg and one with a 5-deg field of view. Together with the camera a small gray-scale LCD display of 5×6.5 cm is used.

3.1.2 *Scenery*

The images should be relevant for military applications and they therefore had the following properties. 1. The scenery had the edge of woods in the background and a field in the foreground. 2. The edge of the wood should be so distant that a large object located there, for instance a tank, looked small in the images. The distance to the edge of the woods should be from 1 to 4 km. 3. The images must be taken from ground level to simulate the military use of these kinds of cameras. 4. There must be varying spatial frequencies in the images. 5. The middle of the images must be at a place where it would be realistic to detect an inserted object.

3.1.3 *Preparation*

The images made from the QWIP camera were prepared so that they could be shown to the observers in the experiment. One of the tools used was AGEMA Research 2.1, an IR image processing program from FLIR Systems. The program gives the opportunity to choose which part of the registered temperature information should be shown.

The selection of areas to be presented was done in Adobe Photoshop 5.0. To get a good match between the images, the gray scale sometimes had to be changed to compensate for the sliding temperature in the camera. The relative temperature ΔT was never changed. For further details on the preparation of the images, see Jacobson.[7]

3.1.4 *Experimental images*

Six images were used in the experiment, five for the experiment proper and one for practice. Figures 5–9 show the experimental images. The white dot in the middle of each image is the object to be detected. In the images shown here the dot is given a maximum luminance value.

Image 1, "timber," and image 2, "sky," show the same view except that image 1, had its center a little bit lower so that the object was placed at the edge of a wood. In image 2, the object was placed in the black sky. Both these images have some details in the foreground and ΔT is 10 °C. In image 3, "road," the object was placed on a road just in front of the edge of a wood. This was hypothesized to make the object more difficult to detect. ΔT here was also 10 °C. Image 4, "house," was the brightest image and the object was placed so that it should be very hard to detect. ΔT is 1.9 °C. Image 5, "poles," was the only image taken with a 5-deg lens and showed power-line poles close to the object. ΔT is 3.1 °C.

3.1.5 *Object*

The gray scale of the object had to be uniform but still be capable of changing. The object should be small, since the task to be solved was detection, and should not involve any higher process of visual or cognitive processing of the observer, such as recognition or identification of objects. Detection is assumed only to require objects with a small

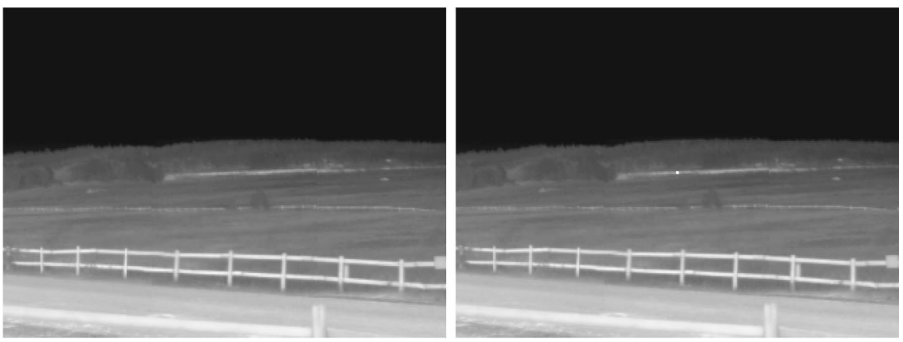

Fig. 7 Image 3, "road," used in the experiment. The target object is added to the right.

Fig. 8 Image 4, "house," used in the experiment. The target object is added to the right.

amount of information. If they contain more information, it is possible that other processes would be involved. Therefore, a small target was used.

The form of the object was a small square of four pixels with a uniform gray scale. This is the smallest spatial object that can be resolved by an IR camera in an MRTD test (where two of the pixels are black and two are white). In other words, to present a square object on the screen that would be able to present a spatial cycle, one needs at least four pixels.

3.2 *Method*

3.2.1 *Participants*

The experiment included ten test persons, seven men and three women. Their ages ranged from 19 to 31, the median being 29 years. All of the test persons had normal vision; three of them compensated with lenses.

3.2.2 *Experimental setup*

The experimental setup was a 17-in CRT display placed on a small computer table, a keyboard, and a head-and-chin rest. To simulate a smaller screen measuring 0.16 $\times$0.12 m, cardboard with a hole in the middle was used in the tests. The screen just presented the image, which was 320$\times$240 pixels.

The viewing distance was 0.47 m and controlled with a head-and-chin rest (Fig. 10). The viewing angle was 14.5 deg in height and 19 deg in width. These angles are the same as when looking at a 6$\times$8-in display at a distance of 0.60 m.

The display was an EIZO T563-T with a resolution of 640$\times$480 pixels and a screen refresh frequency of 85 Hz. The active screen size was 0.32$\times$0.24 m. The relation between the gray-scale values (0 to 255) and the luminance from the display was determined as a gamma function. This function was measured with a spectrophotometer, Photo Research Spectrascan PR 702 AM, and is described in Jacobson.[7] The level of the background light, which came from a diffuse light source placed behind the observers, was about 1 lux measured in the horizontal direction from the middle of the screen. This background light level was intended to give a comfortably dark, but not too dark, environment. The photo in Fig. 10 gives a much brighter impression than it actually was during the experiment.

3.2.3 *Experimental procedure*

The experimental procedure was a method based on a two-alternative forced-choice methodology. The observer was forced to chose in which of two sequentially presented images the object was located. The presentation time per image was 0.7 s, with an interstimuli interval of 0.7 and 1.0 s between the trials. The contrast of the object was changed so that the detection threshold could be found. A staircase procedure was used, which was built up in two steps. At first the object was very bright so that the observer could see it clearly. From that level the contrast was reduced in large steps. When the contrast was approaching the threshold value, smaller steps were taken. These steps were taken upward (increased contrast) if the answer was wrong, and downward (reduced contrast) if three answers in a row were right. With this kind of tuning procedure, it is possible

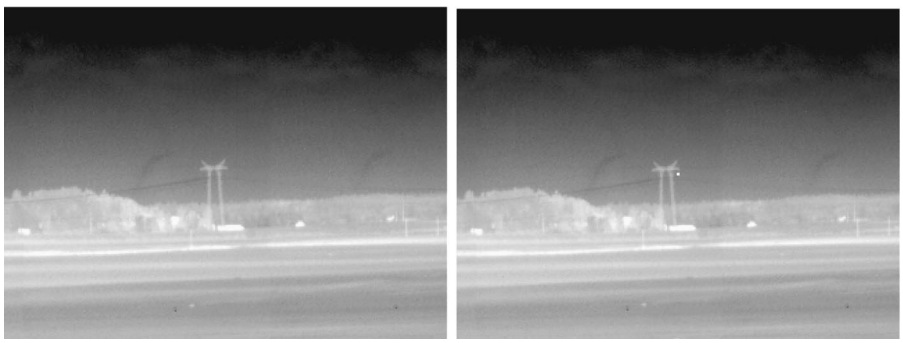

Fig. 9 Image 5, "poles," used in the experiment. The target object is added to the right.

Fig. 10 The experimental setup.

to find an observer's detection threshold. The total number of steps was 50 for each image series. The order of presentation of the five images was randomized for each test person. The verbatim instructions to the test persons may be found in Jacobson.[7] The observers were given feedback on a wrong answer by a tone signal.

Technically, the contrast changes were controlled by changing the image intensity in gray levels, and the responses were also recorded in this unit. The stimulus was always brighter than the background. It was ensured with test trials that the procedure would not result in passing through zero contrast and thus start presenting stimuli darker than the background.

4 Results

During the experiment all the answers for each test person and each image sequence were recorded. The proportion of correct answers on each gray level was then calculated. Figure 11 shows the results for one of the test persons at one of the scenes. The imposed psychometric function is also shown.

The contrast detection threshold, i.e., when a test person can detect objects with a probability of 50%, corresponding to 75% correct answers was then estimated. Pure guessing will result in 50% correct answers. We are looking for the

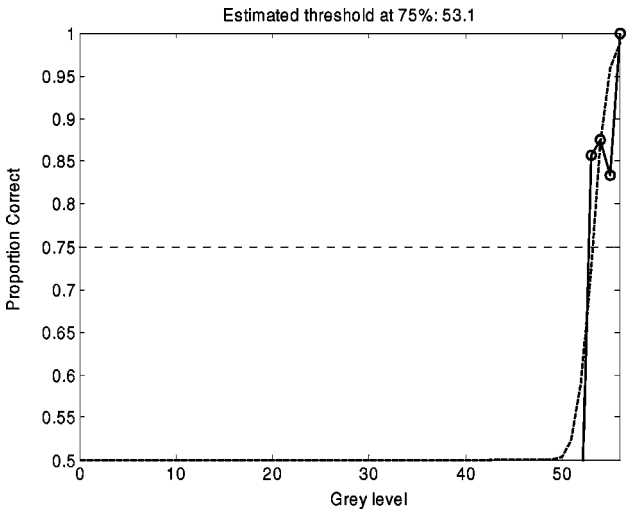

Fig. 11 Proportion of correct answers on each gray level for one of the test persons and one scene, together with the estimated psychometric function.

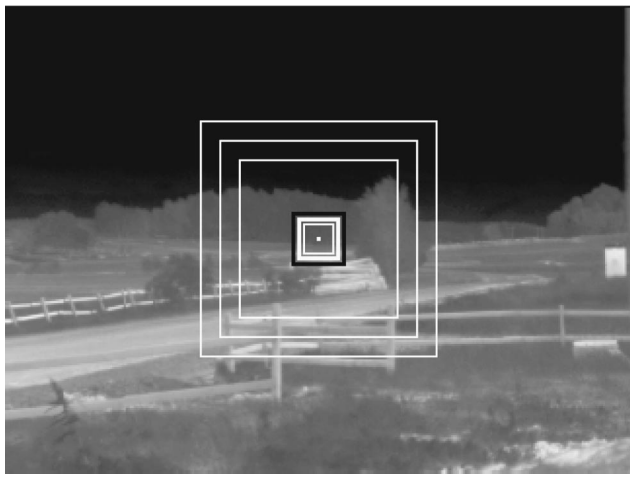

Fig. 12 The white-sided squares show how much of each image six test persons reported that they used for the detection task. The black-sided square shows foveal vision extending to about 3 deg of the view angle.

midpoint between the baseline of guessing and that of always being correct, therefore the 50% probability threshold corresponds to the midpoint between 50 and 100% correct answers, i.e., 75%. In Fig. 11 this corresponds to a threshold of a gray level around 53. The data collected from the experiment are measurements of the underlying psychometric functions of the test persons. These were estimated by fitting cumulative normal distributions to the data.[22,23] In the estimations of the psychometric functions, the first ten trials have not been included because they were only used to enable the test persons to find the object.

In some cases the lowest percent level of correct answers has been lower than 50%, which is theoretically unreasonable. This was caused by the limited number of trials used and the way the staircase procedure handles the different levels of contrast of the target during the experiment. For example, if a level was only presented once and the response was incorrect, then this gave a percent correct of 0%. These values have been kept in the fitting procedure, because they will add valuable information about the location of the threshold. However, the fitting procedure gives more weight to levels that have a higher number of responses, so these levels will influence the resulting threshold the most. The threshold values were then converted into contrast energy. On a few occasions a threshold was not obtained from the observers and these data points have been excluded. They comprise a total of four thresholds out of 50.

The detection probability P_{det} can be estimated from proportion correct P_c by

$$P_{\mathrm{det}} = 2P_c - 1. \tag{12}$$

The response in a two-alternative forced-choice method always involves a 50% chance of being correct.[23] After each test the test persons were asked what they saw in the images. Most of them found it difficult to tell what the images depicted. It was easy to find different characteristics in the middle of them, e.g., the power-line poles in image 5, ''poles'' (Fig. 9), and the white line in image 3, ''road''

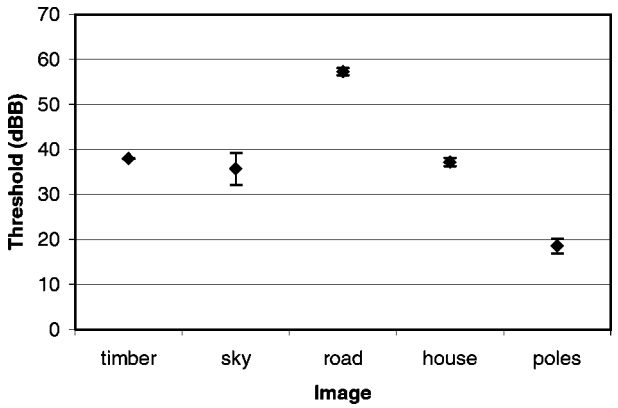

Fig. 13 The average thresholds for the test persons. The error bars represent 95% confidence intervals.

(Fig. 7), but nobody saw the house in the left part of image 4, ''house'' (Fig. 8). Six of the test persons were asked how much of each image they used for the task (see Fig. 12). The answers resemble the foveal vision, which is about 3 deg of the view angle.[24] The assumption that only foveal vision was used is tenable. There was no movement in the images, and the object was always in the center so no eye movements were needed.

The gray level thresholds are converted into luminance using Eq. (3) Thereafter their contrast against the background is computed using

$$C = \frac{L_t - L_b}{L_b}, \tag{13}$$

where L_t is the luminance of the target and L_b is the luminance of the background. The background luminance was computed as the average value over a square region of 50 $\times$50 pixels, which corresponded to about 3 deg of visual angle (see Fig. 12). The contrast energy is then computed using Eqs. (10) and (11), where the area of a pixel A was $0.036\ \mathrm{deg}^2$ in our case, and the duration of the stimuli as mentioned before was 0.7 s.

The individual thresholds have been summarized with the mean and 95% confidence intervals for each image (see Fig. 13 and Table 2), based on all the values for all test persons. However, the number of observers used for each mean varies for the different images, since the number of estimated thresholds for each image were different. The model responses are estimated as described in Sec. 2 and are shown graphically in Fig. 14 and numerically in Table 3.

Table 2 The individual thresholds in dBB for each observer at each scene. Empty cells mark excluded data points.

Observer	I_1 ''timber''	I_2 ''sky''	I_3 ''road''	I_4 ''house''	I_5 ''pylon''
1	38.01	38.29	57.14	36.66	22.19
2	37.88	40.85	56.50	38.61	20.49
3	38.02	31.47	56.91	37.67	21.03
4	37.77	43.48	56.17	36.07	19.93
5	37.88	34.52	56.95	38.18	20.72
6	38.10		57.02	35.84	15.65
7	38.04	33.31	57.17	34.64	14.90
8	38.11		60.37		18.16
9	37.91	36.43		38.53	17.93
10	37.78	27.23	57.12	38.30	14.72

Table 3 Observers average thresholds and model predictions of thresholds in dBB.

	I_1 ''timber''	I_2 ''sky''	I_3 ''road''	I_4 ''house''	I_5 ''pylon''
Observers	37.95	35.70	57.26	37.17	18.57
PD	32.73	32.66	57.32	37.27	28.46
SCR	36.41	28.25	57.35	36.22	18.98
PSNR	36.44	67.14	56.84	31.75	18.57
IDM1	37.33	37.33	37.33	37.33	37.33
IDM2	36.85	75.14	27.65	39.85	39.46

The correlation between the contrast energy of the models and the average observer were for the models PD, SCR, PSNR, IDM1, and IDM2 equal to -0.40, -0.29, 0.64, 0.91, and 0.97, respectively. This gives the explained variance R^2 0.17, 0.09, 0.41, 0.83, and 0.95. The models IDM1 and IDM2 clearly had a better performance than the others did. Interestingly, PSNR also had quite a high correlation with the empirical data. One may also note that SCR had the worst performance of all the models, but we want to remind the reader that this model was not originally developed for the present kind of tests. Visual inspection of the data showed that the advantages of the visually based models, i.e., IDM1 and IDM2, compared to the physical model PSNR, were most apparent at low luminance levels.

5 Discussion

We investigated a number of alternatives that could be substituted to replace the traditional way of testing IR cameras by the Johnson criteria. To do so, we had to take into consideration a number of visual conditions that may exist in a real, natural situation, e.g., in a battlefield. Important conditions for detecting a target within an IR scene are the brightness of the target, its contrast to the background, its internal structure, and the texture of the background or ''clutter'' (see e.g., Rotman, Tidhar, and Kowalcyk[25]).

We see four different routes that a specification of an IR camera could take:

1. using the Johnson criteria (see Fig. 2)
2. using the Johnson criteria together with a visual model (see Fig. 3)
3. using a visual model together with a set of identical images with known characteristics that are tested against different IR systems
4. using a visual model together with images with varying backgrounds.

Route 1 is the common way today of testing IR cameras. It supposes that the background is constant. Route 2 can handle images with varying backgrounds. Route 3 assumes that the model can predict human observer detection. If this assumption is correct, then the model could be used to test

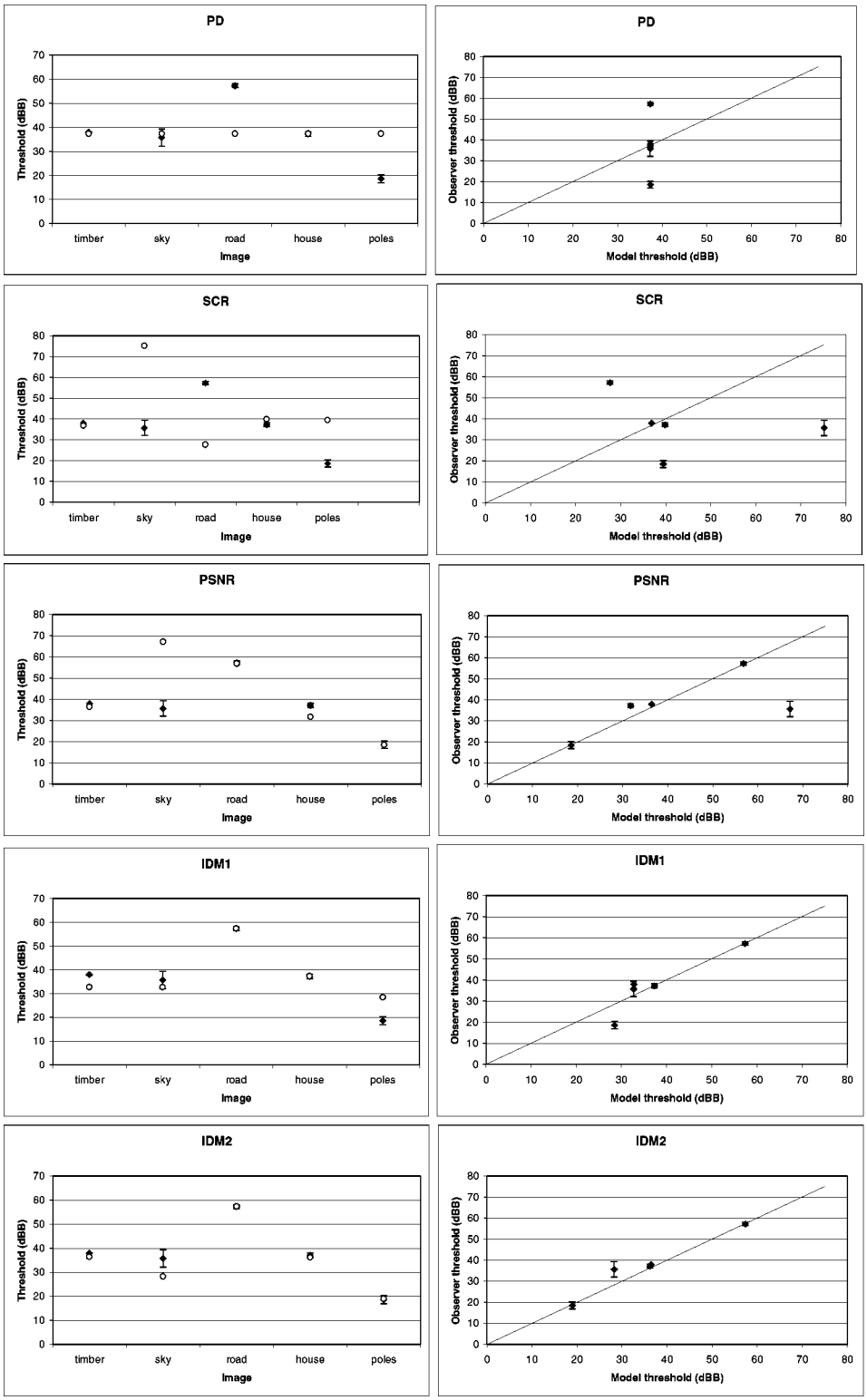

Fig. 14 The average thresholds of the observers (diamonds) and the model predictions (circles) for each scene (left) and the relation between predicted model value and observer threshold (right). The line represents perfect correspondence between prediction and outcome. The error bars show the 95% confidence interval around the mean. From top to bottom are shown the models probability of discrimination, signal to clutter ratio, peak signal to noise ratio, image discrimination model 1 (IDM1) and image discrimination model 2 (IDM2).

images recorded in similar conditions by different IR systems, or to test the differences between IR systems. Route 4 is a variation of route 3, since instead of finding the result for a number of images, one instead varies the background till a certain criterion is reached.

One main purpose of this study was to examine if the use of different visual models could be of assistance when specifying and testing IR camera systems. On the whole, the visual models well fulfils the task of simulating human behavior in detection experiments. They can thus be used as an alternative to human observers in testing and developing IR camera systems. Threshold data indicating 50% detection probability could be established for nearly all images and test persons. Another result of the experiment, due to its design, was that the test persons probably only used a small part of each image to fulfil the detection task. These small parts appear to coincide with human foveal vision.

The often-used PD technique does not take the background scenery of the image into consideration. The PD technique was changed by transforming the input ΔT into a perceptual ΔT calculated by the vision model. Other models such as PSNR and SCR were also studied.

With this improvement of the PD technique, the value of the Johnson detection criterion can be set at 1 and be independent of the clutter ratio of the background. However, it is important to point out that this is a specific case. For a more general improvement, the range and the size of the object must also be modified into perceptual data. The vision model chosen must be expanded by adding, for example, an atmospheric model.

It was possible to complement and improve the PD technique by a vision model. Thus, a new way of looking at the problem of detecting objects in IR images was created. An IR detection task is a complicated matter containing many parameters, ranging from the atmosphere and the camera resolution to complex perceptual processes in the human brain. Besides the fact that the PD does not function very well in the case of images with clutter, other problems have to be investigated. One of these is MRTD. The measurement of MRTD has disadvantages, since the method is dependent on the observer's subjective decision criterion, and the bar pattern stimulus is theoretically and practically unsuitable for focal plane array cameras.[26] Since IR light is not visible to the human eye, how to show the IR image has to be decided by the camera manufacturer and the operator. A common way to show the image is to have a gray scale represent the different temperatures, where black represents cold and white represents warm. An image displayed in this way resembles a common black and white photo, except that some parts seem to be inverted. An IR image is somewhat similar to images produced by x-ray detectors or electron microscopes. These kinds of images also have an unnatural character and can be displayed in many ways.

An IR image can be shown to the observer in many ways, since the created image is not natural. This gives the operator plenty of room to choose how the image should be presented. Finding the optimal way to present an image in a detection task is very hard. This is another area where vision models might be useful.

We note that the present experiment has some similarities to the concept of minimum findable threshold differences (MFTD).[27] Differently sized squares with different intensities were randomly placed in images and the observers were asked to detect these squares. In the present experiment, the object had the same size, and was always located at the center of the image. MFTD was proposed as a means of characterizing thermal imaging performance under scene clutter limited conditions.

One way of improving the vision model for IR images would be to simulate human perception of contrast resolution for different background luminances, since a higher contrast is needed for darker images. The model should then be able to predict scenes like image 2, ''sky,'' better. However, a vision model improved in these ways would still be specialized for detection tasks. To expand the model to encompass also recognition and identification, more changes have to be made. For recognition and detection the objects need to be larger. It is not sufficient to use virtual objects like four squares; the objects have to be real. They should also be able to change in luminance and size. New images are also needed where the background is closer, so that large objects can fit in a natural way. An expanded model should no longer look just at the masking effects of the background; it should also look at the masking effects of the objects and see how they mask each other. To test this new model, more psychophysical experiments are needed in which the tasks of the test persons are based on recognition or identification. The tests should include different backgrounds and objects of different types, sizes, and luminances.

There has been proposals that it should be possible to use undersampled IR images.[28] It should thus be possible to see ''beyond'' the Nyquist frequency and still get a reasonable image quality. Instead of using measures like MRTD, Wittenstein[28] suggests one could use minimum temperature difference perceived (MTDP). Undersampled images may provide information, albeit distorted primarily by aliasing. All four bars in a test pattern are not required to be resolved. Wittenstein[28] proposed that one should replace the modulation transfer function with average modulation at optimum phase. A development of a spatial model, taking into account the less stringent ideas of undersampling, may make the resulting model more realistic.

For models of the perception of medical images, there has been much progress done in recent years. For a review, see Eckstein.[29] These images have many similarities to IR images. They are intended for an observer to detect structures in images that have been made through nonvisible radiation. They are made by x-rays and not by IR radiation. Among the successful approaches is the close linkage to classical signal detection theory. There is also a connection for some of the models in the present study to statistical detection theory and to signal detection theory, but it is not as explicit as for the theories of image quality in medical images. One missing aspect in our models is the explicit modeling of the internal noise that the observer has when he or she is looking at the image. We believe that elaboration of models for IR images would benefit from an assimilation of the experiences of what has been learned regarding image quality and object detection in medical pictures.

6 Conclusions

The main result of this study is that the influence of the background on the contrast detection of IR images can be

predicted well with models, taking into account the human contrast sensitivity function and masking. These models perform well for these images, compared to models based on purely physical considerations. This would make it possible to improve the currently used PD technique, which is based on the Johnson criteria, but does not consider the influence of the background on detection.

The current study was done under several simplifying assumptions in order to have a controlled experiment. Further investigations and developments are needed before these models can be of practical use.

Acknowledgments

Jean M. Bennett, then visiting professor at Acreo, gave much helpful advice. Åke Arbrink at Försvarets materielverk (Swedish Defense Materiel Administration) helped taking the IR images. Hans Hallin at FLIR Systems supplied the IR image software. Mårten Nilsson at Bonnier Lexikon assisted in the preparation of the experimental images. SaabTech Electronics, FMV, Ericsson Saab Avionics, Ericsson, Telia Research, and Vinnova (The Swedish Agency for Innovation Systems) supported this work. Marie-Claude Béland, Acreo, provided additional funds. Finally, we thank the test persons.

References

1. J. Johnson, ''Analysis of image-forming systems,'' *Proc. Image Intensifier Symp.*, 249–273 (1958).
2. G. Holst, *Electro-Optical Imaging System Performance*, SPIE Press, Bellingham, WA (1995).
3. R. Driggers, S. Pruchnic, C. Halford, and E. Burroughs, ''Laboratory measurement of sampled infrared imaging system performance,'' *Opt. Eng.* **38**, 852–861 (1999).
4. R. Driggers, P. Cox, J. Leachtenauer, R. Vollmerhausen, and D. Scribner, ''Targeting and intelligence electro-optical recognition: a juxtaposition of the probabilities of discrimination and the general image quality equation,'' *Opt. Eng.* **37**, 789–797 (1998).
5. A. J. Ahumada, Jr. and B. L. Beard, ''Object detection in a noisy scene,'' *Proc. SPIE* **2657**, 190–199 (1996).
6. D. A. Schmieder and M. R. Weathersby, ''Detection performance in clutter with variable resolution,'' *IEEE Trans. Aerosp. Electron. Syst.* **19**, 622–630 (1983).
7. B. Jacobson, ''Detection of objects in IR images,'' Master thesis, TRITA-FYS 2116, Royal Institute of Technology, Stockholm (2000).
8. R. Eriksson and B. Andrén, ''Modelling the perception of digital images,'' Technical Report TR 315, Institute of Optical Research, Stockholm (1997).
9. J. Lubin, ''The use of psychophysical data and models in the analysis of display system performance,'' in *Digital Images and Human Vision*, A. B. Watson, Ed., MIT Press, Boston, MA (1993).
10. S. Daly, ''The visible difference predictor: An algorithm for the assessment of image fidality,'' in *Digital Images and Human Vision*, A. B. Watson, Ed., MIT Press, Boston, MA (1993).
11. B. Watson and J. A. Solomon, ''A model of visual contrast gain control and pattern masking,'' *J. Opt. Soc. Am. A* **14**, 2378–2390 (1997).
12. R. Eriksson, K. Brunnström, and B. Andrén, ''Evaluation of image discrimination models for static images,'' Technical Report 330, Institute of Optical Research, Stockholm (1998).
13. M. P. Eckstein, A. J. Ahumada, Jr., and A. B. Watson, ''Image discrimination models predict signal detection in natural medical image backgrounds,'' *Proc. SPIE* **3016**, 44–56 (1997).
14. A. M. Rohaly, A. J. Ahumada, Jr., and A. Watson, ''Object detection in natural backgrounds predicted by discrimination performance and models,'' *Vis. Sci.* **37**, 3225–3235 (1997).
15. A. J. Ahumada, Jr., A. B. Watson, and A. M. Rohaly, ''Models of human image discrimination predict object detection in natural backgrounds,'' *Proc. SPIE* **2411**, 355–365 (1995).
16. G. E. Legge and J. M. Foley, ''Contrast masking in human vision,'' *J. Opt. Soc. Am.* **70**, 1458–1471 (1980).
17. J. M. Foley, ''Human luminance pattern-vision mechanisms: masking experiments require a new model,'' *J. Opt. Soc. Am. A* **11**, 1710–1719 (1994).
18. P. G. J. Barten, ''The square-root integral (SQRI): A new metric to describe the effect of various display parameters on perceived image quality,'' *Proc. SPIE* **1077**, 73–82 (1989).
19. A. J. Ahumada, Jr. and B. L. Beard, ''A simple vision model for inhomogeneous image quality assessment,'' *Soc. Info. Displays Intl. Symp.* **29**, J. Morreale, Ed., paper 40.1 (1998).
20. K. Brunnström, R. Eriksson, B. Schenkman, and B. Andrén, ''Comparison of predictions of a spatio-temporal model with responses of observers for moving images,'' TR-338, Inst. Optical Research, Stockholm (1999).
21. A. B. Watson, H. B. Barlow, and J. G. Robson, ''What does the eye see best?'' *Nature (London)* **302**, 413–422 (1983).
22. G. A. Gescheider, *Psychophysics: Method, Theory and Application*, Lawrence Erlbaum, HillsDale, NJ (1985).
23. P. G. J. Barten, *Contrast Sensitivity of the Human Eye and Its Effects on Image Quality*, SPIE, Bellingham, WA (1999).
24. G. Skinner and P. Connell, *Notes from course PHM41D in image processing*, University of Birmingham, UK (2000).
25. S. R. Rotman, G. Tidhar, and M. L. Kowalcyk, ''Clutter metrics for target detection systems,'' *IEEE Trans. Aerosp. Electron. Syst.* **30**, 91 (1994).
26. P. Bijl and M. Valeton, ''Triangle orientation discrimination: the alternative to minimum resolvable temperature difference and minimum resolvable contrast,'' *Opt. Eng.* **37**, 1976–1983 (1998).
27. J. D'Agostino and J. R. Moulton, ''Minimum findable temperature,'' *Infrared Imag. Syst. Design, Analysis, Modeling, and Testing* **2224**, 79–94 (1994).
28. W. Wittenstein, ''Minimum temperature difference perceived—a new approach to assess undersampled thermal images,'' *Opt. Eng.* **38**, 773–781 (1999).
29. M. P. Eckstein, ''The perception of medical images 1941–2001,'' *Opt. Photonics News* **12**, 34–40 (2001).

Kjell Brunnström received his MS in engineering physics and PhD in computer science from the Royal Institute of Technology. Stockholm, Sweden, in 1984 and 1993, respectively. From October 1985 to April 1987 he was a visiting research student at Tokyo University, Japan. During 1995 he was a postdoctoral associate at the University of Surrey, Guildford, United Kingdom. He is currently holding a research position at the research institute Acreo, previously called the Institute of Optical Research, Stockholm. His current main research interest is image discrimination models for still images and video.

Bo N. Schenkman received a BA degree in psychology and philosophy from Hebrew University, Jerusalem, Israel, and a PhD degree in psychology from Uppsala University, Sweden, in 1985. From 1985 to 1996 he worked as a human factors specialist in research and development departments at the Swedish computer divisions of Ericsson, Nokia, and ICL. During 1996 he did research at the Royal Institute of Technology, Stockholm, on image quality issues. From 1997 to 1998 he worked with telecommunication research at Telia, Stockholm. In 1999 he joined the Institute of Optical Research in Stockholm, later named Acreo. His present research interests are human perception. image quality, human factors, and psychophysics.

Bengt Jacobson received his MS in engineering physics from the Royal Institute of Technology, Stockholm, Sweden, in January 2001. His diploma work, ''Detection of objects in infrared images,'' was done at Acreo during 2000. In March 2001 he joined Acreo as a Development Engineer. His current research interests are signal light modulators and image quality of CRT and LCD displays.