To be published in B. Rogowitz, T. N. Pappas, eds., *Human Vision and Electronic Imaging III,* Proceedings 3644, Paper 42, SPIE, Bellingham, WA, 1999.

Spatio-temporal discrimination model predicting IR target detection

Kjell Brunnström[a], Robert Eriksson[b] and Albert J. Ahumada, Jr.[c]

aInstitute of Optical Research, Electrum 236, SE-164 40 Kista, Sweden

bTelia Research AB ,Vitsandsgatan 9, SE-123 86 Farsta, Sweden

cNASA Ames Research Center, Mail Stop 262-2, Moffett Field, CA 94035

Correspondance: Email: kjell.brunnstrom@optics.kth.se, Tel: +46-8-6327732, Fax: +46-8-6327710, URL: www.optics.kth.se

Email: Robert.A.Eriksson@telia.se

Email: al@vision.arc.nasa.gov, URL: http://vision.arc.nasa.gov/~al/ahumada.html

Abstract

Many image discrimination models are available for static images. However, in many applications temporal information is important, so image fidelity metrics for image sequences are needed as well. Ahumada *et al.* (1998) [Ref. 1] presented a discrimination model for image sequences. It is unusual in that it does not decompose the images into multiple frequency and orientation channels. This helps make it computationally inexpensive. It was evaluated for predicting psychophysical experiments measuring contrast sensitivity and temporal masking. The results were promising. In this paper we investigate the performance of the above-mentioned model for a practical application – surveillance with infrared (IR) imagery.

Model evaluation is based on two-alternative forced choice experiments, using a staircase procedure to control signal amplitude. The observer is presented with two one-second-duration IR-image sequences, one of which has an added target signal. The observer's task is to guess which sequence contained the target. While the target is stationary in the image centre, the background moves in one direction, simulating a tracking situation in which the observer has locked on to the target. The results shows that the model qualitatively, in four out of five cases, have the desired behaviour.

1. video, image quality, target detection, spatio-temporal, vision model, masking, infrared image sequence

1. Introduction

Many image display system tasks involve the detection or identification of objects in the displayed imagery. For instance a medical diagnosis may depend on the identification of structures in X-ray images. A pilot aided with a camera in the nose of an aeroplane and a display in the cockpit must not miss obstacles on the runway. In surveillance it is important to detect something which should not be present or discriminate that something is missing. During system development, subjective testing for assessing user's ability to perform the detection task, is both expensive and time-consuming. An objective measure based on visual perception is therefore valuable, especially during the early development stages.

Traditionally, infrared systems are evaluated for object detection by the Johnson criterion, giving detection probability of an object as a function of parameters such as the target distance. (There are criteria for recognition, orientation and identification as well.) However, this criterion does not take into account masking by the distribution of contrast in a neighbourhood of the object. Thresholds for object detection are higher in a complex, high contrast background [Refs. 3, 4].

Studies show that image discrimination models can predict object detection in still medical images [Ref. 5], noisy military images [Ref. 6], and in natural images [Ref. 7]. In this study we investigate the performance of a spatio-temporal image discrimination model in a practical application – surveillance with infrared (IR) imagery. The model is unusual in that it does not decompose the images into multiple frequency and orientation channels. This helps make it computationally inexpensive. When it was evaluated for predicting psychophysical experiments measuring contrast sensitivity and temporal masking, the results were promising [Ref. 1]. Here we evaluate the model's ability to predict the ability of observers to detect stationary targets in moving backgrounds.

We begin by presenting the main features of the model (the details are available elsewhere [Refs. 1, 2]). Next we describe the psychophysical experiment. The model predictions are then presented, discussed, and conclusions are drawn.

1. Model

It is convenient to describe the model as simulating the physiological stages of the human visual system. These stages are optics of the eye, retina, LGN and the cortex, see Figure 1. Actually, the model is a lumped parameter model, so that the low pass filtering attributed to the optics of the eye, for example, is actually the combined effect of low pass filtering at all stages. The model is an image sequence discrimination model. It takes as input two-luminance image sequences and outputs a single number, the predicted discriminability of the two image sequences in just-noticeable-differences. After "optical" low pass filtering and a "retinal" temporal smoothing and contrast conversion, separate sustained ("parvo") and transient ("magno") spatio-temporal channels then process each image sequence. The masking rules are different in magno and parvo. The differences between the processed channel outputs for each sequence are

accumulated to generate an overall difference.

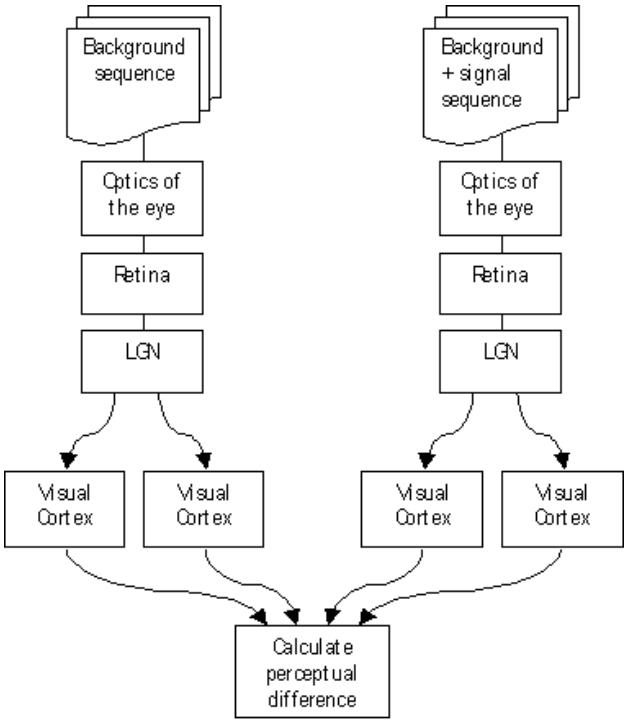

Figure 1: Schematic overview of the model

1. Psychophysical Experiment

1. Observers

Ten novice observers participated in the experiments. All had normal or corrected-to-normal visual acuity. Ages ranged from 21 to 47, with the median age being 25.

2. Experimental setup

The experimental display was a Sony Trinitron Multiscan 20sh, driven by a Matrox Millenium graphics card. The resolution was set to 1024 pixels horizontally and 768 pixels vertically, with a screen refresh rate of 60 Hz. The luminance profile of the central parts of the display was measured and a gamma function estimated. The estimated gamma function was:

$$L(x, y) = 6.1+0.037\, g(x, y)^{1.97} \quad (1)$$

where $g(x, y)$ is the grey value at location x, y and $L(x, y)$ is the luminance value in cd/m^2.

The background was held at a luminance of 27 cd/m^2. The image sequences were displayed in the centre part of the display in an area subtending 2 degrees of visual angle, with a resolution of 63 pixels per degree, shown at a frame rate of 30 Hz. The viewing distance was 140 cm, controlled with a head-and-chin-rest.

3. Experimental Procedure

A 2-alternative-forced-choice procedure was used to determine the target visibility threshold in infrared sequences, using a staircase procedure to control the target amplitude. On each trial the observer was presented with two one second duration IR-image sequences, one with the target added. The observer's task was to guess which sequence contained the target. The target was stationary in the image centre, and the background moved in one direction, simulating a tracking situation in which the observer had locked on to the target. The size of the target was 19 arc min horizontally and 5 arc min. vertically.

Before, between, and after the sequences are shown the observer viewed a grey screen (27 cd/m^2) containing a faint fixation mark in the centre. The observer commanded the start of each trial. Between each sequence was a 1 sec pause, and between trials a pause of at least 1.5 sec. The observer entered the guess by pressing 1 (first sequence) or 2 (second sequence) on the keyboard. A beep indicated a wrong guess. After 3 correct answers the contrast was decreased, and 1 incorrect answer raised the contrast. A contrast gain factor was multiplied with the grey values of target. The gain factor was changed with large steps (10/255) in the first three steps and then with small steps (1/255). The initial gain factors were selected individually for each sequence to obtain more steps close to the threshold. Thresholds were determined by a maximum likelihood probit analysis [Ref. 9] that estimated the gain level leading to 75% correct. The analysis assumed the psychometric function was a cumulative normal distribution. Three 100 trial staircases were run on each of five masking sequences.

4. Stimuli

Five IR sequences were obtained with an IR camera of high image quality. They were processed with software simulating an IR system of much lower quality [Ref. 8], see Figures 2, 3 and 4. The target was an elongated object that when added to the backgrounds brightened them, indicating a warmer area. The target image was also processed with the IR system simulation software so as to contain noise similar to that of the backgrounds.

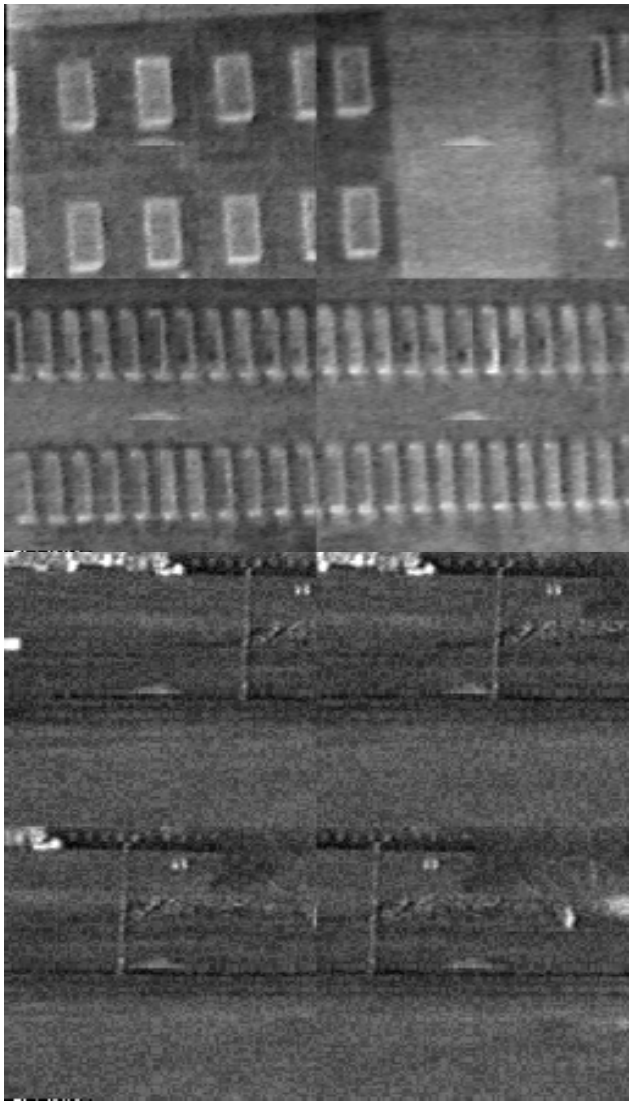

Figure 2: Sequence 1 and 2, displaying four images out of 30, with the target added.

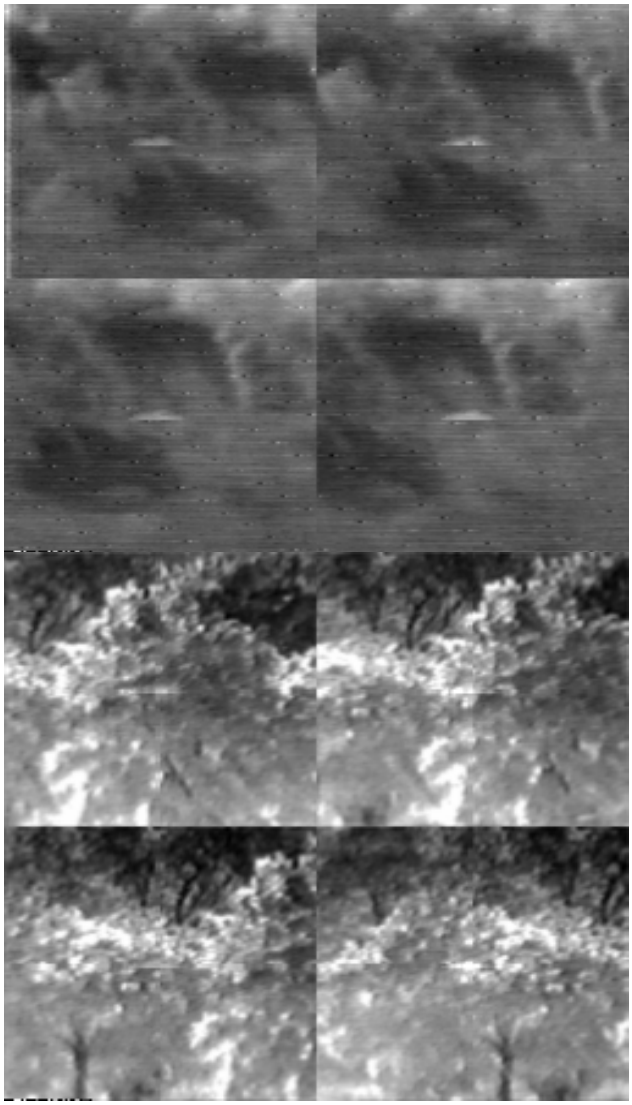

Figure 3: Sequence 3 and 4, displaying four images out of 30, with the target added.

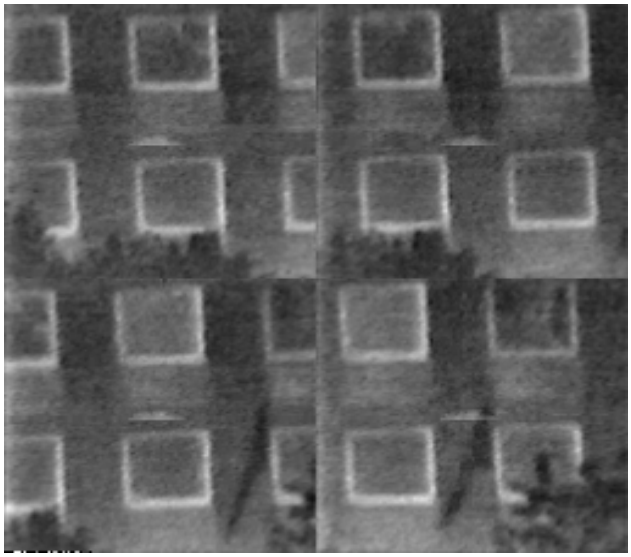

Figure 4: Sequence 5, displaying four images out of 30, with the target added.

5. Experimental Results

The experimental thresholds are reported in decibels of contrast energy (*dBB*), which is defined as

dBB = 10 log10 *E_c* + 60 . (2)

The zero level of this unit is related to the highest sensitivity for a pattern measured by Watson *et al.* [Ref. 10]. The contrast energy is determined by

$$E_c = A \cdot t \cdot \sum_{x,y \in T} c(x,y)^2 \left[deg^2 \cdot \sec \right] \quad , (3)$$

where *A* is the area of single pixel in deg^2, *t* is duration the target is visible, *c(x, y)* is the contrast at location *x, y* and *x, yÎ T* means *x, y* should belong to the target signal. The contrast is defined as

c(x, y) = *L_T(x, y)* / *L_B* - 1 . (4)

L_T is the luminance of the target and *L_B* is the mean luminance of the background.

The experiments were conducted in three sessions. The first session had a clearly higher average threshold than the other two, indicating a learning effect, see Figure 5.

The individual observer thresholds were estimated from the combined data of sessions 2 and 3, see Figure 6. From these thresholds an average value and a 95% confidence interval were calculated for each sequence, see Figure 7.

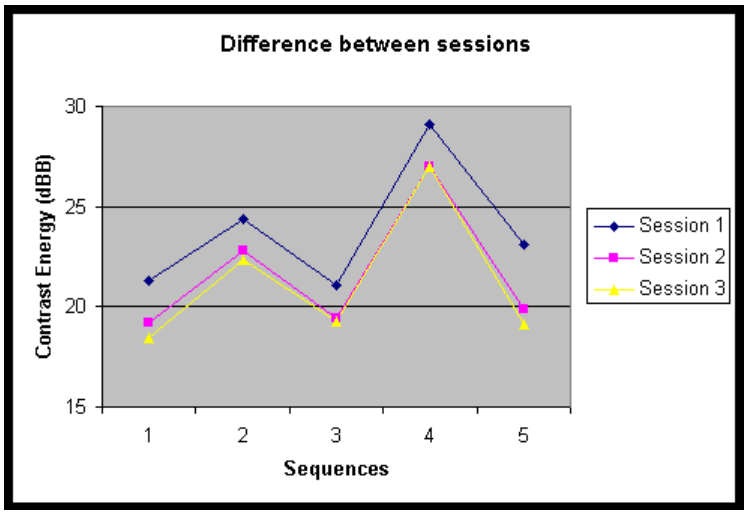

Figure 5: The estimated thresholds for the different sessions

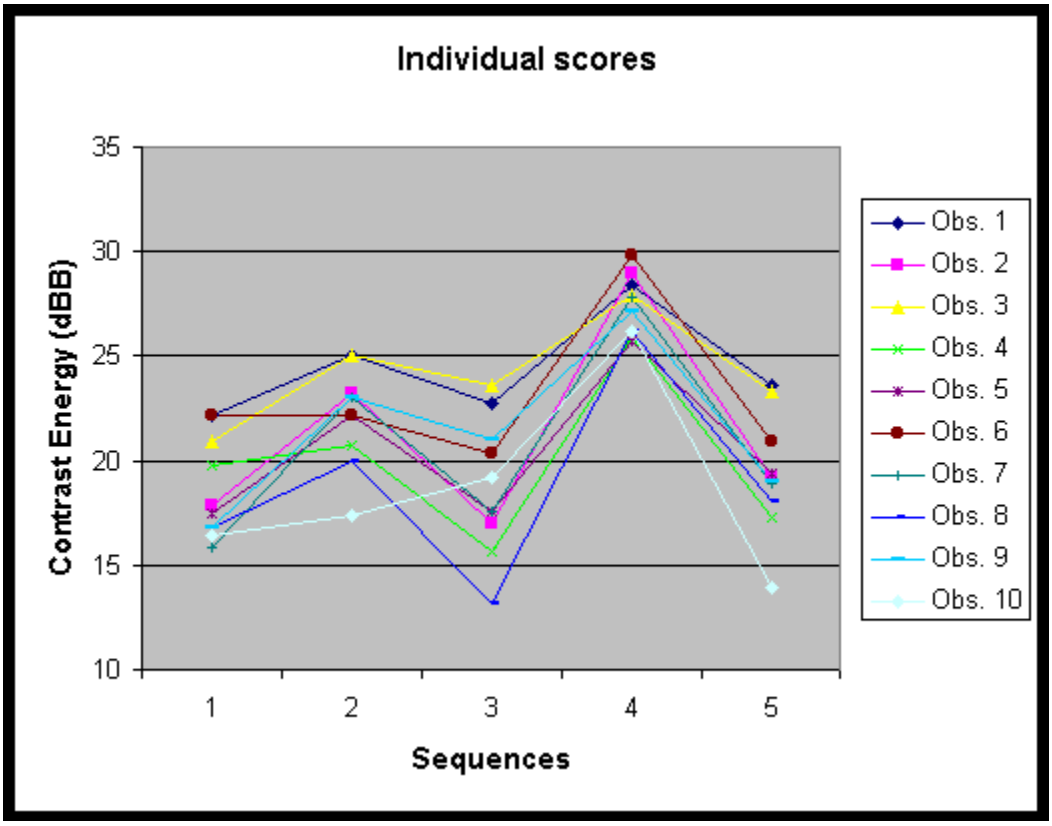

Figure 6: The thresholds of each observer

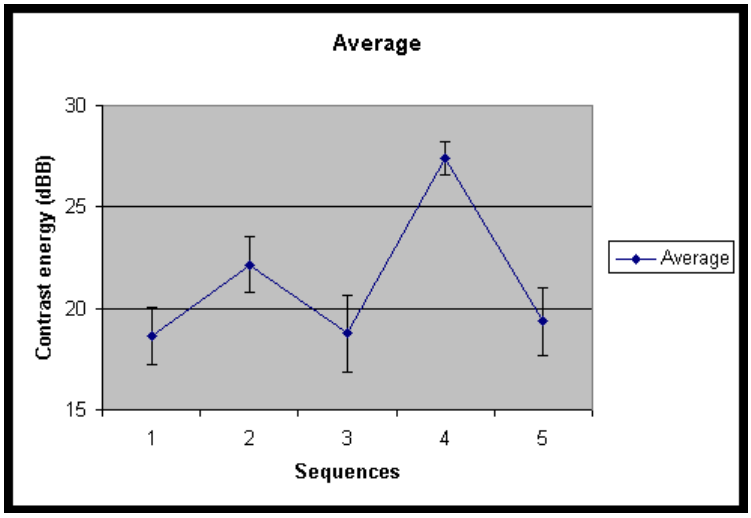

Figure 7: Average thresholds for the observers and 95% confidence intervals.

1. Model Predictions

To evaluate the spatio-temporal model, predicted threshold values were computed for each sequence. The model output is given in just-noticeable-differences *d'*. For comparison with the experimental data, we need to find the contrast energy *E_c*, which makes the model respond with *d'* = 1.

The model parameters used are the same as in Ahumada *et al.* [Ref. 8], except for a sensitivity parameter which sets the overall response level

$$d' = a\, M(s\_back, s\_signal, E\_c) \,, \quad (5)$$

where *a* is the sensitivity parameter, *M* is the model, *s_back* is a background sequence, *s_signal* is the target signal and E_c is the contrast energy that the signal is added with. The best estimate of the gain parameter in the least square's sense is

$$\text{mean}(\log(a\ \mathbf{m}) - \log d'\_true\ ) = 0 \,, \quad (6)$$

where **m** is a vector with model responses. Since *d'_true* = 1, we have

log *a* = - mean(log **m**) . (7)

To minimize the effect of outliers, we used the median instead, giving an estimate for *a* of 8.4x10^7.

To find the thresholds in contrast energy, we should find the value *E_c* that makes *d'* = 1. This is equivalent to log *d'* = 0. The values for each sequence have been numerically computed to:

Sequence	1	2	3	4	5
Threshold	19.8	17.2	20.7	24.5	20.7

For comparison, we have computed the peak-signal-to-noise-ratio (PSNR), for each sequence, which is defined in dB as

PSNR = 10 log10((max(*s_back*) – min(*s_back*))^2 / mean((*s_back+signal* – *s_back*)^2) . (8)

The rule of thumb is that if the response is about 32 dB the difference is on the threshold. However, if we add the target to the backgrounds with the contrast energies of the thresholds of the average observer, we get values far from 32. Since there is a free parameter in spatio-temporal model to adjust the overall level, a direct comparison is not fair. Therefore, the same type of parameter has been added to the PSNR measure.

PSNR_*a* = 20 log10 a + PSNR = *A* + PSNR , (9)

where *A* = 20 log10 a. We can now adjust the level in a similar way as above, which gives

PSNR_*a* = 32 – mean(**psnr**) + PSNR , (10)

where **psnr** is a vector of the PSNR values for all sequences.

The results for the average of the observers, the spatio-temporal model, and PSNR_*a*, is plotted in Figure 8.

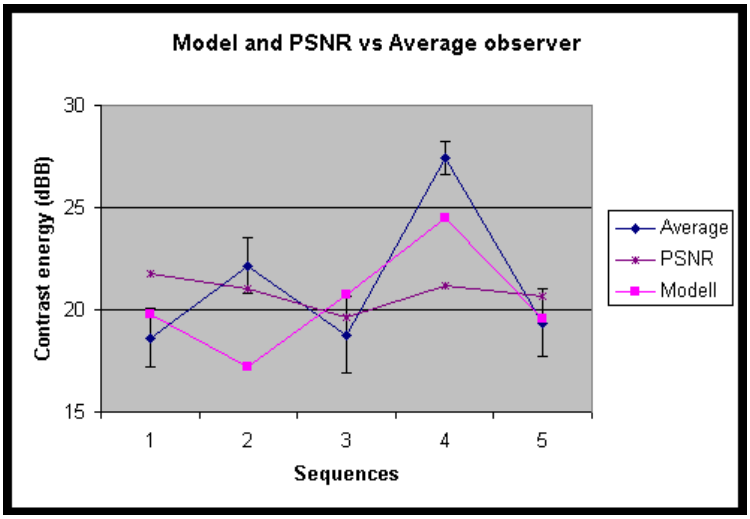

Figure 8: Thresholds estimated with the spatio-temporal model and PSNR, plotted together with the threshold of the average observer.

2. Discussion and Conclusions

An evaluation has been performed of a spatio-temporal model in a practical application – surveillance in IR imagery. We have investigated the model's ability to predict the ability of observers to detect stationary targets in moving backgrounds.

The results indicate that the model, at least qualitatively, in four out of five sequences have the desired behaviour. The result of sequence 2 is poor. The dominating masking effect in this sequence is dynamic noise. The calibration of the model [Refs. 1, 2] were performed with data covering mainly the low temporal frequency range. The movement of the background is of low temporal frequency. The noise, on the other hand, contains mainly high temporal frequencies. The background movement makes the parvo channel dominate the response of the model. The dynamic noise does not activate the masking in the parvo channel and the model will, therefore, underestimate the threshold.

PSNR does not take masking into account, but the dynamic range of the background, which makes it adjust the threshold slightly for different backgrounds. When the masking effects are strong, as in sequence 4, this is not enough. However, the sample set is

too small to say which of the measures is the better. In Eckstein *et al.* (1997) [Ref. 5] a simple measure of contrast energy adjusted to the background with the mean luminance, did perform comparably and even better than many of the vision models in the investigation. Compared to the Johnson criteria, which does not take the background into consideration, both PSNR and the spatio-temporal model are better.

3. Acknowledgements

Börje Andrén carried through the experiments. FMV (the Swedish Defence Materiel Administration) gave us the raw infrared imagery. FOA (the Swedish Defence Research Establishment) provided the IR-system simulation software. All the observers contributions are gratefully acknowledged. This work was supported by the companies Telia Research AB, Ericsson Telecom AB, Ericsson SAAB Avionics AB, CelsiusTech Electronics AB, Bofors AB, FMV and NUTEK (the Swedish Board for Industrial and Technical Development).

4. References

1. A. J. Ahumada Jr., B. L. Beard, and R. Eriksson, "Spatio-temporal image discrimination model predicts temporal masking function", *Human Vision and Electronic Imaging III,* B. Rogowitz, T. N. Pappas, eds., Proc. 3299, pp. 120–127, SPIE, Bellingham, WA, 1998.
2. R. Eriksson, "A spatio-temporal vision model for digital video", Technical Report, TR 331, Institute of Optical Research, Kista, Sweden, 1998.
3. G. E. Legge and J. M. Foley, "Contrast masking in human vision", J. Opt. Soc. Am, 70, pp. 1458–1471, 1980.
4. J. M. Foley, "Human luminance pattern-vision mechanisms: masking experiments require a new model", J. Opt. Soc. Am., 11, pp. 1710–1719, 1994.
5. M. P. Eckstein, A. J. Ahumada Jr., and A. B. Watson, "Image discrimination models predict signal detection in natural medical image backgrounds", *Human Vision and Electronic Imaging II*, Proc. 3016, pp. 44–56, SPIE, Bellingham, WA, 1997.
6. A. J. Ahumada Jr and B. L. Beard, "Object detection in noisy scene", *Human Vision and Digital Display VII,* B. Rogowitz, J. Allebach, eds., Proc. 2657, pp. 190–199, SPIE, Bellingham, WA, 1996.
7. A. J. Ahumada Jr, A. B. Watson, and A. M. Rohaly, "Models of human image discrimination predict object detection in natural backgrounds", *Human Vision and Digital Display VI,* B. Rogowitz, J. Allebach, eds., Proc. 2411, SPIE, pp. 355-362, Bellingham, WA, 1996.
8. C. Wigren, "IGOSS Model of image generation in optronic sensor systems", Technical report, FOA-R–97-00582-616–SE, Linköping, Sweden, 1997.
9. J. Finney, "Probit Analysis: A statistical treatment of the sigmoid response curve", University Press, Cambridge, Cambridgeshire, 1952.
10. A. B. Watson, H. B. Barlow, and J.G. Robson, "What does the eye see best?", Nature 302, pp. 419–422, 1983.