GUIDELINES FOR ACCURATE TOD MEASUREMENT

Piet Bijl[1] and J. Mathieu Valeton
TNO Human Factors Research Institute
P.O. Box 23, 3769 ZG Soesterberg, The Netherlands

ABSTRACT

Recently[1,2,3], we presented the Triangle Orientation Discrimination (TOD) threshold as an alternative method to characterize electro-optical system performance. The TOD has a number of theoretical and practical advantages over the MRTD and MRC. The method is based on an improved test pattern, a better-defined observer task and a solid psychophysical measurement procedure, and allows accurate and comparable results from different measuring teams.

Guidelines to perform TOD measurements are given in the present paper. The optimal range of test pattern sizes and contrasts are specified, as well as the required number of presentations for a threshold estimate. Special attention is paid to the statistical analysis. A standard frequency-of-seeing curve is fitted to the observer data in order to obtain 75%-correct thresholds. A χ^2-statistic provides an objective criterion for acceptance or rejection of the threshold estimates. Finally, a complete TOD curve is obtained by fitting a weighted least-square polynomial through the 75%-correct thresholds. Further, a simple Go-NoGo screening procedure with objective pass/fail criteria, based on the TOD methodology, is proposed.

With the TOD methodology, accurate sensor performance measuring and Go-NoGo testing have become very easy to carry out. Therefor, the investment in a thoroughly designed measurement setup will pay itself back easily.

Keywords: electro-optical system performance testing, TOD methodology, standard measurement procedure, objective acceptance/rejection criteria

1. INTRODUCTION

The current standards to characterize Electro-Optical (EO) system performance, the MRTD (Minimum Resolvable Temperature Difference) and MRC (Minimum Resolvable Contrast) have several well-known disadvantages which hinder accurate measurement. One is the subjective measurement procedure: the observer indicates when he is just able to resolve the test pattern. Disadvantages are that the answer cannot be verified, and that the result depends on the observer's internal decision criterion. Criterion differences occur between observers, but also the criterion of the observer changes, even during a session. Holst[4] reports a number of typical problems that may occur during a measurement, such as internal observer variability, learning, or distraction. Solutions to reduce the effects of subjective measurement (or observer bias) on the MRTD are: training and the use of a large number of observers. Both are time-consuming.

In vision science, bias-free psychophysical measurement procedures are widely used[5]. The advantages of such procedures over an adjustment procedure (as used in the MRTD/MRC) are: (i) thresholds are independent of the observer's internal decision criterion, (ii) the observer task is relatively easy, and (iii) the reliability of the observer responses can be checked statistically. Usually, a bias-free procedure requires a relatively large number of stimulus presentations. On the other hand, many observer responses can be collected in a short time frame because the observer's decisions are easy to make and require less time than the judgement process in an adjustment procedure. Further, less training is required and accurate threshold estimates can be obtained with a smaller number of observers.

The TOD (Triangle Orientation Discrimination threshold)[1,2,3], which we presented recently as an alternative to the MRTD and MRC, makes use of a bias-free psychophysical measurement procedure. We use the TOD methodology for two purposes: sensor performance specification, and Go-NoGo testing. The basics of the method, the advantages over the MRTD and MRC, and a validation of the TOD have been presented elsewhere[1,2,3]. In the present paper, guidelines are given to perform a TOD measurement.

In chapter 2, definitions are given. Chapter 3 describes the methods, and chapter 4 the statistical analysis. Chapter 5 gives guidelines for Go-NoGo testing.

[1] Further information: bijl@tm.tno.nl; tel +31 346 356277; fax +31 346 353977

2. DEFINITIONS

2.1 TEST PATTERN

The test pattern is an equilateral triangle on a uniform background. The triangle has four possible orientations: one of the apexes is directed up (U), down (D), left (L) or right (R) (see Fig. 1).

2.2 TRIANGLE SIZE

Triangle size S is defined as the *square-root area* (in visual angle) and is expressed in mrad. Note that the triangle base (the length of a side) is:

$$Base = \left(\frac{2}{\sqrt[4]{3}} \right) \cdot S \approx 1.5 \cdot S \tag{1}$$

We introduce the quantity *Reciprocal Angular Subtense:* $S_R = S^{-1}$. S_R will be used as ordinate of the TOD. Its unit ($mrad^{-1}$) is the same as for spatial frequency, which makes the TOD curve easily comparable to the MRTD and MRC. Further, S_R is proportional to target acquisition range.

2.3 TRIANGLE CONTRAST

Thermal contrast $\Delta T = T_T - T_B$, where T_T is the temperature of the test pattern and T_B of the background, expressed in *K*. Visual contrast $C = (L_T - L_B) / L_B \cdot 100\%$, where L_T is the luminance of the target, and L_B is the luminance of the background. Contrast will be used as the coordinate of the TOD.

2.4 OBSERVER TASK

In each presentation, a test pattern of certain (angular) size and contrast is presented to an observer through the imaging system under test; the orientation of the test pattern (U, D, L, R; see Fig. 1) is pseudo-randomly chosen. The observer task is Four-Alternative Forced-Choice (4AFC): the observer has to indicate which of the four orientations is perceived, even if he is not sure.

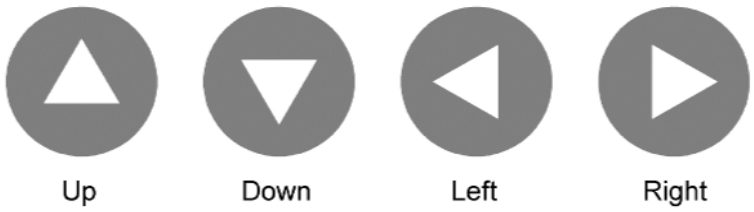

Fig. 1: The test pattern or stimulus in the TOD method is an equilateral triangle and has one of four possible orientations: apex Up, Down, Left or Right. The observer has to indicate which orientation he perceives (a 4AFC task, see 2.4).

2.5 RELATIONSHIP BETWEEN OBSERVER PERFORMANCE, STIMULUS SIZE AND CONTRAST

The fraction of correct observer responses is a function of both stimulus size and contrast. This relationship is called the psychometric function. It is hypothesized that the fraction of a correct response increases monotonically with stimulus size or contrast (in general: with stimulus strength), and that the psychometric function can be described by a Weibull function of the form:

$$P_{\alpha\beta\gamma\delta}(x) = (1-\delta) - (1-\gamma-\delta)\cdot 2^{-(x/\alpha)^{\beta}} \tag{2}$$

where x = stimulus strength (either contrast or size), α is a stimulus strength threshold value (see below), β determines the steepness of the function, γ is the guess rate, i.e. in a 4AFC task $\gamma = 0.25$, and δ (usually set at 0.02) is called 'finger', i.e. the probability that the observer erroneously pushes a wrong button or misses a presentation because he blinked with his eyes at that moment. The Weibull function is shown in Fig. 2.

The exact probability level at stimulus strength $x = \alpha$ depends on the values of β, γ and δ. Therefore, we define the stimulus strength α_{75} at which 75% of the responses is correct. The relation between α_{75} and α is given by:

$$\frac{\alpha_{75}}{\alpha} = \left[-^{2}\log\left(\frac{0.25-\delta}{(1-\gamma-\delta)} \right) \right]^{1/\beta} \tag{3}$$

With $\gamma = 0.25$, $\delta = 0.02$ and typical values for β: 3 – 8, (α_{75}/α) varies between 1.18 and 1.06.

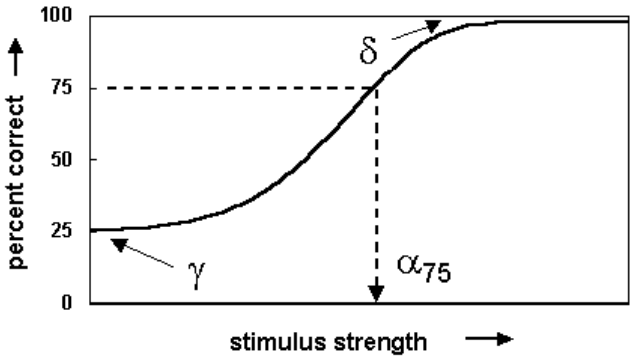

Fig. 2: Example of a psychometric function. The fraction correct gradually increases with stimulus strength (contrast or size) from chance (γ A 100% = 25% in a 4AFC task) to approximately 100% (in this example: (1-δ)A100% = 98%). The threshold α_{75} is defined as the stimulus strength at which the observer scores 75% correct, and is independent of the decision criterion of the observer. The threshold is obtained by fitting the Weibull function (equation 2) to the frequency-of-seeing data of the observers at different stimulus strengths and calculating the 75% correct point (equation 3).

2.6 TOD THRESHOLD

At a given stimulus size S (or reciprocal angular subtense S_R, see 2.2), threshold contrast ΔT_{75} or C_{75} is defined as the contrast at which the observer judges the triangle orientation correctly in 75% of the presentations.

Similarly, at a given contrast ΔT or C, threshold stimulus size S_{75} (in mrad) and reciprocal angular subtense $S_{R,75}$ (in mrad^{-1}) are defined as the stimulus size and its reciprocal at which the observer judges the triangle orientation correctly in 75% of the presentations.

The TOD threshold is obtained by fitting the Weibull function (equation 2) to the frequency-of-seeing data of the observers at different stimulus strengths (the fit procedure is given in section 4.2) and calculating the 75% correct point (equation 3).

2.7 TOD SENSOR ACUITY

TOD *sensor acuity SA* is a measure of the smallest detail that can be resolved with the sensor at high contrast (comparable to resolution limit).

SA is defined as the threshold $S_{R,75}$ (in mrad^{-1}; see 2.6) at a predefined high contrast (averaged over positive and negative contrast). For *visual cameras*, *SA* is defined at contrast $C = \pm 100\%$, for *thermal cameras* it is defined at $\Delta T = \pm 2$ K.

2.8 TOD CURVE

The TOD curve is defined as the relationship between the reciprocal angular subtense S_R and contrast ΔT or C at the 75% correct level. S_R (in mrad^{-1}) is plotted on a linear scale, and contrast on a log scale. An example of a TOD curve for a CCD camera at two luminance levels is given in Fig 3. How the TOD curve can be obtained is described chapters 3 and 4.

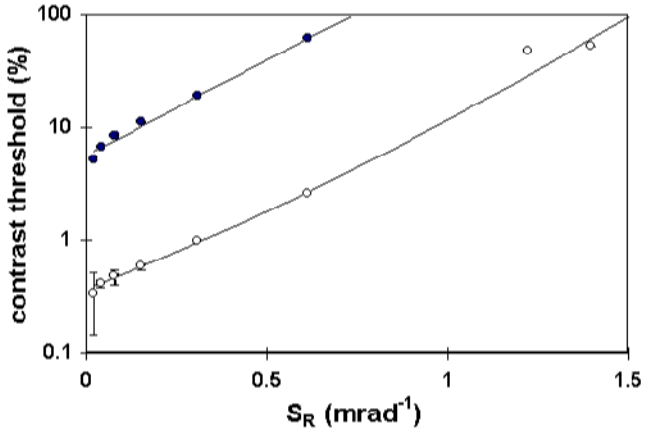

Fig. 3: Example of a TOD curve for a CCD camera at two background luminance levels. Open circles: L = 165 cd/m^2; Filled circles: L = 0.33 cd/m^2; Solid lines: best polynomial fits (see section 4.2). For this sensor, *SA* (sensor acuity) is about 1.50 mrad^{-1} at the higher background luminance and 0.73 mrad^{-1} at the lower background luminance.

3. METHODS

This chapter is organized as follows. In section 3.1, a number of considerations for accurate TOD measurement are given. Design choices are discussed and explained. Section 3.2 describes an adjustment procedure that preceeds the 4AFC procedure. This procedure is used to obtain initial threshold estimates and to optimize the sensor settings for the final threshold estimate. In section 3.3, additional measurements for visual devices (dependence on background luminance) are described. A step-by-step description of the TOD measurement procedure is given in section 3.4.

3.1 CONSIDERATIONS

3.1.1 *Sensor settings*

During the 4AFC measurement, it is *not allowed to change the sensor settings*. This is essential because doing so would change the shape of the psychometric function (Fig. 2) during the test and yield erroneous threshold values. Sensor settings are optimized during an adjustment procedure (section 3.2), and changes during the measurement are not necessary since the variation in stimulus strength is relatively small. Further, it is *not allowed* (and not necessary) *to change the sensor viewing direction* during a measurement. In the TOD procedure, the effect of the relative position of the test pattern with respect to the imaging plane on performance (e.g. caused by sampling) is automatically taken into account (see section 3.1.6).

In this respect, the procedure differs significantly from the traditional MRTD where the observer is allowed to change sensor settings and viewing direction during a measurement in order to optimize the image. Advantages of the present procedure are that it eases the observer task and considerably reduces the duration of a threshold estimate (see 3.1.11).

3.1.2 *TOD threshold measurements: vary contrast or size?*

Like the MRTD, a TOD curve is obtained from a number of threshold estimates at different stimulus sizes. However, these thresholds may be obtained either by measuring the fraction correct observer responses as a function of contrast at fixed stimulus sizes (and fitting a psychometric function through the frequency-of-seeing data to obtain the 75% correct point) or vice versa. The role of contrast and size as variables is essentially equivalent, however in some regions of the *size-contrast space* it is most efficient to vary contrast (for large stimulus sizes), and in other regions it is most efficient to vary size (at high contrasts, near the sensor acuity limit). In principle, a complete 2-D approach is possible in which both stimulus size and contrast are varied and stimulus strength (e.g. in equation (2)) is defined as a combination of these two variables. Such a procedure is under development.

The approach here is as follows. First, sensor acuity *SA* (see 2.7) is determined by varying stimulus size at a fixed, high contrast. Based on *SA*, a number of suitable triangle sizes are selected for which the TOD contrast thresholds are measured using contrast as variable. It is advised to select at least 7 stimulus sizes, ranging from 1.2-30 times the acuity stimulus size.

3.1.3 *Number of stimulus sizes or contrasts*

Each threshold estimate is obtained by measuring fraction correct as a function of stimulus strength (size or contrast). It is advised that the number of different stimulus strengths I for a threshold estimate is at least 5.

When triangle size is the variable, the size difference between the targets should be about 0.06 log units (about 15%). When contrast is the variable, contrasts should differ by about 0.1 log units (about 25%). In general, the optimal differences between adjacent stimulus levels depend on the slope of the psychometric function.

3.1.4 *Number of presentations*

The accuracy of a threshold estimate depends on the number of presentations per stimulus size or contrast. It is advised that the number of presentations per stimulus strength is at least $N = 16$. Thus, with $I = 5$ different stimulus strengths (see section 3.1.3), a single threshold estimate is based on I A $N = 80$ stimulus presentations.

3.1.5 *Orientation balance*

It is advised that the number of presentations of stimuli with the same size and contrast are *balanced* with respect to triangle orientation. For example, if the number of presentations of a certain stimulus pattern is 16, show each orientation 4 times. Balancing is necessary in order to average out the effect of response assymmetries (e.g. an observer guesses up or down more often than left or right), or assymmetries caused by system properties.

3.1.6 *Pseudo-random position shift*

In each presentation, *a small pseudo-random shift* has to be added to the horizontal and vertical position of the test pattern. It is advised that at least 4 horizontal and vertical positions are used, differing in position by about $0.5*S$.

The position shift is an essential part of the procedure in order to correctly account for sampling effects. Note that in this respect the procedure differs from the MRTD measurement where the observer may optimize the position of the test pattern with respect to the sampling array. In the TOD procedure outlined here, shifting the test pattern across the imaging plane is taken care of automatically.

3.1.7 *Presentation order*

Thresholds might depend on the stimulus presentation order. It is advised to start the procedure with the largest stimuli or highest contrast.

3.1.8 *Stimulus duration*

Stimulus presentation time can be short, typically in the order of 2 s.

3.1.9 *Positive and negative contrast*

It is advised that a threshold measurement is performed for both positive and negative triangle contrast. Especially in the case of low thermal contrasts, measurements for opposite contrasts should be performed quickly after each other to cancel out a possible offset in temperature difference between target and background.

3.1.10 *Observers*
It is advised that at least three observers with normal visual acuity participate in the test.

3.1.11 *Duration of a TOD measurement*
With N = 16 presentations per stimulus size or contrast, I = 5 sizes or contrasts per threshold estimate, and a presentation time of 2 s, an observer response time of 1 s, measuring positive and negative contrasts alternately after the same adjustment procedure, and an adjustment time of 3 min, a single threshold estimate for positive and negative contrast together will take about (16 * 5 * 3 s)* 2 + 3 min = 11 min.

With 8 threshold estimates for a complete curve, the complete measurement time for a single observer is about 8 * 11 min = 88 minutes. This can be performed in two or three separate sessions. Remeasurements of some points might be necessary (see section 4.1). With visual devices, measurements should be performed at different luminance levels (3.3) which will of course increase the duration for a complete TOD measurement.

3.2 ADJUSTMENT PROCEDURE

Each threshold measurement starts with an *adjustment procedure*. The purposes of the adjustment procedure are (1) to optimize the sensor settings (focus, sensor gain and level, viewing direction), and (2) to obtain a initial TOD threshold estimate in order to suitably choose the sizes or contrasts near the 75% correct point for the 4AFC task.

The adjustment procedure is an iterative process:

1. Show a fixation cross. It is advised to present both a □+□ and a □x□. The observer focusses the sensor.

2. Continuously show a triangle test pattern. The observer varies stimulus size or contrast (whichever is desired) until the pattern seems near threshold. The observer optimizes gain and level settings of the sensor for this pattern. Again he adjusts stimulus size or contrast, and iterates this procedure until he feels satified about sensor settings and threshold estimate. During the procedure, he is allowed to move the viewing direction slightly to correctly judge the effect of sampling. Repeat the procedure a few times, each time with opposite contrast, starting with the most recent threshold estimate. Stop if the observer is satisfied about the sensor settings and threshold estimate for both positive and negative contrast.

3. If a more accurate threshold estimate is desired, the procedure can be followed by a number of presentations of the stimulus pattern having the same temporal envelope as in the 4AFC procedure. Proceed if the observer is satisfied about the sensor settings and the threshold estimate. Otherwise, return to 2.

4. The most recent threshold estimate is used as a starting point for the 4AFC procedure; Changing the sensor settings and viewing direction is no longer allowed (see 3.1.1).

3.3 SENSITIVITY TO BACKGROUND LUMINANCE (VISUAL DEVICES ONLY)

The performance of a visual device depends on background luminance. Usually, the TOD (and also the MRC) shifts vertically when background luminance is varied[1,6].

For a quick estimate of the sensor sensitivity to background luminance, it is advised to measure contrast threshold as a function of background luminance for a single, relatively large test pattern (at least 10 times the acuity threshold) using the adjustment procedure described in section 3.2[6]. An example of the result for a CCD camera system is shown in Fig. 4.

Now measure the TOD curve at two luminance levels. One at a high background luminance where thresholds are independent of background luminance, and one at a low background luminance, where contrast threshold is about 10 times the value for high luminances. On the basis of Fig. 4, it was decided to measure the TOD for this sensor at L = 165 cd/m^2 and 0.33 cd/m^2, respectively. These TOD curves are shown in Fig. 3. TOD estimates for other background luminances can be estimated using Fig. 4.

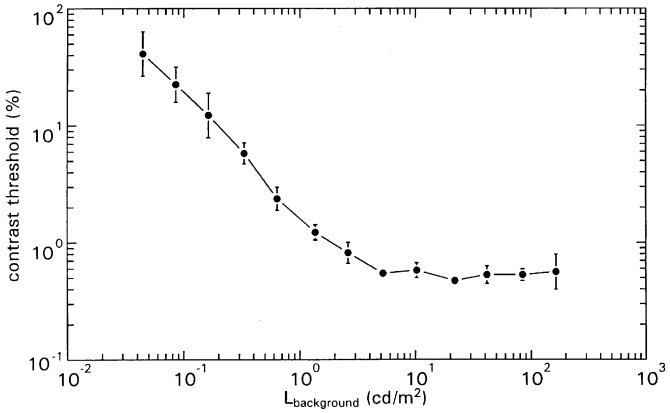

Fig. 4: Contrast threshold as a function of background luminance for a CCD camera system. Above L = 5 cd/m^2, thresholds are independent of background luminance. Below L = 1 cd/m^2, thresholds vary inversely proportional to background luminance. It is sufficient to measure a TOD curve at only two luminances (see text).

3.4 MEASUREMENT PROCEDURE

In the TOD measurement procedure, two main steps can be discerned. First, sensor acuity *SA*, the threshold at fixed high contrast ($\Delta T = \pm$ 2 K or $C = \pm$ 100%) is determined (step A). Then, a number of larger test patterns are selected on the basis of the *SA*, and contrast thresholds are determined for these test patterns (step B). The procedure is as follows:

Step A: determination of SA

1. Approximate *SA* (see 2.7) using an adjustment procedure (3.2). The procedure is illustrated in Fig. 5a.

2. Carry out the 4AFC procedure to obtain a final estimate of *SA* (Fig. 5b). It is not allowed to change the sensor settings and viewing direction (3.1.1). It is advised that the number of different triangle sizes *I* is at least 5, differing in size by about 0.06 log units (3.1.3), around the initial threshold estimate determined with the adjustment procedure. Use at least 16 presentations per stimulus size (3.1.4), balance the number of presentations of stimuli with the same size with respect to triangle orientation (3.1.5) and use the pseudo-random stimulus position shift described in section 3.1.6. Start the procedure with the largest triangle size (3.1.7). Perform the measurement for both positive and negative stimulus contrast (3.1.9).

3. Calculate sensor acuity *SA* (the 75% correct threshold) using the procedure described in section 4.1.

Step B: determination of other TOD thresholds

4. Select a number of suitable triangle sizes S_j (j = 1 to J, $J \exists 7$), ranging from about 1.2 to at least 30 times the acuity stimulus size SA^{-1} (section 3.1.2). The selection process is illustrated in Fig. 5b. Use the same stimulus sizes for different observers.

5. For each stimulus size, estimate the contrast threshold and optimize the sensor settings using the adjustment procedure described in 3.2 (Fig. 5c).

6. For each stimulus size, carry out the 4AFC procedure (Fig. 5c). It is not allowed to change the sensor settings and viewing direction (3.1.1). It is advised that the number of different contrasts *I* is at least 5, differing in contrast by about 0.1 log units (3.1.3). Use at least 16 presentations per stimulus size (3.1.4), balance the number of presentations of stimuli with the same contrast with respect to triangle orientation (3.1.5) and use the pseudo-random stimulus position shift described in section 3.1.6. Start the procedure with the highest contrast (3.1.7). Perform the measurement for both positive and negative stimulus contrasts (3.1.9).

7. Calculate the 75% correct thresholds (4.1), geometrically average over observers (4.1) and fit a polynomial curve through the data (4.2). The result is shown in Fig. 5d.

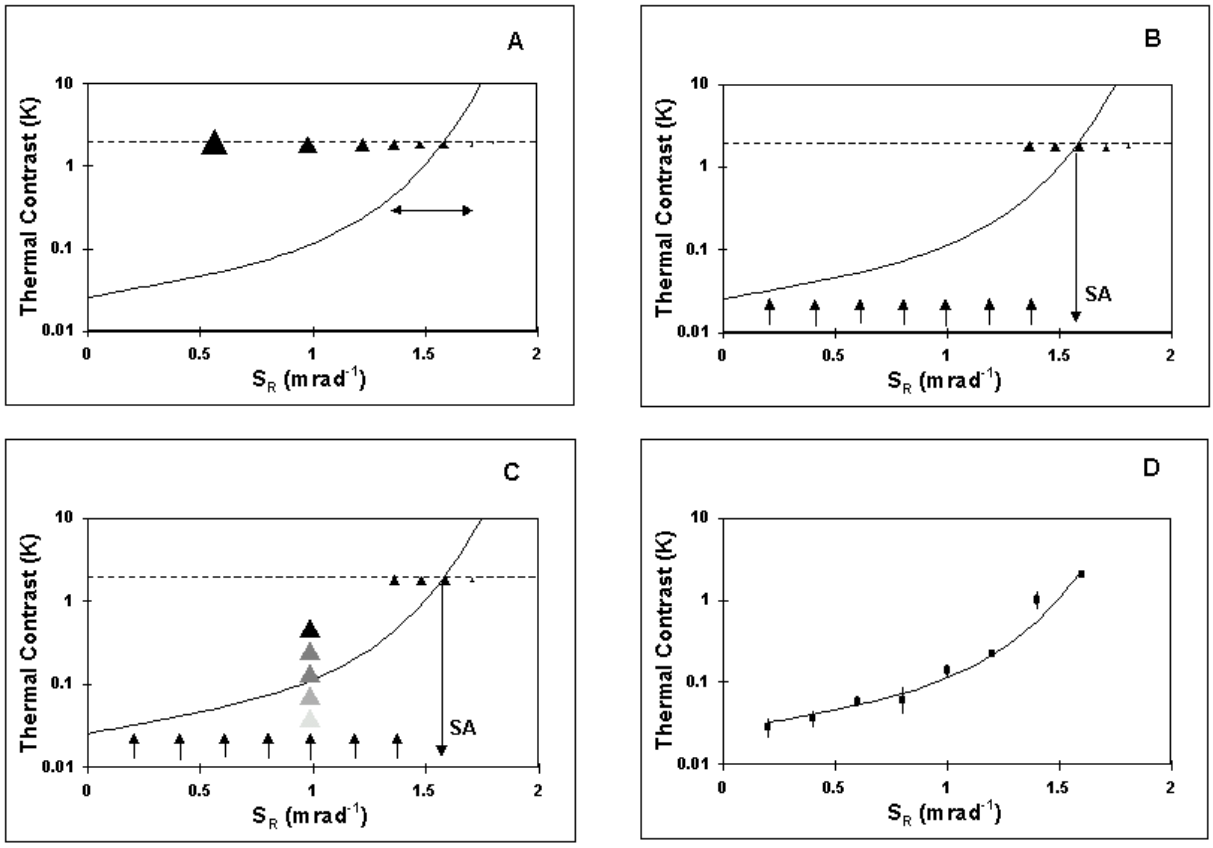

Fig. 5: TOD measurement procedure. A: Obtain an initial estimate of sensor acuity *SA* at high contrast using an adjustment procedure; B: Measure *SA* using the 4AFC procedure, and select a number of suitable triangle sizes based on *SA*; C: Measure the contrast threshold for the selected stimuli; D: Calculate the 75% correct thresholds and fit a polynomial curve through the data (chapter 4).

4. ANALYSIS

4.1 THRESHOLD CALCULATION

A maximum-likelihood estimate of the 75% correct level is obtained by fitting the Weibull function (equation 2) through the observer data using a minimalization procedure proposed by Watson[7]. The procedure is as follows:

1. Analyze the data for each threshold estimate, positive or negative contrast, and observer separately.

2. Calculate the two functions M_0 and M_1:

$$M_0 = \sum_{i=1}^{I} n_i \left[p_i \cdot \ln\left(p_i\right) + \left(1 - p_i\right) \cdot \ln\left(1 - p_i\right) \right] \tag{4}$$

$$M_1 = \sum_{i=1}^{I} n_i \left[p_i \cdot \ln\left(P_i\right) + \left(1 - p_i\right) \cdot \ln\left(1 - P_i\right) \right] \tag{5}$$

where I is the number of different stimulus strengths (3.1.3); n_i (= N) is the number of presentations at stimulus strength a_i (3.1.4); p_i is the measured fraction correct at stimulus strength x_i, and $P_i = P_{\alpha\beta\gamma\delta}(x_i)$ is the predicted fraction correct at stimulus strength a_I according to the Weibull function given in equation (2).

3. Maximize M_1 by varying α and β ($\gamma = 0.25$ and $\delta = 0.02$ are fixed). This gives the maximum likelihood estimate of α and β. If the steepness of the psychometric function β is known (e.g. from earlier measurements), β may be held fixed and M_1 may be maximized by varying α.

4. Calculate $2[M_0 - M_1]$. Asymptotically, (i.e. with a sufficient number of presentations) this value approaches χ^2 with degrees of freedom df = $I - F$, where I is the number of stimulus strengths in the experiment (e.g. $I = 5$), and F is the number of free parameters in the psychometric function (i.e. $F = 2$ if α and β are free, and $F = 1$ if α is free and β fixed).

5. Choose a confidence level of $p = 0.95$.

6. Reject the data if $2[M_0 - M_1] > \chi^2_{p,\mathrm{df}}$. $\chi^2_{p,\mathrm{df}}$ can be obtained from a χ^2-table or a probability calculator in a statistics program. For example, if the number of contrasts in a measurement is $I = 5$, and a Weibull function is fitted to the data with both α and β free ($F = 2$), accept the data if $2[M_0 - M_1] < \chi^2_{0.95,3} = 7.82$. With a confidence level of 0.95, about 5% of the fits will be rejected on the basis of this criterion.

7. Calculate the 75% correct threshold α_{75} (equation (3)).

8. Also reject the measurement if α_{75} falls outside the range of presented stimulus strengths x_i.

9. Average over positive and negative contrast: $\alpha_{75} = (\alpha^+_{75} + \alpha^-_{75})/2$. In the case of low thermal contrasts, a possible temperature offset between target and background is cancelled out. The standard error reduces by about a factor 1//2.

10. Calculate mean and standard error in the mean of the *log*-values of the thresholds for different observers (the average of the log-values is equivalent to the geometrical average of the thresholds).

Example. Suppose we have presented the following stimulus sizes x_i ($i = 1$ to I, $I = 5$), differing 0.06 log units in size: 1, 1.15, 1.32, 1.51, 1.74 mrad. The number of presentations for each stimulus size is $N = 16$. The observer correct scores p_i ($i =$ 1 to 5) are respectively: 6/16, 10/16, 9/16, 14/16, 15/16. Then $M_0 = -41.9$ (equation 4). Choose $\gamma = 0.25$ and $\delta = 0.02$, and maximize M_1 by varying α and β. This yields: $M_1 = -42.8$, and the maximum likelihood estimates $\alpha = 1.26$ mrad and $\beta =$ 4.59. Equation 3 yields: $\alpha_{75} = 1.41$ mrad. The degree of freedom is 3, and $\chi^2_{0.95,3} = 7.82 > 2[M_0 - M_1]$ which means that the fit is accepted. α_{75} Falls within the range of presented stimulus sizes x_i, thus the threshold measurement is valid.

4.2 MATHEMATICAL DESCRIPTION OF THE TOD CURVE

A polynomial function is fitted to the data. The lowest order polynomial function that adequately decribes the relationship between the reciprocal angular subtense S_R ($= S^{-1}$) and log contrast (within the experimental error) is determined by the following procedure:

1. The observer data (calculated in 4.1) are a table of the form: $S_{R,j}$, $\log C_j$, $\sigma_{m,j}$ ($j = 1$ to J), where $\log C_j$ is the average over observers of the log values of the contrast thresholds for triangle test pattern j, and $\sigma_{m,j}$ is the standard error in the mean of those log values ($= \sigma/\sqrt{M}$, where M is the number of observers).

2. Plot the thresholds and the standard errors in the mean on a graph with S_R (in mrad^{-1}) on the ordinate (linear scale) and contrast on the abscis (log scale).

3. The polynomial is defined as:

$$\log C(S) = \sum_{k=0}^{K} b_k \cdot S^{-k} \tag{6}$$

where C is contrast, S is stimulus size, K is the order of the polynomial, and b_k are the coefficients of the function.

4. Start with the lowest order polynomial ($K = 0$) and find the maximum likelihood coefficients (b_j) by minimizing:

$$\chi^2 = \sum_j \left(\frac{\log C_j - \log C(S_j)}{\sigma_{m,j}} \right)^2 \tag{7}$$

which is a measure of the difference between measured and calculated values.

5. Repeat step 4 for the next order polynomial, and calculate $\Delta\chi^2$, i.e. the difference between the two χ^2-values for the two polynomials.

6. Test if $\Delta\chi^2_{p,df}$ is significant at the p = 0.95 level and with df = 1, using a table or a probability calculator. If the difference is not significant, the higher order polynomial does not descibe the data significantly better than the lower order polynomial, and the latter is the best fit. If the χ^2 difference is significant, repeat 4 and 5 with the next order polynomial.

7. The polynomial fit is valid between the minimum and maximum stimulus size in the set. This is called the TOD curve. Plot the TOD curve together with the data (see Fig. 5d). Usually, a first to third order fit adequately describes the data.

5. GO-NOGO TESTING

A Go-NoGo test is a (usually) quick measurement to determine whether sensor performance is still sufficient. Applications are e.g. routine tests performed by maintenance personnel. In our laboratory a Go-NoGo test apparatus was developed named TIPI: Thermal Imager Performance Indicator. This apparatus tests sensor performance at two points: (i) TOD sensor acuity *SA* (threshold stimulus size at a fixed, high thermal contrast), and (ii) the TOD contrast threshold (75% correct contrast) for a large stimulus size. The test procedure is equivalent for both points.

Essential in a Go-NoGo test are:

- The performance for a well-functioning sensor of the type (the reference sensor) must be known. For example, α_{75} and β for the sensor were measured shortly after it was manufactured, and are used as reference for routine testing during its life cycle.

- A criterion ε (> 0) is set: the rejection threshold for the sensor under test (α'_{75}) may be higher (performance lower) than for the reference by a certain amount: $\alpha'_{75} = (1 + \varepsilon)$ A α_{75}. If high performance is desired (e.g. if a mission depends critically on sensor performance), a low value of ε is chosen.

- There is always a (statistical) probability that a bad sensor passes or a good sensor fails the test. These probabilities depend on the number of stimulus presentations used. The smaller the criterion ε, the larger the possibility that a normal sensor will be rejected and the higher the number of presentations should be to avoid this.

The Go-NoGo test proposed here is extremely simple since no fit procedure is required. The total number of correct responses (over different stimulus strengths) is simply counted and compared to the number of correct answers that is expected on the basis of the psychometric function of the reference sensor (α_{75} and β) and the criterion ε. This also makes the test procedure more reliable: especially if the sensor under test is poor and the stimulus strengths fall outside the transition region of the psychometric function, a fitting procedure would be unstable while counting correct responses is not.

The procedure is as follows:

1. For the reference sensor, α_{75} and β, and thus the relationship between fraction correct and stimulus strength (contrast or size) are known.

2. Choose a criterion. For example, use $\varepsilon = 0.25$ for normal use, and $\varepsilon = 0.10$ for critical purpose.

3. Calculate: $\alpha'_{75} = (1 + \varepsilon)$ A α_{75}.

4. Choose stimulus strengths x'_i (i = 1 … I, according to 3.1.3) *around the rejection or criterion threshold* α'_{75} (*not around* α_{75}). If I is odd, choose the middle stimulus strength at or slightly above α'_{75}. The reason is that in that case the measurement procedure is most sensitive to a sensor which operates near the criterion threshold. An additional advantage is that with a good sensor system the task is easy, which is pleasant for the observer.

5. Choose a number of presentations per stimulus strength N. The total number of stimulus presentations will be I A N.

6. Do the measurement. Start with an adjustment procedure as given in 3.2 in order to optimize the sensor settings (it is not necessary to find an initial threshold estimate since the criterion threshold α'_{75} is already known). Start the 4AFC test procedure with the highest stimulus strength.

7. Calculate the *total number* of correct answers $N_{correct}$.

8. For a sensor that just satisfies the rejection threshold α'_{75}, the *expected total number* of correct answers will be:

$$N_{crit} = N \cdot \sum_{i=1}^{I} P_{\alpha'\beta\gamma\delta}(x'_i) \tag{8}$$

   (note that N_{crit} is not an integer).

9. The sensor passes the test (Go) if the number of correct answers $N_{correct} \geq N_{crit}$. Otherwise it fails (NoGo).

10. Since correct scores are binomially distributed, the standard error in the expected number of correct answers N_{crit} will be:

$$\sigma = \sqrt{\sum_{i=1}^{I} \sigma_i^2} = \sqrt{\sum_{i=1}^{I} \left[\frac{P_{\alpha'\beta\gamma\delta}(x'_i) \cdot (1 - P_{\alpha'\beta\gamma\delta}(x'_i))}{N} \right]^2} \tag{9}$$

    The standard error is used to calculate the probabilities of undesired acceptance or rejection (see the example).

11. The procedure may be speeded up for poor sensors by first calculating the number correct for the highest stimulus strength. A low number suggests that the sensor is poor or the settings are not optimal. In that case, repeat the adjustment procedure and start again. If the result for the highest stimulus strength is again poor, the sensor fails the test.

Example. Suppose a well functioning sensor has α_{75} = 1.41 mrad and β = 4.58 (the result from the example in section 4.1). We choose a strict rejection criterion: ε = 0.10 or 10%. Thus, α'_{75} = 1.1A1.41= 1.55 mrad. We select 5 stimulus strengths x'_i around α'_{75}, differing 0.06 log units: 1.18, 1.35, 1.55, 1.78, 2.04 mrad. The number of presentations per stimulus size N = 16, thus the total number of presentations is I A N = 80.

For a sensor that just satisfies the criterion, the expected number of correct answers N_{crit} can be calculated from equation (8): N_{crit} = 58.5. Thus the criterion is: $N_{correct}$ ∃ 59: Go, and $N_{correct}$ # 58: NoGo. Note that for a sensor that just satisfies the criterion, the probabilities of Go and NoGo both are 50%.

How large is the probability that a good sensor fails or a bad sensor passes the test? For the reference sensor (α_{75} = 1.41 mrad, 10% better than the criterion), the expected number of correct answers is 64.5 ± 3.21 (use equations (8) and (9) with ε = 0). The probability that $N_{correct} \leq 58$ (NoGo) can be calculated using a table or probability calculator for Z-scores: Z = (64.5-58)/3.21 = 2.02 yields p = 0.021, or only 2.1%. For a sensor with α_{75} = 1.71 mrad (10% worse than the criterion), the expected number of correct answers is 51.9 ± 3.89 (from equations (8) and (9) with ε = 0.21). The probability that $N_{correct} \geq$ 59 (Go) is p = 3.5%. Thus, even for the strict criterion used here (ε =10%), the probability of undesired acceptance or

rejection is very low with only 80 presentations. With a normal criterion, usually 40 presentations (taking only a few minutes, see 3.1.11) are sufficient.

6. DISCUSSION

The present paper describes a number of essential guidelines for objective measurement of EO system performance using the TOD[1,2,3] methodology. The purpose is to make this methodology, which makes use of procedures that were developed in vision science, accessible to the EO systems testing community. Another goal is to contribute to a new international agreement on a method to characterize EO system performance in order to ensure compatibility between the results from different measuring teams.

The paper is organized in such a way that the reader will be able to apply the measurement procedures step by step to his own testing facility. Both the assessment of complete TOD curves and quick Go-NoGo testing are discussed. Whichever level of testing is applied, objectivity of the criteria and thorough statistical analysis are emphasized.

It is not the objective to be complete. Just as with the measurement of the MRTD or MRC, the test engineer should be aware of specific test requirements and pitfalls. Measurements should always be performed with care, and a number conditions that need to be satisfied are not mentioned here. For example, with thermal imagers the background temperature should be specified (for the MRTD it should be approximately 20 °C, or otherwise a conversion factor has to be used), the emissivities of the paint should be specified, etc. For visual camera's, the color temperature of the light source and the spectral sensitivity of the sensor should be taken into account, care should be taken when the camera has AGC (automatic gain correction), etc. These type of details are described elsewhere. For example, the STANAGs for MRTD and MRC give a number of guidelines, and procedures for EO systems testing including good advise for test engineers is given by Holst[4].

The bottom line is: With the TOD methodology, sensor performance measuring and Go-NoGo testing have become accurate, fast and very easy to carry out. The price to pay is only a very thorough design of the measurement setup. This is not a trivial thing to do, but it has to be done only once. Work is under way to put TOD measurement equipment on the market, in which case only the advantages of the TOD method remain.

7 REFERENCES

1. P. Bijl and J.M. Valeton, "TOD, a new method to characterize electro-optical system performance, *Proc. SPIE* 3377, 182-193, 1998.
2. P. Bijl and J.M. Valeton, "TOD, the alternative to MRTD and MRC", *Optical Engineering* 37 (7), 1976-1983, 1998.
3. P. Bijl and J.M. Valeton, "Validation of the new triangle orientation discrimination method and ACQUIRE model predictions using observer performance data for ship targets", *Optical Engineering* 37 (7), 1984-1994, 1998.
4. Holst, G.C. *Testing and Evaluation of Infrared Imaging systems*. Second edition. SPIE Press, Bellingham, WA, 1998.
5. C.J. Bartleson and F. Grum. *Optical Radiation Measurements. Volume 5: Visual Measurements.* Academic Press, Inc., Orlando, FL, 1984.
6. P. Bijl and J.M. Valeton, "Bias-free procedure for the measurement of MRTD and MRC", in press, *Optical Engineering.*
7. Watson (1979). Probability summation over time, *Vision Research*, 19, 515-522, 1979.