A critical evaluation of test patterns for EO system performance characterization

Piet Bijl[*], J. Mathieu Valeton and Maarten A. Hogervorst
TNO Human Factors, P.O. Box 23, 3769 ZG Soesterberg, The Netherlands

ABSTRACT

The traditional test pattern for end-to-end EO system performance testing in the laboratory has been the static 3- or 4-bar target. This choice was governed by linear systems approach. The introduction of under-sampled imagers such as IRFPAs (infrared focal plane array cameras) has challenged the testing community to develop an alternative test, because the occurrence of aliasing has a completely different effect on periodic targets (such as the bar target) and real, non-periodic targets. A new test should at least have the following properties: lab testing is objective and easy, the measure is representative for field performance, and modeling (sensor and human) the test should be relatively easy. Several alternative test methods and test patterns have already been proposed. An example is the TOD method that uses non-periodic test patterns. Other examples are the dynamic MRT that uses a moving 4-bar target, and the MTDP that uses the traditional static target but allows that not all four bars have to be present in the image. The development of real-time scene projection allows testing with real infrared targets under controlled conditions. The authors will discuss a large number of test patterns and methods and show their advantages and disadvantages for end-to-end EO system performance testing. They conclude that simple non-periodic spatial test patterns, such as used in the TOD, are the best choice for sensor performance characterization.

Keywords: Electro-Optical system performance testing, TOD, MRTD, Dynamic MRT, MTDP

1 INTRODUCTION

A standard laboratory test to determine human performance with Electro-Optical (EO) viewing systems is required for several purposes, e.g. (i) to verify if a sensor is working properly, (ii) to compare competing sensor systems, (iii) to verify if a new type of sensor meets the expectations based on its design, or (iv) to predict field performance. Field performance can be target acquisition (TA) performance in a military or a civilian environment (security), or reading performance with the unaided eye for example.

Basically, there are two approaches for such a test. The first is to measure essential physical parameters of the sensor system or parts of the system (e.g. the MTF or the NETD), and then use a model to predict human performance. However, this requires a thorough understanding of human vision with sensors, and current vision models are not sophisticated enough to cope with all target and sensor factors that affect visual performance.

The second approach is to measure end-to-end sensor performance including the human observer. It is often not very practical (or even possible) to perform such measurements for a real task in a real situation. Therefore, laboratory Sensor Performance Measures (such as the MRTD[1]) have been introduced. In the laboratory, circumstances can be controlled and measurements can be performed quicker, easier, and with higher accuracy. It is obvious that the results of such a lab test need to be representative for the corresponding tasks in the field, so that field performance can be predicted from the laboratory measurements.

Until now, end-to-end EO system performance tests in the laboratory are usually performed with simple test patterns on uniform backgrounds. Examples of simple test patterns that are often used for EO system testing and in optometry are shown in Fig. 1. One of the patterns is the four-bar target that is used in the current standard measure, the MRTD (Fig. 1a). Other examples are an equilateral triangle in four possible

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

orientations, used in the TOD[2] (Triangle Orientation Discrimination threshold, Fig. 1e), and the Landolt-C that is often used to measure the visual acuity of the unaided eye (Fig. 1f).

Recently, the development of dynamic IR scene generators has greatly extended the possibilities to generate IR test patterns. It has now become possible to present in the laboratory e.g. (stationary or moving) real thermal targets in real backgrounds under controlled conditions (see Fig. 2a, for example). These may seem to be the ultimate test patterns from a field performance point-of-view, but they also have disadvantages, as will be shown in this paper. Other possibilities with the scene generator are sine wave gratings (an example is shown in Fig. 2b), stationary or moving in arbitrary directions and with an arbitrary temporal envelope. A sine wave grating may be better suited for system performance characterization than a bar target because it contains no higher harmonics.

In this paper, we will examine a large number of test patterns and methods and show their advantages and disadvantages for end-to-end EO system performance testing. Chapter 2 describes the scope of the survey. Chapter 3 describes essential properties for a laboratory Sensor Performance Measure, and in Chapter 4 the actual test pattern analysis is performed. The results are summarized in Chapter 5.

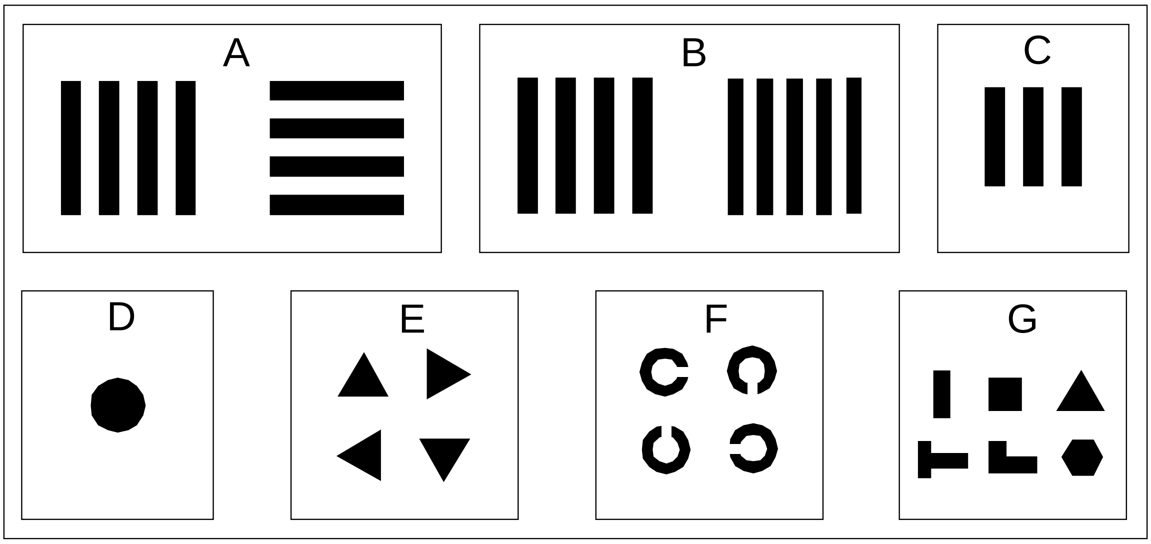

Fig. 1: Examples of test patterns that are often used in end-to-end EO system performance tests or in optometry: (a) standard 4-bar vertical and horizontal test pattern, used for MRTD[1], MTDP[3] and Dynamic MRT[4], (b) 4-bar test pattern and 5-bar reference pattern used for the bias-free MRTD[5] (c) 3-bar USAF target used for MRC (Image Intensifiers), (d) circular disc, used for MDTD, (e) Triangle test pattern (in 4 orientations) used in the TOD[2],(f) Landolt-C (4 or 8 orientations) used by optometrists, (g) various simple shapes used by Fairhurst and Lettington[6].

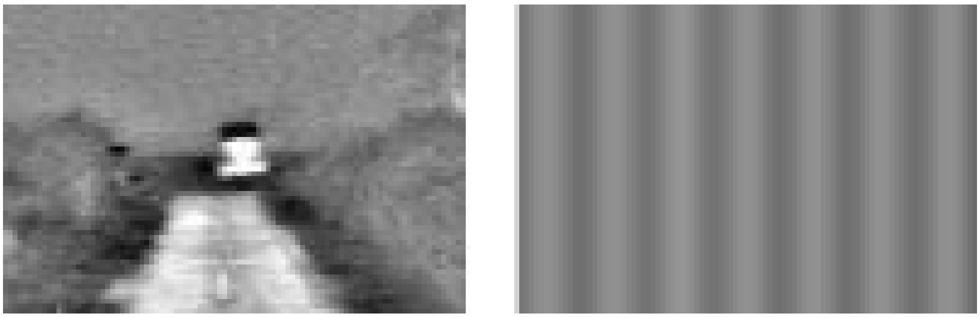

Fig. 2: Examples of possible test patterns with a dynamic IR scene generator: (a) real thermal target in real background, (b) sine wave grating.

2 SCOPE

We limit ourselves to end-to-end system performance measures, i.e. measures that cover the overall system, including the human observer. Automatic target recognition (ATR) is not considered. The survey is further restricted to certain field tasks (2.1), sensor types (2.2), and to specific (families of) test patterns (2.3).

2.1 FIELD TASK

The laboratory sensor performance measure should be representative for the real application or task you want to perform with the system. These tasks can be widely different. In this paper we limit ourselves to *detection, recognition and identification of targets in a military environment by human observers*. Everyday life recognition tasks, including reading, may be described very well by the same laboratory measure but are not explicitly taken into account. Visual search is excluded.

2.2 SENSORS

The image forming systems that are considered operate in a single spectral band (visual, short-wave IR, mid-wave IR or long-wave IR), may be well-sampled or under-sampled, and may have microscan. These include: Optical systems, Scanners (in 1-D or 2-D), Focal plane Array cameras, Image Intensiers (1st, 2nd or 3rd Gen) or CCD cameras. Not considered are: systems that perform complex non-linear operations such as image enhancement techniques, systems that use image fusion, and radar. The test patterns may be applicable to these systems, but this is not required.

2.3 TEST PATTERNS

We only consider single test patterns on a uniform background and real targets. For most of the simple test patterns discussed in this paper, temperature or luminance is constant over the test pattern. The observer usually knows the location of the test pattern in the image.

3 EVALUATION CONSIDERATIONS

For an end-to-end Sensor Performance Measure, three things are important[7]: (i) the lab testing procedure, (ii) the relationship with field performance, and (iii) modeling.

3.1 LAB TESTING PROCEDURE

A laboratory test should be unambiguous, objective, and should yield reproducible results including a measure of accuracy. Preferably, the test is easy and efficient and the test procedure is independent of the type of sensor under test. In the next chapter, it will be shown that the choice of the test pattern(s) has an important impact on the possibility to realize such a test.

Since a Sensor Performance Measure (SPM) represents human performance, the test is usually performed with the human-in-the-loop. The alternative is automated measurement. It is difficult to develop a reliable automated test because it requires a vision model and current models do not yet cope with all factors that affect visual performance.

A human observer introduces a number of variance sources[8]. Examples are: visual acuity and contrast sensitivity differences between observers, criterion differences, differences in training level and expertise, differences in motivation, day-to-day variations within observer, fatigue, differences in sensor settings (within and between observers). The effect of a number of factors on the results can be minimized when the correct psychophysical measurement procedure and test patterns are used.

3.1.1 *Psychophysical procedure*

Two well-known psychophysical procedures will be discussed here. One is the Adjustment procedure, the other is a multiple-alternative forced-choice (nAFC, with n = the number of alternatives). For accurate measurements the nAFC procedure is highly preferable above the Adjustment method, as will be illustrated in the examples below.

A: nAFC procedure. Suppose we want to determine the range at which a tank can just be identified with a certain sensor. There are six possible tank targets (Leopard 2, M1A1, T-72, etc.) at a range of distances. These are presented to an observer using the sensor. The observer is instructed to name the presented target, even if he is not sure (this is a little different from the military practice, in which an observer can indicate that he does not know which target it is). This is a typical nAFC task, with n=6. With this task, the exact fraction of correct identification can be measured as a function of target range, for example. The observer criterion (i.e. his confidence that he chooses the correct target) has no effect on the results, only his knowledge of the targets. A curve can be fitted through the fraction correct vs. range relationship, the range at an exact probability correct threshold level (e.g. 75% or 50%) can be calculated, and the consistency of the responses can be checked statistically. Exactly the same procedure is used in the TOD, where the observer has to choose which of four possible triangle orientations was presented (see section 4.6). An nAFC procedure also has advantages for modeling and automatic measurement[9,10] (see section 3.3).

B: Adjustment procedure. Now suppose we use the Adjustment procedure. In that case we only need one tank (e.g. a Leopard-2) at different ranges, and the observer knows which tank it is. The observer is asked to indicate the range at which he *thinks* he can identify the tank as a Leopard-2. This is a typical Adjustment task (or a Method of Limits if the range is controlled by the experimenter). The same method is used for the MRTD (section 4.1). It is a good and quick method to obtain first threshold estimates. However, the threshold range depends on the criterion of the observer (his confidence that he would really be able to identify the target as a Leopard 2 in a real situation with other possible targets), and we do not know to what probability it corresponds nor can his choice be checked. Further, we do not obtain a complete fraction correct vs. range relationship, only a single value. Training is required in order to reduce criterion differences between observers, and the task can be stressful for an observer who wishes to make a well-considered choice (in an nAFC-task, he only has to make the best choice he can). In order to obtain accurate thresholds, this procedure may take more time than an nAFC procedure because more observers are required, and more training is needed.

3.1.2 *Test pattern*

The use of simple test patterns has a number of advantages over the use of more complex patterns. For example, it reduces the effect of cognitive factors which means that less training on recognition of the patterns is required. It also facilitates the observer task. Other advantages are that simple test patterns are technically easier to manufacture, and modeling will be easier.

The choice of a test pattern has consequences for the psychophysical measurement procedures that can be used. It will be shown in Chapter 4 that an nAFC procedure cannot be used for some test patterns.

3.2 RELATION WITH FIELD PERFORMANCE

3.2.1 Major target and sensor properties

Although there are many factors of target and background that affect TA performance, two major parameters are target (angular) size and target-background contrast. Fig. 3a illustrates the relationship between these parameters and, e.g., target identification performance (the same applies to recognition). From left to right, target angular size decreases (or range increases: target angular size is inversely proportional to target range). From the top down, target-background contrast decreases. At the top-left corner (large, high contrast targets), target identification is easy and an observer who knows the targets will score a 100% correct. At the bottom-right corner, identification is impossible. The dashed curve indicates the 75%-correct threshold for target identification. This curve runs from the detection limit near the bottom-left corner to the upper-right corner where TA performance is resolution limited. Fig. 3b-d illustrate the effect of some types of image degradation that often occur in EO systems on the 75%-correct threshold: Blur (b) and Sampling (d) result in a curve shift towards larger targets, Noise (c) results in a curve shift towards higher contrasts. Usually, a combination of different types of image degradation occurs in the EO system. For example, a very simple description of a FPA camera would be: blur (optics) + sampling (FPA) + noise + blur (electronics + display).

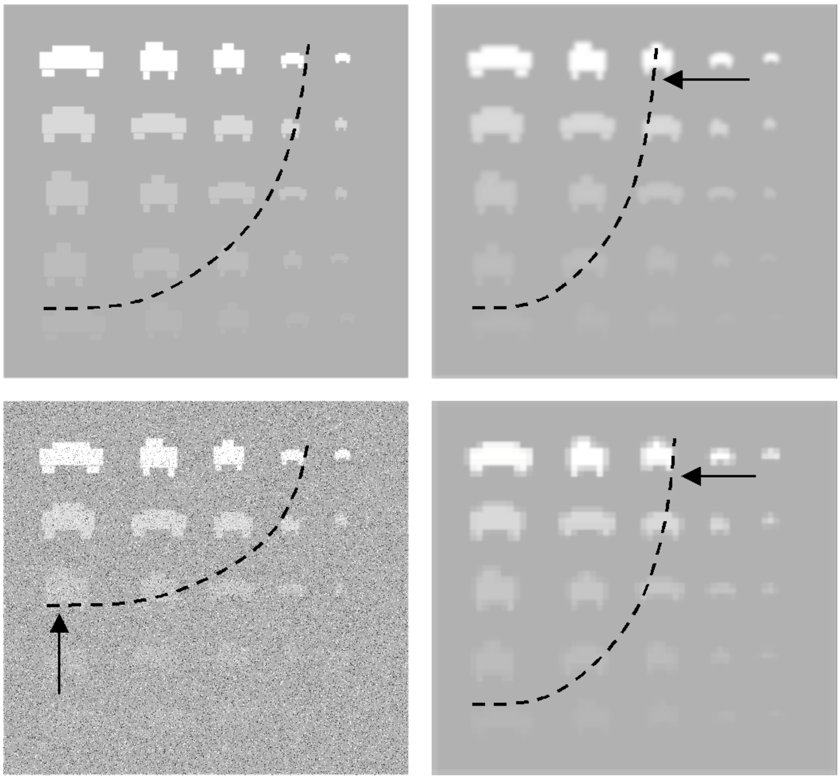

Fig. 3. (a) Target (angular) size and contrast are two major factors determining TA performance. In this figure, target angular size decreases (range increases) from left to right, and contrast decreases from top to bottom. The dashed curve indicates the 75% correct TA threshold. (b, c, d) Blur, Noise, and Sampling are important types of degradation by the sensor that affect TA performance. These figures show the effect of these types of image degradations on the 75% correct threshold. A good lab Sensor Performance Measure or model yields a curve that has the same shape as the curve for real targets, and the effects of image degradation of different types (or a combination) on this curve are the same as they are for real targets.

3.2.2 Assumptions for the TA process

At present, the TA process (target identification and recognition in the field) is not fully understood and modeled. Therefore, current TA models such as ACQUIRE are based on two simple assumptions:

- Only two target and background parameters are used: target angular size (usually √area) and a simple contrast measure (usually the average apparent contrast between target and background).
- Sensor effects are included using a laboratory Sensor Performance Measure (either measured or calculated), and an experimentally derived factor between this lab measure and field performance.

For example, ACQUIRE uses the MRTD which is a laboratory measure that describes the relationship between spatial frequency (or cycle width) and the contrast at which an observer using an EO system is able to resolve a 4-bar test pattern, and the so-called cycle criteria N that relate the MRTD to field performance. Currently, the recommended N_{75} for target identification is 8, which means that the model predicts 75% correct identification performance in the field when a target is at such a range that 8 cycles of the MRTD bar pattern can be resolved over the target.

3.2.3 Consequences for Sensor Performance Measures

The second assumption in section 3.2.2 places important demands to the lab Sensor Performance Measure. It means that:

1. The curve that is produced in the lab (either by measurement or calculation) must have essentially the same shape as the curve for real targets (as shown in Fig. 3a)
2. The effects of image degradation by the sensor like blur, noise, and sampling (Fig. 3b, c, d) or any combination, have to be the same for the lab and field curve.

Only then the laboratory measure is representative for field performance and a comparison between sensors of different types based on the laboratory measure is also valid for field performance.

3.3 MODELING

In principle, modeling the lab measurement is easiest and most reliable if the procedure is unambiguously clear and objective, and if the observer task and test pattern are simple. Modeling and automatic measurement are more straightforward for a forced-choice measurement procedure than with an adjustment method. In that case, the vision model may be equipped with a decision routine that chooses the highest probable alternative, just as the human would do[9,10]. This always leads to a threshold at the same probability correct level, even if the degradation of the images is completely different due to different sensor systems (noise, blur, under-sampling). With an adjustment procedure, a threshold (e.g. a modulation depth in the case of an MRTD bar pattern) has to be set that may work out differently for different sensor systems.

4 OVERVIEW OF TEST PATTERNS AND PROCEDURES

In this chapter, we will discuss a number of test patterns. Most of the examples are currently applied in existing Sensor Performance Measures. The outcome of a test depends not only on the pattern, but also largely on the test procedure used. Some test patterns are used in combination with more than one test procedure. Therefore, we will discuss the applicability of the test pattern in combination with the procedure.

For each combination of test pattern and procedure, the three items mentioned in Chapter 3: Lab measurement, Field performance, and Modeling will be discussed separately. The test patterns in this section can be subdivided into three categories: simple periodic (4.1-4.5), simple non-periodic (4.6-4.8), and real targets (4.9).

4.1 STANDARD FOUR-BAR PATTERN 1: CONVENTIONAL MRTD

Lab measurement: In the conventional MRTD, the test pattern is the standard 4-bar pattern (vertical and horizontal) shown in Fig. 1a. The task of the observer is to indicate the contrast at which he is just able to resolve the four bars. The 2D-MRTD is defined as the geometric average of the horizontal and vertical MRTD. The measurement procedure is described in STANAG 4349[1].

A major disadvantage of the MRTD is that for under-sampled imagers the method is only applicable up to the Nyquist frequency (i.e. half the sampling frequency) of the sampling array. Above this frequency the MRTD is not defined, and usually the Nyquist frequency is taken as the cut-off of the MRTD curve for under-sampled imagers. An additional disadvantage is that an artificial frequency limit has to be set, and that knowledge of the system (i.e. the sampling frequency) is required. Higher spatial frequencies cannot be reproduced by the system and are aliased back into lower frequencies. As a result, the image may be a pattern of 4, 3, 2 or 1 unevenly spaced low frequency bars of different widths and contrasts that depends very much on spatial frequency and phase (i.e. the relative position of the test pattern with respect to the sampling array). It is clear that the original observer task (resolving the four bars) is no longer possible, and that the judgement of such an image has little to do with resolution of the original test pattern. Even below half the sampling frequency problems may occur[2,11].

The second disadvantage of the MRTD method is the subjective adjustment procedure (see 3.1). The disadvantages of such a procedure are pointed out in section 3.1. For application of a 2AFC procedure, see 4.2.

Field performance: The MRTD for under-sampled imagers is *not* a good representative for field performance, because it cuts off at the Nyquist frequency. For real targets, a strong cut-off at a certain size or range has not been observed. This is illustrated in Fig. 4a. Driggers et al.[7] clearly point out that the MRTD makes an unfair comparison between well-sampled and under-sampled sensors, and also between different types of under-sampled cameras, e.g. between cooled and uncooled cameras. This is also shown experimentally by Bijl, Valeton & de Jong[12].

Modeling: The bar pattern is convenient when linear systems approach is applicable to a sensor system, and when the process is separable in horizontal and vertical direction. This is often the case for well-sampled systems, and sometimes for under-sampled systems up to the Nyquist frequency (but not always for cameras with micro-scan). The FLIR92 model predicts the MRTD based on the calculation of the system MTF. The test pattern is mathematically separable in the horizontal and vertical direction, and orientation and spatial frequency are well defined (although a sine-wave grating would be more appropriate because it contains no higher harmonics; see section 4.5).

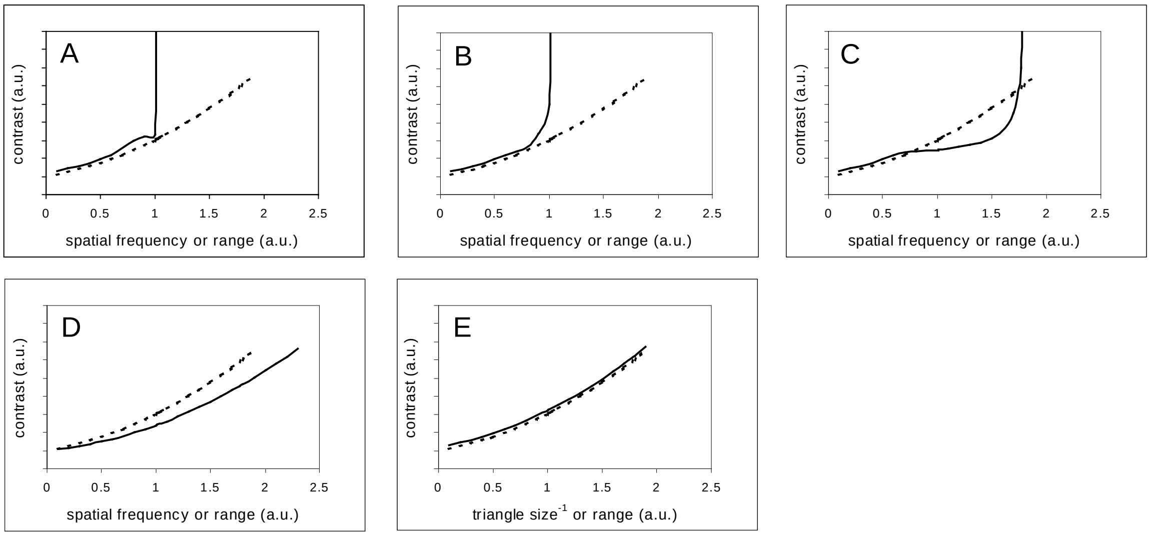

Fig. 4: Illustration of the relation between a number of laboratory Sensor Performance Measures and field performance for *under-sampled* imagers. Dashed line: Typical Contrast *vs*. TA range relationship for real targets (as in Fig. 3). Continuous lines: Laboratory curves. (a) Conventional MRTD (4.1), (b) Bias-free MRTD (4.2), (c) MTDP (4.3), (d) Dynamic MRTD (4.4), (e) TOD (4.6). In these examples, it is assumed that the lab curves for a *well-sampled* imager are optimally scaled to the corresponding field data (e.g. by using the correct cycle criteria, see 3.2.2). See text for details.

4.2 STANDARD FOUR-BAR PATTERN 2: BIAS-FREE MRTD

Lab measurement: In this test, the observer has to choose whether a standard 4-bar test pattern or a 5-bar reference pattern is shown. The test and reference pattern are shown in Fig. 1b. As with the standard MRTD, the test can be performed in horizontal and vertical direction. The test is described in detail by Bijl & Valeton[5].

The major difference with the conventional MRTD (section 4.1) is the 2AFC procedure, having all the advantages mentioned in section 3.1. An additional advantage is that the Bias-free MRTD-curve for under-sampled imagers cuts off near Nyquist so that it is not necessary to have a priori knowledge of the system or to set an artificial frequency limit. This makes the entire procedure independent of sensor type.

Field performance: The Bias-free MRTD for under-sampled imagers is too steep and cuts off near the Nyquist frequency[13] (see Fig. 4b). Therefore it not a good representative for field performance (in fact it has the disadvantages of the MRTD which are pointed out in section 4.1).

Modeling: Modeling is comparable to the MRTD, except that the 2AFC procedure is probably easier to model than the adjustment procedure.

4.3 STANDARD 4-BAR PATTERN 3: MTDP

Lab measurement: The test pattern and measurement procedure are identical to the conventional MRTD, except that the observer task is to indicate the contrast at which he is just able to resolve the bars. The number of bars is not necessarily 4, but may also be 3 or 2 (for under-sampled systems). The measurement procedure is described in detail by Wittenstein[3].

The advantage of the MTDP over the MRTD is that the procedure is well-defined for both well-sampled and under-sampled imagers also beyond the Nyquist frequency. No a priori knowledge of the system (such as the sample size) is required.

A disadvantage is that the method is still an adjustment procedure. To our knowledge, it is not possible to define a sound forced-choice procedure for the MTDP (such as the bias-free MRTD).

Field performance: For well-sampled imagers, the MTDP and MRTD curves are almost identical. An important advantage of the MTDP over the MRTD is, that the curve for under-sampled imagers does not cut-off at the Nyquist frequency. Validation studies show that the MTDP predicts the performance differences between scanning systems and sampling array cameras at high thermal contrasts well[3,14].

The shape of the MTDP curve for under-sampled imagers does *not* match the contrast vs. range relationship for real targets. Fig. 4c shows an MTDP curve for an under-sampled imager. The slope of the curve is very shallow between 1 and 1.5 times the Nyquist frequency: this is the region where the number of bars decreases from 4 to 2. Near the cut-off frequency, the curve is very steep. Such a shape has not been observed and is also not expected for real targets, because these targets contain many spatial frequencies.

Modeling: The mathematical advantages of the 4-bar pattern where given in section 4.1. TRM3[3] is a comprehensive model that predicts the MTDP for well-sampled and under-sampled imagers. The model is not linear, and calculation is performed partly by simulation.

4.4 STANDARD FOUR-BAR PATTERN 4: DYNAMIC MRTD (DMRT)

Lab measurement: The only difference of the Dynamic MRTD (DMRT) with the conventional MRTD is that the test pattern is moving at a certain speed in the direction perpendicular to the bars. The task of the observer is to indicate the contrast at which he is just able to resolve the four bars.

The advantage of the DMRT is that motion with the correct speed reduces the effects of sampling on the image of the test pattern, and this makes the method applicable to under-sampled imagers. Another advantage is that the method closely resembles the conventional MRTD.

A disadvantage is that the psychophysical method is still an adjustment procedure. It is probably possible to apply the bias-free MRTD procedure (section 4.2) to the dynamic MRTD. Another disadvantage of the measurement procedure is that the optimum speed of the test pattern depends on the sample size of the system. This means (i) that the measurement method is system dependent, and (ii) that a priori knowledge of the sensor system is required.

Field performance: The DMRT does not include sampling effects and therefore it does *not* directly relate to field performance (actually it is too good for under-sampled imagers, and largely independent of the sample size of the Focal Plan Array). See Fig. 4d. A correction is necessary to account for the effects of sampling on target identification and recognition. These effects are experimentally and theoretically well characterized.

Modeling: NVTherm[7] calculates the DMRT. DMRT mainly characterizes pre- and post-sampling parts of the sensor system and these can be described reasonably well with linear systems approach (very similar to FLIR92). In order to be able to predict TA performance, NVTherm models sampling as an additional blur MTF (the MTF squeeze approach).

4.5 SINE WAVE GRATING

A sine wave grating (see Fig. 2b) is commonly used in vision science, for example to measure the Contrast Sensitivity Function (CSF) of a human observer. The advantage over a bar pattern is that it contains no higher harmonics, which may be convenient for both measurement and modeling.

4.6 TRIANGLE TEST PATTERN: TOD

Lab measurement: In the TOD, the test pattern is an equilateral triangle in four possible orientations (apex Up, Down, Left, Right) shown in Fig. 1e. The task of the observers to indicate which orientation was presented (orientation discrimination). The method is described in detail by Bijl & Valeton[15].

The psychophysical procedure is 4AFC and has all the advantages of a forced-choice procedure outlined in section 3.1. The TOD test procedure for EO viewing systems is very well-defined.

The test pattern is simple, which means that little training is required to learn to recognize the alternatives. The triangular shape is not essential, but has some advantages over alternative simple spatial test patterns (see 4.7 and 4.8).

The main advantage of the TOD method is that it is applicable without limitation to all the sensor types listed in section 2.2. The test is independent of sensor type, no a priory knowledge of the sensor is required, and the underlying physical mechanisms of the sensor are irrelevant. There is no problem with the Nyquist frequency limit. The test can easily be performed with moving test patterns to characterize the dynamic behaviour of a sensor.

Field performance: Validation studies have shown that (i) the shape of the TOD curve matches the contrast vs. range relationship for real targets very well[13] (see Fig. 4e), and (ii) that the TOD predicts the performance differences between scanning systems and sampling array cameras at high thermal contrasts well[12,14].

The rationale of the TOD method is that the test patterns represent features or the relation between features of real targets. If an observer is able to discriminate between the different triangle orientations, he also has information about the target features necessary to identify a target. Note that the limits of current sensor systems are clearly determined by the TOD method: the observer is not able to determine the correct orientation of the original test target if (i) the test pattern cannot be detected (because the SNR is too low), (ii) its shape cannot be determined (because corners and edges disappeared due to blur or sampling), or (iii) the shape is incorrectly judged (relative positions are disturbed due to under-sampling or to phase shifts introduced by electronics). Because the TOD method has such a close relationship to real target acquisition, it is likely that the method can be used (maybe in a slightly adapted form) for future imaging systems.

Modeling: The test pattern is not separable in the horizontal and vertical direction, and contains many spatial frequencies (similar to a real target). This is not a problem for a model that calculates TOD performance by computer simulation, but it is difficult to make a closed-form TOD sensor model (such as NVTherm).

Hogervorst et al.[10] developed a model that predicts the TOD for well- and under-sampled imagers. A convenient property of the model is that it consists of a separate sensor part and a plausible human vision model. The decision routine is simple because the measurement procedure is forced-choice (see section 3.1). The vision model may also be used separately for automatic TOD measurement.

4.7 LANDOLT-C

The Landolt-C pattern (Fig. 1f) is a good alternative to the triangle test pattern used in the TOD with only some minor disadvantages: (i) phase shifts do not lead very easily to misjudgements in the position of the gap, (ii) a Landolt-C is less easy to manufacture and (iii) the triangle has a higher degree of symmetry.

4.8 VARIOUS SIMPLE SHAPES

Fairhurst and Lettington[6] use a set of various simple spatial test patterns to characterize the effect of image degradation on recognition (see Fig. 1g for some examples). This set of shapes may be a good alternative to the triangle test patterns but has some disadvantages: (i) the patterns are less easy to learn, (ii) it is more difficult to the observer to indicate his choice, (iii) some of the test patterns are more easily confused than others, which makes the analysis and modeling more complicated, and (iv) the triangle has a higher degree of symmetry.

4.9 REAL TARGETS

Lab measurement: Real thermal targets (Fig. 2a) may be used now or in the near future to measure IR sensor performance in the laboratory with a dynamic IR scene generator. The observer task is target recognition or identification.

An advantage of using real targets is that it is easy to design an nAFC (n-alternative forced-choice) test procedure, with n being the number of targets or target classes in this set. All the advantages of such a procedure are outlined in section 3.1.

A disadvantage is that cognitive factors play an important role. Even with a limited target set, observer training may take days. Performance depends strongly upon the observer, which means that they have to be trained exactly up to a certain level. Other problems are that observers may get to know some targets too well, or that picture recognition may occur, which means that the observer knows the background rather than the target. Further, the definition of size and contrast of the targets is a problem.

Using real targets for sensor performance characterization requires an exact definition of scenarios, target images in the set, the specifications of the experimental setup, and of observer training. These will have to be used everywhere over the world (which is not yet feasible or cost-effective because infrared dynamic scene generators are still very expensive).

Field performance: Of course, a laboratory measurement with real targets is expected to be representative for field performance. However, in practice there are an infinite number of scenarios and targets, and only a very limited part of these can be used in the lab test. The predictive power of a limited test for other scenarios has to be assessed.

Modeling: There is not yet a human vision model that accurately predicts TA performance.

5. SUMMARY OF THE RESULTS

5.1 PERIODIC TEST PATTERNS

The use of periodic test patterns for Electro-Optical system performance characterization is understandable from a historic point-of-view because early sensors could often be treated as linear systems, at least for small signals. In that case, the response to an arbitrary stimulus signal can be predicted from the system MTF (Modulation Transfer Function), which is the (amplitude and phase) response of a system to a sinosoidal signal as a function of frequency[1]. In the thermal domain, a sine wave grating was very difficult to manufacure, so that a bar pattern was an obvious alternative.

With the increasing complexity of sensor systems the conventional approach becomes more and more difficult to maintain. The difficulties have become apparent for the MRTD after the introduction of (under-) sampled imaging systems[11]. In the present analysis, two things have become clear with respect to periodic

[1] Note that the linearity refers to the *viewing system*, not to the human observer. Human vision is highly non-linear, and it is not easy to predict vision thresholds for non-periodic patterns from thresholds for gratings[16]. This may be another reason to avoid testing with periodic patterns.

test patterns. First, none of the presented patterns and measurement procedures leads to a curve that is directly representative for TA performance. Second, with periodic targets (in combination with under-sampled imagers) it is very difficult to apply a multiple-alternative forced-choice psychophysical procedure due to spurious response. Such a procedure is important for two reasons: (i) accurate measurement in the laboratory, and (ii) correspondence with the field task.

Nevertheless, recently introduced methods that use the standard four-bar test pattern better predict TA performance in the field than the conventional MRTD does, be it in an indirect way. The DMRT suppresses the effects of under-sampling by motion (with the objective to be able to apply the linear systems approach), and needs a correction to account for sampling effects. The MTDP seems to predict the differences between well- and under-sampled systems correctly, but the curve for under-sampled systems has an unusual shape due to the interference of the test pattern with the pixel array.

Sampling is only one step in a process of increasing sensor complexity. Further digitization at several stages of sensor systems (e.g. image enhancement algorithms, scene-based sensor calibration, image fusion) or other techniques (e.g. diagonal dither or microscan) will make is almost impossible to consider a system as linear and separable in two directions, and to use concepts as MTF or bandwidth. This greatly reduces the advantage of periodic, separable test patterns.

5.2 REAL TARGETS

Using real targets in real backgrounds of course has a close relationship to field performance. These targets, however, are not very suitable for routine testing. One reason is that cognitive factors and training level play an important role in the lab measurement. Another reason is that modeling is very difficult.

5.3 NON-PERIODIC SIMPLE SHAPES

Non-periodic, symbolic test patterns share a number of advantages of real targets and periodic test patterns. They can be considered to represent details or features of real targets, so there exists a natural link with real targets. If there are more than one alternative shapes in the set, than the task is close to a target acquisition process. Their simplicity on the other hand makes it easier to characterize, standardize, measure, and model the process than with real targets.

There is an enormous freedom of choice in symbols (Triangle, Landolt-C, various simple shapes, letters, etc.), and there is also freedom in test procedure (adjust or nAFC). Not all shapes are applicable, and some symbols are more useful for TA performance characterization, while others may be more suited for different tasks (e.g. reading).

On the basis of simplicity of the observer task (and modeling), it is convenient to use a Landolt-C or a Triangle test pattern in four possible orientations, as is used in the TOD. Learning such a task requires almost no training for the observer. Furthermore, using a single test pattern in different orientations has a number of practical advantages. Theoretically, the link with real targets is better for the triangle test pattern than it is for the Landolt-C. Finally, the TOD has been thoroughly developed, tested and evaluated for Electro-Optical viewing systems, and a TOD sensor performance model is under development. Therefore, to our knowledge the Triangle test pattern and the TOD method are currently the best available alternative for end-to-end Sensor Performance Characterization.

6. CONCLUSIONS

1. An end-to-end EO system laboratory measure that is representative for TA performance is required because current vision models are not sophisticated enough to directly predict field performance from sensor and (complex) target properties.

2. With increasing sensor complexity it will be more difficult to understand and describe sensor performance. Therefore, the best strategy is to design a laboratory measure that is close to the real TA task, but yet as simple as possible.

3. A number of *simple, non-periodic* test patterns are well-suited for (i) laboratory testing, (ii) field performance predictions, and (iii) modeling of advanced sensors. Real targets are not so suited for routine laboratory testing because cognitive factors play an important role, and the process is difficult to model. Periodic patterns are not suited for laboratory testing and not directly representative for TA performance of under-sampled imagers.

4. The TOD is a good example of a Sensor Performance Measure that uses well-suited simple, non-periodic test patterns.

REFERENCES

1. STANAG 4349.
2. P. Bijl and J.M. Valeton, "TOD, the alternative to MRTD and MRC". Optical Engineering 37(7), 1984-1994 (1998).
3. Wittenstein, W. (1999). Minimum temperature difference perceived – a new approach to assess undersampled thermal imagers. Optical Engineering 38, 5, 773 – 781.
4. C.M. Webb (1993). Dynamic MRTD for staring Focal Plane Arrays. IRIS Passive Sensors Vol II.
5. P. Bijl and J.M. Valeton, “Bias-free procedure for the measurement of MRTD and MRC”, Optical Engineering 38, 10, 1735-1742.
6. Fairhurst M.F. & Lettington, A.H. (1998). Method of predicting the probability of human observers recognizing targets in simulated thermal images. Optical Engineering 37(3), 744-751.
7. Driggers, R.G., Vollmerhausen, R., Wittenstein, W., Bijl, P., Valeton, J.M. (2000). Infrared Imager Models for Undersampled Imaging Systems. *Proc. Fourth Joint International Military Sensing Symposium*, 45, 1, 335-246.
8. Webb, C.M. & Holst, G. (1992). Observer variables in MRTD. SPIE Proc. 1689, 356-367.
9. De Lange, D.J., Valeton, J.M. & Bijl, P. (2000). Automatic characterization of electro-optical sensors with image-processing, using the Triangle orientation Discrimination (TOD) method. *SPIE Proceedings, Vol. 3701*, 104-111.
10. Hogervorst, M.A., Bijl, P. & Valeton, J.M. (2001). Capturing the sampling effects: a TOD sensor performance model. SPIE Proc. 4372 (in press).
11. C.M. Webb, “MRTD, how far can we stretch it?” In: Infrared imaging systems: design, analysis, modeling, and testing V, SPIE Proc. 2224, 297-307 (1994).
12. Bijl, P., Valeton, J.M. (2000) & de Jong, A.N. TOD predicts target acquisition performance for staring and scanning thermal imagers, *SPIE Proc.* 4030, 96-103.
13. Bijl, P.& Valeton, J.M. (1998c). Validation of the new TOD method and ACQUIRE model predictions using observer performance data for ship targets. *Optical Engineering 37, 7,* 1984 - 1994.
14. Bijl, P. (2000) Validation of the TOD and MTDP Sensor Performance Measures for staring and scanning thermal imagers. (Report TNO-TM-01-A020). Soesterberg, The Netherlands: TNO Human Factors.
15. Bijl, P.& Valeton, J.M. (1999). Guidelines for accurate TOD measurement, *SPIE Proceedings, Vol. 3701*, 14-25.
16. Koenderink, J.J. and van Doorn, A.J. (1978). Visual detection of spatial contrast; Influence of location in the visual field, target extent and illuminance level. Biol. Cybern. 30, 157-167.