

TOD versus MRT when evaluating thermal imagers that exhibit dynamic performance

Joseph Kostrzewa, John Long, John Graff, John David Vincent
Indigo System Corporation, 50 Castilian Drive, Goleta, CA 93117

ABSTRACT

While it is universally recognized that image quality of a thermal sensor is a strong function of spatial uniformity, the metrics commonly used to assess performance do not adequately measure the effectiveness of non-uniformity correction (NUC). Image uniformity is generally not static, particularly if correction terms are updated intermittently (with periodic shuttering) or gradually (with scene-based NUC). Minimum Resolvable Temperature (MRT), the most prevalent test for characterizing overall imaging performance, is poorly suited for characterizing dynamic performance. The Triangle Orientation Discrimination (TOD) metric proposed by Bijl and Valetton, because of its short observation window, provides better capability for evaluating sensors that exhibit non-negligible uniformity drift. This paper compares the effectiveness of MRT and TOD for measuring dynamic performance. TOD measurements of a shutter-based thermal imager are provided immediately after shutter correction and 3 minutes later. The drift in TOD performance shows excellent correlation to drift in system noise.

1 BACKGROUND

Just five years ago, achieving noise equivalent temperature difference (NE Δ T) of 100 mK with an uncooled detector was considered a daunting challenge; today, several uncooled cameras tout sensitivity below 25 mK. Furthermore, pixel sizes have been continuously shrinking while focal plane array (FPA) formats have grown larger. As a result of the rapid progress, the uncooled infrared community places considerable attention on these three attributes – temporal NE Δ T, pixel size, and array format. Indeed, these parameters have emerged as *de facto* criteria by which uncooled systems are compared and judged. Unfortunately, these criteria completely neglect spatial uniformity (also called spatial noise) as a component of image quality. Consequently, the importance of effective NUC is often disregarded despite experimental evidence that spatial noise is in fact more detrimental to overall performance than temporal.¹

A more complete method of evaluating a Forward Looking Infrared (FLIR) system is MRT, which is commonly held as the best measure of overall imaging performance. One of the positive attributes of MRT is that it includes spatial noise in the assessment of performance. However, this advantage is undermined by the fact that no standard methods have been defined for adequately representing true NUC effectiveness when measuring MRT. For example, there are no definitive guidelines prescribing how often to update NUC terms when evaluating a system that periodically employs a shutter-based correction or how to handle performance drift between updates. Furthermore, MRT test conditions are poorly suited for assessing systems that use scene-based NUC, and there are no defined procedures for resolving this issue either. So while MRT does not completely disregard spatial noise, it does not always capture true performance in real-world conditions. Therefore, results can be very deceiving.

Significant resources are being directed by the infrared community towards the improvement of uncooled sensors; however, efforts to improve image uniformity are lagging. Considering that the prevalent performance metrics either downplay or completely ignore NUC effectiveness, the lack of emphasis on spatial-noise reduction is hardly surprising. There is a clear need for more accurate reporting of true performance under typical NUC conditions. The body of work described in this paper suggests that TOD is a more effective tool than MRT for realizing this goal.

2 MRT AND TOD EXPLAINED

2.1 MRT

MRT has evolved as the most prevalent metric in the infrared community for assessing overall imaging performance of FLIR systems. The outcome of the test is a curve of minimum thermal contrast versus target size, as depicted in Figure 1. Each data point in the curve is obtained by finding the minimum contrast required for a human observer to resolve a standard four-bar target pattern of a given size. This determination is usually made by slowly increasing the temperature difference of the target (relative to a uniform background) until the observer makes a subjective judgment that the contrast is sufficient to discern all four bars. This step is usually repeated for both target polarities (i.e., bars hotter than background and bars colder than background), and it is often repeated for both target orientations (i.e., horizontal bars and vertical bars). Ideally, the entire test is repeated for multiple observers.

The fundamental value of the MRT measurement is that it incorporates the effects of sensitivity (temporal and spatial), resolution, and the human observer into the assessment of system performance. Although the measurement is time-consuming, it can be performed in a controlled laboratory environment in a fraction of the time and expense of full-blown field testing. Moreover, MRT curves can be used to predict target-acquisition range for various observation tasks, such as recognition, classification, and identification, using standardized Johnson cycle criteria. Because it is the basis for range predictions, MRT is often the fundamental imaging requirement that is specified for a FLIR system.

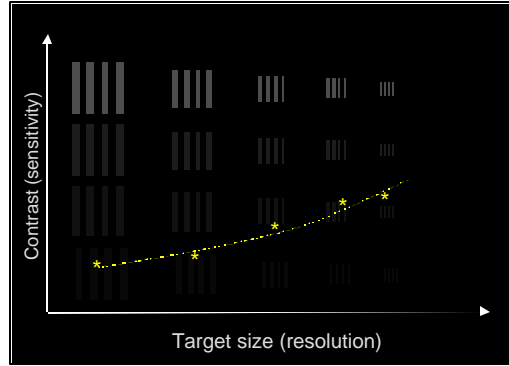


Figure 1: Typical MRT curve.

Despite its deep entrenchment in the infrared community, numerous shortcomings of MRT have been cited repeatedly.^{2,3,4,5,6,7} Two main focal points of criticism are the subjective nature of the test and the periodic test pattern, which exacerbates aliasing phenomena when evaluating under-sampled imagers. However, another weakness of MRT testing that has been ignored thus far in industry literature is the implicit assumption that imaging performance is static over the course of the test. For sensors that use a shutter or scene-based NUC to update correction terms, performance may in fact vary with time. Since the time required to measure a single data point of the MRT curve can span several minutes, performance drift over the course of the measurement can produce variability and errors.

Figure 2 illustrates the measurement problem caused by gradual drift of image uniformity. As spatial noise degrades, the four-bar pattern becomes more difficult to discern, which translates to a true degradation in MRT performance (the solid curve in Figure 2). If the blackbody used to generate target thermal contrast is slewed at rate R_1 (the dotted line), the bars are resolved with a thermal contrast of ΔT_1 . However, if the blackbody is slewed instead at a slower rate R_2 (the dashed line), then the bars are not resolvable until thermal contrast reaches ΔT_2 . The difference between ΔT_1 and ΔT_2 represents a variation in MRT measurement caused by a simple change in blackbody slew rate. While this situation is clearly undesirable, it is difficult to avoid when the performance of the system-under-test is changing with time.

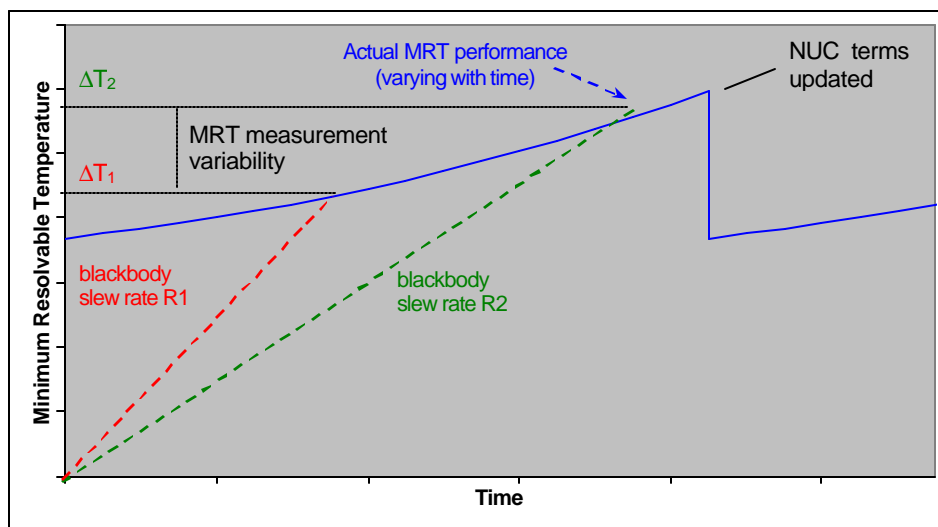


Figure 2. Variation in measured MRT value when true performance is not static.

Another large weakness of the MRT test that has not been sufficiently addressed by the infrared imaging community is the difficulty of measuring systems that employ scene-based NUC. The MRT observation is generally performed with a static target pattern against a uniform background that subtends most of the image plane. These are worst-case conditions for most scene-based NUC algorithms, which usually require some degree of scene motion. Webb and Halford have proposed a dynamic MRT test with a moving test pattern⁵, and this variant of the standard test is likely to be more favorable for scene-based NUC. However, the fundamental issue when measuring MRT on a system employing scene-based NUC is that most algorithms are optimized for the expected scene conditions the system will see in its real-world application. Therefore, even a moving test pattern will not necessarily yield true performance of a scene-based-NUC algorithm. Unfortunately, there are no provisions or guidelines in the MRT test procedure or in industry literature for solving this problem.

2.2 TOD

Recently, Bijl and Valetton proposed an alternative to MRT called the Triangle Orientation Discrimination (TOD) threshold.⁷ Fundamentally, TOD is similar to MRT in that the result is a curve of minimum contrast versus target size. However, there are several key differences between the two metrics. First, the four-bar target of the MRT test is replaced with an equilateral triangle. Instead of slewing thermal contrast gradually, the test pattern is presented at fixed contrast in one of four possible orientations – pointing left, right, up, or down. The observers' task is to discern the orientation, and an answer must be provided to each observation, even if it is merely a guess. The target is presented several times at the same contrast, each time with random orientation. After many iterations at a given contrast value, the frequency of correct responses typically ranges from 25% (complete guess) to 100% (complete certainty). Figure 3 shows a representative curve of percent-correct after a large number of iterations for six different contrast values. After fitting a curve to the data points, the contrast corresponding to a 75% correct-answer rate is selected as the threshold value. This process is repeated for multiple target sizes to generate a curve of contrast threshold versus target size. As exemplified in Figure 4, the TOD curve is fundamentally similar to that produced by MRT.

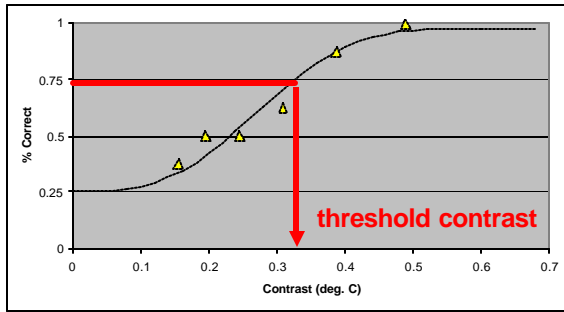


Figure 3. Typical TOD curve of correct-answer percentage versus target contrast.

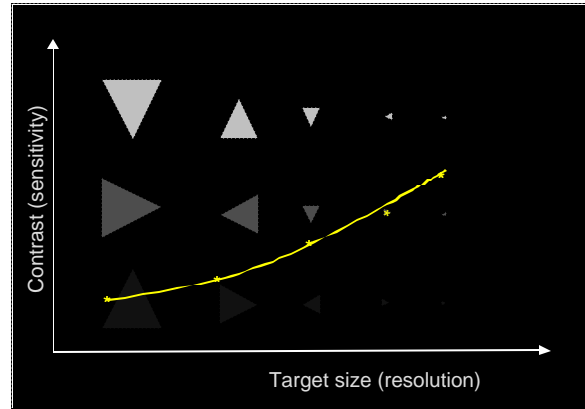


Figure 4: Typical TOD curve.

One of the stated advantages of TOD over MRT is the choice of test pattern – the triangular pattern is not periodic and therefore does not alias when the sensor-under-test is undersampled. Furthermore, the observers’ task is simple and objective, with results that are statistical in nature. In comparison, MRT is based on highly subjective criteria. Finally, TOD can also be used as the basis for range predictions, and in limited trials, has actually produced better correlation to target-acquisition range than bar-target testing.^{8,9}

An important distinction between MRT and TOD is the duration of the observers’ task. When measuring MRT, the target contrast is gradually adjusted as the observer stares at the imagery. The interval between the start of the observation and the declaration of a resolved target can span several minutes. At any time during this interval, the observer is permitted to modify sensor settings such as display brightness and contrast. For TOD measurement, the target is presented at a single fixed the rmal contrast, and the observer is not permitted to adjust sensor settings during the measurement.¹⁰ Consequently, the observation time is comparatively short, typically less than 5 seconds, and any drift in sensor performance during observation is likely to be negligible. An analogy can be made between observation window and exposure time of a camera. TOD allows measurement of performance “snapshots” whereas MRT will “blur” any drift in performance. By providing more temporal resolution to the measurement, TOD is potentially a better alternative to MRT when evaluating systems that exhibit dynamic performance.

3 METRIC ASSESSMENT

To evaluate the effectiveness of MRT and TOD as performance metrics, the following assessment criteria must be considered:

Relationship to target-acquisition range. The ultimate measure of system effectiveness for many FLIR systems is standoff range at which imaging tasks (e.g., detection, classification, recognition, identification) can be performed reliably. The ideal metric will serve as a basis for accurate range predictions.

Capability to measure dynamic performance. Imaging performance of a FLIR is generally not static but rather can be a function of many variables including irradiance, ambient temperature, and time. The ideal metric is flexible enough to characterize performance over a range of operating conditions, not merely at a single operating point.

Compatibility with scene-based NUC. Effectiveness of scene-based NUC is highly dependent on scene dynamics. The ideal metric adequately captures the performance of scene-based NUC algorithms in their true operating conditions.

Repeatability. The ideal metric shows little or no variation in results if the test is repeated for the same set of operating conditions or even if repeated using different observers.

Resources required to perform the measurement. The ideal metric will not require prohibitively expensive test equipment and can be measured in a reasonable amount of time.

Table 1 provides a qualitative summary rating of MRT versus TOD against these criteria. The text following the table provides a more detailed assessment.

Table 1: Evaluation of existing metrics against the assessment criteria.

	MRT	TOD
Relationship to standoff range	Good. Correlation to range performance is well established using Johnson criteria.	Potentially Good Correlation to range performance is excellent but based on only a few sets of experimental data.
Capability to measure dynamic performance	Poor. The long observation window makes it difficult to quantitatively measure performance drift versus time.	Potentially Good Dynamic performance might be measured by repeating the test at different “snapshots” in time.
Compatibility with scene-based NUC	Poor. Test conditions are not well-suited for scene-based NUC. Performance drift during measurement is an issue if the algorithm is exercised prior to measurement.	Potentially Good Test conditions are not well-suited for scene-based NUC, but completing observation immediately after exercising the algorithm in typical conditions might facilitate scene-based NUC assessment.
Repeatability	Poor. Subjective criteria; too many extraneous variables (e.g. target phasing, observer experience, drift, etc.).	Potentially Good The test is objective and statistical. Variability might be reduced by more iterations.
Resources required	Fair. Time-consuming to measure.	Fair. Time-consuming to measure with considerable post-test analysis.

MRT: Focusing first on MRT, several conclusions can be drawn from examination of Table 1. One of its major benefits is that there are standard, validated criteria for converting the results into range predictions. However, there are several significant shortcomings of MRT. First, it has already been noted that the performance of the system-under-test is liable to drift significantly during the period of time it takes to complete the MRT observation. This problem leads to measurement variability and does not support measurement of performance drift over short periods of time. It also has been noted that MRT test conditions are not compatible with most scene-based NUC algorithms. Finally, there tend to be large variations in results even when MRT is repeated with the same set of observers and especially when repeated with a different observer set. All of these disadvantages suggest that MRT, despite its deep entrenchment in the infrared imaging community, is far from an ideal metric for evaluating system effectiveness.

TOD: TOD has significant potential as a performance metric. One of the big advantages of TOD is the short observation window – each judgment call is made within seconds. The short observation window makes it possible to measure TOD at different snapshots in time (e.g. immediately after shutter correction and again some later time), a strategy that allows performance drift to be quantitatively characterized. The short observation window also makes it feasible to assess the performance of scene-based NUC algorithms because measurement can be performed immediately after exercising the algorithm. For example, the system-under-test can be exposed to “real-world” imagery before quickly inserting the test target into the field of view for a snapshot measurement. Yet another advantage of TOD is that it is objective and statistical in nature, whereas MRT relies on subjective criteria. Consequently, it is hypothesized that TOD is more prone to repeatable results. The primary drawback of TOD as a performance metric is that it requires many observations and iterations followed by significant post-test data reduction. Generating a TOD curve is time-consuming for even a single set of measurement conditions, and repeating for a multitude of variables (e.g. versus time, irradiance, scene temperature) would be very time-intensive.

Rather than repeating TOD for a large number of conditions, a more practical approach to system assessment is to correlate measurements of system noise over a variety of conditions to measurements of TOD at a subset of those conditions. Figure 5 exemplifies this strategy. The first step is to measure noise versus one of the system variables that affect spatial noise, such as elapsed time since update of NUC terms as shown in Figure 5a. The next step is to measure TOD for two or more known values of system noise, as shown in Figure 5b. The TOD curves are then used to predict range performance using standard load-line analysis. As shown in Figure 5c, the degradation in range resulting from uniformity degradation can thus be quantitatively assessed. The premise behind this strategy is that a correlation between total perceived noise and TOD exists and can be measured.

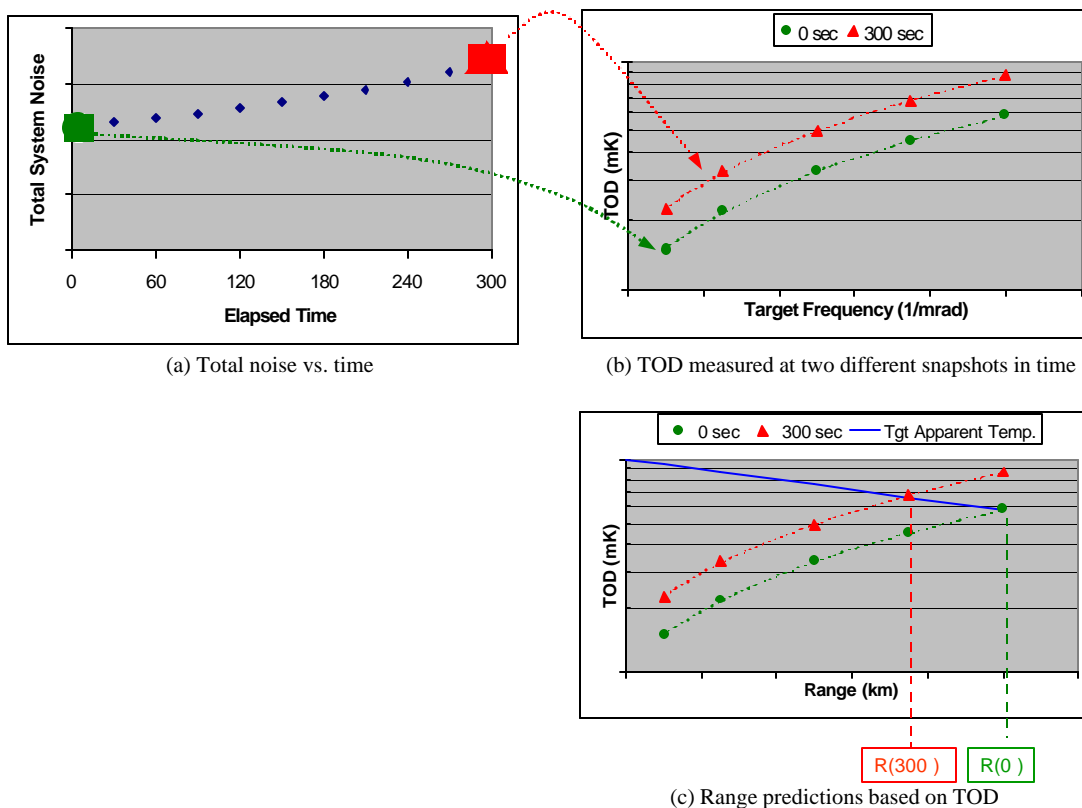


Figure 5: Example curves of TOD for various value of total perceived noise.

4 EXPERIMENTAL VALIDATION

4.1 Procedure and Results

The primary goal of the series of experiments described in this section was to validate the combined use of system noise and TOD as a means of characterizing dynamic sensor performance. A secondary goal of the experiments described herein was to compare variability in TOD and MRT measurements. The experiment summarized in the following paragraphs was designed around these two goals.

4.1.1 Measure total system noise versus time.

The sensor selected for evaluation was the Indigo Omega™ uncooled microbolometer camera. This camera uses an internal shutter for periodic NUC update and generally shows non-static uniformity over short periods of time. Temporal and spatial noise for this camera were measured while imaging a 20 °C blackbody at room ambient temperature. Three minutes after shutter correction, spatial noise was observed to be 25% greater than the value measured immediately after shutter correction. Conversely, the temporal noise measurement did not vary appreciably with elapsed time.

Because the human eye/brain temporally integrates multiple frames of data from a video sequence, the perception of temporal noise is reduced by the square root of the number of integrated frames. Spatial noise on the other hand is unaffected by temporal frame integration. When random spatio-temporal noise (σ_{tvh}) and random spatial noise (σ_{vh}) are the dominant noise components (which is generally the case for staring infrared FPAs), then total perceived noise can be expressed mathematically as

$$\Omega_{perceived} \propto \sqrt{\frac{\mathbf{s}_{tvh}^2}{F_R \cdot t_{eye}} + \mathbf{s}_{vh}^2}, \quad (1)$$

where F_R is the frame rate and t_{eye} is the eye integration time. The generally accepted value of t_{eye} is 100 msec; and thus the perception of temporal noise in 30 Hz imagery is a factor of $\sqrt{3}$ less than the measured value. Figure 6 shows normalized total perceived noise as a function of elapsed time for the system-under-test. At $t = 3$ minutes, the total noise is 11% higher than the value at $t = 0$ minutes.

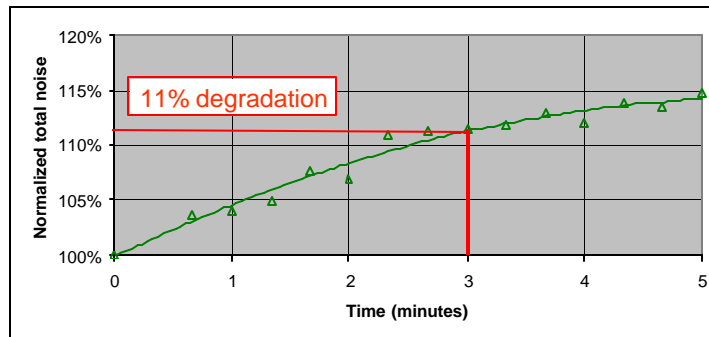


Figure 6: Total noise versus elapsed time since shutter correction for the evaluated camera.

4.1.2 Acquire TOD test sequences.

The guidelines for measuring TOD prescribe viewing five or more target sizes, with five or more contrast values per target size. For each contrast value, at least 16 observations of the target are recommended. Assuming three observers, a rigorous TOD curve therefore requires a minimum of 1200 total observations ($5 \times 5 \times 16 \times 3$). To reduce test time, two modifications were made to the standard TOD test. First, the scope was limited to 2 target sizes rather than 5. The first target was selected to measure performance near the theoretical sampling limit of the evaluated camera ($0.25 \text{ mrad}^{-1} = 85\%$ of f_{max}), and the second target was selected to measure performance at a lower spatial frequency ($0.10 \text{ mrad}^{-1} = 35\%$ of f_{max}). A second modification of the standard TOD test designed to reduce test time was the use pre-recorded digital sequences.

One benefit of using pre-recorded sequences is that all observers witness the same imagery in parallel. This is not only a time-saving measure but also a mechanism for eliminating variations in results due strictly to variations in observation conditions. Another considerable benefit of pre-recorded sequences is that they make most efficient use of the observers' time. Observers are not required to wait while the experimenter rotates and/or changes target plates or while the sensor is reaching the "snapshot" of time at which it is to be evaluated. With pre-recorded sequences, the observers that participated in this experiment could perform all observations of a single target plate (160 total sequences) in less than two hours. Without pre-recording, the same test would easily have taken 5 to 10 times longer. The primary disadvantage of using digital sequences is that the system is not evaluated exactly as used in the field – after video processing, digital-to-analog conversion, and display on an analog monitor. Since the goal of this test was not to make absolute measurements of the system-under-test but rather to identify relative changes in performance, this disadvantage was not an issue.

For acquisition of TOD sequences, the triangular target was manually rotated to the desired orientation (up, down, left, or right), and the blackbody temperature was set to the desired contrast value. Then camera phasing relative to the target was varied randomly, and the camera was commanded to perform shutter correction. Immediately after shutter correction, a 150-frame sequence was acquired, then a second 150-frame sequence was acquired three minutes later. For each target contrast, this procedure was repeated a total of 16 times – four times for each of the four target orientations. Two of the sequences per target orientation were acquired at positive contrast (blackbody hotter than target plate) and two at negative target contrast (blackbody colder than target plate). 5 contrast values were measured for each target size, for a total of 320 sequences.

4.1.3 Acquire MRT test sequences.

One of the stated goals of the experiment was to compare repeatability of TOD and MRT measurements. However, it was recognized that using pre-recorded sequences to measure TOD would eliminate variations in observation conditions; therefore, to make a fair comparison between the two metrics, it was necessary to use pre-recorded digital sequences for measurement of MRT. The MRT sequences for each target size (0.1 cycles/mrad and 0.25 cycles/mrad) were acquired as follows:

1. Starting at target thermal contrast of 0 mK, a 100-frame sequence was acquired.
2. The target contrast was increased by 10 mK, and another 100-frame sequence was acquired.
3. Step 2 was repeated many times until target contrast reached a value deemed sufficient to easily resolve the four bars in the target.
4. All of the 100-frame sequences were stitched together to create a single, continuous "movie" of continuously increasing target contrast.
5. The entire process was repeated for negative contrast. In other words, starting at 0 mK contrast, target contrast was *decreased* 10 mK for each 100-frame sequence. Repeating the process resulted in a second "movie" with decreasing target contrast.

4.1.4 Observe TOD sequences

Four observers were selected for the experiment. All had previous experience making MRT measurements but no experience measuring TOD. The testing was performed in a darkened room, with the only significant ambient light coming from the video display. The larger of the two targets was observed on the first two days of the experiment, and the smaller was observed two weeks later. For each trial, every observer was seated comfortably around the display, and the experimenter started with the sequences recorded at $t = 0$ min. at the largest contrast value (i.e., the easiest target to see). After playback of each pre-recorded sequence, each observer marked a private scorecard indicating whether the target was deemed to be pointing up, down, left, or right. To maintain independent scoring, no conversation regarding target observation was permitted. After all 16 sequences for the largest contrast value were scored, the experimenter proceeded to the next highest contrast value and so forth down to the dimmest thermal contrast. The observers then turned in scorecards, and after a short break, repeated the entire procedure with the sequences recorded at $t = 3$ min. A second identical trial was repeated the following day using the same observers, sequences, and test conditions.

4.1.5 Observe MRT sequences

The guidelines for conducting MRT tests state that each observer is permitted to modify display settings (contrast and brightness) as well as the display polarity (white-hot or black-hot) at any point during target observation. Consequently, MRT observations could not be performed in parallel, but rather each observer had a turn alone in front of the display. The previously-described MRT movie recorded with positive thermal contrast was played back for the observer, who stopped playback when he deemed contrast sufficient to resolve all four bars of the target. The experimenter would then note the frame number and repeat the test with the negative-contrast sequence. One of the advantages of using pre-recorded sequences for MRT testing is that they allow the observer to overshoot the minimum resolvable contrast, quickly rewind, and thus fine-tune the judgment call by gradually narrowing in on it. One potential problem with using digital sequences is that the observer might recall the frame number at which he stopped playback on a previous trial and bias his current observation (unintentionally or otherwise) based on that memory. For example, if the observer called the contrast sufficient at frame number 1200 on iteration 1, he might be biased to find sufficient contrast at frame number 1200 on iteration 2. To avoid that potential bias, each time a sequence was played back, the experimenter would intentionally modify the correlation between frame number and actual target contrast by either removing frames at the start of the sequence (prior to playback) or adding additional “fill frames”. The observer was then informed that the frame number from a previous playback would not necessarily correspond to the same frame number for the current playback.

4.1.6 Analyze raw data.

After all of the observations had been completed, it was necessary to convert raw TOD observations into minimum threshold values. The prescribed mechanism for performing this conversion is to determine a threshold value for each observer independently, considering positive-contrast and negative-contrast as separate data sets and averaging the results. Minimum threshold contrast is determined by plotting the percentage of correct answers as a function of contrast value, mathematically fitting a Weibull function to the data, and interpolating the curve to find the contrast value at which 75% correct answers are obtained. If the Weibull curve fit fails a $\chi^2_{0.95}$ test or if the 75%-correct threshold value falls outside the range of observed contrast values, then the data is rejected*. Figure 7 shows analyzed TOD results for all four observers for both iterations of both target plates. Figure 8 shows MRT data.

* See Ref. 10 for detailed procedures and guidelines for analyzing raw TOD data.

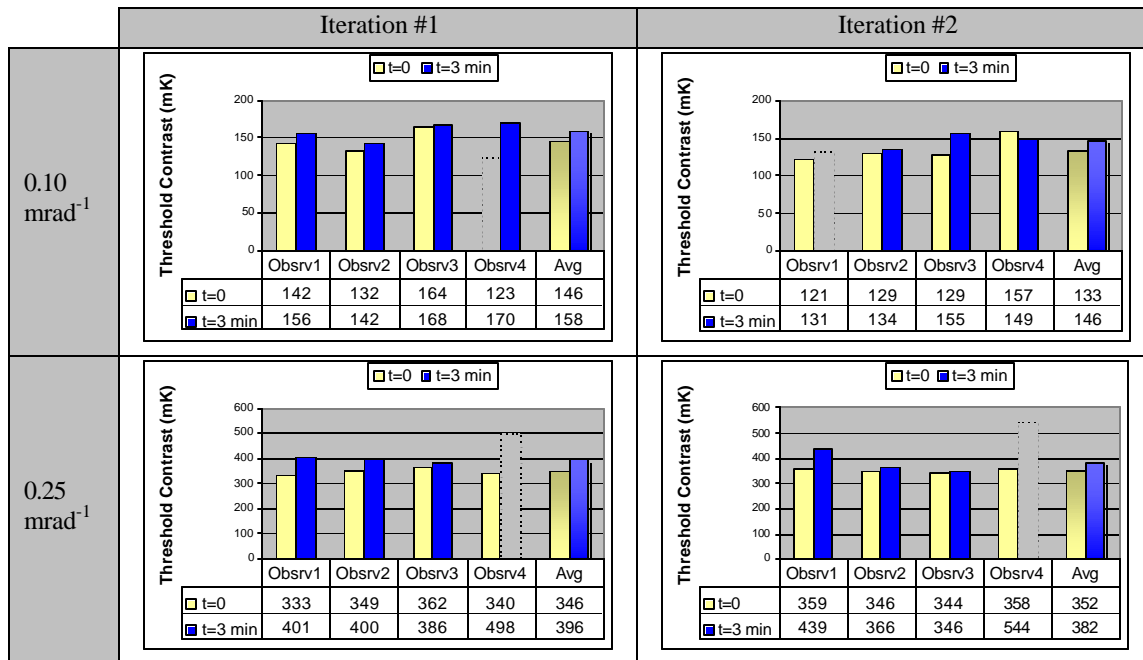


Figure 7: TOD threshold data.

(Note: data shown in gray and outlined with dashed lines are rejected for failing to meet the required curve-fit criteria and are not included in the average.)

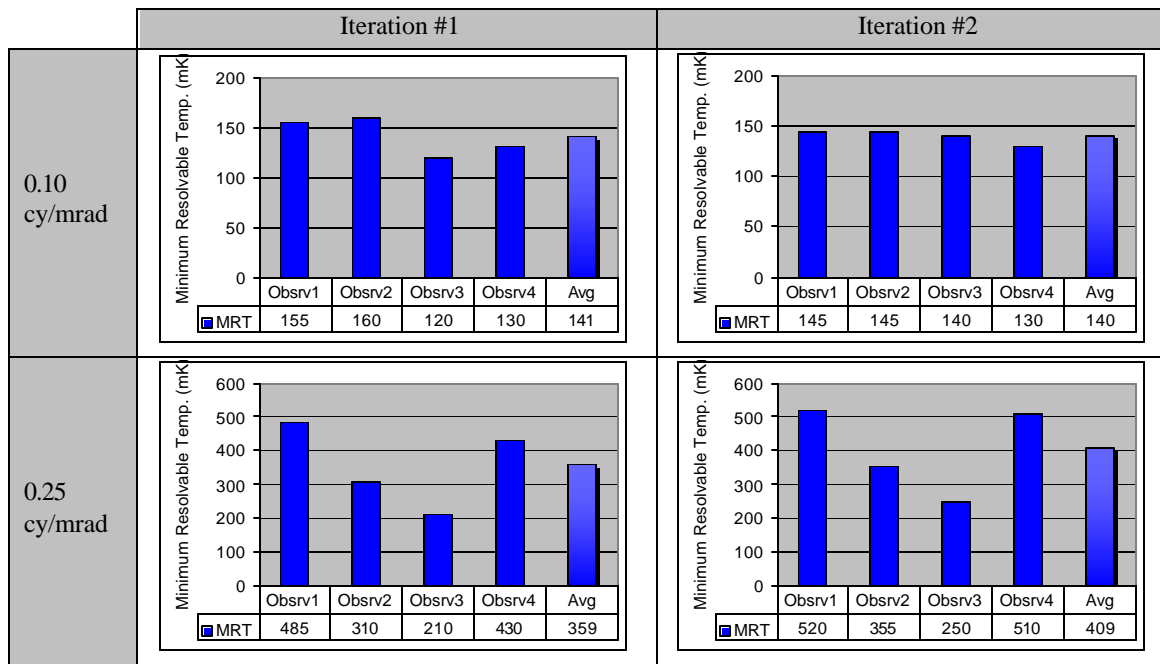


Figure 8: MRT data.

4.2 Discussion of Results

4.2.1 Correlation between Total perceived noise and TOD

One of the stated goals of the experiment was to determine a correlation between total perceived noise and TOD. Referring back to Figure 6, the degradation in total perceived noise from $t = 0$ min to $t = 3$ min is approximately 11% for this camera. As shown in Figure 9, the degradation in measured TOD for the higher spatial frequency (0.25 mrad^{-1}) is exactly 11%, and for the 0.10 mrad^{-1} target, the degradation is 9%. This data implies a linear proportionality between total noise and TOD. In other words, based on measurement of total noise as a function of elapsed time, the degradation of TOD performance at each point in time can be accurately predicted by simple linear scaling.

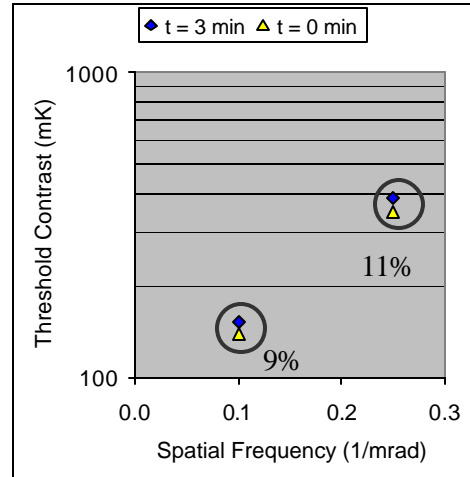


Figure 9: TOD measurements, averaged.

4.2.2 Observer-to-Observer Variation, TOD versus MRT

A second goal of the experiment described herein was to compare variation in TOD results to variation in MRT results. Figure 10 examines variation from observer-to-observer. Shown in the left-hand panel is the spread between minimum and maximum observer TOD threshold values, while the right-hand panel shows the same for MRT measurements. The uncertainty due to observer-to-observer variation represents 80% of the mean MRT value at the upper end of the MRT curve. In contrast, observer variation is an order of magnitude lower for TOD. This data corroborates the hypothesis that TOD results, because they do not rely on observers' subjective criteria, exhibit far greater uniformity from individual to individual. This suggests that selection of observers is not nearly as likely to affect TOD results as it might MRT.

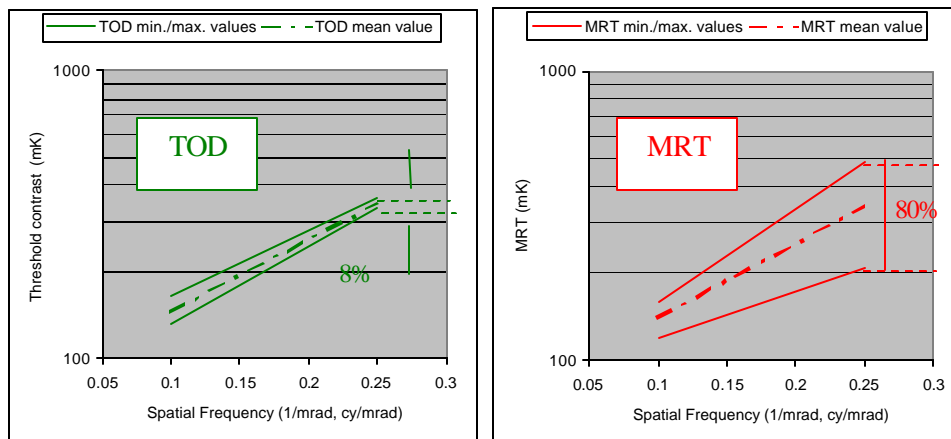


Figure 10: Observer-to-observer variation, TOD vs. MRT.

(Data shown is for iteration 1, TOD at $t=0$ min.)

4.2.3 Trial-to-Trial Variation, TOD versus MRT

Another measure of repeatability is the amount that results vary from one trial to the next. Figure 11 shows absolute difference between data obtained in trials 1 and 2. (Data is averaged across all observers.) MRT data shows slightly more variation from trial to trial than TOD at the higher spatial frequency, but less at the lower frequency. Note that using identical pre-recorded sequences in both trials almost certainly reduced trial-to-trial variation for both metrics, compared to the likely results had different sequences been used in successive trials. However, as illustrated previously in Figure 2, MRT is more prone to variations in results stemming from variations in test conditions. Consequently, the data shown in Figure 11 would likely show a more compelling case for TOD if the tests had been performed in a more standard manner, using live camera imagery rather than pre-recorded sequences.

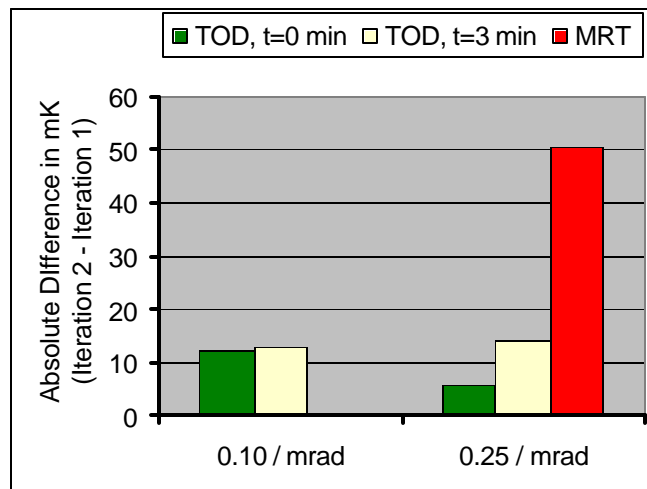


Figure 11: Trial-to-trial variation, TOD vs. MRT.

5 CONCLUSIONS

TOD has a number of notional advantages over MRT that have been described previously in industry literature:

1. non-periodic target – no aliasing
2. objective observer task
3. equal or better correlation to range performance (based on limited data)

In this paper, the following additional features of the TOD metric were identified as significant advantages:

1. short observation window – negligible performance drift during measurement
2. can be readily measured at multiple snapshots in time
3. compatible with scene based NUC – measurement can be made immediately after exercising the algorithm with real-world imagery
4. less variation in results, both from observer-to-observer and from trial-to-trial

The goal of the study described herein was to define and validate a metric for measuring dynamic performance of thermal imaging systems, particularly those that use a shutter for periodic update of NUC correction terms. A strategy for quantitatively characterizing performance drift with TOD was verified. System noise was measured as a function of elapsed time and then correlated to TOD measurements made at two different snapshots in time. The data suggests that a small but non-negligible degradation in total noise produces a comparable degradation in TOD. The data also validates the hypothesis that variations from observer-to-observer are far lower for TOD measurements than for MRT. These results suggest that TOD is a highly effective means of evaluating performance of modern thermal imagers. Because it is capable of accurately characterizing performance drift under dynamic operating conditions, TOD provides a more realistic measure than MRT of how a system will perform in the field, not just in a laboratory environment.

6 ACKNOWLEDGEMENTS

The authors wish to acknowledge the support of the U.S. Navy Space and Naval Warfare Systems Command (SPAWAR) and DARPA Microsystems Technology Office (MTO) for the body of work described in this paper. This paper is largely derived from a more extensive report generated under contract N66001-01-C-8054.

7 REFERENCES

-
- ¹ R. Driggers, R. Vollmerhausen, and K. Krapels, "Target identification performance as a function of temporal and fixed pattern noise," *Proc. SPIE* **4030**, 144-150 (2000).
 - ² J. Mooney, "On the future of MRT," *Proc. SPIE* **1969**, 177-183 (1993).
 - ³ P. Bijl and M. Valeton, "Bias-free procedure for the measurement of minimum resolvable temperature difference and minimum resolvable contrast," *Opt. Eng.* **38**(10), 1735-1742 (1999)
 - ⁴ W. Wittenstein, "Minimum temperature difference perceived – a new approach to assess undersampled thermal imagers," *Opt. Eng.* **38**(5), 773-781 (1999).
 - ⁵ C. Webb and C. Halford, "Dynamic minimum resolvable temperature testing for staring array imagers," *Opt. Eng.* **38**(5), 845-851 (1999).
 - ⁶ I. Bendall, "Automated objective minimum resolvable temperature difference," *Proc. SPIE* **4030**, 50-59 (2000).
 - ⁷ P. Bijl and J.M. Valeton, "Triangle orientation discrimination: the alternative to minimum resolvable temperature and minimum resolvable contrast," *Opt. Eng.* **37**(7), 1976-1983 (1998).
 - ⁸ P. Bijl, M. Valeton, A. Jong, "TOD predicts target acquisition performance for staring and scanning thermal imagers," *Proc. SPIE* **4030**, 96-103 (2000).
 - ⁹ P. Bijl, M. Valeton, J. Mathieu, "Validation of the new triangle orientation discrimination method and ACQUIRE model predictions using the observer performance data for ship targets," *Opt. Eng.* **38**(7), 1984-1994 (1998).
 - ¹⁰ P. Bijl and J.M. Valeton, "Guidelines for accurate TOD measurement," *Proc. SPIE* **3701**, 14-25, (1999).