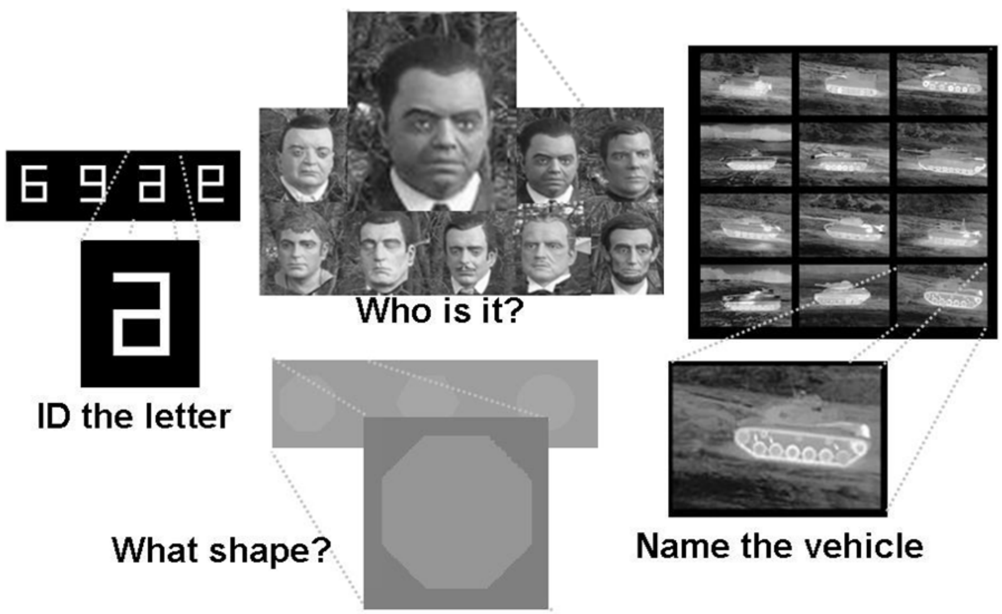
RHV Electro-Optics, Lake Mary, Florida

The Targeting Task Performance (TTP) Metric and Models

1. **Introduction:** This paper provides a summary of the TTP metric concept, what it predicts, and the experimental data that supports it. A performance model with a straightforward user interface will be available on this website.

   The models currently distributed by the Army do not use the validated TTP metric. The differences between the original TTP and the TTP in current Army models are discussed at the end of this paper.

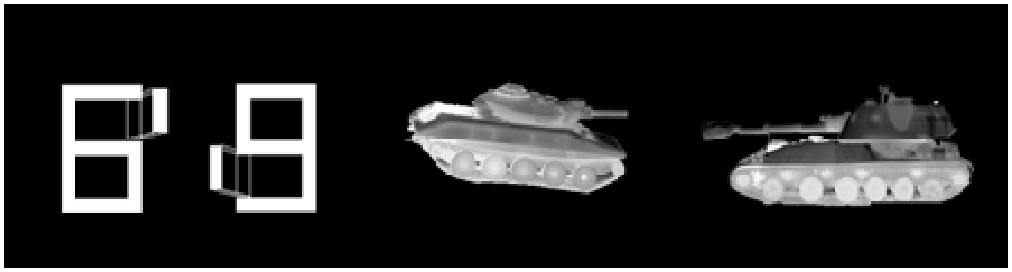
2. **What is the TTP metric?** TTP predicts the probability of identifying (PID) objects in a test set when viewed through a camera. The test set might be letters, faces, tactical vehicles, shapes, or other things. The observers are trained and know what is in the test set. The model predicts average PID for the set.



3. **Why a set and not a tank or a letter or a face?** One reason is that observer consistency is better with forced choice experiments. Also, we need multiple calls to get a good average, and continuously repeating the same image does not work with humans.

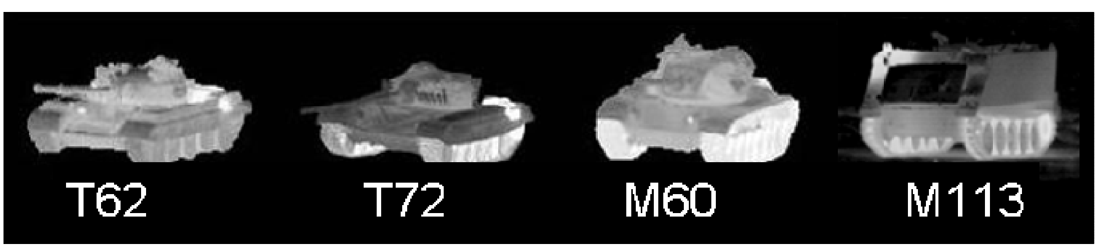
   Another reason for using a set is that task difficulty is based on a comparison. For example, if the visual task is “identify Richard,” does that mean discriminating me from my wife or me from my look-alike brother Bob? Those two visual discriminations are totally different, thank goodness. When using a model to quantify the probability of identifying an object, we must always be able to specify “from what?”

4. **Understanding Probability of Identification:** The set of characters in the first figure above share a probability curve, but the set of tracked tactical vehicles do not. The two target sets, characters and tracked tactical vehicles, represent two extremes. With characters we know exactly what the discrimination cue is, and it is always the same. As illustrated in the figure below, the difference between an “8” a “9” a “6” an “a” or an “e” is the location of the rectangle.



With tactical vehicles, each vehicle has different features at every vehicle aspect. For tactical vehicles, trained observers look at the location of engine heating, the number of road wheels, fender and hatch shape, the location and size of the gun barrel, and other cues. Discriminating between tactical vehicles involves seeing and comprehending a diversity of cues that vary in size and contrast. That is why a set of tactical vehicles makes a good eye chart when selecting a camera for future and unknown situations.

The next figure shows T62 and T72 Russian tanks, an American M60 tank, and an M113 Personnel Carrier. The Russian tanks look a lot alike and get confused at close range. The Russian tanks are the main reason that the PID versus range curves in this paper drop to 0.9 at close range. Discriminating the Russian from the American tanks reliably occurs at mid range. With most but not all imagers, the M113 is identified at four or five times the range of a T62/T72 discrimination.



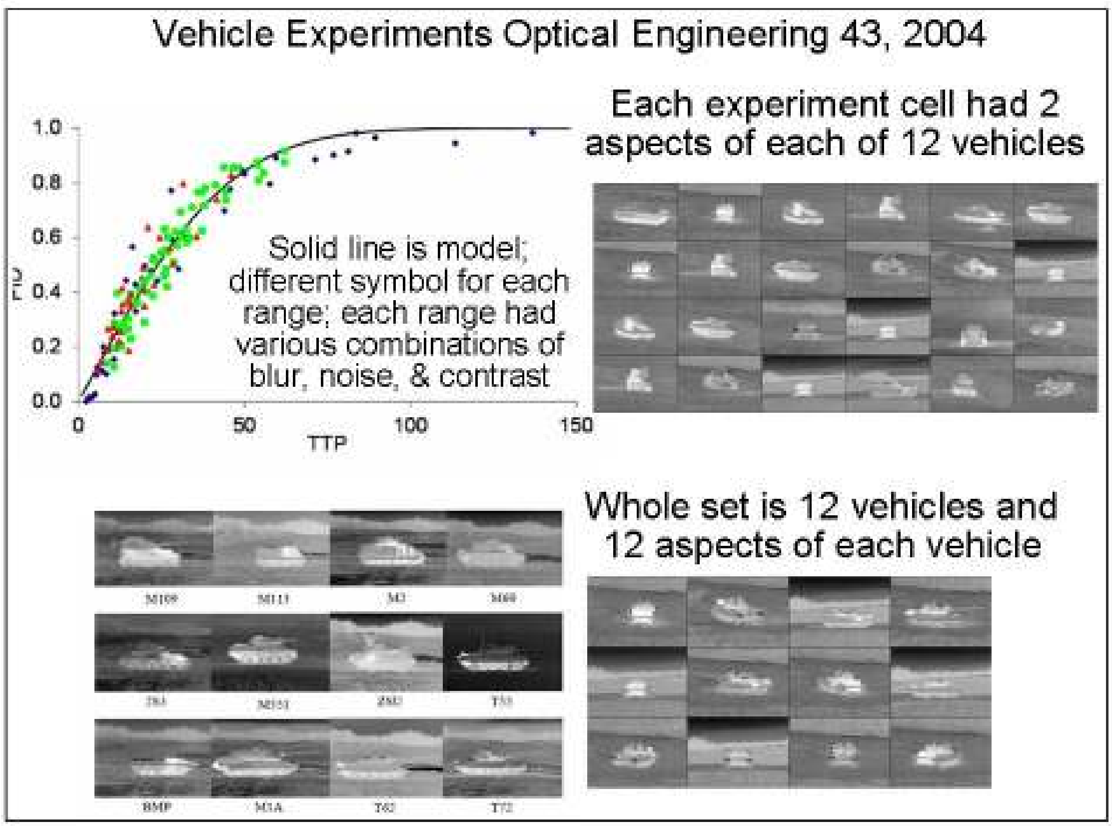
When the PID curve is corrected for chance, then the tail of the PID curve sorts out the bad apples, not the 0.9 PID range. The M113 and some aspects of other vehicles are reliably identifiable at very long range provided the imager is capable of displaying the target structure. If the tactical vehicle PID curve falls off rapidly at mid-range, that is an indication of a problem with the imager. Some of the vehicles in the set should be identified at long range.

5. **What is the function of the TTP metric and model?** Primarily, the TTP metric lets us compare the range performance of competiting camera designs. Some cameras are purchased for a specific purpose like facial identification or catching irregularities on an assembly line. Evaluating those cameras for a specific function

requires using a Specific Object Model, where the items to be discriminated are characterized by their spatial Fourier Transform.

Most cameras purchased today will be used by someone, somewhere, to do an as yet unspecified visual task at sometime in the future. To evaluate a camera design for its ability to render scene detail, we use a target set with diverse spatial cues. That is, a target set where the spatial discrimination features vary from small to large and low contrast to high contrast. The average size and contrast of the target set can be changed in order to evaluate the sensitivity and sampling characteristics of the imager. That is, we can reliably apply the empirical observer data to mathematically generated scale models of the target set.

6. **Does TTP work?** The original published model does. Much of the test data is summarized in the figures below and references are given. These figures do not apply to the recent models published by the Army such as NVIPM. The recent changes made by the Army to the TTP model are not supported by published experimental data.

7. 

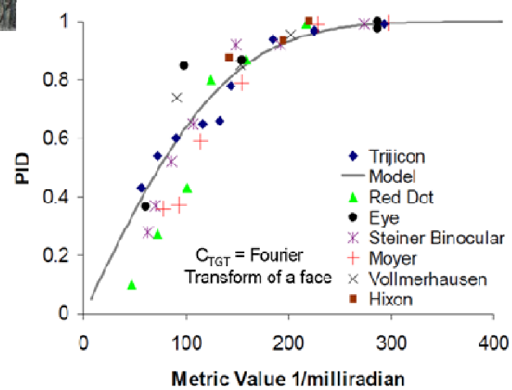
Applied Optics 47, 2008.

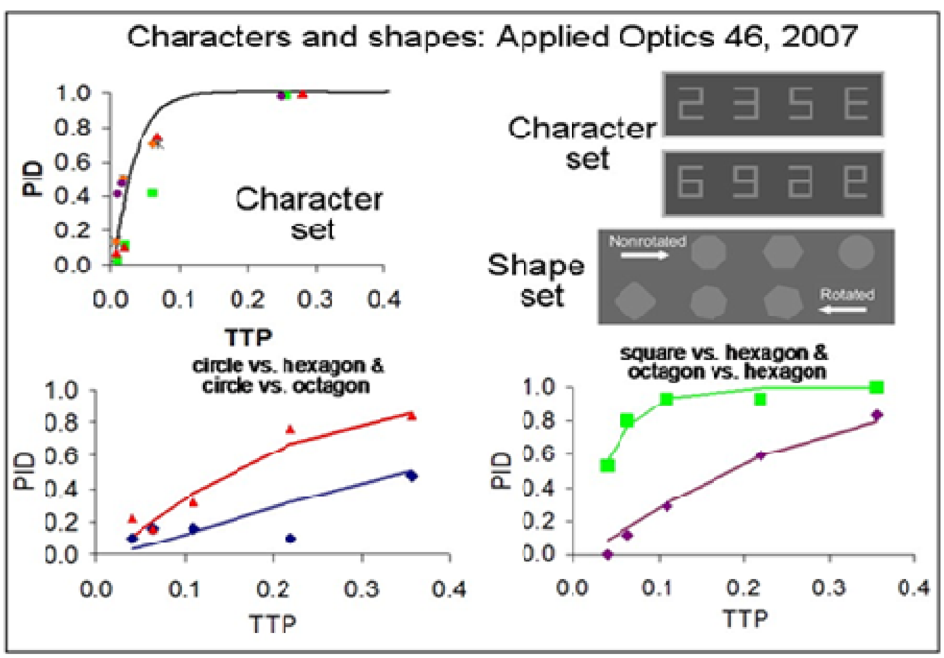
Top two rows show 12 faces

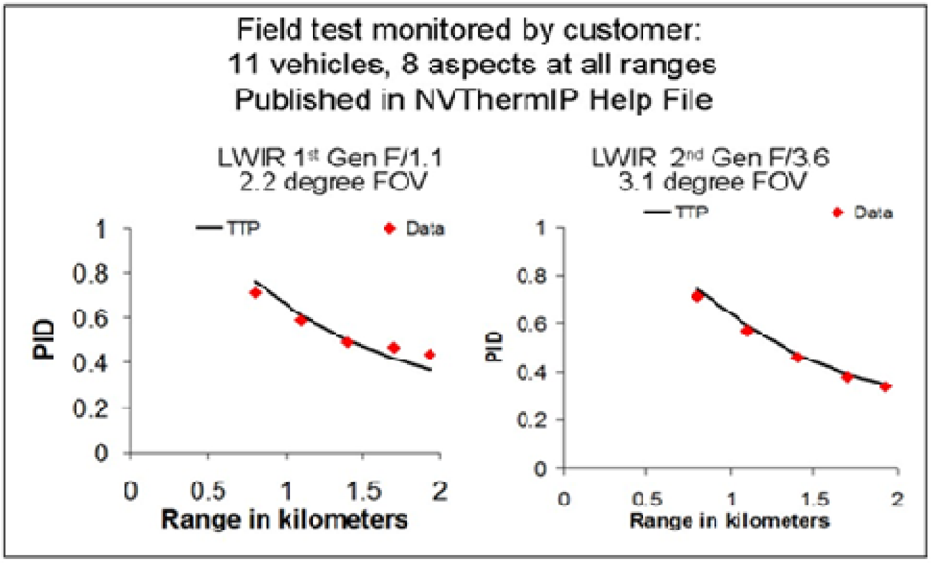
Bottom row shows 6 aspects

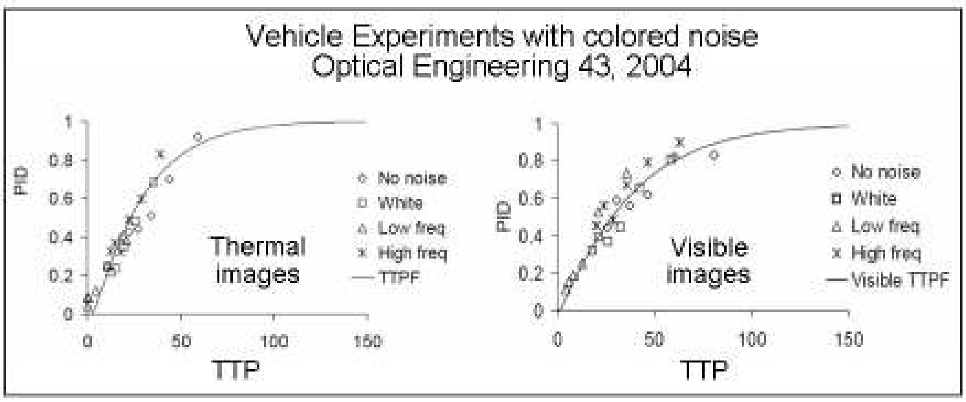
These two rows show wax replicas of ten famous faces

- **Naked eye data used famous faces in background shown.**
- **All other experiments used the actual faces with brick or plain wall background.**
- **Experiments were done with the naked eye, four spotting scopes, and two used digitized magery displayed on computer monitors**









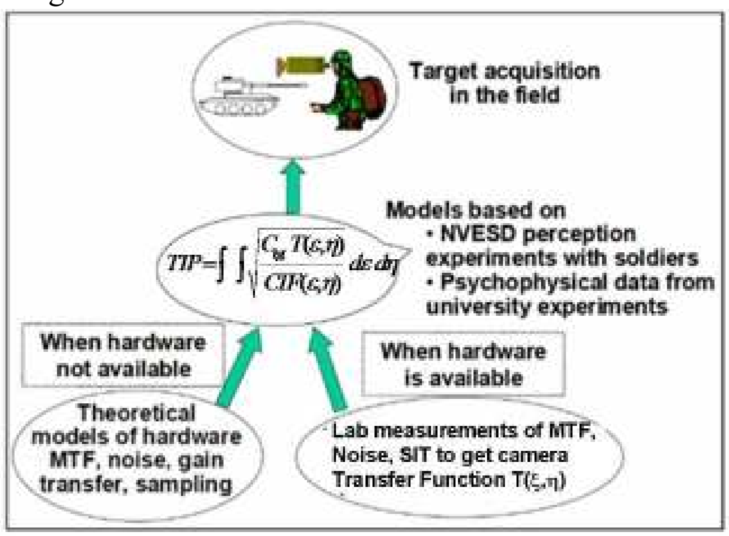
8. **Other than PID experiments, what factors support the validity of TPP as a visual task performance metric?**

Resolution metrics like TTP are calculated based upon factors like imager blur, noise, and sampling together with target contrast and structure. The equation that relates a resolution metric value to PID is called the Target Transform Probability Function (TTPF). Most researchers consider the TTPF to be an empirical fit to all available empirical data. It turns out that the TTP TTPF is the Error Function. TTP accurately predicts PID for four kinds of target sets and a wide variety of imager characteristics for each set using the error function to calculate PID from TTP value.

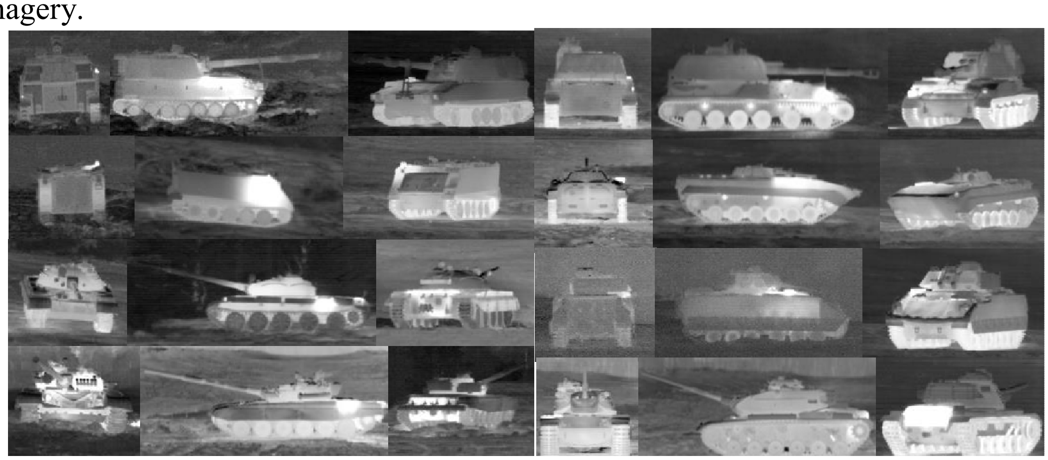
Second, TTP predictions are consistent with aviator surveys on the effect of environment on pilotage system performance. See the Proceedings of SPIE Volume 6207.

Third, it seems reasonable to assume that a diverse target set like the tactical vehicles with many shapes and spatial features would be represented in the model as a flat spatial frequency spectrum, and using a flat frequency spectrum for that set leads to accurate PID predictions. However, it also seems reasonable, a priori, that the spatial features that result in visually discriminating between look-alike objects must reside in their spatial frequency spectrums. Experimentally, the TTP model requires that the spatial frequency spectrum of look-alike objects be input in order to accurately predict PID using the Error Function. Far from being a problem with the model, the need to include spatial information for most target sets actually supports metric validity.

9. **Lab performance testing.** The concept published with NVThermIP is shown in the figure. We suggested that data from laboratory hardware measurements of MTF, NETD, and SIT be input into the model along with the specified C_{tgt} to see if the range specification was met. Our feeling was that the government was buying hardware, so do hardware testing.



10. **What field test corresponds to the TTP performance model?** The eight vehicle and three aspect target set was designed for field evaluations that would match TTP tactical vehicle predictions like those generated by NVThermIP. Digital imagery was to be collected for the entire set at five or six ranges. Trained observers with applicable military specialties (like tank gunners and photo interpreters) would ID the imagery.



Problems with current Army models (i.e., NVIPM):

Disagreement 1: Target angle has replaced display angle when calculating spatial frequency contrast thresholds.

The change is justified in NVIPM documentation by the fact that facial ID is not accurately predicted by the TTP when a spectrally flat frequency spectrum represents a group of faces. (Facial ID is not mentioned in the justification, but the few data points presented are from a facial ID experiment.)

There are two fundamental points of disagreement. First, the a priori belief that size alone should allow us to predict the probability of discriminating between objects, like faces, that share spatial features is not supported by experimental data on face or shape discriminations. The model calculates imager resolution based on the frequency transfer characteristics of the imager. When the discrimination cues lie within a limited spatial frequency spectrum, it is reasonable to expect that it is the resolution of those particular spatial frequencies that allows us to discriminate between objects. That conclusion is supported by experimental results.

It is hard to understand the contrary argument, that the spatial cues that allow us to discriminate between look-alike objects do not lie within the spatial frequency spectrum of the set. In that case, what are we looking at?

It seems that the two types of PID are being confused. The distinction between identifying one member of a like set (a face, a character) is being equated to discriminating between disparate objects (a tank or a self propelled howitzer). In the case of tactical vehicles, PID is the fraction of vehicles correctly identified based on a variety of spatial features, large and small, associated with each vehicle. With tactical vehicles, the observers are trained to look for engine heat location, the size of the gun, the shape of the hatch, the number of road wheels, fender shape, and other specific features of each target. In the case of faces, the observer must distinguish based on the shape of the face, eye separation, and other features that all humans share. The two visual tasks are not the same.

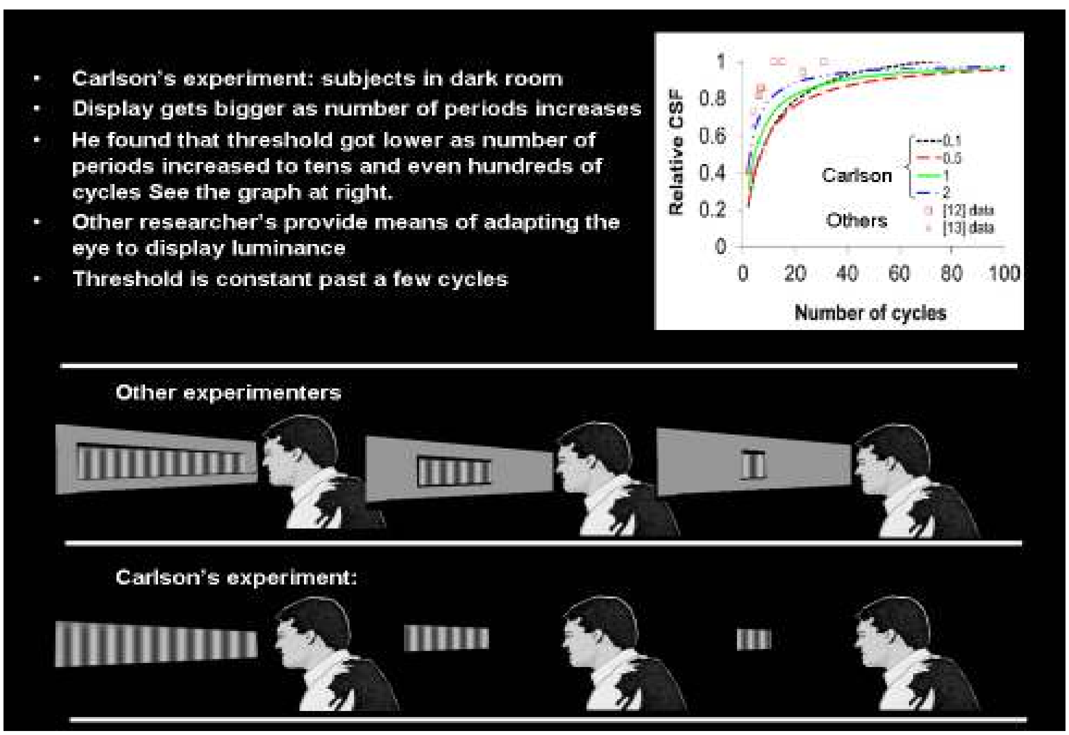
Further, the recent model change is based on a misunderstanding of Barten's numerical Contrast Threshold Function (CTF) approximation. The Barten numerical CTF was originally selected because a professor at (previously) the Virginia Polytechnic Institute found Barten's fit to CTF data to be as accurate as any then available and easy to use. When the adapting luminance angle at the eye is kept to a few degrees minimum, the CTF approximations are representative of people with very good eyesight.

However, for his own stated reasons, Barten based his small display angle CTF on the data collected by Carlson. The figure below illustrates the difference between Carlson's experimental approach and that conducted by other researchers. This is not a matter of questioning Carlson's approach; the discussion here is about whether the particular CTF data should be used in the manner currently adopted by the Army.

As illustrated by the top row of observers, most CTF data is collected by keeping the observer's adapting luminance the same as the luminance of the sine wave pattern. The sine wave pattern shrinks, but sufficient area of display remains illuminated to keep the eye adapted to the display luminance. In the Carlson experiment, however, the illuminated part of the display shrank along with the sine wave pattern.

The experiments were performed to find the increase in contrast threshold as the number of cycles in a sine wave pattern decreases. Since the TTP metric is calculated in the frequency domain, we are interested in CTF measured with many cycles. We share that interest with most of the experimenters; they are trying to establish how many cycles must be presented to the observer to get the many-cycle CTF. They, like us, want a threshold that corresponds to a spatial frequency. They, like us, want to ensure that the eye is adapted to the luminance level associated with the CTF data.

How might Carlson's data be used? Many people like watching television in a dark room. For many years, the display area on televisions was rather small, so the viewer's sensed picture might degrade because his eye adapts to the average room luminance and not the video picture. How much does the sensed picture degrade, and what can the television manufacturer do to improve the viewer's experience? Those are reasonable questions, and Carlson's data can help answer them.

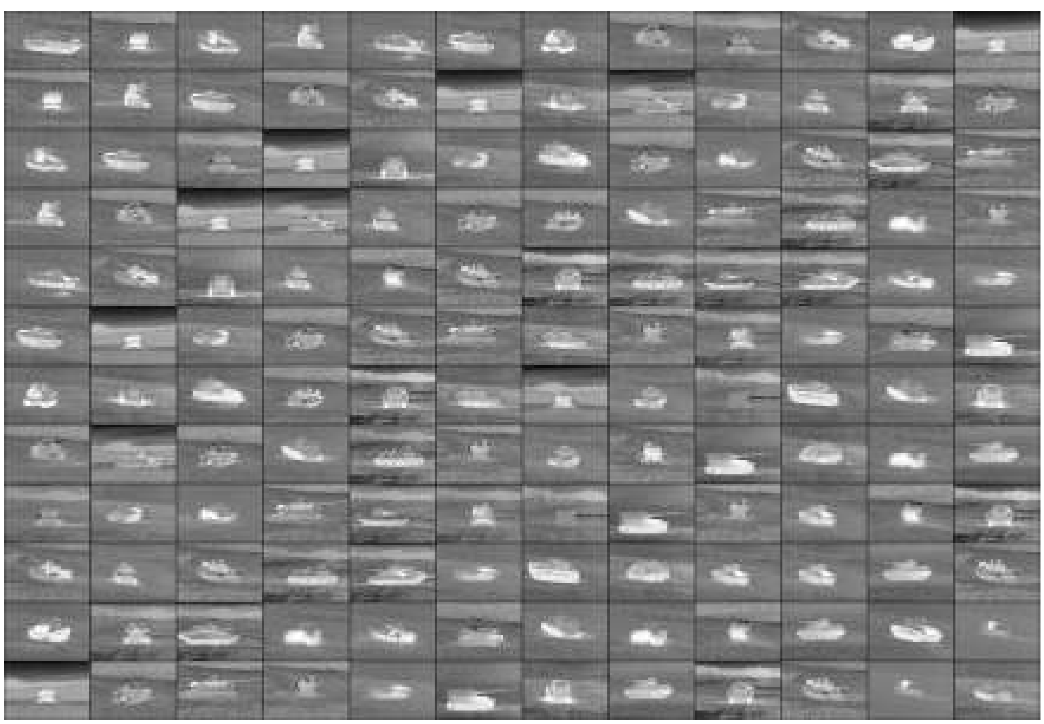


On the other hand, TTP uses CTF to represent the observer's sine wave threshold in the frequency domain, and using a CTF based on a few sine wave cycles is not correct, Further, the target in NVIPM is now viewed against a black background, because the display goes black everywhere except the target location. That is not correct either.

Disagreement 2: The current model development personnel consider the vehicle set used in NVThermIP validation to be too large and too hard on observers; recent experiments use the field test set instead. It is true that the original ID experiments were tough on observers. Identifying targets at long range is difficult, and that is why the Army provided training. The test participants were all active military with appropriate MOS, and the experiments were approved by their command as part of the soldier's formal training.

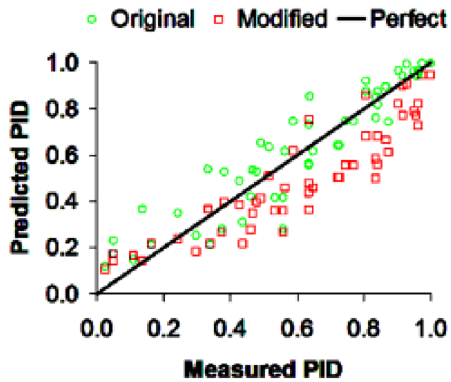
The training required to reliably interpret imagery taken at long range is time consuming and difficult, and that limits the number of suitable observers. We understand the problem posed by perception experiments that can only be performed by a few military specialists. Getting observers for the original set of TTP experiments was a lengthy and frustrating process.

However, the current approach is to use the 24 picture field test set pictured in Paragraph 10 instead of the 144 picture set shown below. The 144 picture set allowed us to use 24 pictures per cell, with each cell having about the same average PID. The 24 picture set does not provide enough pictures to both avoid learning during the test and provide cells that have equal average probabilities. While we understand the problem, we do not understand accepting an unworkable solution.

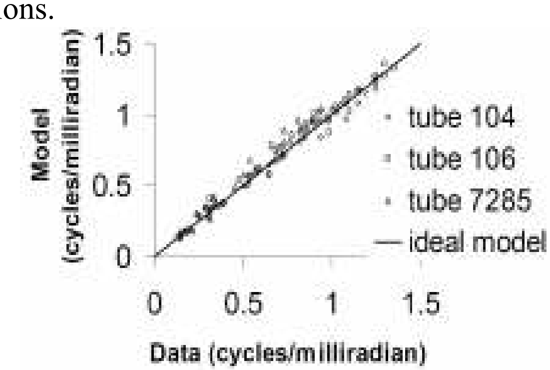


Disagreement 3: Recent Army models do not treat eye noise correctly. The change in the TTP noise model is described in Proceedings of SPIE 801406-1.

The change in the TTP noise model does not improve the accuracy of PID predictions; the opposite is true. In the chart, the abscissa is measured PID and the ordinate is predicted. Perfect accuracy is represented by the straight line. The Army model (NVIPM) predictions (red squares) are pessimistic when compared to the original TTP predictions (green circles).



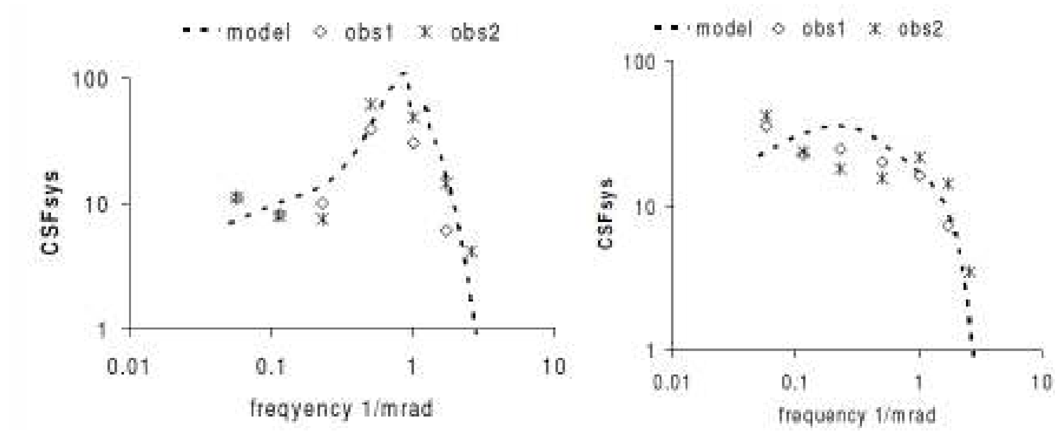
Also, unlike the NVIPM noise model, the original noise model was verified in several different ways. For example, the noise model in the original TTP predicted image intensifier performance very well. In the plot below, light level to the eye ranged from 3.6E-4 foot Lamberts to 35 foot Lamberts. One axis of the plot is prediction and the other measured, and the straight line represents perfect predictions.



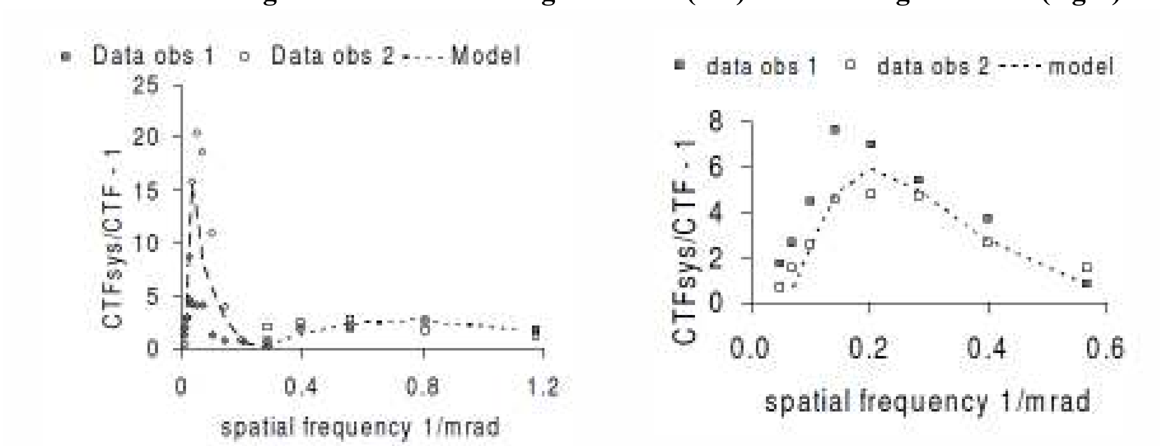
Also, unlike the NVIPM noise model, the original TTP CTF in noise predictions were compared to data from van Meeteren and Valeton, JOSA A **5**(3), and Stromeyer and Julesz, JOSA A 62(10). The noise bands are summarized in the table. More detail is provided in Optics Express 17(20), page 17253. Data for observers 1 and 2 are labeled obs1 and obs2, respectively.

It seems clear from these CTF in noise comparisons and the noise data from perception experiments that the original TTP does not have a problem predicting the effect of noise on CTF and therefore the effect of noise on PID.

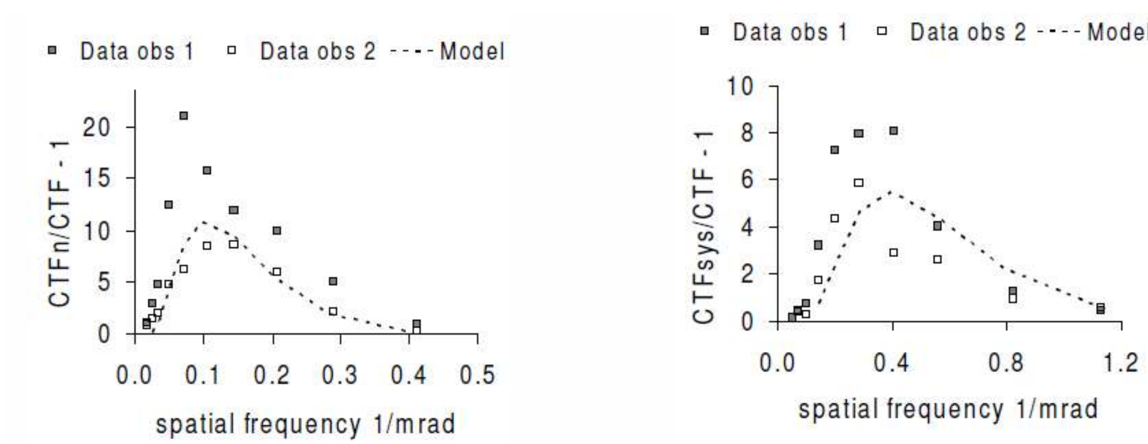
band number	low pass 3dB mrad^{-1}	high pass 3dB mrad^{-1}
1	0.036	0.072
2	0.14	0.29
3	0.57	1.15
4	0.072	0.14
5	0.29	0.57



These figures are for medium grain noise (left) and coarse grain noise (right).



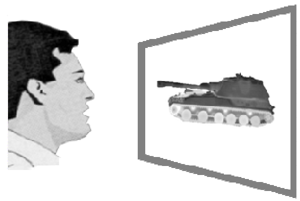
These figures are for noise bands 1 and 3 (left) and noise band 2 at (right).



These figures are for noise bands 4 and 5 at left and right respectively.

Disagreement 4: In-band aliasing (i.e. aliasing below Nyquist) does not affect NVIPM model predictions.

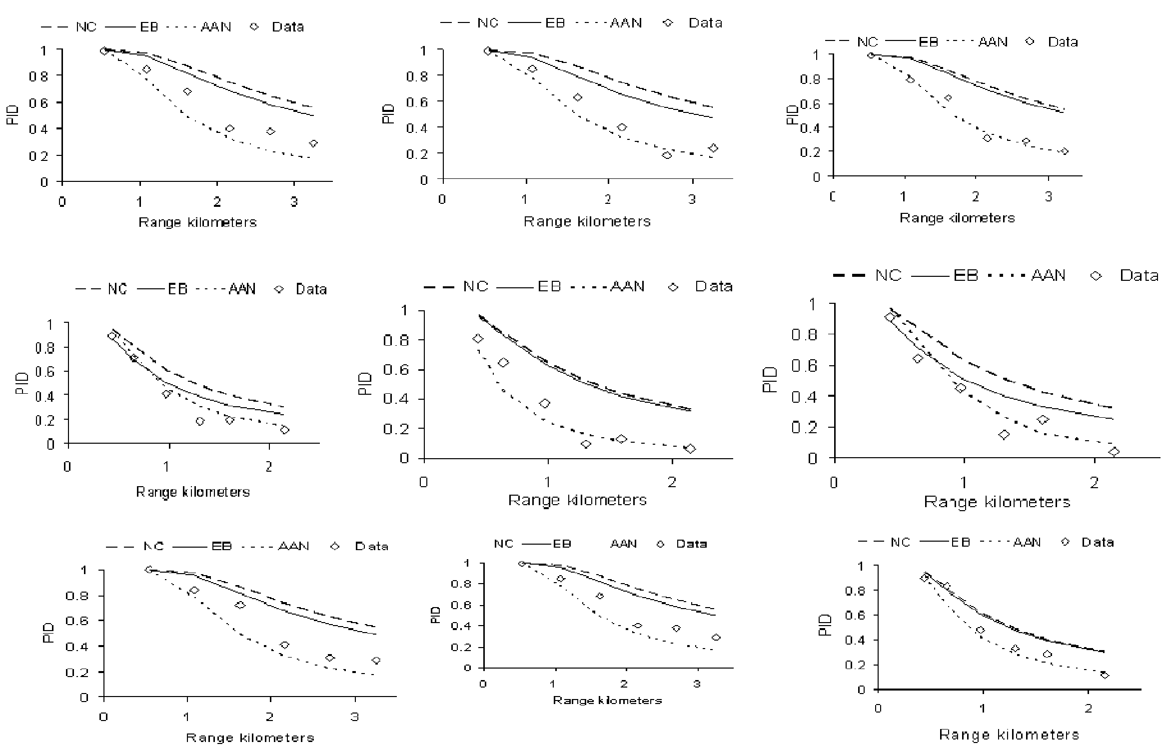
The current Army models use an early version of sampling correction. The first set of sampling experiments occurred before we had characterized the displays for perception experiments. Therefore, targets were displayed like the one shown in the figure; the target filled much of the display area, and that made the unknown display blur unimportant. It was easy to make the digitally applied blurs and noise big enough to obscure the target and get a probability curve.



When those experiments were analyzed, in-band aliasing had no effect on the probability of identifying the tactical vehicles. The Equivalent Blur model resulted; that model penalized only out of band aliasing that resulted from processes like pixel replicated image magnification.

The problem with the initial experiments was that the cues that differentiate one tactical vehicle from another are not small. One visual cue that helps to identify a vehicle is where the engine heats the frame, another is fender shape, or the number of roads wheels, or the size of the gun barrel. Those cues are not aliased when the vehicle is displayed as shown n the figure. Bolts and other small details are aliased, but they are not critical to target identification.

Later experiments showed definitively that in-band aliasing degrades visual discriminations. The charts below compare no correction (NC) for aliasing of TTP model predictions, Equivalent Blur (EB) correction, and Aliasing as Noise (AAN) corrections. The effect of aliasing on the probability of target identification is obvious, as is the improved accuracy of the AAN model.



Summary remarks about NVIPM model changes.

- Any changes in the TTP model should be checked against all of the validation data and not just selected data points from one face experiment.
- The decision to base CTF calculation on target angular subtense is hard to understand theoretically and is not supported by experiment.
- The modified noise model incorporated into NVIPM degrades PID model prediction accuracy. Further, the Army's new noise model is a numerical fit to one set of CTF in noise experiments. The Army researchers apparently did not check whether their CTF experimental results are consistent with the data collected by other experimenters. Since their CTF data has not been published, we cannot compare their data to the literature. The fact that their model fits their CTF data is not surprising, since their model is a numerical fit to that CTF data.
- Most Army modeling effort is directed at thermal imaging where aliasing tends not to be a problem. However, in recent years, facial identification has received increased interest. Facial ID requires comparing direct view spotting scope range with camera range. Unlike optical blur, aliasing is seldom a major factor in establishing range performance. It is almost always true that blurring an image to get rid of aliasing degrades performance much more than the aliasing. However, it is still true that significant amounts of aliasing affects range performance, and incorporating aliasing into the model necessitates no extra work on the part of the model user.
- There are three fundamental problems with the current Army modeling approach. The first is the failure to recognize that discriminating between a set of diverse objects like tanks and Howitzers is not the same as discriminating between a set of faces, or a set of sedan cars, or a set of pickup trucks where members of each set all share many spatial features.

  The second problem is that model changes are not vetted against available data, but rather changes are based on curve fitting a single experiment. Whether the modified model still predicts the original validation data is not checked, and their experimental results are not vetted against available literature. In fact, the model changes have degraded PID predictions. The third problem compounds the second; the experiments now performed by Army modelers use too few targets. If the cells are balanced to give the same probability, then the observer sees each picture so often that learning during the experiment is certain to happen.
- When done correctly, identifying targets using imagery taken at long range is a difficult visual task. In the past, we asked active military with appropriate MOS to do one of the tasks that the Army had trained them for and expected them to do in the field. None of what we did was hidden from the soldier's command; in fact, we needed command permission to get on the soldier's busy training schedule. We did those experiments in the hope of developing better models, so that the Army can buy the appropriate equipment for those same soldiers. The commanders understood what we were doing or they would not have hosted us and put us on the training schedule Somehow, that past endeavor on our part is now deemed inappropriate by our successors. What is actually inappropriate is ignoring the data provided by those soldiers and their command.