In G.C. Holst (Ed.), *Infrared Imaging Systems: Design, Analysis, Modeling, and Testing XV*, SPIE-5407 (pp. 96-103) Bellingham, WA., USA: The International Society for Optical Engineering.

Identifiability: a fast way to measure identification performance

Maarten A. Hogervorst [a], Alexander Toet [a], Piet Bijl [a], Brian Miller [b]
[a] TNO Human Factors, P.O. Box 23, 3769 ZG Soesterberg, The Netherlands
[b] US Army RDECOM CERDEC Night Vision and Electronic Sensors Directorate

ABSTRACT

We present a new measure called target *identifiability*, as an efficient alternative for measuring identification scores. Identifiability is operationally defined as the amount of blur required to reduce the target signature to its identification threshold. It can quickly be determined using a simple adjustment procedure. To validate the new measure, we measured the identifiability of targets in a set of real and simulated thermal images. The identification scores for these targets were available from a previous study. Our results show that identifiability indeed determines identification performance. Sufficient accuracy can be obtained with only a few (typically 2 or 3) trained observers. The associated measurement procedure is simple and requires only a limited amount of time.

Keywords: Identification, simulation, modeling, validation

1. INTRODUCTION

Military scenarios usually involve visual search, detection, and identification tasks. A large number of different electro-optical (EO) devices are available to perform these tasks. Some of these devices even enable the observer to see in the dark (image intensifiers, thermal cameras). In principle it is possible to compare the performance of different EO devices through well-conditioned observer experiments. In practice, visual search and detection experiments are often difficult since they inherently require numerous repetitions and/or a large number of observers to obtain statistically significant data. In field experiments it is usually impossible to perform the same scenario more than once, simply because one has no control over the environmental conditions. Also, field trials are usually extremely elaborate, and often very costly or even dangerous. But even in the laboratory, where one has in principle complete control over all experimental parameters, search experiments are very time consuming because they involve a large number of observations. Hence, there is a need for efficient metrics to quantify observer detection and identification performance with different types of EO devices, both for development and procurement purposes.

Nowadays virtual environments are increasingly used to train military observers in search, detection and identification tasks. A range of different visualization tools and algorithms is available to simulate both normal daytime imagery and imagery obtained with EO devices. Generated imagery is only fit for training purposes if it has a high degree of fidelity in those characteristics that are relevant for the task that will be trained. Thus, targets in imagery generated for scenarios involving search, detection and identification, should require the same amount of effort to find and should yield the same level of identification accuracy as their real-world counterparts. The fidelity of simulated imagery can in principle be assessed by performing visual search experiments. However, for the same reasons mentioned before, efficient task-related metrics are required to assess the fidelity of simulated imagery.

TNO previously developed a simple psychophysical measure called target *conspicuity*, that determines visual search and detection performance[1-3]. Target conspicuity is operationally defined as the maximal lateral distance between target and eye-fixation at which the target can be distinguished[4]. It characterizes the extent to which a target stands out from its immediate surroundings. The associated measurement procedure is simple. As a result, conspicuity measurements

can easily and quickly be performed in the field or in complex environments. Only a few observers (typically 2-3) are needed to achieve sufficient accuracy. Toet and colleagues (from TNO) have shown that conspicuity predicts human visual search performance in realistic and military relevant complex scenario's [3,5]. Also, conspicuity measured on photographic slides agrees with conspicuity measured in the field. This implies that the conspicuity measure can be used to validate simulated imagery[6-8].

Here we present a new measure called target *identifiability*, as an alternative to costly and time consuming identification experiments. Identifiability is operationally defined as the amount of blur required to reduce the target signature to its identification threshold. It can be determined quickly using a simple adjustment procedure. It determines identification performance and requires only a few (typically 2 or 3) observers and a limited amount of measuring time to achieve sufficient accuracy.

2. THE IDENTIFIABILITY CONCEPT

Although the mechanisms mediating visual target identification are still poorly understood, one thing is evident: a target will be easier to identify when its details are well resolved and clearly visible. It is therefore a priori likely that a measure that captures a target's visual articulateness should correlate with human visual identification performance.

The identifiablility of a target in a complex background can be determined by quantifying the visual articulateness of the target. Here we operationally define target identifiability as the amount of Gaussian blur that is required to reduce the target signature to its identification threshold. The rationale for the choice of a low-pass signature degradation filter is the fact that all spatial frequencies contribute to target identification[9]. The Gaussian blurring process is easy to implement. A simple adjustment procedure can be applied to quickly determine the threshold blur. This makes the target identifiability metric an attractive alternative for intricate identification experiments.

The method is similar to a well established method to determine legibility of (e.g. traffic) signs[10]. With this method the distance to the image is varied and a threshold identification distance is determined. In that case the blur stems from blur in the visual system. In our situation a scene is viewed with a sensor system, and the main source of blur originates from the sensor and not from the visual system. Therefore it makes sense to apply an external blur to the image. In this way, the (extra) variation between subjects due to differences in visual acuity will be eliminated.

In the next sections we describe an experiment that was performed to verify the new target identifiability metric. The results of this experiment will show that identifiability indeed determines identification performance, and that it therefore provides an efficient alternative for measuring identification scores.

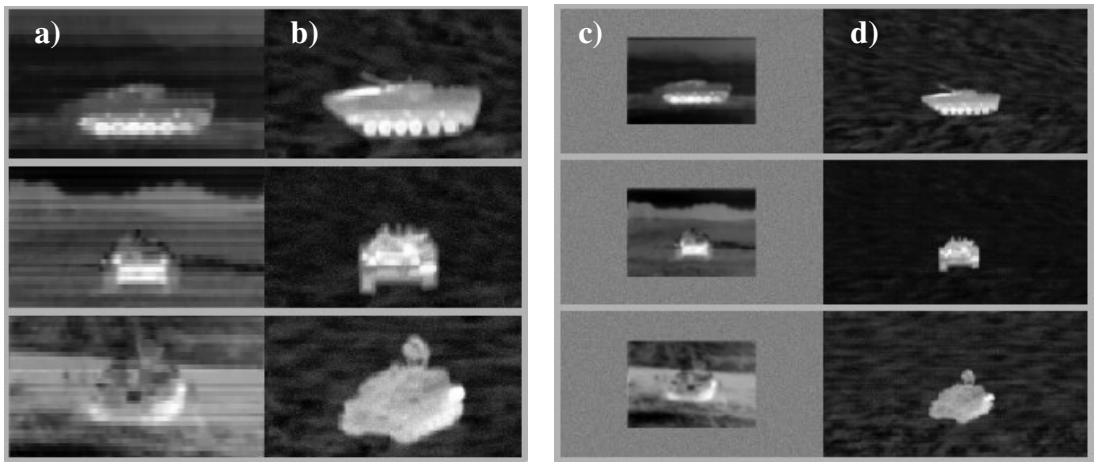

Figure 1. Example reference images (a and c) and synthetic (uncorrected) versions (b and d) for first generation FLIR (left: a and b) and advanced scanning FLIR (right: c and d). For details see Jacobs et al.[11]

3. VERIFICATION EXPERIMENT

Jacobs et al.[11] recently performed an experiment in which they compared identification performance for realistic and synthetic thermal imagery. Realistic imagery representing first generation and advanced scanning sensor systems taken from a set of vehicles at various ranges was derived from thermal imagery taken at a short distance. Parameters derived from the sensor data were used to generate synthetic imagery using the NVESD Paint the Night simulation (see http://www.tec.army.mil/TD/tvd/survey/Paint_the_Night.html). Both image sets were then used in a target identification experiment with trained human observers. An example of these images is shown in Figure 1. For details about the image registration and generation procedures we refer to Jacobs et al.[11] In this study we use a subset of their imagery and the corresponding identification data to validate our new identifiability metric.

3.1. Experimental setup

Two observers participated in the experiment. The experiment was run on a PC and controlled by Matlab software. The images were shown on a 22 inch CRT-monitor, displaying 1280 x 1024 pixels in a screen area of 29.7 x 40.5 cm^2. The observer was seated at a regular distance from the monitor (around 50 cm) and was free to adjust his/her distance to the screen. The experiments were performed in a dimly lit room.

3.2 Imagery

Reference image sets representing realistic imagery as obtained with *first generation* and *advanced scanning* long wave infrared (LWIR) sensors were derived from real images of tanks and armored personnel carriers taken at a nominal range of 125 meters with an Agema LWIR camera. Each set of 50 images represented 5 different military vehicles (2S3, BMP-1, T-62, T-72 and ZSU) in natural terrain backgrounds, registered at 5 different viewing distances, for two different aspects (viewing angles) for each vehicle. At each distance all 10 different aspects were represented across the various vehicles. The resulting 100 reference images were also simulated twice with Paint the Night, using two different parameter settings (see Jacobs et al.[11] for more details on how the images were created). In their experiment Jacobs et al. [11] found that the set of parameters that was initially used to generate the synthetic images was not entirely complete, and additional parameters were needed. They therefore generated a second set of corrected synthetic images. In the rest of this paper we will refer to the original and corrected synthetic image sets as the “synthetic” and “corrected synthetic” images. The entire image set therefore consisted of 100 nominally identical triplets of reference, synthetic and corrected synthetic thermal images.

3.3 Procedure

Prior to participating in the test, each observer was trained in target identification using the Recognition of Combat Vehicles (ROC-V) training software (see http://www.peostri.army.mil/PRODUCTS/ROCV/). ROC-V is a Windows-based thermal sight training program developed by CECOM NVESD. It helps soldiers learn to identify the thermal signatures of combat vehicles through the use of an interactive curriculum that teaches the unique patterns and shapes of vehicle ' hotspots,' and overall vehicle shapes.

At the start of each trial a randomly selected image from the test set was presented and the observer had to indicate the type of the vehicle that was displayed. After the observer had identified the target, he/she pressed a button on the keyboard, and written feedback was given by displaying the name of the actual vehicle type on the screen. After another second button press the name disappeared and the observer started the actual identifiability measurement procedure. This process involved the adjustment of the width of a Gaussian blur kernel. By pressing the up and down arrow keys on the keyboard the observer could increase or decrease the amount of blur of the displayed image. The amount of blur could be altered in steps of 15% or 3.5% (for fine tuning), amounting to an increase in sigma (width of the Gaussian blur kernel) by a factor 2 in 5 respectively 20 steps. In a typical measurement the observer initially increased the amount

of blur until he could no longer identify the target. He then decreased the blur until he felt certain he could identify the target again. This process of blurring and de-blurring was repeated until the subject felt confident that he had reached the threshold.

In case the observer could not identify the un-blurred image the threshold was set to a value of 0.76 pixels, corresponding to 2 steps below a blur of 1 pixels, i.e. $\sigma = 2^{-i/5}$, i = -2 (Note that, in principle, the image should be de-blurred in this case to obtain a threshold image). By artificially setting the value to a low value whenever the un-blurred image could not be identified, we assured that the average (e.g. of all images corresponding to a given distance) was affected by the amount of images that could not be identified.

We obtained identifiability measures for each of the 300 images. The experiment was run in 8 sessions that lasted approximately 7 min. per session per observer (amounting to a total of 56 minutes per observer). Identifiability is expressed in image pixels; each pixel corresponds to a (fixed) angle of about 0.065 mrad (the same for both sensors).

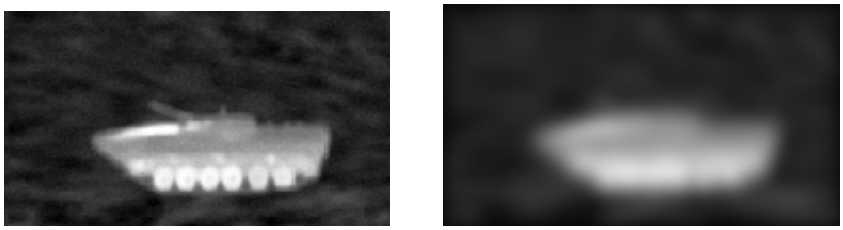

Figure 2. Example of an original image (a) and its blurred version (b). The task for the observer was to set the amount of blur to a value for which the vehicle could just be identified. This threshold value is referred to as the *identifiability* measure.

4. RESULTS

The results of the verification experiment are shown in Figure 3. Figures 3a and 3b reproduce the results from Jacobs *et al*[11]. These figures show the proportion of correctly identified targets (PID) as a function of the viewing distance for a first generation and an advanced scanning FLIR. Results for realistic ("reference") imagery is compared with results for original synthetic imagery ("synthetic"), and synthetic imagery for which the parameter setting has been corrected ("corrected synthetic") such that PID performance for these images matches PID performance for realistic images. The results clearly show that PID for the original synthetic imagery was much higher than the PID for realistic imagery. After correction PID performance for the synthetic images matches PID performance for realistic images.

Figures 3c and d show identifiability expressed in pixels (averaged over the two observers) for the same images. The results show a similar pattern as the PID-performance results. In both cases the thresholds corresponding to the original synthetic data are (significantly) higher than the reference data, but the corrected synthetic data matches the reference data (not significantly different[1]).

[1] A One-Way ANOVA-test shows a significant difference between the corrected synthetic data and the reference data and no significant difference between synthetic data and the reference data.

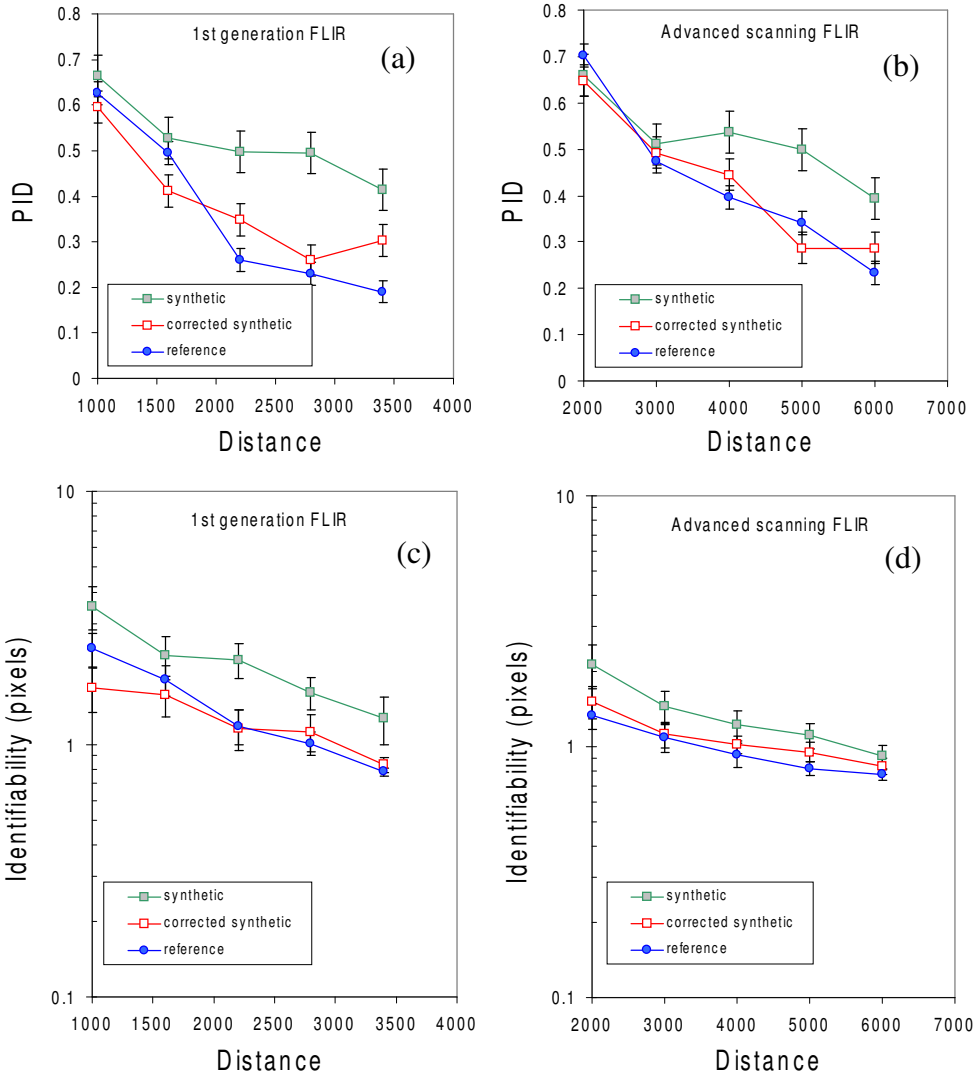

Figure 3. The top two figures (a and b) show the identification score (PID) measured by Jacobs *et al.*[11] as a function of distance for 1st generation (a) and advanced scanning (b) FLIRs. The curves correspond to realistic imagery ("reference", filled circles), synthetic (filled squares) and corrected synthetic (open squares) imagery. Error bars indicate the variability due to the probabilistic process (binomial distribution). The lower two figures (c and d) show identifiability (blur threshold) in pixels (1 pixel = 0.065 mrad) for the same images, where the judgments of two observers have been averaged. Error bars indicate the variability in the judgments for different vehicles and viewing angles at a given distance.

Figure 4 shows the relationship between PID (measured by Jacobs *et al.*[11]) and identifiability (in pixels, as well as in mrads). PID-score is plotted against identifiability for the two imagers (first generation and advanced scanning FLIR), for the two observers (Fig 4a and 4b) and the average over the two observers (Fig.4c and Fig 4d). The data are fitted by logarithmic lines. The correlation between PID and identifiability is high, with an average correlation coeffient (R) of 0.91 for first generation FLIR images and 0.88 for advanced scanning FLIR images (for the individual observers the correlation coefficients are respectively: 0.92 and 0.90 for observer MH, and 0.89 and 0.80 for observer PB)[2].

[2] R represents the linear correlation correlation coefficient between PID and log(identifiability).

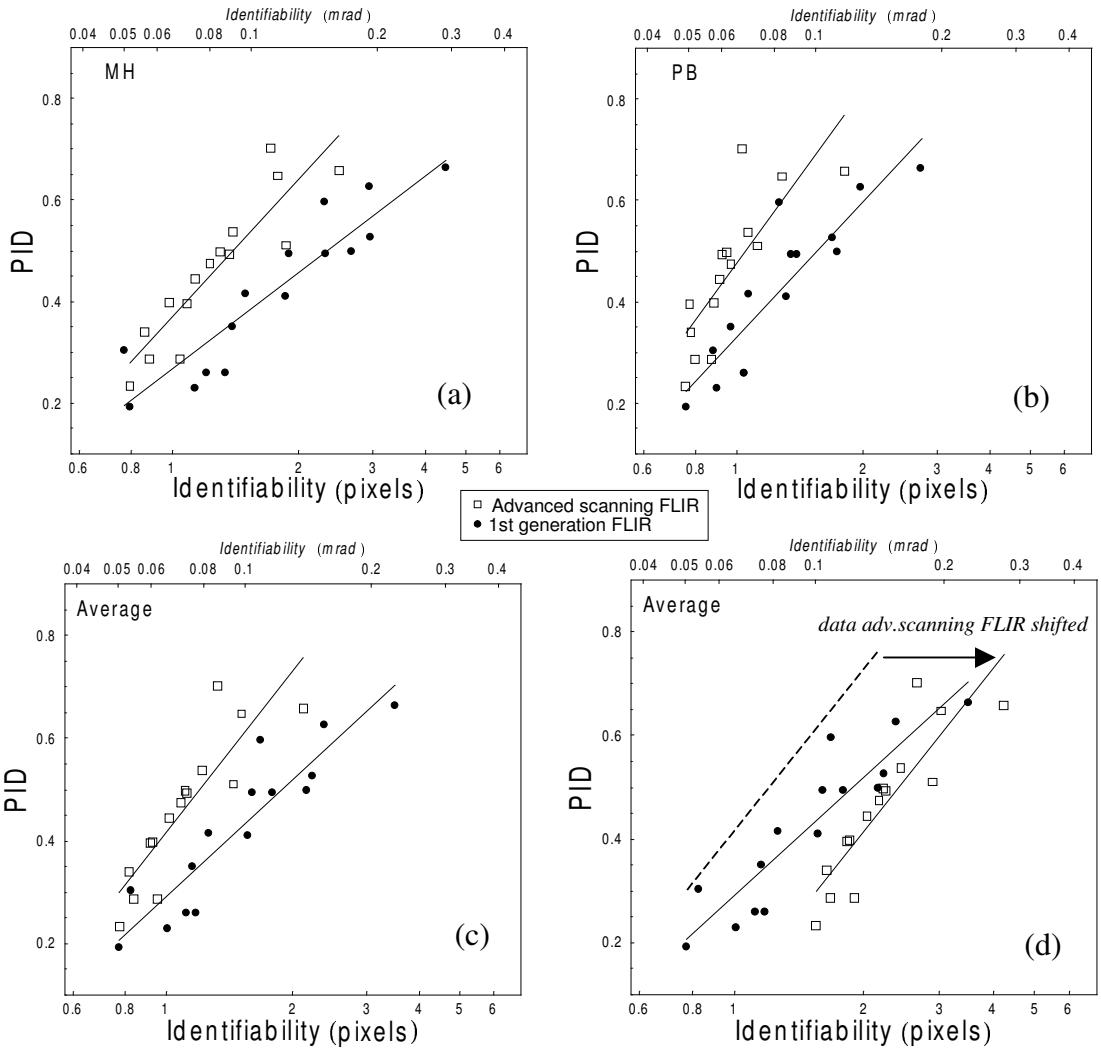

Figure 4. PID-score, as measured by Jacobs *et al.*[11] versus identifiability for each of the two observers (a and b) and the average over the two observers (c and d) for first generation (open squares) and advanced scanning (closed circles) FLIR. Figure 4d shows the same data as Fig. 4c, but now the identifiability measures for the images of the advanced scanning FLIR are multiplied by two, taking the difference in image size (in pixels) into account. Identifiability measures are averaged over all images registered at the same viewing distance (averaged over sensor type, vehicle type and orientation). The drawn lines represent logarithmic fits to the data. Correlation coefficients are given in the text.

The relationship between PID and identifiability depends on the type of sensor. This can be explained from the fact that the detector size of the advanced scanning FLIR spans about twice as many pixels as that of the first generation FLIR (see Fig.1). This means that the images of the first generation FLIR are approximately scaled versions of the advanced scanning FLIR images, scaled by a factor of two. When the identifiability measures for the advanced scanning FLIR images are multiplied by two the data of the advanced scanning and the first gen FLIRs become more alike (Figure 4d). The fact that these data do not coincide can be attributed to the fact that the images of the two sensor types are not scaled versions. For low PIDs the contribution of the sensor to the image degradation is large relative to the external blur. Therefore, the data of the two sensors differ for low PIDs. The data comes together at high PIDs where the image degradation due to sensor quality is small relative to the external blur.

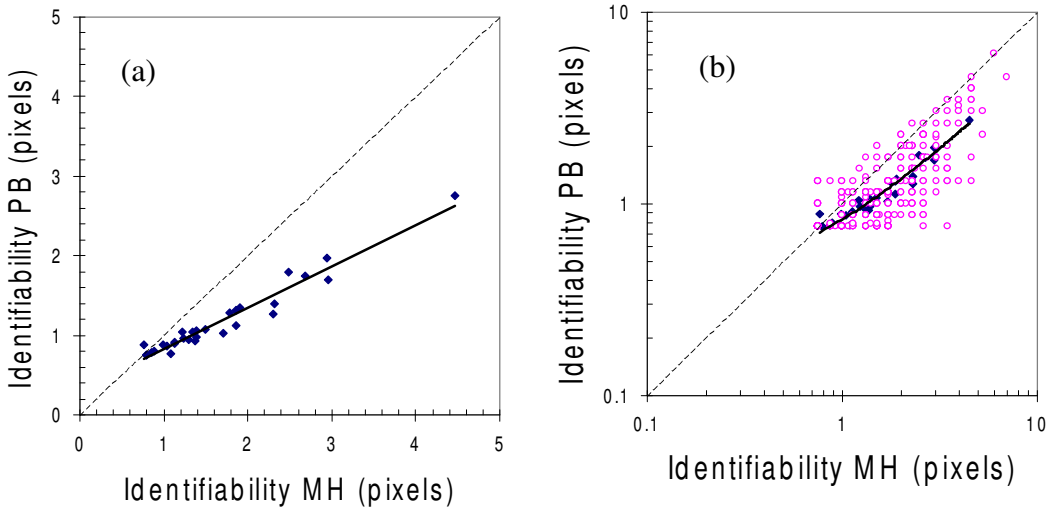

Figure 5. Correlation between the identifiability judgments of two observers. Fig 5a shows the average data (averages over all images registered at the same viewing distance and with the same sensor type) on linear axes, Fig 5b shows the values of all images (open symbols) together with the average data from Fig 5a on log-log axes. The drawn line in both figures corresponds to the linear fit. The dotted line corresponds to the condition in which both observers are in complete agreement.

Figure 5a shows the correlation between the identifiabilty judgments of two observers. The data points correspond to averages over all images registered at the same viewing distance and with the same sensor type. The judgments of both observers are highly correlated, with a linear correlation coefficient of $R = 0.97$. The judgments of both observers show a linear relationship, in which the estimations of observer MH are about twice as high as the judgments of observer PB. The fact that for low identifiability values the results for both observers become more similar can be attributed to the fact that the minimum value of identifiability that could be chosen was 0.76 pixels (see Procedure section). The results show that observers agree well on the order of how well targets can be identified, but may differ in the absolute value of their judgments. It is therefore advisable to use the same observers when comparing synthetic with real imagery.

Figure 5b shows identifiability for each image separately, as well as the averaged data from Fig 5a, for both observers. (This figure clearly shows the lower boundary effect, i.e. the fact that all values are higher than 0.76.) The data shows a relative standard deviation from the linear fit of 23%.

5. CONCLUSIONS AND DISCUSSION

We have devised a method to gain insight into human identification performance without having to resort to elaborate and costly experiments. Identifiability gives reliable estimates with a small number of observers and limited measuring time. We have shown that identifiability is directly related to PID-performance. Identifiability is therefore well suited for comparing synthetic and realistic imagery.

The fact that the relationship between PID-score and identifiability depends on the sensor makes it less suited for comparing the performance of the different sensors. When identifiability is measured the image is blurred until a threshold is reached corresponding to a fixed (low) PID-score. The assumption is that the PID-score of the unblurred image is related to the amount of blur required to reach the threshold PID value. The fact that the relationship between PID-score and identifiability differs for the two sensors indicates that the rate at which PID-score decreases with increasing blur differs between the two systems. When this assumption does hold identifiability can be used to compare performance for different sensors. Then it seems best to express identifiability in terms of visual angle (similar to

resolution). In our case the identifiability in visual angle (e.g. mrads) is simply a factor times the blur in pixels (see Figure 4).

The general idea is to degrade an image until perceived identifiability is at a threshold level. We chose to use blur as the degradation method. In principle other degradation methods can be deployed just as well, such as the addition of (white) noise. However, blur is one of the major factors to determine identification distance with and without a sensor, while noise determines identification distance to a lesser extent. Therefore it makes sense to use blur to estimate ID-performance.

To obtain reliable estimates of identifiability it is best to use trained observers, since they are capable of judging for what amount of blur a vehicle can still be discriminated from other vehicles.

There are several advantages to using identifiability over performing an experiment in which PID-scores are recorded. Apart from the fact that it can be obtained easy, fast and with a limited number of observers, identifiability makes it possible to estimate PID performance of single images. Synthetic imagery/models can be evaluated and adjusted (in a continuous loop) until identifiability matches that of realistic imagery.

6. REFERENCES

1. Toet, A. and Bijl, P., Visual conspicuity, In: R.G. Driggers (Ed.), *Encyclopedia of optical engineering*, pp. 2929-2935, Marcel Dekker Inc., New York, USA, 2003.
2. Toet, A., Bijl, P., Kooi, F.L. and Valeton, J.M., Quantifying target distinctness through visual conspicuity, In: W.R. Watkins & D. Clement (Ed.), *Targets and Backgrounds: Characterization and Representation IV*, SPIE-3375, pp. 152-163, The International Society for Optical Engineering, Bellingham, WA, 1998.
3. Toet, A., Kooi, F.L., Bijl, P. and Valeton, J.M., Visual conspicuity determines human target acquisition performance, *Optical Engineering,* 37*(7),* pp. 1969-1975, 1998.
4. Wertheim, A.H. *A quantitative conspicuity index; theoretical foundation and experimental validation of a measurement procedure,* (Report C-20), TNO Human Factors, Soesterberg, The Netherlands, (1989).
5. Toet, A. and Kooi, F.L., Conspicuity: an efficient alternative for search time, In: A.G. Gale, I.D. Brown, C.M. Haslegrave & S.P. Taylor (Ed.), *Vision in Vehicles*, VII, pp. 451-462, Elsevier Science Ltd, Oxford, UK, 1999.
6. Toet, A. *Visual conspicuity of targets in synthetic IR imagery,* (Report TNO-TM 1999 C044), TNO Human Factors, Soesterberg, The Netherlands, (1999).
7. Toet, A., de Vries, S.C., Bijl, P. & Kooi, F.L. *Validation of military target representation in a simulator through conspicuity,* (Report TNO-TM 1996 A-061), TNO Human Factors, Soesterberg, The Netherlands, (1996).
8. Toet, A., Visual conspicuity of targets in synthetic IR imagery, In: A. Toet (Ed.), *Camouflage, Concealment and Deception Evaluation Techniques*, RTO-TM-017 AC/323(SCI-012)TP/32, pp. 137-152, North Atlantic Treaty Organization, Neuilly-sur-Seine Cedex, France, 2001.
9. Driggers, R.G., Vollmerhausen, R.H. and Krapels, K., Target identification performance as a function of low spatial frequency image content, *Optical Engineering,* 39*(9),* pp. 2458-2462, 2000.
10. CIE. *Roadsigns,* (Report CIE 74), International Commission on Illumination CIE, Vienna, Austria, (1988).
11. Jacobs, E., Edwards, T., Miller, B. and Hodgkin, V., Comparison of ID performance using real and synthetic imagery, In: R. Appleby, G.C. Holst & D.A. Wikner (Ed.), *Infrared and Passive Millimeter-wave Imaging Systems: Design, Analysis, Modeling, and Testing*, SPIE-4719, pp. 34-41, The International Society for Optical Engineering, Bellingham, WA, 2002.