Search by photo methodology for signature properties assessment by human observers

Gorm K. Selj*[a], Daniela. H. Heinrich[a],
[a]Norwegian Defence Research Establishment, P.O. Box 25, N-2027 Kjeller, Norway.

ABSTRACT

Reliable, low-cost and simple methods for assessment of signature properties for military purposes are very important. In this paper we present such an approach that uses human observers in a search by photo assessment of signature properties of generic test targets. The method was carried out by logging a large number of detection times of targets recorded in relevant terrain backgrounds. The detection times were harvested by using human observers searching for targets in scene images shown by a high definition pc screen. All targets were identically located in each "search image", allowing relative comparisons (and not just rank by order) of targets. To avoid biased detections, each observer only searched for one target per scene. Statistical analyses were carried out for the detection times data. Analysis of variance was chosen if detection times distribution associated with all targets satisfied normality, and non-parametric tests, such as Wilcoxon's rank test, if otherwise. The new methodology allows assessment of signature properties in a reproducible, rapid and reliable setting. Such assessments are very complex as they must sort out what is of relevance in a signature test, but not loose information of value. We believe that choosing detection times as the primary variable for a comparison of signature properties, allows a careful and necessary inspection of observer data as the variable is continuous rather than discrete. Our method thus stands in opposition to approaches based on detections by subsequent, stepwise reductions in distance to target, or based on probability of detection.

Keywords: Camouflage, Human observers, Relative comparison, Visual search, Conspicuity, Detection times.

1. INTRODUCTION

Camouflage and signature suppression is a very important part of force protection. Furthermore, the ongoing development in such areas makes it progressively more difficult to detect and recognize targets. New technology for detection of targets as well as for improving concealment demands a continuous development of assessment methodologies, given by reproducible and well documented procedures. This is an important, but still very difficult task as there is always a risk that the final recommendation (of, say, a camouflage pattern) depends on the test method, the evaluation criteria, or any other parameter that has to be chosen by the research team. Ideally, we would want available some broadly applicable, low-cost, unbiased and reliable signature evaluation methodology.

Several methods for signature properties assessment have already been developed, all aiming to rank the targets under consideration as well as possible. Photo-simulation by using human observers as an assessment method of camouflage effectiveness has been used in various forms in the recent decades [1-4]. Other methodologies have also been tested out, involving video surveillance [5], simulation of human vision [6-11], similarity measures of target-background by image analysis techniques [2,11], assessment by simulation of target vehicles against different backgrounds [12], and image sequences taken by approaching sensors [13].

In this paper we describe an alternative approach to the photo-simulation methodology for the evaluation of signature properties, mainly in the visual spectrum, but with potential extension to thermal and infrared parts of the electromagnetic spectrum. The methodology, being objective in nature, was developed during a full test of camouflage effectiveness of various targets, and has thus been carried out and evaluated. A similar study to ours has earlier been carried out by DSTL, UK [14]. The methodology, using humans in a search by photo observer trial, rests upon a large number of detection times of targets in a variety of natural backgrounds of high operational importance and relevance. It aims to be an objective and reliable measure of camouflage effectiveness among several targets. By handling observer data carefully, targets are not simply ranked by order. Hence, the methodology allows for a relative measure of performance, that is, how much better some target is compared to another. This kind of relative comparison (amongst targets) is very important as it narrows the gap between test and operational performance.

Signal Processing, Sensor/Information Fusion, and Target Recognition XXIV,
edited by Ivan Kadar, Proc. of SPIE Vol. 9474, 947411 · © 2015 SPIE
CCC code: 0277-786X/15/$18 · doi: 10.1117/12.2176983

We will in this paper focus on the methodology and discuss possibilities, pitfalls and future improvements. Hence, results from the camouflage evaluation tests we carried out recently will only be presented for illustration purposes.

2. METHODS

2.1 Preparations for observer trials

Before the observer trial can be onset, targets have to be recorded (*e.g.* photographed) in several different terrain backgrounds of relevance to the purpose of the test. A set of those images will in turn be presented to the observers during the trial conduction itself. The methodology description in the following section describes how we carried out our recently finished trial campaign, with the aim to evaluate signature effectiveness for 6 generic targets, consisting of slightly different camouflage test patterns (T0 to T5) intended for green forest backgrounds. The patterns were tested as mannequin jackets (ref Fig. 1) in our specific trial, but are not restricted to that.

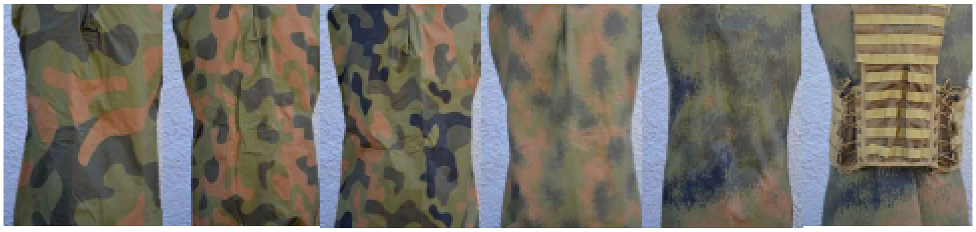

Figure 1. Close-up image of the 6 different targets to be evaluated by our observer trial methodology. One of the targets wore a dummy-west playing the role as a reference, due to its different visual appearance.

Image capture in field of operation

The quality of the observer trial outcome will in general depend on the quality of the photographs that are presented to the observers. Image quality in this context is very broad, involving physical parameters such as resolution of target and illumination conditions [15,16], but also operational relevance of the images to their non-biased appearance to the observer. In the section below, we give a short description of the steps during the image capture of test targets that we needed to run through before the observer trial was carried out.

Background data capture

The scenes intended for our trial were chosen to span the areas of operation in which the test targets were thought to be used operationally. This means that most relevant types of natural (green forest) backgrounds were covered, which in principle means that the higher the number of unique scenes, the better. When we carried out our trial, the 6 individual targets, which we wanted to compare relatively, were placed in each scene. We recorded a large number of 14 scenes, ensuring a span in different (and relevant to the main purpose of our trial) backgrounds.

We recorded the targets in as identical conditions as we were able to (same target orientation, position and area exposed, stable illumination conditions) as we were able to. This means that we assured that all targets had the same position and orientation from the observer's perspective, as well as that the illumination conditions of the exposed area was as stable as possible. To achieve that, we carried out a near-continuous (within minutes) recording of the targets in each scene. Furthermore, only one target was recorded per image to avoid any confusion about what is actually to be assessed by the observer.

The scenes were chosen so that whenever an observer was subjected to a sequence of images, we wanted a minimal bias in the way they were presented to the observers. Targets were located randomly in the image frame (not always centered) in order to avoid observers' expectations on where to start to search. Also humans tend to start searching at the center of a display [17] which would bias the results. The physical distance to the targets in the field was varied between 8 and 150 m in our trial. When the scene images were recorded, our aim was to construct the scenes so that the target actually was possible to detect, whenever the observer's eye focus was at the target's spot in the image frame. Still, it was not easy to decide in advance, when situated in the fields, whether a target was going to be possible to detect or not for the observers. Therefore some control-testing of scenes was conducted prior to the trial; some scenes were discarded in that process (Scene discardment also happened whenever variations in illumination conditions between targets were found unacceptable).

2.2 Methodology – conducting the trial

Based on the *outdoor* target recordings, we carried out the trial *indoors* where observers' conditions were kept stable. The observers, which were 148 recruit soldiers, performed the test singly without the possibility of interacting during or after the test to avoid any undesired bias to the final results.

Preparing observers for the trial

Prior to the observer trial, each soldier was asked to fill in a form about his/her military background, vision anomalies (such as colour blindness), experience with hunting or target recognition etc. Thereafter, each soldier was given a word by word *identical* introduction to the observer trial by an instructor as illustrated in Fig. 2. This reduced the bias in the results by the instructions [18]. Each observer was adjusted to have an optimal and identical distance to the widescreen (ca 40 cm), as the screen is to fill most of the observers' field of view. Also, the observer's eyes were approximately leveling the center of the screen. Thereafter, each of the observers conducted a test run consisting of two scene images were the observer were to search for a target in different natural backgrounds. During this test run, the observers were allowed to ask questions to the instructors, reducing the risk of misunderstandings before the main trial started. (During trial itself, observers were not allowed to ask questions, but left to find targets solely by themselves). Finally, the observers were free to choose their own search strategy, and were not instructed to scan the image or start the search at some particular position in the image frame.

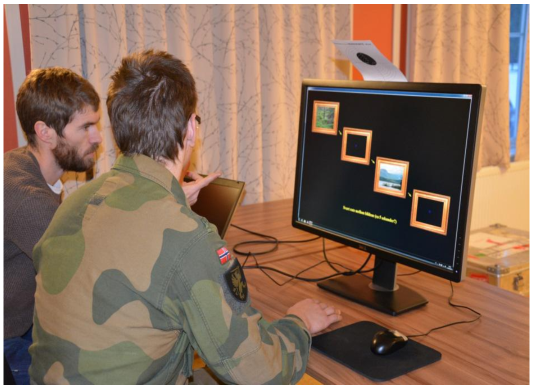

Figure 2. The observer trial preparations consist of a short, consistent brief to each observer. Then the observer has to carry out some test images, searching for targets, before the trial itself is started.

Trial conduction

The observer search trial normally starts with the presentation of a new scene. During the trial, that we carried out recently, each observer was shown a randomized sequence of photographs of different scenes, one at the time in a high definition (HD) wide screen (2560 x 1600 pixels) in a dimly lit room. Each observer ran through all the 14 unique scenes in the trial. Each photograph represented a scene, collected in operationally relevant areas, with either one or no target. An example of a scene is shown in Figure 3. The observers searched for a target and indicated detection by mouse-clicking at the target as soon as he or she felt confident that it was a proper target and not an anomaly (e.g. a target-shapet bush etc). The corresponding detection time was then stored. There was a small tolerance surrounding each target (see Figure 3). Hence the observer had to click close to the target to indicate a "hit", but not necessarily spot on (the observer trial method is a detection test, not a mouse-clicking accuracy test). Each observer was exposed no more than one single target per scene, as targets were identically positioned in each scene.

The total duration of each scene image presentation was limited to 60 s in the trial we carried out recently, but can be adjusted on demands. The time limit was set not be too long, which would increase the risk of reducing the observer's concentration trough tedious searches for a well hidden target. Whenever our fixed time limit was exceeded, the target (in that particular scene) was considered not detected. Furthermore, in order to carry out the trial we developed a dedicated software tool, presenting the trial-scenes in a randomized order and keeping data in a sorted manner for further analysis.

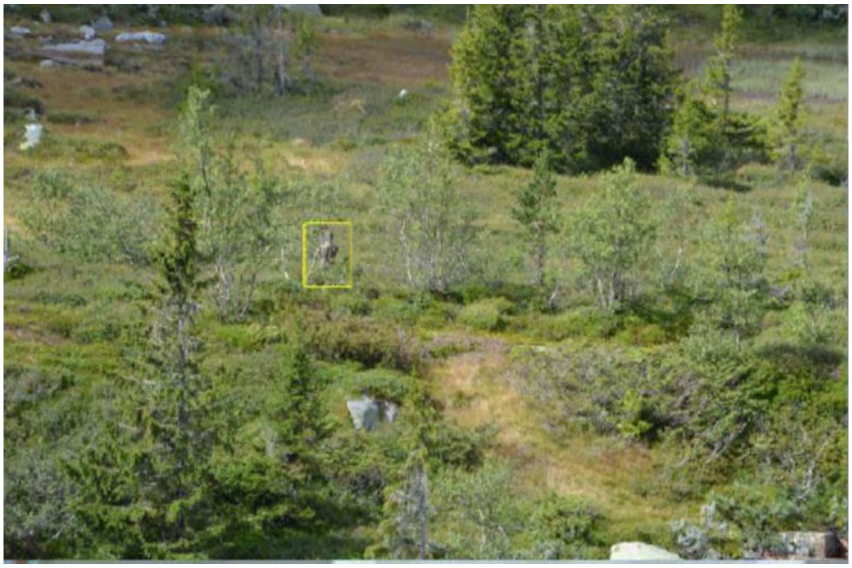

Figure 3. Illustration of a target in a scene with tolerance (yellow rectangle), indicating what is considered to be a proper detection. The tolerance was not visible to the observers (for obvious reasons) during the trial.

In short, the procedure of the observer trial ran as follows:

- Detection time and location (co-ordinates of mouse click) of the detection were both automatically logged for each distinct target in each individual scene for each of the observers.
- Any non-detections or miss-detections were also logged.
- There was a pre-defined search time limit for each scene.
- Before showing a photograph of a scene, the HD computer screen was black for about 5 seconds. Hence, the observer had to focus on some reference point (e.g. circular, black target above screen in Figure 2), ensuring equal search conditions within each scene.

3. STATISTICAL ANALYSIS OF OBSERVER DATA FOR RANKING OF TARGETS

In this section the purpose is to give an overview of how we in this paper suggest the sorting of the observer data from a trial to be carried out. The aim is to be able to extract the vital (significant) information of signature properties, collected during a trial, from all other (non-significant) "noise".

3.1 Comparison of signature properties within a certain scene

An important issue we needed to establish ahead of any comparison and subsequent rank by order of the test targets was the ranking criteria. In the trial that has recently been carried out by the authors, we did choose *detection time* as the primary criterion with probability of detection as a possible secondary criterion. This important issue is considered further in the Discussion section.

On a general basis detection times for a target will be a distribution, that is, a spread of numbers from low to high. Often the spread of detection times will not be symmetric around its mode (*i.e.* the most frequent detection time), but shifted towards higher values (*i.e*. a right-shifted distribution) [19-21]. As detection times can be non-negative, limited space for deviations exist below the mode. On the contrary, above the mode there is room for high deviations (at least until the search time limit fixed by the trial operator). A mean value of a test target's detection times can thus be misleading, simply because single outliers potentially will shift the mean detection time towards a high value (but almost never towards a much lower value). The median, however, has the advantage of not being much right-shifted by one or two high detection times (outliers). Hence, in our methodology we believe the median reflects the signature properties of the target in a better way than the mean as also suggested in earlier works [18, 22]. In addition, and importantly, the median also accounts for the non-detections as it simply counts them as high numbers.

Non-detections

The number of non-detections during our trial was handled carefully to include their value in a rank of targets based on detection times. By definition, the non-detections were not assigned any time value in the trial, but were treated as some undefined value above the search time limit. We note that if the non-detections outnumbered the detection times, the median of that particular target turned out to be a "non-detection" which still can be used in a comparative test with other targets. In special cases, if the distinct number of observers for a target was even, in combination with the rare event that the median turned out to be the average of a (well defined) physical detection time and the "first" non-detection, this particular non-detection was assigned the value of the search time limit, enabling a well-defined median, albeit with a conservative estimate. In any case, the median preserved the valuable information, represented by non-detections, about the test targets.

3.2 Test for significance among targets

We will now describe in short how we tested for differences of significance among the targets during our trial. We did this by carrying out statistical tests which supplied us with a *p-value*. The *p-value* is then the probability that the two targets we consider had a non-distinguishable performance. Hence, the lower the *p-value,* the more probable it was the test targets of consideration being different in performance. From literature, it is common to say that *p-values* lower than 0.05 indicate a difference of significance [22]. Below, we give a short description in the statistical methods we used when analyzing our data from the trials

Comparing targets

To be able to test whether targets differed significantly from one another or not, we followed one of the two paths; that of parametric tests and that of non-parametric [22]. The parametric route is recommended whenever the distribution of detection times follows a well-defined mathematical description, such as the normal distribution. Then statistical packages (such as ANOVA) are available to test significant differences between targets based on the mathematical shape of the distribution. The outcome is a *p-value* telling us whether the two test targets were likely to be significantly different or not.

If the distribution of detection times for a target fails to follow any known distribution, it is considered non-parametric. The lack of a mathematical description of how the detection times are distributed complicates the safe establishment of significant differences as there is no optimal test statistic. The most common procedure is based on the chronologic rank of each of the detection times for the individual targets [22, 23]. Normally Wilcoxon's rank test can be used if there are only two targets to be compared in the test and Kruskal-Wallis if the number exceeds two (e.g. 6 targets in our trial) [18, 22]. The parametric tests are commonly based on the median value as this, too, is found by sorting the detection times chronologically. The outcome is, as for parametric tests, a *p-value* telling us whether the two targets were likely to be different or not. Figure 4 gives a simplified overview of the steps in testing for significant differences among target's performance based on human observers' detection times.

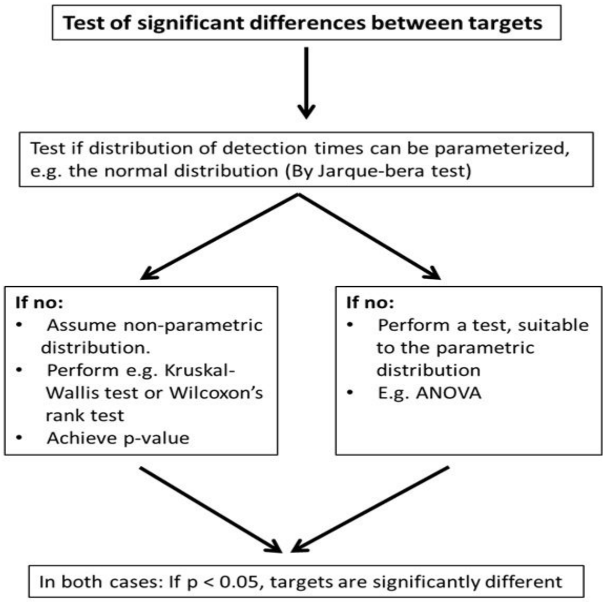
Test of significant differences between targets
Test if distribution of detection times can be parameterized, e.g. the normal distribution (By Jarque-bera test)
If no:
• Assume non-parametric distribution.
• Perform e.g. Kruskal-Wallis test or Wilcoxon's rank test
• Achieve p-value
If no:
• Perform a test, suitable to the parametric distribution
• E.g. ANOVA
In both cases: If p < 0.05, targets are significantly different

Figure 4. Schematic, and simplified, overview of the steps that are carried out in order to establish whether some target's distribution of detection times, harvested during our observer trial methodology, are significantly different or not.

3.3 Finding the overall result and ranking in a trial

In order to find the overall result for each target (over all scenes), we did the following:

- A normalization of the median detection time for each target in each scene, representing the performance of the target relative to the other targets in a particular scene. A numeric value above 1.0 then reflected signature performance above average, whereas a value less than 1.0 reflected the opposite. Such an approach also accounted for the relative difference (and not just their order) of the test targets in a ranking.
- An assigned weight (higher, equal to or less than one) for each scene.

4. RESULTS – EXAMPLES FROM A SCENE

For illustration purposes we present now an example of how the distribution of detection times turned out for the 6 test targets (T0-T5) in one particular scene (forest background at Kjeller, Norway) during our recent trial campaign. Figure 5 shows detection times of 148 observers, distributed among the 6 targets. The red squares indicate the median detection time for each of the targets, and we note that it varied with a factor up to about 4 (target 5 vs target 4). The numbers in the horizontally oriented rectangle, right above the distributions of detection times, show the corresponding number of non-detections for each target. A high number of non-detections generally indicated that the target was difficult to detect within the time limit of 60 seconds.

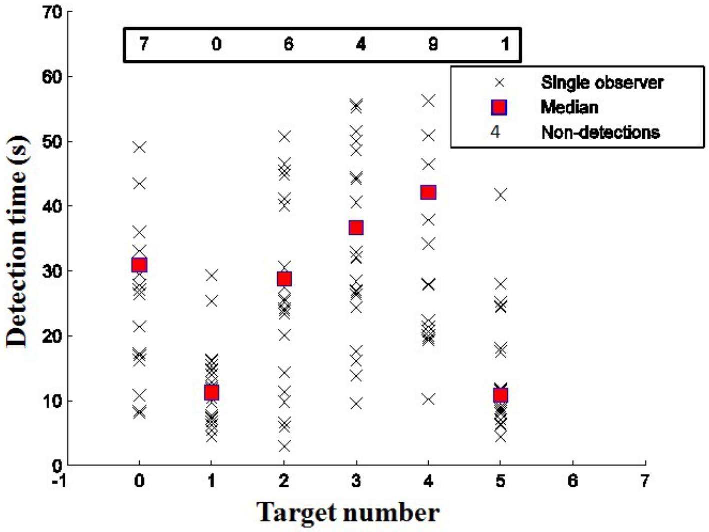

Figure 5. Distribution of detection times for 6 different camouflaged targets for one certain natural scenery. The targets were all located identically in the scenery and recorded by digital camera prior to the observer trial.

In Table 1 we illustrate which combinations of targets that were significantly different ($p < 0.05$) when distributions of detection times were tested by using the Kruskal-Wallis approach. The paired combinations of targets that are highlighted in red, show that the distributions of detection times of the corresponding targets were significantly different. As an example, we see that T1 had significantly different distribution of detection times to all the remaining targets, apart from T5. This corresponds well with the visual impression of the detection times distributions given in Figure 5 above.

Table 1. Overview of which targets that turned out significantly different in camouflage effectiveness when compared to the other test targets by the Kruskal-Wallis approach.

	T0	T1	T2	T3	T4	T5
T0	▨	🟥				🟥
T1	🟥	▨	🟥	🟥	🟥	
T2		🟥	▨			🟥
T3		🟥		▨		🟥
T4		🟥			▨	🟥
T5	🟥		🟥	🟥	🟥	▨

5. DISCUSSION

In this paper we have described the steps in a systematic search by photo evaluation method, based on human observers. This methodology makes it possible to compare signature properties of different targets in a controlled and reproducible way. Furthermore, our methodology is, through our suggested statistical analysis procedure, capable of an evaluation and ranking of the targets by a procedure that, in our opinion keeps most of the valuable data on the targets' signature properties.

One important feature of the observer assessment trial, described in this paper, is that it is capable of telling us the relative strengths among targets. This is illustrated by the relative difference in value of the medians (red squares) in Fig 5. This means that not only do we get to know whether some target performed significantly better than another, but the method also serves us with numbers on *how much better* it was. Such kind of relative comparison allows the adding of results from several scenes without losing valuable information about the signature suppression effectiveness of the targets.

The observer trial method brings the natural scenes to the human observers and not the opposite, which can be cost-effective. Furthermore, we believe the observer trial methodology can be cost-effective compared to other tests of signature properties such as image sequences at different target distances (requires helicopter [13]) or transportation of a large number of soldiers for signature testing in the field (several different locations, at different seasons), but still reliable and easy to use. As an example the methodology allows signature solutions intended for winter and summer to be tested simultaneously, and is therefore quick. Furthermore, the software, that was developed to run the trials, enables a rapid, first order evaluation of the targets already during the observer trial.

We believe the observer assessment method will be suitable for test and evaluation not only of camouflaged combat suits (as was shown in Fig 5), and not only in the visual band of the electromagnetic spectrum. As the method is founded on photographs of targets in a natural setting, ranking them based on their corresponding detection time, we may picture this approach to be feasible also for thermal images, where targets can be ranked accordingly. Assessment of thermal patterns, induced by patches with different emissivities has been carried out earlier [24]. However, a test of any thermal application of our method needs to be tested out further.

5.1 The importance of rank criteria

There is no gold standard regarding the optimal criterion parameter for evaluation and ranking of signature properties. In our method we have chosen detection time as our primary criterion, with probability of detection as a potential secondary. Any secondary criteria will act as a minor control (albeit not independent) of the rankings based on the primary criteria. Whenever a rank based on these two criteria are in conflict, the primary should be emphasized. Still, such conflicting issues will act as a reminder to the test team to take a closer look at the raw data for further analysis if possible. The choice of ranking parameter is inextricably intertwined with set up and purpose of the test. In our methodology targets are recorded (e.g. photographed) in a natural setting, which commutes well with criteria such as detection time and probability of detection, but rules out detection range, even if the latter can be found well suited as ranking parameter in other test set-ups [13, 18]. The choice of detection time over probability of detection is a decision we had to make ourselves, and our rationale was the following:

In our opinion the detection time, being a continuous variable, will capture more of the signature properties of the targets than probability of detection which is binary (“yes” or “no”). The probability of detection may depend too much on the criteria for how non-detections are registered, and is more sensitive to the test set-up as we see it. A distribution of detection times, as we retrieved by our methodology, allows both test of significant differences among targets, and does also capture *relative* differences in effectiveness to a larger extent than binary variables as we see it.

Finding overall result and ranking of targets

The overall best target, in our study, was found by adding the results and their corresponding weights over all scenes. As the individual N scenes were (spectrally and structurally) very different as well as in level of difficulty, we found large spread in characteristic detection times (i.e. median) among the scenes. To be able to sum up, we calculated normalized medians (to the average median in a scene) for each scene, and weighted the scenes equally. In our opinion, a normalization of the medians, as described above, preserves the valuable information of the *relative* strength among the targets to be evaluated. The normalized medians can then finally be summed up over all scenes, to give an overall performance value for each target in our study.

We believe this approach (normalized medians) allows for a more realistic description of a group of target’s signature effectiveness in a number of different natural backgrounds potentially making improvements to earlier works where mean detection time over all scenes was calculated or where the mean hit rate (i.e. probability, binary variable) has been found [3, 25]. Interestingly, this issue touches some of the core questions regarding signature evaluation, and we do not dare to state that detection time should be the preferred ranking variable in all search-by-photo trials, involving human observers. Furthermore, plotting the primary rank parameter (in our case: detection time) versus some secondary parameter (e.g. probability of detection) in a phase plot, as was done with two other rank parameters in a similar study by Toet et al. [25], can strengthen the assessments further.

Sometimes it will be important to discriminate between high performing targets and poorly performing targets, as also discussed in similar studies [25]. Our methodology enables that important feature, by the preservation the relative measures in effectiveness, and we see from Fig 5 that target 0, 2, 3, and 4 belonged to the category of well performing targets, whereas 1 and 5 belonged to the category of poorly performing targets in that particular scene.

When performing an overall ranking among various signature targets, the default approach would be to identify the target with the overall highest score. However, it might happen that this particular ("winner") target performed very well in 7 of 9 scenes, and poor in 2. Will it then be wise to consider it better than another target that performed average or better in all 9 scenes, but with a slightly lower overall score? Our method does not give an answer to that, other than if the scenes are all considered to be of high relevance, one should try to avoid that a preferred and recommended solution performs poor in parts of its core operational area.

Handling non-detections in our method

The non-detections harvested in our trial, due the observer's search time running out, indicated that the target of evaluation was hard to detect. A high number of such non-detections for a certain target will be valuable to us in a ranking of that target relative to another test target. On the contrary, if the miss-detections (i.e. clicking at the wrong location in a scene) posed a large fraction of the total number of non-detections for a certain target in a scene, this would lead to a ranking on false premise. The particular target can then be assigned too good signature properties as the reason that it was not detected was due to the observers' incomplete search rather than good signature. However, it can be difficult to separate these two effects from one another, as illustrated by the following example: A miss-detection after two seconds surely should be assigned little value as the observer most likely was jumping to conclusions. On the other hand, a miss-detection after 55 seconds (of a total search time of 60 seconds), contains valuable information as the target was searched for a long period of time and still was not detected.

As our methodology stands today, it does not automatically differ between the types of non-detections (other than by a manual inspection of observer data). However, there are potential approaches that possibly solve this issue:

- Rule out all miss-detections that are unreasonably quick in time. This could for instance be all miss-detections less than 10 seconds in a scene where the maximal search time was set to 60 seconds. However, there is no fool-proof recipe to follow in this manner.
- Rule out all miss-detections by not accounting for mouse clicks that are not within the target tolerance.

5.2 Choosing scenes and their corresponding weights

It is obvious that the higher the number of scenes, the lower impact a single scene has on the final outcome. Whenever a signature test involves several kinds of terrain backgrounds (wood, open, dry, wet etc) considered to be of operative importance, we find it vital that scenes of each kind should be included in the test for reducing the sensitivity of individual scenes on the final ranking. Not only will the number of unique scenes be important, but equally much the local background surrounding the target as the revealing contrast between target and background is generated therefrom.

Furthermore, it is important that the test conditions are as identical as possible for each of the targets as they are to be compared to one another (same target area exposed, stable illumination conditions, each target placed at exactly the same spot and with controlled orientation). It is important that the observer is tested against *only one* of the test targets in each scene. This is due to the fact that all the targets in a scene were located at exact the same spot in the local background. Thus, detecting one target would automatically lead to an immediate (and strongly biased!) detection, by one unique observer, of all other targets in this particular scene. Our choice of placing all targets in identical position in a scene also reduces the image edging effect on the target rankings. It has been reported that observers tend to focus on the center of an image [17], resulting in differences in detection performance by the observers [2, 17, 19].

Scene images were shown in a random order for each observer. We chose this approach to minimize an overall (over all observers) effect of learning. Learning has been reported to occur [19] for individual observers as the search procedure will be similar in each of the scene images in a sequence, and hence a learning effect cannot be omitted completely. As our scenes were unique and different, we believe there was little or no effect of familiarity between them, which can affect observers by recognizing local backgrounds in a scene where targets are likely to be – or not to be – located [4,19].

Finally, for the observer trial to work optimally, the targets in the pictures should always be possible for the observer to detect as long as the eye is focused at the correct spot in the image surface. Hence, the observer should not be in doubt, when focusing at a target, whether it is actually a target or not. This reduces the undesired effect that the observers making guesses during the trial. Finally, it will most often be beneficial that there a spread in physical distance to the target among the scenes. This allows different aspects of signature to be tested as well as preventing the observers to expect the same target distance in each image of the sequence.

In our recent trial, where camouflage patterns were assessed in the visual spectrum, we chose each scene to count equally. The simplest solution in this context is to leave all scenes unweighted especially if it is not obvious that some scenes are less important than others and all are operationally relevant.

The role of empty scenes

Some scenes were chosen as "empty scenes" (a natural background with no target. This was not vital for the trial to be conducted, but we believe it may have the effect that the observers had to verify, at least for themselves, that a potential target was real. Thus, it may strengthen the value of the result.

5.3 The importance of illumination conditions

Of particular importance for the quality of the observer trial is the illumination conditions during the photography in the field. For a visual based observer trial, as an example, changes in illumination conditions will alter the luminance contrast between target and background [15,16]. We therefore emphasize our opinion that all targets (within a certain scene) must be photographed under as identical conditions as possible. The illumination conditions need not be similar from one scene to the next. On the contrary, different types of illuminants (clear sky, overcast, target in full shade etc) among the scenes in a full trial will broaden the validity of the ranking of targets as it will cover more than one standard illuminant. Important, too, is that such an approach minimizes the risk of a reduction of a target's signature effectiveness due to an increase in contrast between target and background, induced by metamers, when switching from one illuminant to another [15,16]. From a tactical point of view, a location in shaded areas will often be preferred and such positioning of targets should also be included in a signature test. In the trial that was carried out by the authors, stable illumination conditions (in a scene) were ensured by a rapid recording of the unique targets, combined with a visual inspection of images afterwards. An improvement to our approach would be to perform a calibration of the scene images by using a colour board as a reference [26].

5.4 Controlling variations in observers' prerequisites

A spread in the observers' qualifications will always be present. This may influence on how the detection times data are distributed (see Fig 5 for an example), and - of more importance – may influence on the final conclusions of the observer trial as well. As it is difficult to estimate how different the skills and prerequisites among the observers are, we recommend the following steps when a trial is carried out:

A large number of observers (per target per scene) is recommended as then eventually the statistical fundament of all individual targets will be sufficiently identical. However, such high numbers of observers are not achievable in many cases and variations among the observers must be controlled. Therefore, a most homogenous group of observers is advisable, for example recruit soldiers as we used in our study. Still, observers may have very different prerequisites (*e.g.* of a personal character) for the trial. Some are easy to jump to a conclusion, whereas others are very thorough and will not indicate detection (by mouse-clicking) until they have been given it some thought. It has also been reported, from similar studies, that tactical knowledge of the observers influenced on their detection skills in a trial [27] as well as a learning effect of observers during a set of images of scenes [19].

We cannot rule out the possibility that the trial itself may induce stress for the observer as he or she has to perform under surveillance of an instructor. Such extremes will occur from time to time and are difficult to foresee and nearly impossible to correct for when the trial is over. On the bright side, however, each observer will be exposed to *different* targets from one scene to the next, and the order of scenes will be randomized among the observers. Therefore, any differences in the observers' prerequisites will be smoothed out if the trial contains a high number of distinct scenes.

5.5 Future possibilities

One aspect that we see as beneficial would be to introduce some independent control-test to the human based observer trial. Although the observer trial, as it is today, does not seem to have too many obvious pitfalls, we still believe it would be valuable to run some independent control-test in parallel with the main observer trial. At the moment, we do not have such a method, although a software tool known as CAMAELEON seems to be a candidate, at least within the VIS and NIR spectrum [7]. It has been used with apparent success in our recent camouflage study [28] as well as in others [29]. We found a good correlation between the rankings by humans and CAMAELEON, although more work is needed to establish the strength of such a correlation. Ideally, we believe the control-test should assess the test targets with different parameters than our assessment trial (where detection time is used). How to ensure the observer trial commutes well with the reference test is not obvious, but as a rough guide we may say the following: As long as the reference test results in the same trend in the ranking as in the main trial, it brings much value to the strength of the recommendations, ruling out the possibility of some un-known systematic error in the trial itself.

It would also be interesting to extend the methodology so that it not only tests signature effectiveness of a group of existing targets, but also allows for a generation of new targets (with new, and pre-optimized, signature properties) as suggested in similar studies within camouflage [3, 30]. Finally, we might want to test out the trial methodology in the near-infra red (NIR) or thermal part of the electromagnetic spectrum, following the ideas of Bobo et al [24].

6. CONCLUSIONS

In this paper we have presented an observer based methodology for evaluation of signature effectiveness based on optical images in the visual spectrum. The methodology allows for relative comparisons amongst test targets by logging detection times of targets located in operatively highly relevant scenery. Not only does the methodology tell us which target that is assumed to be the best in overall, it also tells us *how much better* one target is compared with another.

Finally, the methodology is cost-effective, reliable and easy to use which allows for rapid testing of signature properties, in most areas where digital scene imagery can be captured.

ACKNOWLEDGEMENTS

The authors would like to thank C. Jones and T. Evans, DSTL, for valuable support during preparations of the observer trial. The authors would also like to thank Morten Søderblom for valuable discussions and Keith Cooke for proof - reading parts of the manuscript as well as for his useful inputs.

REFERENCES

[1] Gretzmacher, F. M., Ruppert, G. S. and Nyberg, S., "Camouflage assessment considering human perception data ," Proc SPIE 3375 (1998).

[2] Chang, C. C., Lee, Y. H. and Lin, C. J., "Visual assessment of camouflaged targets with different background similarities," Percept Motor Skills, 114, 527-541 (2012).

[3] Toet, A. and Hogervorst, M. A., “Urban camouflage assessment through visual search and computational saliency,” Opt Eng 52 (2013).

[4] Toet, A., Bijl, P. and Valeton, J. M., “Image dataset for testing search and detection models,” Opt Eng 40(9), 1760-1767, (2001).

[5] Boult, T. E., Micheals, R. J., Gao, X. and Eckmann, M., “Into the woods: Visual surveillance of noncooperative and camouflaged targets in complex outdoor settings,” Proc IEEE 89(10), 1382-1402, (2001).

[6] Tamura, H., Mori, S. and Yamawaki, T., “Textural features corresponding to visual perception,” IEEE Trans Syst Man and Cybern, 8, 460-473, (1978).

[7] Hecker, R., “CHAMELEON-CAMOUFLAGE ASSESSMENT BY EVALUATION OF LOCAL ENERGY, SPATIAL-FREQUENCY AND ORIENTATION,” Proc SPIE 1687, 342-349, (1992).

[8] Kilian, J. C. and Hepfinger, L.,”Computer based evaluation of camouflage,” Proc SPIE 1687, 359-369 (1992).

[9] Birkemark, C. M., “CAMEVA, a methodology for computerised evaluation of camouflage effectiveness and estimation of target detectability,” Proc SPIE 3699, 229-238, (1999).

[10] Nyberg, S. and Bohman, L., "Assessing camouflage using textural features," Proc. SPIE 4370, 60-71 (2001).

[11] Nyberg, S. and Bohman, L., “Characterizing low signature targets in background using spatial and spectral features,” Proc SPIE 5152, 139-149, (2003).

[12] Houlbrook, A. W., Moorhead, I. R., Filbee, D., Stroud, C., Hutchings, G. and Kirk, A., “Scene simulation for camouflage assessment,” Proc SPIE 4029, 247-255, (2000).

[13] Schoene, R., Meidow, J. and Mauer, E.,”Feature evaluation for target/background discrimination in image sequences taken by approaching sensors ,” Proc SPIE 7697, (2010).

[14] Jones, C., “Now you see them – now you don’t,” Desider, 10 Feb 2010, www.gov.uk/government/uploads/system/uploads/attachment_data/file/33843/desider_22_Feb2010.pdf

[15] Ohta, O. and Robertson, A.R., [Colorimetry], John Wiley & Sons Ltd, West Sussex, 92-93 (2005).

[16] De Marsh, L. E. and Giorgianni, E. J. “Color science for imaging systems,” Physics Today 42, 44-52, 1989.

[17] Mannan, S. K., Ruddock, K. H. and Wooding, D. S., “The relationship between the locations of spatial features and those of fixations made during visual examination of briefly presented images,” Spatial Vision 10(3), 165-188,1996.

[18] Peak, J., Hepfinger, L., Balma, R., Christopher, G., Fleuriet, J., Honke, T., Huebner, G, Mauer, E, Dotoli, P, Ronconi, P. and Jacobs, P., "Guidelines for camouflage assessment using observers," AG-SCI Rapport 095, 2006.

[19] Toet, A., Bilj, P., Kooi, F. L. and Valeton, J. M. “A high-resolution image data set for testing search and detection models,” TNO-report TM-98-A020 1998.

[20] Williams, L. G., “Target conspicuity and visual search,” Human Factors, 8, 80-92 1966.

[21] Cooke, K., “The sources of variability in the search process,” Search and Target Aquis., RTO-MP-45, 14-1, (2000).

[22] Bickel, P. J. and Doksum, K. A., [Mathematical statistics - Basic ideas and selected topics], Holden-Day, Oakland, USA, 344-390 (1977).

[23] Sawilowski, S., S. “Misconceptions leading to choosing the t test over the Wilcoxon Mann-Whitney test for shift in location parameter,” J Mod Appl Statist Method 4, 598-600 (2005).

[24] Bobo, G., Gonda, T. and Bacon, F., “Thermal camouflage pattern prediction using PRISM and PMO, “Proc SPIE 4370, 84-93 (2001).

[25] Toet, A. and Hogervorst, M. A., “Design and evaluation of (urban) camouflage,” Proc SPIE 7662, (2010).

[26] Jones, C., DSTL (https://www.gov.uk/government/organisations/defence-science-and-technology-laboratory), personal communication, April 2013.

[27] Ruppert, G. S., Beichel, R. and Gretzmacher, F. M., “Robust measure for camouflage effectiveness in the visual domain,” Proc SPIE 4029, 286-393, (2001).

[28] Heinrich, D. H. and Selj, G. K., “The effect of contrast in camouflage patterns on detectability by human observers and CAMAELEON,” Proc SPIE DSS (submitted), (2015).

[29] McManamey, J. R., “Validation plan for the German CAMAELEON model,” Proc SPIE 3062, 300-310, (1997).

[30] Friskovec, M., Gabrijelcic, H. and Simoncic, B., “Design and Evaluation of a Camouflage Pattern for the Slovenian Urban Environment,” J Imag Sci Technol 54(2) (2010).

*gorm-krogh.selj@ffi.no; phone 004763807615; www.ffi.no