AN INVESTIGATION INTO THE APPLICABILITY OF COMPUTER-SYNTHESISED IMAGERY FOR THE EVALUATION OF TARGET DETECTABILITY

M. Ashforth
Defence Clothing & Textiles Agency
Science & Technology Division
Flagstaff Road
Colchester
Essex
CO2 7SS
United Kingdom
E-mail: mlashforth@dcta.demon.co.uk

1. SUMMARY

In the course of an earlier study of the influences on an observer's performance in target detectability assessments, the statistical analysis of the data suggested that there was a difference in the influences at work on an observer between the detection of targets in a real scene and the detection of targets in computer-generated (synthetic) images. Since synthetic imagery is increasingly used in this field, this is an important result. The work described in this report is a further analysis of the original data with the aim of studying more closely this difference. Analysis showed that there is indeed a marked difference between the influence of the observers' visual acuity on their performance in the two types of detection task. The reason is that there is less detailed clutter in synthetic images, which alleviates much of the decision-making an observer has to undergo in detecting a target in a real-scene image. In the synthetic case, the target is either seen or not seen and there is much less uncertainty. This uncertainty, which attends real target detection, swamps any measurable influences on an observer's relative performance in the real-scene case. The conclusion is that computer-generated images used for the evaluation of low-contrast target detection should contain much more clutter detail than at present.

Keywords: Target detection, camouflage evaluation, observer tests, visual acuity, synthetic imagery, visual perception.

2. INTRODUCTION

Evaluation of the effectiveness of camouflage, or, more generally, the measurement of the detectability of low-contrast targets in a cluttered environment, is not a trivial task. Although there are models of human perception, they are at present limited in their applicability, and the case of low-contrast targets in a cluttered environment is the most difficult. Many unquantifiable influences are at work in a human search for inconspicuous targets.

For this reason, the NATO camouflage research community has always relied on the use of numbers of human observers in their evaluation of camouflage effectiveness. This has usually involved photosimulation tests (ref. 1), whereby observers are shown projected photographic images within which a target is concealed. The simulated range at which the target is detected becomes the variable to be tested in the subsequent statistical analysis, whereby individual camouflage measures can be evaluated and compared. Despite the various problems and inadequacies of the test (ref. 2), this remains the most reliable method of camouflage evaluation.

In recent years, computers have made it easy to construct images of targets that do not exist, such as new vehicles in development, or to construct images which are less variable than are real scenes, so that one parameter at a time (e.g. gloss) can be varied, to evaluate its effect on target detectability. These possibilities offer the prospect of an improvement in the method of photosimulation by removing the variability found in real imagery, such as that caused by variations in imaging position, natural illumination, and so on, and also by allowing measurement of the effect of otherwise minor influences on target detectability.

Implicit in the use of computer-generated images in this way is the assumption that the search task for the human observers is the same as for a real scene. Therefore an analysis of observers' performance on computer-generated imagery should show a correlation with their performance on real imagery. An opportunity to test this hypothesis arose during a photosimulation exercise held at the Defence Clothing and Textiles Agency (DCTA) Science and Technology Division (S&TD), in Colchester, United Kingdom, recently.

3. DESIGN OF THE PHOTOSIMULATION TEST

The photosimulation test was set up primarily to evaluate developmental camouflage measures within specific projects, such as for helicopters; for hot, arid environments; and so on. The opportunity was taken to make measurements of other observer-specific attributes that may affect the performance of each observer relative to the pool of observers. It had been hoped that this would enable any quantifiable influence on observer performance to be accounted for, and thereby limit spread in the detection data generated in the photosimulation test.

Some of the imagery used in the test was computer-generated. Although it was not considered at the design stage, this meant that the test also lent itself to the analysis of any difference between real and computer-generated imagery in terms of the dependence of observer performance on any of the measured attributes.

The choice of attributes to include was restricted to those which were intuitively likely to influence observer response and were easy to measure. A brief questionnaire was designed to record details of the observers' age, rank, relevant training, and their normal job within the unit. Tests were devised, with advice from a local optometrist, to measure visual acuity with a Snellen Chart and colour perception with a series of Ishihara Colour Plates.

Paper presented at the RTO SCI Workshop on "Search and Target Acquisition", held in Utrecht, The Netherlands, 21-23 June 1999, and published in RTO MP-45.

Past experience had suggested that some observers were consistently "good" or "bad" in their ability to detect targets in the recorded image. There had been evidence (Annex D of ref. 3) to suggest that observer ability could be accounted for by adjusting the raw data according to how well an observer performed relative to the other observers, and that spread could be reduced as a result.

Observers are familiarised with the nature and procedures of the photosimulation experiment by being shown a pre-test image similar to those that will be shown in the experiment proper. If all observers were shown the same image, and their performance on this image was recorded, it should give a guide to their relative ability. Therefore the final factor to be incorporated was the observer's performance on this pre-test image. This would have the disadvantage that the observers would be learning the procedure at this stage, but the advantage that all observers would see this same image before any of the others, so all saw it under equal terms.

Of the four sets of imagery used in the photosimulation, one consisted of computer-generated imagery. Because the observers were likely to be less familiar with this type of imagery than with real-life photographs, it was decided that the familiarisation image should be computer-generated too.

The photosimulation test was designed so that each observer saw several slides (taken in different locations), some of which had more than one target. This provided data for between 5 and 7 target detections per observer, of which one was in a synthetic image, plus the familiarisation image (also synthetic) that all observers saw. In analysing these detections individually, the assumption is made that they are independent (i.e. one detection does not influence another in the interactive cueing effect). This is not always the case for slides containing more than one target, but no trends were noticed that might have suggested that detections were not independent. Unfortunately 5 to 7 is not a high enough number to conduct a test on the independence of target detections.

4. PRE-TEST DATA

A total of 104 observers were conducted through the trial, all of whom were army personnel from the Colchester Garrison. Their questionnaire responses were coded for entry into an analysis of variance, which would establish how significant each factor was in its contribution to the variance observed in performance. Reference to individuals was made by their Observer Index, which was the number given according to the order in which they were conducted through the whole test. Age was recorded as a whole number of years. Military rank was coded with an integer to represent each level. The military unit to which the observers belonged was recorded, as was the category of job each performed within that unit.

Visual acuity was measured under test conditions and codified in a way suitable to the statistical analysis. Two observers were considered outliers in the visual acuity data. Both of these observers normally wore spectacles, but did not have them available for the test.

Seven of the 104 observers had defective colour vision, and were diagnosed according to the type and degree of deficiency. From a statistical viewpoint, however, so small a sample could not be further subdivided. Colour vision was therefore characterised simply as normal or abnormal. The last category recorded for each observer was the amount of relevant training he had received. All appropriate training was recorded on the questionnaires and was graded by the supervisors with a subjective score out of ten for relevance to the photosimulation task.

5. PHOTOSIMULATION DATA

In order to make detections of different targets comparable, each observer's detection range for a given target was normalised with respect to the mean and standard deviation of all detections made on the same target, as follows.

Z-score for observer against target =

$$\frac{\text{observer's score} - \text{mean score for target}}{\text{sample standard deviation for target}}$$

Thus the Z-score is the amount by which the observer's score exceeds the mean score in units of the standard deviation. A positive Z-score represents a better-than-average result and a negative Z-score represents a worse-than-average result. This removes the differences that exist between the detection difficulty of different targets and allows a comparison to be made of the performance of each observer, relative to the relevant subgroup of observers, i.e. those who detected the same target.

Consistently good observers would be expected to get consistently high Z-scores, so the mean Z-score, averaged over all targets seen by each observer, should be an indication of that observer's ability to detect targets in photosimulation. This, along with the Z-score of the familiarisation slide result, makes two independent measures, designed to be of the same thing.

6. STATISTICAL TESTING

A regression analysis was conducted to determine the correlation between the two sets of Z-scores. A high correlation would confirm that the familiarisation test gives a guide to the ability of the observers. The resulting correlation coefficient was 0.165, which for samples of this size is significant at the 90% confidence level, but no higher. This is not very high and does not give much confidence in the usefulness of the familiarisation slide results as a monitor of observer ability.

Further tests that were conducted to evaluate the effect of the different attributes on observer performance highlighted more differences between the mean Z-scores and the familiarisation Z-scores. These were principally analysis-of-variance (ANOVA) tests, designed to show which of the factors under consideration were contributing to the variance in simulated detection range.

Table 1: ANOVAs on Z-scores (101 Observers)

Factor	df	p (Mean Z)	p (Fam Z)
Age	1	0.432	0.911
Rank	1	0.610	0.637
Unit	3	0.349	0.108
Job	3	0.596	0.021
Colour Vision	1	0.685	0.863
Visual Acuity	1	0.010	0.001
Training	1	0.387	0.269
Error	89		

The three observers who came from training units had to be excluded from the ANOVA because they formed too small a data subgroup. This left 101 observers in the data set. Table 1 shows the results of two separate ANOVAs on the mean Z-scores and the familiarisation Z-scores respectively. This gives a comparison of the relative contribution of each of the

factors to the variance in observer Z-score between the overall mean of the 5 to 7 target detections (the column headed p(Mean Z)) and that for the familiarisation slide (headed p(Fam Z)). The figure in the "df" column gives the number of degrees of freedom for each factor within the analysis. The error term refers to the residual variance. The figures in the "p" columns are the significance levels for each factor: less than 0.05 denotes a significant result, i.e. that the factor has a significant effect on the observers' Z-scores.

Most of the factors included in the analysis have not had a significant influence on either of the sets of Z-scores. In the column for mean Z-scores, only visual acuity has shown a significant effect. It is obvious that in the broadest sense visual acuity will be significant, because if an observer has very poor eyesight, he will not be able to distinguish the targets at all. However, people with very poor eyesight are unlikely to be of interest in a simulation of military target detection and the reason for including this factor was to see if there was an influence even among observers with good eyesight, as mainly used here. There are two observers within the pool who are outliers in the distribution of visual acuity, and they will be exercising a large leverage on the data and its analysis. To check this effect they were removed from the analysis, which was conducted again, exactly as above, but now on the remaining 99 observers. Table 2, below, gives the results of this second analysis.

Table 2: ANOVAs on Z-scores (99 Observers – Visual Acuity Outliers Removed)

Factor	df	p (Mean Z)	p (Fam Z)
Age	1	0.337	0.933
Rank	1	0.526	0.421
Unit	3	0.384	0.060
Job	3	0.667	0.026
Colour Vision	1	0.700	0.983
Visual Acuity	1	0.297	0.001
Training	1	0.400	0.205
Error	87		

Some of the figures in the table have changed, most notably the visual acuity figure for the mean Z-score column, but, importantly, not the visual acuity figure for the familiarisation Z-score column. This is the result that first highlighted the possibility of a difference between the requirements of a search of real imagery and that of synthetic imagery.

Removal of the visual-acuity outliers had the expected effect on the analysis of mean Z-scores, i.e. it removed the apparently significant influence of visual acuity on observer performance (within the narrow spread of visual acuity scores still in the analysis). Remarkably, the same effect was not apparent in the analysis of familiarisation Z-scores; a very significant influence remaining. Note also the other two apparently significant effects; "unit", at 90% confidence; and "job", significant at the 95% confidence level.

If there really is a difference between the requirements of real and synthetic imagery searches, then a closer correlation would be expected between the familiarisation Z-scores and the synthetic-imagery photosimulation Z-scores than that measured earlier between the familiarisation scores and the overall mean ones. This is easily tested. The correlation coefficient for familiarisation Z-scores against the synthetic imagery Z-scores was 0.305, which is significant at the 99.8% confidence level. This is therefore a much more significant correlation than was found with the overall mean results.

Further, if this highly-correlated set of results formed part of the data making up the overall means, then another correlation test should be conducted on the familiarisation Z-scores against the mean of all real-scene Z-scores (that is all except the synthetic-imagery scores). This produces a correlation coefficient of 0.085, which equates to a confidence level of 61%, i.e. not at all significant, or no correlation.

This is a striking result. There is no correlation between the relative performance of observers on the familiarisation slide with that on the 6 real-scene targets, but there is a high correlation with their performance on the other synthetic-image target.

7. DISCUSSION

The statistical work has proved that there is an important difference between target detection from real-scene imagery and detection from computer-generated imagery. This difference has been detected through the relative performance of observers in the target detection task. This infers that some observers are particularly good at detection of targets in real scenes and others are better on synthetic imagery. There must, therefore, be a difference in the demands of each.

The analyses of variance, reported in Section 6, gave a clue when they produced different figures for the significance of the influence of various factors on observers' relative performance. The most notable difference was recorded in the case of visual acuity, which, for the limited spread of acuity found in the 99 observers tested, was not a significant factor in observer performance on real-scene imagery, but was highly significant in the case of synthetic imagery. This implies that detection of targets in synthetic imagery demands good visual acuity, more than does detection of targets in real-scene imagery.

This can be tested specifically, by calculating the correlation coefficient between the visual acuity score and both the mean Z-score for real-scene imagery and the mean Z-score for synthetic imagery. Table 3 shows the results of such an analysis.

Table 3: Correlation of Visual Acuity with Z-Scores

Image Type	Correlation Coeff	p
Real Scene	0.147	0.137
Synthetic	0.381	0.000067

This is an emphatic result. The "p" column gives the probability that the correlation coefficients given could occur by chance if there was no real correlation. It is therefore the significance figure. Within the range covered (by all 104 observers), visual acuity has no significant correlation with the observers' performance in detecting targets in real-scene imagery, even at the 90% confidence level (which would require that $p<0.1$). By the same token, visual acuity is significantly correlated with observer performance in synthetic-imagery target detection at the 99.99% confidence level. Visual acuity would therefore seem to be the main cause of differences in observer performance between the two types of imagery.

There was a suggestion evident in Table 2 that "job" and "unit" may also contribute something to the difference between observers' performance on real and synthetic imagery. One way to test this is to run single analyses of

variance on each data set for each of these two factors. This would produce significance values for each effect. The resulting values are shown below in Table 4.

Table 4: ANOVA for "Job" and "Unit"

Type	p(unit)	p(job)
Real Scene	0.664	0.520
Synthetic	0.555	0.093

The non-significant figures for "unit" suggest that the slightly-significant result in Table 2 (p=0.060) was a rogue. Such a value would be expected by chance roughly once in twenty occasions, so this is quite likely, given the number of tests conducted. The new results above are more reliable than the one in Table 2 because all of the data are used here, whereas some elements had to be removed to do the earlier multiple ANOVA.

Note, however, that there is still a minor difference apparent in the data for "job". There is no significance at all in the effect of "job" on the real-scene data, whereas 0.093, for the synthetic-image data, represents a significant result at the 90% confidence level, though this is not very high and could have occurred by chance.

It would appear that visual acuity is the factor that accounts for almost all of the difference between the demands of real and synthetic imagery in the search for inconspicuous targets. Comparison of the visual appearance of the two types of imagery is necessary in order to attempt to explain this difference.

The reason for the difference is probably that the artificial scene was very homogeneous, using a large number of almost identical-looking trees with a very plain "grass" base. There were few opportunities to be mistaken about the target's whereabouts: it could either be seen or it could not. In real imagery, trees and bushes differ more. There are shady clumps that can look like a camouflaged vehicle. There is much more scope to be mistaken.

In other words, the visual acuity is much more important in synthetic imagery, because there is very little other decision-making to do. When a target is found, it is found with some certainty. In real imagery, there may be many potentially "false" targets, and the observer has to decide how certain he is that he has indeed found a real target. In this case, though visual acuity might be equally important as in the former case, it is swamped by the vagaries of human decision-making in the detection data. Indeed, for real-scene imagery, no factor has been shown in this investigation to have a significant effect on the performance of an observer relative to the pool of observers who detected the same target. The "random error" of the decision process is greater than the effect of any of the individual influences considered here.

8. CONCLUSIONS

An important, and potentially far-reaching, conclusion has emerged from work that was originally designed to evaluate the effect of various potential influences on the performance of observers in the detection of low-contrast targets in a cluttered environment. It is that there is a major difference in the influence of observers' relative performance within the group of observers between target detection in real-scene images and that in computer-generated images.

In essence, the problem is that synthetic images are not sufficiently cluttered to simulate the search task presented by a low-contrast target in a real scene. Computer-generated images are increasingly being used in target detectability studies, on the assumption that such imagery is a sufficiently realistic simulation of real scenes. The work reported here throws doubt on that assumption. In particular it has shown that there is a difference in the demand on observers in the detection task, i.e. that visual acuity is more important in synthetic imagery than it is in real-scene imagery.

The effect of this problem in detectability evaluations will be to introduce a bias that would not show in real-scene work. The observers' visual acuity would influence their own performance. The choice of observers and their distribution across comparative groups would need to be done very carefully with regard to their visual acuity, which would of course need to be tested. Alternatively, by measuring the size of this influence of visual acuity, it could in principle be accounted for by adjusting observers' responses, according to their acuity score.

As computers advance in power, so it should be possible to generate more and more realistic synthetic imagery that would approach the degree of clutter found in photographs of real scenes. This work suggests that that position has probably not yet been reached, and certainly suggests that as much realistic clutter as possible should be included in any synthetic imagery intended for use in an evaluation of the detectability of low-contrast targets.

9. REFERENCES

1. Ashforth, M. and Collins, J.H., "Determination of Detection Range by Analysis Recorded Imagery" (Technical Memorandum SCRDE 91/6, NATO AC/243/CCD/WG(D) 1/91), DCTA – S&T Division, Colchester, UK, 1991.

2. Ashforth, M., "Camouflage Evaluation: Improvements in the Conduct and Analysis of Photosimulation" (DCTA S&TD Research Report 96/02), DCTA – S&T Division, Colchester, UK, April 1996.

3. Ashforth, M., "Evaluation of Handling and Camouflage Properties of Commercially Available Camouflage Nets with Regard to Requirements of SCST 005" (DCTA S&TD Technical Report 94/08), DCTA – S&T Division, Colchester, UK, December 1994.