Establishment of Human Performance Baseline for Image Fusion Algorithms in the LWIR and SWIR Spectra

CHRISTOPHER HOWELL
STEVE MOYER

This research is complementary to research presented in "Establishment of Human Performance Baseline for Image Fusion Algorithms in the LWIR and MWIR Spectra" by Moyer and Howell in which we established a baseline performance candidate for image fusion comparison by investigating the impact of different display formats on the probability of identification, P(ID), performance of a human observer. We advance this line of research by measuring the inherent ability of the human observer to perform an identification task using source band imagery, long-wave (LW) infrared and short-wave (SW) infrared that was not algorithmically fused prior to human observation. A multi-part experiment was conducted where human observers were asked to identify displayed military targets using a standard set of tracked military vehicles. The observers performed the identification (ID) visual discrimination task using source band imagery concatenated or presented side-by-side on a single monitor, temporally interlaced source band imagery on a single monitor. Observers' performances using source band imagery fused with the superposition fusion algorithm was also included as a reference because it is a well understood algorithm and shares an assumed similarity with the temporal interlaced display format. This research proposes that a forced choice human perception experiment is a more meaningful evaluation of an image fusion algorithm's performance, specifically when the application of the algorithm is for dimensionality reduction to a single image designed for human observation. The results from this research indentify a clear performance baseline when assessing human observer P(ID) performance employing an image fusion algorithm.

Manuscript received March 19, 2013; released for publication June 14, 2013.

Refereeing of this contribution was handled by Alexander Toet.

Authors' address: U.S. Army REDECOM CERDEC Night Vision and Electronic Sensors Directorate, 10221 Burbeck Rd, Fort Belvoir, Va 22060-5806, USA, E-mail: (info@nvl.army.mil).

1. INTRODUCTION

The performance benchmark process for discriminating between image fusion algorithms presented by Howell [3] and again by Moyer and Howell [6] establishes a performance goal for any aggregate function merging the source band imagery under investigation. It was reported by Moyer and Howell, in "Establishment of Human Performance Baseline for Image Fusion Algorithms in the LWIR and MWIR Spectra," that increased P(ID) performances could be realized dependent upon how the source band information was presented. In that work, human observers were asked to identify displayed targets using a standard set of tracked military vehicles. The observers performed the ID tasks using LW and MW source band imagery concatenated on a single monitor, presented side by side on a single monitor, temporally interlaced source band imagery on a single monitor, and source band imagery presented to each eye of the observer in parallel. The research presented in this paper explored the impact dissimilar source band information had on the observer's ability to identify targets without the aid of image fusion algorithms. The spectral bands under consideration for this effort were the LW and SW spectral bands. It was hypothesized that the different display techniques using dissimilar source spectral band information would better allow the observers to choose the portions of the source band images they needed to perform the visual discrimination task of identification above that achievable using the fused superposition images. It was determined after comparing the observers P(ID) performances using the superposition fused images to the resultant P(ID)s' using these display techniques, that the performances using the superposition fused images were well below that which was achieved by the observers benchmark source band performance. Allowing the observer to view spectral source band imagery in different display formats without the aid of image fusion algorithms, which we refer to as "self-fusion," allows the experimenter to establish an absolute benchmark for discriminating between image fusion algorithm performances.

It was our intent to perform a mirror analysis of the research performed in Moyer and Howell [6] using LW and SW imagery; however the experiment where LW and SW source band imagery was presented to each eye of the observer in parallel could not be completed due to the effects of binocular rivalry [1] caused by the competing information presented in each eye independently. The remaining experiments performed using the LW and MW imagery were repeated in this work using LW and SW imagery.

This paper is outlined as follows: a background section describes some common approaches to image fusion along with some common image quality metrics and their shortcomings regarding predicting human task

performance; a section describing the imagery and experimental set-ups used in this study followed by a section showing the results of each experiment and a section discussing the results; followed by a summary of the conclusions.

2. BACKGROUND

Many military operations require soldiers to perform visual discrimination tasks, such as detection, recognition and identification (DRI) of targets. These tasks are conducted in a wide range of environments and in both daytime and nighttime settings. Information apparent in one spectral band might not be present in another. To this end, the military continuously seeks to improve its imaging capabilities for both day and night operations. As a result, many methods and practices have been employed to assess image fusion algorithm performances with hopes of improving DRI tasks performance [2, 5, and 9]. Combining multi-sensor data onto a single display or into a single image supports the need to provide each soldier or end user with as much relevant and high quality information in the most efficient manner possible.

In general, image fusion techniques can be categorized into three categories or levels: pixel level fusion, feature level fusion and decision level fusion [4]. Pixel level fusion requires an algorithm to first register the source band imagery before combining their information on a pixel-by-pixel basis. Feature level fusion requires fusion of features extracted from the images such as edges or textures to obtain new feature sets. Decision level fusion requires an initial judgment be made on extracted features of a target from multiple sensors and then renders a decision based on the aggregate result whether or not the correct target was identified. Several factors exist that complicate image fusion algorithm assessments, including but not limited to: a lack of proper registration between the source images; the non-linearity with which many image fusion algorithms operate on image data; the absence of an available reference image; and often large disparities exist between spectral bands being fused, types of backgrounds, the target sets and scenes used in comparison studies reported throughout the literature. Because of diverse image characteristics coupled with the lack of a standardized image database, it is a difficult task to identify common requirements and capabilities of image fusion algorithm performance. However, there are three fundamental requirements that should be achieved by any fusion algorithm: (1) the fusion algorithm should preserve as much of the salient information in the source images as possible; (2) the fusion algorithm should not generate artifacts that affect the human observer's ability to perform the task; and (3) the fusion algorithm must be tolerant of imperfections in the source imagery such as noise or improper registration.

Taking into account the fundamental requirements for all fusion algorithms, it seems reasonable then to approach the assessment of image fusion quality by: (1) evaluating the transfer of relevant information content from the spectral source band images to the fused images; (2) quantifying how much degradation or artifacts can exist before human performance is affected. Traditionally, image quality metrics are applied to fused imagery to discriminate between algorithms based on comparisons with other algorithms. The research in this paper places more emphasis on the "self-fusion" performance capability of the human. This "self-fusion" measurement provides an appropriate baseline to compare human task performance using image fusion algorithms. Understanding the human's performance capacity to exploit spectral source band images without the aid of a computer algorithm will ultimately contribute to understanding the relationship between measures of image quality and measures of task performance. Truly understanding the impact of "self-fusion" is necessary to make the best cost decisions and identify the best direction for future image fusion research.

3. DESCRIPTION OF IMAGES AND EXPERIMENTS

The target set used in this study was a standard set of tracked military vehicles referred to as the "8-target set." The "8-target set" was constructed, based on a history of research [8], so that certain vehicles were highly similar to other vehicles while at the same time subsets of vehicles still had distinct characteristics. As a review, the target set consists of the 2S3 and M109 self-propelled artillery pieces, the BMP, M113, and M2 armoured personnel carriers, and the M60, T-62 and T-72 main battle tanks. All observers were trained using the recognition of combatant vehicles (ROC-V) training package [7]. Prior to participating in the perception experiments each observer needed to obtain a 96% probability of correctly identifying the eight previously named vehicles at different aspects in both the reflective and emissive wavebands.

Figure 1 shows a sample of the imagery used in the experiments reported in this paper. The long-wave (LW) infrared and short-wave (SW) infrared spectral signatures for the targets in the scenes appear to be very different and one can immediately see significant differences exist between them. The source spectral bands shown in Figure 1 were selected specifically to test how human observers processed fused imagery from complementary sources.

Figure 2 shows a snapshot of one particular target imaged at its 3 different aspects and all the respective test ranges. The test ranges were distributed between 100 m and 2 km and all targets were imaged at each range. However, publishing specific information regarding the sensors used in this study is prohibited; as a result only relative ranges are reported throughout this

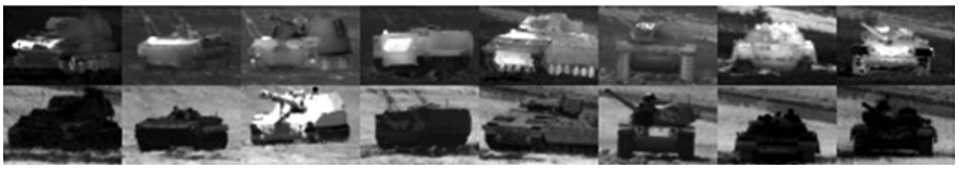


Fig. 1. Sample test imagery: Eight different targets; Two different spectral bands; One range; LW targets top row; SW targets bottom row.

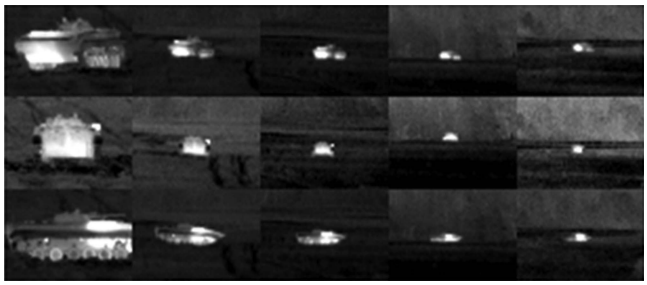


Fig. 2. Single target; three different aspects; five different ranges.

Fig. 3. Example of a concatenated image used in side-by-side experiment, LW image on left and SW image on right.

paper. The interested reader should contact the corresponding author to request a copy of the experimental imagery.

3.1. Side-by-side Imagery

Images were concatenated so that each image rendered on the computer display was the height of each source band image but twice the width of each original image, as shown in Figure 3.

This experiment was designed as an 8-alternative forced choice (8-AFC) experiment to test the human observers' ability to identify targets in each source band image as a concatenated image and then a spectral combination of LW as the left image and SW as the right image. The only time a LW image appeared as the right image was when both images were LW. This was necessary to reduce the size of the experiment, and it was believed that any biases related to testing the SW imagery only on the right would have minimum impact.

The imagery was grouped by range into experimental cells. When the imagery was presented for the experiment, the experimental cells in the experiment were randomized and the images in each experimental cell were also randomized. This ensured that no observers saw the images in the same order.

3.2. Temporally Interlaced Imagery

Source band images were written as individual fields of a movie frame. The movie was played with a field rate of 30 Hz which produced a frame rate of 15 Hz. This simulated a progressive scanned display. Each movie was looped until the observer made a selection. As mentioned in the target set section these images were spatially registered to ensure that no additional blur was added when the movie was viewed. This experiment was designed as an 8-AFC experiment to test each source band image as a 2-field movie. A spectral combination of LW and SW was produced as the other 2-field movie. The movies were grouped by range into experimental cells. When the movies were presented for the experiment, the experimental cells were randomized and the movies in each experimental cell were also randomized. This ensured that no observers saw the same sequence of movies at any range or even the same sequence of ranges. While the movies were being played, there was discernible flickering between the images of the different spectral source bands.

3.3. Algorithm Fused Imagery

This experiment was also designed as an 8-AFC experiment to test each spectral source band image as an algorithmically fused image. The imagery was grouped by range into experimental cells. When the imagery was presented for the experiment, the experimental cells in the experiment were randomized and the images in each experimental cell were also randomized. This ensured that no observers saw the same sequence of images or ranges.

Each image was shown at its native format. However, each image was either a spectral source band image or an image fused using the super position algorithm with a ratio of 0.5. This fusion algorithm was chosen because it is well understood, easily implemented and shares an assumed similarity with the temporal interlaced display format. Additionally, in order to account for experimental variances regarding future image fusion studies, the superposition algorithm results presented in this work can be used as a normalization factor, thereby providing clear performance comparisons between future algorithm research and this research.

4. EXPERIMENTAL RESULTS

In each experiment, each observer's ID performance was calculated for the experiment. This performance probability was corrected for the probability of guessing

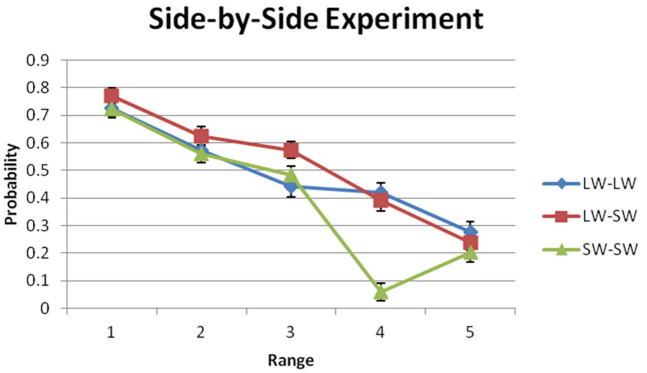


Fig. 4. Experimental results of concatenated (side-by-side) LW and SW spectral images. Error bars are one standard deviation from the mean normalized to the square root of the observer population.

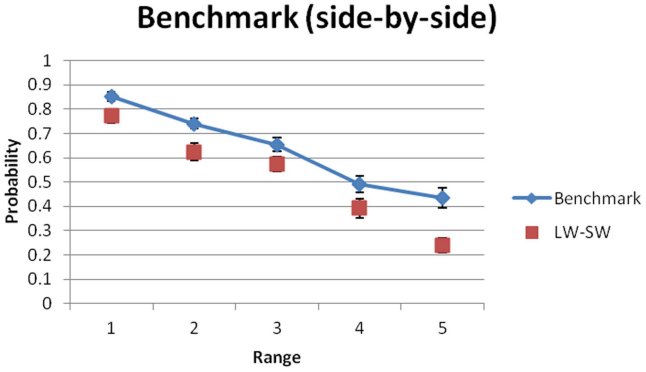


Fig. 5. Benchmark performance results of concatenated (side-by-side) LW and SW spectral images. Error bars are one standard deviation from the mean normalized to the square root of the observer population.

the correct answer according to the following relationship

$$P_C = \frac{P_M - \left(\frac{1}{8}\right)}{\left(\frac{7}{8}\right)} \tag{1}$$

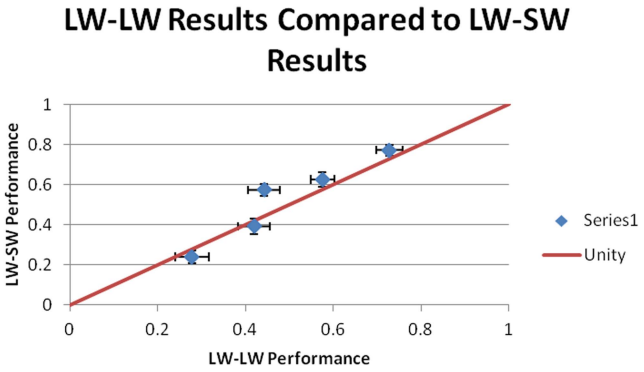
where P_M is the measured probability of the observer's response and 1/8 is the probability of guessing the correct answer for an 8-AFC experiment. These probabilities were then ordered and two statistical tests were performed to test each observer's performance relative to the distribution of the ensemble. The first test was an inter-quartile distance test. If any observer's performance was less than three times the distance from the first quartile to the median, those observer's results were removed from the data set for the entire experiment. If any observer's performance was more than three times the distance from the median to the third quartile, those observers' results were removed from the data set for the entire experiment. The second statistical test was Chavenault's criterion. Since Chavenault's criterion assumes that measurements follow a Normal distribution, it is sensitive to the distribution of the data. An inter-quartile test based on the median value and the distance between the first and third inter-quartile to the median makes no assumptions on the distribution of the data and was found to be more useful initially to find outlying observer results. If an observer result was rejected by either test, that observer result was removed from any further analysis.

4.1. Side-by-side Experiment Results

Twenty-three observers participated in the experiment and five observers were removed because their overall results in the experiment were rejected as outliers by the previously mentioned statistical tests. The observer responses were then averaged over all images at each specific range for each waveband combination. These probabilities were then corrected for guessing using the same algorithm discussed earlier. The experimental results of the remaining eighteen observers are shown in Figure 4.

As may be seen in Figure 4, there are only 2 occurrences in which the average values differ by more than the standard error. LW-SW imagery outperformed both the LW-LW and SW-SW images at Range 3 and SW-SW underperformed both the LW-LW and LW-SW imagery at Range 4. The results using the LW-SW images show that the observers effectively used the available source band information as needed when making their decisions. This is evident by observing that the LW-SW curve tracks the better performing source band across range. This is an important characteristic that any image fusion algorithm should obtain if the intent is to judge how it affects task performance. This characteristic, the best combined spectral source band performance achievable, is referred to as benchmarking performance [3] and can be calculated by recording when the observer correctly identified the target using either spectral image. A performance benchmark indicates the optimal level of performance capability based on the fact that the information present in both source bands is the same across image fusion techniques; with differences being attributed to how the information is merged. The benchmark source band ID performance for the side-by-side LW and SW experiment is shown in Figure 5.

The benchmark source band performance was calculated from the LW-LW and SW-SW images. If the observer responded with the correct answer for a target and aspect in either spectral source band, then the image was graded as correct even if the observer recorded an incorrect answer in the other spectral band. A subtle but important point should be noted that range 4 produced the worst performance in the SW-SW performance with a 0.06 probability, yet the benchmark performance at range 4 is 0.06 greater than the better performing LW-LW source band. Therefore, the few images that were correctly identified in the SW-SW case were those vehicles and aspects that were incorrectly identified in the LW-LW source band. This subtlety of the data is a result that is missed if we assume that the observer is not allowed to select between different source bands



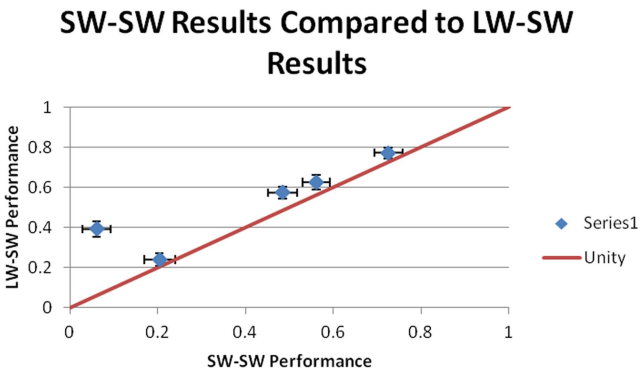


Fig. 6. Comparison of concatenated (side-by-side) LW-SW performance to individual source band performance.

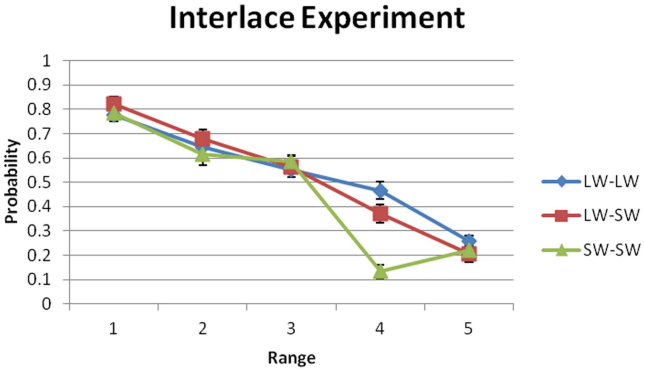


Fig. 7. Experimental results of LW and SW spectral 2-field movies. Error bars are one standard deviation from the mean normalized to the square root of the observer population.

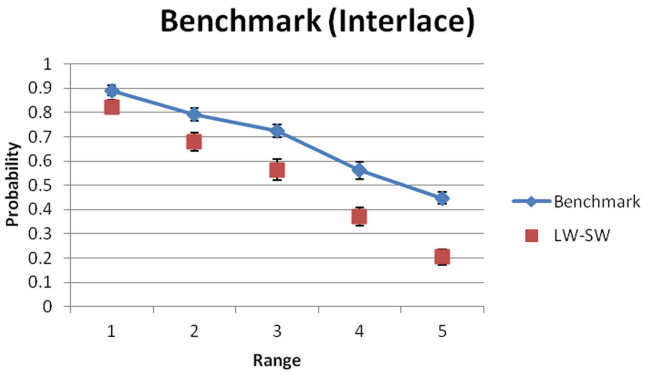


Fig. 8. Benchmark performance results of LW and SW spectral 2-field movies. Error bars are one standard deviation from the mean normalized to the square root of the observer population.

and must choose only a source band image (LW or SW) or the fused image. A possible explanation for the degraded performance in the concatenated LW-SW imagery is that the observers may have cued exclusively onto one spectral image even though both were available. In order to test this explanation we can compare the LW-SW results to both the LW-LW and SW-SW results. These comparisons are shown as Figure 6.

As may be seen in Figure 6, LW-SW results are almost identical to the LW-LW results. In fact the correlation coefficient between these data sets is 0.9522, whereas the correlation coefficient between the LW-SW and SW-SW results is 0.8991. As in the LW-MW experiment, the LW-SW results are more like the LW-LW results than the SW-SW results, implying that the observers made more targeting decisions using the LWIR spectral image when both spectral images were present. This effect might be addressed in better training for this display format.

4.2. Temporally Interlaced Imagery Experiment Results

Twenty-three observers participated in the experiment and five observers were removed because their overall results in the experiment were rejected as outliers by either an inter-quartile distance test or Chavenault's criterion. The experimental results of the remaining eighteen observers are shown in Figure 7.

Figure 7 shows there is one occurrence in which the averages were separated by more than the standard error of the data. Range 4 LW-SW outperformed SW-SW and the LW-LW data outperformed the other two. For all other ranges, observers performed equally well on all movies regardless of spectral content. The results using the LW-SW images show that the observers effectively used the available source band information as needed when making their decisions. This is evident by observing that the LW-SW curve tracks the better performing source band across range. This is an important characteristic that any image fusion algorithm should obtain if the intent is to judge how it affects task performance. However, if either source band spectral image is available to the observer and viewed at will by the observer, the benchmark performance achievable for this experiment is shown in Figure 8.

As in the previous experiment, the calculated benchmark source band performance was calculated from the LW-LW and SW-SW movies. If the observer responded with the correct answer for a target and aspect in either spectral source band, then the movie was graded as correct even if the observer recorded an incorrect answer in the other spectral band. A possible explanation for the degraded performance in the LW-SW movie is that the spectral image least useful to the observer may have masked or diminished the spectral information in the

TABLE 1
Comparison of observer performance between the different display formats while viewing only LW source band imagery. Avg. is the average probability of identification P(ID) of all observers and ±Error is the standard error for each P(ID).

LW	Range									
Display Type	1		2		3		4		5	
	Avg.	±Error	Avg.	±Error	Avg.	±Error	Avg.	±Error	Avg.	±Error
Side-by-Side	0.728	0.030	0.575	0.027	0.442	0.037	0.420	0.036	0.278	0.037
Temporal Interlace	0.778	0.026	0.644	0.030	0.553	0.033	0.466	0.035	0.257	0.024
Algorithm Fused	0.730	0.054	0.448	0.046	0.405	0.046	0.325	0.046		

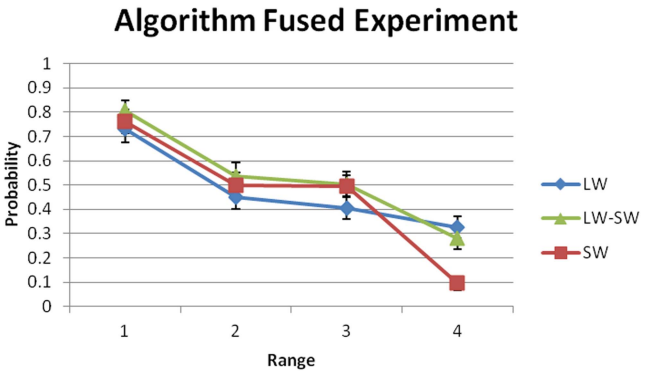


Fig. 9. Experimental results for the LW-SW superposition algorithm fused imagery. The error bars are one standard deviation from the mean normalized to the square root of the observer population.

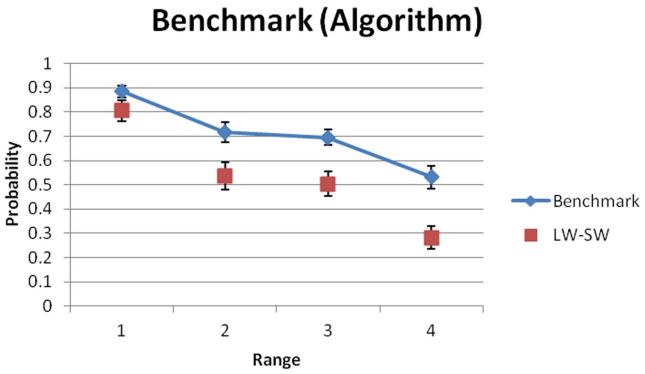


Fig. 10. Benchmark performance results for the LW-SW superposition algorithm fused imagery. The error bars are one standard deviation from the mean normalized to the square root of the observer population.

resulting movie. This effect cannot be addressed with training as it is an artifact of the display format.

4.3. Algorithm Fused Image Experimental Results

Fifteen observers participated in this experiment and two of them were removed as outliers. The remaining thirteen observer responses were then averaged over all images at each specific range for each waveband combination. Range 5 was omitted from this experiment. The probabilities were corrected for guessing as described previously, Equation 1, and the results are shown in Figure 9.

Figure 9 shows the LW and SW images are never separated further than the standard error at ranges 1 and 2. However, at range 3 SW outperforms LW and at range 4 LW outperforms SW. The LW-SW superposition fused imagery performed as well as the best performing source band spectral images within the standard error of the data sets. This was not the case in the LW-MW superposition experiment, where the LW-MW superposition fused imagery underperformed both spectral source bands greater than can be accounted for by the standard error in the data sets. When the source band information was similar (LW-MW) the superposition algorithm degraded the source band information enough that task performance suffered; in comparison, when the source bands contain different information (LW-SW) about the same scene the superposition algorithm retained enough source band information such that task performance did not suffer.

Figure 10 shows the calculated benchmark source band performance obtained using the LW and SW images. If the observer responded with the correct answer for a target and aspect in either spectral source band, then the image was graded as correct even if the observer recorded an incorrect answer in the other spectral band. A possible explanation for the degraded performance in the LW-SW algorithm case is that the superposition algorithm may have masked or diminished the spectral information provided by the spectral image most useful to the observer. This effect cannot be addressed with training as it is an artifact of the algorithm.

5. COMPARISON OF EXPERIMENTS

In order to draw wider ranging conclusions, it was important to assess if the display format contributed to differences in performance. As a first order analysis, a direct comparison of observer responses for each display format when viewing single spectral source band images will be done followed by a comparison of performance between the display formats with mixed spectral image content.

Table 1 shows a comparison between the average observer performances using only LW images at each of the five ranges for each display format.

TABLE 2
Comparison of observers' LW P(ID) performance between the various display formats for each range which exceeded the standard error for the data set. (+) column exceeds row, (−) row exceeds column, (X) no comparison.

LW	Side-by-Side	Temporal Interlace
Temporal Interlace	−0.012 R2 −0.041 R3	X
Algorithm	+0.054 R2 +0.013 R4	+0.120 R2 +0.069 R3 +0.060 R4

Table 2 compares the observers' LW P(ID) performance between the various display formats for each range. The format in the column is being compared to the format or algorithm listed in the row. Hence reading the first column first row shows that LW Side-by-Side underperformed the temporal interlace format at ranges 2 and 3 (labeled as R2 and R3 respectively) greater than what is accounted for in the data set error.

Of the 13 paired comparisons made between the different experiments, 7 comparisons produced differences not accounted for by the error associated with calculating the mean value and the largest value was 0.120 greater than the error associated with calculating the average value. However, 5 of the 7 comparisons involved the temporal interlace format outperforming the other two formats.

Table 3 shows a comparison between the average observer performances using only SW images at each of the five ranges for each display format.

Table 4 compares the observers' SW P(ID) performances between the various display formats for each range. The format in the column is being compared to the format or algorithm listed in the row. Hence reading the first column first row shows that SW Side-by-Side underperformed the temporal interlace format at ranges 3 and 4 greater than what is accounted for in the data set error.

TABLE 3
Comparison of observer performance between the different display formats while viewing only SW source band imagery. Avg. is the average probability of identification P(ID) of all observers and ±Error is the standard error for each P(ID).

SW	Range									
Display Type	1		2		3		4		5	
	Avg.	±Error	Avg.	±Error	Avg.	±Error	Avg.	±Error	Avg.	±Error
Side-by-Side	0.725	0.032	0.561	0.031	0.484	0.033	0.061	0.032	0.204	0.035
Temporal Interlace	0.783	0.032	0.616	0.045	0.585	0.025	0.132	0.029	0.220	0.022
Algorithm Fused	0.762	0.048	0.500	0.053	0.496	0.044	0.095	0.027		

TABLE 4
Comparison of observers' SW P(ID) performance between the various display formats for each range which exceeded the standard error for the data set. (+) column exceeds row, (−) row exceeds column, (X) no comparison, (ND) no statistically significant difference for all ranges.

SW	Side-by-Side	Temporal Interlace
Temporal Interlace	−0.043 R3 −0.010 R4	X
Algorithm	ND	+0.018 R2 +0.020 R3

Of the 13 paired comparisons made between the different experiments, only 4 comparisons produced differences not accounted for by the error associated with calculating the mean value and the largest value was 0.043 greater than the error associated with calculating the average value. However, all comparisons involved the temporal interlace format outperforming the other two formats.

An analysis was conducted to further investigate any effect display format may have had on the observer's performance for their benchmark performance given either source spectral band imagery; Table 5 shows these comparisons.

TABLE 5
Comparison of observer's benchmark source band performance between the different display formats.

Benchmark	Range									
Display Type	1		2		3		4		5	
	Avg.	±Error	Avg.	±Error	Avg.	±Error	Avg.	±Error	Avg.	±Error
Side-by-Side	0.852	0.020	0.741	0.019	0.654	0.029	0.492	0.034	0.434	0.042
Temporal Interlace	0.892	0.021	0.793	0.027	0.725	0.027	0.561	0.035	0.447	0.024
Algorithm Fused	0.885	0.025	0.715	0.042	0.694	0.032	0.531	0.047		

Table 6 compares the observers' benchmark P(ID) performance between the various display formats for each range. The format in the column is being com-

TABLE 6
Comparison of observers' benchmark source band P(ID) performance between the various display formats for each range which exceeded the standard error for the data set. (+) column exceeds row, (−) row exceeds column, (X) no comparison, (ND) no statistically significant difference for all ranges.

Benchmark	Side-by-Side	Temporal Interlace
Temporal Interlace	−0.006 R2 −0.015 R3	X
Algorithm	ND	+0.009 R2

TABLE 8
Comparison of observer's performance between the different display formats using the LW-SW spectral source bands together for each range which exceeded the standard error of the data set. (+) column exceeds row, (−) row exceeds column, (X) no comparison, (ND) no statistically significant difference for all ranges.

Fused	Side-by-Side	Temporal Interlace
Temporal Interlace	ND	X
Algorithm	+0.025 R4	+0.052 R2 +0.004 R4

TABLE 7
Comparison of observer's performance between the different display formats using the two spectral source bands together. Avg. is the average probability of identification P(ID) of all observers and ±Error is the standard error for each P(ID).

Display Type	Range									
	1		2		3		4		5	
	Avg.	±Error	Avg.	±Error	Avg.	±Error	Avg.	±Error	Avg.	±Error
Side-by-Side	0.772	0.030	0.625	0.036	0.574	0.030	0.392	0.039	0.239	0.030
Temporal Interlace	0.823	0.028	0.680	0.036	0.563	0.044	0.370	0.038	0.204	0.031
Algorithm Fused	0.806	0.044	0.536	0.056	0.504	0.051	0.282	0.046		

pared to the format or algorithm listed in the row. Hence reading the first column first row shows that Benchmark Side-by-Side underperformed the temporal interlace format at ranges 2 and 3 greater than what is accounted for in the data set error.

Of the 13 paired comparisons made between the different experiments, only 3 comparisons produced differences not accounted for by the error associated with calculating the mean value and the largest value was 0.015 greater than the error associated with calculating the average value. However, all comparisons involved the temporal interlace format outperforming the other two formats. Although not nearly as significantly as in either source band comparison. This result provides us confidence that the benchmark concept is valid and should be reproducible regardless of display format.

Table 7 shows a comparison of observer performance between the different display formats using only LW-SW images at each of the five ranges.

Table 8 compares the observers' "self-fused" P(ID) performance between the various display formats and superposition fused P(ID) performance for each range. The format in the column is being compared to the format or algorithm listed in the row. Hence reading the first column second row shows that fused Side-by-Side outperformed the fusion algorithm format at range 4 greater than what is accounted for in the data set error.

Of the 13 paired comparisons made between the different experiments, only 3 comparisons produced differences not accounted for by the error associated with calculating the mean value and the largest value was 0.052 greater than the error associated with calculating the average value. However, all comparisons involved the fusion algorithm underperforming the other two formats.

6. DISCUSSION AND RECOMMENDATIONS

The research presented in this paper utilized a standard set of multi-spectral images suitable for image fusion and image fusion algorithm performance evaluations. The results obtained from using both the display formats and the superposition algorithm correlated with the best performing individual source band. However, test results show the superposition fusion algorithm underperformed the temporal interlace format. In fact, temporal interlacing the imagery allowed for task performance closest to the observers' benchmark. Overall experimental results show that observers P(ID) performances using the superposition fused images are well below that which was achieved by the observers' benchmark source band performances. It is therefore recommended that superposition fusion not be used as a baseline when assessing image fusion P(ID) performance.

Benchmark performances were measured for a variety of "self-fusion" display techniques. Considering the diverse observer pool and small errors associated with the resultant data, the comparisons reported in this paper show that benchmark performances were relatively unaffected due to these changes in the display format. As a result, we are further recommending that human observer performance using fusion algorithms to fuse together the LW and SW spectral band imagery needs to achieve the optimal values shown in Table 9.

These recommendations are based on the measurements made in the experiments in which the human observer was viewing only single source band imagery.

TABLE 9
Recommended optimal human performance for fusion algorithms when fusing this LW and SW imagery.

	Range 1	Range 2	Range 3	Range 4	Range 5
Average	0.88	0.75	0.69	0.53	0.44
±Error	0.02	0.03	0.03	0.04	0.03

TABLE 10
Recommended optimal human performance for fusion algorithms when fusing this LW and MW imagery.

	Range 1	Range 2	Range 3	Range 4	Range 5
Average	0.88	0.66	0.50	0.33	0.38
±Error	0.03	0.03	0.03	0.03	0.03

That is, the image was LW or SW but not both simultaneously. For comparison, the recommended human performance given in Moyer and Howell using fusion algorithms to fuse together the LW and MW spectral band imagery needs to achieve the optimal values shown in Table 10.

It should be noted that the LW imagery used within the LW-MW tests had an average contrast of 0.32 while the LW imagery used within the LW-SW tests had an average contrast of 0.21. This is important if comparisons between the LW imagery used in the LW-MW and LW-SW experiments are conducted. The values in Table 9 and Table 10 are greater than any results achieved using the original bands exclusively. It is therefore reasonable with regards to benchmarking performance to expect this type of performance if the human observer is provided both source bands when making the targeting decision.

These benchmark performances exceeded the superposition algorithm performance for both the LW-MW and LW-SW experiments. By extension, if a fusion algorithm achieves an image quality metric similar to the superposition algorithm, then we can expect human performance using that image fusion algorithm to be less than the human performance achievable if the observers had the ability to select between the original source band images. This research effectively bridges the knowledge gap between "self-fusion" and algorithmically fused performance assessments and has also identified a standard set of source band imagery suitable for image fusion and image fusion algorithm performance assessment.

REFERENCES

[1] R. Blake
A Primer on Binocular Rivalry, Including Current Controversies.
Brain and Mind, 2, (2001), 5–38.

[2] K. Byrd, H. Szu, and M. Chouika
A subspace learning approach to evaluating the performance of image fusion algorithms.
Proceedings of SPIE, 7703, (2007).

[3] C. Howell
Benchmarking image fusion algorithm performance.
Proceedings of SPIE, 8355, (2012).

[4] T. Hui and W. Binbin
Discussion and Analyze on Image Fusion Technology
International Conference on Machine Vision, (2009).

[5] D. Kim, S. Choi, and K. Sohn
Depth adjustment for stereoscopic images and subjective preference evaluation
Journal of Electronic Imaging, 20(3), (2011).

[6] S. Moyer and C. Howell
Establishment of Human Performance Baseline for Image Fusion Algorithms in the LWIR and MWIR Spectra.
Journal of Advances in Information Fusion, (2013), *submitted for publication*.

[7] J. O'Connor
Infrared imagery acquisition process supporting simulation and real image training
Proceedings of SPIE, Modeling and Simulation for Defense Systems and Applications VII, 8403, 3 (May 2012).

[8] J. O'Connor, P. O'Shea, J. Palmer, and D. Deaver
Standard target sets for field sensor performance measurements
Proceedings of SPIE, 6207, (2006).

[9] A. Toet and M. Hogervorst
Progress in color night vision
Optical Engineering, 51(1), (2012).

Christopher Howell received his M.S. and Ph.D. degrees in electrical engineering from the University of Memphis, Memphis, TN in 2007 and 2010 respectively. He is currently working as an electronics engineer for the U.S. Army's Night Vision and Electronics Sensors Directorate. His current research interest involves investigating the effects of image fusion on human visual task performance and image quality assessment. Other research interests include model development for image fusion systems and target tracking.

Dr. Steve Moyer is currently working at the U.S. Army Night Vision Electronic Sensors Directorate (NVESD) as the Lead Researcher modeling EO/IR sensor performance in detecting hastily emplaced explosives from a moving ground vehicle. He has a B.S.E.E. from the Pennsylvania State University, a Master of Science from Georgia Institute of Technology, and completed his Ph.D. in August 2006 also from the Georgia Institute of Technology. Previously, Dr. Moyer was the Lead Researcher conducting experiments to further the application of sensor models, such as NVThermIP, to tasks associated with the urban battlespace. Most recently Dr. Moyer has been working on characterizing human performance in searching for and detecting small easily concealed explosives from moving vehicles using infrared sensors.