Evaluation of fused imagery using eye movement-based measures of perceptual processing

Mackenzie G. Glaholt
Grace Sim
Philips Laou
Simon Roy

DRDC – Toronto Research Centre

Bibliographic information: Information Fusion, Volume 39, Pages 186–203

Date of Publication from Ext Publisher: April 2017
IMPORTANT INFORMATIVE STATEMENTS

Disclaimer: This document is not published by the Editorial Office of Defence Research and Development Canada, an agency of the Department of National Defence of Canada, but is to be catalogued in the Canadian Defence Information System (CANDIS), the national repository for Defence S&T documents. Her Majesty the Queen in Right of Canada (Department of National Defence) makes no representations or warranties, expressed or implied, of any kind whatsoever, and assumes no liability for the accuracy, reliability, completeness, currency or usefulness of any information, product, process or material included in this document. Nothing in this document should be interpreted as an endorsement for the specific use of any tool, technique or process examined in it. Any reliance on, or use of, any information, product, process or material included in this document is at the sole risk of the person so using it or relying on it. Canada does not assume any liability in respect of any damages or losses arising out of or in connection with the use of, or reliance on, any information, product, process or material included in this document.

This document was reviewed for Controlled Goods by Defence Research and Development Canada (DRDC) using the Schedule to the Defence Production Act.
Full Length Article

Evaluation of fused imagery using eye movement-based measures of perceptual processing✩

Mackenzie G. Glaholt*, Grace Sim, Philips Laou, Simon Roy

Defence Research and Development Canada

ARTICLE INFO

Article history:
Received 27 September 2016
Revised 22 March 2017
Accepted 16 April 2017
Available online 17 April 2017

Keywords:
Sensor fusion
Performance evaluation
Eye movements
Infrared
Visual search

ABSTRACT

Human performance measures were used to evaluate the perceptual processing efficiency of infrared and fused-infrared images. In two experiments, eye movements were recorded while subjects searched for and identified human targets in forested scenes presented on a computer monitor. The scenes were photographed simultaneously using short-wave infrared (SWIR), long-wave infrared (LWIR), and visible (VIS) spectrum cameras. Fused images were created through two-way combinations of these single-band images. In Experiment 1 the single band sensors were contrasted with a simple average fusion scheme (SWIR/LWIR). Analysis of subjects’ eye movements revealed differences between sensors in measures of central processing (gaze duration, response accuracy) and peripheral selection (detection interval, saccade amplitude). In Experiment 2 this methodology was applied to compare three two-way combinations of sensors (SWIR/LWIR, SWIR/VIS, VIS/LWIR), produced by state-of-the-art fusion methods. Peripheral selection for fused images tended to exhibit a compromise between the performance levels of component sensor images, while measures of central processing showed evidence that fused images matched or exceeded the performance level of component single-band sensor images. Stimulus analysis was conducted to link measures of central and peripheral processing efficiency to image characteristics (e.g. target-contrast, target-background contrast), and these image characteristics were able to account for a moderate amount of the variance in the performance across fusion conditions. These findings demonstrate the utility of eye movement measures for evaluating the perceptual efficiency of fused imagery.

Crown Copyright © 2017 Published by Elsevier B.V. All rights reserved.

1. Introduction

Electro-optical (EO) sensor technology has advanced greatly over the past three decades and portable EO cameras that capture electromagnetic (EM) radiation with wavelengths outside the visible spectrum are increasingly available. In addition, a variety of software and hardware techniques have been developed for merging images from multiple sensors to create a single image. This area of research, known as sensor and image fusion, has introduced a wide variety of synthetic imagery that can be presented to the human viewer. In contrast to the rapid pace of development for sensor fusion technology, advancement in methods for quantitative evaluation of the performance of fused imagery has been slow. The purpose of the present study is to explore the use of eye movement-based techniques for evaluating the performance of sensor imagery. In particular, we analyzed viewers’ eye movements and quantitatively compared the perceptual efficiency of different sensor images in terms of central and peripheral visual processing. We begin by introducing some of the characteristics of modern EO sensor imaging devices as they pertain to visual search performance in the military context. We then briefly review the literature on human performance evaluation of EO sensor imagery, with a focus on prior research that used eye movements for this purpose. Finally we provide the rationale for the present analytic approach, which is motivated by the central/peripheral functional distinction in human vision.

Infrared imaging devices are widely used in military and law-enforcement to enhance the operational vision capabilities of human operators. Infrared cameras detect EM radiation of wavelengths outside the human visible spectrum and convert this radiation into a visible spectrum image that is presented to the human viewer (for reviews see [1,2]). The visual information carried by infrared EM radiation is potentially useful in the context of visual search. For example, long-wave infrared (LWIR, also called “thermal”) imaging detects EM radiation in the 8–14 μm spectral band. Objects that are warm relative to the ambient background temperature exhibit high contrast in LWIR. Because LWIR radiation is
emitted by objects, LWIR cameras do not need a separate source of illumination and thus can enhance human vision capabilities when ambient light levels are low (e.g. at night). More importantly, warm objects are often relevant for visual search in the military context (e.g. humans, vehicles), thus LWIR can provide imagery in which relevant search targets exhibit high target-background contrast. Short-wave infrared (SWIR) sensors detect EM radiation in the 1.1–3.0 μm spectral band. In contrast to LWIR, SWIR radiation is reflected by objects rather than emitted by them (as is visible spectrum light), and therefore objects must be externally illuminated to be photographed in SWIR. Importantly, because some surfaces reflect EM radiation in the SWIR band but not the visible band (and vice versa), SWIR imaging sensors can provide additional visual information about objects that is not available in the visible spectrum. Imaging sensors that detect EM radiation in other bands (e.g. medium-wave infrared; 3–5 μm) may also provide complementary information about objects in the visual world and therefore have potential to aid or enhance visual search of natural scenes.

Certain practical problems emerge with the presentation of information from multiple sensor imagers to the human viewer. The user can only view one image at a time, and therefore when multiple imaging sources are available (e.g. short-wave, long-wave, visible spectrum images), the viewer must either toggle between sensor images, or else the source images must be merged in some way to create a composite image. This latter approach is known as image fusion. The goal of image fusion is to combine images from two or more sources in such a way that all salient information from the source images is preserved, while any noise in the source images is muted, and no additional artifacts have been introduced as a result of the fusion process [3,4]. Moreover, the fused image might contain emergent information resulting from the contrast between the two source images [5]. A considerable body of research has been devoted to developing methods for sensor/image fusion (e.g. [3,4,6–10]). While a complete review is outside of the scope of the present paper (but see [11,12]), fusion methods vary according to the source sensors that are fused (e.g. SWIR, LWIR, visible spectrum, image-intensified visible to near infrared) and many aspects of the fusion process such as: fusion at the level of sensor or image, the method of warping or transforming images into the same space, image decomposition techniques, the use of monochrome or colour for the fused image, etcetera.

Given the numerous ways to combine information from multiple sensors to create a single image, it is of critical importance to have methods for evaluating the quality of fused image products. A variety of computational image metrics have been developed to quantify the relationship between a fused image and its component sensors [see [13–16]], but researchers have increasingly sought to employ behavioural measures in order to evaluate the efficiency of perceptual processing of infrared and fused images [5,17]. In particular, performance on operationally-relevant viewing tasks has been used as a measure of perceptual processing efficiency. The most common viewing task used for this purpose is visual search, where the dependent measure is the latency or likelihood to detect a target within an image. For example, studies have examined subjects’ performance in detecting and/or localizing human targets [5,18–23], vehicles [5,23–25], and buildings or other terrain features [17,23] within fused images. In addition to visual search, some studies have required subjects to make orientation judgements about objects in the scene [5] or perform scene categorization [5,19]. Within this body of research are cases for which fused images demonstrated superior performance to their component images [5,17–19,21,22], and also cases where the fused images were no better [24] or even worse than their component images [5,20]. These differences across studies are likely due to differences in component image source sensors (e.g. SWIR, LWIR, visible spectrum), fusion algorithms, the content of the images, the viewing task, as well as the method of assessing behavioural performance.

Eye movement recordings have proved to be an important tool in the study of visual search of natural scenes (for reviews see [26–29]), but there have been only a few applications of this technique to the domain of infrared sensors and image fusion [15,30–33]. In a study by Krebs, Scribner, and McCarley [33], eye tracking was used to obtain measures of scanpath length during visual search of infrared and fused scenes. By monitoring the sequence of spatial locations of eye fixations on the scene, the total length of the path of these eye fixations (i.e., scanpath length) prior to fixating the visual search target was computed. Scanpath length was interpreted as reflecting search efficiency, and the authors observed that this variable differed between sensors and fusion schemes. In subsequent studies, Dixon et al. [15,30] monitored eye movements while subjects viewed video sequences in which a target character moved across the scene. The authors also conducted and analyzed the sequence of raw fixation locations and found that subjects’ gaze tracked the target more accurately with fused images than with component infrared and visible spectrum video sequences. This performance measure was also sensitive to the method of sensor fusion (simple averaging vs. wavelet-based fusion).

Lanir and Maltz [31] recorded eye movements while subjects searched fused infrared images (fusing SWIR, MWIR, or LWIR images) for military vehicles. The images were composite such that one half of the image contained imagery from one fusion method and the other half of the image contained imagery from another fusion method. Three fusion methods were directly contrasted in this way and the authors computed the percent of fixation duration directed to each image half (i.e. fusion condition). Fusion conditions were shown to differ in the percentage of fixation duration, and the authors interpreted a greater viewing duration as being indicative of greater information content. The authors also measured the cumulative fixation duration while fixating the visual search target, and found a marginally significant difference between image fusion conditions. This was interpreted as reflecting differences in target salience across fusion conditions.

Toet et al. [32] examined eye movement behaviour during visual search of colour-fused near-infrared/visible spectrum images as well as component single-band images. The authors examined the total fixation duration, number of fixations, and fixation rate (i.e. number of fixations per second) but found no differences in these measures as a function of sensor condition. However, in a subsequent analysis of scanpaths, they did observe an effect of sensor on the order in which objects in the scenes were fixated, and concluded that this was a result of differences in the salience of those objects across image conditions.

One important aspect of the human oculomotor system that has been overlooked in research on sensor fusion assessment is the distinction between central and peripheral visual processing. The human visual system is specialized such that the central visual field, including the fovea (i.e., the central few degrees about the point of gaze), contains a high concentration of cone photoreceptors and is optimized for visual acuity and the perception of high spatial frequencies. Outside of the fovea the density of cone photoreceptors drops off steeply and consequently the peripheral visual field has low acuity and reduced sensitivity to chromatic information. Nevertheless, the peripheral visual field remains sensitive to luminance transients and motion [34–36]. During visual search, areas outside of central vision are selected (peripheral selection) for detailed processing in central vision (central processing), and eye movements (saccades) align central vision to fixate those areas so that detailed visual processing can occur [37–40]. This functional division between central and peripheral vision is potentially important in the context of electro-optic sensor imagery, as certain
sensors might facilitate peripheral selection (e.g. low resolution, monochromatic but high luminance contrast) while others might promote efficient central processing (e.g. high resolution, chromatic information).

The purpose of the present study was to investigate the use of eye movement-based measures of central and peripheral visual processing for evaluating the perceptual efficiency of infrared and fused images. Accordingly, a multi-spectral image set was created featuring human targets posed against a forested background. The scenes were photographed with a visible spectrum camera (VIS), as well as with SWIR and LWIR cameras. In Experiment 1 we introduced measures of peripheral selection and central processing in order to characterize the perceptual efficiency of the single-band imagery as well as SWIR/LWIR fused imagery created with a simple averaging method. To anticipate the findings of Experiment 1, the single-band sensors were found to differ substantially along central and peripheral processing measures, and the fusion condition was found to produce performance that was a compromise between the performance levels of its component single band sensors. In Experiment 2 we investigated the generality of this finding by testing a broader set of image fusion schemes (SWIR/LWIR, VIS/LWIR, SWIR/VIS) created using state-of-the-art fusion techniques. Finally, we conducted stimulus modeling to determine whether image characteristics could predict the behavioural measures of peripheral selection and central processing.

2. Experiment 1

In Experiment 1 we demonstrated the use of eye movement-based measures for assessing the peripheral selection and central processing efficiency of single-band (VIS, SWIR, and LWIR imagery) and fused imagery. Fused images (FUSE) were created by combining SWIR and LWIR images according to a simple method of alignment and pixel-level averaging. The LWIR images were expected to exhibit strong target-background contrast (due to thermal differences) which we hypothesized would facilitate peripheral selection of the target in these images. The VIS and SWIR sensors capture reflected EM radiation and were expected to provide more target detail than the LWIR sensor, which we hypothesized would result in a relative advantage for VIS and SWIR in central processing of the target. We speculated that the FUSE images might inherit some or all of the properties of its component sensors, and this condition was used as a test case to determine whether our measures of central and peripheral processing were sufficiently sensitive to quantify the relative efficiency of a fused sensor image compared to its component images.

2.1. Method

2.1.1. Subjects

Sixteen male members of the Canadian Armed Forces participated in the experiment (mean age = 28.3 years, s.d. = 5.9, all right-handed). Subjects provided informed consent and were remunerated according to Government of Canada Treasury Board guidelines for a total of $12.72 CAD for their one hour of experiment participation. All subjects had normal or corrected-to-normal vision. The research protocol was reviewed and approved by the Human Research Ethics Committee at the DRDC Toronto Research Center.

2.1.2. Apparatus

Eye movements were measured with an SR Research EyeLink 1000 system with high spatial resolution and a sampling rate of 1000 Hz. Viewing was binocular, but only the subject’s dominant eye (self-reported) was monitored. Following calibration, gaze position error was less than 0.5°. The stimuli were presented on a BenQ 2420TX monitor (viewable area = 531 × 298 mm) with a refresh rate of 120 Hz and a screen resolution of 1920 × 1080 pixels. The experiment room was dimly lit and a chin rest with a head support was used to minimize head movement and ensure a consistent viewing distance of 68.5 cm. The experiment was implemented in SR Research Experiment Builder.

2.1.3. Materials and design

The stimuli for the visual search task were scenes photographed in rural Quebec, Canada. The photography occurred on a partly cloudy day in November, between the 1000 and 1500 hours, and the ambient temperature was 2–3°C. In each scene (see Fig. 1c) a single visual search target was present: a male model dressed in one of four ways (see Fig. 1a): military uniform holding a weapon, military uniform unarmed, civilian clothing holding a weapon, civilian clothing unarmed. These dress configurations were chosen because they constitute a military-relevant discrimination. The target character was approximately 1.75 m tall and stood at one of three distances (100 m, 200 m, 300 m) from the camera (see Fig. 1a). Stimuli used in Experiment 1: a cut-out of each of the four characters in each of the four sensors at Distance 1 (panel a); cut-outs of Character 4 displayed at Distances 2 and 3 (panel b); a sample scene with Character 1 at Distance 1 in the VIS sensor (panel c).
Fig. 1a, b). At the time of presentation during the experiment, the three target distances (Distances 1, 2, and 3) appeared as 3.18, 1.41, and 0.88° of visual angle (vertically), which correspond to apparent target distances of 31 m, 71 m, and 114 m, respectively. For each character at each Distance setting there was a unique scene background (12 total), and for each of these scene backgrounds the target was photographed in three randomly chosen positions, resulting in a total of 36 original scenes.

The design of the experiment required that each scene be photographed simultaneously in three different spectral bands. Accordingly, a camera mounting system was created that housed three cameras: a VIS camera, a SWIR camera, and a LWIR camera. The technical parameters of the VIS, SWIR, and LWIR cameras are presented in Table 1. Each camera was oriented to capture (as closely as possible) the same field of view of the distal scene. The cameras filmed concurrently and still images were later extracted from the raw footage. The images were subjected to several pre-processing steps in Adobe Photoshop CS4. The images were initially encoded as 16-bit greyscale and were subsequently resampled to 8-bits. The LWIR images were up-sampled to the resolution of the SWIR images multiplied by the ratio of their fields of view (1381 × 1106), were then aligned with the SWIR images (target overlap matched by hand) and then cropped down to the SWIR image resolution (1280 × 1024). From these two aligned images, a fused sensor condition (FUSE) was created by taking the arithmetic mean of the SWIR and the LWIR images at each pixel. VIS images were down-sampled by a factor of two to a resolution of 1224 × 1025, in order to approximately match the resolution of the SWIR and LWIR images. Each image was then contrast-adjusted (via histogram stretching) manually to subjectively maximize the contrast of the target figure in the scene. Images were mounted in the center of a 1920 × 1080 grey background (R= G= B= 70). Finally, in order to increase experimental power, a copy of each image was created by mirroring it in the vertical axis; this produced a total of 6 target positions (3 target positions mirrored) for each character at each distance for a total of 72 images per sensor condition (288 images total).

To summarize, the experiment was a within-subject design with four Sensor conditions (VIS, SWIR, LWIR, FUSE) crossed with three Target Distances (1, 2, 3). Each subject saw each scene four times, once in each sensor condition and in a random order, though the order of scene presentation was counterbalanced across every four subjects such that each scene was equally likely to appear in each sensor condition first, second, third, and fourth.

### 2.1.4. Procedure

The experimenter described the general procedure to the subject and the subject provided informed consent to participate. A 9-point eye-tracker calibration and accuracy test was carried out to ensure an average gaze-prediction error less than 0.5° and a maximum error less than 1°. The subject was instructed to search for and identify the character present in the scenes as quickly and accurately as possible. As part of the instructions the subject saw a familiarization screen showing all of the characters in each of the four sensor conditions (shown at Distance 1), along with the mapping of each of the four characters to one of the buttons on a gamepad. The familiarization screen was re-presented before each trial as a reminder of the character appearances and response mappings.

On each trial the subject pressed the ‘space’ bar to advance past the target familiarization screen. The next screen that was displayed was uniform grey except for an oval containing the words ‘look here’ in one of the four corners of the screen. This start position was used to confirm eye-tracking calibration and to control the subject’s initial gaze position. The four start positions were equally likely to appear across trials and were randomly assigned to each of the 72 scene images (this assignment was constant across subjects). After the subject had fixated the start position for 300 ms, the search scene was presented on the screen. The trial was terminated by the subject’s response or else after a 10 s interval. There were 288 trials in the experiment broken up into 8 blocks of 36 trials. Subjects were offered a break between trials and an eye tracker calibration was carried out if necessary. The entire procedure lasted approximately an hour.

### 2.2. Results and discussion

The purpose of our analysis was to compare sensor performance according to measures of peripheral selection and central processing. Our analysis approach was as follows. First we sought to document any differences between the single band sensors. We expected increasing target Distance to have an adverse effect on all measures of performance, and therefore we included this factor in our analysis. Accordingly, each pair of sensors was analyzed in a within-subject Sensor × Distance (2 × 3) ANOVA. We were primarily interested in main effects of Sensor, and where the two-way interaction was found to be significant we conducted follow-up paired t-tests comparing Sensors at each level of Distance. Next, we compared the FUSE condition with each of its component sensors (i.e., SWIR, LWIR), in separate Sensor × Distance (2 × 3) ANOVAs, and conducted follow-up paired t-tests for any significant two-way interactions. In order to control the family-wise error rate, we adjusted the alpha criterion as follows. There were five ANOVAs carried out, and accordingly we adopted an alpha of .01 in order to produce a family-wise error rate of .05. Where follow-up t-test were applied to assist in the interpretation of significant two-way interactions, three t-tests were conducted comparing the two Sensor conditions at each Distance setting. Accordingly, we applied an alpha of .016. ANOVAs were carried out in SPSS 22 (IBM), and where Mauchly’s test of Sphericity indicated a violation of the sphericity assumption we applied Greenhouse-Geisser corrections. When analyzing the eye movement record for each visual search trial we considered only fixations and saccades (identified by the EyeLink parser) that began after the appearance of the visual search scene and prior to participant’s button response. Trials in which the subject did not respond were excluded from analysis (1% of trials).

### 2.2.1. Peripheral selection

We employed two behavioural measures of peripheral selection that were expected to differentiate sensor conditions. The primary measure was the detection interval, which was intended to measure the time elapsed until the search target entered central vision. For each experimental trial, the detection interval spanned from the presentation of the search scene until the target detec-

---

**Table 1**

Technical specifications of the cameras.

<table>
<thead>
<tr>
<th>Camera</th>
<th>Spectral Band (μm)</th>
<th>Image Resolution (pixels)</th>
<th>Horizontal Field of View (°)</th>
<th>Vertical Field of View (°)</th>
</tr>
</thead>
<tbody>
<tr>
<td>VIS</td>
<td>0.4–0.7</td>
<td>2448 × 2050</td>
<td>9.68</td>
<td>8.14</td>
</tr>
<tr>
<td>SWIR</td>
<td>0.9–1.7</td>
<td>1280 × 1024</td>
<td>9.17</td>
<td>7.33</td>
</tr>
<tr>
<td>LWIR</td>
<td>8–12</td>
<td>1024 × 768</td>
<td>9.91</td>
<td>7.45</td>
</tr>
</tbody>
</table>
tion event, which was defined as the beginning of the first eye fixation to fall within 3° of visual angle of the target center of mass. This measure differs from the scanpath length measure employed in previous work [33], in that it is sensitive to both the number of fixations prior to target detection as well as their duration. As a secondary measure of peripheral selection we computed the amplitude of the saccade (eye movement) that produced the detection event (henceforth saccade amplitude). For sensor conditions that facilitate target detection in the non-central visual field, we expected that the target would be detected at greater retinal eccentricities, and consequently the saccade that brought the target into central vision would have greater amplitude.

Detection intervals are displayed in Fig. 2 (panel a). The ANOVAs on detection interval confirmed that LWIR produced faster target detection than VIS and SWIR (effect of Sensor: both \( F(1,15) > 318.2, \ p < .001 \) and the magnitude of this effect differed as a function of Distance (Sensor x Distance interaction: both \( F(2,30) > 31.2, \ p < .001 \). Follow-up t-tests indicated significant differences for each Distance (all \( t(15) > 11.2, \ p < .001 \). The ANOVA comparing SWIR and VIS indicated that detection intervals for SWIR were significantly shorter than VIS (effect of Sensor: \( F(1,15) = 69.7, \ p < .001 \). This difference tended to decrease as target distance increased (Sensor x Distance interaction: \( F(2,30) = 5.3, \ p = .01 \), and follow-up t-tests indicated that the difference was significant at Distance 1 and 2 (both \( t(15) > 6.2, \ p < .001 \) but not Distance 3 (\( t(15) = 2.0, \ p = .068 \). As can be seen in the Figure, the FUSE condition produced intermediate performance compared to its component sensors. The ANOVA confirmed that FUSE produced shorter detection intervals when compared with SWIR (effect of Sensor: \( F(1,15) = 165.8, \ p < .001 \), and this advantage increased with increasing target distance (Sensor x Distance interaction: \( F(2,30) = 50.8, \ p < .001 \) with significant differences at each Distance in the follow-up t-tests (all \( t(15) > 2.9, \ p < .01 \)). Finally, the ANOVA comparing FUSE with LWIR indicated that FUSE detection intervals were slower than LWIR (effect of Sensor: \( F(1,15) = 43.2, \ p < .001 \), but this difference did not change significantly as a function of target distance (Sensor x Distance interaction: \( F(2,30) = 1.8, \ p = .18 \).

---

1 The visual field has typically been divided into three concentric circles based on eccentricity (e) from the point of fixation (see [28]): e < 1.5° (fovea); 1.5° < e < 5° (parafovea); e > 5° (periphery). In the present study we divided the parfoveal region to obtain two regions. We operationalized ‘central’ processing of the target as occurring during eye fixations within 3° of the target center pixel and ‘peripheral’ processing of the target occurred when it was outside the 3° radius.

---

Fig. 2. Experiment 1: detection interval (panel a), saccade amplitude prior to detection (panel b), gaze duration on target (panel c), and response accuracy (panel d), as a function of Sensor and Distance.
Mean values for saccade amplitude are presented in Fig. 2 (panel b). The ANOVAs indicated that LWIR produced longer detection saccades than SWIR and VIS (effect of Sensor: both \(F(1,15) > 13.4, \, p < .01\), but the difference between LWIR and VIS did not reach significance (effect of Sensor: \(F(1,15) = 2.2, \, p = .16\), nor did any of the two-way interactions (all \(F < 1\)). While detection saccades in the FUSE condition were numerically shorter than the LWIR condition and longer than the SWIR condition, these differences did not reach statistical significance (effect of Sensor: both \(F(1,15) < 5.9, \, p > .027\)) nor did the two-way interactions (both \(F < 1\)).

The detection interval and saccade amplitude findings are consistent with prior research reporting superior detection performance for LWIR imagery compared to visible spectrum imagery [22], and together they confirm that peripheral selection was more efficient for the LWIR images than the SWIR and VIS images. Images in the FUSE condition produced a pattern intermediate to its component sensors: detection took longer than in the LWIR condition but less time than in the SWIR condition, and detection saccades in the FUSE condition tended to be longer than in the SWIR condition, while it did not differ from the LWIR condition. Overall, peripheral selection performance in the FUSE condition reflects a compromise between its component sensors.

2.2.2. Central processing

We considered two behavioural measures of central processing: gaze duration on the target and response accuracy. Gaze duration on target was designed to capture the interval with which participants processed the target in central vision, and was computed as the cumulative duration of all fixations falling within 3° of the target center pixel. Sensor imagery that presents clear, discriminable, target information was expected to result in shorter processing times and consequently shorter gaze duration on target. Note that this interpretation of gaze duration differs from prior work [31] where longer gaze duration was interpreted as reflecting higher information content of the scene. Our interpretation is based on the notion that a shorter identification interval is consistent with rapid decision making and is therefore desirable from the point of view of evaluating a sensor fusion scheme. Of course, it is of critical importance that a fused sensor supports accurate target identification as well, and accordingly we computed the proportion of target identification responses that were correct (hereinafter ‘response accuracy’). Average response accuracy is another measure of the discriminability of the target information for a given sensor image condition.

The ANOVAs on total gaze duration on target, shown in Fig. 2 (panel c), revealed that gaze duration was longer in the LWIR condition than the SWIR condition (effect of Sensor: \(F(1,15) = 17.0, \, p < .001\)) and this difference varied in magnitude as a function of Distance (Sensor x Distance interaction: \(F(1,4,216) = 12.7, \, p = .001\)). Follow-up t-tests to assist interpretation of this interaction showed that LWIR had longer gaze duration than SWIR at Distance 1 and 2 (both \(t(15) > 2.8, \, p < .012\), but not Distance 3 (\(t(15) < 1\)). In the ANOVAs comparing LWIR and VIS, and SWIR and VIS, there were no significant differences between sensors (effect of Sensor: both \(Fs(1,15) < 6.8, \, p > .020\)), nor did these differences significantly differ as a function of Distance (Sensor x Distance interaction: \(Fs(2,30) < 4.8, \, p > .016\)). In comparing the FUSE and LWIR conditions, the FUSE sensor condition did not differ from the LWIR condition (effect of Sensor: \(F(1,15) = 1.5, \, p = .24\)) nor was there a Sensor x Distance interaction (\(F < 1\)). However, the FUSE condition did produce longer gaze duration than the SWIR condition overall (effect of Sensor: \(F(1,15) = 18.6, \, p = .001\), and this effect depended on target Distance (Sensor x Distance interaction: \(F(2,30) = 6.7, \, p = .004\)). In particular, gaze duration for SWIR exceeded that for FUSE at Distance 1 and 2 (both \(ts(15) > 2.9, \, p < .01\), but not Distance 3 (\(t(15) < 1\)).

The mean response accuracy for each sensor as a function of target distance is plotted in Fig. 2 (panel d). The ANOVA comparing LWIR and VIS confirmed that accuracy in the LWIR condition was lower than the VIS condition (effect of Sensor: \(F(1,15) = 8.9, \, p = .009\), with no significant interaction with target distance (\(F(2,30) = 2.2, \, p = .13\)). The LWIR condition also had lower accuracy than the SWIR condition (effect of Sensor: \(F(1,15) = 33.3, \, p < .001\), and the two-way interaction was not significant (\(F(2,30) = 4.4, \, p = .020\)). Accuracy for the VIS condition was in turn lower than the SWIR condition (effect of Sensor: \(F(1,15) = 19.0, \, p < .001\), and this difference increased with target Distance (sensor x Distance interaction: \(F(2,30) = 5.7, \, p = .008\), with follow-up t-test indicating a significant difference at Distances 2 and 3 (both \(t > 2.5, \, p < .006\)) but not Distance 1 (\(t(15) = 2.5, \, p = .025\)). The ANOVA comparing FUSE with SWIR confirmed that accuracy in the FUSE condition was poorer than in the SWIR condition (effect of Sensor: \(F(1,15) = 36.1, \, p < .001\)) and this difference increased over Distance (Sensor x Distance interaction: \(F(2,30) = 12.6, \, p < .001\), with follow-up t-tests pointing to significant differences at Distances 2 and 3 (both \(t > 4.6, \, p < .001\)) but not Distance 1 (\(t = 2.1, \, p = .050\)). However, the ANOVA comparing FUSE and LWIR indicated that accuracy in the FUSE condition was superior to that of the LWIR condition (effect of Sensor: \(F(1,15) = 11.3, \, p = .004\), while this effect did not interact with target distance (Sensor x Distance interaction: \(F(2,30) = 2.1, \, p = .14\)).

When considered together with the findings from the analysis of peripheral selection, it appears that the LWIR condition constitutes a trade-off where peripheral selection is facilitated at the expense of central processing. The poor central processing performance is likely due to target detail being poorly captured in the LWIR imagery. However, it should also be noted that the LWIR camera had the lowest resolution of the three cameras, which might have contributed to poor central processing efficiency. The SWIR condition showed the highest response accuracy and the shortest gaze durations, suggesting that central processing was most efficient in this condition. Somewhat surprisingly, the VIS condition exhibited lower response accuracy on average than the SWIR condition. The reason for this is not immediately clear. It is possible that the SWIR sensor captured more diagnostic information than the VIS sensor given the target discrimination that was required for the present search task.

Overall the FUSE condition in this experiment appeared to represent a compromise between the benefits and costs of its component sensors: the peripheral selection advantage of LWIR was diluted but still superior to SWIR, and the central processing advantage for SWIR was also muted, but there was still evidence for superiority to the LWIR condition. This outcome might reflect the simple fusion method used in Experiment 1. This possibility was further investigated in Experiment 2.

3. Experiment 2

Experiment 1 demonstrated the utility of eye movement-based measures for evaluating the perceptual efficiency of single-band and fused images. In Experiment 2 we applied this methodology to the evaluation of three different fusion schemes created using state-of-the-art fusion techniques. More specifically, we contrasted performance of fused sensory imagery with single-band
component imagery. This was done separately for each of three two-way fusion schemes (SWIR/LWIR, VIS/LWIR, SWIR/VIS) and for two different modern-image fusion methods. Of particular interest was whether or not the performance compromises for fused images that were observed in Experiment 1 would hold for state-of-the-art fusion techniques, or if instead these methods would produce fused imagery that matched or surpassed single-band performance. As a secondary goal of Experiment 2 we directly contrasted the two-way fusion schemes with one another, as well as the fusion methods, in order to determine whether particular schemes and methods are superior in terms of central processing and peripheral selection.

3.1. Method

3.1.1. Subjects

Eighteen members of the Canadian Armed Forces (17 males) participated in Experiment 2a (mean age = 26.9, s.d. = 7.3, 16 right-handed). A separate group of eighteen members of the Canadian Armed Forces (17 males) participated in Experiment 2b (mean age = 32.2 years, s.d. = 8.5; 14 right-handed). None of the subjects in Experiment 2 participated in Experiment 1. Subjects provided informed consent and were remunerated according to Government of Canada Treasury Board guidelines for a total of $12.72 CAD for their one hour of experiment participation. All subjects had normal or corrected-to-normal vision. The research protocol was reviewed and approved by the Human Research Ethics Committee at the DRDC Toronto Research Center.

3.1.2. Apparatus

See Experiment 1.

3.1.3. Materials and design

The visible spectrum, short-wave infrared, and long-wave infrared imagery used in Experiment 1 were also used in Experiment 2. However, in Experiment 2 the single-band images were subjected to pre-processing and fusion according to state-of-the-art fusion methods. In Experiment 2a the single-band images were fused using an implementation of the Jang method [41] (“Jang”). It consisted of a sub-band decomposition operator applied to each Retinex frequency plane, with an automatic fusion rule based on relative image content. In Experiment 2b, we applied an experimental fusion method (“Custom”) where a weighted-average fusion rule was applied to the two layers produced by the tone-mapping algorithm proposed by Durand and Dorsey [42]. The input images were first decomposed into a ‘base’ layer and a ‘details’ layer with an edge-preserving bilateral filter applied in the log domain. ‘Base’ and ‘details’ layers of both sensors were then combined with a weighted-average fusion rule, where the fusion weights were chosen to subjectively maximize the quality of the image (the result tended to be biased towards LWIR content). Finally, fused ‘base’ and ‘details’ layers are recombined with a base contrast factor of five as proposed in [42] and [43]. This base contrast factor compresses the ‘base’ layer (lower spatial frequencies), and consequently, enhances image details (higher frequencies).

For both Experiment 2a and 2b there were six conditions: three single-band conditions (VIS, SWIR, LWIR) and three fusion conditions (SWIR/LWIR, VIS/LWIR, SWIR/VIS). Because contrast enhancement is integral to both fusion methods, it was imperative to produce single-band imagery that underwent the same contrast enhancement in order to provide a baseline against which to compare the fused images. Accordingly, the single-band conditions in Experiment 2a and 2b differed according to the pre-processing (including contrast enhancement) that they had undergone. Samples of these images are shown in Fig. 3. The size of the fused images varied according to the fusion scheme: SWIR/LWIR = 1280 × 1024; VIS/LWIR = 1280 × 1024; SWIR/VIS = 1220 × 1024. Images were mounted in the center of a 1920 × 1080 grey background (R = G = B = 70). As in Experiment 1, a second version of each image was created by mirroring it in the vertical axis; this produced a total of 6 target positions (3 target positions mirrored) for each character at each distance for a total of 72 images per sensor condition. In summary, for Experiments 2a and 2b were within-subjects designs with six Sensor conditions (VIS, SWIR, LWIR, SWIR/LWIR, VIS/LWIR, SWIR/VIS) and three Target Distances (1, 2, 3). There were a total of 432 trials in each experiment. Each subject saw each scene six times, once in each sensor condition and in a random order, though the order of scene presentation was counterbalanced across every six subjects such that each scene was equally likely to appear in each sensor condition first, second, third, etcetera.

3.1.4. Procedure

The experimental procedure was analogous to Experiment 1, but with additional conditions and consequently a greater number of trials, Experiments 2a and 2b each took just over an hour.

3.2. Results and discussion

The analyses of the data for Experiments 2a and 2b are presented in two sections. In the first section, performance of fused imagery from each two-way fusion scheme was compared against its component single-band sensors (e.g. SWIR/LWIR vs. SWIR and LWIR). As in Experiment 1, we employed measures of peripheral selection and central processing efficiency. We applied a convention to summarize the outcomes of fusion: the fused condition might perform better than its two component sensors (‘enhancement’), not different from the better of the two component sensors but better than the worse of the two component sensors (‘improvement’), worse than the better of the two component sensors and better than the worse of the two (‘compromise’), worse than the better of the two and not different from the worse of the two (‘inefficiency’), or worse than both component sensors (‘interference’). Another possible outcome is that a fusion condition could not be statistically differentiated from either of its component sensors for a particular behavioural measure (‘inconclusive’). In the second section, we directly contrasted fusion schemes against one
another in order to identify the scheme with superior peripheral selection and central processing efficiency. We also included a between-subjects analysis that contrasted the fusion methods (Jang method, Custom method) to determine if one method or the other produced superior performance. Trials in which the subject did not respond were excluded from analysis (3% in Exp. 2a, 4% in Exp. 2b).

3.2.1. Fused versus single-band imagery

Each two-way fusion scheme (SWIR/LWIR, VIS/LWIR, SWIR/VIS) was compared against each of its component single-band sensors in a 2 × 3 Sensor (Fused, Single-band) × Distance (Distance 1, 2, 3) within-subjects ANOVA. Therefore there were six ANOVAs per Experiment, and accordingly we set the alpha to .0083 in order to obtain a family-wise error rate of .05. Where follow-up t-tests were applied to assist with the interpretation of significant Sensor × Distance interactions (comparing Sensor conditions at each of the three Distance settings), we applied an alpha of .016. The analyses are symmetrical for Experiments 2a and 2b and accordingly the results are discussed together. Measures of peripheral selection (detection interval, saccade amplitude) and central processing (gaze duration on target, response accuracy) are discussed in turn, followed by a summary of the outcomes for each fusion scheme.

Detection intervals are presented in Fig. 4 and as can be seen in the Figure, the rank order of single band sensors was the same as for Experiment 1 (LWIR < SWIR < VIS). The ANOVAs revealed that, consistent with the findings of Experiment 1, for both Experiments 2a and 2b (Fig. 6, panels a, b) the detection interval for SWIR/LWIR was significantly longer than LWIR (effect of Sensor: both Fs(1,17) > 105.6, ps < .001), and this difference increased with target distance for both experiments (Sensor × Distance interaction: Exp. 2a: F(2,34)= 40.7, p < .001; Exp. 2b (F(1.5,25.3)= 24.1, p < .001). Detection times for SWIR/LWIR were also significantly shorter than SWIR for both Experiments (effect of Sensor: both Fs(1,17) > 28.9, ps < .001) and this difference increased with target distance (Sensor × Distance interaction: both Fs(2,34)= 12.4, ps < .001). Follow-up t-tests between sensors at each distance revealed that for the Jang method (Exp. 2a), detection time for SWIR/LWIR was significantly shorter than SWIR at Distance 3 (t(17)= 5.2, p < .001) but not significant at Distance 1 or 2 (both ts < 2.4, ps > .033). For the Custom method (Exp. 2b), SWIR/LWIR had a significantly shorter detection interval at Distance 1 and Distance 3 (both ts > 7.0, ps < .001) but not Distance 2 (t(17)= 1.9, p = .070). Detection intervals for VIS/LWIR were longer than the LWIR condition in both Experiments (effects of Sensor: both Fs(1,17) > 165.7, ps < .001) and this difference increased as a function of target distance (Sensor × Distance interactions: both Fs(2,34)= 32.1, ps < .001). For both Experiments, VIS/LWIR (Fig. 4, panels c, d) produced detection intervals that were shorter than the VIS condition (effects of Sensor: both Fs(1,17) > 19.9, ps < .001), and this difference changed with target distance for both Experiments 2a (Sensor × Distance interaction: F(1.5,25.0)= 22.7, p < .001) and 2b (Sensor × Distance interaction: F(2,34)= 21.0, p < .001). However, on inspection of the Figure (panel c) it appears that for the Jang method (Exp. 2a) the difference between the VIS/LWIR and VIS conditions is primarily driven by Distance 3 and this was confirmed by post-hoc t-tests (Dist. 1, 2: both ts(17) < 1.5, ps > .17; Dist. 3: t(17)= 5.3, p < .001). For the Custom method (Exp. 2b), post-hoc t-tests revealed significant differences for all Distances (all ts(17) > 3.2, ps < .005). For both Experiments the SWIR/VIS scheme (Fig. 4, panels e, f) produced detection intervals that were significantly shorter than the VIS condition (effect of Sensor: both Fs(1,17) > 48.7, ps < .001), and this difference increased as a function of target distance (Sensor × Distance interaction: Exp. 2a: F(1.5,25.2)= 18.9, p < .001; Exp. 2b: F(1.2,20.0)= 36.0, p < .001). Importantly, the SWIR/VIS condition did not differ significantly from the SWIR condition for either Experiment 2a (effect of Sensor: F(1,17)= 4.9, p = .041; Sensor × Distance interaction: F(2,34)= .11, p = .34) or 2b (effect of Sensor: F(1,17)= .011, p = .92; Sensor × Distance interaction: both F(1,12.20.4)= .38, p = .58).

Mean values for the amplitude of the saccade prior to the detection event are plotted in Fig. 5. Overall, the saccade amplitude prior to detection did not strongly differentiate the fused and single-band sensors. For both Experiments, saccade amplitude for SWIR/LWIR (panels a, b) was not significantly different from LWIR (effect of Sensor: both Fs(1,17) < 6.4, ps > .021; Sensor × Distance interaction: both Fs(2,34) < 3.0, ps > .068), nor did it differ significantly from SWIR (effect of Sensor: both Fs < 1; Sensor × Distance interaction: both Fs(2,34) < 1.2, ps > .34). For VIS/LWIR (Fig. 5 panels c, d), the outcome was very similar across fusion methods. For the Jang method (Exp. 2a), VIS/LWIR did not differ from VIS or LWIR (effect of Sensor: Fs < 1; Sensor × Distance interaction: F(2,34) < 3.7, ps > .034). For the Custom method (Exp. 2b), saccade amplitude for VIS/LWIR was not significantly different from LWIR overall (effect of Sensor: F(1,17)= 3.9, p = .064), and while there was a significant Sensor × Distance interaction (F(2,34)= 6.3, p = .005), follow-up t-tests did not reveal significant Sensor differences for any of the target distances (all ts < 1.9, ps > .30). The VIS/LWIR condition for Experiment 2b also did not differ significantly from VIS (effect of Sensor: F = 1; Sensor × Distance interaction: F(1,17)= 4.6, p = .017). For SWIR/VIS (Fig. 5 panels e, f), saccade amplitudes were very similar for all sensor conditions and indeed SWIR/VIS did not differ significantly from the single-band conditions for either experiment (effect of Sensor: all Fs(1,17) < 2.7, ps > .12; Sensor × Distance interactions: all Fs(2,34) < 5.1, ps > .012). Total gaze duration on target is presented in Fig. 6. For both Experiments, gaze duration for SWIR/LWIR (panels a, b) was much shorter than LWIR (effect of Sensor: both Fs(1,17) > 21.0, ps < .001) but this difference did not depend on Distance (Sensor × Distance interaction: both Fs(2,34) < 5.2, ps > .011). In contrast, SWIR/LWIR did not differ from SWIR for both Experiments (effect of Sensor: both Fs(1,17) < 6.1, ps > .025; Sensor × Distance interaction: both Fs(2,34) < 1). Gaze duration in VIS/LWIR (panels c, d) was substantially shorter than LWIR in both Experiments (effect of Sensor: both Fs(1,17) > 21.9, ps < .001) and this difference did not vary as a function of Distance (Sensor × Distance interaction: both Fs(2,34) < 2.9, ps > .071). However when compared against VIS, VIS/LWIR was not significantly different from VIS in either Experiment (effect of Sensor: both Fs(1,17) < 6.2, ps > .024; Sensor × Distance interaction: both Fs(2,34) < 1.5, ps > .25). For SWIR/VIS (panels e, f) all conditions were quite similar. In the comparison of SWIR/VIS against VIS there were no significant effects for either Experiment 2a (effect of Sensor: F(1,17) < 1; Sensor × Distance interaction: F(1,4,24.6)= 3.0, p = .082) or 2b (effect of Sensor: F(1,17)= 1.5, p = .24; Sensor × Distance interaction: F(2,34) < 1). When compared against SWIR, for the Jang (Exp. 2a) and Custom methods (Exp. 2b) there was no significant main effect of Sensor (both Fs(1,17) < 2.1, ps > .16). However, for Experiment 2b there was a significant Sensor × Distance interaction (F(2,34)= 5.6, p = .0077) and a similar marginally significant effect in Experiment 2a (F(2,34)= 5.5, p = .0087). Post-hoc t-tests at each Distance revealed that for both fusion methods, gaze duration in SWIR/VIS was significantly shorter than the SWIR at Distance 2 (both ts(17) > 2.9, ps < .0076), but for the other Distance settings the difference was not significant (all ts(17) < 2.0, ps > .060). This finding, though subtle, is potentially important because it provides evidence of a fused sensor condition that outperforms its single band component sensors.

Response accuracy is presented in Fig. 7. For both Experiments 2a and 2b, response accuracy in SWIR/LWIR (panels a, b) was sig-
Fig. 4. Experiment 2: Detection interval for each fusion scheme compared to its single-band component sensors, plotted separately for Experiment 2a (panels a, c, e) and 2b (panels b, d, f). Error bars represent standard error across subjects.

Response accuracy in VIS/LWIR (panels c, d) was also higher than LWIR for both Experiments (effect of Sensor: both $F(1,17) > 100.6$, $p < .001$). For Experiment 2b this difference increased over Distance settings (Sensor x Distance interaction: $F(2,34) = 7.4$, $p = .002$), while for Experiment 2a the interaction was not significant ($F(1.5,25.1) = 3.4$, $p = .065$). In contrast, when compared against VIS, response accuracy for VIS/LWIR did not differ significantly for either experi-
Fig. 5. Experiment 2: Saccade amplitude prior to detection for each fusion scheme compared to its single-band component sensors, plotted separately for Experiment 2a (panels a, c, e) and 2b (panels b, d, f). Error bars represent standard error across subjects.

ment (effect of Sensor: both $F_{(1,17)} < 8.1, p > .011$; Sensor x Distance interaction: both $F_{(2,34)} < 2.3, p > .12$). In the evaluation of response accuracy for SWIR/VIS (Fig. 7 panels b, c), the comparison with VIS (panels b, c) failed to produce a significant effect of Sensor for either fusion method (both $F_{(1,17)} < 7.4, p > .015$), though there was a Sensor x Distance interaction for the Custom fusion method (Expt. 2b: $F_{(2,34)} = 9.1, p = .001$) but not the Jang fusion method (Expt. 2a: $F_{(2,34)} = 5.3, p = .010$). This interaction in Experiment 2b was driven by the higher accuracy for SWIR/VIS than VIS at Distance 2 ($t_{(17)} = 4.8, p < .001$) but not for the other Distances (all $t_{(17)} < 1.6, p > .13$). In comparing SWIR/VIS against SWIR there were no significant main effects of Sensor for both Experiments (both $F_{(1,17)} < 1$), though once again there were significant Sensor x Distance interactions, this time for both Experiments.
Fig. 6. Experiment 2: Gaze duration on target for each fusion scheme compared to its single-band component sensors, plotted separately for Experiment 2a (panels a, c, e) and 2b (panels b, d, f). Error bars represent standard error across subjects.

In both Experiments this interaction was driven primarily by increased accuracy for SWIR/VIS compared to SWIR at Distance 2 (both ts(17) > 3.3, ps < .004) but not for the other Distances (all ts(17) < 2.7, ps > .018). This corroborates the evidence from the gaze duration measure of enhanced performance for the SWIR/VIS fused sensor.

A summary of the comparisons of fusion schemes against component sensors for Experiments 2a and 2b is presented in Table 2. Overall the pattern of findings was remarkably similar across fusion methods (Expts. 2a and 2b), though there were subtle differences. As can be seen in the Table, fusion schemes involving LWIR tended to result in ‘compromised’ peripheral selection performance. This is consistent with the outcome of Experiment 1.
In contrast, the SWIR/VIS fusion schemes tended to have ‘efficient’ peripheral selection performance. This might be due, in part, to the fact that the difference in detection interval between SWIR and VIS is small, and thus ‘compromised’ outcomes may be less statistically detectable. In addition, for both Experiments the saccade amplitude measure did not statistically differentiate fused sensor conditions from single band sensors, which might indicate that it is a relatively noisy measure. With regards to central processing efficiency, in contrast to Experiment 1 where central processing was ‘compromised’ for fused images, in Experiment 2 both fusion methods (Jang, Custom) produced fused images whose performance matched the better of the two component sensors (‘efficient’). Moreover, for SWIR/VIS, the fused images were shown to outperform both single-band conditions in some cases. This ‘en-
hancement’ outcome is clearly a desirable feature of a sensor fusion method and scheme.

3.2.2. Comparison of two-way fusion schemes and fusion methods

In this section we directly contrasted the three two-way fusion schemes, as well as the two fusion methods (Jang, Custom). For each behaviour measure, we conducted a ‘grand’ 3 × 3 × 2 mixed ANOVA crossing Sensor (SWIR/LWIR, VIS/LWIR, SWIR/VIS) and Distance (1, 2, 3) as within-subject factors and Fusion Method (Expt. 2a: Jang, Expt. 2b: Custom) as a between-subjects factor. Of particular interest were effects and interactions of Sensor and Fusion Method. Significant interactions were interpreted via follow-up ANOVAs and t-tests. Each of the behavioural measures were analysed separately and are discussed in turn.

Detection interval for each fusion scheme and each fusion method separated by experiment are presented in Fig. 8 (panels a, b). The grand ANOVA on detection time revealed a significant main effect of Sensor (F(2,68) = 84.9, p < .001) as well as a Sensor x Distance interaction (F(2.99,60) = 10.6, p < .001). Neither of the interactions with Fusion Method approached significance (Sensor x Fusion Method: F(2,68) = 1.6, p = .20; Sensor x Distance x Fusion Method: F < 1). In order to clarify the cause of the Sensor x Distance interaction, we conducted a one-way ANOVA on Sensor at each Distance setting, collapsing across the Fusion Method factor. In order to control the family-wise error rate for these three tests, we adopted an alpha of .016. This revealed a significant effect of Sensor at each Distance setting (Distance 1: F(1.6,55,1) = 36.1, p < .001; Distance 2: F(1.5,51.6) = 19.9, p < .001; Distance 3: F(2,70) = 41.4, p < .001). Three follow-up t-tests were conducted comparing Sensors at each Distance setting, and we also adopted an alpha of .016 for these tests. These tests revealed that in each case, detection time in SWIR/LWIR was shorter than VIS/LWIR (all ts(35) > 6.0, ps < .001) and SWIR/VIS (all ts(35) > 4.8, ps < .001), but that VIS/LWIR and SWIR/VIS did not differ from one another (all ts(35) < 2.6, ps > .016). Finally, in the grand ANOVA there was a main effect of Fusion Method (F(1,34) = 8.4, p = .007), where the Custom method (Expt. 2b) produced shorter detection intervals than the Jang method (Expt. 2a).

Saccade amplitude prior to detection is presented in Fig. 8 (panels c, d). On inspection of the Figure, saccade amplitude was similar for all fusion schemes, and indeed the fusion conditions were not found to differ significantly in the ANOVA (effect of Sensor: F(2,68) = 2.8, p = .067), nor were there any significant interactions with Sensor (Sensor x Distance: F(4136) = 1.5, p = .19; Sensor x Distance x Fusion Method: F < 1). In addition, there was no significant effect of Fusion Method (F < 1).

Gaze duration on target is presented in Fig. 8 (panels e, f). While there was no main effect of Sensor (F(2,68) = 1.4, p = .26), there was a significant Sensor x Distance interaction (F(3.1106,8) = 7.9, p < .001). The interactions between Sensor and Fusion Method did not approach significance (both Fs < 1), and the between-subjects effect of Fusion Method was not significant (F < 1). To clarify the cause of the Sensor x Distance interaction, we conducted follow-up one-way ANOVAs on Sensor at each Distance setting. This revealed a significant effect of Sensor at Distances 1 (F(1,6,56.2) = 6.2, p = .005) and 2 (F(1,7,59.7) = 6.3, p = .005), but not Distance 3 (F(2,68) = 3.9, p = .024). Follow-up t-tests revealed that for Distance 1 the SWIR/LWIR fusion scheme produced shorter gaze durations than the VIS/LWIR scheme (t(35) = 3.0, p = .005) and the SWIR/VIS scheme (t(35) = 2.6, p = .014), while VIS/LWIR, SWIR/VIS comparison was not significant (t < 1). In contrast, for Distance 2, gaze durations for the SWIR/LWIR scheme were longer than the VIS/LWIR scheme (t(35) = 2.8, p = .008) and the VIS/VIS scheme (t(35) = 3.4, p = .002), and the VIS/LWIR, SWIR/VIS comparison was not significant (t < 1).

Response accuracy for each fusion scheme is presented in Fig. 8 (panels g, h). In the ‘grand’ ANOVA on response accuracy the main effect of Sensor was not significant (F < 1), however there was a significant Sensor x Distance interaction (F(3,6122.2) = 11.9, p < .001). The Sensor x Fusion Method interaction was not significant (F < 1) nor was the Sensor x Distance x Fusion Method interaction (F(4136) = 3.0, p = .020). The between-subjects factor of Fusion Method was also not significant (F < 1). To interpret the Sensor x Distance interaction, we conducted follow-up one-way ANOVAs on Sensor at each Distance setting. This revealed a significant effect of Sensor at Distance 1 and Distance 2 (both Fs(2,70) > 5.1, ps < .008), but not Distance 3 (F(2,70) = 2.1, p = .13). Follow-up t-tests between Sensor conditions at each Distance revealed that at Distance 1, SWIR/LWIR had higher response accuracy than VIS/LWIR (t(35) = 35, p = .010), and also tended to have a higher accuracy than SWIR/VIS (t(35) = 2.5, p = .017), while VIS/LWIR and SWIR/VIS did not differ significantly (t < 1). For Distance 2, the pattern was different: SWIR/LWIR had a lower response accuracy than SWIR/VIS (t(35) = 5.4, p < .001), but was not significantly different than VIS/LWIR (t(35) = 1.8, p = .078). Finally, at Distance 2 the SWIR/VIS scheme had significantly higher accuracy than the VIS/LWIR scheme (t(35) = 3.0, p = .005).

To summarize, we found robust differences between fusion schemes and fusion methods across measures of peripheral selection and central processing. In terms of peripheral selection, detection intervals were shortest for the SWIR/LWIR fusion scheme compared to the other two schemes (VIS/LWIR, SWIR/VIS), which did not differ from each other, and detection intervals were also shorter for imagery generated using the Custom method vs. the Jang method. Saccade amplitude did not discriminate between fusion schemes or fusion methods. For central processing measures, the relative performance of the fusion schemes depended on target range. In particular, for Distance 1 targets, the SWIR/LWIR fusion scheme tended to have shorter gaze duration on target and higher response accuracy than the other two, while at Distance 2 the SWIR/LWIR condition had longer gaze durations and lower response accuracy than the other two schemes. This change in the ranking of fusion schemes across target distances indicates that stimulus characteristics play an important role in determining sensor performance. The two fusion methods (Jang, Custom) did not differ along either measure of central processing.
Fig. 8. Experiment 2: Comparison of fusion schemes for each behavioural measure within an experiment. Detection interval (panels a, b), saccade amplitude prior to detection (panels c, d), gaze duration on target (panels e, f), and response accuracy (panels g, h).
4. Stimulus modeling

Experiments 1 and 2 demonstrated the utility of eye movement-based measures of performance for determining the efficiency of fusion, and also in comparing different fusion schemes and methods. However, given that human experimentation, and eye movement recording in particular, is technical, time-consuming, and costly, it would be useful to be able to predict these behavioural outcomes based on image characteristics. Accordingly, we conducted a correlation analysis that explored the extent to which various image characteristics could predict performance variables.

4.1. Image characteristics

We considered image characteristics that measure the target-background contrast and the discriminability of target information. In order to capture target-background contrast, we computed target-background brightness contrast, target-background structural similarity (TSSIM), target-background correlation coefficient, and target-background salience contrast. To capture the discriminability of the target information we computed root-mean-squared (RMS) target contrast, target-target TSSIM, and target-target correlation coefficient. Each image characteristic will be described in turn followed by a correlation analysis that estimates the extent to which these variables predict performance data from Experiment 2.

4.1.1. Target-background brightness contrast

Target-background brightness contrast was computed using the Weber method. For each visual search image in each condition, a target cut-out was created by hand that included only the pixels occupied by the search target. We then computed the average pixel value for that target cut-out (\( \text{Brightness}_{\text{Target}} \)) and the average pixel brightness for the remainder of the image (\( \text{Brightness}_{\text{Background}} \)), and then applied the following equation to obtain target-background brightness contrast:

\[
\frac{\text{Brightness}_{\text{Target}} - \text{Brightness}_{\text{Background}}}{\text{Brightness}_{\text{Background}}}
\]

4.1.2. Target-background TSSIM

The TSSIM metric was developed in order to provide a measure of image quality within the domain of digital imagery, and has been shown to predict certain aspects of performance in natural scene images [24,44]. TSSIM is a measure of target-background similarity, and is computed between image areas of the same size (in pixels). Accordingly, target cut-outs were obtained as in 4.1.1. We then divided the non-target area of each image into non-overlapping cells (width x height in pixels = 80 x 150, 40 x 70, and 26 x 40 for Distance 1, 2 and 3, respectively). For each of these non-target cells, we extracted a target-sized area defined using the target cut-out for that image as a mask. We then computed the TSSIM value (arithmetic mean) for each target cut-out/non-target area cut-out pair within each image, and averaged across all pairs to obtain a target-background TSSIM value for each image.

4.1.3. Target-background correlation coefficient

As an additional measure of target-background similarity, we computed the average correlation between the target and the background. As in Section 4.1.2, target cut-outs and non-target area cut-outs were obtained for each image and we computed the Pearson correlation (computed across pixels) and then averaged the correlation across all target/non-target area pairs.

4.1.4. Target-background salience contrast

Salience algorithms have been developed to account eye movement behaviour during the viewing of natural scenes, and have been shown to robustly predict certain aspects of performance (for a review see [45]). Target-background salience contrast was computed as follows: we applied the Signature Salience algorithm [46] to each image (as implemented by [47]) which produced a salience image with values ranging from 0–255. We then computed the average salience value for the target-cut-out region and the average value for the remainder of the image. The target-background salience contrast was then computed using the Weber method as in Section 4.1.1.

4.1.5. Target RMS contrast

Target RMS contrast for each image was defined as the standard deviation of brightness values within the target cut-out.

4.1.6. Target-target TSSIM

As a measure of target-target similarity, we computed the average TSSIM for the target cut-out with all other different character target cut-outs that were within the same sensor condition and Distance setting (i.e. average similarity with three other target Characters at each of 3 positions). Because target cut-outs varied slightly in size across Characters, the target-target TSSIM for every target pairing was computed for the subset of pixels for which the two target cut-outs overlapped (i.e. the intersection) when aligned by the pixel in the center of the target’s head.

4.1.7. Target-target correlation coefficient

As an additional measure of target-target similarity, we computed the pixel-wise Pearson correlation between target cut-outs (as in Section 4.1.6).

4.2. Correlations with performance

We computed correlation coefficients to quantify the extent to which each of the image characteristics accounted for the performance of sensor conditions. For each Sensor (6) x Distance (3) setting in Experiment 2a and 2b, we computed the average image metric and performance value (averaged across Position and Character, and then averaged across subjects). We then computed the Pearson correlation between each image metric and each performance measure, computed across all Sensor x Distance settings from both experiments (n = 36). The correlation for each image metric with each performance variable is presented in Table 3. Associations with measures of peripheral selection and central processing are discussed in turn.

Inspection of Table 3 shows that several of the image metrics had significant correlations with detection interval. Among these, the strongest predictor of detection interval was target-target TSSIM. Conditions with higher target-target similarity were associated with longer detection intervals. While somewhat counterintuitive, we speculate that measures of target-target similarity might capture a camouflage effect: when targets are similar to one another, they are also likely to blend into the background and be difficult to detect. Consistent with this interpretation, target-background TSSIM was also positively correlated with detection interval. We also noted that salience contrast was also a moderately strong predictor of detection interval: images with high target-background salience contrast had shorter detection intervals.

There were also several significant predictors of saccade amplitude prior to detection. The strongest predictor was target-background salience contrast, where images with higher salience contrast had longer saccades prior to detection. This is consistent with salience as a model of saccadic selectivity in scene viewing [45]. Target-background salience contrast was unique among the
measures examined here in that it produced a significant association with both measures of peripheral selection. High target-background correlation was also associated with long saccades. This reason for this is less clear, as one might expect a correlation to with the background to impair peripheral selection and result in shorter target detection saccades. In addition, there was a significant association between target-target correlation and saccade amplitude, where high correlations predicted short saccade amplitudes.

There were three significant predictors of central processing measures: target-background brightness contrast, target-background TSSIM, and target-target correlation. In particular, increases in target-background contrast and target-target correlation were associated with increased gaze duration and reduced response accuracy. We speculate that these effects might be driven in part by the presence of LWIR information. Sensor conditions that present LWIR information would tend to have high target-background brightness contrast and high target-target correlation. High target-target similarity would be expected to increase target confusion and consequently lengthen gaze duration on target and impair response accuracy. In contrast, images with high target-background TSSIM tended to have high accuracy and short gaze duration. While somewhat difficult to interpret, this association might be related to the presence of visual detail; sensors that are rich in detail (e.g. SWIR, VIS) might facilitate target processing and also increase the structural similarity of target to the background (foliage in this case).

In summary, none of the image metrics tested provided a significant association with all performance measures, but rather image metrics tended to associate with either peripheral selection or central processing performance. Target-background salience contrast provided a robust prediction of both measures of peripheral selection. In contrast, target-background brightness contrast, target-background TSSIM, and target-target correlation were associated with central processing performance. Therefore, these image metrics might be considered as candidate estimators of peripheral selection and central processing performance for fused images. However, note that the observed correlations tended to be modest in magnitude, and consequently in the present context these image metrics would be unlikely to serve as a replacement for behavioural assessment of fused imagery.

5. General discussion

Eye movement analysis is an important empirical tool for research in the domain of scene perception. Recently this technique has been applied in the domain of sensor fusion, in order to provide quantitative measures of the perceptual processing efficiency of fused imagery. In support of this goal, the present study developed analyses of eye movements that distinguished between central and peripheral processing during visual search of infrared and fused images. In Experiment 1 we demonstrated the capability for eye movement-based measures of performance to differentiate sensor conditions. In particular, we contrasted single band visible spectrum (VIS), short-wave infrared (SWIR), long-wave infrared (LWIR) images, and a simple average SWIR/LWIR fusion scheme (FUSE). In Experiment 2, we applied this methodology to a wider array of two-way fusion schemes (SWIR/LWIR, VIS/LWIR, SWIR/VIS) and fusion methods (Jang method, Custom method). Stimulus modeling was also conducted in order to determine the extent to which image properties could predict the behavioural outcomes.

The results of Experiment 1 demonstrated that single-band and fused imagery conditions could be differentiated according to measures of perceptual efficiency. More specifically, the LWIR single-band imagery had more efficient peripheral selection than SWIR and VIS: detection intervals were shorter and saccades prior to fixation tended to have greater amplitude. In contrast, the SWIR single-band imagery produced more efficient central processing than VIS and LWIR: gaze duration on target tended to be shorter and response accuracy was higher. These differences also confirm the importance of separating measures of processing duration for central processing and peripheral selection (i.e. detection interval vs. gaze duration on target). For example, LWIR exhibited a short detection interval and long gaze duration on target while SWIR exhibited the opposite pattern. Measures of processing duration that sum across these two epochs, as have been applied in prior research (e.g. [31,32]), would have tended to mask these differences. Importantly, Experiment 1 demonstrated the sensitivity of these measures for the purpose of evaluating the effectiveness of image fusion. In this case, the FUSE condition produced both peripheral selection and central processing performance that was intermediate to that of its component sensors (SWIR, LWIR). This fusion outcome might be considered sub-optimal as it constitutes a compromise between the performance levels of the two single-band conditions.

In Experiment 2 we tested the generality of this finding across different fusion schemes and methods. Importantly, each fused sensor condition was compared against single-band component sensor conditions that were subject to the same contrast enhancement and pre-processing as the fused conditions. We introduced a taxonomy describing the performance outcome of fusion relative to its component sensors ranging from ‘interference’ to ‘enhancement’. We found that, consistent with Experiment 1, fusion schemes that incorporated LWIR (SWIR/LWIR, VIS/LWIR) exhibited ‘compromised’ peripheral selection performance for both fusion methods (Jang, Custom). Therefore, it would appear that compromised peripheral selection performance is common to fusion schemes that incorporate LWIR, though further research is required to confirm this. In contrast, peripheral selection performance for the SWIR/VIS fusion scheme was not distinguishable from the better of the two component sensors (SWIR), an outcome that was labeled ‘efficient’. However, in the case of SWIR/VIS, performance in the two component sensors was more similar to one another than for SWIR/LWIR and VIS/LWIR, which might reduce the chance of statistically detecting a performance compromise.
Considering central processing performance, we found that the fusion methods employed in Experiment 2 outperformed the simple averaging method employed in Experiment 1. In particular, for both the Jang and Custom methods, all two-way fusion schemes showed ‘efficient’ or ‘enhanced’ central processing performance. In particular, for the Jang method (Expt. 2a) there was evidence of central processing enhancement in the SWIR/VIS fusion scheme, and for the Custom method (Expt. 2b) for the VIS/LWIR and SWIR/VIS fusion schemes. These effects were small, yet are encouraging for applications of sensor fusion as they represent a clear benefit of fused imagery over single-band imagery. While further research is required to determine the cause of this performance enhancement, we speculate that it might reflect either the retention of complementary information from the component sensors, or the emergence of new information due to the contrast between information from the component sensors.

When directly contrasting the fusion schemes with one another, we observed only small differences in overall performance. In terms of peripheral selection, the SWIR/LWIR fusion scheme was superior, producing robustly shorter detection intervals. For central processing performance the ranking of fusion schemes depended on stimulus characteristics. In particular, for Distance 1 targets, SWIR/LWIR produced the shortest gaze durations and the highest response accuracy, whereas for Distance 2 targets, SWIR/VIS produced the shortest gaze durations and highest response accuracy. We speculate that this is because SWIR/VIS better preserved the high spatial frequencies required to make target discriminations at further target distances compared to SWIR/LWIR. This pattern of results underlines the importance of considering the characteristics of the viewing task and stimuli when assessing fusion performance. In fact, the stimuli used in the present study represent a narrow selection from the range of possible stimulus variations that are relevant in the context of visual search of infrared and fused images (e.g. day vs. night, ambient temperature, character apparel, human targets vs. vehicles or object, rural vs. urban scenery, etc.). Accordingly, the present findings regarding central and peripheral processing for these sensors must be interpreted with caution, and further research is required to explore the influence of stimulus variables on the differences between sensors observed in this study.

Experiment 2 also provided a comparison of two fusion methods. We documented a significant difference between the two methods, where detection intervals were shorter overall for the Custom imagery than for the Jang imagery. We did not detect any significant differences between the methods in the measures of central processing efficiency. This information may be useful for guiding the development of fusion algorithms. However, it must be noted that the fusion method entails multiple processing steps. For example, the Custom fusion method employed in Experiment 2b involved a weighted-average fusion rule. The choice of weights will likely influence the perceptual processing efficiency of the fused imagery. In addition, both fusion methods involve contrast enhancement prior to fusion which is likely to impact upon performance. Based on the present findings it is unclear which of these processing steps contribute to the observed difference in perceptual efficiency. Further research might seek to compare the perceptual efficiency of fusion methods in greater depth, and with attention paid to the effect of individual processing steps. Information emerging from such experimentation might be useful, in particular, for the development of fusion methods.

Finally, stimulus modeling was conducted to determine the extent to which image metrics could predict behavioural outcomes. A variety of image metrics were computed that captured target-background contrast and target-target similarity. We observed moderate correlations between several image metrics and performance for peripheral selection and central processing measures. In particular, target-background salience contrast predicted both measures of peripheral selection, while target-background brightness contrast, target-background TSSIM and target-target correlation were predictive of central processing efficiency. These metrics are therefore expected to be of some utility in predicting the perceptual efficiency of fused images. However, because these correlations were only moderate in magnitude, they are unlikely to be a suitable replacement for behavioural evaluation of the perceptual efficiency of fused images.

In conclusion, the present findings clearly demonstrate the potential utility of eye movement recordings for assessing the performance of infrared and fused sensor images. Eye movements yielded measures of central and peripheral processing that differentiated single-band and fused sensor conditions. More specifically, these measures were sensitive to differences between the fused sensor condition and its component sensors, and we proposed a classification scheme to express these outcomes. In addition, the present data demonstrated that eye movement-based measures can be used to directly compare the efficiency of sensor fusion methods, and this information could provide feedback during their design cycle. Finally, while the computational image metrics employed here were found to predict central and peripheral processing performance to some extent, the present findings imply a unique role for eye movement recordings in obtaining diagnostic information on the perceptual processing efficiency of infrared and fused images.

Acknowledgements

We would like to thank David Alain and Louis Durand for collecting the imagery that was used in this study.

References


**DOCUMENT CONTROL DATA**

(Security markings for the title, abstract and indexing annotation must be entered when the document is Classified or Designated)

<table>
<thead>
<tr>
<th>1. ORIGINATOR (The name and address of the organization preparing the document. Organizations for whom the document was prepared, e.g., Centre sponsoring a contractor's report, or tasking agency, are entered in Section 8.)</th>
<th>2a. SECURITY MARKING (Overall security marking of the document including special supplemental markings if applicable.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DRDC – Toronto Research Centre Defence Research and Development Canada 1133 Sheppard Avenue West P.O. Box 2000 Toronto, Ontario M3M 3B9 Canada</td>
<td>CAN UNCLASSIFIED</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>2b. CONTROLLED GOODS</th>
</tr>
</thead>
<tbody>
<tr>
<td>NON-CONTROLLED GOODS DMC A</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>3. TITLE (The complete document title as indicated on the title page. Its classification should be indicated by the appropriate abbreviation (S, C or U) in parentheses after the title.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evaluation of fused imagery using eye movement-based measures of perceptual processing</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>4. AUTHORS (last name, followed by initials – ranks, titles, etc., not to be used)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glaholt, M.G.; Sim, G.; Laou, P.; Roy, S.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>5. DATE OF PUBLICATION (Month and year of publication of document.)</th>
<th>6a. NO. OF PAGES (Total containing information, including Annexes, Appendices, etc.)</th>
<th>6b. NO. OF REFS (Total cited in document.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>April 2017</td>
<td>18</td>
<td>48</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>7. DESCRIPTIVE NOTES (The category of the document, e.g., technical report, technical note or memorandum. If appropriate, enter the type of report, e.g., interim, progress, summary, annual or final. Give the inclusive dates when a specific reporting period is covered.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>External Literature (P)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>8. SPONSORING ACTIVITY (The name of the department project office or laboratory sponsoring the research and development – include address.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DRDC – Toronto Research Centre Defence Research and Development Canada 1133 Sheppard Avenue West P.O. Box 2000 Toronto, Ontario M3M 3B9 Canada</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>9a. PROJECT OR GRANT NO. (If appropriate, the applicable research and development project or grant number under which the document was written. Please specify whether project or grant.)</th>
<th>9b. CONTRACT NO. (If appropriate, the applicable number under which the document was written.)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>10a. ORIGINATOR’S DOCUMENT NUMBER (The official document number by which the document is identified by the originating activity. This number must be unique to this document.)</th>
<th>10b. OTHER DOCUMENT NO(s). (Any other numbers which may be assigned this document either by the originator or by the sponsor.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DRDC-RDDC-2018-P007</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>11a. FUTURE DISTRIBUTION (Any limitations on further dissemination of the document, other than those imposed by security classification.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unlimited</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>11b. FUTURE DISTRIBUTION OUTSIDE CANADA (Any limitations on further dissemination of the document, other than those imposed by security classification.)</th>
</tr>
</thead>
</table>
Human performance measures were used to evaluate the perceptual processing efficiency of infrared and fused-infrared images. In two experiments, eye movements were recorded while subjects searched for and identified human targets in forested scenes presented on a computer monitor. The scenes were photographed simultaneously using short-wave infrared (SWIR), long-wave infrared (LWIR), and visible (VIS) spectrum cameras. Fused images were created through two-way combinations of these single-band images. In Experiment 1 the single band sensors were contrasted with a simple average fusion scheme (SWIR/LWIR). Analysis of subjects’ eye movements revealed differences between sensors in measures of central processing (gaze duration, response accuracy) and peripheral selection (detection interval, saccade amplitude). In Experiment 2 this methodology was applied to compare three two-way combinations of sensors (SWIR/LWIR, SWIR/VIS, VIS/LWIR), produced by state-of-the-art fusion methods. Peripheral selection for fused images tended to exhibit a compromise between the performance levels of component sensor images, while measures of central processing showed evidence that fused images matched or exceeded the performance level of component single-band sensor images. Stimulus analysis was conducted to link measures of central and peripheral processing efficiency to image characteristics (e.g. target contrast, target-background contrast), and these image characteristics were able to account for a moderate amount of the variance in the performance across fusion conditions. These findings demonstrate the utility of eye movement measures for evaluating the perceptual efficiency of fused imagery.

n/a

Infrared imagery, sensor fusion, eye movements, peripheral vision