Information Fusion 39 (2018) 186–203

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

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.

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

☆ Her Majesty the Queen in Right of Canada, as represented by the Minister of National Defence, 2016; Licensed material is licenced pursuit to the *Creative Commons Attribution-Noncommercial-NoDerivative Works 3.0 Unported License* (CC-BY-NC-ND); Provisions of Section 5 of the above noted license apply to the licenced material.

* Corresponding author.
E-mail address: Mackenzie.Glaholt@drdc-rddc.gc.ca (M.G. Glaholt).

http://dx.doi.org/10.1016/j.inffus.2017.04.003
1566-2535/Crown Copyright © 2017 Published by Elsevier B.V. All rights reserved.

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 composites 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