Assessing Multispectral Image Fusion with Systems Factorial Technology

Hanshu Zhang and Joseph W. Houpt
Wright State University, Dayton, OH, USA, 45435

Despite the rich literature on techniques for creating a single image from multispectral sensors, there is relatively little research on methods for assessing these techniques based on human performance. We propose the use of Systems Factorial Technology (SFT), a nonparametric, mathematical modeling framework for analyzing human cognition. Previous work has demonstrated the use of SFT in evaluating human perception of multi-spectral imagery, although on relatively contrived tasks. In this work, we extend the approach to a task in which observers must determine whether a person in the image is holding a gun or a tool. We found that all observers processed the information from each spectrum less efficiently when images based on two different spectra were presented together, regardless of whether the information was fused into a single image or kept separately. Furthermore, when images from the two different spectra were presented side-by-side, some observers were able to use both sources in parallel.

Image fusion is the process of combining information from two or more images of a scene into a single composite image that ideally is more informative and is more suitable for visual perception or computer processing. By combining the output of different sensors that capture complementary information in the environment, image fusion offers the potential of extending the information available to users without overwhelming their perceptual systems. Despite the potential advantage for human cognition and perception, relatively little research has focused on human performance with fused images. Indeed, Muller and Narayanan (2009) stressed that the cognitive aspects of multiple-sensors had not received as much attention as technical aspects of combining multisensory image information.

Nonetheless, cases in which researchers have explored the potential perceptual and cognitive advantages of image fusion span many domains including medical imaging (Deshmukh & Bhosale, 2010), night-time driving (McCarley & Krebs, 2000), and concealed weapon detection (Xue, Blum, & Li, 2002; Xue & Blum, 2003). Although these studies often indicate advantages of fused imagery, improved performance of observers is not guaranteed, and it is unclear in which situation image fusion would be helpful. The three main factors that vary across these studies that may have contributed to heterogeneous conclusions are: algorithm used, task performed and scene content.

In most cases these studies have focused on the comparison between single-sensor images and fused imagery but we argue that side-by-side presentation of images from different sensors should also be considered as a baseline of comparison. For example, Nogami et al., (2007) demonstrated that although fused images were clinically valuable, side-by-side presentation showed equivalent performance. More recently, Fox (2015) found that side-by-side presentation of images results in equivalent or even enhanced performance as fused images. While Fox (2015) focused on more generic tasks (e.g., discriminating the direction which a subject in the scene is facing), in this paper we extend the investigation to a more applied task: weapon/non-weapon discrimination.

In the current design, we compared whether multi-sensor images presented side-by-side or fused by algorithmic method led to more efficient performance. We refer to the side-by-side presentation as cognitive fusion because the combination of information across sensor images occurs in the cognitive processing by the observer. Different styles of cognitive fusion are possible across observers—e.g., focusing on one-side-at-a-time or simultaneously.

To examine both how efficiently observers are using multispectral imagery and to examine processing strategies for cognitive fusion across observers, we used systems factorial technology (SFT). SFT is a framework for studying how different sources of information combine in cognitive processing (Townsend & Nozawa, 1995; Houpt, Blaha, McIntire, Havig, & Townsend, 2014). SFT includes measurements that are informative with regards to architecture, stopping rule, workload capacity and stochastic dependence. Here we use architecture to mean the temporal organization of information processing; whether from each source could is used one at a time in a sequence (serial) or simultaneously (parallel). Workload capacity is how the processing rate of each source changes as more sources are added. The stopping rule refers to the whether one (OR rule) or both sources (AND rule) are processed before responding. Stochastic dependence refers to how the processing of each source of information interacts with the processing of others.

The SFT measure of workload capacity, the capacity coefficient, is a comparison between the predicted performance of an unlimited-capacity, independent and parallel (UCIP) system and the observed performance. The predicted UCIP performance is based on the summed cumulative hazard function for responding to each sensor image alone (see Houpt, et al., 2012 for details). The equation for the capacity coefficient for OR systems is given by:

$$C_{OR}(t) = \frac{H_{Fused}(t)}{H_{LWIR}(t)+H_{Visible}(t)} \quad (1)$$

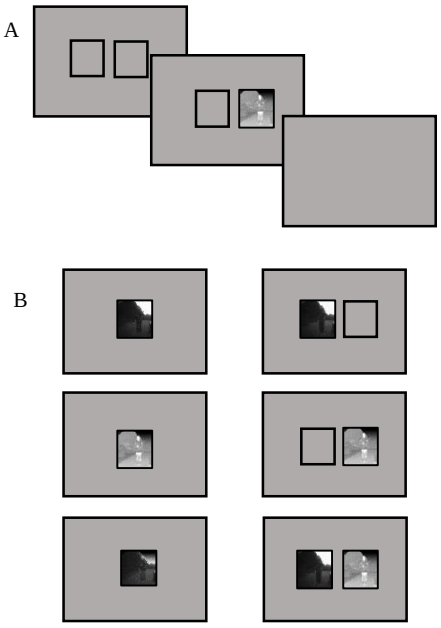

Figure 1. A: The procedure of a trial (example was given for side-by-side presentation). B: three center blocks (left) and three side-by-side blocks (right) used in the experiment. From top to bottom: visible blocks, LWIR blocks and fused blocks.

If the capacity coefficient is larger than one, the system is super-capacity. If the ratio is equal to one, the system is unlimited-capacity. If the ratio is smaller than one, the system is limited-capacity. In this paper, we use the capacity coefficient to assess the effect of presenting a single sensor image compared to two sensor images on weapon (i.e. "gun") and non-weapon (i.e. "tool") discrimination performance.

The SFT measures of architecture and stopping rule are the Mean Interaction Contrast (MIC) and more general Survivor Interaction Contrast (SIC). The interaction contrast is between speeds manipulations factorially applied to each source in information processing. In current study, we factorially added Gaussian luminance noise to each sensor image.

$$SIC(t) = [S_{SS}(t)-S_{SF}(t)] - [S_{FS}(t) - S_{FF}(t)] \quad (2)$$

Each term within the brackets should be positive (slower processing implies higher mean response times and higher survivor functions across time). However each combination of architecture and stopping-rule imply a different SIC shape: A parallel model with an OR stopping rule has an entirely positive SIC; a parallel model with an AND stopping rule has an entirely negative SIC; a serial process with an OR stopping rule has an SIC always equal to zero; a serial process with AND stopping rule is first negative, then positive, with equal positive and negative areas-under-the-curve; a coactive process also has a negative-then-positive SIC but with more positive area.

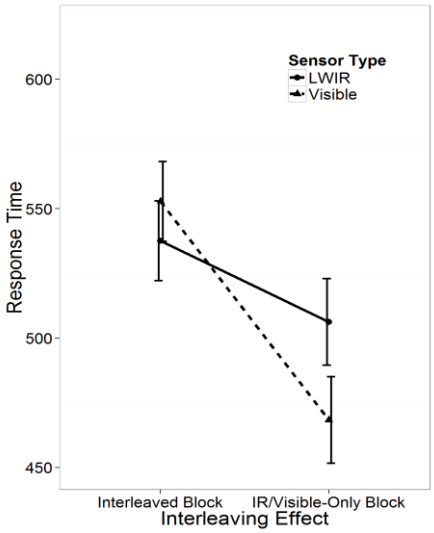

Figure 2. Response times for interleaving effect and sensor type. The confidence intervals in repeated measures designs were calculated according to methods by Jarmarsz and Hollands (2009).

Experiment 1

The goal of Experiment 1 was to measure the capacity coefficient with both side-by-side imagery and algorithmically fused imagery.

Methods

Observers. Ten observers whom gave informed consent participated in the study (6 male; average age=23.8). All had normal or corrected-to-normal visual acuity, and normal color vision. After finishing the study, each participant received 10 dollars as compensation.

Stimuli. Stimuli were ten images taken of a female holding either a gun or a tool using both a long-wave infrared (LWIR) sensor and a standard visible-spectrum sensitive camera (see Pinkus, Toet, and Task, 2009, for details of the image collection). Both visible and LWIR images were mapped to grey scale for presentation to the observers. Fused images were created using simultaneously captured LWIR and visible images that were fused image using a Laplacian pyramid algorithm (see Yang, Jing, & Zhao, 2010 for a review). All images were 256 × 256 pixels in size. Stimuli were presented in the center of a 19" monitor with resolution of 1280 ×1024 pixels and a refresh rate of 85 Hz.

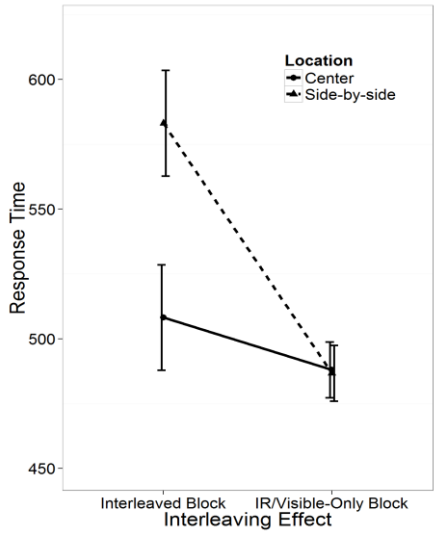

Figure 3. Response times for interleaving effect and location. The confidence intervals in repeated measures designs were calculated according to methods by Jarmarsz and Hollands (2009).

Procedure. Figure 1 gives an overview of the trial procedure. The observer's task was to click mouse to indicate whether the person in the image was holding a gun or tool. Each trial began with a square indicating where images might appear that lasted a uniformly random duration between 400ms and 500ms. Next, the stimuli appeared for 350ms followed by a blank screen while waiting for the subjects' response. There were six different blocks totaling 1200 trials. The whole session lasted approximately one hour.

Results

Accuracy and correct response times (RT) were analyzed with a repeated measure ANOVA using the ez package (Lawrence, 2015) in R (R Core Team, 2015). One ANOVA was run to assess the difference between a single sensor image in isolation and two sources presented together (averaged across both centered and side-by-side presentation blocks). A second ANOVA was applied the subset of the data for which both sensor were presented to assess the difference between fused and side-by-side images. The third ANOVA checked all the other factors that would lead to differences for non-fused trials.

The difference between single-source images (M = 515.83, SD = 176.33; M = 0.93, SD = 0.26) and two-source images (M = 530.92, SD = 183.89; M = 0.91, SD = 0.29) was not significant for either RT (F (1, 9) = 3.93, p = 0.08) or accuracy (F (1, 9) = 1.48, p = 0.26).

The difference between performance with cognitive fusion (M = 542.19, SD = 179.78; M = 0.94, SD = 0.24) and algorithmic fusion (M = 518.87, SD = 187.52; M = 0.88, SD = 0.33) was not significant for either RT (F (1, 9) = 1.21, p = 0.3) or accuracy (F (1, 9) = 2.92, p =0.12).

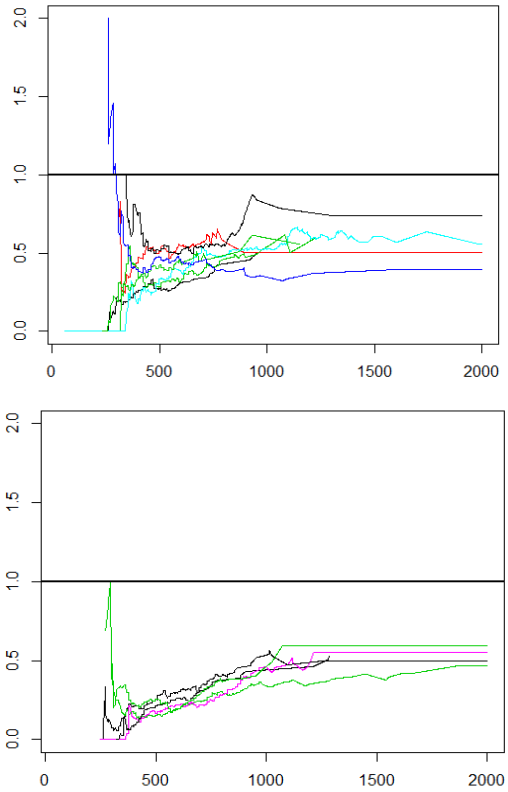

Figure 4. Plots for capacity coefficient analysis for algorithmic fusion (top) and cognitive fusion (bottom).

An ANOVA tested sensor type of images, location presented, and interleaving effect in non-fused trials. For accuracy, only a main effect of interleaving effect was significant ($F(1,9) = 19.6, p < 0.01, \eta_G^2 = 0.06$). When images were presented in the center of the screen, subjects responded faster (M = 497.98, SD = 169.40) than when images were presented side-by-side (M = 534.05, SD = 181.33; F (1, 9) = 19.68, p < 0.01, η_G^2= 0.10). Performance was better in the no interleaving (only LWIR or visible images) blocks (M = 487.39, SD = 159.50) than interleaved blocks (M = 545.20, SD = 187.71; F (1, 9) = 40.68, p < 0.01, η_G^2 = 0.10). The main effect of sensor type, LWIR (M = 509.90, SD =180.47) or visible (M =521.71, SD =171.93), was not significant (F (1, 9) = 1.63, p = 0.23). But there was a significant interaction with interleaving (Figure 2, F (1, 9) = 24.53, p < 0.01, η_G^2 = 0.02). And interleaving effect also interacted with the location presented (Figure 3; F (1, 9) = 42.87, p < 0.01, η_G^2 = 0.04).

Figures 2 and 3 indicate that when LWIR and visible images were interleaved in presentation, subjects couldn't predict the next image's type, thus the advantage of visible image (M = 468.40, SD = 152.54) over LWIR image (M = 506.28, SD = 164.00) disappeared. Also, in the interleaved block, subjects needed to switch attention between the left side and right side, so people responded slower (M = 583.12, SD = 194.28) compared to that when they only need to focus on a fixed position of the screen (M =508.18, SD = 173.28; F (1,9) = 35.83, p < 0.01, η_G^2 = 0.80).

Capacity analyses were only applied to observers who had at least 90% accuracy in all conditions within the algorithmic blocks (7/10 participants) or cognitive-fusion blocks (5/10 participants). Figure 4 depicts those participant's capacity

Figure 5. Example of visible image without noise (left) as fast processing task and with noise (right) as slow processing task.

coefficients. Individual level capacity analysis indicated 7/7 participants were limited capacity with cognitive fusion (z-scores between -11.2 and -10.3) and 5/5 were limited capacity with algorithmic (z between -10.98 and -6.13). Group level t-tests indicated both fusion methods led to limited capacity (Algorithmic: t (6) = -12.64; Cognitive: t (4) = -73.36). Cognitive fusion was statistically more limited compared to algorithmic fusion (t (6.62) = 4.02, p <0.01, d = 2.16).

Discussion

Both algorithmic fusion and cognitive fusion led to limited capacity, consistent with previous work (Fox, 2015). However, unlike Fox (2015), there was a difference in the degree of capacity limitation between the two methods. The process of switching attention between locations increased reaction time and reduced the advantage of visible images over IR images in our task. The side-by-side presentation method increased reaction times but resulted in equal performance compared to the images developed by algorithmic fusion in the center of screen.

There are various potential explanations for the limited capacity performance. For the side-by-side imagery, observers may have only used one source and hence lost out on the redundancy gain predicted by independent parallel processing (cf. Raab, 1962). This explanation may also apply to performance with fused images: Because the information in both sensor images is essentially the same, the fused image is not necessarily more informative than either individual source image.

If observers were using a strategy of focusing on only one sensor image, then they should have a flat SIC and 0 MIC when examined with the factorial salience conditions. Alternatively, observers could be using both sensor images but, due to limited resources (e.g., attention, foveation), be processing each source slower when they are together. In this case, observers would have a positive SIC and MIC. To discriminate between these two possible explanations of performance with the side-by-side imagery, we ran a follow-up experiment with the factorial manipulations necessary to calculate the MIC and SIC. Because the manipulation of each sensor image is no longer selective after algorithmic fusion, we were not able to apply the SIC and MIC analysis to the algorithmically fused imagery.

EXPERIMENT 2

The goal of Experiment 2 was to measure the SIC and MIC from participants when they were making discrimination judgments with side-by-side imagery.

Methods

Observer. Ten new observers who gave informed consent participated in the study (male = 3, average age=23.7). All had normal or corrected-to-normal visual acuity, and normal color vision. After finishing the study, each participant received 40 dollars as compensation.

Stimuli. The base images were the same as images in Experiment 1 (although the algorithmically fused images were not used). For slow processing trials, the base images were displayed with zero mean Gaussian luminance noise added. The variance of the noise was chosen individually as described in the next subsection.

Procedure. Each trial was the same as Experiment 1. For each trial, images appeared on left side, right side or both. There were two blocks in each session; the first block used the **Ψ**-method (Kontsevich & Tyler, 1999) to find the noise level corresponding to 90% accuracy for each image. In the second block, each source image could be presented with (slow) or without (fast) the added noise (see example in Figure 5). All combinations of fast, slow, and absent on each source were used except there were no trials in which both were absent. There were four by 1 hour sessions, each session consisting of 1320 trials.

Results

Allowing for the possibility that participants processed gun images different than tool images, we analyzed the SICs separately. Two subjects' data were excluded from the tool analysis because their data indicated violations of selective influence (Formula 2). A summary of the SIC shapes indicated by the Houpt-Townsend (2010) statistic is given in Table 1.

Table 1
SIC Analysis for Experiment 2

Subject	Tool			Weapon		
	SIC	MIC	Strategy	SIC	MIC	Strategy
1	+	+	P-O	+	+	P-O
2	NA	NA	NA	+	+	P-O
3	=	=	S-O	-	=	P-A
4	NA	NA	NA	+	=	P-O
5	+	+	P-O	+	+	P-O
6	+	+	P-O	=	=	S-O
7	+	+	P-O	-	=	P-A
8	+	+	P-O	=	+	S-O
9	+	+	P-O	=	=	S-O
10	=	+	S-O	-	-	P-A

Note. For strategy, P-A means parallel-AND, P-O means parallel-OR, and S-O means serial-OR.

Discussion

The SIC analysis indicated two observers employed the same parallel-OR processing strategy for identifying both guns

and tools. Among the six remaining subjects for whom both tool and gun SIC data were interpretable: three subjects used parallel-OR strategy for identifying tool and serial-OR for identifying gun, two subject used serial-OR for identifying tool and parallel-AND for identifying gun, one subject used parallel-OR strategy for identifying tool and parallel-AND for identifying gun. Based on the simulation studies reported in Houpt and Townsend (2010) and the estimated SICs, it is unlikely that those four observers used the same strategy for both tool and gun.

GENERAL DISCUSSION

Our first experiment indicated that the information from additional source does not gain advantage in information processing as expected. Our second experiment tried to explore the reason for limited capacity in cognitive fusion in Experiment 1. It suggested searching strategy might lead to inefficiently processing when images from different sensors were presented side-by-side. When identifying gun, the observers may be limited capacity because they are waiting to process both images rather than responding as soon as they have identified gun in either image. When identifying tool, the observer may be limited capacity due to serial-OR processing (i.e., only using one image source) or they may have been parallel-OR (and the SIC was not significant due to lack of power) and the limitation may be due to limited perceptual resources.

Both algorithmic and cognitive fusion result in limited capacity, but cognitive fusion led to more limited capacity in current study. Given that Fox (2015) found no difference in capacity using a discrimination task with different stimuli, the performance differences between cognitive and algorithmic are likely task dependent. Based on our research thus far, and that it would be infeasible to catalog all possible stimuli that an operator might need to discriminate among, we suggest that operators should have a choice between algorithmic and cognitive fusion displays.

ACKNOWLEDGMENTS

This work was supported by a grant from the Air Force Office of Scientific Research, FA9550-13-1-0087.

REFERENCES

Blasch, E. (2000). Assembling a distributed fused information-based human-computer cognitive decision making tool. *Aerospace and Electronic Systems Magazine*, *15*(5), 11-17.

Deshmukh, M., & Bhosale, U. (2010). Image fusion and image quality assessment of fused images. *International Journal of Image Processing (IJIP)*, *4*, 484-508.

Dixon, T. D., Li, J., Noyes, J. M., Troscianko, T., Nikolov, S. G., Lewis, J. J., ... & Canagarajah, C. N. (2007). Scanpath assessment of visible and infrared side-by-side and fused video displays. *Proceedings of the 10th International Conference on Information Fusion, Canada,* 1-8

Fox, E.L. (2015). *Cognitive analysis of multi-sensor information* (Unpublished Master's thesis). Wright State University, Dayton, OH.

Houpt, J. W., Blaha, L. M., McIntire, J. P., Havig, P. R., & Townsend, J. T. (2014). Systems factorial technology with R. *Behavior research methods*, *46*, 307-330.

Houpt, J.W. and Townsend, J.T. (2010). The statistical properties of the survivor interaction contrast.Journal of Mathematical Psychology, 54, 446-453

Jarmasz, J., & Hollands, J. G. (2009). Confidence intervals in repeated-measures designs: The number of observations principle. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *63*, 124-138.

Kontsevich, L. L., & Tyler, C. W. (1999). Bayesian adaptive estimation of psychometric slope and threshold. *Vision Research*, *39*, 2729-2737.

McCarley, J. S., & Krebs, W. K. (2000). Visibility of road hazards in thermal, visible, and sensor-fused night-time imagery. *Applied ergonomics*, *31*, 523-530.

Michael A. Lawrence (2015). ez: Easy Analysis and Visualization of Factorial Experiments. R package version 4.3. https://CRAN.R-project.org/package=ez

Muller, A. C., & Narayanan, S. (2009). Cognitively-engineered multisensor image fusion for military applications. *Information Fusion*, *10*, 137-149.

Nogami, M., Nakamoto, Y., Sakamoto, S., Fukushima, K., Okada, T., Saga, T., ... & Sugimura, K. (2007). Diagnostic performance of CT, PET, side-by-side, and fused image interpretations for restaging of non-Hodgkin lymphoma. *Annals of nuclear medicine*, *21*, 189-196.

Pinkus, A. R., Toet, A., &Task, H. L. (2009). Object recognition methodology for the assessment of multi-spectral fusion algorithms: Phase 1. In *SPIE Defense, Security, and Sensing* (pp. 73360X-73360X). International Society for Optics and Photonics.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Raab, D. H. (1962). division of psychology: statistical facilitation of simple reaction times*. *Transactions of the New York Academy of Sciences*, *24*(5 Series II), 574-590.

Townsend, J. T., & Nozawa, G. (1995). Spatio-temporal properties of elementary perception: An investigation of parallel, serial, and coactive theories. *Journal of Mathematical Psychology*, *39*, 321-359.

Xue, Z. Blum, R.S. (2003), Concealed weapon detection using color image fusion. *Proceedings of the 6th International Conference on Image Fusion, Australia, 1,* 622-627

Xue, Z., Blum, R. S., & Li, Y. (2002). Fusion of visual and IR images for concealed weapon detection. *Proceedings of the Fifth International Conference on Information Fusion, USA*, *2*, 1198-1205

Yang, B., Jing, Z. L., & Zhao, H. T. (2010). Review of pixel-level image fusion. *Journal of Shanghai Jiaotong University (Science)*, *15*, 6-12.