Object recognition methodology for the assessment of multi-spectral fusion algorithms: Phase 1

Alan R. Pinkus*[a], Alexander Toet[b], H. Lee Task[c]
[a]Air Force Research Laboratory, 2255 H St., Wright-Patterson AFB, OH 45433
[b]TNO Defence, Security and Safety, PO Box 23, NL-3769 ZG, Soesterberg, The Netherlands
[c]Task Consulting, 817 S. Bill Martin Dr., Tucson, AZ 85745

ABSTRACT

In this effort we acquired and registered a multi-spectral dynamic image test set with the intent of using the imagery to assess the operational effectiveness of static and dynamic image fusion techniques for a range of relevant military tasks. This paper describes the image acquisition methodology, the planned human visual performance task approach, the lessons learned during image acquisition and the plans for a future, improved image set, resolution assessment methodology and human visual performance task.

Keywords: multi-spectral, resolution, fusion, visual performance, triangle orientation detection, image enhancement

1. INTRODUCTION

The fusion of multi-spectral images and video sources is emerging as a vital technique for surveillance purposes, object detection, vehicle navigation, and object tracking applications[1]. Purportedly, the main goal of image fusion is to provide a single compact representation of the input images that is more informative than each of the individual inputs. There are several potential benefits of multi-sensor, multi-spectral image fusion: wider spectral, spatial and temporal coverage, extended range of operation, decreased uncertainty, improved reliability, and increased robustness of the system's performance.

In many applications, the human visual perception of the fused image is of fundamental importance. As a result, image fusion results are mostly evaluated by subjective human visual inspection, i.e., "Does the fused image look better." Subjective A-B evaluation tests are often substituted for the more complex, time-consuming, and expensive objective test methodologies[2]. Moreover, subjective methods often do not correlate well with how people actually perform in visual tasks utilizing fused images. This has led to an increasing demand for more efficient objective tests that allow rapid comparison of the results obtained with different fusion, registration, and enhancement algorithms, the automatic selection of the appropriate fusion algorithm for a given scenario, and/or to obtain the optimal settings for a specific fusion algorithm.

Effective image fusion systems should provide a more complete representation (with increased visually useful information content) of the scene, which is easier to interpret and understand (ergonomic value). A range of different image fusion algorithms is currently available, many of which can be implemented in real-time. In practice, many image fusion algorithms merely produce fused images with an increased amount of detail (compared to the original input images) without taking into account the information content (the meaning) of the resulting combined details. As a result, the perceptibility of relevant features in the fused representation of the scene may be degraded, and visual task performance may be adversely affected. For instance, when clutter is included in the fusion process a large number of spurious (i.e., non-informative or task-irrelevant) details may appear in the resulting fused image. As a result, human or machine performance of object recognition and object classification may be severely degraded.

*alan.pinkus@wpafb.af.mil; phone 1-937-255-8767; fax 1-937-255-8366

Signal Processing, Sensor Fusion, and Target Recognition XVIII, edited by Ivan Kadar,
Proc. of SPIE Vol. 7336, 73360X · © 2009 SPIE · CCC code: 0277-786X/09/$18 · doi: 10.1117/12.820006

Image quality is task related. A fused image can be said to be of good quality if it allows the observer to achieve a task performance that is similar to or better than the performance that can be achieved with the original, individual images. However, the quantification of this performance assessment is often difficult and time consuming. Hence, there is a need for efficient and reliable methods to quantify the operational effectiveness of image fusion systems.

The multi-spectral night vision imagery collected in this effort was intended to be used to (1) evaluate existing image fusion schemes, (2) to design and optimize new dynamic image fusion schemes, and (3) to develop new image fusion quality metrics. In the following sections, we will provide a detailed account of the equipment, the scenario, the objects, the registration location, and the environmental conditions under which this imagery was collected. We will also describe our approach to human visual performance testing, the lessons learned from this effort, and our plans for a future effort in this area.

2. METHOD

Human Visual Performance Assessment Investigation - It is necessary to first select a human visual performance task that will be conducted with the multi-spectral imagery in order to establish the parameters under which the imagery is collected. There are several possible methods that could be employed to conduct a human visual performance study. The originally selected method for this effort has been used successfully in the past to study human visual target recognition capability[3,4,5,6]. The concept is to have an image that dynamically increases in size over time until the observer can correctly identify the object from the set of objects selected. The intent is to simulate a vehicle-mounted sensor closing on a target area with the operator having to decide if the object he/she is viewing is a legitimate target of interest. As the image of the object increases in size, the amount of object detail available to the observer increases due to the added number of pixels across the object. In general, the better the image quality, the smaller the object image size required for recognition (meaning the object is farther away). Therefore, the dependent variable for this method is the angular subtense of the object at the point of recognition/decision, which directly relates to target distance (slant range). Note that this is only a useful dependent variable if the error rate is relatively constant across whatever parameters are under investigation (e.g., different fusion algorithms or different spectral bands). In order to improve the probability of a relatively constant error rate across subjects and conditions, past studies revealed[3,4] that instructing the subjects to make their response "as soon as possible" but only if they are "virtually certain" that they know what the object is and produces about a 95% correct response rate (5% error rate).

In order to investigate the efficacy of different fusion algorithms,[1] the approach was to record the raw imagery from each of three sensors at 25 frames per second. These sequences of images could then be played back in an AVI file, producing a dynamic image for subjects to observe. This approach allows one to apply fusion algorithms or enhancement algorithms to the sequence of images in non-real time (since some enhancement and fusion algorithms are too computationally intensive for real time processing). Also, having a set of baseline test runs that can be processed in non-real time allows investigations of the effectiveness of additional fusion and enhancement algorithms without having to generate more stimulus material.

2.1 Equipment

The sensor suite used to capture the multi-spectral images as shown in Figure 1 include: (A) Lion Advance 8-12 μm long-wave infrared (LWIR) camera (Thales Optronics), field of view 7.8°x5.9°, focal point distance 81.71 mm, detector pitch 35 μm, and NETD<80mK, (B) digital image intensifier for near infrared (NIR), field of view 8.1°x6.1°, and (C) Raytheon Radiance High Speed, 256x256 pixels, InSb focal-plane array mid-wave infrared (MWIR) camera, field of view 8.8°x8.8°, focal point distance 50 mm, detector pitch 30 μm, and NETD<25mK.

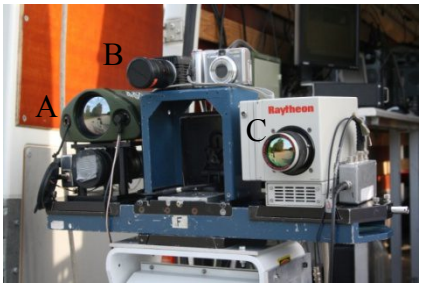

Figure 1. The tripod-mounted sensor suite used to capture multi-spectral images.

For field data collection purposes, a research van was used to house and transport the sensors, computers, monitors and power generation equipment for monitoring and recording multi-spectral imagery. The analog signals from all cameras were digitized at a frame rate of 25 Hz, using a Solios Matrox frame grabber, running under MIL Lite. The Lion and image intensified charge-coupled device images were acquired frame by frame. The resulting difference in time was approximately 1-2 frames. The MWIR camera images were acquired using another program. Thus, the images were not synchronized. In practice, this means that for the image sequences representing the same run (same experimental condition), images with corresponding numbers represent approximately the same moment in time (there can maximally be a difference of 2 between the frame numbers of images representing the same moment in time).

GPS signals were continuously (Figure 2) registered both at the location of the sensor suite and at the location of the objects. During the experiment the soldiers carried a backpack with a laptop (Dell Inspirion) that was attached to a BU-353 USB SiRF Star III GPS receiver. An identical combination of GPS receiver and laptop was placed next to the sensor suite. The difference between these two GPS signals at a given time corresponds to the distance between the object and the sensor suite at that time.

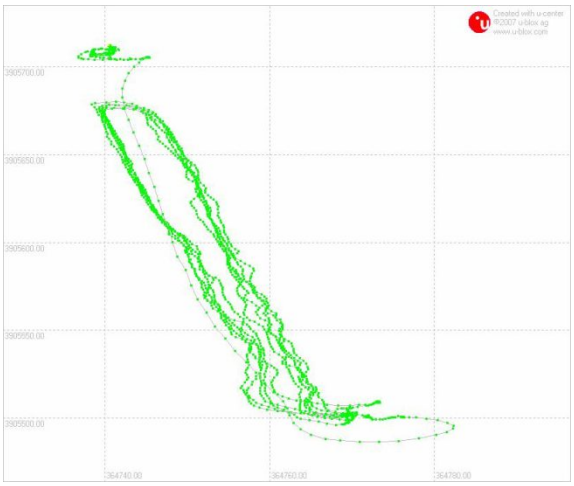
Figure 2. Multiple GPS tracks of soldiers approaching the cameras (lower position) from afar (upper position) and then returning to their starting position.

2.2 Stimuli

A total of eight objects were selected as stimuli (Figure 4). These were held (one at a time) and carried by a soldier who walked toward the sensor array at a relatively constant gait. The intent was for the soldier to start at a distance such that the objects were not identifiable (approximately 300m) and then approach to a distance at which the objects were then easily identifiable (see sample start and end pictures in Figure 3).

Figure 3. LWIR M-16 with grenade launcher at start of sequence (left) and end of sequence (right).
The eight objects selected were: Glock pistol, M-16 rifle, M-16 rifle with grenade launcher attached, mini SAW (squad automatic weapon), heavy machine gun, hammer, wooden stake, and an axe (see Figure 4).

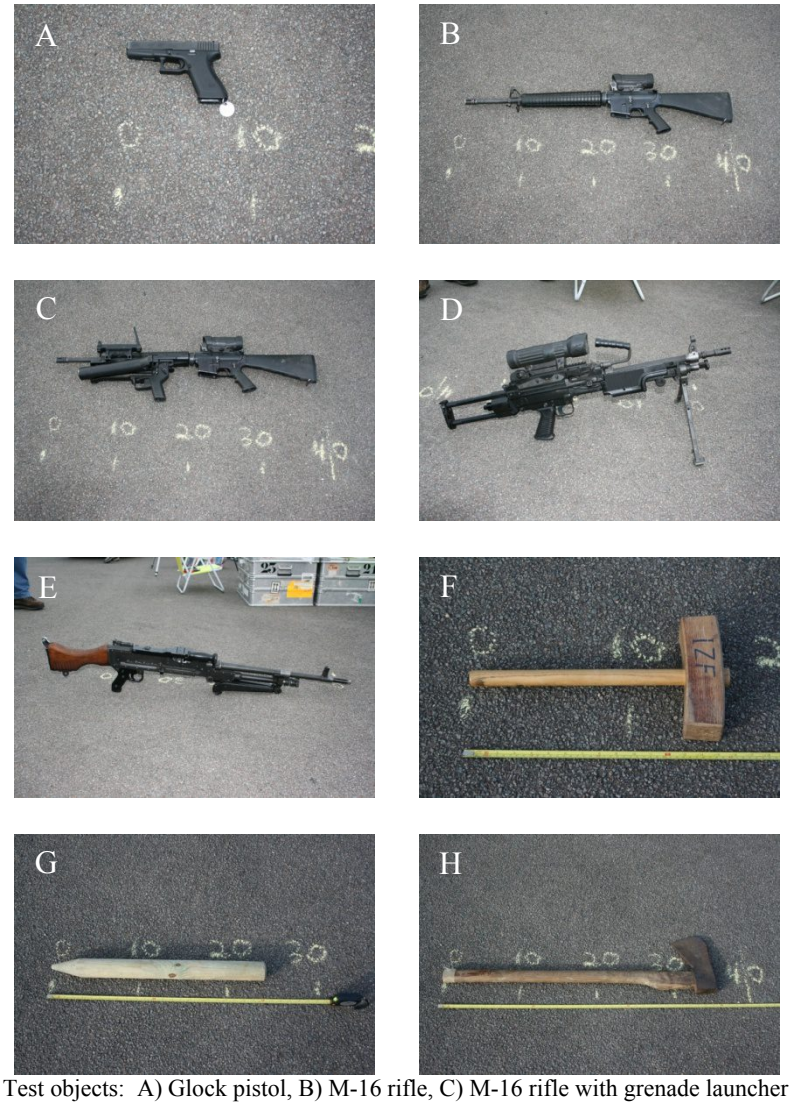

Figure 4. Test objects: A) Glock pistol, B) M-16 rifle, C) M-16 rifle with grenade launcher attached, D) mini SAW (squad automatic weapon), E) heavy machine gun, F) hammer, G) wooden stake, and H) axe.

Figure 5 shows the pistol held high for each of the three sensor bands (the pictures have been cropped to show primarily the soldier with the pistol at the closest distance).

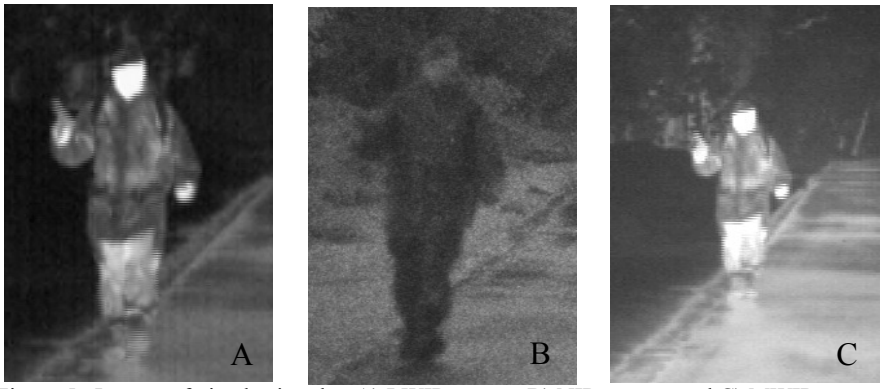

Figure 5. Images of pistol using the: A) LWIR sensor, B) NIR sensor, and C) MWIR sensor.

Figure 6 shows the M-16 rifle held by the soldier for each of the three sensor bands (the pictures have been cropped to show primarily the soldier with the rifle).

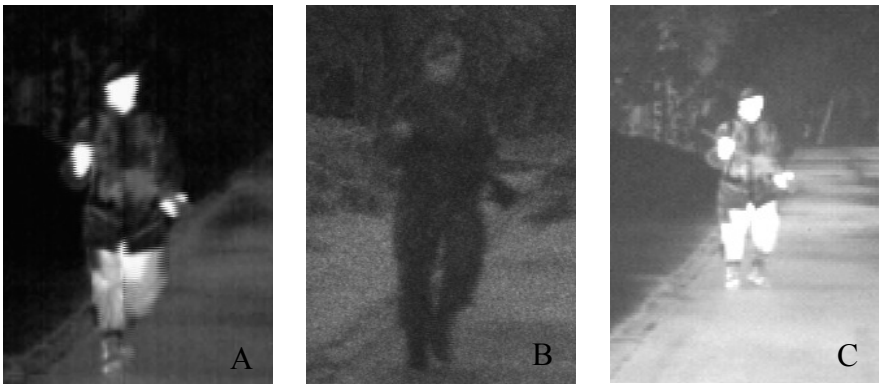

Figure 6. Images of M-16 using the: A) LWIR sensor, B) NIR sensor, and C) MWIR sensor.

As seen in Figures 5 and 6, it is somewhat difficult to determine the object held by the soldier even when comparing these images to the set of objects shown in Figure 4.

2.3 Environmental conditions

The image collection was performed during the night of Tuesday, 9 September, 2008 in Amersfoort, NLD. The weather was quite variable during the image acquisition period.

At the start of the image collection period, it was dry and the visibility was good. After completing five consecutive runs, it started to drizzle. The next 11 runs were performed during the rain. At the end of the experiment it was dry again. Table 1 lists the meteorological data at the time and location of the field trial.

Table 1. Weather data of Tuesday 9 September 2008 at de Bilt, near the registration site.
Source: The Royal Dutch Meteorological Institute.

Weather data of tuesday 9 September 2008 at De Bilt					
Temperature		**Average**	**Precipitation**		
Mean	17.3 °C	15.2 °C	24h sum	2.9 mm	
Maximum	23.3 °C	20.0 °C	Duration	1.5 hours	
Minimum	10.0 °C	10.4 °C			
Sun, cloud cover & visibility			**Wind**		
Duration sunshine	10.1 hours		Mean	3.0 m/s	= 2 Bft
Relative sunshine duration	77 %	37 %	Maximum hourly mean	5.0 m/s	= 3 Bft
Average cloud cover	4 octa's partly cloudy		Maximum gust	11.0 m/s	
Minimum visibility	0.4 km		Prevailing direction	151 °	= SSE
Relative atmospheric humidity			**Air pressure**		
Mean	79 %	82 %	Mean air pressure	1014.1 hPa	

Figure 7 is a daytime picture of the area where the images were collected. The soldier used the edge of the blacktop as a guide in walking a straight line from the starting point toward the sensor suite.

Figure 7. Daytime image of the soldier's walking path.

3. RESULTS

3.1 Lessons Learned

The image sequences from all three sensors and the nine test runs (eight objects plus a no object run) were processed and converted to AVI files for dynamic viewing. Also, several fusion algorithms were applied to the images. After carefully reviewing the nine test run sequences and the results of the image enhancement, registration, and fusion algorithms, it was decided that the imagery was marginal for the original stated goal of conducting human visual performance studies. This was a difficult decision but it was determined that there were enough issues regarding the imagery and the procedures that conducting a human performance study with the imagery would probably not be productive. However, obstacles encountered in this effort have resulted in a number of valuable lessons learned and guidance about ways to improve field data collection procedures that will be applied to a future effort. The remainder of this paper addresses the specific lessons learned and improved procedures that will be implemented for the next iteration of this project.

3.1.1 Methodology

Using a *dynamic zoom* method for investigating image enhancement, registration, and fusion algorithms has the unique quality of presenting the subject with continuously increasing object image sizes. This approach is computationally intensive but could be executed by multiple computers, processing individual frames that make-up the motion sequences and then playback the new imagery at real-time frame rates. An alternative approach is to select a set of fixed distances (sensor to object) to limit the number of images that must be processed. The images could still be presented in a semi-dynamic fashion by starting with the smallest object image first, presenting it for a fixed amount of time (e.g., 3 to 5 seconds), then moving to the next larger image for the same time interval. By selecting the appropriate object sizes and time intervals one could simulate a zoom sequence that was continuous but would move in jumps. This method would, in effect, be a cross between the dynamic zoom method and a traditional tachistoscopic presentation. The advantages are that it would still have the same dependent variable (angular subtense of the object at recognition, which relates to object size and range) and would require considerably less image processing. Another advantage is that the object could be set at known, fixed distances and not depend on the soldier trying to maintain a consistent gait from sequence to sequence. The disadvantages are that it would no longer give the appearance of a smooth object approach and the effects of noise (if the sensor/ambient lighting produces significantly noise images) and dynamic blur would not be properly captured. A third approach would be to use multiple fixed positions but shown in a random order[7]. Employing Probit analysis[8] the percentage of correct object identifications (adjusted for chance) could be used to construct a probably of seeing S-curve (e.g., percent correct as a function of distance per fusion algorithm, per object type).

3.1.2 Objects

Eight objects were selected for image acquisition for this effort. A somewhat smaller set of test objects that have more equal discriminability would probably produce more uniform results. Some of the objects selected required the soldier to hold the object in both hands (e.g., the rifle and axe) while others required only a single hand to hold the object (e.g., pistol and hammer). Subjects would most likely be able to separate out the two types of objects earlier, which may result in some subjects developing strategies to correctly identify the object based on other artifacts in the sequence run. This is also why it is advantageous to have multiple runs for the same object to prevent the subjects from learning the specific stimulus material as opposed to responding to the perceived detail of the object.

One must also make sure that the endpoints of the zoom sequence are such that at the farthest point no subject can recognize any of the objects and at the nearest point all subjects can recognize all of the objects (see Figures 5 and 6). This may be easier to double-check in the field if one is using still images and can view the results on site before completion of the image collection.

In addition to collecting a set of images of the object at different distances, it is also possible to include in the set of still images a resolution type of object to serve as a standard object for assessing the quality of the sensor. For example, a Landolt C, an equilateral triangle,[9,10,11] or combination, of a specific size could be included just off to the side of the soldier holding the object in the revised methodology. These resolution objects could be used to determine the *effective resolution* of each of the sensors under the environmental conditions in effect during the image collection through either a separate psychophysical study or perhaps through a software recognition algorithm (currently in development). One approach would be to collect at least four images of each object at each distance with the resolution object in the image oriented in a different direction (up, down, left, right) in each of the four images. This would aid in correlating standard sensor resolution assessment methods and subject performance with respect to object recognition and with image fusion algorithms. Note that the resolution object must result in a good contrast image for all spectral band sensors included in the sensor array, which means it needs to be thermally active.

3.1.3 Sensors

In order to maximize the potential success of enhancement, registration, and fusion algorithms, all sensors should be aligned axially and rotationally. Using the approach of collecting images at fixed distances, all sensors should be capable of being activated for image captured at the same time.

3.1.4 Environmental conditions

As noted in the introductory section of this paper, the potential advantage of image fusion algorithms is to successfully combine the best parts of the images from each of the sensors. If the environmental conditions under which stimulus material is collected are poor for all sensors, then one cannot really expect image fusion to have any chance of success. Thermally neutral times of night (e.g., predawn or during wet conditions) with no thermally emissive objects in the image result in thermal images with poor information content. Similarly, near infrared images taken on an overcast night will most likely produce dark images with low contrast. Therefore, it is necessary to collect images on multiple days and/or nights at different times and under different weather conditions. This would result in a range of conditions that may benefit each type of spectral sensor, which should better mimic the real world and therefore overall expected sensor performance and image fusion success.

One specific problem that occurred with the previously described dynamic image collection was the occasional occurrence of unplanned wildlife visitations. On two of the dynamic runs, wildlife entered the image from the left, out of the woods. In one case it appears that a pair of rabbits came close to the soldier's pathway, then scurried back out of the picture as the soldier approached (see Figure 8). In another run, a frog hopped across the pathway. These occurred without the knowledge of those collecting the images but it makes the imagery unusable for later psychophysical studies as these events allow subjects to learn the imagery (because of the artifacts) rather than respond to the details of the object.

Figure 8. Wildlife (white spots, lower left; probably rabbits) ran in and out of the scene.

CONCLUSIONS

A second field data collection will be undertaken to collect multi-spectral imagery that can be used for testing image enhancement, registration, and fusion algorithms. These lessons learned will serve as guidance for the collection of new imagery and the ensuing psychophysical studies including the assessment of sensor resolution using the Landolt C and equilateral triangle resolution objects[9,10,11]. A separate effort, currently under way, will attempt to ascertain sensor resolution through the use of the collected imagery and a software algorithm (currently under development at the US Air Force Research Laboratory) designed to detect the orientation of the resolution objects. The ultimate goal is to combine the results of the image fusion psychophysical studies with the software-based sensor/fusion resolution assessment methodology to be able to predict the probable improvement (if any) of various current and future image fusion algorithms for different spectral bands.

ACKNOWLEDGMENTS

This research was funded, in part, by the European Office of Aerospace Research & Development (EOARD)
US Air Force, under Air Force Contract FA8655-06-1-3017.

REFERENCES

[1] Neriani, K. E., Pinkus, A. R., and Dommett, D. W., "An investigation of image fusion algorithms using a visual performance-based image evaluation methodology," Wright-Patterson AFB, OH: Air Force Research Laboratory (in-press).

[2] Neriani, K. E., Herbranson, T. J., Pinkus, A. R., Task, C. M. and Task, H. L., "Visual Performance-based Image Enhancement Assessment Methodology," Proc. SPIE, 5802, 92-101 (2005).

[3] Task, H. L., "An evaluation and comparison of several measures of image quality of television displays," (Report No. AMRL-TR-79-7). Wright-Patterson AFB, OH: Aerospace Medical Research Laboratory, (DTIC No. A069690) (1979).

[4] Pinkus, A. R., "The effects of color and contrast on target recognition performance using monochromatic television displays," (Report No. AFAMRL-TR-82-9). Wright-Patterson AFB, OH: Air Force Aerospace Medical Research Laboratory, (1982).

[5] Task, H. L., Pinkus, A. R., and Hornseth J. P., "A comparison of several television display image quality measures," Proc. SID, 19(3), 113-119 (1978).

[6] Task, H. L. and Pinkus, A. R., "Contrast sensitivity and target recognition performance: A lack of correlation," SID Int. Symp. Digest Tech. Papers 18, 127-129 (1987).

[7] Pinkus, A. R. and Task, H. L., " Measuring observers' visual acuity through night vision goggles," SAFE Symposium Proceedings 36th Annual Symposium, 1-11 (1998).

[8] Finney, D. J., [Probit Analysis], Third Edition, Cambridge University Press, Cambridge, (1980).

[9] Bijl, P. and Hogervorst, M.A., "A test method for multi-band imaging sensors," Proc. SPIE 5076, 208-219 (2003).

[10] Hogervorst, M.A., et al., "Capturing the sampling effects: a TOD sensor performance model," Proc. SPIE 4372, 62-73 (2001).

[11] Bijl, P. and Valeton, J.M., "Guidelines for accurate TOD measurement," Proc. SPIE 3701, 14-25 (1999).