University of Memphis
University of Memphis Digital Commons

Electronic Theses and Dissertations

11-18-2010

Visual Task Performance Assessment using Complementary and Redundant Information within Fused Imagery

Christopher Leonard Howell

Follow this and additional works at: https://digitalcommons.memphis.edu/etd

Recommended Citation

Howell, Christopher Leonard, "Visual Task Performance Assessment using Complementary and Redundant Information within Fused Imagery" (2010). *Electronic Theses and Dissertations*. 92.
https://digitalcommons.memphis.edu/etd/92

This Dissertation is brought to you for free and open access by University of Memphis Digital Commons. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of University of Memphis Digital Commons. For more information, please contact khggerty@memphis.edu.

To the University Council:

The Dissertation Committee for Christopher Leonard Howell certifies that this is the approved version of the following electronic dissertation: " Visual task performance assessment using measures of complementary and redundant information within fused imagery."

_____
Carl Halford, Ph.D.
Major Professor

_____
Eddie Jacobs, D.Sc.

_____
Aaron Robinson, Ph.D.

_____
Keith Krapels, Ph.D.

Accepted for the Graduate Council:

_____
Karen D. Weddle-West, Ph.D.
Vice Provost for Graduate Programs

VISUAL TASK PERFORMANCE ASSESSMENT USING COMPLEMENTARY AND REDUNDANT INFORMATION WITHIN FUSED IMAGERY

by

Christopher Leonard Howell

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

Major: Engineering

The University of Memphis

December 2010

EPIGRAPH

Listen to those whom you seek counsel, and receive instruction, with hope that you may be wiser tomorrow than you are today

Personal interpretation, *The Bible*, King James Version, Proverbs 19:20

DEDICATION

This dissertation is dedicated to my wife, Aneicia Howell, whose love, understanding and encouragement guarded me against doubt and complacency.

ACKNOWLEDGEMENTS

First and foremost I would like to thank Dr. Carl Halford for giving me the opportunity to study under his leadership. Under his guidance I learned to seek out the value in my contributions, whether as part of a research team or some other endeavor life places before me. This lesson I will always remember and try to instill in others.

I would like to personally thank Dr. Ronald Driggers, Dr. Eddie Jacobs, Dr. Steve Moyer, Dr. Keith Krapels and Dr. Aaron Robinson for their invaluable insights and personal discussions offered. I would especially like to thank Dr. Srikant Chari for his insight and guidance offered throughout my graduate education experience. Last, I would like to thank Night Vision and Electronic Sensors Directorate (NVESD) for allowing me to participate in their co-op program. Having the opportunity to converse with so many intellectual minds on a daily basis has been beneficial in strengthening me as an engineer as well as a person.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Christopher Leonard Howell, Ph.D. The University of Memphis. December/2010. Visual task performance assessment using complementary and redundant information within fused imagery. Major Professor: Dr. Carl E. Halford.

Image fusion is a process of combining information from a set of source images to obtain a single image with more relevant information than any individual source image. The intent of image fusion is to produce a single image that renders a better description of the scene than any of the individual source images. Information within source images can be classified as either redundant or complementary. The relevant amounts of complementary and redundant information within the source images provide an effective metric for quantifying the benefits of image fusion. Two common reasons for using image fusion for a particular task are to increase task reliability or to increase capability. It seems natural to associate reliability with redundancy of information between source bands, whereas increased capability is associated with complementary information between source bands. The basic idea is that the more redundant the information between the source images being fused, the less likely an increase in task performance can be realized using the fused imagery. Intuitively, the benefits of image fusion with regards to task performance are maximized when the source images contain large amounts of complementary information. This research introduces a new performance measure based on mutual information which, under the assumption the fused imagery has been properly prepared for human perception, can be used as a predictor of human task performance using the complementary and redundant information in fused imagery.

The ability of human observers to identify targets of interest using fused imagery is evaluated using human perception experiments. In the perception experiments, imagery of the same scenes containing targets of interest, captured in different spectral bands, is fused using various fusion algorithms and shown to human observers for identification. The results of the experiments show a correlation exists between the proposed measure and human visual identification task performance. The perception experiments serve to validate the performance prediction accuracy of the new performance measure. The development of the proposed metric introduces into the image fusion community a new image fusion evaluation measure that has the potential to fill many voids within the image fusion literature.

1 INTRODUCTION

It is no surprise that image fusion has gained a fair amount of research activity over the past few years. The process of combining image information from multiple sensors can increase the information content made available to the end user, giving the end user a competitive edge. It should be noted, the relevance of the information being fused depends highly on the intended visual perception task. An illustration of the potential benefits of image fusion can be seen in Figure 1. In the figure are a short-wave infrared (SWIR), long-wave infrared (LWIR) and an image showing the fusion of the two source images. A quick comparison between the SWIR and LWIR images shows various amounts of complementary and redundant information exists. In both images, ample visual information is available regarding the vehicle in the scene. That is, both the LWIR and the SWIR contain enough information to recognize the presence of the vehicle. In contrast, only one human target can be seen lurking in the tree line for the LWIR, while for the SWIR two human targets can be seen sitting in the vehicle. The fusion of the LWIR and SWIR images results in an image that contains more information than any individual source image with regard to human targets in the scene. This illustration shows the possible benefits image fusion can have on visual task performance. At the same time it shows the benefits of image fusion are often dependent on the intended task. For tasks such as identifying human targets in a scene, it is safe to assume using this imagery, visual task performance would be best using the fused image.

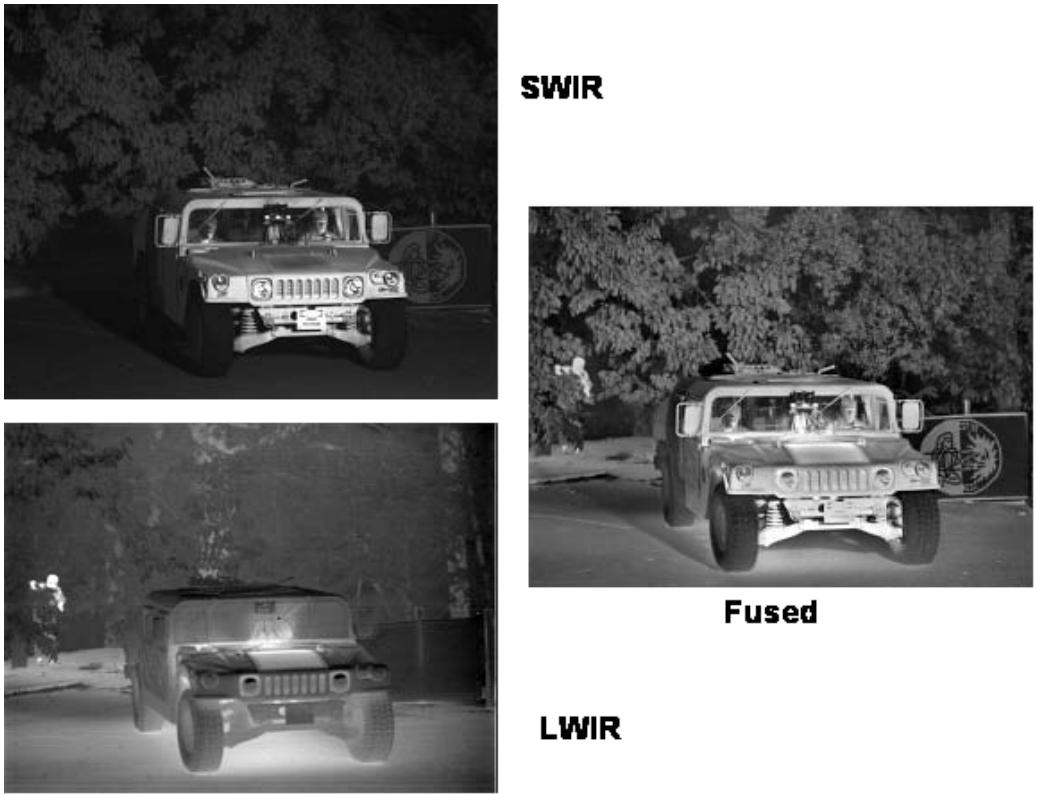


Figure 1. Infrared imagery used to illustrate possible benefits achievable with image fusion.

However, determining the effectiveness of an image fusion algorithm is not a trivial task. Several reasons complicate the evaluation of image fusion algorithms (to include but not limited to): lack of proper registration between the source images; non-linear image fusion algorithms (thereby negating the use of familiar linear tools); lack of an available reference image; and differences in the target sets and scenes used in comparison studies reported in the literature. In addition to these reasons, image fusion is used for a variety of different tasks which make it difficult to identify common requirements for image fusion algorithm development. However, there are three fundamental requirements generally accepted for any fusion algorithm before other considerations are entertained: (i) the fusion algorithm should preserve as much of the

salient information in the source images as possible; (ii) the fusion algorithm should not generate artifacts which affect the human observer's ability to perform the task; and (iii) the fusion algorithm must be tolerant of imperfections in the source imagery such as noise or mis-registration. Besides the fusion algorithm itself, other pre and post processing operations such as rescaling of the source data and manual gain and level for optimum contrast in the fused data, along with display quality, may affect the quality of the fused image. These considerations may have as much influence on the fusion quality as the actual fusion algorithm if not attended to properly.

Quantitative image fusion quality assessments usually consist of calculating the similarity between the fused image and either a reference image (Smith, Chari, Halford, Fanning, & Reynolds, 2009) or the source images. Usually there is not a fused reference image available for comparison, therefore the development of non-reference objective measures for predicting image fusion quality is needed. There are many objective measures being reported in the literature. Wang, Yu, and Shen (2009) presented an overview of some common image fusion metrics. Accurate assessment of human performance using image fusion algorithms in response to system changes is needed for the development of better fusion algorithms and their implementations.

In general, proving objective measures useful with assessing whether a fused image has better quality over its source images is an important task, but for the target acquisition community, the objective measure of interest is human task performance. The most accurate means for evaluating how a human observer will perform a particular visual task using various image fusion algorithms is to measure performance through direct testing. The value of interest obtained from perception testing is the average performance of the

observers for each particular visual task. Given a proper amount of test participants, the uncertainty associated with the average performance measure can be minimized and a sufficient estimate of expected visual task performance can be measured. Perception tests, albeit an ideal measure of human task performance, are time consuming and very expensive. Identifying an objective measure that correlates to human visual task performance would alleviate the need for exhaustive subjective testing, along with its associated high cost and demanding time requirements. I present in this dissertation the results from an investigation into the effects fused source band complementary and redundant information have on the identification task.

2 BACKGROUND

Many military operations require soldiers to perform visual tasks, such as detection, recognition and identification (DRI) of targets. The military is very interested in improving human observers' ability to perform DRI tasks. The military is also very interested in improving situational awareness for their military observers. These tasks are conducted in a wide range of environments and in both daytime and nighttime settings. To this end, the military continuously seeks to improve its imaging capabilities for day and night operations. Information apparent in one spectral band may not be present in another. Combining multi-sensor data onto a single display or into a single image supports the need to provide each soldier or end user with as much high quality information in the most efficient manner as possible. To this end, in this dissertation, the quality of information transferred to the fused images is evaluated and used to quantify the benefit of image fusion.

An accurate model of image fusion based task performance will depend on how much the atmosphere, along with an imager's noise and blur affects the reproduction of the scene. These environmental effects and imaging system imperfections affect the human observer's ability to extract information about a target within the imagery. In addition, the behavior of image fusion algorithms in response to factors like registration errors, noise, blur, and contrast reduction also affects the human observer's ability to extract information about a target. As such, there are many factors that affect image fusion quality; the need for an objective metric is clear.

2.1. Target Acquisition Models

Target acquisition (TA) models serve as a means for quantifying the quality of an image (or sensor) by predicting the field performance of quantities such as the probability of detection, recognition or identification of a set of targets. As the image information propagates from the scene through the atmosphere to the optical instrumentation and then display, a loss of information about the target occurs. The U.S. Army RDECOM CERDEC Night Vision and Electronic Sensors Directorate (NVESD) TA model, NVThermIP[1], models this process well for a single sensor system. The objective for the model NVThermIP is twofold: it models sensor performance and task performance. The theory behind NVThermIP is based on the following two assumptions:

1. Target acquisition performance is related to perceived image quality, that is, how well a human sees a target depends on the perceived quality of the imagery.

2. Perceived image quality is related to threshold vision. This basically states that if a human observer cannot perceive degradations within the imagery then these degradations have no effect on TA performance.

These two assumptions provide the basis for determining the limits on TA performance imposed by both the sensor and the human observer. The objective of sensor performance modeling is to predict the probabilities of an ensemble of military observers using electro-optic/infrared (EO/IR) sensors to accomplish a predefined military task as a function of range. The objective of task performance modeling is to

[1] NVThermIP (U.S. Army NVESD, 2005) is a computer program that models thermal imaging systems in the mid to far infrared spectral bands (3-12 μm). This model is used for thermal systems that sense emitted infrared light. NVThermIP predicts target acquisition performance that closely matches the capability of the sensor being modeled.

determine the sensitivity and resolution requirements for successful completion of military tasks, based upon a description of target size and signature, along with an empirical measure of the task difficulty for a human observer.

Based on the above assumptions, if the contrast of a target exceeds some threshold, target and background discriminations can be made allowing the target acquisition task to be performed. Vollmerhausen, Jacobs, and Driggers (2004) gave an equation for quantifying a human observer's vision while looking through EO/IR sensors, expressed as:

$$CTF_{SYS}^{2}(f)=\frac{CTF_{EYE}^{2}(f)}{MTF_{SYS}^{2}(f)}\left\{1+k^{2}\sigma_{N}^{2}(f)\right\} \tag{2.1-1}$$

where f is the spatial frequency in cycles per milliradian at the target, CTF_{EYE}^{2} is the contrast threshold function (CTF) of an unaided eye, MTF_{SYS}^{2} is the system MTF, k is a calibration factor and σ_{N}^{2} is the perceived noise.

Vollmerhausen *et al.* (2004) has shown that the TTP metric given in equation 2.1-2 defines the effective bandwidth, weighted by the sensitivity of the system, that is available to the human observer, under the assumption that an ensemble of targets has an average target contrast C_T at all spatial frequencies. The TTP metric is given as:

$$TTP=\int_{f1}^{f2}\sqrt{\frac{C_T}{CTF_{SYS}^{2}}}df \quad (cycles/mrad) \tag{2.1-2}$$

where f1 and f2 are the lower and upper frequency intercepts, respectively, between the target contrast and the system CTF. The TTP metric determines the excess modulation as

a function of spatial frequency, weighted by the system CTF. This weighted bandwidth can be viewed as a measure of perceivable information available to the human observer for a particular sensor.

When the TTP is multiplied by the ratio $\left(\frac{\sqrt{Area_{tgt}}}{Range} \right)$, the TTP metric yields what is referred to as V. The V value to obtain 50% probability of task performance is determined through experimentation. This value is referred to as V50. V and V50 are used to obtain the actual performance prediction curve Target Transfer Probability Function (TTPF) that NVThermIP outputs. The expression used to obtain the curve has the form:

$$TTPF(V) = \frac{\left(\frac{V}{V_{50}} \right)^{\beta}}{1 + \left(\frac{V}{V_{50}} \right)^{\beta}} \tag{2.1-3}$$

the coefficient $\beta = 1.54 + 0.24\left(\frac{V}{V_{50}} \right)$.

According to NVThermIP model, V is a measure of image quality and V50 is a measure of task difficulty. The relationship between human observer performance and image quality is obtained through the TTPF.

A measure is needed that can accurately account for the loss or increase in information as a direct result of image fusion. Once a measure is validated using this criterion, the metric's outcome needs to be integrated into TA models.

The concept of weighted bandwidth being viewed as a measure of perceivable information available to the human observer for a particular sensor served as a motivation for this research. The questions I sought to answer were: given two separate sensors to be fused, how should information within competing frequency band intervals between sensors be handled? How much redundant information exists between overlapping as well as unique frequency bands? I decided to explore the latter question as the former seems to be more related to fusion algorithm development. Determining the complementary or redundant information within a specified frequency interval provides invaluable insight for developing better image fusion algorithm comparison methods as well as aiding the prediction of visual task performance with image fusion.

2.2. Complementary and Redundant Information

Information is being defined as spatial frequency information for the purposes of this dissertation. Complementary information exists when imagery of the same scene is captured in different spectral bands. Likewise redundant information exists when imagery of the same scene is captured in similar spectral bands. In this context, redundant or complementary information can be described as overlapping or unique spatial frequencies between spectral bands, respectively. This definition of complementary and redundant information leads to the conclusion that different spatial frequency bands within the same sensor possess only complementary information. Although the spatial frequency information is complementary, the perceivable information is often similar enough to perform task such as identification of targets. Driggers, Vollmerhausen, and Krapels (2000) conducted a study investigating the importance of low spatial frequency information versus high spatial frequency

information for an identification task. The results of this study indicated that low spatial frequency information does contribute to overall performance. This means that the entire target spectrum, complementary and redundant target information, is used in the identification task. It should be noted that the importance of the target spectral regions remains uncertain. With regards to image fusion, when fusing the information from different imagers, the final resultant image will be made up of various amounts of complementary and redundant information. Common spatial frequency bands between different sensors can contain redundant information or they can contain complementary information. Likewise complementary spatial frequency bands between sensors can contain redundant information or they can contain complementary information. In general, the fused image information can be described by the following equation:

$$\text{Fused information} = \text{Complementary information} + \text{Redundant information} \qquad (2.2\text{-}1)$$

The fusion of complementary and redundant information from sensors with different characteristics allows for a more complete understanding of the surroundings, thus allowing for improved situational awareness and better planning and decision making. This concept is not new. Humans gather information from multiple senses like sight, touch and smell to make decisions necessary to perform everyday tasks. The sense of hearing can be combined with the other senses to allow for decisions to be made in environments where sight, touch and smell may be severely degraded or impeded altogether.

As early as the 1950s, data fusion was used to obtain a better quality composite image more suitable for identifying natural and manmade objects. Image fusion is a branch of data fusion, where data fusion is defined as the process describing the means

and tools for the alliance of data originating from different sources. In the 1980s, the development of image fusion algorithms began to evolve from primitive techniques to more advanced methods. Harper, Lilliesand, and Keifer (1990) and Chavez, Sides, and Anderson (1991) recognized that obtaining a single image with more information than the source images was achievable. Towards the end of that decade, Hall and Liinas (1997) and Pohl and Genderen (1998) gave two very good introductions to multi-sensor data fusion. These papers gave good explanations of the concepts, methods and possible applications pertaining to image fusion. Although these papers certainly do not mark the origin of image fusion precisely, they provide a comprehensive description of where image fusion was up to that point in time. Image fusion has gained a great deal of attention over the years since then.

2.3. Image Fusion

The concept behind image fusion is easily understood, however a broad definition for image fusion, like the one offered in the introduction, is needed because of the different tasks and types of information being combined. Some situations require a single image that contains high spatial resolution and high spectral information, as in remote sensing or pan sharpening, in an attempt to optimize human visual task performance. Image fusion offers a solution for situations such as these. If we think of the information in each source image in terms of sets, then the fusion of image information can be thought of as a set reduction as opposed to an integration of information. Set theory concepts can be evoked to help understand this concept.

When dealing with problems concerning the interrelationship of sets, Venn diagrams are useful visual tools that provide assistance with understanding the

relationship of sets. Referring to the Venn diagrams in Figure 2, if we let A represent the set that contains all the information within image 1 and we let B represent the set that contains all the information in image 2, then the union, intersection and difference operations of the sets can be explained as follows:

- The union (or sum) of two sets, A and B, is the set of all those elements that belong to A or B, or both.
- The intersection (or product) of two sets A and B is the set of all the elements that belong to both A and B.
- The difference A – B of any set B relative to the set A is a set consisting of all elements of A that are not elements of B. This can also be understood as the relative complement of B in A.

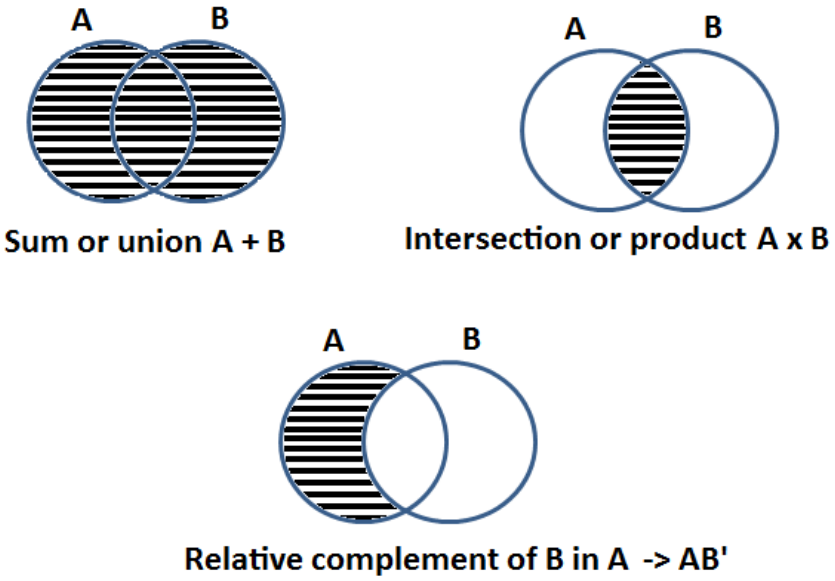


Figure 2. Venn diagram illustrating the union, intersection and difference operations on sets.

The relative complements of B in A and the relative complement of A in B are used to obtain the complementary information in both A and B; that is $\overline{\mathrm{A}} = AB'$ and $\overline{\mathrm{B}} = BA'$, is found to be:

$$\overline{\mathrm{A}} = \mathrm{A} - (\mathrm{AB}) \tag{2.3-1}$$

$$\overline{\mathrm{B}} = \mathrm{B} - (\mathrm{AB}) \tag{2.3-2}$$

The sum (or union) of equations 2.3-1 and 2.3-2 yields:

$$\overline{\mathrm{A}} + \overline{\mathrm{B}} = (\mathrm{A} + \mathrm{B}) - (\mathrm{AB}) \tag{2.3-3}$$

To see that equation 2.3-3 holds, the difference relation $A - B = AB'$ given in set theory and DeMorgan's law $(AB)' = A' + B'$ may be used. The verification of equation 2.3-3 is as follows:

Using the right hand side of the equation 2.3-3,

$$(\mathrm{A} + \mathrm{B}) - (\mathrm{AB}) = (\mathrm{A} + \mathrm{B})(\mathrm{AB})' \tag{2.3-4}$$

by virtue of the difference law previously given. An application of DeMorgan's law can be applied to one of the terms on the right hand side of this equation to obtain:

$$(\mathrm{A} + \mathrm{B})(\mathrm{AB})' = (\mathrm{A} + \mathrm{B})(\mathrm{A}' + \mathrm{B}') \tag{2.3-5}$$

$$(\mathrm{A} + \mathrm{B})(\mathrm{A}' + \mathrm{B}') = (\mathrm{AB}') + (\mathrm{A}'\mathrm{B}) = \overline{\mathrm{A}} + \overline{\mathrm{B}} \tag{2.3-6}$$

Rewriting equation 2.3-3, replacing sums with unions and products with intersections:

$$\overline{\mathrm{A}} \cup \overline{\mathrm{B}} = \mathrm{A} \cup \mathrm{B} - (\mathrm{A} \cap \mathrm{B}) \tag{2.3-7}$$

Using set theory notation, equation 2.3-7 can be interpreted as the union of the complementary information present in the source bands is equivalent to the union of the total information in A and B minus the intersection of information shared between them. By letting the union of A and B represent the fusion process, then it can be seen that set reduction occurs when the redundant information between the inputs is filtered out. This line of reasoning will be continued in chapter 4 regarding the development of the proposed information measure.

As previously stated, image fusion, in general, is a process of combining the complementary and redundant information contained in a set of source images to obtain a single resultant image with more relevant information than any individual source image. The information involved in a linear fusion process is shown in Figure 3. A fusion algorithm that obeys linearity principles does not introduce additional information into the fused image that was not originally present in either source image. It is the intent of image fusion to combine the complementary and redundant information from multiple images into a single image that renders a better description of the scene than any of the individual source images. For image fusion to be beneficial, the resultant image (usually a composite image) is required to be more suitable for the purposes of human visual perception task performance. It should also be noted, the relevance of the information being fused depends highly on the intended visual perception task (Krebs & Sinai, 2002). It is the goal of image fusion to preserve and combine all the relevant information present in the source imagery into a single output image. The output image should improve information content making it easier to perform target acquisition tasks such as DRI tasks as well as situation awareness. Furthermore, the task performance using the fused

imagery should be, at a minimum, at least as good as that using the higher performing individual source band for a particular task. Three common motivations for considering the use of image fusion for a particular task are to increase the reliability and capability of a task or to increase the range of operation.

Combining spectral signatures from multiple spectral bands is thought to be a process that gives the end user a single image with the best quality of information so that the performance on a particular task can be optimized. However, quantifying the effect complementary information has on image fusion algorithm performance remains open research.

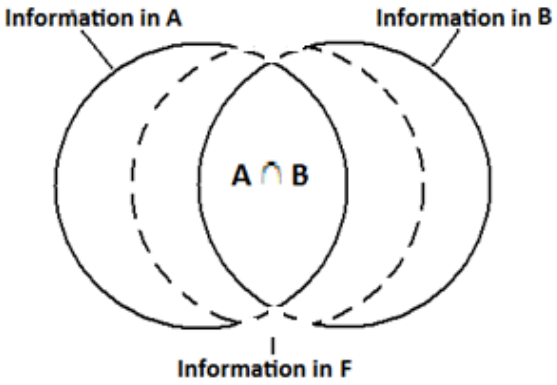


Figure 3. Information diagram for the total information in a fused image F. An ideal fusion process would capture and transfer all the complementary and redundant information in A and in B to the fused image.

2.3.1. Image Fusion Methods

As a general classification, image fusion methods can be described as being either spectral or spatial domain techniques, with three different levels of information represented as: pixel level, feature level and decision level (Pohl & Genderen, 1998):

1. Pixel based fusion is the process of combining information on a pixel-by-pixel basis from the pixels in a set of source images with the goal of achieving improved performance with various image processing tasks, such as DRI tasks.

2. Feature level fusion is the process of extracting and combining the salient features present in a set of source images such as pixel intensities, edges and textures.

3. Decision level fusion is a process performed at the highest level of abstraction. Salient information is extracted from each source image separately and then combined applying decision rules to achieve a final fused image.

Pixel based fusion methods require operation directly on the images without the need of any transformations or decompositions to the images. These techniques are often relatively simple and fast thereby lending themselves favorably to real time processing, but not without costs. Pixel level fusion algorithms such as pixel averaging greatly reduce the contrast in the fused image and are sensitive to registration accuracy. Feature level methods require images to undergo some amount of transformation or decomposition to extract features from the image. The input images are transformed or decomposed into multiple scales or levels and various fusion rules are applied at each level to select which image component will be represented in the corresponding fused image's level. An inverse transform is then carried out to obtain the fused image. Some associated problems with feature based methods are, at times, spectral content of small objects have the potential to not be reproduced in the fused image; the algorithm developer is left to determine appropriate threshold values; also, there is a bit more computational complexity in feature based methods compared to pixel based methods.

Feature based methods can be referred to as pre-detection methods while decision level methods can be referred to as post-detection methods. I defer the reader to Gunatilaka and Baerlein (2001) for more insight into decision level methods, as the majority of the algorithms used in this dissertation are feature level fusion processes.

Image pyramid approaches and wavelet based methods are some popular spectral domain techniques. In image pyramid approaches, an image pyramid is constructed consisting of multiple scales of low-pass or band-pass copies of an image, where each level usually is a factor of two smaller than its predecessor. Furthermore, each successive level is created from a decomposition of the lower spatial frequencies in the image. All the original information needed to reconstruct the original image is contained in the image pyramid, therefore, to obtain the original image an application of fusion rules and an inverse transform is needed.

Wavelet based methods are similar to pyramid based approaches in that wavelet methods also decompose the image into multiple scales, which are constructed by a selection of the salient wavelet coefficients. Various selection rules can be applied to select the coefficients, and then an inverse wavelet transform is conducted to recreate the original image. Multi-resolution approaches, such as those based on pyramids and wavelets, are constructed to take advantage of the human visual system (HVS) being primarily sensitive to local contrast changes, i.e., edges.

2.3.2. Image Fusion Quality

It is a goal of image fusion to ultimately generate a composite image containing a better quality of information; the interpretation of ‘better quality’ will depend on the intended application (Wald, 1999). The quality of different image fusion schemes is

historically determined by subjective evaluations (Ryan & Tinkler, 1995; Toet & Franken, 2003). A great deal of research is available on investigating image fusion quality, both objectively and subjectively (Arathi & Soman, 2009; Canga *et al*. 2005; Cvejic, Loza, Bull, & Canagarajah, 2005; Cvejic, Loza, Bull, & Canagarajah, 2006; Hossny, Nahavandi, & Creighton, 2008; Howell, Halford, & Krapels, 2010; Moore, Halford, & Howell, 2007; Nava, Escalante-Ramirez, & Cristobal, 2007; Neriani, Pinkus, & Dommett, 2008; Petrovic, 2007; Petrovic & Xydeas, 2005; Piella & Heijmans, 2003; Qu, Zhang, & Yan, 2002; Wang & Bovik, 2002; Wang & Bovik, 2004; Wang & Bovik, 2009; Waxman *et al*., 1997). As previously stated, image fusion offers the end user several benefits such as: reduction in uncertainty, improvement in reliability, robustness and capability of the system performance, along with a broader range of operation.

2.4. Performance Benefit of Image Fusion

Moore (2007) investigated the performance benefits achievable with image fusion by conducting a controlled case study in which human observers were tasked to identify basic numbers between 0-9 in a human perception identification experiment using various image fusion algorithms. Experimental details are given in Figures 4a and 4b. In this experiment Moore created an image of white Gaussian noise and applied a threshold to the noise image using Laplacian pyramid decomposition. The binary noise image was then subtracted from a matrix of ones and both were applied to a set of ideal images. This process created purely complementary source images to be fused. The degraded images were used as inputs into various image fusion algorithms. Observers were asked to simply identify the numbers presented to them without knowledge of which fusion algorithm was used to recombine the numbers. Shown in Figure 4c are the results of the

performance evaluation using discrete wavelet transform (DWT) fused imagery. Human performance using the DWT fused imagery clearly outperformed each of the individual source images. However, the issue surrounding the ability to predict the increase in performance remained unclear. Nonetheless, this work suggests that, at least with complementary spatial information, a fusion performance increase over source band performance can be realized.

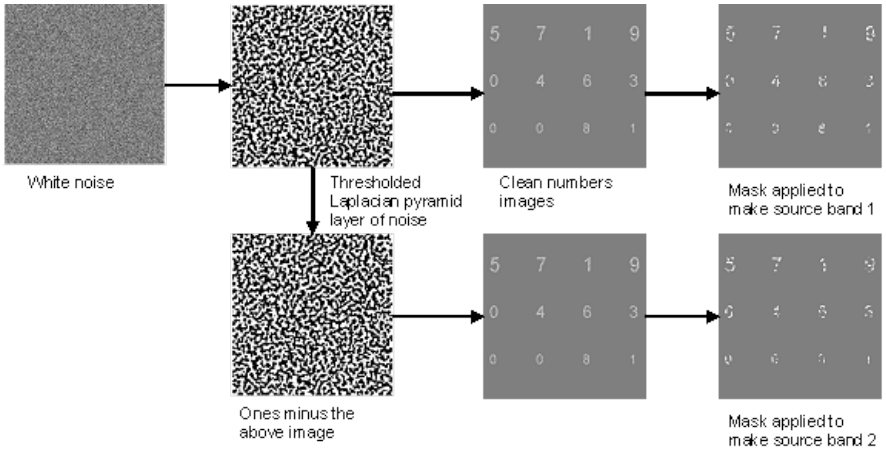


Figure 4a. Process to generate a contrived set of complementary source images

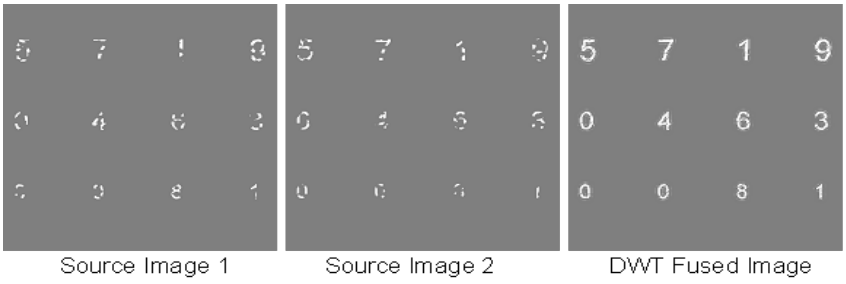


Figure 4b. Example test imagery. Shown are two complementary source images and their fused result using a DWT fusion algorithm.

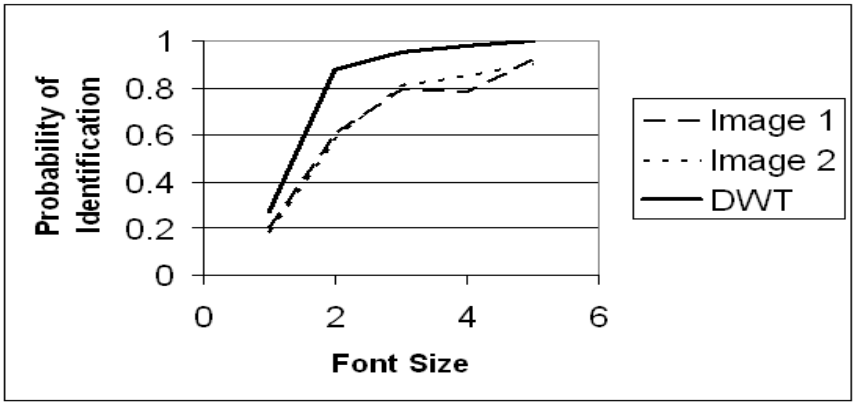


Figure 4c. PID results from perception tests evaluating all the DWT fused imagery along with the source imagery.

2.5. Objective Measures of Image Fusion Quality

Image fusion has been applied to a large variety of multi-sensor applications and, as a result, selecting an optimum fusion algorithm for each application is extremely difficult. In general, it is a difficult task to quantify the benefit associated with using image fusion. Some possible reasons for this are: there usually is not a reference image available for comparison; registration errors, pre and post processing techniques vary depending on who registers the imagery; and quite often there are vast differences in the scenes and targets being fused for comparison. Additionally, the task of quantifying image fusion algorithm performance is complicated further because so many objective measures for image fusion only work well for certain types of algorithms. A common medium with which to compare the quality of fusion algorithm performance is necessary due to factors such as the variety of scenes, variability in registration error, type of spectral bands fused, and differences in the target sets. So how does one go about assessing image fusion performance? To begin answering this question, one needs to focus on the intent of image fusion. The ultimate purpose of image fusion is to provide

information to the end user, i.e., information not readily available in individual source images that allows for an increase in task performance.

Many applications call for a human as an end user of the fused result, therefore human perception of the fused result is very important. Human perception experiments are the primary means for evaluating DRI task performance and are essential in developing and validating TA models. NVESD has identified standard procedures for the analysis of DRI experiments, so that the output from the experiments (typically the probability of correct response) can be repeated and validated with field tests. While perception experiments are ideal for evaluating human task performance, they are time consuming and expensive to conduct. Many objective measures are being investigated to perform trade studies regarding the performance of different image fusion algorithms. In addition to correlating with human performance, a non-reference measure of image fusion quality should address some of the short comings human perception evaluations present. For instance, measurements using objective metrics should be less expensive, faster than human evaluations and repeatable. The main area of interest surrounding objective metrics is whether they possess the capability to predict human task performance. Therefore, an objective performance metric that correlates with human visual perception is extremely desirable. Ultimately we desire a fused image that contains more relevant information than any individual source image. It seems reasonable to approach assessing image fusion performance by evaluating the information content transferred between the source and the fused images, or by evaluating the reduction in uncertainty of image information.

Using Shannon's assumption that one element of a large number of messages from an information source is just as likely as another and, thereby, drawing a comparison to quantifying information in communication theory, many useful methods have been established to quantify the information content in an image. By satisfying this well known assumption, a measure of the information within an image can be achieved using the concepts of information theory. To obtain a measure of the average information (entropy) in an image estimated in the units of bits per pixel, the following equation can be used:

$$H = -\sum_{i=1}^{N} f(i)\log_2 f(i) \tag{2.5-1}$$

Here, N represents the number of grayscale levels in the image's histogram and $f(i)$ is the normalized frequency of occurrence of each grey level. For the purpose of this research, entropy is understood to be a measurement of the uncertainty in the information a single random variable contains. This definition differs from that of physics entropy, in that physics entropy is a measure of the probability of the state of a particular system. Physics entropy is governed by the second law of thermodynamics, which states that the total entropy of the universe will never decrease. However, this research is concerned with statistical entropy, which is a probabilistic measure of the uncertainty or ignorance one has regarding a random variable. Intuitively, an increase in image information is desired if there exists ignorance or uncertainty regarding what is being observed. Therefore, knowledge is gained with increased information, and as a result, a reduction in uncertainty surrounding the topic in question is achieved. This implies that

measurements of reduction in uncertainty allow for measurements of information and vice versa.

Entropy can be used as a measure to estimate the change in information between the source and the fused images. However, entropy as a direct measure of information transfer should be used with caution. If the entropy in the fused image is higher than the entropies in the source image, one may erroneously conclude the fused image contains more information. In fact, it is not clear if the "higher information" is due to noise or an actual increase in information. Additionally, with regards to understanding human perception of information, entropy may fall short on this task as well. As an example, Figure 5 displays an image multiple times processed in such a manner that each image contains identical pixel intensities and therefore identical histograms. The pixel intensities of a portion of each image were successively randomly redistributed and entropy measurements were taken for each image. Figure 5 shows that each image had the same entropy regardless of the perceivable information shown. Although the amount of perceivable information obviously changed, entropy was insensitive to the change. Additionally it can be observed that mutual information (MI) reacts in a desirable manner. The MI measure for each successive image, unlike entropy, gives higher and higher values as more of the information is made perceivable. The question that remains with this type of an analysis is can this type of relationship between human perception of visual information and MI be exploited to predict visual task performance?

MI is an established, non-reference image quality metric and is a reasonable metric for evaluating image fusion performance (Hossny *et al*., 2008; Nava *et al*., 2007; Qu *et al*., 2002;). MI is a well known concept of information theory and measures the

amount of information that one image contains about the other by measuring the similarity of their intensity distributions. Given two random variables X and Y, MI measures the degree of dependence between X and Y by measuring the distance between their joint statistical distributions and that of total independence. Higher MI values indicate higher similarity between the distributions of the corresponding image pairs and therefore a larger reduction in uncertainty. Smaller values indicate smaller reductions in the uncertainty between the random variables, with zero representing independence. MI also captures some of the slight differences of the spectral characteristics of the images. MI can be defined as (Reza, 1961):

$$I_{XY}(x, y) = H_X(x) + H_Y(y) - H_{XY}(x, y) \tag{2.5-2}$$

$H_X(x)$ and $H_Y(y)$ represent the entropies of image x and y; $H_{XY}(x, y)$ is the cross entropy and this is the term that maximizes MI. MI between two random variables can be found with knowledge of their cross entropy.

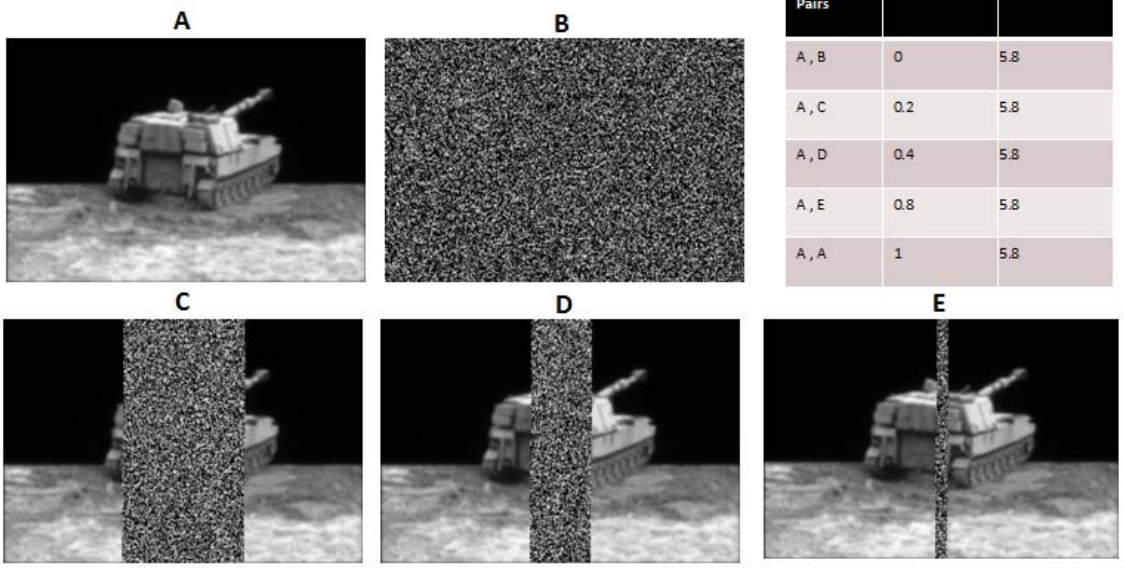


Image Pairs	MI Values	Entropy
A , B	0	5.8
A , C	0.2	5.8
A , D	0.4	5.8
A , E	0.8	5.8
A , A	1	5.8

Figure 5. Comparison between MI and entropy. Entropy appears insensitive to changes in perceivable image information while MI shows sensitivity to changes in the perceivable image information.

Spatial domain techniques akin to pixel averaging will score high using MI. This is because the fused image is simply equal to a linear combination of the input images plus some constant term. This characteristic does not serve MI well as a metric for comparing different types of image fusion algorithms. However, with MI, there is no assumption regarding the relationship of the pixel intensities of the images, so MI is useful for comparisons between similar algorithms. It is a measure of the degree of dependence between the random variables under investigation.

Case studies similar to that shown in Figure 5 justify, or at the very least strongly suggest, the need for a "smarter" measure of information when attempting to evaluate human perception of image information. Many investigations into MI, or other information theory based metrics, are carried out for this purpose.

There are a variety of other measures currently reported in the literature that are information theory related or statistics feature related, each with their own pros and cons. The difficulty surrounding this area of research is due to the variety of application requirements and the lack of a reference image to compare to the composite image. Often an ideal composite image is created and used as a reference for comparing the fused image results (Li, Manjunath, & Mitra, 1995; Rockinger, 1997; Smith *et al*. 2009). The creation of a reference image allows the use of some widely used metrics such as mean squared error (MSE), signal-to-noise ratio (SNR), peak-signal-to-noise ratio (PSNR), and standard deviation to somewhat validate the fusion algorithms. However, in practical situations, a reference or test image is not available, which in turn motivates the development of objective non-reference image fusion quality assessment measures.

Q. Wang *et al.* (2009) gives an overview of other image fusion metrics found within the literature. Z. Wang and Bovik (2002) introduced an image quality index (IQI) measure to quantify the structural distortion between two input images. The IQI can be described as a product of three components: (i) a measure of the correlation coefficient between the two input images; (ii) a kind of average luminance distortion measure; and (iii) a contrast distortion measure. The IQI gives a measure of the similarity between two input images, where one image is meant to be a reference image. Piella and Heijmans (2003) extended the IQI by Wang and Bovik to include a comparison between the fused image and the source images. The new measure is called the fusion quality index (FQI) and it gives an indication of the amount of salient information transferred to the fused image from the source images. This measurement is accomplished by using the local statistics of a $n \times n$ sliding window as the saliency factor for each image. Piella and

Heijmans revised the FQI to assign more weight to the salient regions in the images. The revision is called the weighted FQI (WFQI). This was done in an attempt to model the HVS, which is known to give higher importance to visually salient regions in an image. The authors also modified the WFQI to take into account the edge information available in an image. This was done in an attempt to model the importance of edges in the HVS. Tsagaris and Anastassopoulos (2006) introduced a new global measure for image fusion methods. Their image fusion performance measure (IFPM) uses MI as well as conditional MI to evaluate the amount of information transferred to the fused image from the source images. Tsagaris *et al*. used conditional MI to guarantee that overlapping information within the inputs would only be used once in the evaluation of various fusion algorithms.

3 Image Fusion Algorithms

It is desired to identify image fusion methods suitable for real time implementations. The quality of the spatial registration of the images being fused influences the overall performance associated with each image fusion algorithm. The details and some of the relative strengths and weaknesses of the fusion algorithms used in this research are presented in the following sections. For clarity, when image fusion is referenced, it is understood as fusing two source images only.

3.1. Laplacian Pyramid

The Laplacian Pyramid (LAP) is obtained using an iterative sequence of difference images between two successive levels of Gaussian pyramids. This technique is also referred to as the Difference of Low Pass (DoLP). At each resolution level, the difference image is created by subtracting a weighted, up-sampled low-pass filtered version of the image from itself.

Let G^k be the K^{th} (k=1, …N) Gaussian pyramid level for each of the source images S1 and S2 respectively. Then $G_1{}^0$=S1 and $G_2{}^0$=S2. Each K^{th} level in the source images is convolved with a Gaussian kernel, down-sampled then assigned to the $(K+1)^{th}$ level of the pyramid. The pyramid layers decrease in size due to the down-sampling operation that shifts the spatial frequencies in the image relative to the Gaussian blur kernel. This process is represented by the following:

$$\begin{aligned} G_1{}^0 &= S1 \\ G_2{}^0 &= S2 \\ G_{i=1,2}{}^{k+1} &= [\,\mathrm{w} * \mathrm{G}_{\mathrm{i=1,2}}{}^{\mathrm{k}}\,]_{\downarrow 2} \quad for\ k = 1,\ldots N-1 \end{aligned} \tag{3.1-1}$$

The Gaussian pyramid is constructed with low-pass filtered copies of the original image, each having a cut-off frequency one octave lower than the previous level. Then the Laplacian pyramid is represented as:

$$
\begin{aligned}
L_1^{\,N} &= G_1^{\,N} \\
L_2^{\,N} &= G_2^{\,N} \\
L_{i=1,2}^{\,k} &= G_{i=1,2}^{\,k} - 4\mathrm{w} * [\,\mathrm{G}_{\mathrm{i=1,2}}^{\,\mathrm{k+1}}]_{\uparrow 2} \quad for\ k = 0,\ldots N-1
\end{aligned}
\tag{3.1-2}
$$

The fused image can be obtained by fusing together each corresponding pyramid level of the source images using fusion selection rules such as taking the maximum difference value between the two source images at each level. The Laplacian Pyramid algorithm used is based on Peter Burt's Laplacian Pyramid (Burt & Adelson, 1983; Adelson, Anderson, Bergen, Burt, & Ogden, 1984). Two versions of the Laplacian fusion algorithm are used in this research; LAPA and LAPB only differ by the Gaussian blur kernel used for reconstruction of the final fused image.

3.2. Ratio of contrast

Toet (1989) introduced an image fusion technique that was motivated by the human visual system's dependence on contrast. The local luminance contrast is preserved in the fused resultant image, thereby providing better details to a human observer. Local luminance contrast is defined as:

$$
C = \frac{L - L_b}{L_b} \tag{3.1-3}
$$

where L is the luminance contrast at a certain location in each image and L_b is the local background luminance.

The ratio of contrast (referred to in this research as Contrast Ratio) fusion process is similar to the DoLP described in section 3.1 in that the difference operation in equation 3.1-2 is now represented by a ratio. The K^{th} level $R_{i=1,2}{}^{k}$ of the Ratio of Low-Pass (RoLP) pyramid is obtained as:

$$R^N = G^N \tag{3.1-4}$$

$$R_{i=1,2}{}^{k} = \frac{G_{i=1,2}{}^{k}}{4w * [\,G_{i=1,2}{}^{k+1}\,]_{\uparrow 2}} \quad for\ k = 1,\ldots N-1$$

The fused image is obtained by using a fusion selection rule that retains the image features with high contrast. There are two versions of the Contrast Ratio (CR) fusion algorithm, CRA and CRB, which differ only by the Gaussian blur kernel used for reconstruction of the final fused image.

3.3. Opponent processing

In the opponent processing (OP) image fusion algorithm, the polarity of one of the source bands is reversed before the images are combined. The result is then scaled by the reciprocal of the sum of the original images. $F_{opp}(x, y)$ is found from $S_1(x, y)$ and $S_2(x, y)$ as shown below:

$$F_{opp}(x, y) = \frac{S_1(x, y) - S_2(x, y)}{S_1(x, y) + S_2(x, y) + C} \tag{3.1-5}$$

C is a constant to control the shape of the scaling. In this research, two versions of the OP algorithm are used, OPA and OPB. OPA and OPB represent the cases when the polarity of each source band is reversed, ‘A’ for source band 1 and ‘B’ for source band 2.

3.4. Multi-scale decomposition based fusion

The multi-scale (MS) decomposition based fusion was implemented by an iterative process of obtaining a multi-scale decomposition of the two source images independently while applying a weighting operation at each resolution level. The fused image is obtained by a superposition of the weighted resolution levels of each image. Using multiple scales reduces the possibility of losing image detail in the fused image.

3.5. Shift invariant discrete wavelet transform

The DWT (Huntsberger & Jawerth, 1993) uses multiple scales to decompose the images and thus reduces the possibility of losing image detail. Due to the sub-sampling process, wavelet coefficients are not always properly compared. To solve this problem, a shift-invariant discrete wavelet transform (SIDWT) was developed where the down-sampling of the standard wavelet transform has been omitted. The fused low-pass coefficients are generated by averaging the low-pass coefficients of the source images. The fused high-pass and band-pass coefficients are generated by comparing the magnitude of the high-pass and band-pass coefficients in the source images and selecting the larger magnitude on a pixel-by-pixel basis. To retrieve the fused image, the inverse SIDWT is applied (without the up-sampling). The SIDWT is more computationally expensive than the DWT, but more often results in visibly better results.

3.6. Pixel Averaging and Superposition

Spatial domain methods like pixel averaging are simple and fast enough for real time implementation. Pixel averaging and superposition are almost identical processes, as pixel averaging adjusts the contrast using a mean value for display, whereas

superposition uses a median value. The following can be used to describe the pixel averaging of two images:

Let X and Y represent the two source images to be fused. Then fusion by superposition can be defined as:

$$F(x,y) = \alpha * X(x,y) + \beta * Y(x,y) \tag{3.1-6}$$

Here the constants α and β represent weights whose sum adds to 1. When these weights are both 0.5 (as used in this study), equation 3.1-6 is also known as pixel averaging. Superposition is somewhat analogous to overlaying the images and reducing the contrast by a factor of 1/n, where n represents the number of images to be fused.

Pixel averaging is a widely used and simple approach to image fusion and is based on the assumption of additive noise. If there are N source images and the variance of noise in the N images is equal, then averaging the images reduces the noise by a factor of N. The major drawbacks of pixel averaging are a reduction in the salient feature in the image and the reduction in image contrast.

3.7. Evaluation of fusion algorithms

Preliminary investigations were conducted to gain more insight concerning how fusion algorithms behaved generally, and in response to complementary source information. Information in this context is being referred to as spatial frequencies.

3.7.1. Linearity testing

A study was conducted to determine if any of the fusion algorithms could possibly be approximated as linear. Having the ability to assume linear operation for some of the fusion algorithms could be invaluable when attempting to develop image

fusion metrics. To this end, a Gaussian impulse image was created, filtered using the filters shown in Figure 11 (chapter 5, section 5.1), fused, and then visually compared to the original impulse image. The frequency response of the impulse images contains all frequencies. By filtering the frequency response of an impulse, many combinations of source information can be combined and observed. The fused output images from this process should closely resemble the input image if the fusion algorithm operates under the assumption of linearity. Figure 6 gives some results from this study.

The impulse image at the top of Figure 6 was filtered into low-pass and high-pass components. These images were used as input into the superposition, Laplacian and Ratio of Contrast fusion algorithms. Each impulse response image is the fused output image from each fusion algorithm in response to fusing the filtered inputs. On the left side of the figure, the impulse response of each fusion algorithm is given along with each respective frequency response. The shifted input image was filtered and used as input into the fusion algorithms. The corresponding impulse and frequency responses are given on the right side of the figure. By inspection, it can be seen that the superposition algorithms exhibited linear shift invariant properties as expected. The Laplacian and the Ratio of Contrast fusion algorithms failed to give similar frequency responses when the source impulse was shifted to a different location in the image.

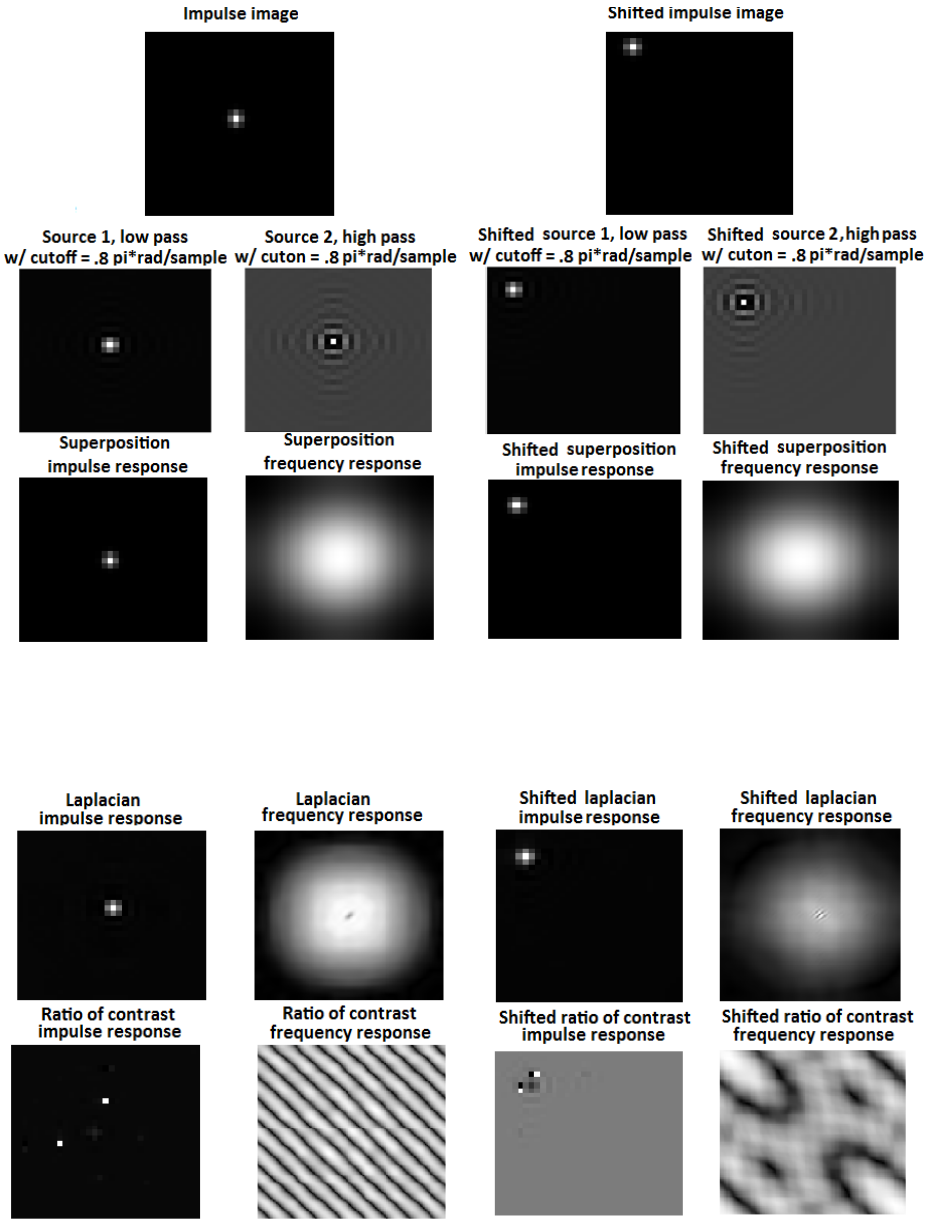


Figure 6. Linearity test for a few common image fusion algorithms. Top images are the original impulse and shifted impulse images. Each impulse image was filtered into low and high pass components and used as inputs for the image fusion algorithms. The fused images are being referred to as impulse response images. The respective frequency responses were obtained using a shifted Fourier transform (the magnification for the shifted ratio of contrast was adjusted due to cropping of image).

3.7.2. Complementary intensity information

As previously shown, many image fusion algorithms are inherently non-linear and often introduce additional information (noise) that could degrade the quality of the information transferred to the fused image. Information in this analysis is not considered spatial frequency information (the source images in this analysis have identical frequency spectrums); however, the analysis does give an indication as to how the fusion algorithms handle complementary spatial information. An analysis of this type allows for assumptions of degradation or enhancement properties concerning each image fusion algorithm.

The performances of a select group of image fusion algorithms were evaluated using purely complementary source images. A single image and its complement, obtained by subtracting the image from an array of ones, were fused and the output of each fusion algorithm was observed. This process can be referred to as "white-hot / black-hot" fusion. As a result of the "white-hot / black-hot" fusion the images fused were perfectly registered, thereby eliminating any errors due to registration. The superposition algorithm was included due to its functional simplicity and linearity; the SIDWT and Laplacian fusion algorithms were chosen to represent wavelet based and pyramidal based fusion schemes in general.

Two source band images were used in this investigation: LWIR and SWIR (given in Figure 7 as LW and SW respectively), where LWc and SWc represent the complements of each source image respectively. Some of the characteristics of each fusion algorithm can be seen by observing the fused images in Figure 7. For instance, the superposition algorithm tends to output constant grayscale values for input pixels that

have complementary intensities. The SIDWT shows its edge enhancement characteristics. The Laplacian fusion algorithm shows the tendency to introduce competing information within the image space.

The analysis is the same for both source images shown in the figure; however, it can be observed that the effects from fusing complementary LWIR imagery in this case were more visually favorable than in the SWIR case. One can make a case in terms of predicting performance, when using complementary source images, based on how the target area is obscured within the images; however, the effects of these characteristics certainly need to be quantified. Figure 7 shows each image fusion algorithm handles complementary information differently: some enhance edge information; some produce constant intensities; while others actually introduce information into the fused image that effectively competes with the source information. It is important to know how each fusion algorithm handles this type of information. Chapter 4 introduces an image fusion performance metric to quantify visual task performance when fusing source complementary and redundant spatial frequency information.

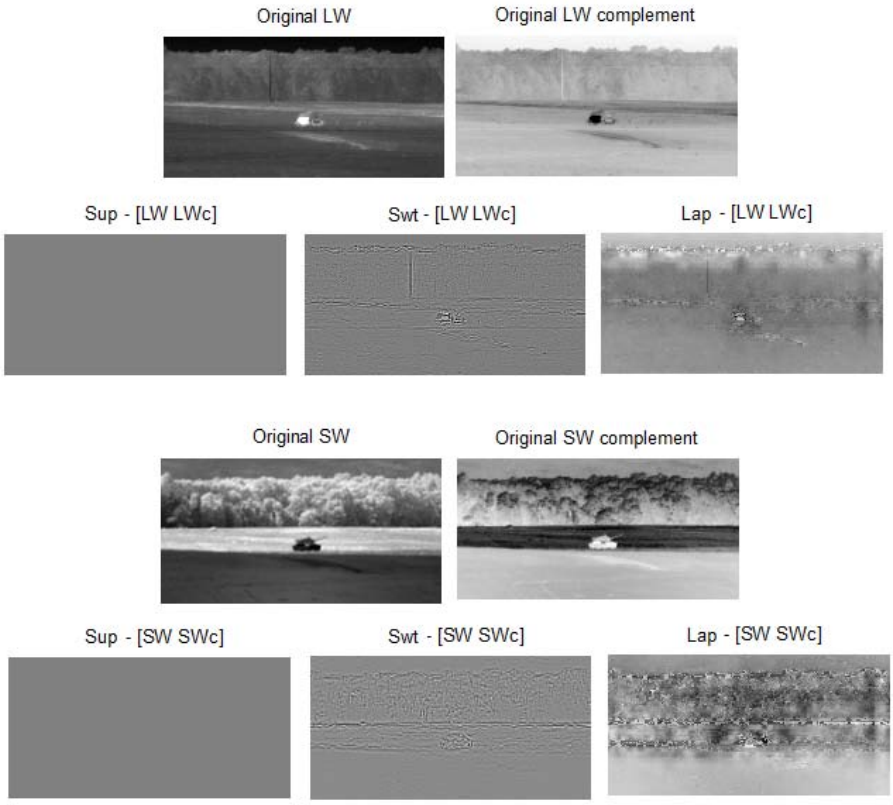


Figure 7. Image fusion using complementary information. LW – long wave; SW – short wave,; LWc and SWc – LW and SW complement respectively. LW and LWc were fused with each fusion algorithm, likewise for SW and SWc. The fused results are shown below each set of original images.

4. Performance Metric

As previous stated there are three performance criteria generally accepted for image fusion algorithms; they are:

(i) The fusion algorithm should preserve as much of the salient information in the source images as possible

(ii) The fusion algorithm should not generate artifacts or inconsistencies which affect the human observer's ability to perform the required task

(iii)The fusion algorithm must be tolerant of imperfections in the source imagery such as noise or mis-registration.

To obtain an effective metric for quantifying the benefit of image fusion, a better understanding of the impact complementary source information has on image fusion algorithm performance is needed. The above criteria serve as a guide for evaluating and comparing multiple image fusion algorithms, however, determining the effectiveness of an image fusion algorithm in terms of visual task performance is not a trivial task. Reasons such as registration errors and lack of a well defined reference image for comparison complicate the evaluation of image fusion algorithms. Section 5.4 gives the results of several analyses investigating the performance of the proposed metric under various types of degradations for a select group of image fusion algorithms.

4.1 Proposed Metric

The quantitative assessment of fusion algorithm performance is complicated by non-linear processes and the lack of an ideal fused image for comparison; therefore, there is a need to identify or develop a non-reference objective metric for determining the

quality of the image fusion process. For this purpose, a new image quality metric referred to as the fused common information measure (FCIM), based on mutual information (MI) is introduced and given as:

$$FCIM = \frac{2 * I_{XY}}{I_{XF} + I_{YF}} \tag{4.1-1}$$

By letting X and Y represent two images with marginal distributions $p_X(x)$ and $p_Y(y)$ respectively, and with joint probability distribution $p_{XY}(x, y)$, the MI between the two images can be described as:

$$I_{XY}(x, y) = \sum_{x,y} p_{XY}(x, y) \log \frac{p_{XY}(x, y)}{p_X(x) p_Y(y)}. \tag{4.1-2}$$

FCIM uses MI to measure the redundant information transferred from the source images X and Y to the fused image F. As complementary information in the source images increases, the FCIM scores decrease. Measuring the transfer of redundant source information to the fused image was the inspiration behind the development of the FCIM metric.

The FCIM can be derived using the definition of MI and a bit of algebra. Referring to Figure 8, the MI between two random variables, say X and Y, can also be defined in the following two ways:

$$I(X, Y) = H(X) - H(X \mid Y) \tag{4.1-3}$$

$$I(X, Y) = H(Y) - H(Y \mid X) \tag{4.1-4}$$

A summation of these two equations yields a separate equation, namely:

$$2 * I(X,Y) = H(X) + H(Y) - H(X \mid Y) - H(Y \mid X) \tag{4.1-5}$$

From this equation, a normalized MI can be derived. It can be seen from equation 4.1-2, only the joint probability distribution is needed to obtain a measure of MI between images X and Y. Referring to equation 2.5-1, the uncertainty in the joint probability distribution can be written as:

$$H(XY) = -\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j)\log p(x_i, y_j) \tag{4.1-6}$$

The maximum value H(XY) may have is H(X)+H(Y); this fact is used in the definition of MI as given in equation 2.5-2. Figure 8 also illustrates this fact. This shows that H(X)+H(Y) is a candidate for normalizing MI.

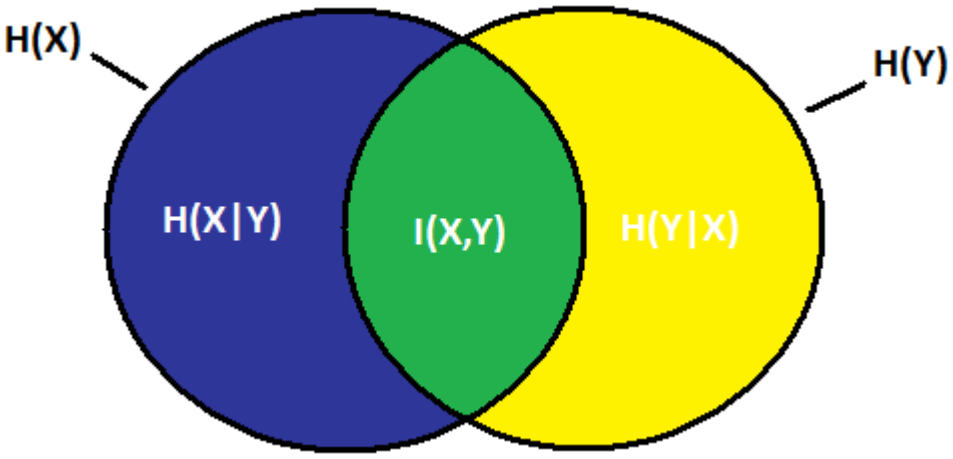


Figure 8. Venn diagram showing the relationship between the uncertainties, conditional uncertainties and mutual information for sets X and Y. H represents entropy.

Normalizing both sides of equation 4.1-5 by H(X)+H(Y) yields:

$$\frac{2*I(X,Y)}{H(X)+H(Y)}=1-\frac{\left[H(X\mid Y)+H(Y\mid X)\right]}{H(X)+H(Y)} \tag{4.1-7}$$

The same analysis used to obtain equation 4.1-7 can be carried out for the MI between two different images, say X and F as follows:

$$I(X,F)=H(X)-H(X\mid F) \tag{4.1-8}$$

$$I(Y,F)=H(Y)-H(Y\mid F) \tag{4.1-9}$$

Figure 9 can be use to help illustrate the above relations. A summation of equations 4.1-8 and 4.1-9 yields a separate equation, namely:

$$I(X,F)+I(Y,F)=H(X)+H(Y)-H(X\mid F)-H(Y\mid F) \tag{4.1-10}$$

Normalizing equation 4.1-10 by H(X) + H(Y) yields:

$$\frac{I(X,F)+I(Y,F)}{H(X)+H(Y)}=1-\frac{\left[H(X\mid F)+H(Y\mid F)\right]}{H(X)+H(Y)} \tag{4.1-11}$$

By taking the ratio of equations 4.1-7 and 4.1-11, the FCIM follows:

$$\begin{aligned}\frac{\dfrac{2*I(X,Y)}{H(X)+H(Y)}}{\dfrac{I(X,F)+I(Y,F)}{H(X)+H(Y)}}&=\frac{2*I(X,Y)}{H(X)+H(Y)}*\frac{H(X)+H(Y)}{I(X,F)+I(Y,F)}\\&=\frac{2*I(X,Y)}{I(X,F)+I(Y,F)}\end{aligned} \tag{4.1-12}$$

The FCIM metric is offered as a new variant of MI eliminating the use of entropy in its performance calculation. FCIM performance is tested using different levels of

white Gaussian noise (WGN) added to the source imagery before fusion. These results are given in chapter 6.

4.2 Variants of MI

There are many variants of MI reported that prove useful with evaluating image fusion performance (Cvejic *et al*., 2006; Nava *et al*., 2007; Piella & Heijmans, 2003; Wang *et al*., 2009). However, Hossny *et al*. (2008) discussed the unsuitability of using classical mutual information as a performance measure for image fusion. Qu *et al*. (2002) presented an image fusion performance measure using mutual information by estimating the joint entropy between two source images x,y and fused image f as follows:

$$M_F^{XY} = I(F, X) + I(F, Y) \tag{4.2-1}$$

where X,Y, and F are the normalized histograms of source images x,y and the fused image f, respectively. Hossny *et al*. reported applying this mutual information measure to image fusion raises a problem concerning the scaling of the metric. Hossny *et al*. argue since $I(X, X) = H(X)$ where $H(X)$ is the entropy of the random variable X, the measures $I(F, X)$ and $I(F, Y)$ are not measured on the same scale. That is, the self information between random variables may be substantially different. Therefore M_F^{XY} mixes two unnormalized quantities which results in a bias toward the source image with the highest entropy. Their proposed solution to resolve the scale imbalance between source information was to use a variant of MI that properly normalizes the measurement. This problem has been the source of the development of several variations of MI to be used as a proper metric. Horibe (1985) proposed a normalized mutual information (NMI) measure as follows:

$$NMI(X,Y) = \frac{I(X,Y)}{\max\{H(X), H(Y)\}} \tag{4.2-2}$$

where $H(X)$ and $H(Y)$ are the entropies of the random variables X and Y, respectively. Kvalseth (1987) was somewhat critical of this normalizing variant, stating the measure in equation 4.2-2 was given without any justification. Kvalseth generalized Horibe's normalized information measure to:

$$NMI(X,Y) = \frac{I(X,Y)}{D_i} \tag{4.2-3}$$

where the denominator D_i is some appropriate normalizing factor. The denominator D_i, however, does not necessarily have a unique solution. Some possible candidates for D_i are given as:

$$D_1 = \min\{H(X), H(Y)\} \tag{4.2-4a}$$

$$D_2 = \max\{H(X), H(Y)\} \tag{4.2-4b}$$

$$D_3 = \frac{H(X) + H(Y)}{2} \tag{4.2-4c}$$

$$D_4 = \sqrt{H(X)H(Y)} \tag{4.2-4d}$$

Kvalseth stated that only D_2 and D_3 are proper normalizing factors and that D_3 is a compromise candidate, as it is an average between D_1 and D_2. Kvalseth further justified the use of D_3 by considering information theory concepts referred to as the coefficients of constraints (Attneave, 1959) defined by:

$$U_{X|Y}^2 = \frac{I(X,Y)}{H(X)} \qquad U_{Y|X}^2 = \frac{I(X,Y)}{H(Y)} \tag{4.2-5}$$

$U^2_{X|Y}$ and $U^2_{Y|X}$ are defined in information theory as one minus the conditional uncertainty of X given Y divided by the uncertainty of X, and one minus the conditional uncertainty of Y given X divided by the uncertainty of Y, respectively. Along with equation 4.1-7, equation 4.2-5 begins to shed some light on why Kvalseth's choice of D_3 as a proper normalizing factor for MI.

It should be noted that the uncertainty in the joint probability matrix between X and Y can be given as:

$$H(X,Y)=H(X)+H(Y\mid X)=H(Y)+H(X\mid Y) \tag{4.2-6}$$

Equation 4.2-6 shows two equivalent ways to obtain the maximum uncertainty between two random variables. In terms of a single random variable, the uncertainty in H(X) and H(Y) can be found by subtracting from H(X,Y) the uncertainty that remains after having knowledge of X and Y respectively. When H(X) and H(Y) are purely redundant, the conditional terms go to zero, resulting in the maximum joint uncertainty equal to H(X) and H(Y) respectively.

To derive the U equations in equation 4.2-5, all that is needed is to normalize equation 4.1-3 and 4.1-4 by their maximum amount of uncertainty. That is, for equation 4.1-3:

$$I(X,Y)=H(X)-H(X\mid Y) \tag{4.2-7}$$

$$\frac{I(X,Y)}{H(X)}=\frac{H(X)-H(X\mid Y)}{H(X)} \tag{4.2-8}$$

$$\frac{I(X,Y)}{H(X)} = 1 - \frac{H(X \mid Y)}{H(X)} = U^2_{X|Y} \tag{4.2-9}$$

In a similar manner, normalization of equation 4.1-4 follows as:

$$\frac{I(X,Y)}{H(Y)} = 1 - \frac{H(Y \mid X)}{H(Y)} = U^2_{Y|X} \tag{4.2-10}$$

It is clear by the definition of MI that the uncertainty in one set can be reduced given knowledge of the other. Equations 4.2-9 and 4.2-10 are asymmetric measures of association between random variables X and Y, and are understood as the proportion of the uncertainty associated with a given set of information that is transmitted from another set.

By taking a weighted average[1] of $U^2_{X|Y}$ and $U^2_{Y|X}$ and using $H(X)$ and $H(Y)$ as the weights, the following symmetric association metric results:

$$U^2 = \frac{2 * I(X,Y)}{H(X) + H(Y)} \tag{4.2-11}$$

which gives $U^2 = \frac{I(X,Y)}{D_3}$, completing Kvalseth's justification for his choice of normalizing factor. Hossny et. al. used the concept of coefficient of constraints (CofC) to modify equation 4.2-1 as:

$$M_F^{XY} = 2 * \left[\frac{I(F,X)}{H(F) + H(X)} + \frac{I(F,Y)}{H(F) + H(Y)} \right] \tag{4.2-12}$$

---

[1]A weighted average is obtained by applying weights to each term to be added, then summing up the quantities. Last, divide the summation by a sum of the weights.

where $H(X), H(Y)$ and $H(F)$ are the entropies of the random variables X, Y and F, respectively. Equation 4.2-12 will be referred to as the Hossny metric for the remainder of the document. Hossny *et al*. went on to show how the error between classical MI and the normalized version increases as the difference between the source image entropies increases, further emphasizing the caution surrounding the use of classical MI as an information measure for image fusion. The use of entropy measures to normalize MI for use as an image fusion metric are frequently reported in the literature.

4.3 Mutual Information using Image Fusion

The concept of reducing the uncertainty in one set given knowledge of another is extended to the case of three random variables. The definition of MI for two source images X and Y and a fused image F follows:

$$I(X,Y,F) = I(X,Y) - I(X,Y \mid F) \tag{4.3-1}$$

Equation 4.3-1 shows that the redundant information shared between the source images and the fused image can be obtained by measuring the redundant information between the source images minus the information that the fused image can explain. This can be readily seen with the help of Venn diagrams, as given in Figure 9.

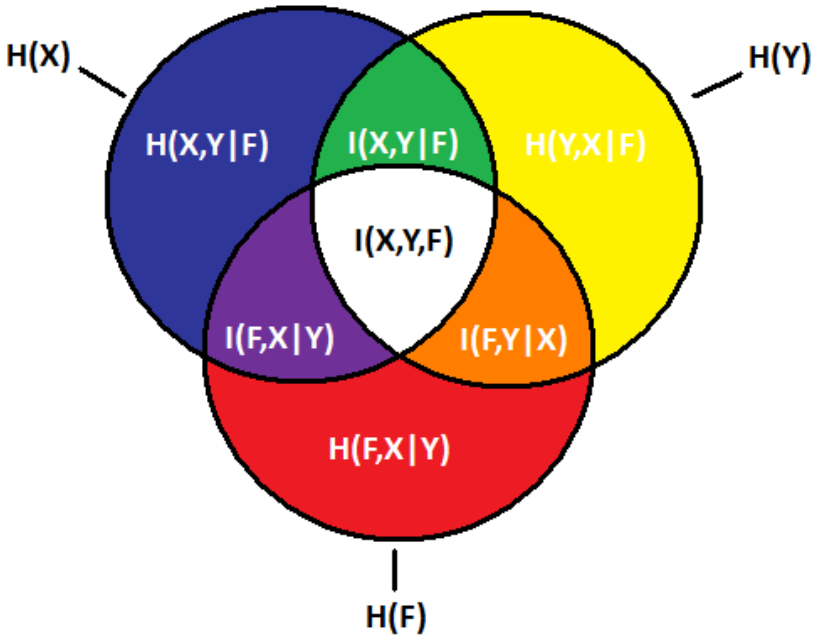


Figure 9. Venn diagram showing the relationship between the uncertainties, conditional uncertainties and mutual information for sets X, Y and F. H represents entropy and I represents mutual information.

To obtain a measure of the information transferred from the source images to the fused image, one needs to measure the shared information between the source images and subtract out the information the fused image explains. Using the definition of the CofC's given in equation 4.2-5, an alternative development of the FCIM measure can be obtained. I sought to develop a metric that tracks the transfer of information between the source and fused images. By definition, the uncertainty in information for source image X is given as H(X), and likewise the uncertainty in information for source image Y is given as H(Y). CofC equations were modified to account for the remaining uncertainty in each of the source images not explained by the fused image, that is:

$$UF^2_{X|Y} = \frac{I(X,Y)}{H(X) - H(X \mid F)} \qquad UF^2_{Y|X} = \frac{I(X,Y)}{H(Y) - H(Y \mid F)} \tag{4.3-2}$$

By subtracting from H(X), the information in X explained by the fused information, H(X|F), what is left is the remaining uncertainty in the information for image X. This process is repeated for image Y. The information in the fused image explains some of the uncertainty in both of the source images (not necessarily the same information in both images). The revisions of equation 4.2-5 yield quantities that satisfy the multivariate constraint property that the amount of information transferred between two sets is equal to the uncertainty of their joint probability matrix minus the actual uncertainty. Each of these measures reduces back to the CofC equations given in equation 4.2-5 when the source images contain exactly redundant information (no complementary information), then. This would indicate the fusion process did not provide any additional information to explain the uncertainty in the source images. Therefore, this measure can be used to evaluate the degree of complementary information transferred to the fused image from the source images.

Recalling the previous definition for MI, $UF^2_{X|Y}$ and $UF^2_{Y|X}$ can be rewritten as:

$$UF^2_{X|Y} = \frac{H(X) - H(X \mid Y)}{H(X) - H(X \mid F)} \qquad UF^2_{Y|X} = \frac{H(Y) - H(Y \mid X)}{H(Y) - H(Y \mid F)} \tag{4.3-3}$$

Using the fact that the MI between the inputs remains constant, a measurement of the reduced uncertainty in the input information as explained by the fused image can be made. $UF^2_{X|Y}$ can be interpreted as the amount of uncertainty in X minus the amount of uncertainty explained by Y, divided by the amount of uncertainty in X minus the amount of uncertainty explained by F. The same can be said of $UF^2_{Y|X}$. In this way, tracking how

much uncertainty the fused image explains about each source image, FCIM can track how much of the information between the source images was actually transferred to the fused image. This process captures each fusion algorithm's ability to reduce the uncertainty in the inputs.

As stated, the denominators in equation 4.3-3 give an indication of how much the fused image explains the uncertainty in information for each of the source bands. They should also be recognized as a measure of information gain, or by remembering the definitions of MI given, one can see the denominators are in fact the MI between each source and fused image separately. Rewriting the equations as:

$$UF^2_{X|Y} = \frac{I(X,Y)}{I(F,X)} \qquad UF^2_{Y|X} = \frac{I(X,Y)}{I(F,Y)} \tag{4.3-4}$$

A weighted average of $UF^2_{X|Y}$ and $UF^2_{Y|X}$ can be obtained and used as a measure of redundant information transferred from the source images to the fused image. Using I(X,F) and I(Y,F) as weights, the weighted average of the two above equations is taken. This procedure results in the proposed information measure referred to as the fused common information measure (FCIM). FCIM serves as another variant of mutual information. For completeness, FCIM is given again below:

$$FCIM = \frac{2 * I(X,Y)}{I(F,X) + I(F,Y)} \tag{4.3-5}$$

Accordingly, using Kvalseth's generalized MI equation:

$$NMI(X,Y) = \frac{I(X,Y)}{D_i} \tag{4.3-6}$$

FCIM is equivalent to Kvalseth's generalized MI equation with $D_i = \frac{I(F,X)+I(F,Y)}{2}$.

4.4 FCIM as an Image Fusion Ranking Metric

FCIM can also be used to rank image fusion algorithm performance. The redundant information between the source images remains constant always; therefore, each fusion algorithm's ability to transfer the source information can be measured in terms of the reduction in the uncertainty of redundant source information as explained by the fused image. Since H(X) is always greater than H(X) – H (X|F), FCIM values are equal to or greater than that of equation 4.2-11, therefore equation 4.2-11should be subtracted from FCIM. The proposed image fusion ranking measure (IFRM) is given as follows:

$$IFRM = FCIM - NMI\ with\ weighted\ average(CofC) \tag{4.4-1}$$

$$IFRM = \frac{2 * I(X,Y)}{(H(X)-H(X \mid F))+(H(Y)-H(Y \mid F))} - \frac{2 * I(X,Y)}{H(X)+H(Y)} \tag{4.4-2}$$

The difference of the CofC and FCIM indicate how complementary the source images are, as a function of each particular fusion algorithm. IFRM is discussed in chapter 8.

4.5 Other Common Metrics

This section describes some of the common image fusion metrics previously referenced in section 2.5. Although the following metric is not useful for image fusion algorithm comparisons by itself, because it only compares two images, it has been used to

develop image fusion metrics; as will be shown later. The image quality index (IQI) (Wang & Bovik, 2002) is used to quantify similarity between two input images (usually when one image is a reference image), and is defined as:

$$Q_0 = \frac{\sigma_{ab}}{\sigma_a \sigma_b} * \frac{2\bar{a}\bar{b}}{\bar{a}^2 + \bar{b}^2} * \frac{2\sigma_a \sigma_b}{{\sigma_a}^2 + {\sigma_b}^2} \tag{4.5-1}$$

where $Q_0(a, b)$ represent a measure of the similarity between input images a and b and has a value between -1 and 1.

The first term of Q_0 measures to what extent input images a and b are linearly correlated. The second term is a measurement of closeness between the mean luminance of images a and b. The third component measures the similarity in contrast between the two images.

The FQI is an extension of the IQI and allows for comparisons between a fused image and two source images. To determine the FQI (Piella & Heijmans, 2003), a comparison is made between the source images and the fused image, identifying which image has the largest amount salient information. This process is accomplished by using the local statistics of a $n \times n$ sliding window as the saliency factor for each image. The FQI is defined as:

$$Q(a, b, f) = \frac{1}{|W|} \sum_{w \in W} (\lambda(w) Q_o(a, f \mid w) + (1 - \lambda(w)) Q_o(b, f \mid w)) \tag{4.5-2}$$

where $\lambda(w)$ indicates the weighting factor that measures which window being compared offers the most relevant information. $\lambda(w)$ is calculated as:

$$\lambda(w) = \frac{s(a \mid w)}{s(a \mid w) + s(b \mid w)} \tag{4.5-3}$$

where $\lambda(w)$ has a value between 0 and 1. s(a|w) and s(b|w) represent the saliency in each window of image a and b respectively.

FQI gives an indication of how much of the salient information has been transferred in to the fused image from both of the input images. The fusion quality index has values between -1 and 1.

The WFQI is similar to the FQI, however, whereas the FQI treats the weighting factors within the sub-images equally, the WFQI assigns more weight to those regions with higher saliency in the input images. This weighting scheme is intended to model the human visual system, which is known to give higher importance to visually salient regions in an image. The saliency of the windows for the WFQI is defined as:

$$C(w) = \max(s(a \mid w), s(b \mid w)) \tag{4.5-4}$$

and the WFQI is defined as:

$$Q_w(a,b,f) = \frac{1}{\sum_{w' \in W'} C(W')} \sum_{w \in W} C(W)(\lambda(w) Q_o(a, f \mid w) + (1 - \lambda(w)) Q_o(b, f \mid w)) \tag{4.5-5}$$

and takes on values between -1 and 1.

The EDFQI uses the WFQI and adds to it by taking into account the edge information present in an image. EDFQI returns favorable results for those algorithms that attempt to model the importance of edges to the HVS. EDFQI is defined as:

$$Q_E(a,b,f) = Q_W(a,b,f)Q_{W'}(a',b',f')^{\alpha} \tag{4.5-6}$$

a`, b`, f` are the edge images corresponding to the source images a, b, f. α is a parameter used to weight the edge information versus the image information. The suggested value of 1 for α was used in this study (Piella & Heijmans, 2003). EDFQI also has values between -1 and 1.

5 Metric Evaluation

The FCIM metric was tested under various levels of noise and the results are given in section 6.1.2.1. The evaluation of information transfer using image fusion needs to be robust in response to such degradations. Additionally, to gain more insight into FCIM's performance, the FCIM was used in a perception experiment where controlled amounts of complementary information were presented to image fusion algorithms and observer identification performance was evaluated. The results of this investigation are given in section 5.2.

5.1 Filter design

The source imagery used in the perception experimented presented in section 5.2 was created using the 2-D filters shown in Figure 8. Figure 10 shows a one dimensional (1-D) representation of the filters. These filters were designed as finite impulse response (FIR) filters and were implemented using the Parks-McClellan algorithm in Matlab. The filters are equiripple linear phase FIR filters, and are optimum in the sense that the maximum error between the desired and actual frequency responses is minimized.

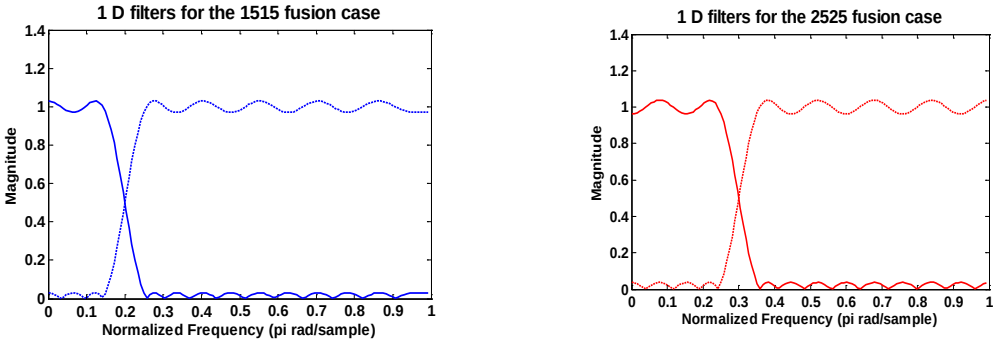


Figure 10. Linear phase FIR equiripple filters used to create imagery with various amounts of complementary information. Left figure shows two filters with cutoff and cuton frequencies of .15. Right figure shows two filters with cutoff and cuton frequencies of .25. Cutoffs and cutons are normalized frequencies defined between 0 and 1, where 1 is the half sample frequency. The transition bands are of length .1. Frequencies are given in pi*rad/sample.

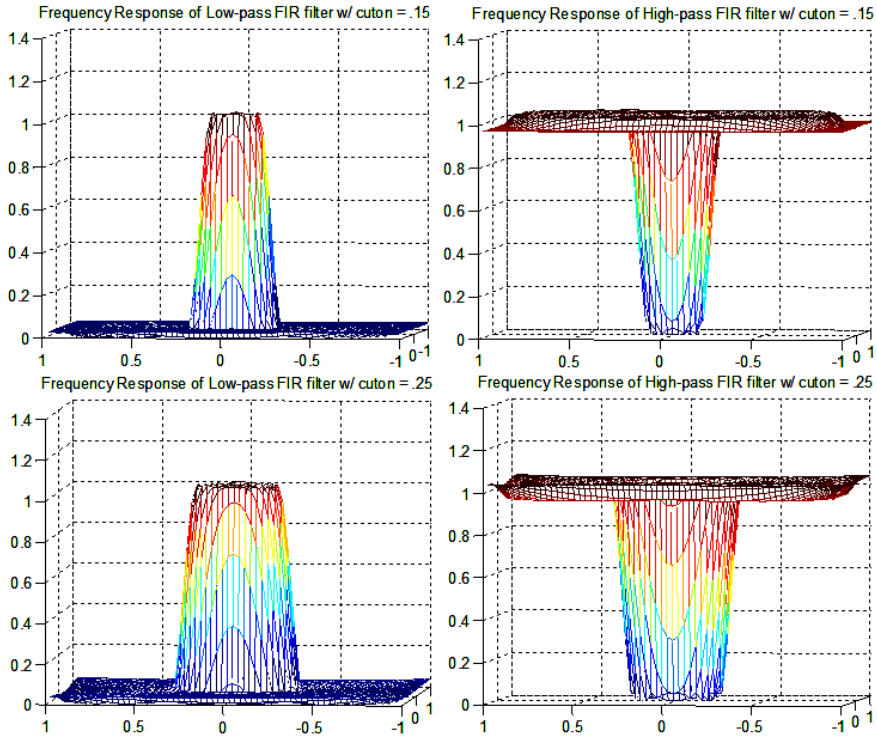


Figure 11. 2-D versions of the 1-D linear phase FIR equiripple filters shown in Figure 10.

Information content in an image, as defined in this research, can be controlled through filtering. Separating an image into various low and high frequency images creates a pair of images that contains different amounts of complementary and redundant spatial frequency information. Even when two images contain exactly complementary spatial frequency information, human performance results have shown that enough perceivable information exists in both bands to perform an identification visual task (Driggers *et al*., 2000). Results such as those in Howell *et al*. (2010) suggest that a measure designed to use image information as a means to predict visual task performance should take into account both the complementary and the redundant information present in an image.

5.2 Complementary and Redundant Information

The evaluation of fusion quality in response to various amounts of complementary information at the input can be simplified somewhat by using filtered imagery as input into fusion algorithms. Generating source imagery from a single image by filtering eliminates any errors due to image registration, thereby eliminating the third performance criterion given at the start of chapter 4. With control of the source information content, image fusion algorithms can be systematically evaluated based on the transfer of salient information between the source images and the fused, and on whether a fusion algorithm introduces any artifacts that affect the task performance. Human perception experiments and the FCIM metric can be used to perform these evaluations.

5.2.1 Human Perception Experiment under Controlled Conditions

Human performance results from an initial investigation to determine if there exists a correlation between the complementary information within source imagery and fusion algorithm performance under controlled conditions are reported in this section.

An eight alternative forced choice (8AFC) ID experiment was conducted in which nine observers participated. The eight targets used in this study are referred to as the 2S3, BMP, M109, M113, M2, M60, T62, and T72 targets. A single source, low noise super resolved long wave infrared (LWIR) image of each target was spectrally filtered into two parts: low spatial frequency and high spatial frequency images. The filters designed in section 5.1 were used in this process. Figures 12 and 13 illustrate the filtering process to obtain the fused test imagery. The high passed (HP) and low passed (LP) filtered images were used to represent the source imagery. The naming convention used to identify the information content in the fused images is the “1515 fusion” and the “2525 fusion” cases. “1515” and “2525” represent the cutoff and cuton frequency for each source image to be fused. For example, the fusion of a LP image with cutoff of .25 pi*rad/sample and a HP cuton of .25 pi*rad/sample will be referred to as a “2525 fusion” case; likewise for the “1515 fusion” case. The source images used in the experiment were fused using pixel averaging, SIDWT and Laplacian fusion. A sample of the original imagery at each range is shown in Figure 14. Only the fused and original images were included in the ID experiment.

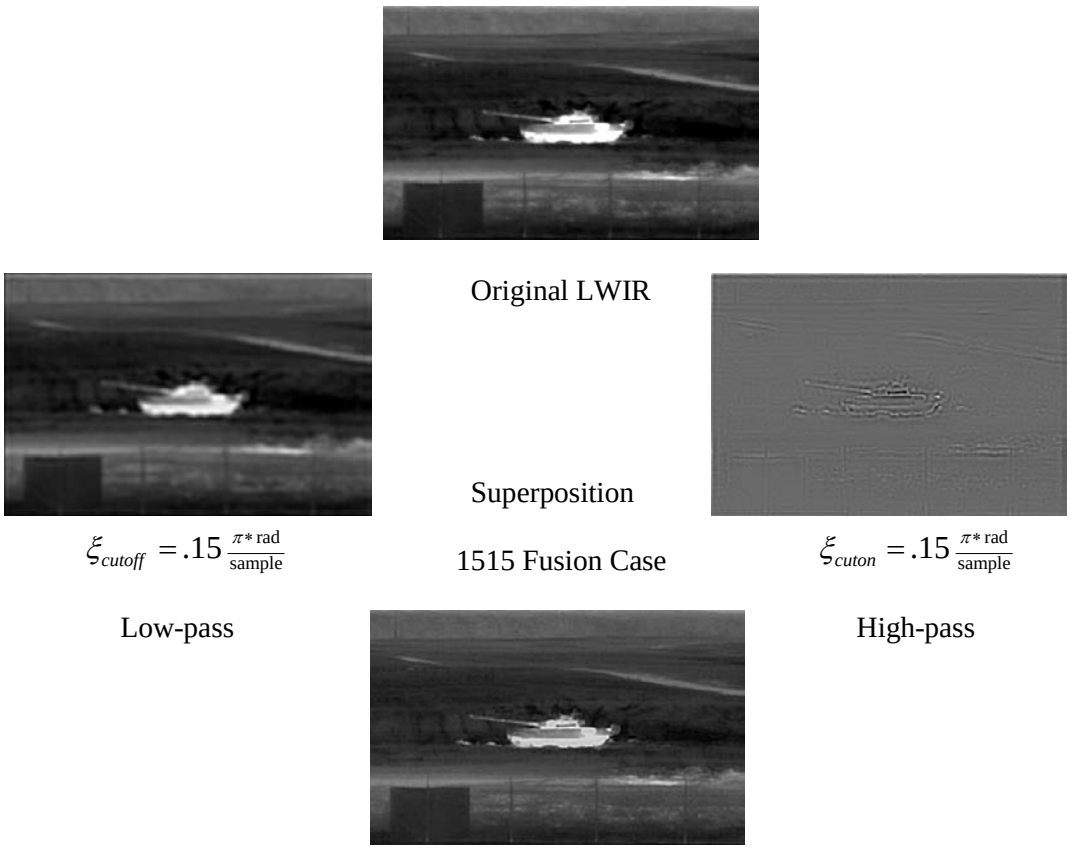


Figure 12. Filtering process to create source band imagery for controlled experiment: one with a low-pass cutoff frequency of .15 pi rad/sample and the other band using a high-pass cuton frequency of .15 pi rad/ sample. The fused imagery shown was created using the superposition algorithm with equal weights of .5.

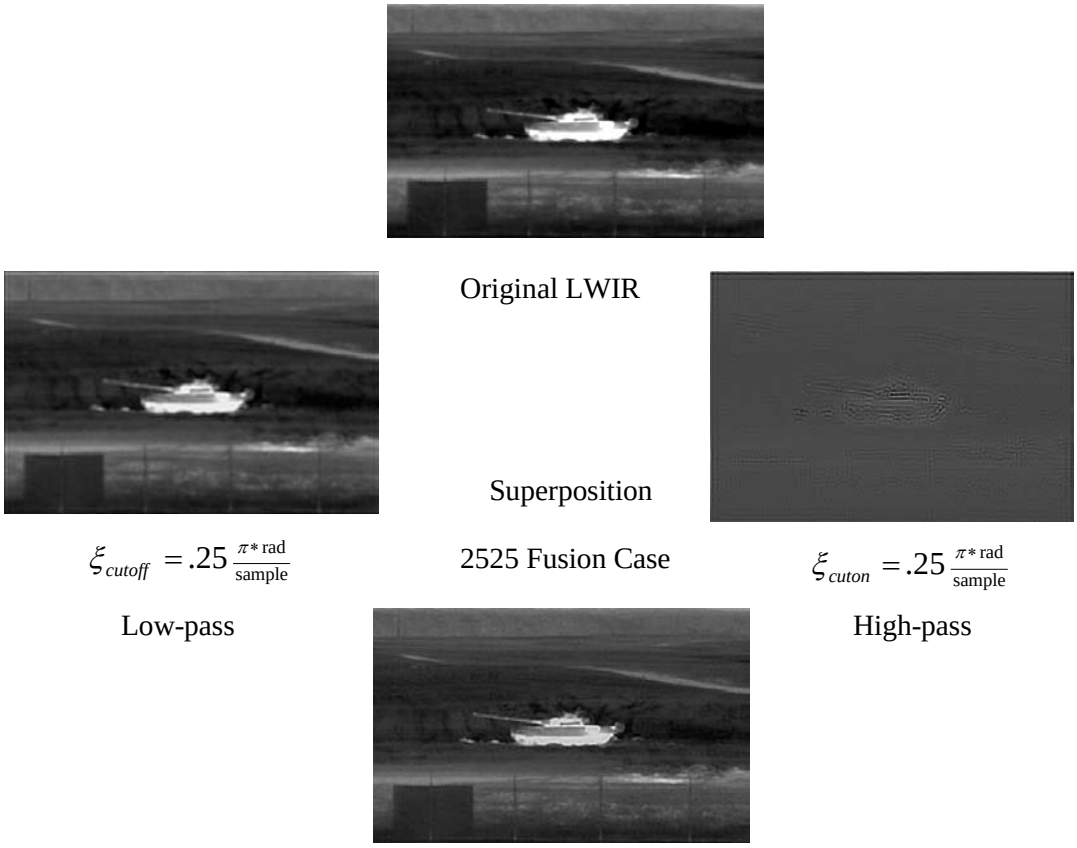


Figure 13. Filtering process to create source band imagery for controlled experiment: one with a low-pass cutoff frequency of .25 pi rad/sample and the other band using a high-pass cuton frequency of .25 pi rad/ sample. The fused imagery shown was created using the superposition algorithm with equal weights of .5.

The observers were shown the original and fused LWIR stimuli of each target at three different aspects and four different ranges. The ranges used in the experiment were 200 meters, 500 meters, 800 meters, and 1500 meters. All observers were trained to identify the standard target set with 95% accuracy using the U.S. Army's Recognition of Combat Vehicles (ROCV) training software. The observers were shown each target in random order and were asked to identify which of the eight targets was presented in the scene; their ability to identify the targets correctly was recorded. A total of the correct and incorrect choices were recorded and the results are plotted in Figure 17.

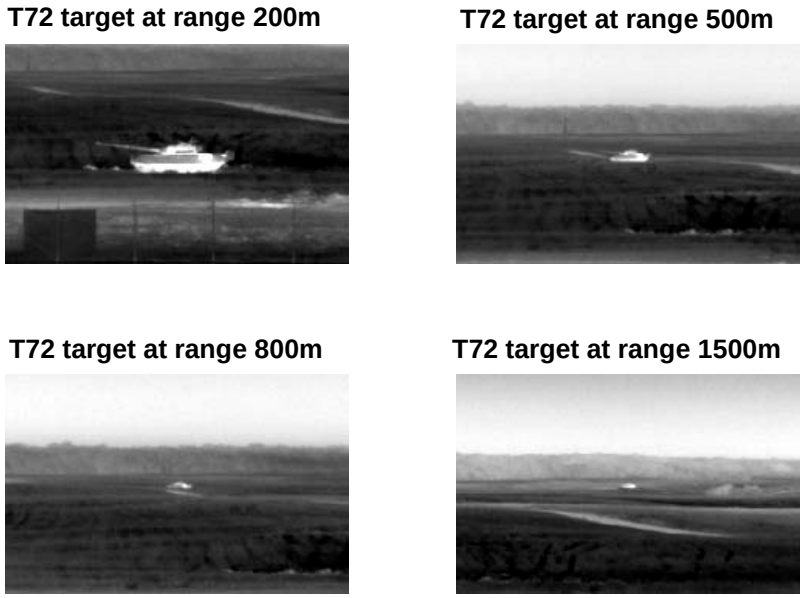


Figure 14. A subset of the experiment test imagery: original imagery of the T72 target at each test range for only one aspect.

Due to the number of images and the high redundancy between the test images, learning effects had to be taken into account. To mitigate inner-test learning, approximately half the test group began with the fused imagery generated with the 1515 fusion case and the rest began with the 2525 fusion case. Given the nature of this experiment, it was expected the performance between the fused and original images is approximately the same.

Initially, the superposition fusion algorithm was used simply because it was simple and was less likely to introduce uncertainties with regards to the analysis of the results. Figure 15 shows the probability of identification (PID) performance results versus range using the original and superposition fused imagery. Although the error bars for the data set was large, the observers mean performances show sensitivity to the

change in complementary information being fused. This conclusion is supported by the increase in the curves between the “1515 fusion” and the “2525 fusion” cases.

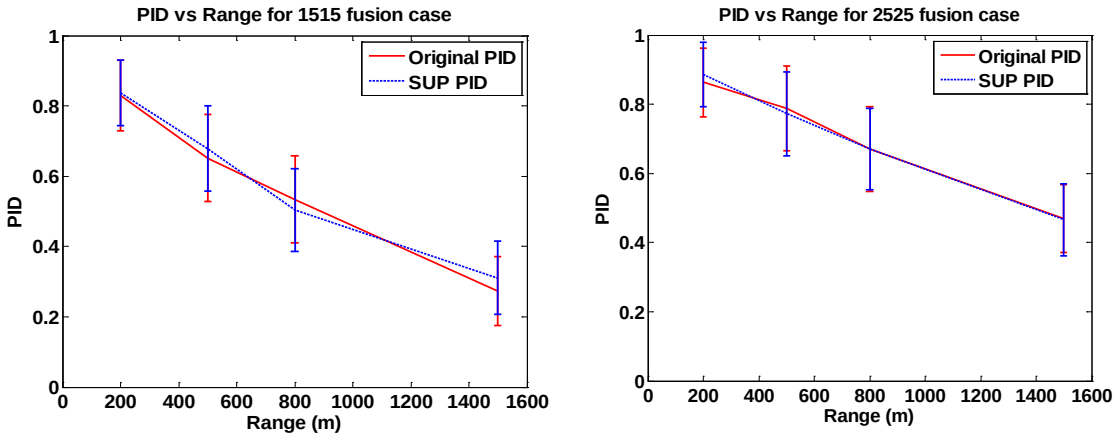


Figure 15. Averaged target performance results from a controlled experiment using superposition fused filtered imagery and the target performance using the original imagery for comparison.

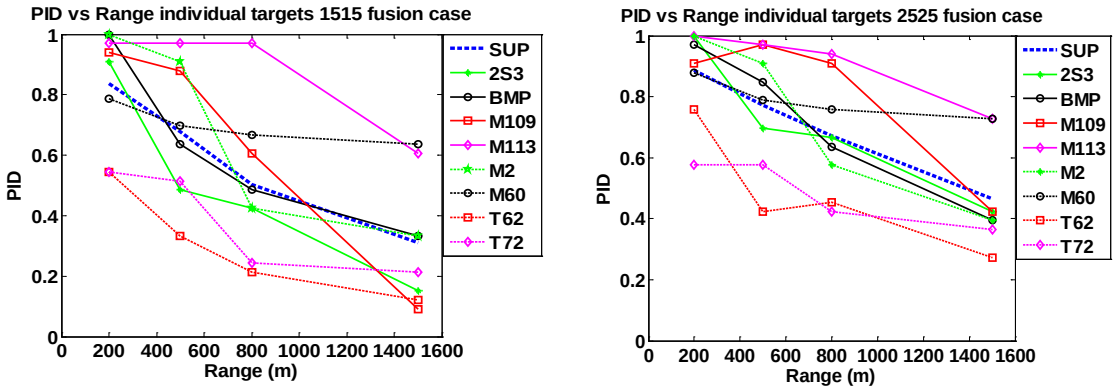


Figure 16. Individual target results from a controlled experiment using the imagery from the 1515 and the 2525 fusion cases.

Figure 16 is given to show how each individual target contributes to the mean performance. Using Figure 16 as a reference, it seems at closer ranges the PID performance using the T62 and T72 targets reduced the mean performance and, as the

range increased, PID performance using the M113 and M60 targets helped maintain a higher mean performance. This type of analysis may lend itself useful for identifying optimal target sets for objective evaluation measures.

The above experimental analysis was extended to include the performance results using the additional image fusion algorithms. The results using all the fusion algorithms are given in Figure 17. Immediately, one sees that the mean performances for all of the image fusion algorithms are approximately equal. And although there is not much separation between the performance curves shown, the Laplacian fusion performance shows a tendency to be the lower performer. As the complementary information increased, the HP source image contributes increasingly less to the fusion process. This means the similarity between the LP image and the original image is increasing, indicating the HP image had less effect on the fused image. An image performance measure using MI should capture this relationship; consequently, the FCIM metric should yield a lower value in these cases.

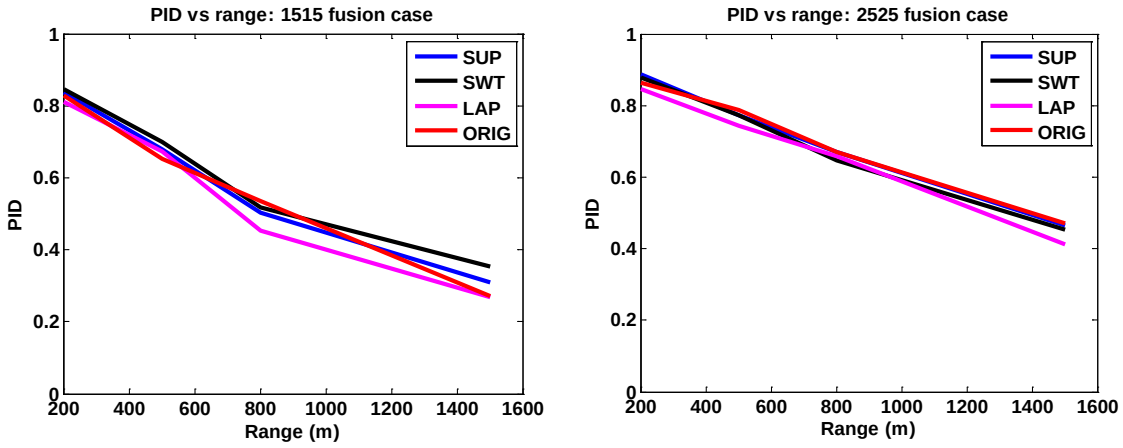


Figure 17. Averaged target performance results from a controlled experiment using fused filtered imagery and the target performance using the original imagery for comparison.

This approach has a couple of benefits that need mentioning. First, controlling the source information content creates an environment for which the sensitivity of the FCIM metric can be systematically evaluated; and second, factors related to errors in image registration can be avoided under these test conditions. This process allows for a single variable to be adjusted while monitoring the performance of the FCIM. In this way, a clearer understanding regarding the relationship between complementary information and fusion algorithm performance is achieved.

This type of evaluation served to establish the relationship between complementary information and task performance using the FCIM metric when evaluating fused imagery. The results of this study show a correlation exists between the proposed metric and identification task performance.

5.2.2 Metric Results

Figure 18 shows partial results of the individual target PIDs along with the FCIM evaluation results; Figure 19 shows the results for the Hossny metric; both figures only show the results for the superposition performance. The T72, M109 and 2S3 targets have been selected to use for the analysis; overall results of both metric performances are given in the Appendix I. The target metric scores are plotted as a function of range for ease of comparison with human visual task performance. The PID increased from the 1515 fusion case to the 2525 fusion case, showing the performance effects of varying the amount of complementary information presented to the fusion algorithm. As the cutoff/cuton frequencies increased, the LP image contained a large majority of the total image information and therefore caused the fusion performance to migrate towards the

PID obtained with the original imagery. For this contrived case, one does not expect the fused performances to surpass that of the originals. This contrived approach does however show as the source imagery became more complementary with respect to each other, the human observers performed better. This conclusion is interesting given that each fusion algorithm was presented the same information in both the 1515 and 2525 fusion cases; the difference being the amount of complementary information in each source band changed. The results of the FCIM support the human performance results. That is, most of the FCIM values decreased as a function of range; in addition the FCIM values decreased going from the 1515 fusion case to the 2525 fusion case. This is a desirable characteristic for the FCIM.

The Hossny metric values appear to exhibit to some degree the same trend as FCIM with regards to decreasing values with increasing range. The variation in the Hossny metric for the 1515 fusion case compared to the 2525 fusion appears to change differently than does the FCIM. The FCIM decreases its values with increasing cutoff/cuton frequencies indicating the redundancy between the source images is decreasing; as seen in Figure 19, the Hossny metric increases its values with increasing cutoff/cuton frequencies.

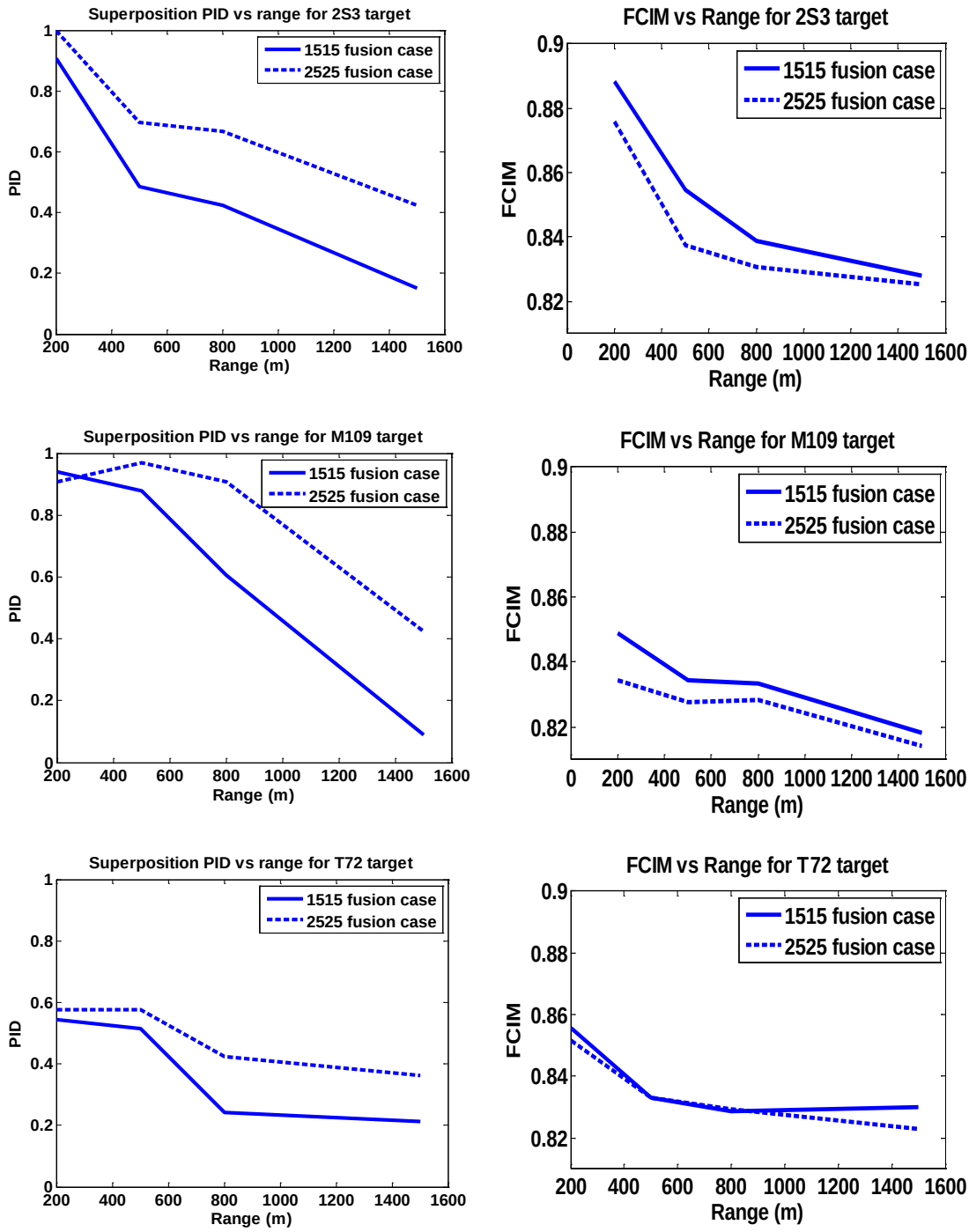


Figure 18. PID performance curves from Figure 13 and FCIM performance for the controlled human perception experiment. The results shown are for the 2S3, M109 and T72 targets using the superposition fusion algorithm.

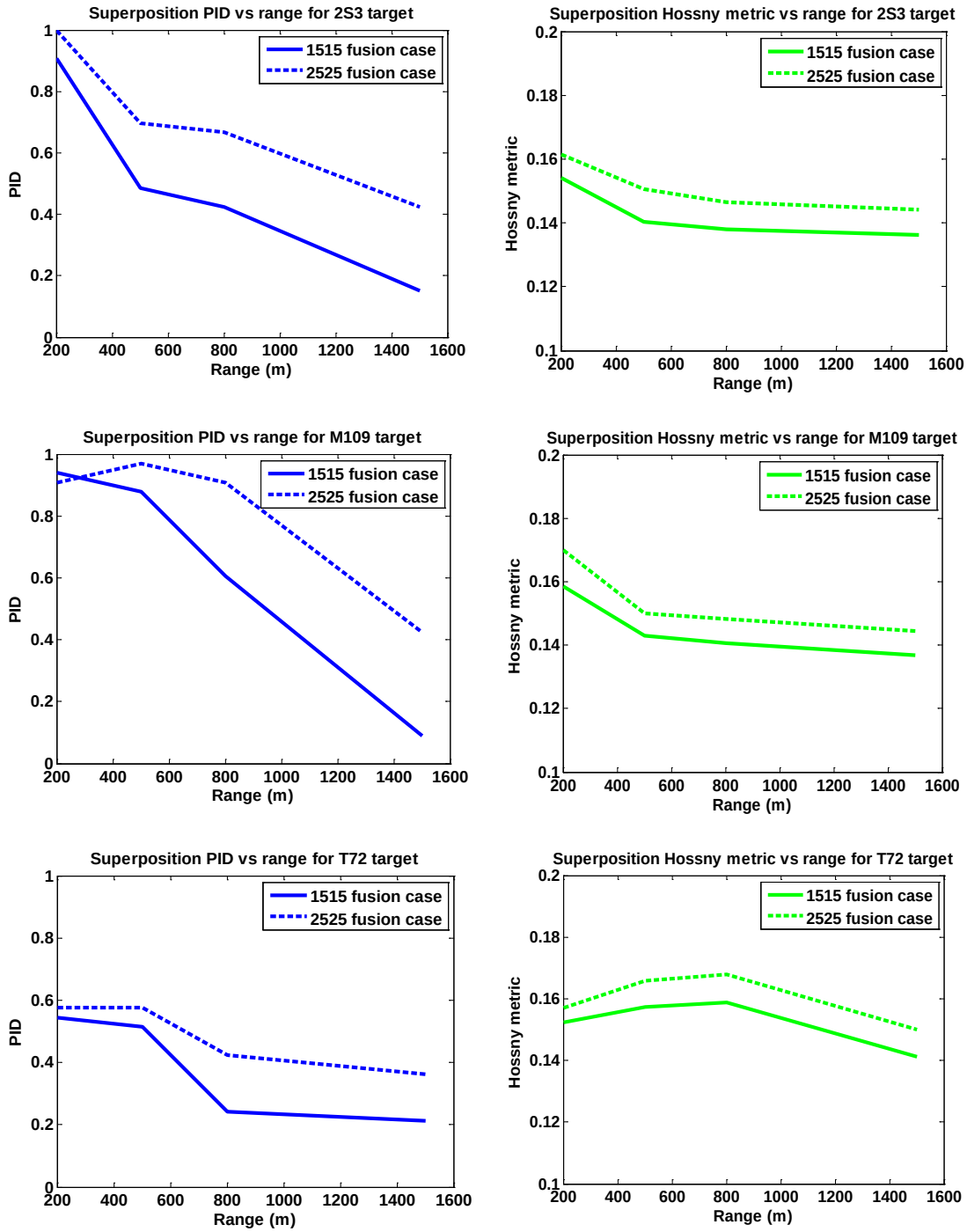


Figure 19. PID performance curves from Figure 13 and the Hossny metric performance for the controlled human perception experiment. The results shown are for the 2S3, M109 and T72 targets using the superposition fusion algorithm.

Pearson's moment correlation coefficient is used to evaluate the relationship between the FCIM, Hossny metric and PID performance. The results of the correlation analysis are given in Table 1. Each cell in Table 1 represents a correlation analysis between the respective targets PID as a function of range and the corresponding metric values for each fusion case. The FCIM shows good correlation with human performance for the 2S3, BMP, M109, M60, and T62 targets for both of the fusion cases. A modest correlation seems to exist for the T72 target. The Hossny metric shows good correlation with the 2S3, BMP, and M2 targets across both fusion cases, but did not show any correlation with the other targets. A point of interest is that the correlation seems to improve for some targets and decrease for others as the complementary information increases (1515 fusion case to the 2525 fusion case), this is true for both metrics evaluated. Another point of interest is that although the Hossny metric did not have good correlation for the individual targets, the overall correlation for the target set as a whole did. This point should raise concerns regarding whether there are other influences causing the correlation to improve (or degrade).

Table 1

Pearson's moment correlation results between the FCIM and human performance for the superposition algorithm performance for each individual target

	FCIM SUP 1515 fusion case Overall = 0.93	FCIM SUP 2525 fusion case Overall = 0.95	Hossny Metric SUP 1515 fusion case Overall = 0.94	Hossny Metric SUP 2525 fusion case Overall = 0.94
2S3	0.98	0.96	0.95	0.96
BMP	0.97	0.99	0.96	0.84
M109	0.91	0.94	0.74	0.46
M113	0.39	-0.05	0.36	0.66
M2	0.44	0.16	0.84	0.91
M60	0.95	0.97	0.59	0.44
T62	0.83	0.93	-0.10	-0.11
T72	0.73	0.82	0.39	0.37

Table 2

Overall correlation values for the average FCIM and Hossny Metric values across the entire data set and human identification performance (PID) for all the algorithms used in the controlled perception experiment

1515 fusion case	FCIM	Hossny Metric
SUP	0.93	0.94
SWT	0.93	0.94
LAP	0.91	0.93

2525 fusion case	FCIM	Hossny Metric
SUP	0.95	0.94
SWT	0.95	0.93
LAP	0.92	0.81

Correlation is a valuable tool when its limitations are understood. Intuitively, we understand correlation as a measure of independence between two variables. The correlation coefficient doesn't indicate an exact functional relationship between two variables; it only gives an indication of the degree of linear relation between the variables. That is, the correlation coefficient expresses the correlation between two variables, telling whether the correlation is positive or negative; it also allows for comparisons between correlations as well as it allows for determining when correlations change.

A common question that arises when dealing with correlation measures is how errors in the data effect the correlation. Errors can increase or decrease the correlation to some value far away from the true value. The scatter plots of Anscombe's quartet can serve as a warning to those who chose to blindly use correlation without examining the data. Figure 20 shows the famous scatter plots of Anscombe's quartet (1973). Each data set has the same mean, standard deviation and correlation value, although not all of the data sets have linear relationships. The functional relationship of the top left data set appears to be normal in nature, while although the top right certainly has an obvious relationship between the data, it doesn't seem to follow an assumption of linearity. Bottom left and bottom right data sets emphasize the need to visually observe the data; the outliers in both of these data sets severely influence the correlation measure. Scatter plots are needed, in addition to correlation, to correctly describe the association between each metric and visual task performance. Refer to Appendix II for a visual illustration of the correlation between PID and the image fusion metrics that resulted in the correlation values given in Table 2.

Another common and important question that arises when using correlation coefficients is how chance affects the correlation measure. One may present the hypothesis: what is the chance, or probability, of getting a correlation of 0.81 or higher when in fact no correlation exists? Hypotheses such as this are easily accepted or rejected with the help of a table of probability levels for the correlation coefficient[1]. For example, using Table 1 in Appendix IV for n = 4 pairs of data, a correlation coefficient of

[1] http://physics.mercer.edu/Younce/pearson.html Table of critical values for correlation coefficients given in Appendix IV.

0.81 has a significance value of 0.05. This means that there is a 95% probability that the correlation measure of 0.81 is significant, therefore the hypothesis can be rejected.

If the concern is with which functional relationship exists between the two variables, then the significance of the correlation measure is important, such as when the FDA says smoking cigarettes greatly increases your chance of getting lung cancer[2]. If a study was conducted that resulted in the findings that smoking cigarettes had a 60% correlation with lung cancer, then 0.6 correlation is a significant measure. If the concern is with forecasting or predicting a relationship between two variables, then their variances are important.

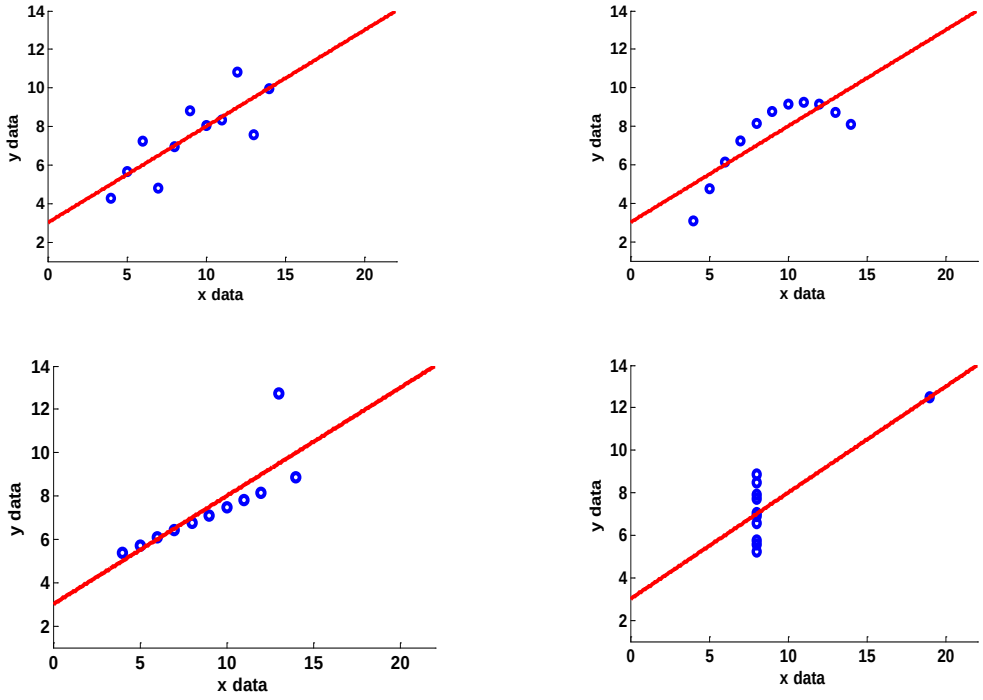


Figure 20. Anscombe's Quartet: a famous data set for illustrating the importance of plotting data in conjunction with correlation measures.

[2] http://www.fda.gov/ForConsumers/ByAudience/ForWomen/ucm118545.htm.

The variation between the two fusion cases for each image fusion metric used in this investigation is given in Figure 21. In the figure, each metric's value for the 1515 fusion case and the 2525 fusion case is plotted against the other. A y = x curve is given in each plot as a reference indicating how far off the metric is from an exact linear correlation when the complementary information changes. If the metric is indifferent to the changes in source information content, then all the points would lie on the y = x curve. Scatter plots far away from the y = x line indicate a large variation in the metric's values between cases and vice versa. Therefore, the distance of the scatter plots relative to the y = x curve suggests the magnitude of variation in the metric scores between the two fusion cases with respect to the input information. The scatter plots in Figure 21 show a large variation in the source information for the Hossny metric. Thus the Hossny metric shows most of the variation between the fusion cases is with the 2525 fusion case. The FCIM shows a smaller variation measured between fusion cases where most of the variation is attributed to the 1515 fusion case (except for the Laplacian fusion algorithm).

Recall, as the cutoff/cuton frequency increases, the HP source image contains less and less edge information, and therefore presents less relative information to the fusion algorithm. This means that more relative complementary information is present in the 1515 fusion case than in the 2525 fusion case. The FCIM supports this conclusion for the superposition and SIDWT fusion algorithms. Also recall, the characteristics of each fusion algorithm performance when presented exactly complementary source information (Figure 7).

In addition, it has already been shown in Figure 5 that noise can influence entropy measurements. It is plausible to assume that the larger variation in the Hossny metric for

the 2525 fusion case, and possibly the FCIM Laplacian 2525 fusion case, is a result of this phenomenon.

When evaluating the correlations given in Table 2 the reader should also remember that the information in the HP band cuton at .25 pi*rad/sample is mostly faint edges to an extent, and therefore contributes very little to the fused resultant image. The FCIM seeks to evaluate the common information content of the source imagery, with one minus the FCIM giving a measure of the complementary information content. The FCIM demonstrates that it is a likely candidate for evaluating imagery that is highly complementary. It is expected, when both source bands posses their maximum relevant complementary information, the FCIM will render its minimum value, indicating the presence of high complementary information in the fused image.

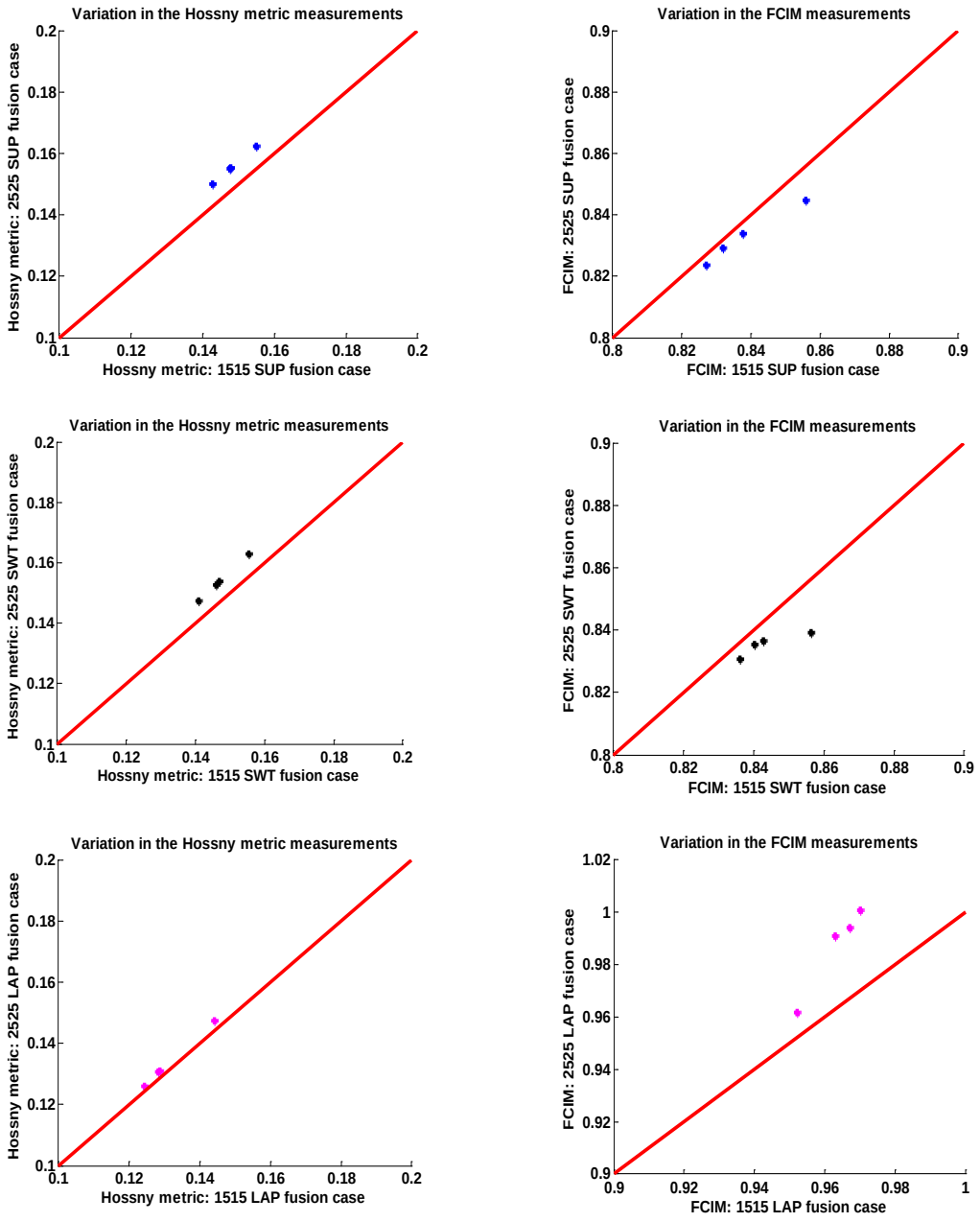


Figure 21. Sensitivity analysis for the FCIM and Hossny metrics evaluating the fusion algorithms used in the controlled human perception experiment. Hossny metric gives more weight to the 2525 fusion case where FCIM gives more weight to the 1515 fusion case (except for the LAP fusion algorithm).

6 Validation of FCIM

The most direct method to evaluate human performance of a visual task such as target identification is to conduct perception experiments. As stated previously, perception experiments are time consuming and very expensive. Therefore, alternatives are sought that reduce the time and financial burdens perception tests create. For this purpose, objective measures that correlate with human visual task performance are needed. Human perception experiments serve to assist with both the development and the validation of objective measures.

The visual task investigated in this research is the ID task, which usually requires an observer to make a decision concerning the target being presented before he/she is allowed to continue testing. This type of experiment is referred to as a forced choice experiment. An N alternative forced choice (NAFC) experiment is conducted to determine how well human performance compares to that of the FCIM using LWIR and SWIR fused imagery. Here, N represents the number of distinct targets to be identified. For the LWIR/SWIR fused perception experiment, N equals 8, because there is a total of eight different targets to be identified. A separate perception experiment evaluating how well humans identified hand held targets is also presented in this chapter; for this experiment N equals 12.

6.1 Human Perception Experiments: tracked vehicles

An 8AFC experiment was conducted using a standard set of tracked military vehicles. The vehicle set was the same vehicle set used in the controlled perception experiment presented in chapter 5. The target set for this experiment was captured in the LWIR and SWIR at four different ranges: 300 meters, 500 meters, 800 meters, and 1500

meters. The source imagery was fused using the superposition, Laplacian and the SIDWT image fusion algorithms. Sample test imagery at a range of 300 meters is shown in Figure 22. Each observer was presented a target from either the source band imagery (LWIR or SWIR) or the fused imagery and was asked to identify which of the N targets was being displayed. Human performance results from this experiment are given in Figure 20.

Figure 22. Sample imagery from the LWIR/SWIR 8AFC experiment. Top left is the LWIR image and top right is SWIR image. From left to right, the bottom row shows the superposition, SIDWT and Laplacian fused images.

6.1.1 Human Performance Results

The performance curves shown in Figure 23 are plotted as a function of range. It should be mentioned that a total of three observers participated in this experiment; as a result, the error bars are large. For de-cluttering purposes, the error bars are not shown in the figure.

The PID performances for the source and fused imagery suggest that it may be possible to obtain human performance using image fusion that is greater than the PID performances using the source imagery. This seems to be the case for the SIDWT performance results up until range 1500 meters where the LWIR performance appears to be better.

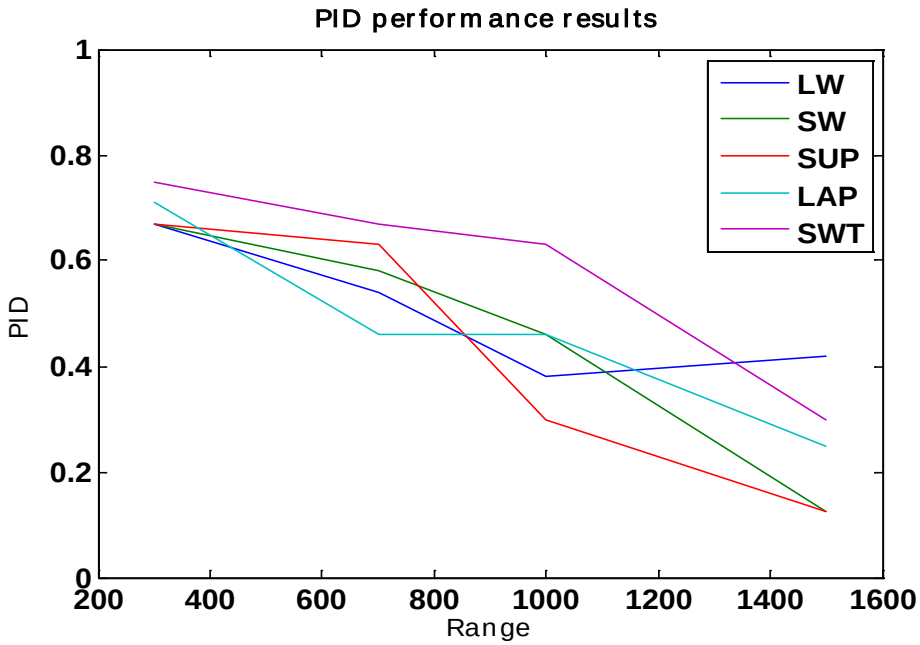


Figure 23. PID performance results for the 8AFC perception experiment using LWIR, SWIR, superposition, Laplacian and SIDWT fused images. The average error (not shown) for the curves was +/- 18%.

The definition of image fusion suggests there are instances where the fused image contains more relevant information than does any source image. However, there is much speculation surrounding the ability to obtain performance increases using image fusion. Identifying an image fusion algorithm that correlates with human performance is a key initial step in image fusion algorithm performance discrimination. Once a metric is

identified that correlates with human performance using image fusion, one would then like to develop a metric that can discriminate between collections of algorithms, selecting a proportion of the algorithms that outperform the rest.

6.1.2 Metric Results

The FCIM and Hossny metric were applied to the test imagery used in the LWIR/SWIR 8AFC perception experiment. The results of this analysis for each image fusion algorithm are given in Figure 24. Table 3 gives the correlation values for both the FCIM and the Hossny metric values plotted in Figure 24 and the PID curves in Figure 23. Comparing Table 2 and Table 3, one immediately sees the correlations are reduced for both metrics when using actual separate spectral bands as inputs to the fusion algorithms. However, the FCIM still has good correlation with PID; the Hossny metric doesn't. The Hossny metric uses entropy as a normalizing factor in its calculations. Entropy was shown in chapter 2 to be an undesirable measure when trying to monitor perceivable visual information.

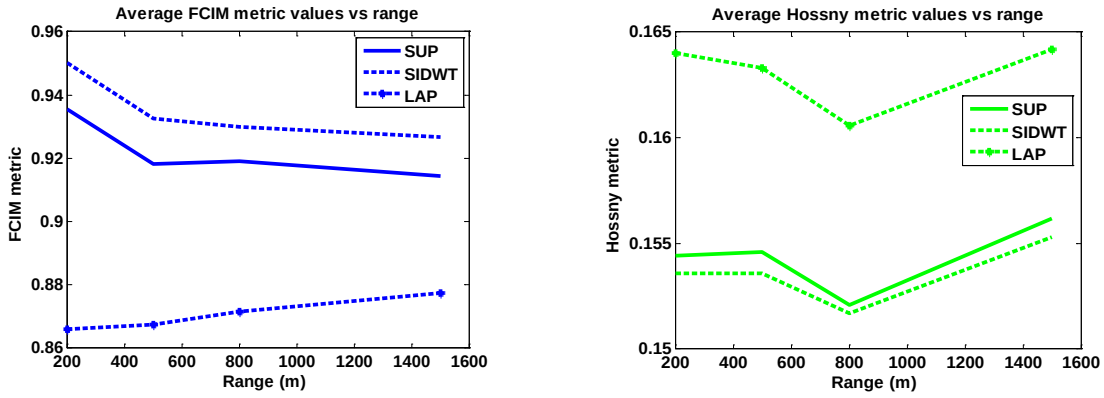


Figure 24. Metric results plotted as a function of range. Both sets of plots shown are averaged across each range for the entire image set.

Table 3

Pearson's moment correlation coefficients between human identification performance given in Figure 23 and both the FCIM and Hossny metric values given in Figure 24 for each of the three fusion algorithms

	Correlation PID versus FCIM	Correlation PID versus Hossny Metric
SUP	0.75	-0.15
LAP	-0.89	0.03
SWT	0.77	-0.62

The FCIM only uses MI in its evaluations of information content in the fused images. The FCIM shows it has good correlation with human performance using fused contrived source imagery and real source imagery. The scatter plots given in Appendix III support the correlations. A third experiment was conducted using hand held targets and different spectral bands to evaluate the FCIM under different conditions. The FCIM shows good correlation in this experiment; section 6.2 gives the results.

6.1.3 Noise analysis

The impact noise may have on the FCIM and Hossny metric is investigated using the imagery from the LWIR/SWIR 8AFC perception experiment. Noise was added to a single image, where the signal-to-noise ratio (SNR) in the image is controlled by applying a ratio of the power spectrum of the signal to the power spectrum of the noise. The signal in this case is to be understood as each individual source image; the noise is an

image of white Gaussian noise (WGN). To fix the SNR within each image, the following procedure was followed:

1. Create an image of WGN
2. Obtain the power spectrum of the noise (Pn), LWIR (Pl) and SWIR (Ps) images
3. Take the ratio of Pl/Pn and Ps/Pn, and divide each ratio by the desired SNR
4. Multiply the square root of the values in step 3 with the noise image
5. Finally, add the noise to each respective image

Figures 25 a, b, and c show sample imagery as a result of fusing the noisy LWIR and SWIR input imagery in this evaluation. In each figure, source images containing the noise are given on top and the fused images for each algorithm are on the bottom. Table 4 gives the evaluation results for both the FCIM and the Hossny metric for the fused images with and without noise applied to the source images.

Figure 25a. Fused LWIR and SWIR imagery using three different image fusion algorithms with a SNR of 40 dB imposed on the imagery. SUP-superposition, LAP-Laplacian and SWT- Shift-invariant DWT

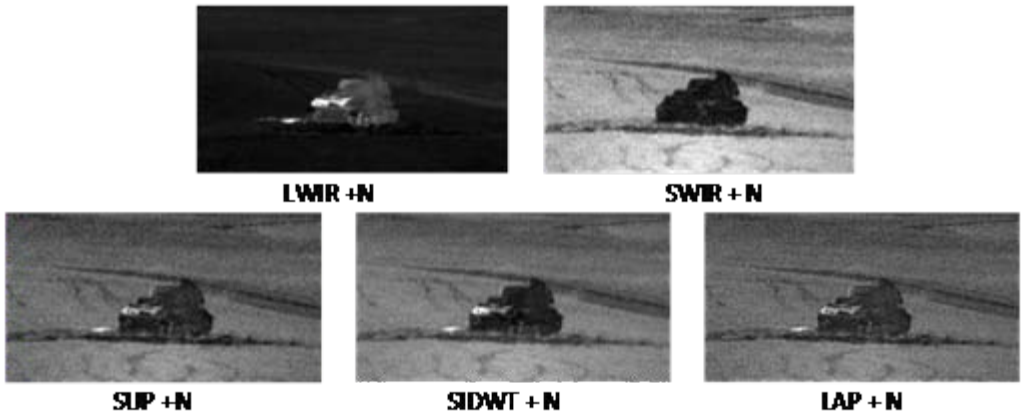


Figure 25b. Fused LWIR and SWIR imagery using three different image fusion algorithms with a SNR of 15 dB imposed on the imagery. SUP-superposition, LAP-Laplacian and SWT- Shift-invariant DWT.

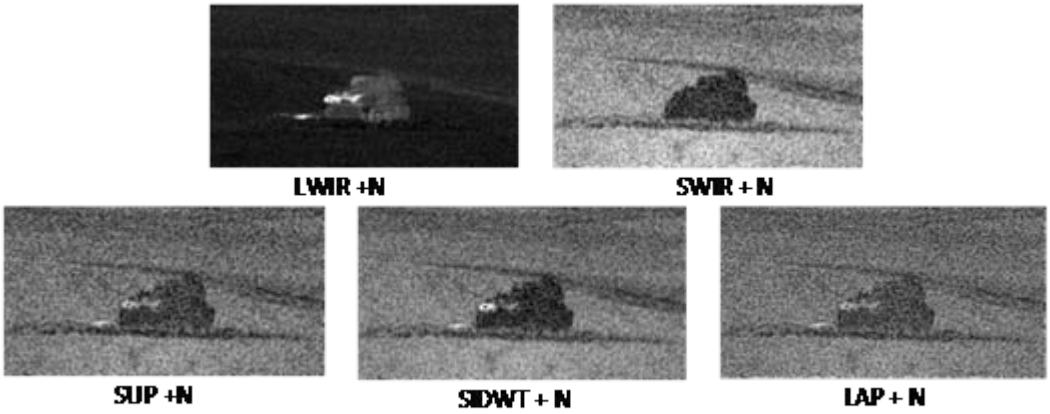


Figure 25c. Fused LWIR and SWIR imagery using three different image fusion algorithms with a SNR of 1 dB imposed on the imagery. SUP-superposition, LAP-Laplacian and SWT- Shift-invariant DWT.

The first column of Table 4 shows the FCIM and Hossny metric results after evaluating the fusion of the source imagery without any noise applied. The following columns show the results for both metrics in response to noise being added to each source image before fusing. Applying various amounts of increasing noise can cause the source images to become redundant. As the information within the source images become more redundant, the FCIM (equation 4.3-5) increases its value with increasing noise. The Hossny metric (equation 4.2-12) values are decreasing while the input sources are becoming more redundant. This is because the two terms in the numerator are becoming more and more similar, while the entropy in the denominator is increasing due to the added noise.

Table 4

Metric evaluation using controlled amounts of noise added to source images before image fusion is carried out

	Original (no noise)	SNR = 40 dB	SNR = 15 dB	SNR = 1 dB
FCIM				
SUP	0.91	0.91	0.91	0.90
SI-DWT	0.97	0.97	0.97	0.95
LAP	0.79	0.78	0.77	0.74
Hossny Metric				
SUP	0.17	0.17	0.17	0.15
SI-DWT	0.17	0.16	0.15	0.14
LAP	0.20	0.19	0.19	0.18

The FCIM and Hossny metric maintained their rankings (in terms of highest to lowest values) as the noise increased for each image fusion algorithm. Both metrics demonstrated a robustness to noise for the superposition algorithm. However, the FCIM values remained constant for the superposition and the SIDWT fusion algorithms up until a SNR of 1. This demonstrates a degree of robustness for the FCIM in its measure of the relevant redundant information; that is, the FCIM values don't react to the added noise appreciably until a SNR of 1 was achieved. The SIDWT and Laplacian fusion algorithms caused a reduction in the Hossny measurement for each level of noise added to the inputs. The results of Table 4 suggest that the FCIM measures similar relevant information

between the inputs and the fused image up until a SNR of 1. The Hossny metric measures differences in information content for the fused images for all noise levels with the SIDWT and Laplacian fusion algorithms. Although human performance results have not been evaluated using the imagery in Figure 25, looking at the images suggests that the quality of the images does not decline drastically until a SNR of 1 is achieved. This is what the FCIM values in Table 4 suggest; however a perception study should be conducted to validate this claim.

6.2 Human Perception Experiments: hand held targets

A 12AFC experiment was conducted using a set of handheld targets. The perception test used LWIR and image intensified (I^2) images of the small handheld objects. There were a total of twelve different objects with each object shown placed in a single hand. The twelve handheld objects displayed in this test were: a gun, hand grenade, rock, cell phone, PDA, camera, flashlight, soda can, coffee mug, knife, brick, and a radio. The 'gun' image set at 7 meters is displayed in Figure 26 as an example of the test images used in the perception test. PID performance was recorded for imagery of the handheld objects captured at ranges of 7m, 10m, and 15m. There was only one aspect (placed in a single hand) of each object shown at the three different ranges.

Ten different fusion algorithms, in addition to the source images, were used in this experiment. The images were shown in random order for each of the three ranges. The trials were divided into three cells to guard against fatigue that could affect the results in an unpredictable manner.

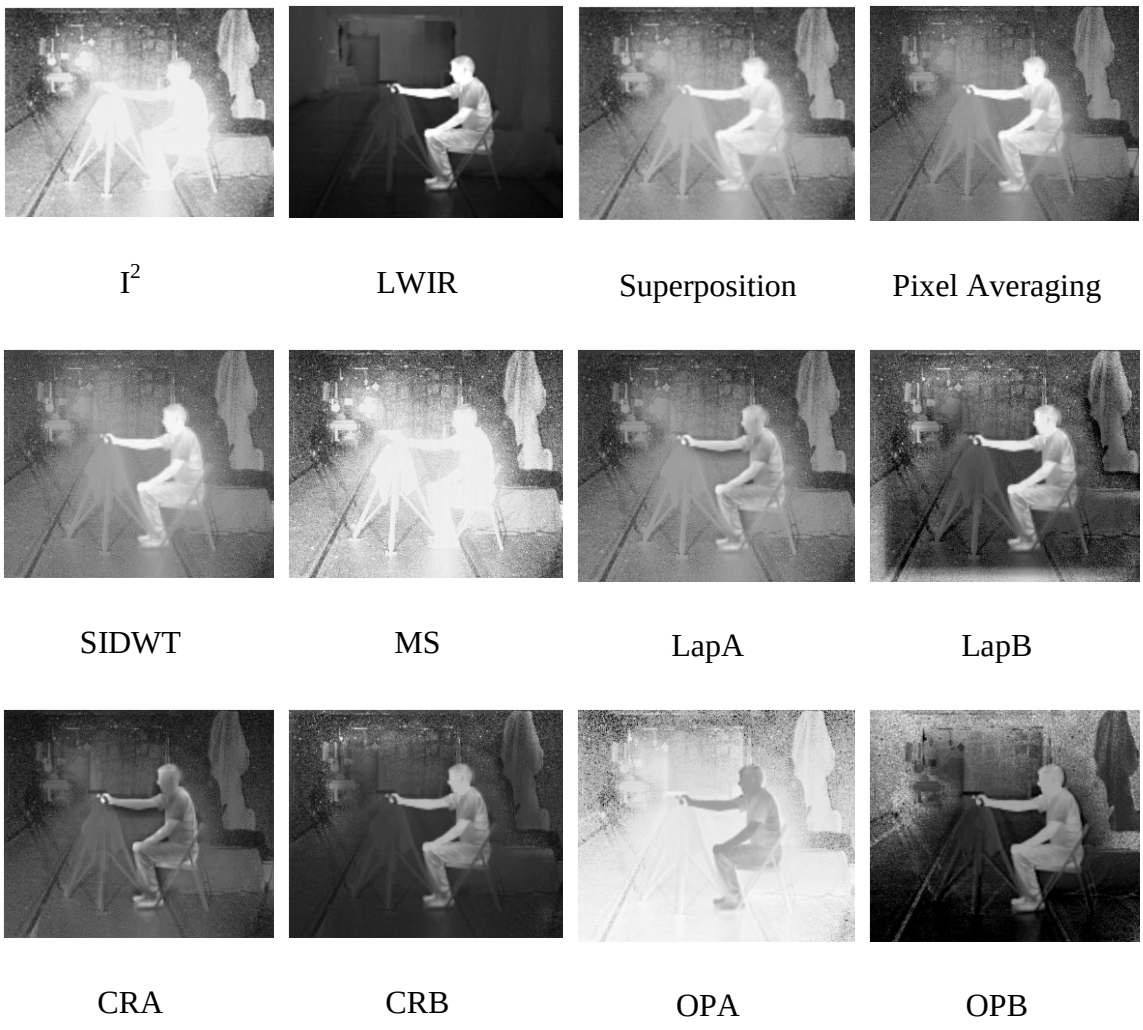


Figure 26. Fused imagery of the "Gun" image set at 7 meters. From top going left to right, LWIR and I^2 are the two source images; superposition and pixel averaging, shift-invariant DWT (SIDWT), multi-scale (MS) decomposition, Laplacian (LAP) fusion: two versions A and B, Contrast Ratio (CR) fusion: two versions A and B and opponent processing (OP): two versions A and B.

Eight observers were trained to identify the 12 different handheld items. A training set of the images was developed and shown to the observers in a PowerPoint presentation to allow time to familiarize themselves with the test images. Once the observers felt they knew the targets in the training set well, a test had to be taken. Each observer was required to obtain 95 percent accuracy with identification on the training set

before being allowed to participate in the actual experiment. The time required for training varied depending upon the individual trainee.

The observers were shown fused images of the handheld objects in random order for a particular range and then asked to identify each object being displayed. The observers' correct responses were annotated for each of the three ranges and the results are given in section 6.2.1.

6.2.1 Human Performance Results

Figure 27 shows the PID versus range results from testing the identification of hand held objects. The curves shown represent the ID performances corresponding to the ten different fusion algorithms; the individual source band performances are not shown.

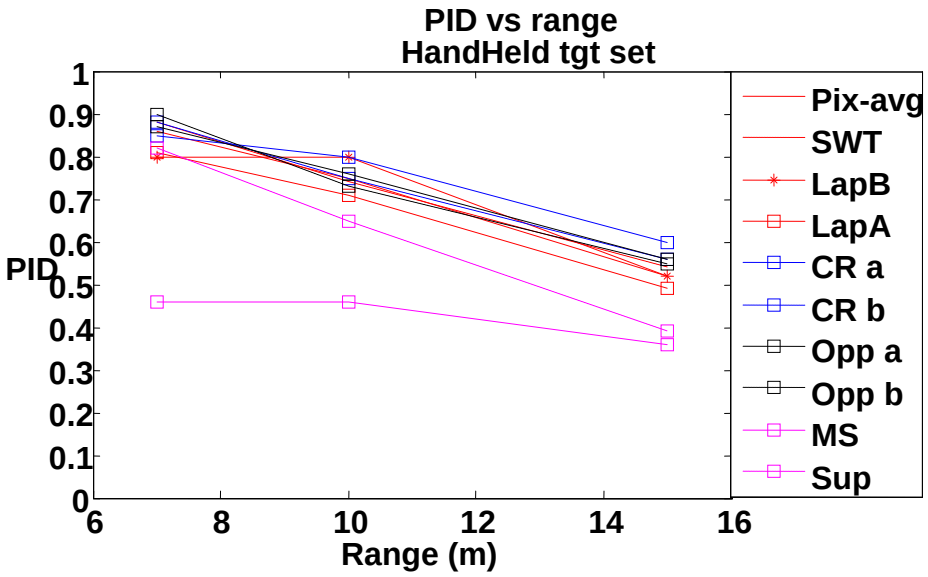


Figure 27. PID versus range results using image fusion of handheld targets.

The performance results in Figure 27 show that the ID performances decreased as the range increased for all the fusion algorithms except the MS algorithm. The performances for the majority of the fusion algorithms are approximately equal. As the range increases the objects become more 'similar', and as a result their differences become more difficult to discriminate. The FCIM metric can be used to monitor relationships between task difficulty and similarity in image information.

6.2.2 Metric Results

There were only three ranges used in the hand held target experiment, which, according to the table of critical values for the Pearson's correlation coefficient given in Appendix IV, means only correlations above 0.8 are significant, resulting in a 10% chance of rejecting the validity of the correlation.

The correlations for the FCIM and Hossny metric with PID of hand held targets for different image fusion algorithms are given in Table 5. Three image quality metrics reported in Piella and Heijmans (2003) are also listed in the table for comparison. The FCIM shows good correlation (0.8 or better) with six out of the ten fusion algorithms.

Table 5

Correlation values for various metrics and the human performance curves given in Figure 27

	LPA	LPB	OPA	OPB	CRA	CRB	MS	SUP	Pix-avg	SWT
FCIM	-0.35	0.91	0.87	0.48	0.27	-0.58	-0.99	0.87	0.96	0.99
Hossny Metric	0.99	-0.45	0.28	0.58	0.89	0.67	0.35	0.49	0.44	0.77
FQI	-0.23	0.00	0.27	0.01	-0.48	0.45	0.91	0.55	-0.04	0.49
WFQI	-0.03	0.87	-0.03	0.24	0.11	0.89	0.42	0.12	0.43	0.21
EFQI	-0.23	0.43	-0.04	0.18	-0.83	0.69	0.41	0.12	0.57	0.29

The FCIM has shown that, by measuring the transfer of information to the fused images, it is an effective metric for predicting human performance identifying military tactical vehicles and non-traditional targets using complementary source band information. More human perception studies are needed to investigate and validate the full potential of the FCIM.

7 Discussion

To be considered a valid metric, the metric must possess three key characteristics: 1) The metric must show that it is consistent with the data it is evaluating; when the trend of the data decreases or increases, the metric values must consistently decrease or increase in the same direction or consistently increase or decrease in opposite directions from the data. 2) The metric must be sensitive to changes in the data; that is, when the data changes significantly in magnitude, the metric changes similarly and vice versa. 3) The metric must be calibrated; for example, a change in metric value of say .5 to .6 translates consistently to an expected data value between .8 and .85. The FCIM has demonstrated it is consistent with human performance and sensitive to variations in performance caused by changes in the information content of the fused imagery. The previous scatter plots between FCIM and PID support an assumption of a linear relation. Therefore, linear regression is used to calibrate the metric to human performance data.

7.1 Evaluation of the FCIM Metric

For the controlled perception experiment, correlation plots of the resulting human performance data and the FCIM are given in Appendix I. From these plots, a linear relation between the PID results and the FCIM values is assumed. Under a linear assumption, a model can be derived for the relationship between the FCIM and PID performance using regression analysis techniques.

A linear model will have the form:

$$Y = a_1 X + a_0 \tag{7.1-1}$$

The idea then is to find the values of a1 and a0 that allows Y to best fit the actual data. In this case, Y is the dependent variable and X is the independent variable. Matlab ® has a function *polyfit* that can be used to determine the unknown coefficients by minimizing the sum of the squares of the model deviations from the actual PID data. This process is also known as least-squares fit to the data. As determined by *polyfit*, the coefficients of the linear models for each image fusion algorithm are given in Table 6.

Table 6

Coefficients for a linear model of the human performance data in the controlled perception experiment

	1515 fusion case		2525 fusion case	
	a1	a0	a1	a0
SUP	17.77	-13.41	18.78	-14.94
SIDWT	22.20	-18.13	50.18	-41.24
LAP	-12.66	12.75	-5.11	5.71

A vector of metric values from 0.8 to 1 was created and used to evaluate the fit between the model and PID results for each image fusion algorithm. The results of the model predictions using equation 7.1.1, the coefficients from Table 6 and the FCIM values for the controlled experiment are given in Figure 28.

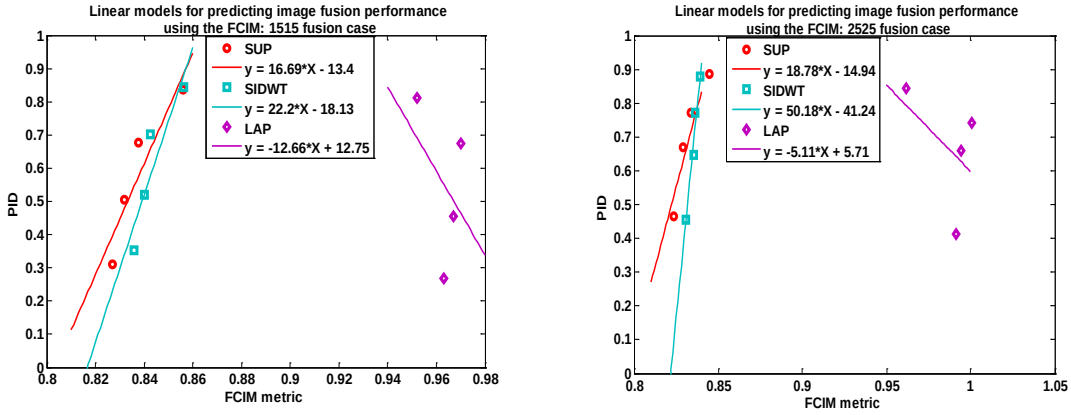


Figure 28. The plots in the left figure show the linear fits to model the PID performances for three different image fusion algorithms for the "1515" controlled perception experiment. Right shows the same analysis but for the 2525 fusion case.

The linear fits given in Figure 28 allow for predictions of identification task performance using images generated from the same system as those shown in Figure 14.

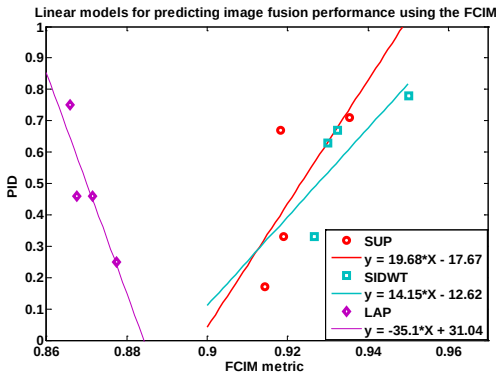


Figure 29. Linear fits to model the PID performances for three different image fusion algorithms for the LWIR/SWIR perception experiment.

Table 7

Coefficients for linear models of the human performance data in the LWIR/SWIR image fusion perception experiment

	Coefficients		Linear models
	a1	a0	
SUP	19.68	- 17.67	y = 19.68 * X – 17.67
SIDWT	14.15	-12.62	y = 14.15 * X – 12.62
LAP	-35.1	31.04	y = -35.1 * X + 31.04

The use of the above models given in Figure 29 should be used under the assumption that pre- and post image processing has been standardized to control the variability in the quality of the fused images. Under these restraints, a clearer understanding of the variations in performance for the FCIM, and human observers alike can be properly assessed.

8 Additional Application of the FCIM

An image quality metric that can be used to rank image fusion algorithms was developed using the FCIM. This metric is the Image Fusion Ranking Metric (IFRM) introduced in section 4.4.

8.1 Ranking metric

The FCIM has shown it correlates well with human performance using different spectral bands and different target sets. Additional research can be conducted to develop an objective metric that can correctly rank the PID performances associated with image fusion algorithms. This would allow hundreds of image fusion algorithms to be evaluated and possibly reduced by an order of magnitude based on a single metric. The IFRM was developed to objectively evaluate image fusion algorithms and rank the fusion algorithms according to how much each fusion algorithm transfers source information to the fused image. Equation 4.4-2 is repeated here:

$$IFRM = \frac{2 * I(X,Y)}{\left(H(X) - H(X \mid F)\right) + \left(H(Y) - H(Y \mid F)\right)} - \frac{2 * I(X,Y)}{H(X) + H(Y)} \tag{8.1-1}$$

The second term on the right hand side of the IFRM equation is Kvalseth's normalized MI measure for two variable cases and the term on the right is the FCIM. Using the fact that Kvalseth's normalized MI measure is a constant regardless of which image fusion algorithm is being evaluated, the variation in IFRM values depends upon the individual image fusion algorithm under investigation. Therefore, the less redundant (that is, more complementary) the information transferred by the fusion algorithm, the lower the IFRM values, and vice versa. This would allow each fusion algorithm to be measured on common terms.

The FCIM can only obtain values greater than or equal to that of Kvalseth's normalized MI measure; therefore, the IFRM will take on values between 0 and 1. Higher values are given higher rank than lower values. As the fused image explains more of the uncertainty concerning each of the source images, the FCIM, and therefore the IFRM, gives larger values. This is a key concept that allows each fusion algorithm to be compared against another. The rankings for the fusion algorithms used in the LWIR/SWIR perception study are given in Table 8. Figure 30 shows plots of the data in Table 8 along with the human performance results from the experiment.

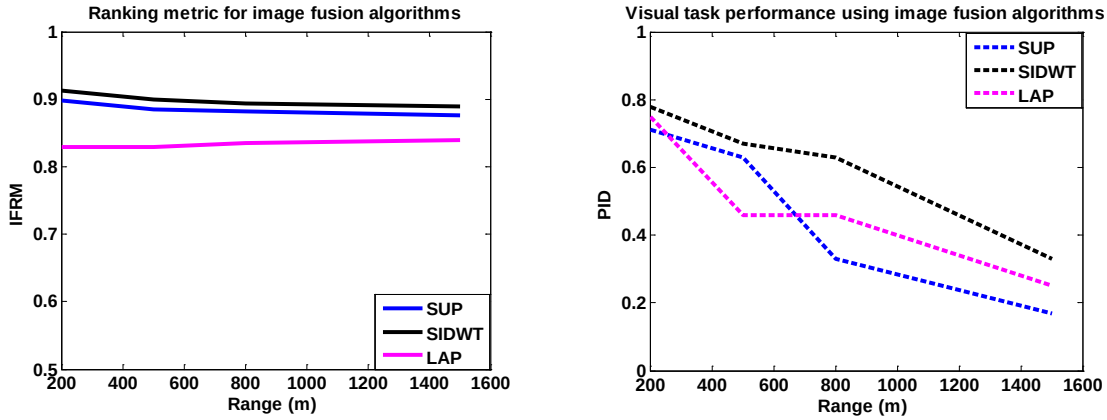


Figure 30. Rankings for three image fusion algorithms as determined by the IFRM. Human performance results from the LWIR/SWIR perception experiment are given for comparison.

Table 8

Image fusion algorithm ranking for the IFRM

	IFRM Values versus range				Rankings
	200m	500m	800m	1500m	
IFRM_SUP	0.899	0.885	0.882	0.877	2
IFRM_SIDWT	0.913	0.900	0.893	0.889	1
IFRM_LAP	0.829	0.829	0.835	0.840	3

The IFRM was able to identify the best performing fusion algorithm. The rankings are given in Table 8; in this case the highest ranking algorithm was the SIDWT fusion algorithm. Human performance data shown on the right of Figure 30 confirms the ranking. Because so few observers participated in the perception experiment, nothing definitive can be mentioned about how IFRM compares with human performance for the other algorithms.

Because the IFRM uses the FCIM and a constant value in its calculations, the IFRM and the FCIM should give identical rankings. The benefit to using the IFRM versus the FCIM for the moment is not yet established; however, it is expected that the IFRM is behaving similarly to when variations in image pixel intensities ride on top of a pedestal; the pedestal is often removed to get a clearer picture of the true variation, i.e., to improve the contrast in the image. This pedestal removal usually results in a higher quality image. The pedestal in the case of the IFRM is Kvalseth's MI measure (a normalized MI measure of the inputs) and the variations that come from each fusion

algorithm can be measured using the FCIM. The IFRM was tested to determine if any improvements in the correlation values given in Table 3 can be further improved upon by removing the 'pedestal' from the FCIM. The correlation analysis results using IFRM and human observer PID are given in Table 9.

Table 9

Comparison between correlations using PID and FCIM and correlations using PID and IFRM for the LWIR/SWIR perception experiment

	Performance correlations	
	PID vs FCIM	PID vs IFRM
SUP	0.75	0.86
SIDWT	0.77	0.85
LAP	-0.89	-0.82

There was a correlation improvement as a result of modifying the FCIM in this way. More research using IFRM should follow to explain the cause in increased correlation between the FCIM and IFRM.

9 Image Quality with Reference Image Generation

One of the reasons why image fusion quality assessment is such a difficult task is because usually there is not a reference image available to compare the fused result. A new method for generating image fusion reference images is presented in this section.

The human visual system (HVS) is known to be sensitive to edge information in images. A method is proposed that uses the high frequency information in each source image to emphasize the edge information. To this end, each source band image is HP filtered and the original source image along with the HP version of that image is used as input to image fusion algorithms. In this way, high frequency information within the respective source bands is "boosted" as a function of individual fusion algorithms. To obtain the fused resultant image, the fused high frequency "boosted" images are simply added to form the fused reference image. This process is depicted in Figure 31.

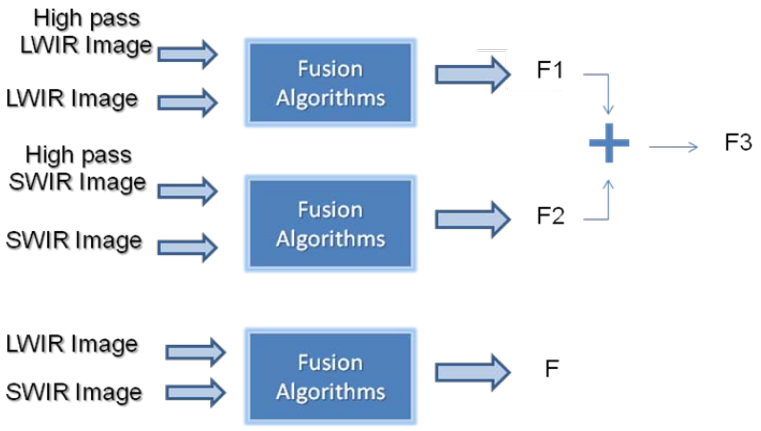


Figure 31. Procedure to generate a reference image using the image fusion algorithm to "boost" the high frequency component of each source image before combining the information.

Adding the fused images requires multiplicative factors that sum to 1 to account for display range differences. This process is nothing more than the superposition fusion defined previously, which is known to reduce contrast in the images. More optimal weighting schemes or combination procedures may be investigated to determine if there exists a higher quality image using the procedure in Figure 31. Given a reference image, popular reference metrics, like UIQI and peak signal-to-noise (PSNR) and mean squared error (MSE) could then be used to evaluate the quality of the image fusion algorithm.

A new metric that uses MI is introduced and proposed to evaluate the similarity between a fused image obtained using the original source images and a different fused image obtained using the "high-frequency boosted" source images. This metric is normalized using a sum of similarity measures between each original source image and the corresponding "high-frequency boosted" image. This measure is given as follows:

$$\text{Fusion Reference Metric} = \frac{I(F, F3)}{I(LWIR, F1) + I(SWIR, F2)} \tag{9.1-1}$$

Evaluating how well an individual fusion algorithm can improve its high frequency information transfer from the source images to the fused image, coupled with human performance evaluations of the data, may lead to the development of methods for which a multitude of image fusion algorithms can be equivalently compared and discriminated against. The FRM can be used to discriminate between the 'best' fusion algorithms according to the similarity between the output of each fusion algorithm and the reference images.

To investigate this claim, let's revisit the performance results from the LWIR/SWIR perception experiment. Rather than look at the PID as a function of range,

I decided to investigate what percentage of correct IDs the ensemble of observers has with the entire target set. In this way I can rank the overall correct ID for an entire image set. To do this, I first used ordered statistics of the FRM values across the image set, after which I plotted an empirical distribution of the data values for each fusion algorithm. This is shown in Figure 32. The blue curve represents the actual data and the magenta curve represents the fit to a normal distribution.

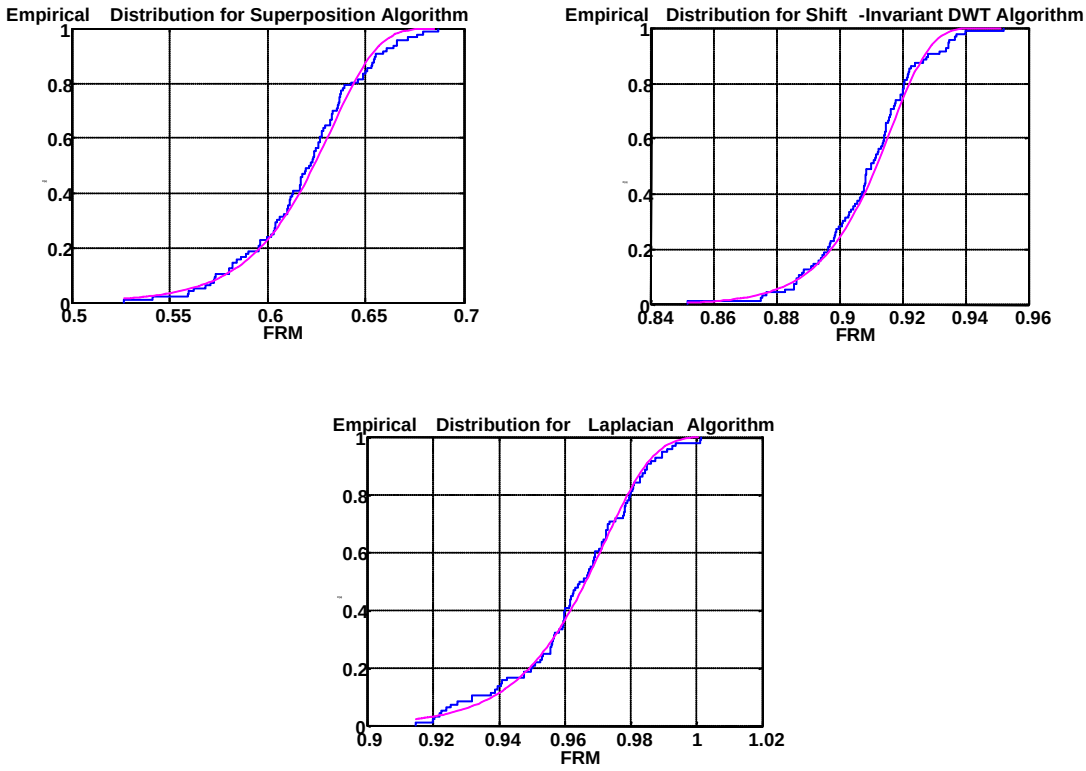


Figure 32. Empirical distributions using ordered statistics of the FRM evaluation of three different fusion algorithms.

The FRM ordered statistics follow a normal distribution for this image set. The question I initially sought to answer was how many of the test images fell within +/- 1 standard deviation (*STD*) from the mean FRM value. Analyzing the data this way would

tell me the distribution of images per range. The thought here is that by applying the FRM to the image set I could determine the proportion of images within a required distance from the mean value, thereby obtaining knowledge of which image at a particular range is included. Depending on the distribution of farthest to closest range imagery, I could impose a rank on the algorithm. The slope of the curves could also be a candidate for ranking the algorithms. For this study, the amounts of images within +/- 1 *STD* for the fused images were used with results given in Table 10.

Table 10

Percentage of images that fell within +/- 1 *STD* of the mean FRM value for the LWIR/SWIR fused images

SIDWT	71/96 = 74 %
LAP	66/96 = 69 %
SUP	65/96 = 68 %

There were 74% of the SIDWT images within a *STD* from the mean FRM value. Of the 26% that were outside the mean, 58% of them were the images at 1500 meters. This suggests that the FRM is grouping the closer range images near its mean value. The other algorithms had similar distributions.

To determine the probability of correct, I plotted the ordered statistics of the FRM versus PID. These plot is given are given in Figure 33.

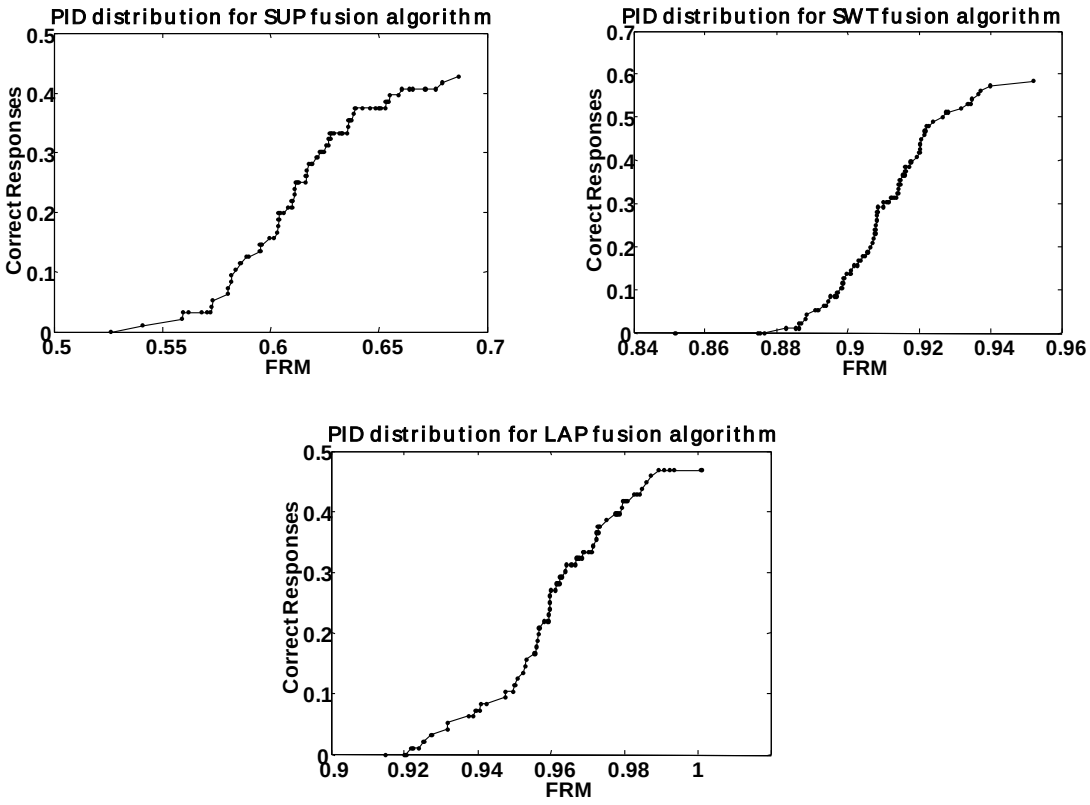


Figure 33. Probability of correct responses recorded as a function of FRM values for the fused image set in the LWIR/SWIR experiment.

The probabilities of correct responses for the superposition, SIDWT and the Laplacian fusion algorithms were 43%, 59% and 47% respectively, as seen in the above figure. If the percentage of images within +/- 1 *STD* of the mean value is used to rank the fusion algorithms, then the ranking of the algorithms would be: SIDWT, Laplacian and then superposition. A small point to note here, the IFRM and FRM rank SIDWT as the better fusion algorithm in this study. Further research should be conducted to determine whether or not this characteristic of the FRM demonstrates consistency. Consistently grouping closer range imagery into one cluster and the farthest range imagery into another cluster gives the FRM a useful classification characteristic for ranking image fusion algorithm performance.

10 Conclusion

This dissertation introduces into the image fusion community a new image fusion evaluation measure, FCIM, which measures the information content within fused imagery and generates an associated performance prediction value. The FCIM is effective in predicting human identification task performance for a particular set of assumptions and conditions. The assumption that the fused imagery evaluated has been properly prepared for human perception is needed to relate human perception to the FCIM measures. The performance of the FCIM is tested using different spectral bands and different target sets, with results showing the FCIM's ability to effectively evaluate the complementary and redundant information content in fused imagery under these conditions for the purpose of predicting human task performance.

Additionally, the FCIM has shown itself to be useful as a tool for comparing various image fusion algorithms and can be used as a measure to investigate image fusion quality. The development of the FCIM is a major step towards the ultimate goal of being able to evaluate human visual task performance without the need of time consuming, expensive perception tests.

REFERENCES

Adelson, E., Anderson, C., Bergen, J., Burt, P., & Ogden, J. (1984). Pyramid Methods in Image Processing. *RCA Engineer*, *29*(6).

Anscombe, F.J., (1973). Graphs in Statistical Analysis. *American Statistician, 27,* 17-21.

Arathi, T., & Soman, K., (2009). Performance Evaluation of Information Theoretic Image Fusion metrics over Quantitative metrics. *International Conference on Advances in Recent Technologies in Communication and Computing,* 225-227.

Attneave, F. (1959). *Applications of Information in Psychology,* New York: Holt, Rinehart, and Winston

Burt, P., & Adelson, E. (1983). The Laplacian Pyramid as a Compact Image Code. *IEEE Transactions on Communications, 31*(4).

Canga, E., Nikolov, S., Canagarajah, C., Bull, D., Dixon, T., Noyes, J., & Troscianko, T. (2005). Characterization of Image Fusion Quality Metrics for Surveillance Applications over Bandlimited Channels. *IEEE 2005 7th International Conference on Information Fusion*, August 2005, 483-489.

Chavez, P., Sides, S., & Anderson, J. (1991). Comparison of three methods to merge multi-resolution and multi-spectral data: Landsat TM and SPOT panchromatic, *Photogrammetric Engineering and Remote Sensing*, *57*, 295-303.

Cvejic, N., Loza, A., Bull D., & Canagarajah, N. (2006). A similarity metric for assessment of image fusion algorithms. *International Journal of Information and Communication Engineering, 2*(3), 178-182

Cvejic, N., Loza, A., Bull D., & Canagarajah, N., (2005). A Novel Metric for Performance Evaluation of Image Fusion Algorithms. *World Academy of Science, Engineering and Technology,* 7, 80-85

Driggers, R., Vollmerhausen, R., & Krapels, K. (2000). Target identification performance as a function of low spatial frequency image content. *Optical Engineering, 39*(9), 2458-2462.

Gunatilaka, A., & Baertlein, B. (2001). Feature-Level and Decision Level Fusion of Noncoincidently Sampled Sensors for Land Mine Detection. *IEEE Transactions on Pattern Analysis and Machine Learning, 23(6)*, 577-589.

Hall, L., & Liinas, J. (1997). An introduction to multi-sensor data fusion. *Proceedings of IEEE, 85*, 6-23.

Harper, W., Lilliesand, T., & Keifer, R. (1990). The use of Intensity-Hue-Saturation transformations for merging SPOT panchromatic and multi-spectral image data. *Photogrammetric Engineering and Remote Sensing, 56*, 459-467.

Horibe, Y. (1985). Entropy and correlation, *IEEE Transactions on Systems, Man, and Cybernetics*, *15*(5), 641-642.

Hossny, M., Nahavandi, S., & Creighton, D. (2008). Comments on 'Information measure for performance of image fusion. *Electronic Letters, 44*(18), 1066-1067

Howell, C., Halford, C., & Krapels, K. (2010). Image fusion algorithm assessment using measures of complementary and redundant information. *Proceedings of SPIE,* 7992.

Huntsberger, T., & Jawerth, B. (1993). Wavelet based Sensor Fusion. *Proceedings of SPIE,* 2059.

Krebs, W., & Sinai, M. (2002). Psychophysical Assessments of Image-Sensor Fused Imagery. *Human Factor, 44(2),* 257-271.

Kvalseth, T. (1987). Entropy and Correlation: Some Comments. *IEEE Transactions on Systems, Man, and Cybernetics. 17,* 517-519.

Li, H., Manjunath, B., & Mitra, S. (1995). Multi-sensor image fusion using wavelet transforms. *Graphical Models and Image Processing, 57*(3), 235-245.

Moore, R., Halford, C., & Howell, C. (2007). Finding a fusion metric that best reflects human observer performance. *Proceedings of SPIE,* 6543, 65430N.

Moore, R., (2007). Investigation into human performance using fused complementary spatial information. Unpublished raw data.

Nava, R., Escalante-Ramirez, B., & Cristobal, G., (2007). Mutual Information improves image quality assessment. *SPIE Electronic Imaging and Signal Processing*, DOI: 10.1117/2.1200708.0824

Neriani, K., Pinkus, A., & Dommett, D. (2008). An investigation of image fusion algorithms using a visual performance-based image evaluation methodology. *Proceedings of SPIE*, *6968*, 696819-696819-9.

Petrovic, V. (2007). Subjective tests for image fusion evaluation and objective metric validation. *Information Fusion,* 8(2), 208-216.

Petrovic, V. & Xydeas, C. (2005). Objective image fusion performance characterization. *Computer Vision*, *2*(10), 1866 – 1871.

Piella, G., & Heijmans, H. (2003). A new quality metric for image fusion. *International Conference on Image Processing*, 173-176.

Pohl, C., & Van Genderen, J. (1998). Multi-sensor image fusion in remote sensing: concepts methods and applications. *International Journal of Remote Sensing, 19,* 823-854.

Qu, G., Zhang, D., & Yan, P. (2002). Information fusion measure for performance of image fusion. *Electronic Letters*, *38*(7), 313-315.

Reza, F. (1961) *An introduction to information theory*, New York: McGraw Hill

Rockinger, O. (1997). Image sequence fusion using a shift invariant wavelet transform. *Proceedings of IEEE, International Conference on Image Processing, 13*, 288-291.

Ryan, D., & Tinkler, R. (1995). Night pilotage assessment of image fusion. *Proceedings of SPIE*, 2465, 50-67.

Smith, F., Chari, S., Halford, C., Fanning, J., & Reynolds, J. (2009). Novel image fusion quality metric based on sensor models and image statistics. *Proceedings of SPIE*, 7300.

Toet, A., & Franken, E. (2003). Perceptual evaluation of different image fusion schemes. *Display, 24*, 25-37.

Toet, A. (1989). Image Fusion by a Ratio of Low-pass Pyramid. *Pattern Recognition Letters, 9*(4), 245-253.

Tsagaris, V., & Anastassopoulos, V. (2006). Global measure for assessing image fusion methods. *Optical Engineering*. 45, 026201. doi: 10.1117/1.2174005

U.S. Army, NVESD (2005). Night vision thermal and image processing performance model, users manual. *U.S. Army RDECOM CERDEC NVESD, Ft. Belvoir, VA* Revision 8.

Vollmerhausen R., Jacobs E., & Driggers R. (2004). New metric for predicting target acquisition performance. *Optical Engineering*, *43*(11), 2806-2818.

Wald, L. (1999). Some terms of reference in data fusion. *IEEE Transactions on Geoscience and Remote Sensing, 37(3),* 1190-1193.

Wang, Q., Yu, D., & Shen, Y. (2009). An Overview of image fusion metrics. *International Instrumentation and Measurement Technology Conference,* 918-923

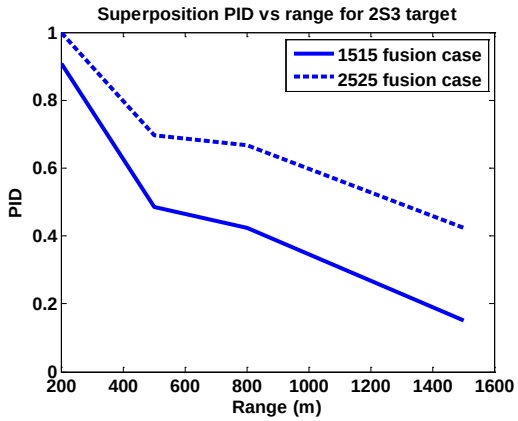
Wang, Z., & Bovik, A. (2002). A universal image quality index. IEEE *Signal Processing Letter*, *9*(3), 81-84.

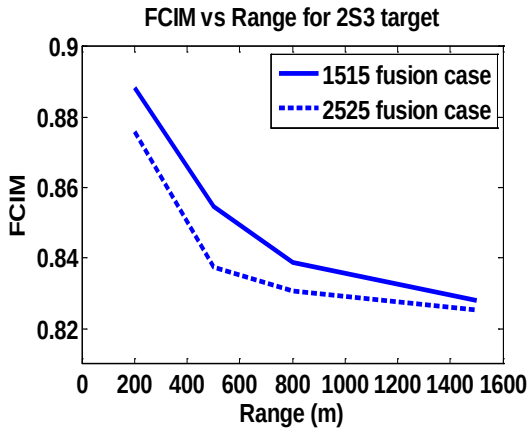
Wang, Z., & Bovik, A., (2009). Mean squared error: love it or leave it? – A new look at signal fidelity measures. *IEEE Signal Processing Magazine, 26*(1), 98-117.

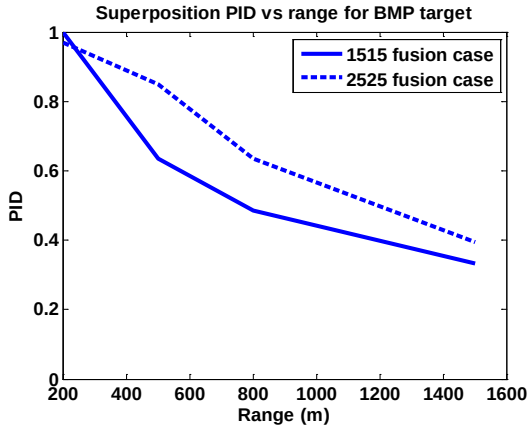
Wang, Z., Bovik, A., Shiekh, H., & Simoncelli, E., (2004). Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*, *13*(4).

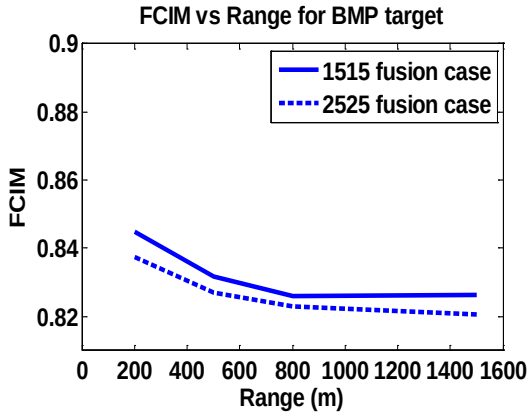
Waxman, A., Gove, A., Fay, D., Racamato, J., Carrick, J., Seibert, M., & Savoye, E. (1997) Color Night Vision: Opponent Processing in the Fusion of Visible and IR Imagery. *Neural Networks*, *10*(1), 1-6.

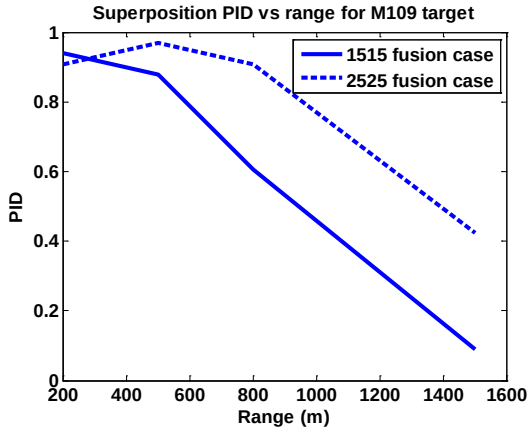
APPENDIX I

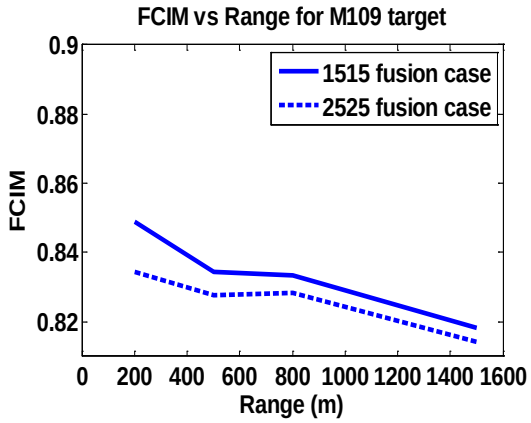

Superposition PID vs range for 2S3 target
1515 fusion case
2525 fusion case
PID
Range (m)

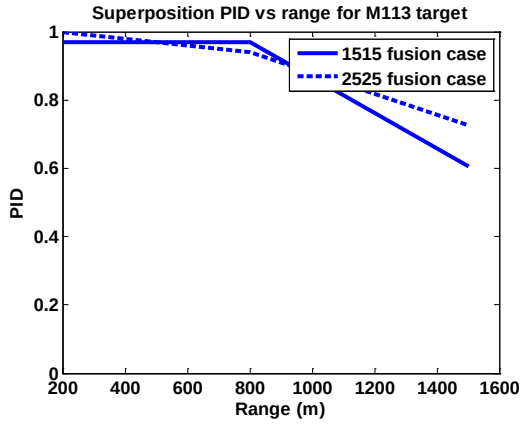

FCIM vs Range for 2S3 target
1515 fusion case
2525 fusion case
FCIM
Range (m)

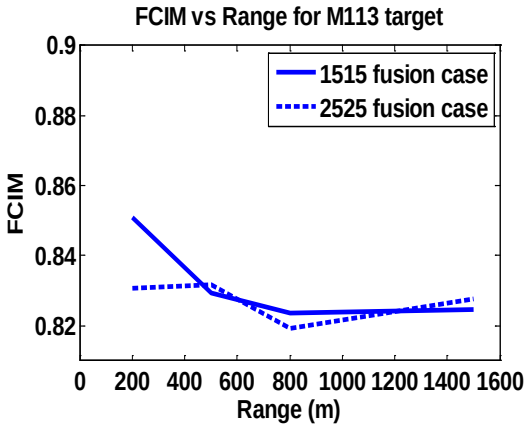

Superposition PID vs range for BMP target
1515 fusion case
2525 fusion case
PID
Range (m)

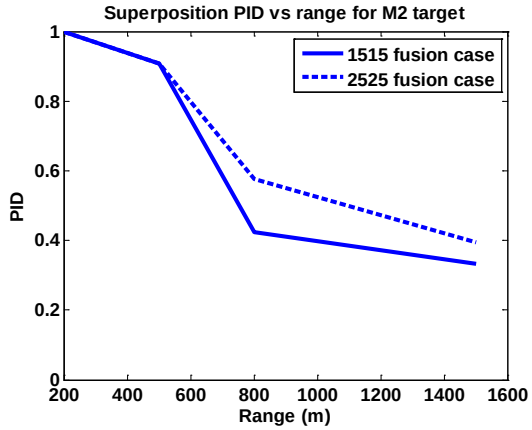

FCIM vs Range for BMP target
1515 fusion case
2525 fusion case
FCIM
Range (m)

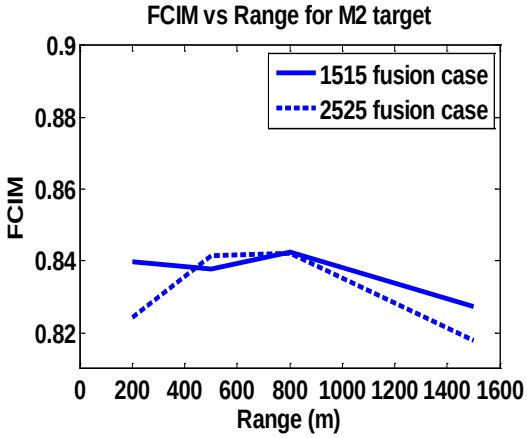

Superposition PID vs range for M109 target
1515 fusion case
2525 fusion case
PID
Range (m)

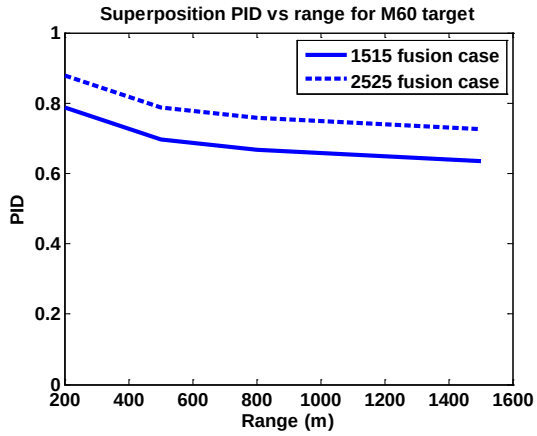

FCIM vs Range for M109 target
1515 fusion case
2525 fusion case
FCIM
Range (m)

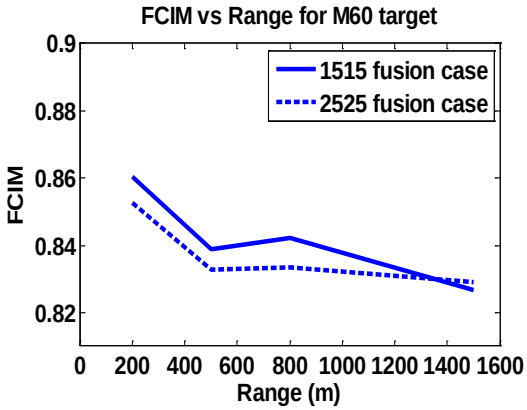

Superposition PID vs range for M113 target
1515 fusion case
2525 fusion case
PID
Range (m)


FCIM vs Range for M113 target
1515 fusion case
2525 fusion case
FCIM
Range (m)


Superposition PID vs range for M2 target
1515 fusion case
2525 fusion case
PID
Range (m)


FCIM vs Range for M2 target
1515 fusion case
2525 fusion case
FCIM
Range (m)


Superposition PID vs range for M60 target
1515 fusion case
2525 fusion case
PID
Range (m)


FCIM vs Range for M60 target
1515 fusion case
2525 fusion case
FCIM
Range (m)

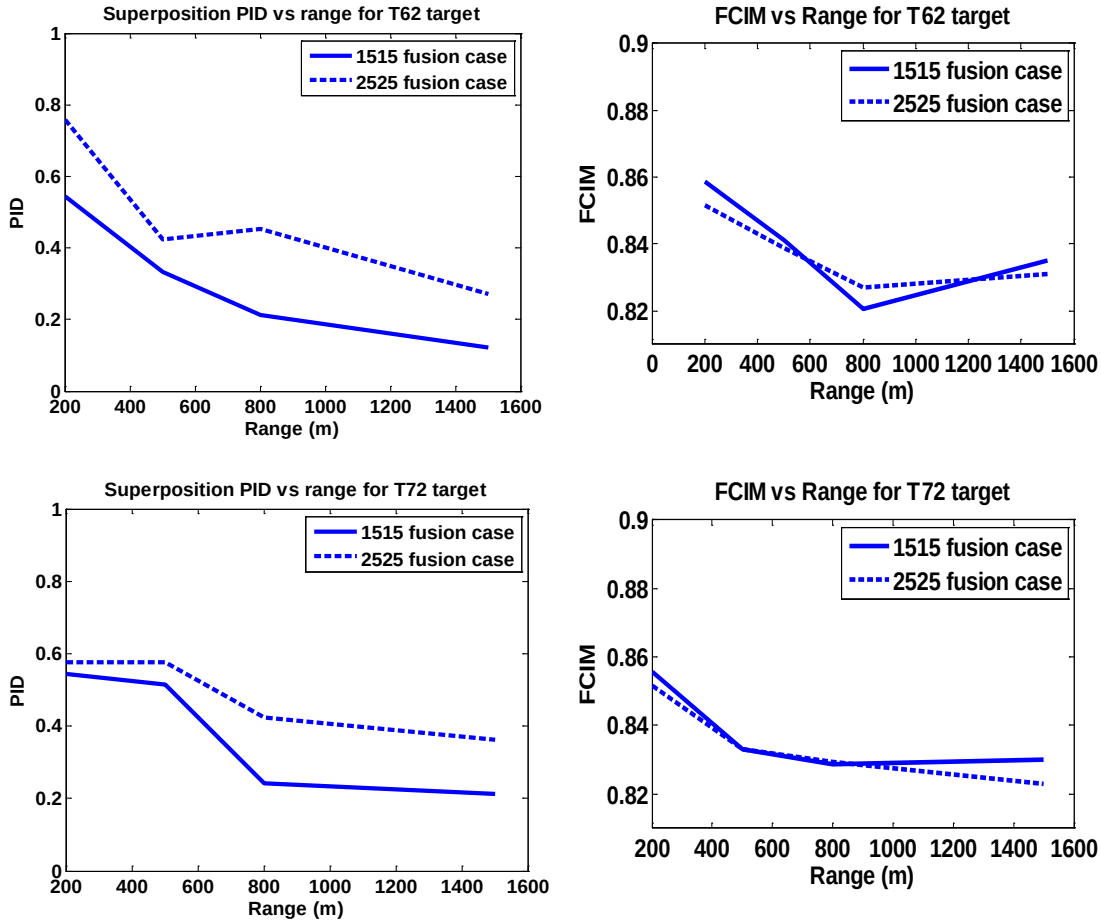
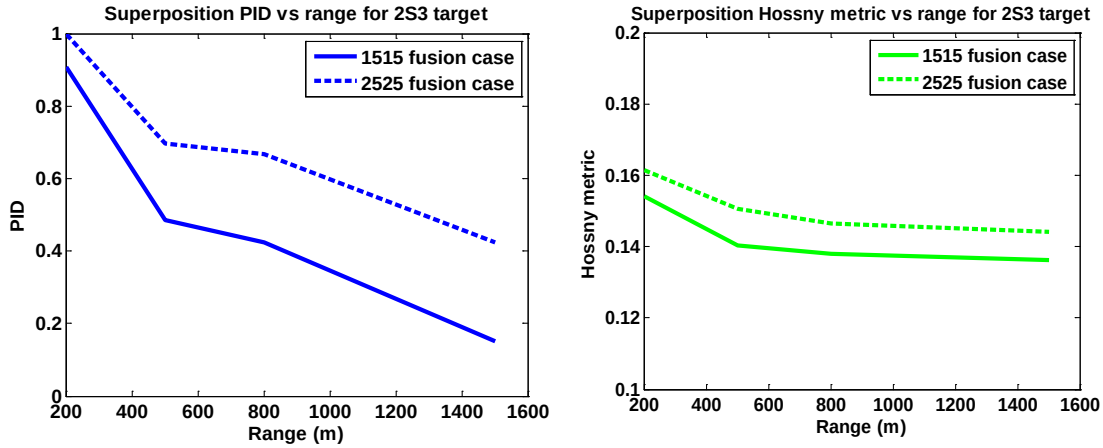
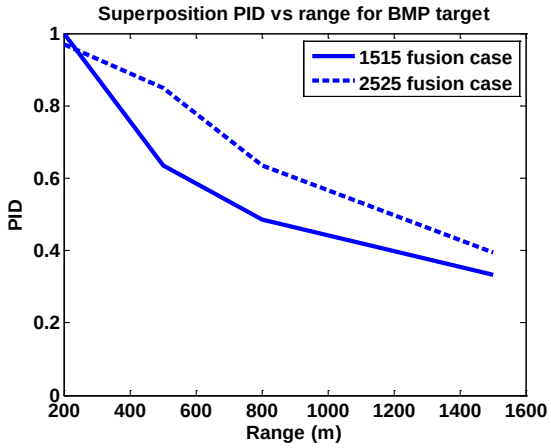
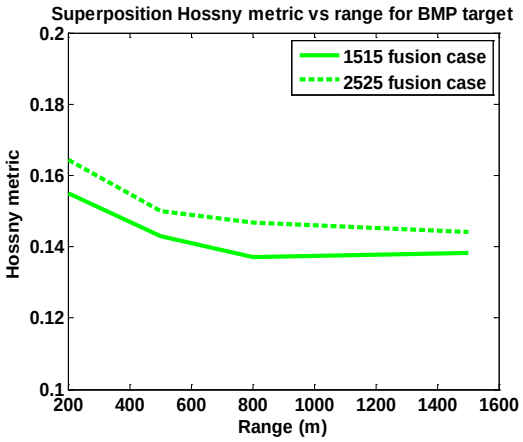
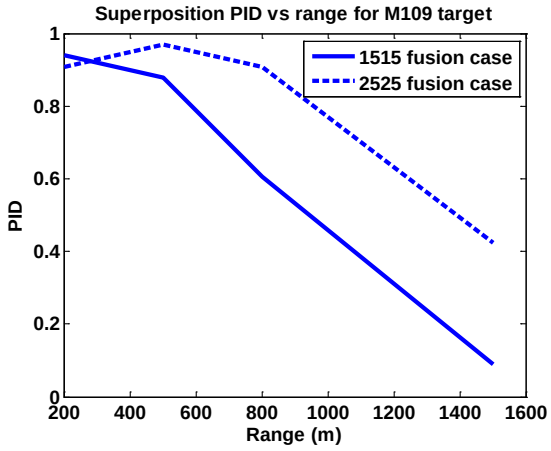


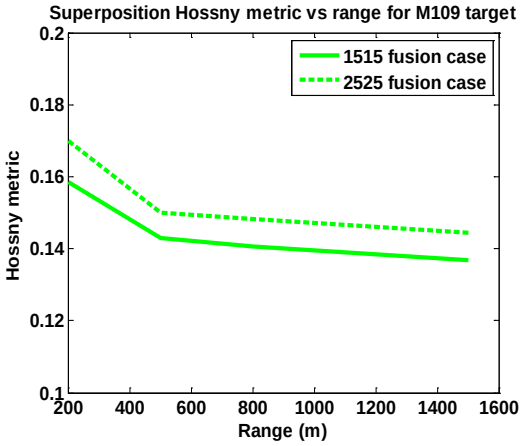
Figure 34. PID performance curves from Figure 13 and FCIM performance for the controlled human perception experiment. The entire target set using the superposition fusion algorithm.

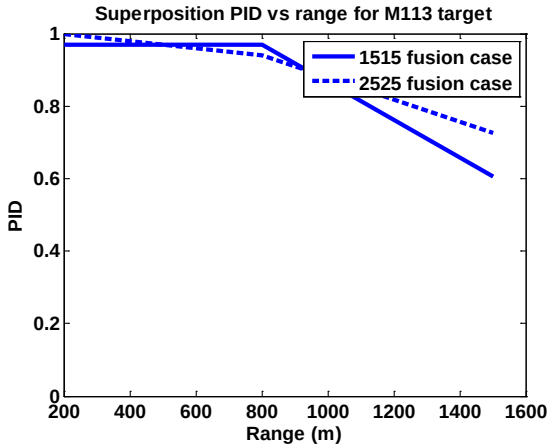


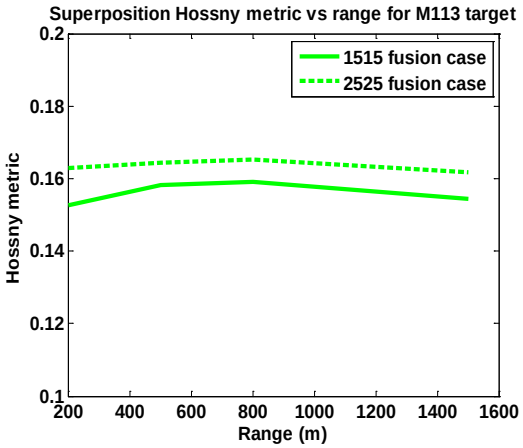

Superposition PID vs range for BMP target
1515 fusion case
2525 fusion case
PID
Range (m)

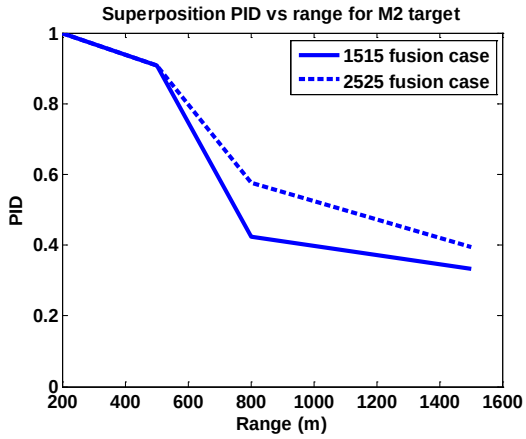

Superposition Hossny metric vs range for BMP target
1515 fusion case
2525 fusion case
Hossny metric
Range (m)

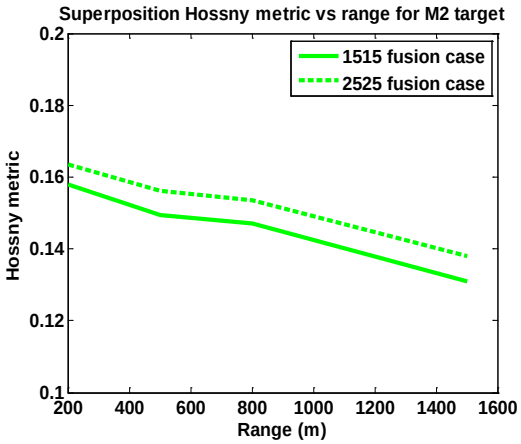

Superposition PID vs range for M109 target
1515 fusion case
2525 fusion case
PID
Range (m)

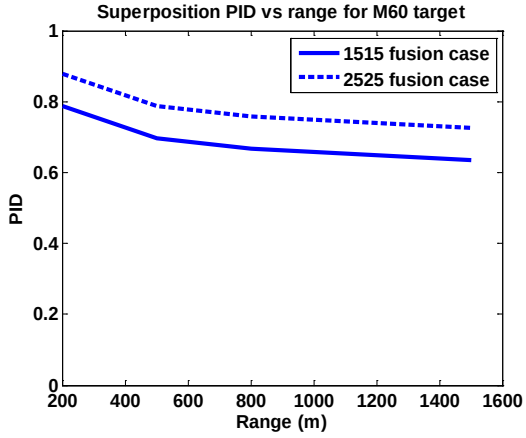

Superposition Hossny metric vs range for M109 target
1515 fusion case
2525 fusion case
Hossny metric
Range (m)

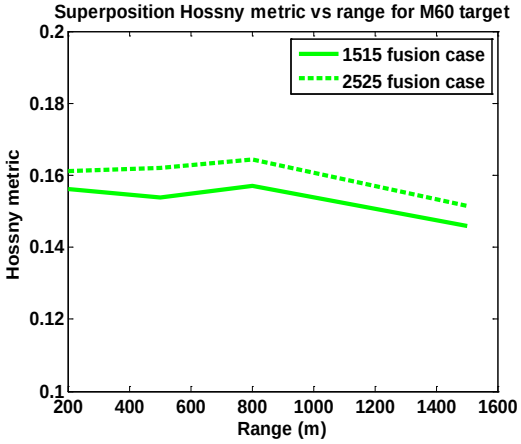

Superposition PID vs range for M113 target
1515 fusion case
2525 fusion case
PID
Range (m)

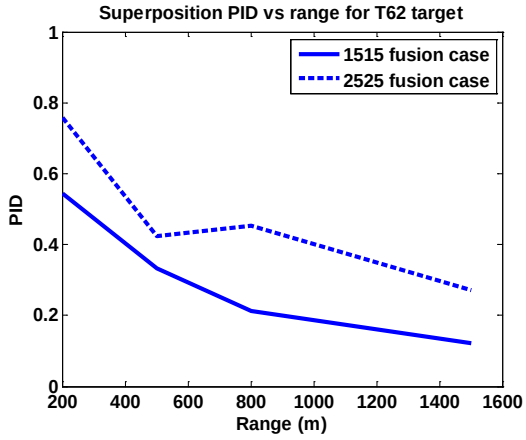

Superposition Hossny metric vs range for M113 target
1515 fusion case
2525 fusion case
Hossny metric
Range (m)

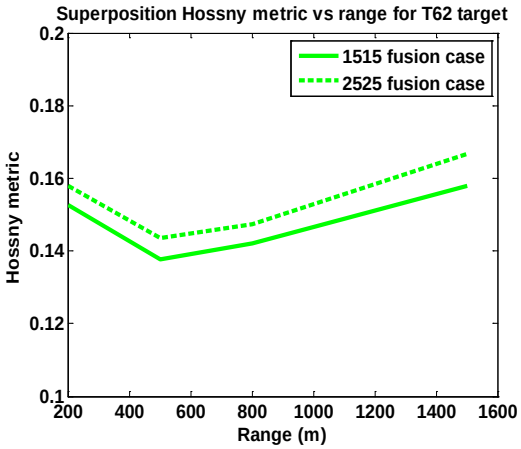

Superposition PID vs range for M2 target
1515 fusion case
2525 fusion case
PID
Range (m)


Superposition Hossny metric vs range for M2 target
1515 fusion case
2525 fusion case
Hossny metric
Range (m)


Superposition PID vs range for M60 target
1515 fusion case
2525 fusion case
PID
Range (m)


Superposition Hossny metric vs range for M60 target
1515 fusion case
2525 fusion case
Hossny metric
Range (m)


Superposition PID vs range for T62 target
1515 fusion case
2525 fusion case
PID
Range (m)


Superposition Hossny metric vs range for T62 target
1515 fusion case
2525 fusion case
Hossny metric
Range (m)

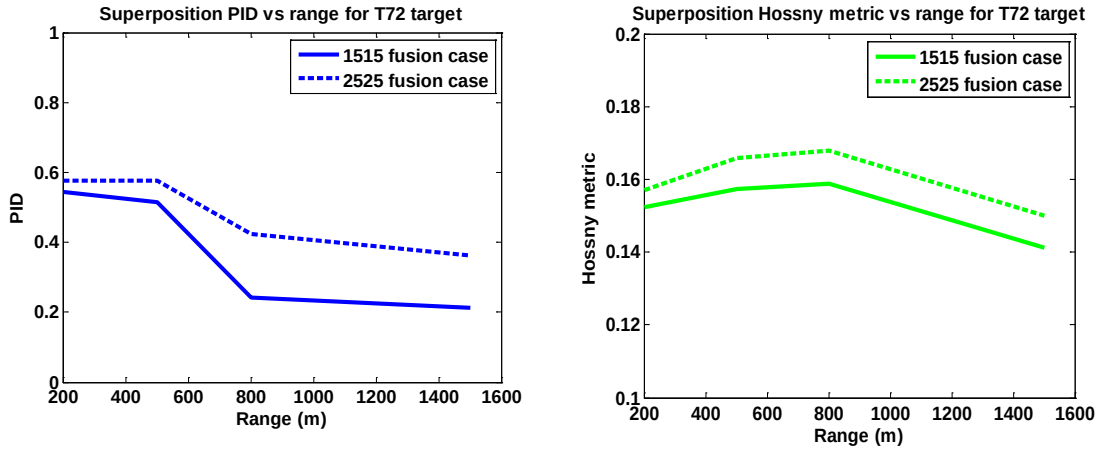
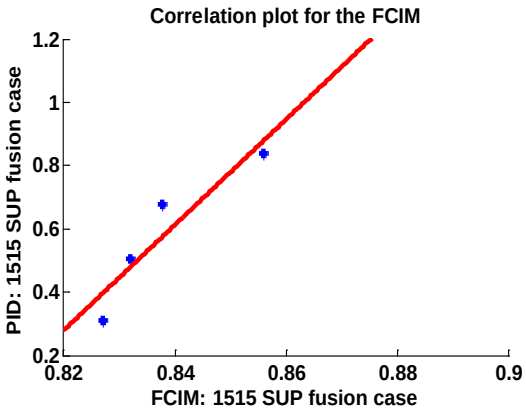
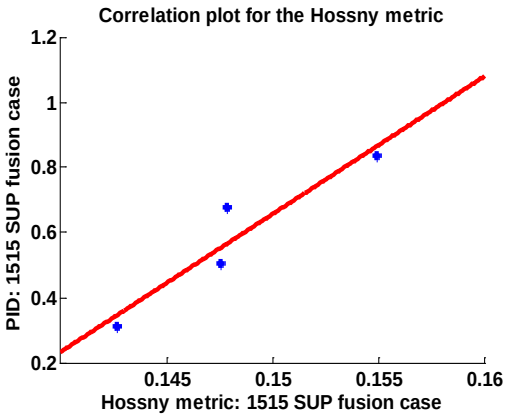
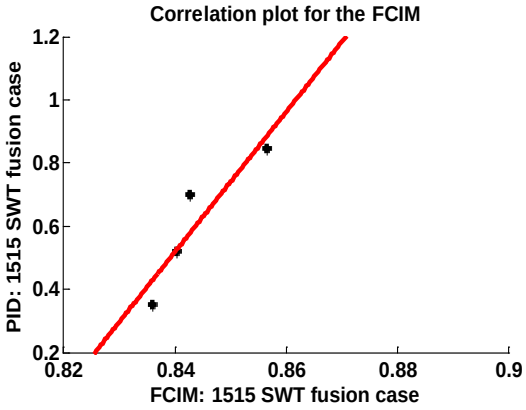


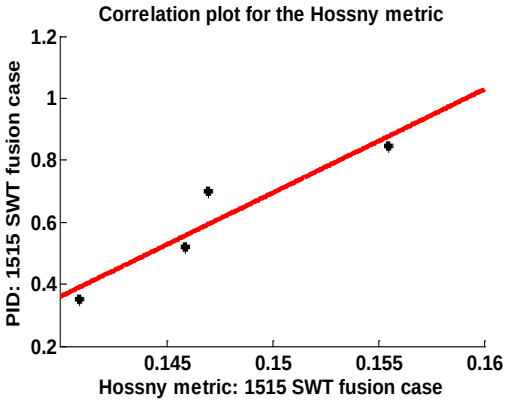
Figure 35. PID performance curves from Figure 13 and the Hossny metric performance for the controlled human perception experiment. The entire target set using the superposition fusion algorithm.

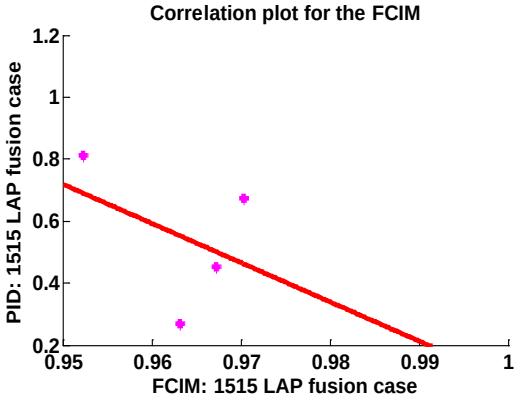
APPENDIX II

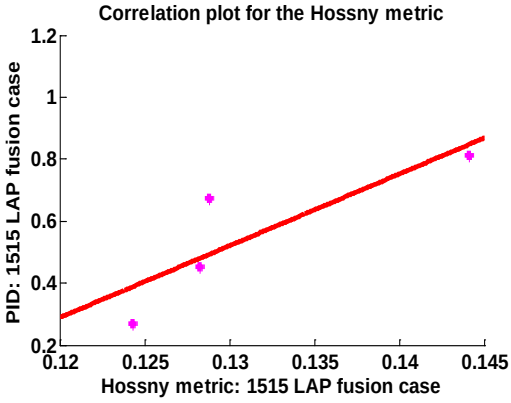

Correlation plot for the FCIM
PID: 1515 SUP fusion case
FCIM: 1515 SUP fusion case
1.2
1
0.8
0.6
0.4
0.2
0.82
0.84
0.86
0.88
0.9


Correlation plot for the Hossny metric
PID: 1515 SUP fusion case
Hossny metric: 1515 SUP fusion case
1.2
1
0.8
0.6
0.4
0.2
0.145
0.15
0.155
0.16


Correlation plot for the FCIM
PID: 1515 SWT fusion case
FCIM: 1515 SWT fusion case
1.2
1
0.8
0.6
0.4
0.2
0.82
0.84
0.86
0.88
0.9


Correlation plot for the Hossny metric
PID: 1515 SWT fusion case
Hossny metric: 1515 SWT fusion case
1.2
1
0.8
0.6
0.4
0.2
0.145
0.15
0.155
0.16


Correlation plot for the FCIM
PID: 1515 LAP fusion case
FCIM: 1515 LAP fusion case
1.2
1
0.8
0.6
0.4
0.2
0.95
0.96
0.97
0.98
0.99
1


Correlation plot for the Hossny metric
PID: 1515 LAP fusion case
Hossny metric: 1515 LAP fusion case
1.2
1
0.8
0.6
0.4
0.2
0.12
0.125
0.13
0.135
0.14
0.145

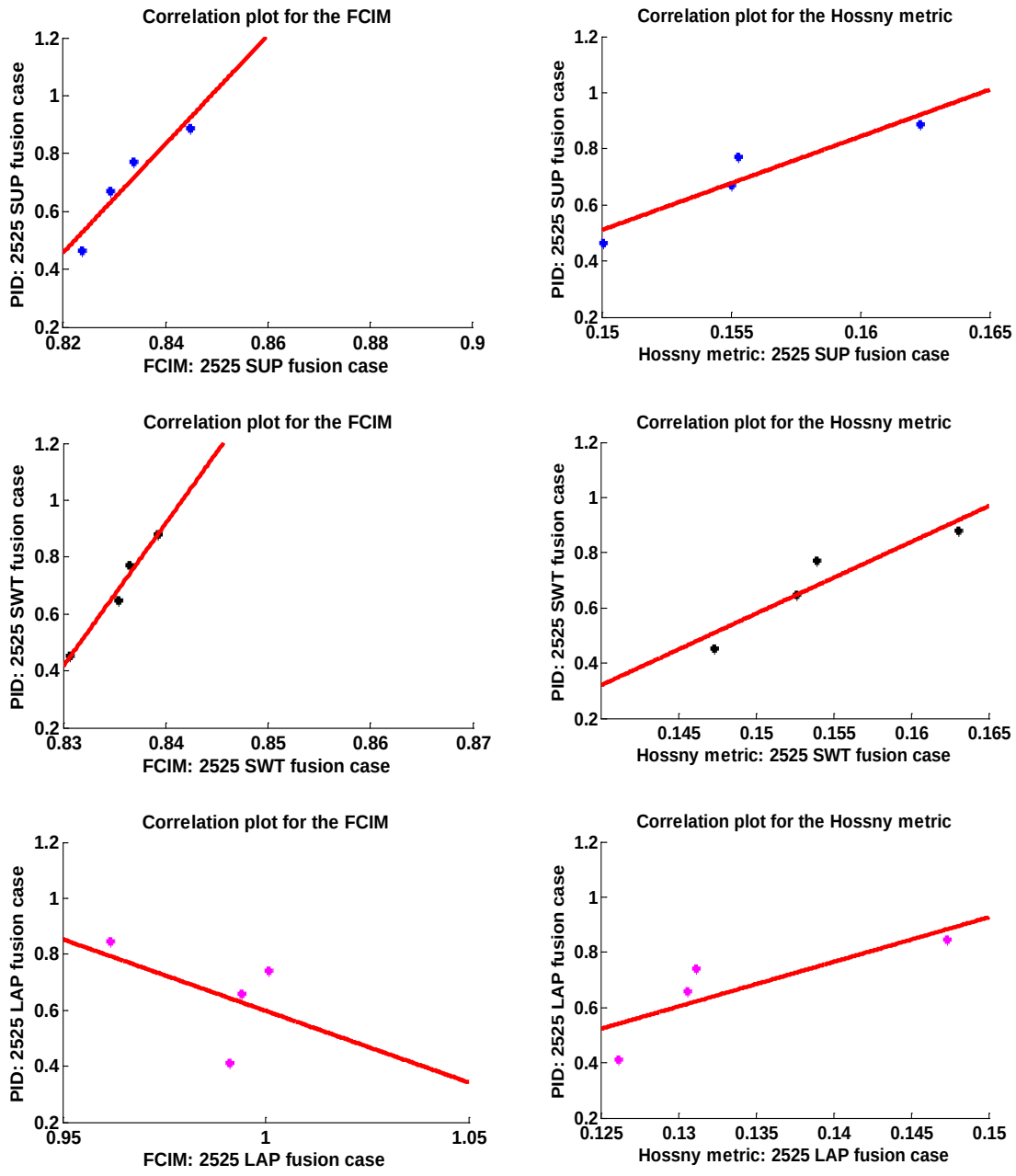


Figure 36. Scatter plots for which correlation measures were obtained between human performance and each of the metrics used in the controlled perception experiment.

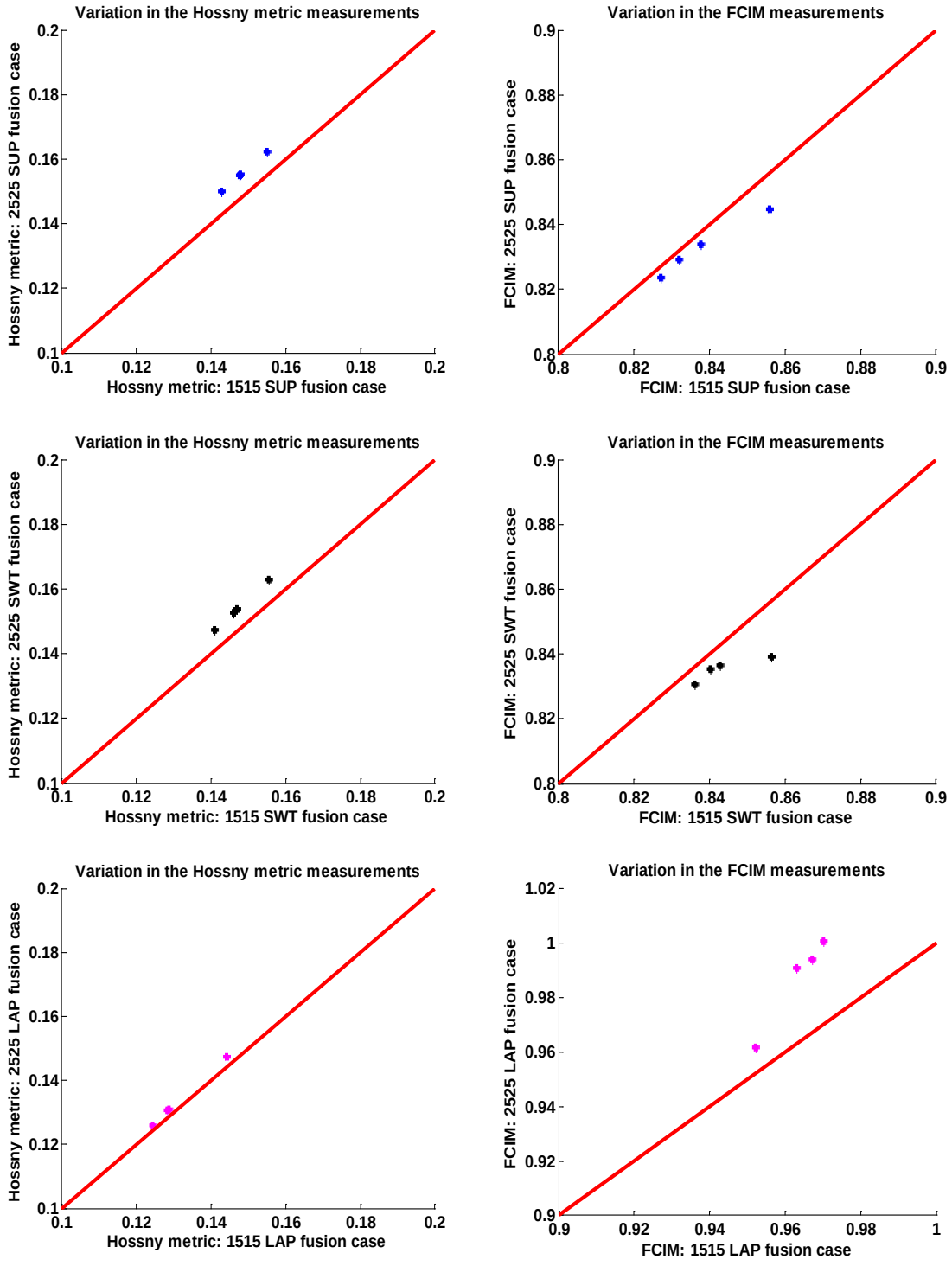


Figure 37. Sensitivity plots for each of the fusion algorithms used in the controlled experiment. The Hossny metric for the 1515 and 2525 fusion cases is given on the left. The FCIM plots are on the right.

APPENDIX III

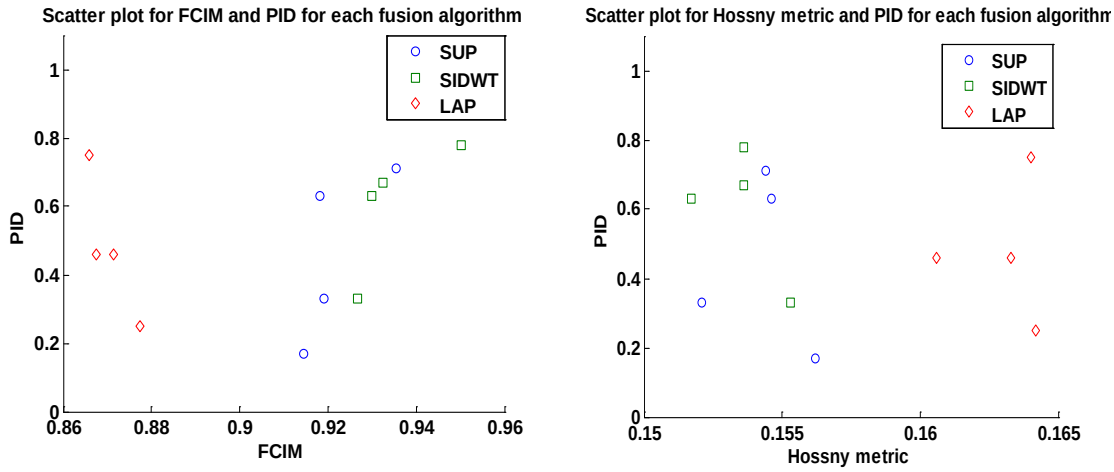


Figure 38. Scatter plots for which correlation measures were obtained between human performance and each of the metrics used in the LWIR/SWIR perception experiment. On the left are the plots for the FCIM and on the right are the plots for the Hossny metric.

APPENDIX IV

Table 11

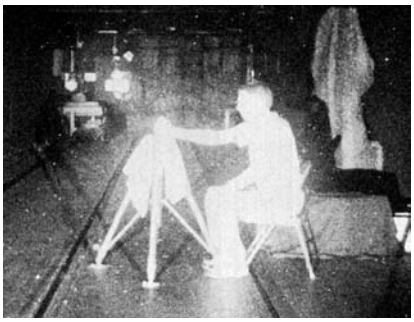
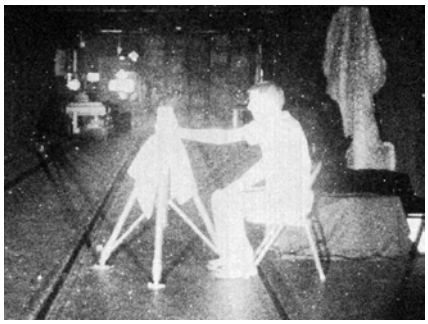
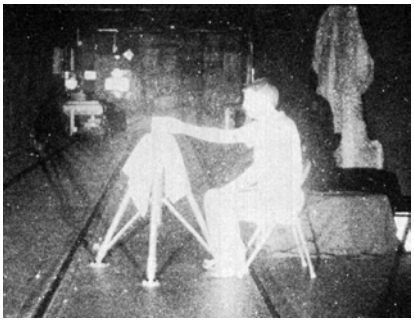
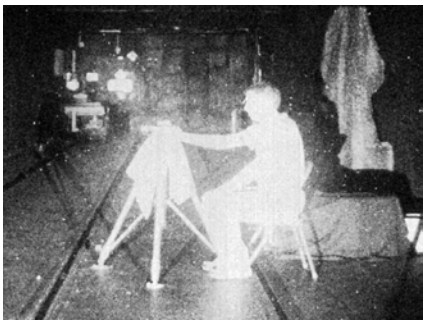
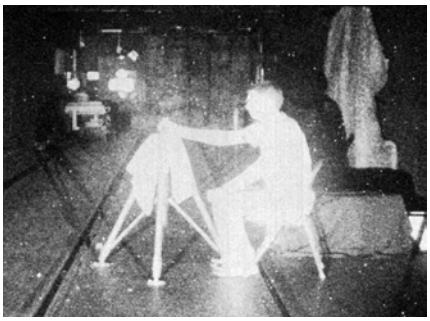
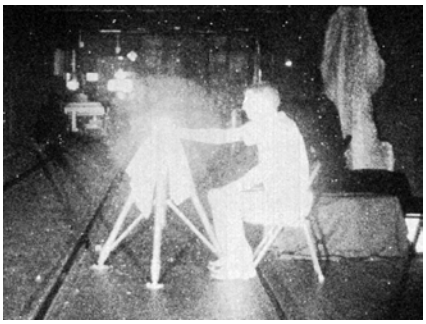
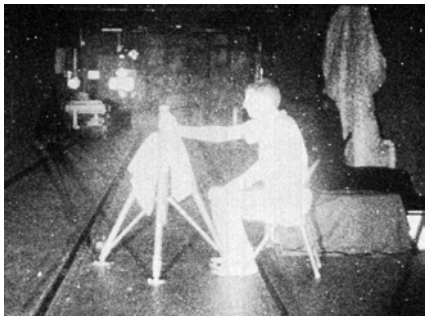
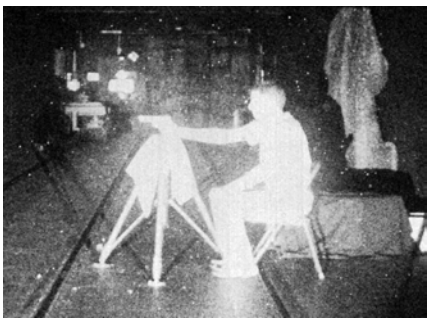
Pearson Product-Moment Correlation Coefficient Table of Critical Values. This table is taken from http://physics.mercer.edu/Younce/pearson.html

df= N-2 N = number of pairs of data	Level of significance for two-tailed test			
	.10	.05	.02	.01
1	.988	.997	.9995	.9999
2	.900	.950	.980	.990
3	.805	.878	.934	.959
4	.729	.811	.882	.917
5	.669	.754	.833	.874
6	.622	.707	.789	.834
7	.582	.666	.750	.798
8	.549	.632	.716	.765
9	.521	.602	.685	.735
10	.497	.576	.658	.708
11	.476	.553	.634	.684
12	.458	.532	.612	.661

13	.441	.514	.592	.641
14	.426	.497	.574	.628
15	.412	.482	.558	.606
16	.400	.468	.542	.590
17	.389	.456	.528	.575
18	.378	.444	.516	.561
19	.369	.433	.503	.549
20	.360	.423	.492	.537
21	.352	.413	.482	.526
22	.344	.404	.472	.515
23	.337	.396	.462	.505
24	.330	.388	.453	.495
25	.323	.381	.445	.487
26	.317	.374	.437	.479
27	.311	.367	.430	.471
28	.306	.361	.423	.463
29	.301	.355	.416	.456
30	.296	.349	.409	.449

35	.275	.325	.381	.418
40	.257	.304	.358	.393
45	.243	.288	.338	.372
50	.231	.273	.322	.354
60	.211	.250	.295	.325
70	.195	.232	.274	.302
80	.183	.217	.256	.284
90	.173	.205	.242	.267
100	.164	.195	.230	.254

APPENDIX V

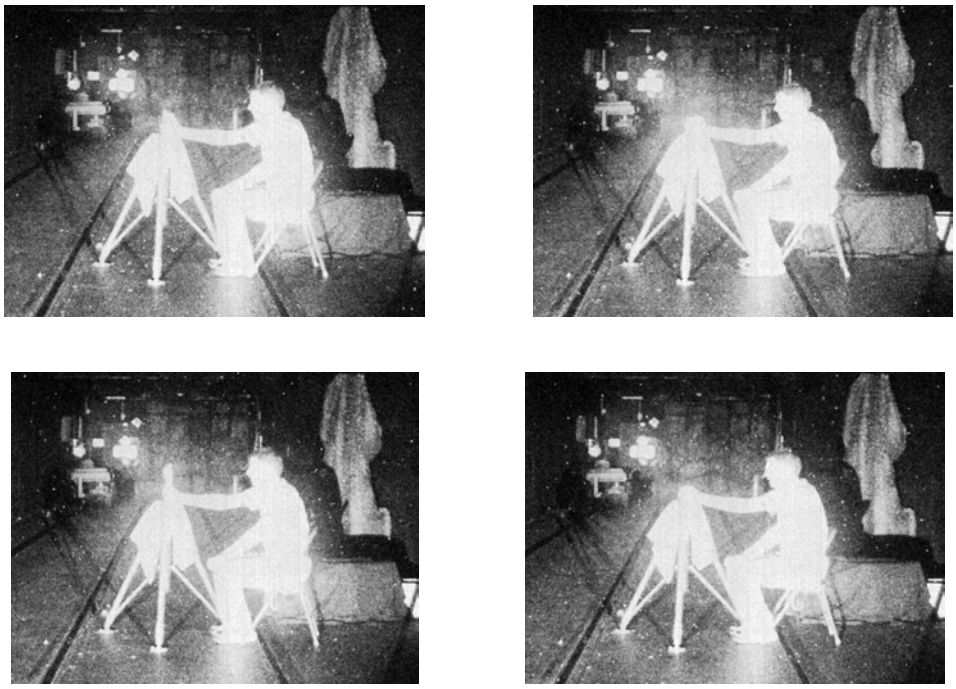
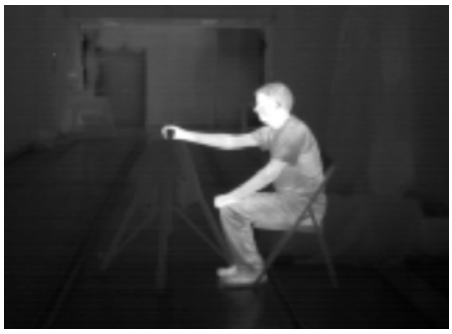


Figure 39. Image Intensifier source imagery for the hand held perception experiment. Starting from top left: Soda Can, Brick, Camera, Cell phone, Coffee Mug, Flashlight, Grenade, Gun, Knife, PDA, Radio, and Rock.







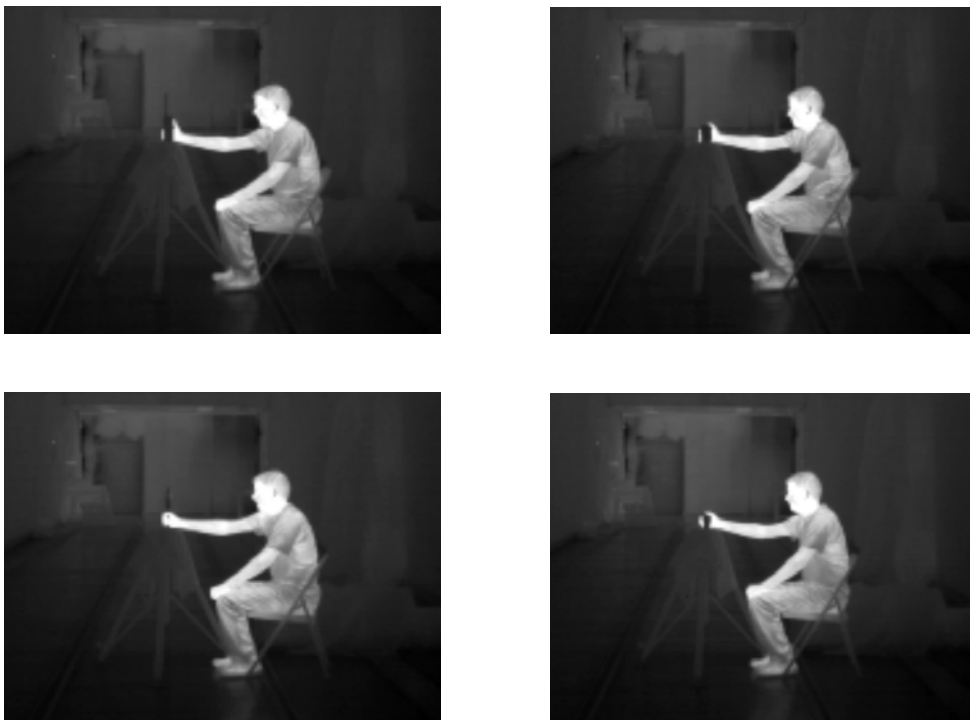



Figure 40. LWIR source imagery for the hand held perception experiment. Starting from top left: Soda Can, Brick, Camera, Cell phone, Coffee Mug, Flashlight, Grenade, Gun, Knife, PDA, Radio, and Rock.

APPENDIX VI

Preface

The following document has been prepared for submission into the journal of Optical Engineering. The contents of this document highlights research results comparing predicted human identification performance using optical systems that have single F-number optics versus optical systems that have dual F-number optics. I am the first author of this document entitled “Performance comparison between dual F-number and single F-number infrared systems”.

Abstract: Dual F-number infrared systems are under development for deployment in various Army platforms and systems in other services. A system incorporating dual F-number technology has the capability to increase system performance by allowing the use of variable field of views when performing military target acquisition tasks such as search and detection, recognition and identification. The focus of this paper is on long-wave infrared (LWIR), mid-wave infrared (MWIR), and Notional Third Generation (3rd Gen) dual band FLIR imaging systems.

In this paper, for a given set of assumptions, improvements in range performance achieved using dual F-number systems compared to single F-number systems is quantified. It has been determined that the use of a dual F-number system can provide a significant increase in overall system performance for MWIR and 3rd Gen FLIR systems. The results presented in this paper are valid for infrared target acquisition systems where the human performs the search and detection process and the identification process.

Keywords: Dual F-number, infrared system design, variable aperture, warm shield design, cold shield design

1. Introduction

The speed in which a system can gather light is described by its F-number and is ultimately determined by the effective focal length and aperture size of the optical system. In optical systems the limiting factor for the narrow field-of-view (NFOV) is commonly the diameter of the optics. In reflective spectral bands, like the visible, a variable F-number is achieved with a non-reflective adjustable iris/pupil in the optical train. In emissive or thermal imaging, an iris like that would be an infrared noise source and adversely affect sensitivity. Fixed aperture/F-number cold shields are typically implemented inside an infrared detector array's dewar. This limitation in turn determines the value of the wide field-of-view (WFOV). Historically, there is a compromise between WFOV and NFOV performance for thermal infrared systems using a single F-number, whereas a more optimal WFOV and NFOV performance can be realized using a system that incorporates dual F-number optics.

A low F-number configuration is used for WFOV search/detection and a high F-number configuration is used for NFOV long range recognition or identification. Dual F-number uses a low F-number in the WFOV and a high F-number in the NFOV. A dual F-number system has the benefit of being able to change its F-number with changing FOV, thus allowing optimization of system parameters such as sensitivity and resolution [1] which result in increased range performance. A single F-number system is limited in this regard in so much as all the FOV's are slaved off each other. Incorporating dual F-number in a system design allows the designer to optimize multiple FOV's thus realizing increases in target acquisition range performance as a result.

Some general benefits of low F-number are: 1) a larger photon flux on the detector permitting higher signal to noise or shorter integration time (less motion blur) and 2) a larger aperture in the WFOV leading to detector-limited resolution. The benefit of high F-number is the increased range due to a smaller instantaneous field-of-view (IFOV sometimes called detector angular subtense or DAS). A high F-number reduces aberrations and increases the diffraction blur diameter; which usually causes the system to be diffraction limited. Dual F-number uses low F-number in the WFOV for Search and Detection (SD) and high F-number in the NFOV for Identification (ID).

An investigation has been conducted to quantify the performance benefit of dual F-number versus single F-number. The results of this study suggest a moderate benefit is expected for LWIR band systems whereas a large benefit is expected for MWIR and 3rd Gen systems. A 3rd Gen FLIR is a dual band thermal imager comprised of a large format staring array that can simultaneously image LWIR/MWIR bands.

The main goal of this study is to quantify the performance benefits associated with utilizing a dual F-number system. Section 2.0 gives a description of system models used in this evaluation along with some underlying assumptions. Performance evaluations between range and FOV for the LWIR and MWIR bands are also given. Section 3.0 discusses the results obtained from analyzing the models in Section 2 quantifies the benefits of dual F-number system performance and provides plots and tabular presentations of results for each infrared system.

2. Implementation approaches

There are currently two approaches to the implementation of dual F-number (see Figure 1); warm shield and a cold shield design. A reflective warm shield is an optic (with a hole in the center that acts as the system stop) on a paddle that rotates into the optical path outside the detector dewar. In the cold shield approach, a variable aperture is placed in the cryogenic cold space and the aperture iris diameter is adjusted to change the system F-number. Having the ability to vary the F-number with changing FOV's and changing cold shield aperture size allows for system performance improvements.

2.1 Dual F-number warm shield

The reflective warm shield is on a paddle that moves in and out of the optical path. The warm shield is actually an aspheric reflector that reflects the cold space radiation band into the dewar so that the detector does not receive unwanted radiation from regions outside the high F-number. The left side of Figure 1 shows the two figures with and without the warm shield in place. When the shield is out of the path, a low F-number is implemented and when the warm shield is in the path, a high F-number is implemented. Since the warm shield is on a swing arm assembly, there are a number of issues such as the failure of the swing arm could block the detector, the alignment repeatability of the aperture, and the space required outside the dewar results in larger optics.

2.2 Dual F-number variable aperture

The variable aperture [3,4,5] implementation is shown on the right side of Figure 1. The variable aperture iris is actually placed in the cryogenic space of the detector dewar. The iris mechanism provides the potential for not just dual F-number, but continuous F-

number designs. Although the moving parts are subjected to extreme temperature cycles, the approach has been proven to be robust enough to be implemented by a number of sensor manufacturing companies. This approach provides compact optical design and is modular so the assembly can fit into a number of optical designs. One drawback is that the added mass in the dewar requires an increased thermal cool down time or more power (and higher capacity cooler) to cool the dewar assembly.

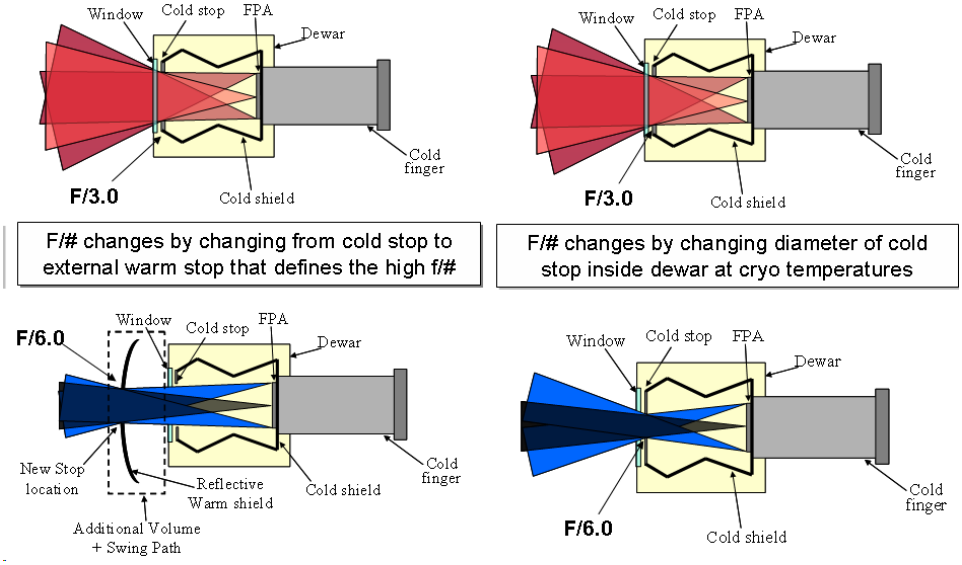


Fig. 1. Dual F-number Implementation Approaches. A more detailed comparison of the two techniques is shown in Table 1

Table 1. Comparison of Dual F-number implementation approaches

Consideration	Warm Shield	Variable Aperture
Modularity	System Specific Design	Packaged in camera system
Compactness	Requires mirror with swing path outside of dewar	Small and Self Contained in Dewar, enables camera package
Alignment	Centering of Aperture and Repeatability is an Issue	Cold stop centered to FPA, tolerances on diameter
Failure	Paddle Can Fail Covering Aperture…Sensor Fails	Failure of Aperture Still Allows Sensor Operation at One F#
Temperature	As system increases in temp, so does warm shield, impacting NUC	Temperature controlled aperture inside of dewar
Cold Shield Efficiency	Off-axis stray light can be difficult to control decreasing efficiency from nominal mirror emissivity	May be as efficient as a fixed cold stop
Stray light	Complex warm shield designs are required to control all stray light, severe penalties for errors	Stray light controlled via two bounce rule, same as existing dewars
Mechanical Design	Dependent on available volume for each independent system, not constrained to small dewar	Complex design must resist low temps, vibration, and vacuum with limited available space
Optical Design	Optics designed for 2 stop positions, space required for warm stop	No impact on optical design other than dual f/# capability
SWAP	No Increase in Power, but definite increase in size	Increase Power or Cool Down Time

3. Performance approach assessment and assumptions

The SD range and the ID range are calculated for two sensor cases: a single F-number sensor and a dual F-number sensor. The performance is also calculated for four conditions: 1) SD set with dynamic range for scene contrast at 1 kilometer, 2) SD is enhanced with Local Area Processing (LAP), 3) cold background, and 4) cold background with a small target contrast (cold target).

Search and detection tasks (low F-number configurations) are performed in the WFOV to locate potential targets. Upon detecting an object of interest in the WFOV the operator switches to the NFOV, where high F-number configurations are used, to determine if the object is indeed a target. An optical system designed with one F-number inherently suffers from trade-offs between the WFOV and NFOV performance. Traditional design of a system's NFOV is limited by the diameter of the optics which in turns limits the system's WFOV. Optical systems designed utilizing dual F-number have the potential to realize performance increases in search and detection tasks and identification tasks within the same system.

There are several sensor and target assumptions and are listed as follows. The NFOV and identification performance is limited by a 6-inch aperture (the FOV is allowed to vary). The WFOV is held constant at 10 degrees diagonal (8 degrees horizontal), so the entrance aperture size is allowed to vary and the focal length is held constant. SD range performance is modeled with an observer search pan of 3 seconds per FOV which results in some level of smear or pan blur. The detector is a 640 by 480 array with 20 micrometer square elements. The maximum charge storage is 10 million electrons and a 50% maximum well capacity is used. A 12 millisecond integration time limit is imposed

and a background temperature of 300K is imposed unless stated otherwise (as in the case with the cold weather performance).

4. Performance calculations

The performance calculations were carried out using the above assumptions and the Night Vision Electronic Sensors Directorate (NVESD) thermal sensor performance model NVThermIP [6]. NVThermIP predicts the contrast threshold function (CTF) of the system as well as predicts the target acquisition range performance achievable by the sensor.

The first performance calculation corresponds to no LAP in a mid-latitude summer environment. The visibility was 23 kilometers and the aerosol content corresponded to a rural situation. The target contrast was 1.25 Kelvin and the scene dynamic range (scene contrast temperature) was 5 Kelvin. The performance calculations shown included no turbulence and no system vibration blur. The FOV associated with the NFOV, identification mode, was allowed to vary with F-number. Recall that the WFOV was constant at 8 degrees horizontal. As the NFOV decreases, the identification range (noted in Figure 2 as ID FOV) increased until the system became diffraction-limited. The eye distance to the system display (resulting in magnification) was optimized in a manner that gave the best range for the given imager. The search and detection range (noted as SD in Figure 2) decreased with F-number due to a number of contributors to include lower signal-to-noise, and/or more smear due to a longer integration time. In Figure 2, dashed lines correspond to the LWIR and solid lines correspond to the MWIR.

One aspect that is immediately apparent from Figure 2 is the small SD range performance compared to ID range performance. The SD range performance is driven by the dynamic range limitations associated with adjusting WFOV gain and offset for temperature variations close to the sensor and then the atmospheric transmission and path radiance reducing the contrast of the target at longer ranges. This reduction in long range target contrast is difficult for the human eye to see since the eye is limited by the CTF. That is, the SD range performance is not limited by signal-to-noise. It is limited by the eye CTF, the atmospheric transmission, and the dynamic range of the image on the display. This effect does not occur with ID since the dynamic range is adjusted where the background is local to the target so that the gain and offset is optimized around the target. Sensors without LAP have these characteristics.

To compare a single F-number system to a dual F-number system, an F-number is selected for the single F-number system where both ID and SD range performances are good. We chose F/5 for the single F-number MWIR system and F/2 and F/6.5 for the dual F-number MWIR system. For the single F-number MWIR system, the SD and ID performance ranges were 3.4 and 11 km, respectively. For the dual F-number MWIR system, the SD and ID performance ranges are 4.2 and 12.2 for a 24% and 11% range increase, respectively. We could have chosen a lower F-number for the single F-number system, where the SD improvement would have been smaller, but the ID performance would have been larger. However, reducing the SD performance range increases the chance for targets slightly pass the SD distance to go undetected and therefore the operator would never switch to the NFOV for ID. As Figure 2 shows, targets above 3.4 km for SD wouldn't be detected using the single F-number MWIR system and therefore

no attempt on ID would be made; but targets between 3.4 km and 4.2 km would be detected using the dual F-number MWIR system, allowing for improved performance. A similar analogy can be carried out for the LWIR system.

For the LWIR system, we chose F/3.5 for the single F-number system and F/2 and F/5 for the dual F-number system. The single F-number SD and ID performance ranges were 4.0 km and 7.0 km, respectively. For the dual F-number system, the SD and ID performance ranges were 4.5 km and 7.7 km for 12% and 10% performance increases, respectively.

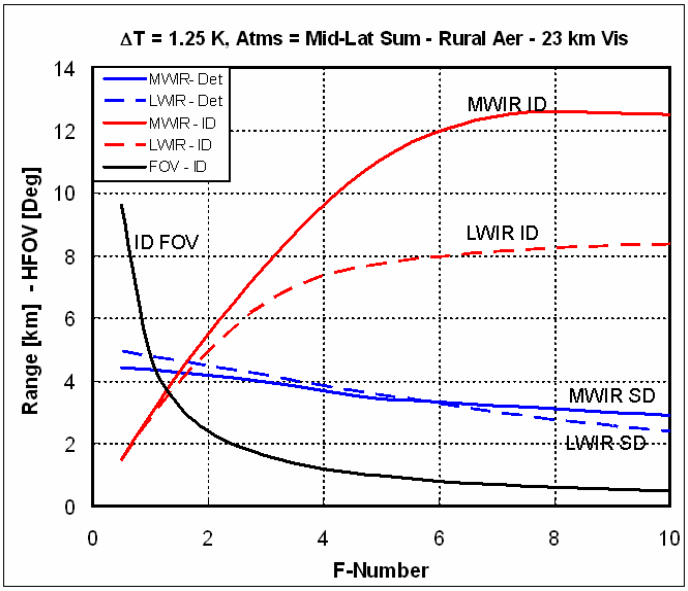


Fig. 2. Detection and Identification Range Performance as a Function of F-number. No Local Area Processing

The second performance calculation included a WFOV that was accompanied by LAP. LAP extends the performance range of SD in the WFOV through the most efficient use of the available display dynamic range. In Figure 3, the left image corresponds to a

WFOV without LAP and the right image corresponds to a WFOV with LAP. The targets are much easier to detect in the right image.

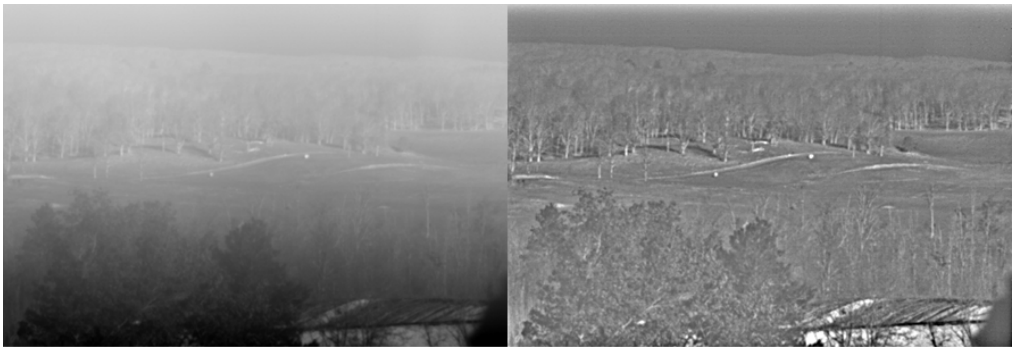


Fig. 3. WFOV without (Left) LAP and with (Right) LAP

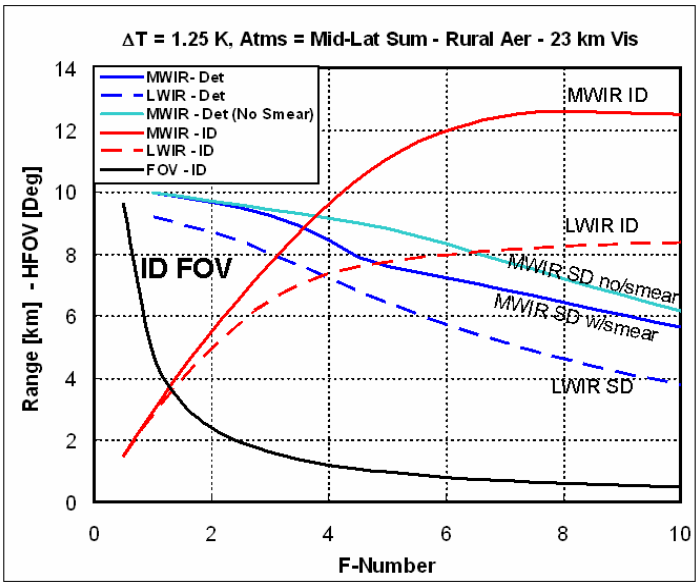


Fig. 4. Detection and Identification Range Performance as a Function of F-number. With Local Area Processing

In Figure 4, the results are provided for a system with LAP. The ID ranges are the same as those in Figure 2 since the performance of ID has a localized gain and offset. The SD range performance shows a dramatic improvement. Two MWIR SD range

curves are shown: one with no MWIR integration time smear associated with the 3 sec/FOV pan rate and one with the MWIR integration time smear. Integration time smear in the LWIR is negligible. The knee in the MWIR SD w/Smear curve is where the integration time maximum is met at 12 seconds. Again, the single MWIR F-number was selected at F/5 and the dual MWIR F-number system was selected as F/2 and F/6.5. The ranges for SD were much larger and the improvement in SD range performance was 28% where the improvement in ID range performance remained at 11%. The single LWIR F-number was selected at F/3.5 and the Dual LWIR F-number system was F/2 and F/5. The LWIR SD range performance for single F-number was 7.6 km and for dual F-number was 8.7 km for a 15% improvement. The LWIR ID range performance improvement remained at 10%.

A stressing situation with MWIR is cold weather performance and an even more stressing situation is cold weather performance with a low contrast target. MWIR is considered photon starved in cold temperatures. We provide the later calculation with a target of contrast of 0.25 Kelvin and a background temperature that ranged from 240K to 320K. LWIR is not affected significantly by cold backgrounds. Hodkins *et. al.* [7] conducted a study examining how the temperature in a scene impacts the noise in infrared sensors and their performances. A cold weather performance comparison between MWIR and LWIR is given that supports the above conclusions. Overall LWIR performance becomes better than MWIR performance when temperatures $\leq$ -24 ^{o}C (-11 ^{o}F).

Figure 5 shows the SD range performance and system noise equivalent temperature difference (NETD) as a function of F-number and background temperature. NETD is given in [7] as:

$$NETD = \frac{4 * \left(\text{F - number}\right)^2 * N_{q,Total}}{\eta * t_{\text{int}} * A_d * \tau * \pi * \int_{\Delta\lambda} \frac{\partial L_{q(\lambda,T_{BCK})}}{\partial T_{BCK}} d\lambda} \quad (1)$$

$N_{q,Total}$ is the total number of noise electrons that can be collected during integration time t_{int}. η is the detector quantum efficiency, A_d is the detector area and τ is the average optics transmission over the waveband $\Delta\lambda$. T_{BCK} is the background temperature and $\frac{\partial L_{q(\lambda,T_{BCK})}}{\partial T_{BCK}}$ is the scene photon radiance contrast evaluated at T_{BCK}. Equation 1 can be used to determine the effect F-number and integration time have on NETD.

Note that the atmospheric model in Figure 5 has been changed to a cold weather arctic type climate with a 23km visibility and rural aerosols. The background temperatures were set to 240K, 260K, 280K, 300K, and 320K (-35F, 1.4F, 37F, 73F, and 109F). We considered 240K extremely cold, 260K moderate cold, and 280K mild cold conditions. The rapid drop in detection range performance at low F-number is due to smear during the search pan rate and the knee in the drop curves correspond to the 12 millisecond integration time limit. To the right of the knee, the NETD increases and a loss in signal-to-noise occurs.

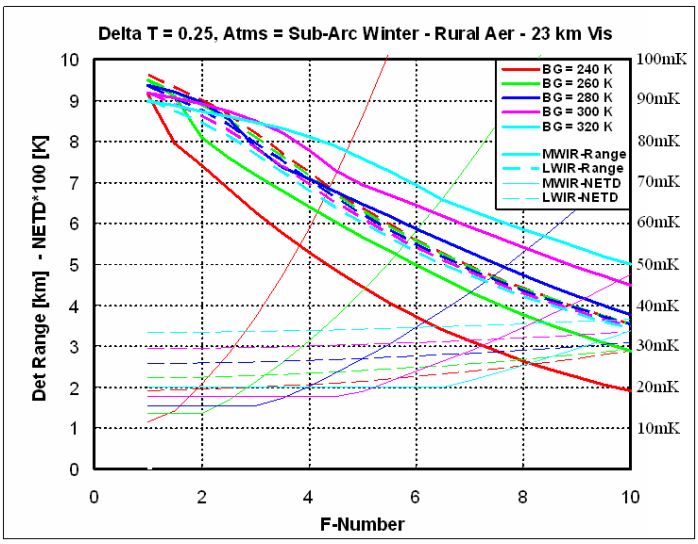


Fig. 5. Detection range and NETD as a Function of F-number and Background Temperature

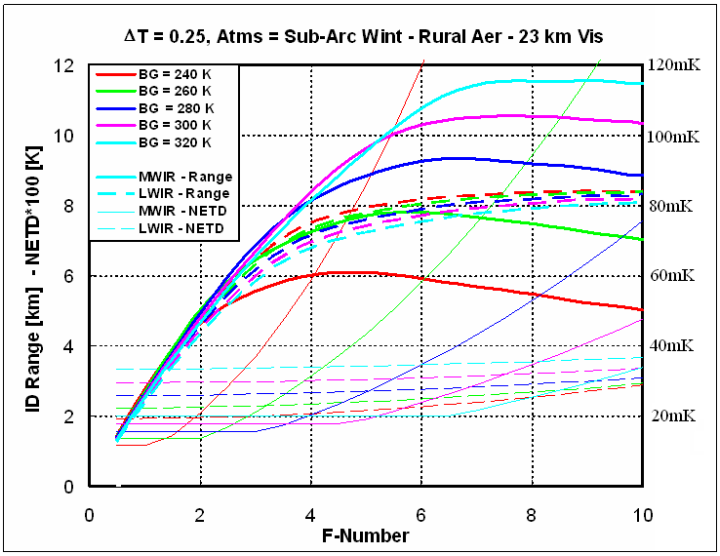


Fig. 6. Identification range and NETD as a function of F-number and background temperature

The cold weather detection calculations include LAP. The cold weather identification results are provided in Figure 6. The identification range is shown as a function of F-

number and the temperature range is varied again. Note that the identification range in the MWIR can also decrease at higher F-number (no smear is included) due to a decrease in signal-to-noise.

Using Figure 5 and Figure 6, a comparison of single F-number to dual F-number systems can be accomplished for cold weather and low contrast targets. In the MWIR, we chose a single F-number system of F/5 and a dual F-number system of F/2 and F/6.5. The single F-number SD and ID range predictions for a 240K background and 0.25K target were 4.5 km and 6.1 km. The dual F-number SD and ID range predictions a dual F-number system was 7.4 km and 5.8 km for a 64% and -5% performance increase respectively. The reduction in ID performance shows that while F/6.5 is a good choice for warmer conditions, it can result in a non-optimized system under photon-starved conditions. From the authors point-of-view, the small reduction in ID performance is worth the large gain in SD performance. In the LWIR, with a single F-number system of F/3.5 and a dual F-number system of F/2 and F/5, the SD and ID performance increase at a 240K background is 17% and 10%.

Figure 5 and Figure 6 can be used to determine the performance benefits of dual F-number at various background temperatures and for any selection of dual F-number.

4.1 General benefits of dual F-number

To compare the general performance of single F-number to dual F-number, consider the summary provided in Table 2. For single F-number, an MWIR F/5 system was chosen and a LWIR F/3.5 system was chosen for all three cases (no-LAP, LAP, and cold weather with low contrast target). This choice appeared to be a good balance between SD

performance and ID performance. A smaller F-number such as F/4 or F/4.5 could have been chosen in the MWIR, however, the dual F-number benefit to ID performance would have been smaller and the benefit to SD would have been larger. The same concept could be applied to the LWIR. Therefore, F/5 appears to be a good compromise for the MWIR and F/3.5 is a good compromise in the LWIR. The choices for dual F-number were F/2 for SD performance and F/6.5 for ID performance in the MWIR and F/2 and F/5 in the LWIR. The authors felt that the F/2 was practical with reasonable implementation. The ID performance selection of F/6.5 and F/5 provided for a diffraction-limited performance realm.

In Table 2, The SD range predictions are shown in the No-LAP/LAP/Cold Weather format in each cell. The ID range predictions are shown in the LAP/Cold Weather format since No-LAP or local automatic gain control did not make sense in the narrow field-of-view. The single F-number performance corresponds to the F/5 MWIR system and a F/3.5 LWIR system. The numbers shown (along with the corresponding F-numbers) are the ranges in kilometers that correspond to 50% probability of detection (with search) and 50% probability of identification. The bottom two rows show the range increase, in percentage, for SD performance and ID performance. The notional 3rd Gen FLIR column takes the higher of the MWIR or LWIR performance for each case of single F-number and dual F-number. The selection (MW or LW) is shown next to the range predictions. These numbers are used to determine the percentage of range performance improvement from single F-number system to dual F-number system.

Overall, the LWIR improvement (for the configuration sensor we used) is on the order of 15% improvement in search and detection performance (mainly due to an improved

MTF at lower F-number) and a 10% improvement for identification (mainly due to a higher diffraction exploitation). The MWIR improvement ranges from 24% to 64% for search and detection and ranges from -5% to 11% for identification. Notional Third Gen FLIR improvement ranges from 22% to 32% for search and detection and from 7% to 18% for identification.

Table 2. Performance Comparison of Single F-number and Dual F-number

Det: No LAP / LAP / Cold ID: LAP / Cold(240K)	LWIR	MWIR	Notional 3rd Gen
Single F# SD Km	4.0 / 7.6 / 7.7 F/3.5	3.4 / 7.6 / 4.5 F/5.0	3.7 LW / 7.9 MW / 6.8 LW F/4.5
Single F# ID Km	7.0 / 7.0 F/3.5	11.0 / 6.1 F/5.0	10.3 MW / 7.7 LW F/4.5
Dual F# SD Km	4.5 / 8.7 / 9.0 F/2	4.2 / 9.7 / 7.4 F/2	4.5 LW / 9.7 MW / 9.0 LW F/2
Dual F# ID Km	7.7 / 7.7 F/5	12.2 / 5.8 F/6.5	12.2 MW / 8.2 LW F/6.5
SD Improve	12% / 15% / 17%	24% / 28% / 64%	22% / 23% / 32%
ID Improve	10% / 10%	11% / -5%	18% / 7%

5. Summary

The performance benefit of Dual F-number (versus single F-number) for infrared target acquisition systems is summarized in Figure 7 (for the configuration described in this paper). The left chart is for search and detection performance and the right chart is for identification performance. Blue bars are LWIR, red bars are MWIR, and white bars are notional 3rd Gen FLIR. The search and detection benefit shows no-LAP, LAP, and cold weather with low contrast target performance. The identification benefit shows LAP and cold weather with low contrast target performance.

Dual F-number can provide a significant increase in overall system performance for MWIR and Third Gen FLIR systems. A moderate performance benefit can be seen in the LWIR. In addition, LAP is extremely important for SD performance. The results presented in this paper are valid for infrared target acquisition systems where the human performs the search and detection process and the identification process. Other assumptions, such as a rapid scan WFOV with Aided Target Detection, would provide a larger benefit for dual F-number in the MWIR. Finally, under cold conditions MWIR detectors require longer integration times than LWIR to achieve good NETD. MWIR systems can be degraded, even with dual F-number systems, if the system is not optimized for cold weather performance.

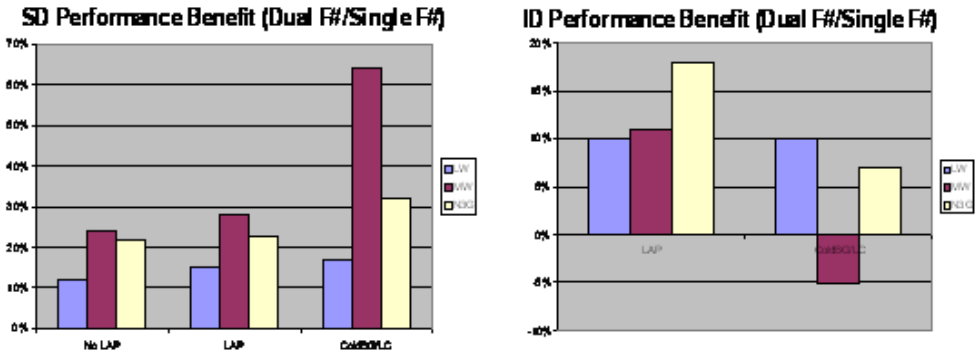


Fig. 7. Performance Benefit of Dual F-number

References

1. J. Vizgaitis, “Dual f/number optics for 3rd generation FLIR systems,” *SPIE Proc.* **Vol. 5783**, (2005).

2. N. Gat, J. Zhang, M.D. Li, L. Chen, and H. Gurrola, "Variable Cold Stop for Matching IR Cameras to Multiple f-number Optics ." *SPIE Proc*, **Vol. 6542**, April 2007.

3. N. Gat, et-al, “Cryogenically cooled adjustable apertures for infrared cameras”, US Patent 7157706

4. M. Shoda, T. Ishizuya, “Thermal infrared camera”, US Patent 6133569

5. R. Chipper, “Infrared zoom lens assembly having a variable aperture”, US Patent 5796514.

6. “Night Vision Thermal and Image Processing Performance Model User’s Manual,” *U.S. Army RDECOM CERDEC, NVESD,* **Revision 8**, March 2005.

7. V. Hodgkin, B. Kowalewski, D. Tomkinson, B. Teaney, T. Corbin and R. Driggers, “Modeling of IR Sensor Performance in Cold Weather”, *SPIE Proc*, **Vol 6207**, 620708, (2006)