



Fusing Infrared and Visible Imageries for Improved Tracking of Moving Targets

by Stephen R. Schnelle and Alex L. Chan

ARL-TR-5552

July 2011

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

Army Research Laboratory

Adelphi, MD 20783-1197

ARL-TR-5552

July 2011

Fusing Infrared and Visible Imageries for Improved Tracking of Moving Targets

Stephen R. Schnelle and Alex L. Chan
Sensors and Electron Devices Directorate, ARL

REPORT DOCUMENTATION PAGE				Form Approved OMB No. 0704-0188	
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.					
1. REPORT DATE (DD-MM-YYYY) July 2011		2. REPORT TYPE		3. DATES COVERED (From - To) May 2010 to January 2011	
4. TITLE AND SUBTITLE Fusing Infrared and Visible Imageries for Improved Tracking of Moving Targets				5a. CONTRACT NUMBER	
				5b. GRANT NUMBER	
				5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S) Stephen R. Schnelle and Alex L. Chan				5d. PROJECT NUMBER	
				5e. TASK NUMBER	
				5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) U.S. Army Research Laboratory ATTN: RDRL-SES-E 2800 Powder Mill Road Adelphi, MD 20783-1197				8. PERFORMING ORGANIZATION REPORT NUMBER ARL-TR-5552	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)				10. SPONSOR/MONITOR'S ACRONYM(S)	
				11. SPONSOR/MONITOR'S REPORT NUMBER(S)	
12. DISTRIBUTION/AVAILABILITY STATEMENT Approved for public release; distribution unlimited.					
13. SUPPLEMENTARY NOTES					
14. ABSTRACT Video surveillance is an important tool for force protection and law enforcement, and visible and infrared video cameras are the most common imaging sensors used for this purpose. In this report, we present a feasibility study on fusing concurrent visible and infrared imageries to improve the tracking performance of an existing video surveillance system. Image fusion was performed using 13 pixel-based image fusion algorithms, including four simple-combination methods and nine pyramid-based methods. The effects of all 13 algorithms on the detection and tracking performance of a given target tracker were examined. Five of the pyramid-based methods were shown to provide superior performance enhancements, three of which also managed to achieve it with relatively low computational costs.					
15. SUBJECT TERMS Image fusion, infrared imagery, visible imagery, target detection, target tracking					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT UU	18. NUMBER OF PAGES 40	19a. NAME OF RESPONSIBLE PERSON Alex Lipchen Chan
a. REPORT Unclassified	b. ABSTRACT Unclassified	c. THIS PAGE Unclassified			19b. TELEPHONE NUMBER (Include area code) (301) 394-1677

Contents

List of Figures	iv
List of Tables	v
1. Introduction	1
2. Image Fusion	5
2.1 Image Registration	5
2.2 Fusion Methods	6
2.2.1 Simple Combinations	7
2.2.2 Pyramid Structures	9
3. FPSS Tracker	14
3.1 Background Modeling.....	14
3.2 Target Detection and Tracking.....	17
4. Experimental Results	18
5. Conclusions	26
6. References	28
List of Symbols, Abbreviations, and Acronyms	30
Distribution List	31

List of Figures

Figure 1. SPOD manufactured by the FLIR Systems.	2
Figure 2. The hierarchy of fusion methods.	3
Figure 3. Extracting edge features from a pair of images using a Sobel edge filter.	3
Figure 4. SIFT on consecutive visible image frames showing strong matches.	6
Figure 5. SIFT on visible and IR image frames showing few and inaccurate matches.	6
Figure 6. Example of a visible (left) and an LWIR (right) image in the FPSS dataset.	7
Figure 7. Fused image through simple average (left) and principal component analysis (PCA)-weighted average (right).	8
Figure 8. Fusion by selecting maximum (left) and minimum (right) pixel intensities.	9
Figure 9. Fusion by selecting the maximum coefficient of Laplacian pyramids (left) and filter-subtract-decimate (FSD) pyramids (right).	10
Figure 10. Fusion by selecting the maximum coefficient of the ROLP pyramids (left) and contrast pyramids (right).	11
Figure 11. Fusion by selecting the maximum coefficient of the gradient pyramids (left) and morphological pyramids (right).	12
Figure 12. Fusion by selecting the maximum coefficient of DWT pyramids (top-left), SIDWT pyramids (top-right), and DT-CWT pyramids (bottom).	13
Figure 13. The background modeling and subtraction process in FPSS tracker.	14
Figure 14. Enhancement of target signatures and suppression of trailing effects and noises via a DPI.	16
Figure 15. Human LWIR signatures reverse polarity in winter (left) and summer (right).	16
Figure 16. The FPSS graphical user interface.	18
Figure 17. Partial content of a typical ground truth file in the FPSS dataset.	19
Figure 18. The performance of four simple-combination methods at low FAR region.	20
Figure 19. The performance of four simple-combination methods at high FAR region.	21
Figure 20. The performance of four inferior pyramid-based fusion methods at low FAR region.	23
Figure 21. The performance of four inferior pyramid-based fusion methods at high FAR region.	23
Figure 22. The performance of five superior pyramid-based fusion methods at low FAR region.	24
Figure 23. The performance of five superior pyramid-based fusion methods at high FAR region.	25

List of Tables

Table 1. CPU time (seconds) needed to fuse 30 images using Matlab code on a Dell T7400 workstation.....	22
Table 2. Performance (hit rate in %/FA per frame) of the 13 fusion methods at low FAR region.	24
Table 3. Performance (hit rate in % / FA per frame) of the 13 fusion methods at high FAR region.	26

INTENTIONALLY LEFT BLANK.

1. Introduction

A picture may be worth a thousand words, but a video sequence may be worth even more. As sensor technology, network communication, computing power, and digital storage capacity have all dramatically improved, still and video imageries have become the most common and versatile forms of media for capturing, analyzing, and disseminating a variety of information. In many scenarios, useful information is derived from the accurate detection, tracking, and recognition of certain targets of interest in a timely manner. Typical applications of this nature include aerial reconnaissance, automatic target recognition, and force protection surveillance systems.

Unfortunately, many of these applications involve monitoring adversarial activity in less than ideal environments, which can be particularly challenging to the imaging systems involved. Visible cameras are the prevailing imaging sensors because they are relatively cheap, easy to use, and capable of producing high-quality imagery under favorable conditions. However, visible cameras can be severely affected by common environmental factors, such as darkness, shadows, fog, clouds, rain, snow, and smoke. Infrared (IR) imaging systems may overcome or alleviate some of these problems, but they are subjected to a number of limitations of their own. IR-specific difficulties include a much lower sensor resolution; drastic diurnal and seasonal changes in target signatures; total loss of non-thermal but important visual features (such as color and text); blockage by visually-transparent thermal signal shields (such as car windshields and glass doors); very low thermal contrast between targets and background under certain combinations of ambient and target temperatures; and much higher costs for purchasing and maintaining the systems. Due to these highly complementary strengths and limitations of visible and IR cameras, more advanced target detection and tracking systems may want to acquire and process both visible and IR imageries concurrently and jointly for critical applications.

To study the usefulness of fusing visible and IR imagery for detecting and tracking moving targets, we have relied on a large collection of concurrent color visible and long-wave IR (LWIR) video sequences that is officially called the Second Dataset of the Force Protection Surveillance System (FPSS) by the data collector (1). These FPSS video sequences were collected using the Sentry Personnel Observation Device (SPOD) manufactured by forward-looking infrared (radar) (FLIR) Systems. As shown in figure 1, the SPOD includes a LWIR microbolometer and a color visible charge-coupled device camera. The LWIR images were acquired with a focal plane array (FPA) of 320 x 240 pixels in resolution, while the color visible images were captured at the resolution of 460 National Television Standards Committee (NTSC) TV lines.



Figure 1. SPOD manufactured by the FLIR Systems.

Both the original color visible and LWIR images were cropped and scaled to attain a coarse level of co-registration between the corresponding color-LWIR images captured at any given time. The image registration step was necessary because the color and LWIR cameras of the SPOD were merely bore-sighted into a ruggedized enclosure. They did not share a common optical lens, having slightly different lines of sight, fields-of-view, and image resolutions. Since the translational shift between these two cameras was only a few inches, while the typical ranges to the targets in the FPSS dataset were 50–200 yards, it was deemed to be acceptable to register the images using only a simple affine transformation, instead of the more general but complicated planar projection method.

Image registering can be done automatically or manually. Although automatic registration is quite accurate and feasible for images of similar electromagnetic spectrum, registering color and LWIR images is a very difficult task. The effects of automatic and hybrid registration schemes were explored by Hines et al., but automatic registration was generally not successful (2). Due to these difficulties, the FPSS dataset was coarsely registered by first manually choosing a large number of salient corresponding markers in many representative pairs of color-LWIR images. The coordinates of these markers were then used to derive the affine transformation between the color and LWIR images through a polynomial fitting process. The maximal usable area could be extracted after applying the affine transformation, and avoiding sensor artifacts in both color and LWIR images. Because the target ranges in FPSS sequences were consistently and immensely larger than the distance between the color and LWIR cameras in SPOD, the same affine transformation and clipping mechanism were used throughout the entire dataset without

introducing additional distortions. The image patches clipped from the original color and LWIR images were scaled to a common image size of 640 x 480 pixels and stored in JPEG format.

As shown in figure 2, image fusion can be handled at several different levels (3). At the lowest levels, the raw image data can be fused. This can either be performed on the original signal or, more likely, after the image has been preprocessed and the resulting pixel values are used. Pixel-level fusion is very common due to its simplicity and universality, and it is the focus of this work as well.

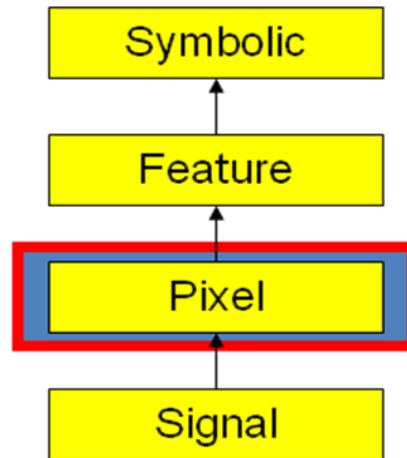


Figure 2. The hierarchy of fusion methods.

At higher levels, feature-based detection uses structural image characteristics, such as edges and corners, to enhance the image. For example, one could extract the edge information from a pair of images using Sobel filter (see figure 3) and fuse the images based on the edge information. However, this approach is much more application-specific, often requiring an understanding of the image itself, either through direct human intervention or automatic object classification algorithms. Therefore, this approach requires much more complex computation or non-real-time intervention. Training these systems appropriately can also be quite challenging.

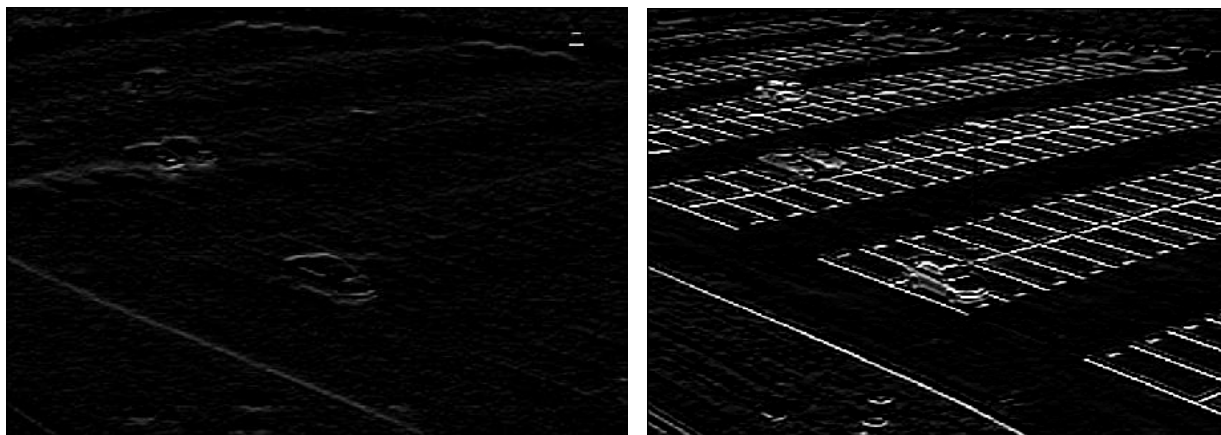


Figure 3. Extracting edge features from a pair of images using a Sobel edge filter.

One example of a higher level fusion system uses Bayesian analysis to sum the probabilities of detected human silhouettes falling within each pair of visible and infrared images. Oftentimes, detections are based on whether the probability exceeds a pre-defined threshold (4). For a stationary camera installed in a specific setting, training such a system may be feasible because its background does not vary significantly. At the highest level of image fusion, symbolic fusion methods are often heavily rule-based and rely on a lot of prior or external knowledge to perform the image fusion. Nonetheless, symbolic image fusion methods carry similar tradeoffs as the fusion methods at the feature-level.

There are many ways to measure performance of image fusion algorithms, including subjective analysis, complex similarity metrics, signal-to-noise ratio (SNR), and tracking performance. Motwani et al. suggested parameters for subjective analysis, but they concluded that subjective measures were not particularly helpful for tracking systems, except in the case of incorporating human feedback into the detection loop (5).

Cvejic et al. discussed a number of objective similarity metrics, including the Piella metric, Petrovic metric, and Bristol metric (6). The Piella metric measures structured similarity (which is based on luminance, contrast, and structure information) over local window regions and then averages these similarity measures over all windows. Weighting is given to the relative importance of each input image toward the fused image, window by window. The Petrovic metric specifically evaluates edge structure (using a Sobel edge operator) by determining the strength of edge information retained from each of the original images in the fused image. The Bristol metric, in contrast to the Piella metric, uses a slightly different weighting scheme based on the ratio of covariances between the original and fused images.

Cvejic et al. compared the tracking performance of a particle filter based on these objective metrics and found that the tracking performance was actually worsened by the fusion of images. Mihaylova et al., of the same research group, later adopted a performance metric of normalized overlapping ground truth and tracking system bounding boxes in their work (7). Their results showed that IR images alone performed just as well or better than most fusion algorithms (including contrast pyramid, dual-tree complex wavelet transform, and discrete wavelet transform) in tracking, while visible spectrum images lagged behind under harsher conditions like occlusions.

There are many possible methods of tracking a moving target, including background subtraction, optical flow, moving energy, and temporal differencing. Because the FPSS dataset was collected with a stationary SPOD with minimal background interference, we decided to use an existing FPSS tracker, which is based on background subtraction method, to examine the tracking performance of various image fusion methods (8). Instead of the FPSS tracker, one of many other moving target tracking algorithms can be used as well. For instance, Trucco and Plakas described a wide range of alternative tracking algorithms in their paper (9).

In the following section, we provide brief discussions on the image registration and 13 image fusion methods of interest. These fusion methods fall into two broad categories, namely, simple combination and pyramid structure. A brief description of the FPSS tracker is provided in section 3, while the experimental results on the tracking performance of various image fusion methods are presented in section 4. Finally, some concluding thoughts are given in section 5.

2. Image Fusion

2.1 Image Registration

Image registration is a key aspect of any image fusion algorithm. Initially, the color-LWIR images in the FPSS dataset were only coarsely-registered using a global affine transformation. In order to verify the accuracy of the existing registration and simultaneously test the quality of automatic registration algorithms applied to these coarsely-registered images, the scale-invariant feature transform (SIFT) was run on some pairs of FPSS color-LWIR images (10).

SIFT decomposes images into features for comparison and association purposes, testing local extreme features over a wide range of scales and orientations to determine the proper transformation control points for an image. Only high contrast points appearing from the difference of Gaussians are retained for robustness. A best-bin-first-search algorithm is used to select matches against previous key points, as new images are added to the set. Additional processing techniques, such as cluster identification, model verification using least squares, and outlier detection, can be included in SIFT to increase its robustness.

SIFT was chosen here due to its relative insensitivity to illumination changes and occlusions as compared to other image registration algorithms. This property is especially important for the FPSS data due to huge variations in intensity between color and LWIR images in this dataset. Unfortunately, even SIFT was incapable of registering the FPSS color and LWIR images due to the lack of corresponding key points in these images. Many salient features in the images appear to be complementary in nature, which greatly confuses the match selection in SIFT.

To illustrate this problem, a pair of visible spectrum images (which were converted from color to grayscale for efficiency reasons) from two consecutive frames of a scene are placed side by side in figure 4. As shown by the many straight lines connecting these two images, many matching key points were found by SIFT, which can be used to determine the proper transform required for the registration of these images. When the two visible images are perfectly aligned, the lines are actually connecting the corresponding key points on these images at identical coordinates. (In the case of the FPSS data set with a stationary camera, consecutive frames should already be aligned, and this is apparent in figure 4.)

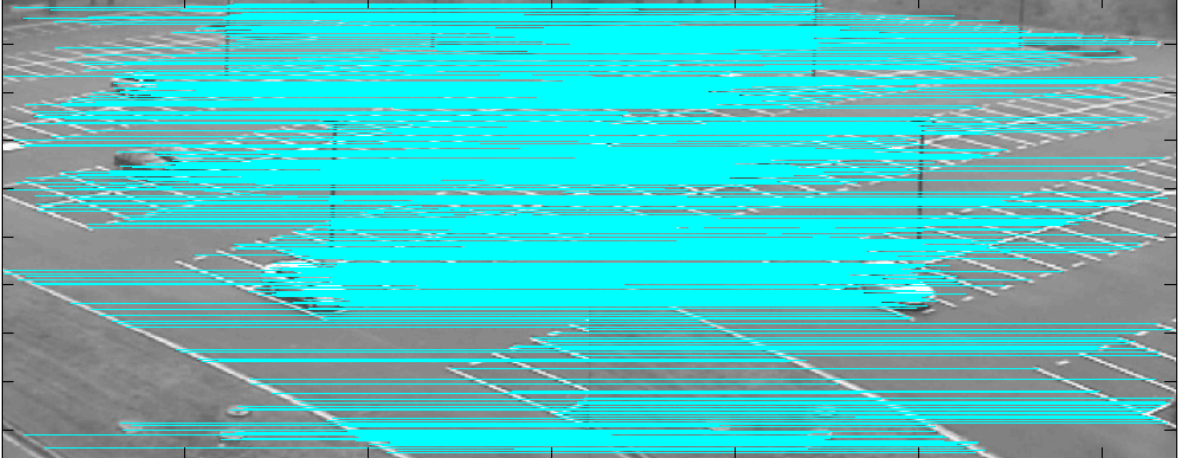


Figure 4. SIFT on consecutive visible image frames showing strong matches.

On the other hand, given a pair of visible and LWIR images from the same scene, SIFT could barely find any matching key points. As shown in figure 5, even the few potential matches suggested by SIFT were all incorrect ones. SIFT was also tried on images pre-filtered by a Sobel filter to emphasize the edge information, but the results were equally unsuccessful. Because the coarse registration of FPSS dataset was deemed visually acceptable, we proceeded on the image fusion work without further pursuing the automatic image registration route.



Figure 5. SIFT on visible and IR image frames showing few and inaccurate matches.

2.2 Fusion Methods

In this work, we focus on 13 pixel-level image fusion methods, ranging from the simplest pixels averaging method to the very complicated dual-tree complex wavelet transform. There are other interesting but less popular image fusion algorithms, including one that relies on factorizing an image (matrix) $\mathbf{V}$ into two non-negative matrix components, $\mathbf{W}$ and $\mathbf{H}$, with $\mathbf{W}$ representing a basis optimized for representing $\mathbf{V}$ (11). Another approach to image fusion is to use training sets

and classifiers, as explored by Chan et al. (12). In the work that follows, however, we assume no prior training data are available.

To evaluate the image fusion algorithms examined here, we used all FPSS coarsely-registered color and LWIR images as input data, a pair of which is shown in figure 6. To allow fusion with LWIR images, the color (RGB) images were converted to grayscale using a simple weighting of $0.2989R + 0.5870G + 0.1140B$, which yielded the intensity value but removed the hue and saturation information. For many automatic target detection and tracking algorithms, it is indeed more efficient to process grayscale images internally, while providing color outputs for human consumption only. The grayscale visible and LWIR images were manipulated using MATLAB functions to produce various fused images (13, 14).



Figure 6. Example of a visible (left) and an LWIR (right) image in the FPSS dataset.

2.2.1 Simple Combinations

The most intuitive pixel-level fusion methods examined here are simple averaging, intelligent weighting, and selecting maximum or minimum pixel values. All these methods involve only simple pixel operations, which require traversing the two input images to be fused pixel-by-pixel, leading to a simple $O(m \times n)$ operations for an image of size $m \times n$. Pixels $(\mathbf{I}_1)_{ij}$ and $(\mathbf{I}_2)_{ij}$ in images $\mathbf{I}_1$ and $\mathbf{I}_2$ need only be compared against each other.

In the first fusion method, a fused image was generated through simple averaging by calculating $(\mathbf{I}_f)_{ij} = [(\mathbf{I}_1)_{ij} + (\mathbf{I}_2)_{ij}]/2$, and the resulting fused image is shown in figure 7 (left). Because the visible and LWIR images have differing resolutions and salient features, this method tends to muddle the details.



Figure 7. Fused image through simple average (left) and principal component analysis (PCA)-weighted average (right).

We can attempt to boost the influence of the better image by using the PCA derived from the covariance matrix between the two input images. A simple way to do this is to consider each image as a single vector $\mathbf{I}_1$ and $\mathbf{I}_2$, creating a 2×2 covariance matrix when we compute the covariance of $[\mathbf{I}_1 \ \mathbf{I}_2]$. The normalized eigenvector for the larger eigenvalue provides the weighting to be used: $(\mathbf{I}_f)_{ij} = (\mathbf{v}_k)_1(\mathbf{I}_1)_{ij} + (\mathbf{v}_k)_2(\mathbf{I}_2)_{ij}$, where $\mathbf{v}_k$ represents the eigenvector corresponding to λ_k , the larger one of the two eigenvalues. Generally, the PCA-weighted averaging method strongly favors the image with the highest variance, which may or may not contain more informative and useful details. In fact, this selection criterion can be a disadvantageous one when dealing with noisy images. As shown in figure 7 (right), the fused image produced by this method closely matches the original visible spectrum image because the visible image has more details and a higher variance.

Choosing the maximum pixel value, $(\mathbf{I}_f)_{ij} = \max[(\mathbf{I}_1)_{ij}, (\mathbf{I}_2)_{ij}]$, from a pair of LWIR and visible images, as shown in figure 8 (left), may be appropriate to find some hidden targets. A man may be occluded in the visible spectrum, for example, but he can still be located in the LWIR image. For a background subtraction method, it may be desirable to boost the relative intensity of targets through this fusion method, if these targets tend to be brighter than their immediate background.



Figure 8. Fusion by selecting maximum (left) and minimum (right) pixel intensities.

Choosing the minimum pixel value, $(\mathbf{I}_f)_{ij} = \min[(\mathbf{I}_1)_{ij}, (\mathbf{I}_2)_{ij}]$, may not be very useful in general because it tends to deemphasize the strong foreground objects, as evident from figure 8 (right). In some rare occasions, this method may be helpful in extracting weak targets (with both weak but detectable visible and LWIR signatures) from busy backgrounds by deemphasizing stronger and brighter neighboring background pixels.

2.2.2 Pyramid Structures

Pyramid decompositions were introduced by Burt and Adelson in 1983 as a compact encoding scheme (15). The original idea is that a Gaussian kernel (low-pass filter) is applied to the top-level image of a pyramid, $\mathbf{I}_1 * \mathbf{G}_1$, representing the convolution of the image $\mathbf{I}_1$ with a Gaussian blurring matrix $\mathbf{G}_1$. This image is then down-sampled to form the next level of these pyramids. The difference between the low-pass version and its previous-level image represents the high frequency or detail information of the previous-level image. At each step down the pyramid, we continue to filter and down-sample in the same manner. A Laplacian pyramid is formed by computing the difference between each level of the pyramid, iteratively separating an image into low and high frequency components, except that the lowest level contains the remaining low-frequency information.

Since each level is a down-sampled version of the previous level, we need to up-sample and interpolate the decimated version in order to compute the difference between the two adjacent levels. For example, the Laplacian image at level k of $\mathbf{I}_m$, denoted as $(\mathbf{L}_m)_k$, can be computed as $(\mathbf{L}_m)_k = (\mathbf{I}_m)_k - f_{k+1}((\mathbf{I}_m)_{k+1})$, where $f_{k+1}(\cdot)$ denotes the function consisting of up-sampling and an interpolation filter with similar blurring response as $\mathbf{G}_k$, while k denotes the level of decomposition. As we proceed down the pyramid, $(\mathbf{I}_m)_k$ denotes the blurred and decimated version of $(\mathbf{I}_m)_{k-1}$. By decomposing each set of the original LWIR and visible images, we form compact representations separated into detail and approximation information. Hence, we can then weight the coefficients in each pyramid. To reconstruct the fused image, we then reverse the decomposition process, starting with a synthesis image at level $k+1$, denoted as $(\mathbf{S}_m)_{k+1}$,

expanding it, and adding it to $(\mathbf{L}_m)_k$ to get $(\mathbf{S}_m)_k$. The initial synthesis image is the background coefficients found at the bottom of the Laplacian pyramid. If we select the maximum coefficients between the two pyramids by taking $\max[(\mathbf{L}_1)_k, (\mathbf{L}_2)_k]$ for each level k and all ij coefficients during this reconstruction process, then a Laplacian fused image is generated (see figure 9 (left)). We could also modify the selection criteria of the algorithm during this reconstruction phase, such as using additional information from neighboring coefficients.



Figure 9. Fusion by selecting the maximum coefficient of Laplacian pyramids (left) and filter-subtract-decimate (FSD) pyramids (right).

Instead of using the maximum coefficients at the lowest level of the pyramid, we may choose to use the LWIR image, the visible image, or a combination of the two at the lowest level, as well. If we choose the lowest level LWIR image, this implies that the background for the fused image is built on the LWIR image, and detail information from the visible image is only included when these details outweigh those of the LWIR. The Laplacian pyramid is a simple decomposition scheme, which assumes very little information about the structure of the image. Implementation details of the Laplacian pyramid include the handling of border effects and ensuring that the image size is a factor of two at each level of decomposition.

A FSD pyramid is similar to the Laplacian pyramid, but the levels are subtracted prior to decimations. This makes the method simpler and reduces delay, therefore, allowing easier real-time implementation. Slight frequency distortions are introduced, thus a correction factor is required for perfect reconstruction. This term can be dropped in practice, though variations can make minor adjustments in the synthesis phase to account for this. Figure 9 (right) shows the result of image fusion based on the original FSD technique proposed by Anderson (16). Both images in figure 9 may look quite similar, except for a slight shading difference, but their differences in tracking performance could be larger than that.

Ratio-of-low-pass (ROLP) pyramid and contrast pyramid use the ratio of levels of the Gaussian pyramid to compute the coefficients at the next level, instead of their differences (17, 18). Otherwise, the decomposition process resembles that of the Laplacian pyramid. Since the stored

coefficients are not used to compute levels of the Gaussian pyramids, the underlying Gaussian pyramid decomposition of the image does not change. The primary difference between ROLP and contrast pyramids is the use of a local background to normalize the ratio. The contrast pyramid computes $(\mathbf{L}_m)_k = [(\mathbf{I}_m)_k / f_{k+1}((\mathbf{I}_m)_{k+1})] - 1$, and the offset of 1 is reversed during reconstruction, whereas the ROLP pyramid computes $(\mathbf{L}_m)_k = (\mathbf{I}_m)_k / f_{k+1}((\mathbf{I}_m)_{k-1})$. Instead of summing coefficients during synthesis (as in the case of Laplacian pyramid), we now reverse-decomposition by expanding $(\mathbf{S}_m)_{k+1}$ and multiplying it with $(\mathbf{L}_m)_k$ to get $(\mathbf{S}_m)_k$. Note that a small epsilon factor can be added to the denominator to prevent division-by-zero issues. Figure 10 shows the resulting fused images from the ROLP and contrast pyramid methods. These decomposition methods are designed to emphasize the contrast in an image. If we are limited in precision due to quantization issues, the contrast and ROLP pyramid decompositions will be much less accurate than the Laplacian pyramid decomposition.



Figure 10. Fusion by selecting the maximum coefficient of the ROLP pyramids (left) and contrast pyramids (right).

The gradient pyramid chooses the largest directional derivative in each of four directions: horizontal, vertical, and the two diagonal directions (19). These derivatives can be computed using simple matrix operators. For example, at each level of the pyramid, the four operators

$$\begin{bmatrix} 1 & & \\ -2 & & \\ & 1 & \end{bmatrix}, \begin{bmatrix} 0 & 0 & 1/2 \\ 0 & -1 & 0 \\ 1/2 & 0 & 0 \end{bmatrix}, \text{ and } \begin{bmatrix} 1/2 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1/2 \end{bmatrix} \text{ can be convolved with a filtered}$$

image $(\mathbf{I}_m)_k + (\mathbf{I}_m)_k * \begin{bmatrix} 1/16 & 1/8 & 1/16 \\ 1/8 & 1/4 & 1/8 \\ 1/16 & 1/8 & 1/16 \end{bmatrix}$, where $*$ represents the convolution operator.

Coefficients are selected for each of the four directions independently during the fusion process and then added together to represent the combined gradient strength at a given pixel location. A synthesized image is reconstructed using the same procedure as in the Laplacian pyramid case. An example of fused image produced by the gradient pyramid method is shown in figure 11

(left). These methods are designed to preserve orientation information, which can be useful in some applications.



Figure 11. Fusion by selecting the maximum coefficient of the gradient pyramids (left) and morphological pyramids (right).

Morphological operations, such as opening and closing, can be applied to the Gaussian pyramid without harmful effects under certain circumstances and result in a morphological pyramid (20). For example, we can apply the following operations to compute the next set of coefficients from $(\mathbf{I}_m)_k$: morphologically open $(\mathbf{I}_m)_k$ by first replacing the value of a given pixel with the smallest pixel value found within a predefined neighborhood of that pixel (erosion), and then on the resulting image, replacing the value of a given pixel with the largest pixel value found in the same neighborhood (dilation). The resulting image can then be closed by reversing the process—namely, first performing a dilation and then an erosion operation. The opening operation will remove small objects, while the closing operation will remove noise and smoothen transitions. We decimate the resulting image to obtain our image for the next level of the pyramid, $(\mathbf{I}_m)_{k+1}$. We obtain the pyramid coefficients of level $k+1$ as the difference between $(\mathbf{I}_m)_k$ and an up-sampled and dilated version of $(\mathbf{I}_m)_{k+1}$. While these morphological operations may produce good-looking results, as shown in figure 11 (right), they are quite computationally intensive in nature, and their usefulness in enhancing tracking performance is not necessarily great.

Finally, many specialized pyramid decompositions, such as contourlets and wavelets, separate an image into approximations and detail. We examined a simple discrete wavelet transform (DWT) using the Daubechies Symmetric Spline wavelet, as well as a Shift Invariant Discrete Wavelet Transform (SIDWT) using the Harr wavelet. The DWT is applied to an input image using two filters, $g_1 = [-2 \ 4 \ -2]$ and $h_1 = [-1 \ 2 \ 6 \ 2 \ -1]$. In this case, g_1 is a high-pass filter and h_1 a low-pass filter. These filters are applied to the columns and rows of an image consecutively in one of these four combinations: $g_1' * g_1$, $g_1' * h_1$, $h_1' * g_1$, and $h_1' * h_1$. The output of $g_1' * g_1$ is the high frequency content of the image, while the output of $h_1' * h_1$ contains only low-pass one.

All four combinations of the outputs are then decimated by two to form four sub-band images. The resulting low-pass image is used for the next iteration of decomposition, while the maximum coefficients from the other three sets are stored in the wavelet tree. An example of the fused images produced by DWT pyramids is shown in figure 12 (top-left).

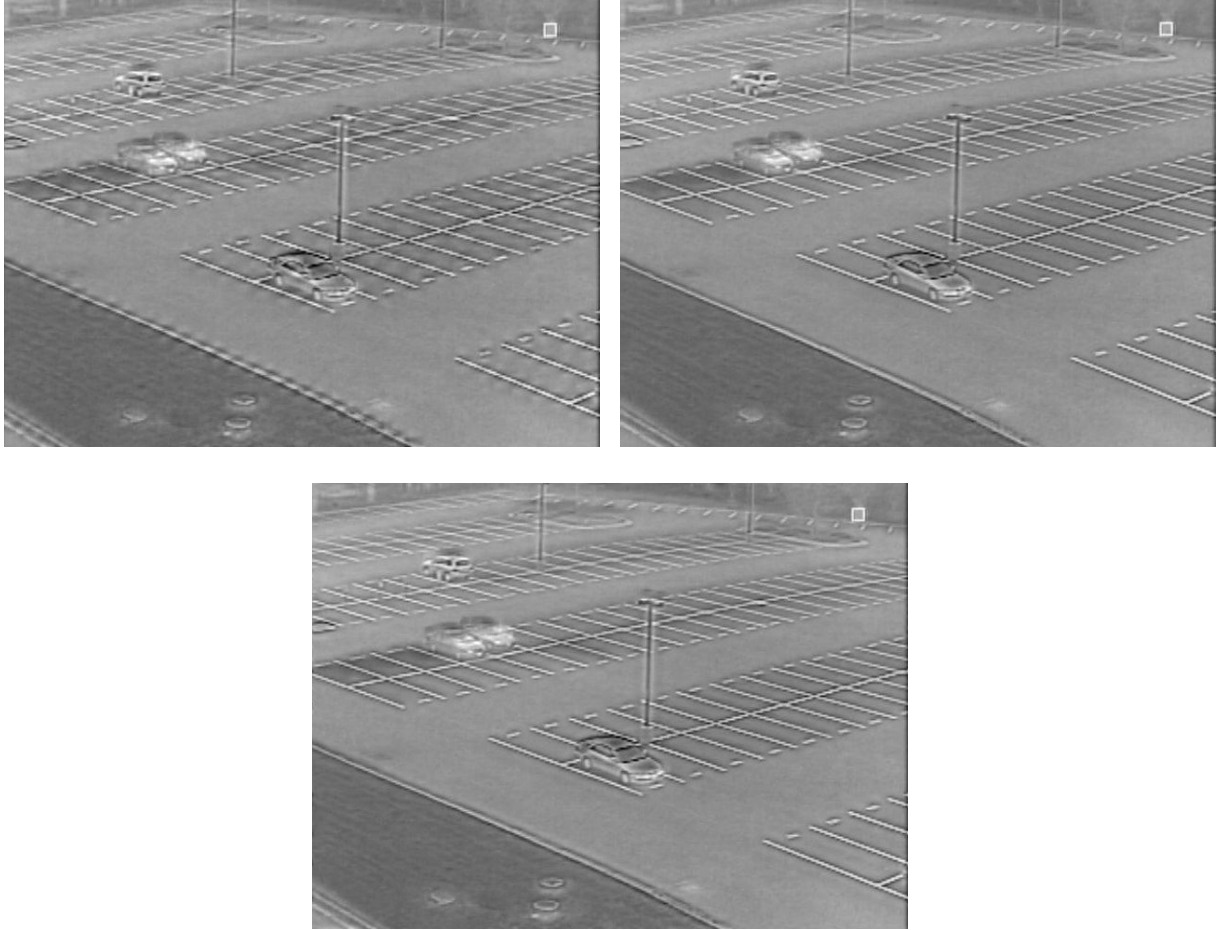


Figure 12. Fusion by selecting the maximum coefficient of DWT pyramids (top-left), SIDWT pyramids (top-right), and DT-CWT pyramids (bottom).

For SIDWT, the filters g_1 and h_1 are defined as $g_1 = [0 \ \dots \ 0 \ 0.5 \ 0 \ \dots \ 0 \ -0.5 \ 0 \ \dots \ 0]$ and $h_1 = [0 \ \dots \ 0 \ 0.5 \ 0 \ \dots \ 0 \ 0.5 \ 0 \ \dots \ 0]$, with $2^{(k-2)}$ zeroes in the first and last set of zeroes, and $2^{(k-1)}$ zeroes in the middle set of zeroes for level k of the pyramid. While the SIDWT is very redundant (because it up-samples the filter response instead of decimating the image at each level of the pyramid), the Dual-tree Complex Wavelet Transform (DT-CWT) can achieve approximate shift invariance and only slight oversampling by filtering the image with a pair of complementary filters. DT-CWT produces real and complex coefficients at each level of the decomposition for a total of 2^d oversampling, where d is the number of levels of decomposition. Figure 12 shows an example of the fused images produced by SIDWT pyramids (top-right) and DT-CWT (bottom), respectively.

The simple DWT can be prone to artifacts as a function of position in the image, which could be particularly problematic when using the FPSS background subtraction tracker to detect motion information. As an object moves slightly, artifacts could shift in the image, resulting in many unnecessary false alarms. Hence, a SIDWT or DT-CWT is expected to perform better in a tracking task. Similar to other pyramid methods, we use the maximum coefficient from either wavelet tree at each level during the image fusion phase.

3. FPSS Tracker

The effects of different image fusion methods were examined and compared using an existing moving target tracking algorithm. Since the FPSS tracker has been developed and adequately tested with the original (non-fused) FPSS dataset, it was chosen for this evaluation work, as well. The FPSS tracker was run on the original color and LWIR images, as well as the fused images generated by all fusion methods described in section 2.

3.1 Background Modeling

The key component of the FPSS tracker is its background modeling and subtraction process, which is depicted in figure 13. Each input image is first filtered by a stability mask and then channeled through four image buffers of equal size and depth. The images in Buffers 2 and 4 are used to generate Background Models 1 and 2, respectively. Instead of being created originally from Buffer 4, Background Model 2 can also be obtained from a buffer of models that is continuously replenished by the outgoing representations of Background Model 1. By subtracting the next input frame from these background models, we obtain two difference images. A difference-product image (DPI) is obtained by multiplying these two difference images pixel by pixel.

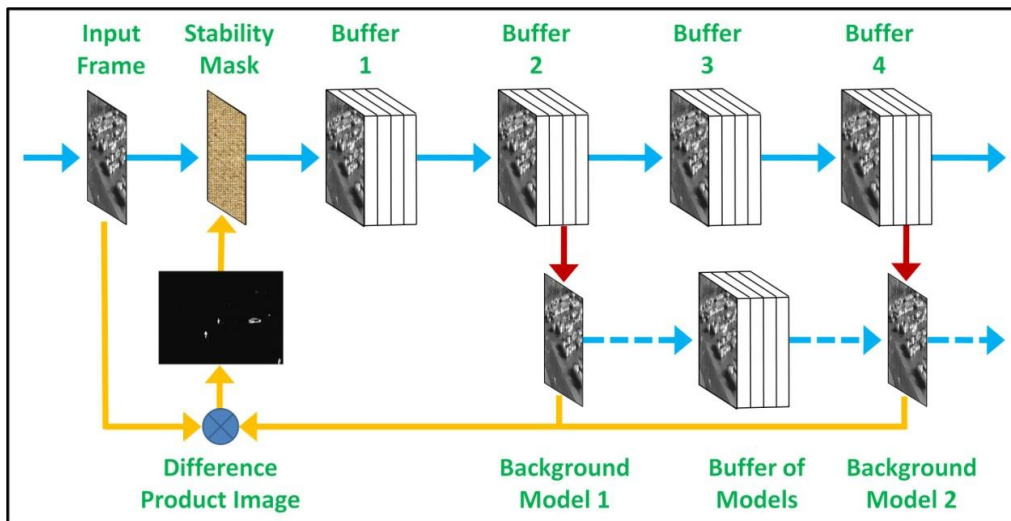


Figure 13. The background modeling and subtraction process in FPSS tracker.

To begin the background modeling process, the first successfully preprocessed input image frame is used to fill up all image buffers and to become the initial background models. For each of the subsequent input image frames, a simple frame registration procedure is used to reduce any potential jitter effects incurred by shaking cameras. Typically, a jitter-free image contains a mostly stable background with a number of small but volatile areas caused by moving objects and other transient events. In order to prevent rapidly changing foreground pixels from ruining the background models, a stability mask is used to filter out all unstable pixels from the input image frame. Updated by the information from DPI, this stability mask looks for significant intensity changes based on a predefined threshold of variability and maintains a record of the stability index at each pixel location. Only those stable pixels on a jitter-free image are fed to Buffer 1, while the once-stable but now active pixels are blocked and substituted by the corresponding stable pixels available from Buffer 1. Without the stable background models, it will be much harder to detect and extract legitimate moving objects in the scene, while additional false alarms will likely be generated.

Each incoming set of pixel values from the stability mask replaces the corresponding pixel values in the oldest frame in Buffer 1 to form the newest frame in Buffer 1, while the oldest frame of Buffer 1 becomes the newest frame in Buffer 2. The same mechanism of first-in first-out (FIFO) frame-shift and update is applied to all image buffers continuously. The role of Buffer 1 is merely a time-delay buffer to induce a noticeable gap in time—and potentially in content—between the current input image and the image frames in Buffer 2. Background Model 1 is derived from the images in Buffer 2, which can be as simple as taking the average of all images in Buffer 2. Similar to Buffer 1, Buffer 3 is just another buffer to separate Buffer 2 and Buffer 4 in time. Background Model 2 can be obtained by either processing (e.g., averaging) the images in Buffer 4 or drawing from the Buffer of Models supplied by Background Model 1. The same background modeling structure depicted in figure 13 can be extended to include four or any larger even number of background models for more stable background representations and higher target enhancement capabilities at the expense of additional computational resources.

One of the advantages of using multiple disjoint background models to generate a DPI is that the problematic “trailing effect,” which is often associated with background subtraction method, can be suppressed effectively; because those gradually fading trails carved out by the moving objects are now showing up in different parts of the two difference images, they are likely to diminish or disappear when the DPI is formed, as demonstrated in figure 14. For the same reason, time-dependent noises on the difference images are also suppressed during the formation of DPI. Another advantage of this method is that the target trails are now clearly detached from the moving objects, which allows the subsequent target detection module to estimate the size and location of those movers more accurately. With improved estimation in target size and location, the target tracking module may also perform better motion estimation and track maintenance.



Figure 14. Enhancement of target signatures and suppression of trailing effects and noises via a DPI.

An even number of background models is needed in the formation of DPI to address the problem of target polarity, which is a common target detection problem. Due to clothing and ambient temperature change, the same type of moving targets may assume different polarity of pixel intensity with respect to their immediate background. Figure 15 shows a pair of LWIR images that exhibit polarity change in human signatures during different seasons of the year. Using a single difference image or a DPI computed with any odd number of difference images to detect the moving targets will have to pick the locations with both positive and negative values simultaneously and appropriately, which is not always easy or straightforward. This problem is alleviated, however, simply as a by-product of forming the DPI using an even number of difference images.



Figure 15. Human LWIR signatures reverse polarity in winter (left) and summer (right).

3.2 Target Detection and Tracking

After a DPI is generated, a morphological operation is used to remove small spikes and to fill up small gaps in the DPI. Furthermore, a pyramid-means method is used to enhance the centroid and overall silhouette of the moving targets. The moving target detection process begins with finding the brightest pixel on the post-processed DPI, which is usually associated with the most probable moving target in the given input frame. The size of this target is estimated by finding all the surrounding pixels that are deemed connected to the brightest pixel. After the first moving target is detected, all the pixels within a rectangular target-sized area of that target are suppressed to exclude them from subsequent detections. The detection process is repeated by finding the next brightest one among the remaining pixels until all the pixels are suppressed, a predefined number of detections are obtained, or other user-defined stopping criteria are reached.

Using the detection results on consecutive input images, tracks of all moving targets are built and maintained. In order to build a meaningful track, a noticeable moving target must appear in multiple contiguous frames in a video sequence. This requirement may not be met when the target is moving across the field of view of the camera at a very short range and/or a very high speed; when the camera is operated at a very low frame rate; when the target is occluded for an extended period of time and/or behind a very large obstacle; or when a combination of these and other detrimental factors occur. The FPSS tracker uses previous locations, velocity, and target size of a moving target to predict the destination of its next movement.

The detection and tracking results can be reported via a graphical user interface (GUI). As shown in figure 16, the GUI of FPSS allows a user to enter or modify a number of parameters related to the file directories, input images, potential targets, tracking characteristics, background modeling, and jittery control. Furthermore, the user may define, activate, deactivate, and remove any “don’t care” zone, “critical” zone, and trip wires by using this GUI, as well. The detected moving or changed targets are annotated or highlighted over the input image frame for easy understanding. The tag number, location on the image, size in pixels, and activation strength of all detected targets on each input frame are displayed at the bottom-right corner of this GUI.

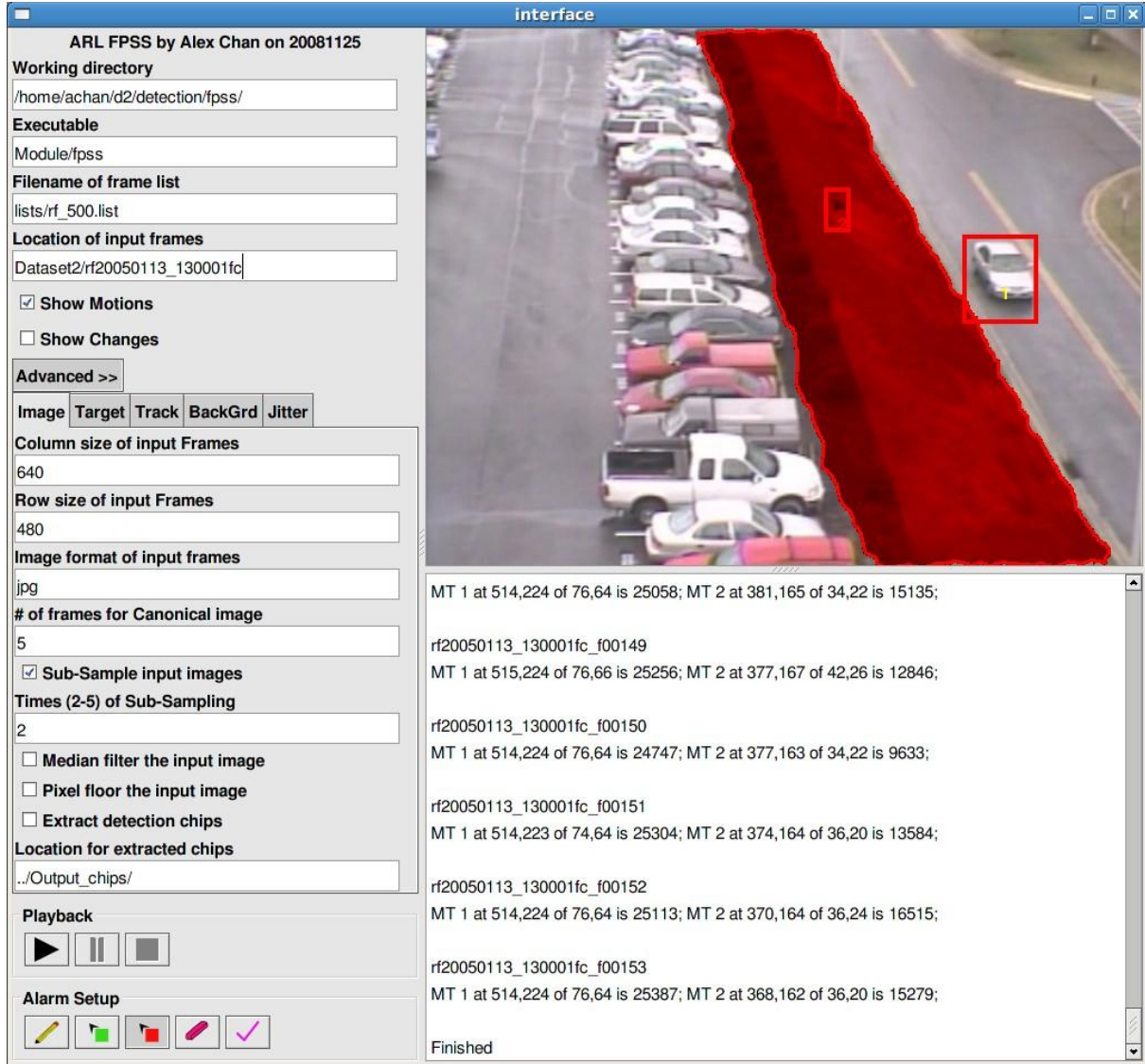


Figure 16. The FPSS graphical user interface.

4. Experimental Results

The Second FPSS dataset consists of 53 short video sequences for a total of 71,236 frames, which depict various staged suspicious activities around a big parking lot. Ground truth information (target type and target location) associated with each observable moving target on each image frame was semi-manually generated using a ground-truthing GUI. Figure 17 shows the partial content of a typical ground truth file in the FPSS dataset. Based on the ground-truth information and the target size estimated by the FPSS tracker, we may compute the tracking performance achievable by the original color and LWIR sequences, as well as the performances pertaining to the fused image sequences generated by different image fusion methods.

```

Frame: 0
People: 2 at x= 259 y= 278 x= 238 y= 105
Vehicles: 1 at x= 129 y= 54
Animals: 0
Others: 0
Unknowns: 0

Frame: 1
People: 2 at x= 256 y= 278 x= 236 y= 105
Vehicles: 1 at x= 127 y= 54
Animals: 0
Others: 0
Unknowns: 0

Frame: 2
People: 2 at x= 253 y= 277 x= 237 y= 104
Vehicles: 1 at x= 127 y= 54
Animals: 0
Others: 0
Unknowns: 0

```

Figure 17. Partial content of a typical ground truth file in the FPSS dataset.

The ground truth files associated with a concurrent pair of color-LWIR sequences may vary slightly in their content, as some moving targets may sometimes be observable in one but not both of the imageries. For example, a man walking in a dark area at night can be noticeable in the LWIR sequence, but is completely obscured in the corresponding color sequence. Because we used the LWIR approximation coefficients during the pyramid decompositions, and because LWIR ground truth files usually contain more information on the targets, we chose the LWIR ground-truths files for the purpose of verifying the detections on fused images.

To be qualified for a correct detection or a hit, the ground-truth location must be included in the bounding box (target size) estimated by the FPSS tracker for the given detection. Multiple detections on the same target were counted as only one hit, but no additional penalty was imposed in this situation. Multiple detections on a non-target, however, were treated as multiple false alarms (FAs), which would decrease the tracking performance. When multiple targets in proximity were covered by a single detection, it would be treated as multiple hits and would boost the tracking performance. Ground-truth targets that were not included by the bounding box of any detection were regarded as misses that would hurt the tracking performance.

An adjustable acceptance threshold was used to vary the tradeoff between hits and FA. By gradually lowering the acceptance threshold, the number of hits and the number of FA would both increase monotonically. By plotting the FA rate (FAR) (average number of incorrect detections per frame) against the hit rate (percentage of true targets that were correctly detected) at different acceptance thresholds, a receiver operating characteristic (ROC) curve results. Values closest to the upper-left corner of an ROC curve are the best results, indicating high accuracy with few FAs. In order to emphasize the critical differences between the ROC curves, we focused on the two end zones of these curves in order to examine the performance at low

FAR and at hit rates exceeding 80%. The ROC curves for the original LWIR and color sequences were first generated. As shown in figure 18, these two ROC curves serve as the benchmark performance curves and are included in all performance-related figures for comparison purposes.

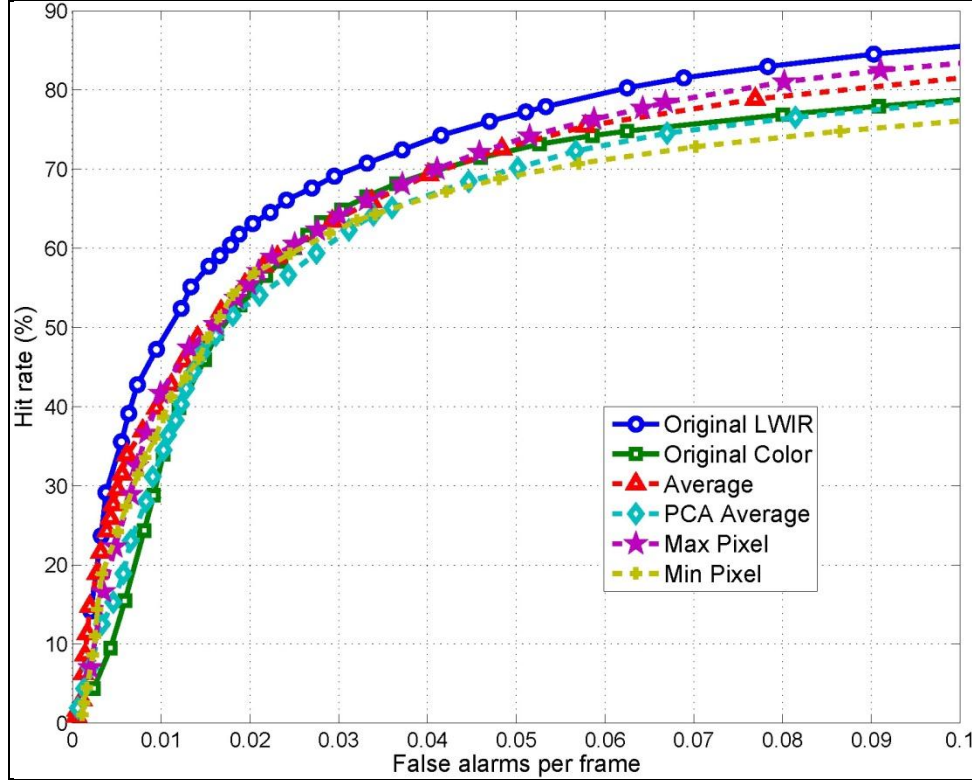


Figure 18. The performance of four simple-combination methods at low FAR region.

Figures 18 and 19 also show the ROC curves associated with the fused images generated by the four simple-combination methods: simple averaging, PCA-weighted averaging, maximum pixel selection, and minimum pixel selection. Their performances at low FAR region are shown in figure 18, while figure 19 shows their performance as more false alarms are allowed. From figure 18, it is clear that the original LWIR images performed the best with a low FAR among this group of six candidates. On the other hand, the original color images were lagging behind their LWIR counterparts consistently due to a significant increase in the number of FAs caused by headlight glares and windshield reflections in the evening hours, and protracted shadows under the slanted sun. As shown in figure 19, the advantage of LWIR sequences over color sequences continues to hold at high FAR region.

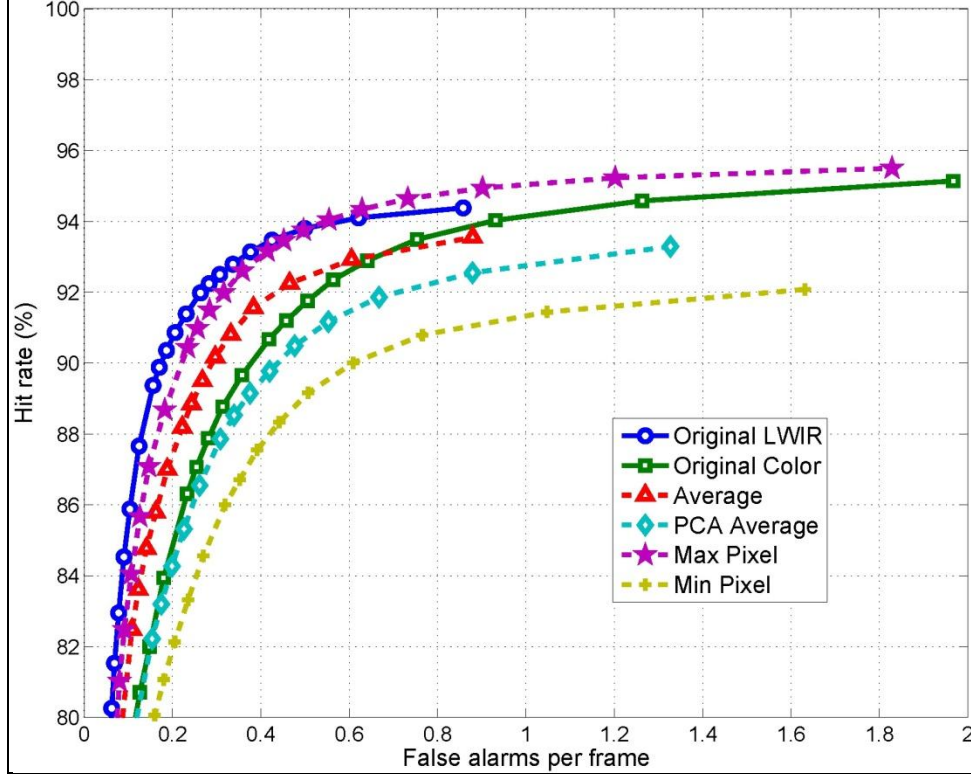


Figure 19. The performance of four simple-combination methods at high FAR region.

Given the nature of simply averaging or selecting the pixels of the original color and LWIR images by the four simple-combination methods, it was expected that their resulting fused images would perform somewhere between the original color and LWIR images. As evident from the far-left region of figure 18, this was, indeed, the case for the FAR region of 0.02 or less FAs per frame. As the allowable number of FAs was increased by lowering the acceptance threshold, as shown in figures 18 and 19, the fused images produced by simple averaging and maximum pixel selection methods continued to yield hit rates that were between those produced by the original color and LWIR images. The performance associated with the fused images generated by the PCA-weighted averaging and minimum pixel selection methods, however, gradually fell below the performance of the original color images. In other words, there was no performance gain in tracking at any FAR by using the images fused with simple combination methods over the original LWIR images. At FARs higher than 0.02 FA per frame, even the original color images outperformed the fused images produced by the PCA-weighted averaging and minimum pixel methods.

The fusion methods based on pyramid structures were performed using an identical set of configuration parameters, such as using five levels of decomposition and a 7×7 neighborhood size when running a saliency/match measure. Based on their resulting ROC curves, these pyramid-based fusion methods were categorized into two groups for subsequent discussions: four inferior methods and five superior methods. As shown in table 1, all nine pyramid-based

methods are much more computationally intensive than the four simple combination methods, especially the SIDWT, gradient, and morphological pyramids. Although the DT-CWT is more than four times more efficient than its more redundant variant, SIDWT, it is still considerably slower than the five simpler pyramid-based methods, three of which are ranked together in the superior pyramid column. More computations do not always generate better results, and as we can see, among the pyramid-based methods there are faster and slower candidates in both the inferior and superior columns of table 1.

Table 1. CPU time (seconds) needed to fuse 30 images using Matlab code on a Dell T7400 workstation.

Simple combinations	CPU time	Inferior pyramids	CPU time	Superior pyramids	CPU time
Simple average	1.280	FSD	21.670	Laplacian	24.040
PCA average	2.030	Gradient	78.970	ROLP	23.050
Maximum pixel	1.560	DWT	22.740	Contrast	23.240
Minimum pixel	1.840	Morphological	62.530	SIDWT	209.600
				DT-CWT	49.940

As shown in figure 20, the FSD, gradient, and DWT achieved slightly worse performance than the original LWIR images at low FARs, whereas the morphological pyramid method clearly lagged behind others under the same conditions. The picture is somewhat different at the other end of these ROC curves, as shown in figure 21, where the DWT and morphological pyramid methods were able to surpass the LWIR curve at the FAR region of 0.7 FA per frame or higher. Since alternative pyramid-based methods offer more consistent gains over the complete range of FAR, we deem these four pyramid-based methods inferior.

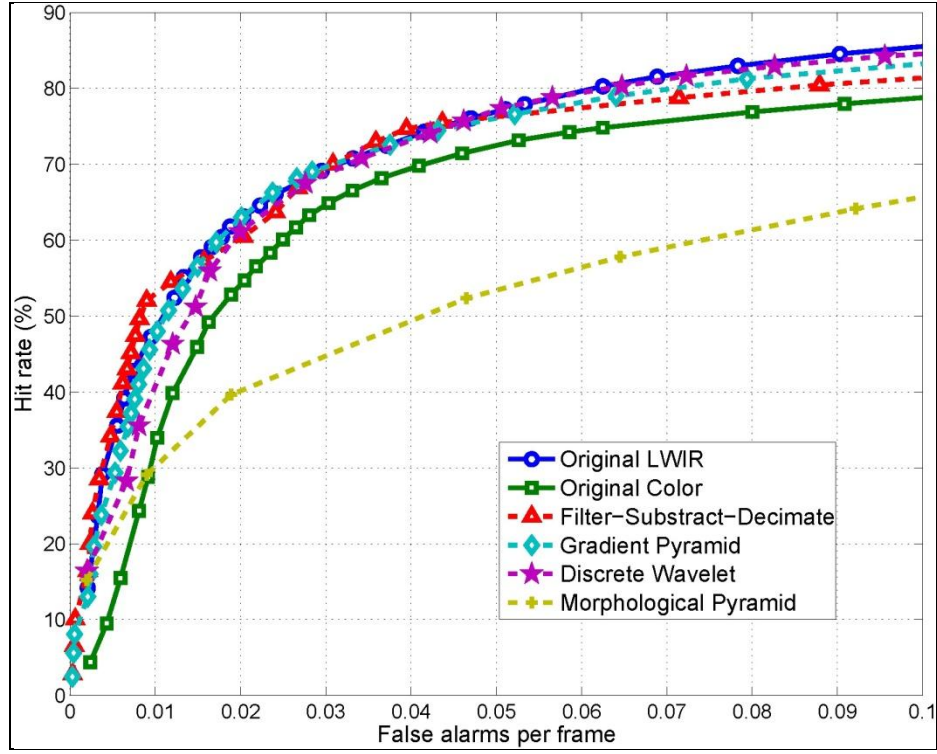


Figure 20. The performance of four inferior pyramid-based fusion methods at low FAR region.

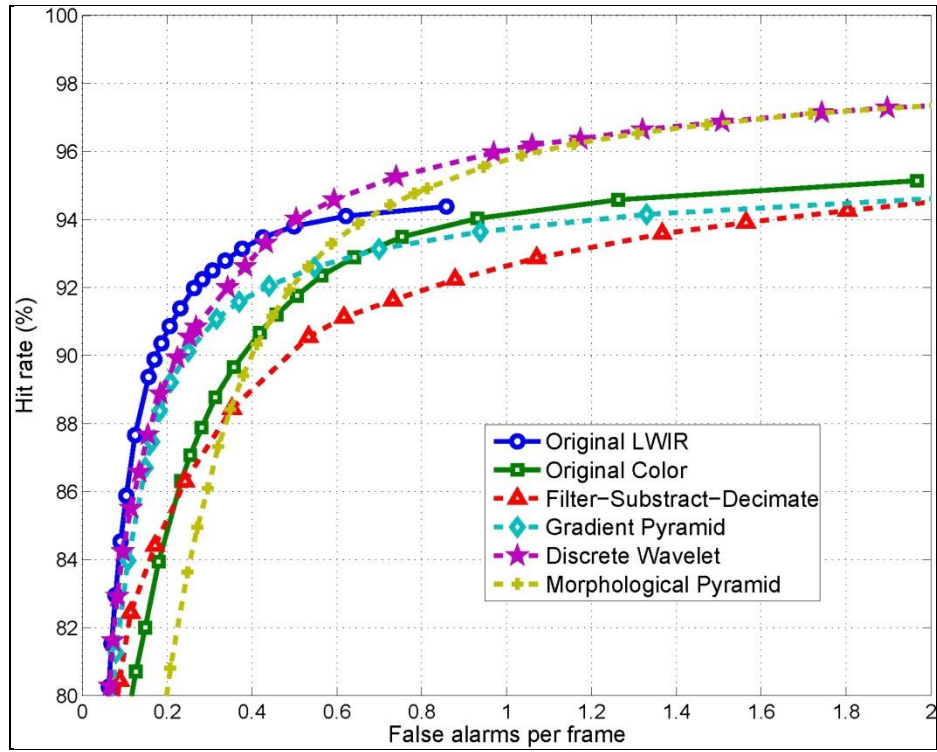


Figure 21. The performance of four inferior pyramid-based fusion methods at high FAR region.

Finally, there are five pyramid-based fusion methods that have achieved good results on both ends of the ROC curves: the Laplacian, ROLP, contrast, SIDWT, and DT-CWT pyramid methods. As shown in figure 22, these five fusion methods clearly outperformed the original color and LWIR images from the beginning and attained the largest advantage at the FAR of around 0.02 FA per frame. At this FAR, the hit rates for the original color and LWIR images are 54.29% and 62.99%, respectively. As shown in table 2, the corresponding hit rates of the images fused by contrast pyramid and ROLP pyramid methods are 76.94% and 75.11%, respectively. With improvements of 12–14% over the LWIR images, the performance gains achieved by these two fusion methods are quite remarkable at this FAR.

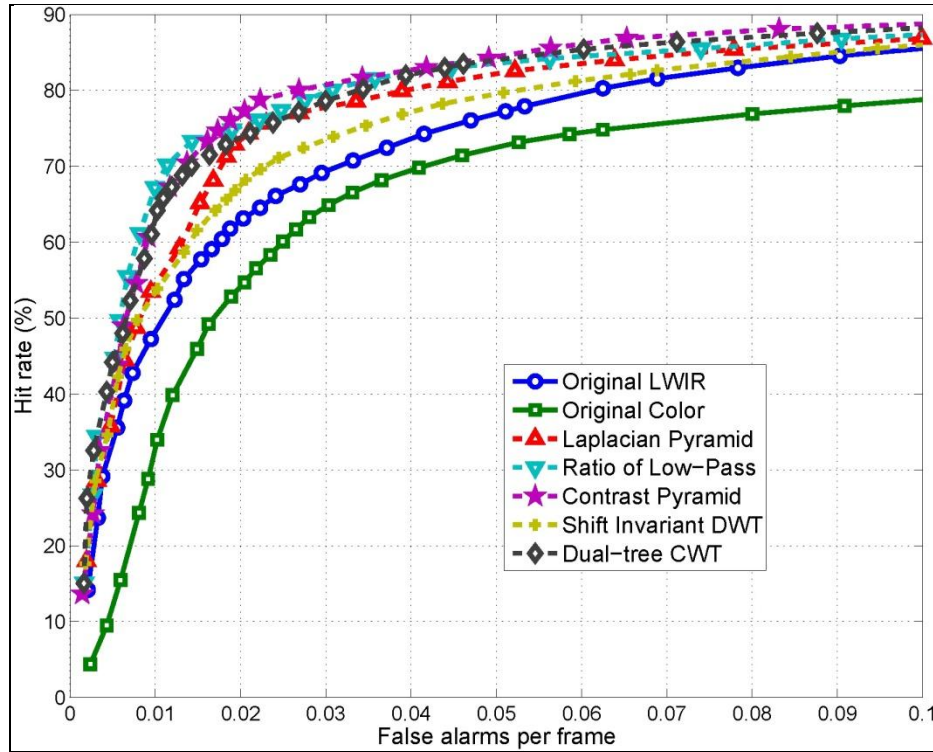


Figure 22. The performance of five superior pyramid-based fusion methods at low FAR region.

Table 2. Performance (hit rate in %/FA per frame) of the 13 fusion methods at low FAR region.

Simple combinations	HR/FAR	Inferior pyramids	HR/FAR	Superior pyramids	HR/FAR
Simple average	56.14/0.02005	FSD	60.34/0.02008	Laplacian	73.51/0.02008
PCA average	53.43/0.02008	Gradient	62.90/0.02005	ROLP	75.11/0.02005
Maximum pixel	55.71/0.02008	DWT	61.47/0.02008	Contrast	76.94/0.02005
Minimum pixel	56.17/0.02008	Morphological	40.27/0.02008	SIDWT	67.53/0.02005
				DT-CWT	73.80/0.02002

Among the five superior pyramid-based methods, SIDWT is clearly lagging behind other methods in performance. Furthermore, the computational complexity of SIDWT is about nine times of that of the contrast pyramid and ROLP pyramid methods. Therefore, SIDWT is the least desirable method among this group. Although the performance of DT-CWT is competitive to those of the contrast, ROLP, and Laplacian methods, it requires more than twice as much CPU time to complete the same image fusion task.

The performance of these superior fusion methods at high FAR region is shown in figure 23 and in table 3. As evident from figure 23, the advantage of these methods over the original color and LWIR images is still maintained at every point in the high FAR region, even though the performance gain is less significant than that in low FAR region. As shown in table 3, the hit rates of the images fused by contrast pyramid and ROLP pyramid methods at a FAR of 0.80 FA per frame are 95.37% and 95.37%, respectively, exceeding those of color (93.66%) and LWIR (94.31%) images by slightly more than 1%.

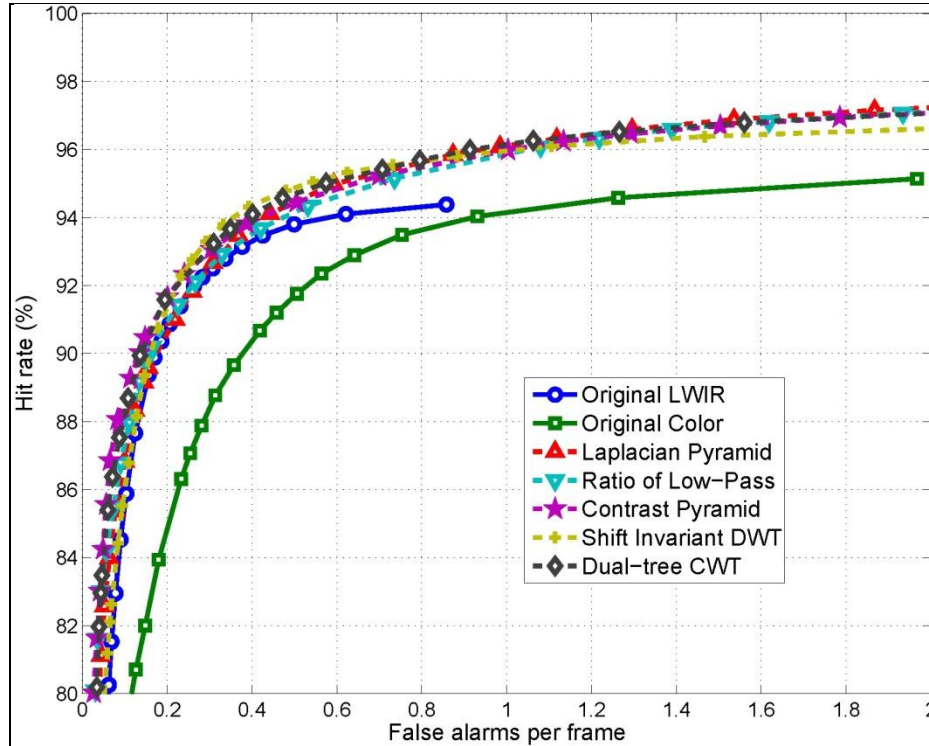


Figure 23. The performance of five superior pyramid-based fusion methods at high FAR region.

Table 3. Performance (hit rate in % / FA per frame) of the 13 fusion methods at high FAR region.

Simple combinations	HR/FAR	Inferior pyramids	HR/FAR	Superior pyramids	HR/FAR
Simple average	93.41/0.80183	FSD	91.93/0.80070	Laplacian	95.69/0.80087
PCA average	92.37/0.80025	Gradient	93.38/0.80343	ROLP	95.37/0.80040
Maximum pixel	94.77/0.80146	DWT	95.48/0.80023	Contrast	95.52/0.80048
Minimum pixel	90.92/0.80343	Morphological	94.88/0.80020	SIDWT	95.66/0.80068
				DT-CWT	95.69/0.80138

5. Conclusions

The sensor fusion community believes that a person may easily fool a sensor sometimes, but nobody may easily fool all the sensors simultaneously at a given time. For this reason, we explored and exploited the rather complementary natures of two common imaging sensors: LWIR and color visible sensors. Instead of harnessing prior background knowledge and external information sources (such as metadata on weather conditions, time of the day, season of the year, site characteristics, number of targets, target ranges, depression angle, speed of movement, and other related information) to perform symbolic level image fusion, we focused on pixel-level image fusion. Therefore, the techniques examined and the results obtained in this work are more readily transferrable to other applications and scenarios that process color and LWIR imageries.

Based on the results generated by the four simple-combination methods examined in this work, we conclude that these simple methods are not useful, because their performances were worse than using the original LWIR images alone. Among the nine pyramid-based image fusion methods, the gradient and FSD methods are the worst candidates, because they required 10–60 times more CPU time than those required by the simple combination methods, but performed even worse at high FAR region. The morphological and DWT methods are slightly better than the gradient and FSD methods, primarily because they managed to outperform LWIR in the high FAR region. Given their performances and computational requirements, these four pyramid-based methods are deemed as inferior methods in general.

The Laplacian, ROLP, contrast, SIDWT, and DT-CWT are found to be superior image fusion methods, because they consistently outperformed LWIR in every FAR region. Contrast and ROLP methods are considered the best image fusion methods to pair with the FPSS tracker because their ROC curves are consistently on top of all other ROC curves produced in this work. Furthermore, the computational requirements of these two methods are almost the lowest among the pyramid-based methods. On the other hand, SIDWT is ranked at the bottom in this group, as

it performed the worst and consumed 4–9 times more CPU time than its counterparts in this group did.

For future work, a potential way of improving image fusion performance is to treat each color image as 3 separate images (R, G, and B images) and fuse these three images with the LWIR image together. The fusion algorithms examined in this work do not limit the number of images that can be fused together. Therefore, short-wave infrared, mid-wave infrared, and hyperspectral imageries could also be considered, if they are properly co-registered. Performance may also be improved by linking the image fusion process with the tracking algorithm, through which the information that is critical to the tracker may be better preserved or enhanced. For instance, a region-based segmentation algorithm may be incorporated into the DT-CWT image fusion process (21, 22). The segmentation algorithm could exploit the limited redundancy in DT-CWT and tie the feature level and pixel level fusion algorithms together. Using a more robust tracking algorithm—perhaps the flux tensors algorithm—may also enhance the image segmentation process and the decision rule (23).

6. References

1. Chan, A. L. *A Description on the Second Dataset of the U.S. Army Research Laboratory Force Protection Surveillance System*; ARL-MR-0670; U.S. Army Research Laboratory: Adelphi, MD, 2007.
2. Hines, G.; Rahman, Z.; Jobson, D.; Woodell, G. Multi-image Registration for an Enhanced Vision System. *Proc SPIE Visual Information Processing* **Aug 2003**, 5108, 231–241.
3. Smith, M.; Heather, J. Review of Image Fusion Technology in 2005. *Proc. SPIE Thermosense* **Mar 2005**, 5782, 29–45.
4. Han, J.; Bhanu, B. Fusion of Color and Infrared Video for Moving Human Detection. *Pattern Recognition* **2007**, 40, 1771–1784.
5. Motwani, M.; Tirpankar, N.; Motwani, R.; Nicolescu, M.; Harris, F. Towards Benchmarking of Video Motion Tracking Algorithms. *Int. Conf. Signal Acquisition and Processing* **2010**, 215–219.
6. Cvejic, N.; Nikolov, S. G.; Knowles, H. D.; Loza, A.; Achim, A.; Bull, D. R.; Canagarajah, C. N. The Effect of Pixel-Level Fusion on Object Tracking in Multi-Sensor Surveillance Video. *IEEE Conf. Computer Vision and Pattern Recognition* **2007**, 372, 1–7.
7. Mihaylova, L.; Loza, A.; Nikolov, S. G.; Lewis, J. J. The Influence of Multi-Sensor Video Fusion on Object Tracking Using a Particle Filter. *Proc. 2nd Workshop on Multiple Sensor Data Fusion* **2006**, 354–358.
8. Chan, A. L. A Robust Target Tracking Algorithm for FLIR Imagery. *Proc. SPIE Automatic Target Recognition* **May 2010**, 7696, 1–11.
9. Trucco, E.; Plakas, K. Video Tracking: A Concise Survey. *IEEE Journal of Oceanic Engineering* **2006**, 31, 520–529.
10. Lowe, D. G. Distinctive Image Features from Scale-invariant Keypoints. *Int. Journal of Computer Vision* **2004**, 60 (2), 91–110.
11. Tsagaris, V.; Anastassopoulos, V. Fusion of Visible and Infrared Imagery for Night Color Vision. *Displays* **2005**, 26, 191–196.
12. Chan, A. L.; Der, S. Z.; Nasrabadi, N. M. Dualband FLIR Fusion for Automatic Target Recognition. *Information Fusion* **2003**, 4, 35–45.
13. Rockinger O. MATLAB Image Fusion Toolbox. <http://www.metapix.de/indexp.htm>, 1999 (accessed 2010).

14. Cai, S.; Li, K. Matlab Implementation of Wavelet Transforms.
<http://taco.poly.edu/WaveletSoftware/references.html> (accessed 2010).
15. Burt, P. J.; Adelson, E. H. The Laplacian Pyramid as a Compact Image Code. *IEEE Trans. Communications* **1983**, 31, 532–540.
16. Anderson, H. A Filter-subtract-decimate Hierarchical Pyramid Signal Analyzing and Synthesizing Technique. U.S. Patent 718 104, 1987.
17. Toet, A. Image Fusion by a Ratio of Low-pass Pyramid. *Pattern Recognition Letters* **1996**, 9, 245–253.
18. Toet, A.; van Ruyven, J. J.; Valetton, J. M. Merging Thermal and Visual Images by a Contrast Pyramid. *Optical Engineering* **1989**, 28 (7), 789–792.
19. Burt, P. A Gradient Pyramid Basis for Pattern Selective Image Fusion. *Society for Information Displays (SID) Int. Symp. Digest of Technical Papers* **1992**, 23, 467–470.
20. Ramac, L. C.; Uner, M. K.; Varshney, P. K. Morphological Filters and Wavelet Based Image Fusion for Concealed Weapon Detection. *Proc. SPIE Sensor Fusion* **1998**, 3376, 110–119.
21. Lewis, J. J.; O’Callaghan, R. J.; Nikolov, S. G.; Bull, D. R.; Canagarajah, C. N. Region-based Image Fusion Using Complex Wavelets. *Proc. Int. Conf. Information Fusion*, Stockholm, 2004.
22. Lewis, J. J.; O’Callaghan, R. J.; Nikolov, S. G.; Bull, D. R.; Canagarajah, N. Pixel- and Region-based Image Fusion with Complex Wavelets. *Information Fusion* **2007**, 8, 119–130.
23. Bunyak, F.; Palaniappan, K.; Nath, S. K. Flux Tensor Constrained Geodesic Active Contours with Sensor Fusion for Persistent Object Tracking. *Journal of Multimedia* **2007**, 2 (4), 20–33.

List of Symbols, Abbreviations, and Acronyms

DPI	difference-product image
DT-CWT	Dual-tree Complex Wavelet Transform
DWT	discrete wavelet transform
FA	false alarm
FAR	false alarm rate
FIFO	first-in first-out
FLIR	forward-looking infrared radar
FPA	focal plane array
FPSS	Force Protection Surveillance System
FSD	filter-subtract-decimate
GUI	graphical user interface
IR	infrared
LWIR	long-wave infrared
NTSC	National Television Standards Committee
PCA	principal component analysis
ROC	receiver operating characteristic
ROLP	ratio-of-low-pass
SIDWT	Shift Invariant Discrete Wavelet Transform
SIFT	scale-invariant feature transform
SNR	signal-to-noise ratio
SPOD	Sentry Personnel Observation Device

NO. OF COPIES	ORGANIZATION
1 ELECT	ADMNSTR DEFNS TECHL INFO CTR ATTN DTIC OCP 8725 JOHN J KINGMAN RD STE 0944 FT BELVOIR VA 22060-6218
1 CD	OFC OF THE SECY OF DEFNS ATTN ODDRE (R&AT) THE PENTAGON WASHINGTON DC 20301-3080
1	US ARMY TRADOC BATTLE LAB INTEGRATION & TECHL DIRCTRT ATTN ATCH B 10 WHISTLER LANE FT MONROE VA 23651-5850
1	US GOVERNMENT PRINT OFF DEPOSITORY RECEIVING SECTION ATTN MAIL STOP IDAD J TATE 732 NORTH CAPITOL ST NW WASHINGTON DC 20402
4	CECOM NVESD ATTN L GRACEFFO ATTN M GROENERT BLDG 305 ATTN J HILGER ATTN C WALTERS BLDG 307 10221 BURBECK RD FT BELVOIR VA 22060-5806
3	COMMANDER US ARMY RDECOM ATTN AMSRD AMR J MILLS ATTN AMSRD AMR K DOBSON ATTN AMSRD AMR W MCCORKLE 5400 FOWLER RD REDSTONE ARSENAL AL 35898-5000

NO. OF COPIES	ORGANIZATION
1 HC 1 CD	DIRECTOR US ARMY RSRCH LAB ATTN RDRL ROI C L DAI ATTN RDRL ROI M J LAVERY (1 CD) PO BOX 12211 RESEARCH TRIANGLE PARK NC 27709-2211
34 HCS 1 CD	US ARMY RSRCH LAB ATTN IMNE ALC HRR MAIL & RECORDS MGMT ATTN RDRL CIO LL TECHL LIB ATTN RDRL CIO MT TECHL PUB ATTN RDRL D J MILLER ATTN RDRL D J CHANG ATTN RDRL SE P AMIRTHARAJ ATTN RDRL SE J RATCHES ATTN RDRL SES J EICKE ATTN RDRL SES M D'ONOFRIO ATTN RDRL SES E R RAO ATTN RDRL SES E N NASRABADI ATTN RDRL SES E A CHAN (20 HCS, 1 CD) ATTN RDRL SES E H KWON ATTN RDRL SES E S HU ATTN RDRL SES E S YOUNG ADELPHI MD 20783-1197
TOTAL:	48 (44 HCS, 3 CDS, 1 ELECT)

INTENTIONALLY LEFT BLANK.